



LEUPHANA
UNIVERSITÄT LÜNEBURG

Influencing Belief in LLM-Based Agent Networks:
An Empirically Validated Simulation of
Bot-Driven Manipulation

Master Thesis

in the degree Program Management & Data Science (M.Sc.)

Leuphana University Lüneburg

submitted by

Julian Burmester

First Examiner: Prof. Dr. Ricardo Usbeck

Second Examiner: Kai Moltzen, M.Sc.

Submission Date: 3 January 2026

Abstract

Large language models enable generative social simulations in which agents interact through natural language rather than predefined behavioural rules, extending classical agent-based modelling toward language-mediated interaction. Prior work has implemented stylised opinion dynamics models and platform-like social simulations augmented with LLM agents. Yet, belief updating is often driven by unconstrained prompting, heuristic rules, or agent self-reports, leaving it unclear whether observed dynamics reflect human persuasion or artefacts of modelling and prompting choices. This thesis asks whether belief change induced by language-capable social bots can be analysed under empirically grounded conditions, and how bot narrative styles shape population-level belief dynamics in a Twitter-like environment. To address this, a Twitter-like generative social simulation, PersuaRealSim¹, is developed in which human-like generative agents and specialised bot agents interact. Stance updating is externalised to a supervised persuasion judge implemented as a RankFormer model trained on 46,846 *r/ChangeMyView* threads containing human-verified belief change outcomes. The resulting persuasion scores are calibrated to plausible stance-shift magnitudes and injected into a continuous stance update mechanism to ground belief change at the message level. Across two misinformation-relevant domains, simulations with four distinct bot narrative styles reveal rapid early belief change, followed by asymptotic convergence, consistent with assimilative influence. Narrative styles primarily affect the speed and efficiency of early shifts. Scientific-authority bots induce the strongest and fastest shifts, while emotional framing underperforms control conditions, although statistical resolution is limited by the number of simulation runs feasible under computational constraints. Beyond substantive findings, the thesis contributes methodologically by demonstrating how external, purpose-aligned validation can be integrated directly into the causal mechanism of a generative social simulation. Overall, the work shows that generative agents can support the principled study of belief change when empirical grounding is embedded into the simulation mechanism itself, and conclusions are drawn within clearly stated methodological bounds.

¹ Code and data available at <https://github.com/JulianBurmester/PersuaRealSim>

Contents

List of Figures	V
List of Tables	VII
List of abbreviations	VIII
1 Introduction	1
1.1 Research Questions	2
2 Theoretical Foundations of Generative Social Simulation	3
2.1 Large Language Models	3
2.2 Prompt Engineering	6
2.3 Agent-based Models in the Social Sciences	9
2.4 Generative Agents	11
2.4.1 Architectural Scaffold of Generative Agents	12
2.4.2 Core Agent Modules	12
2.5 New Paradigm of Generative Social Simulation	16
2.6 Advantages of Generative Social Simulation	17
2.7 Challenges and Limitations of Generative Social Simulation	19
2.8 Validation Principles for Generative Social Simulation	22
2.8.1 Purpose-Aligned and Externally Grounded Validation	23
2.8.2 Micro-Level Agent Design	23
2.8.3 Macro-Level Experimental Design	24
3 Theoretical Foundations of Information Disorder	26
3.1 Types of Information Disorder	26
3.2 Phases of Information Disorder	27
3.3 Agents as First Element of Information Disorder	28
3.3.1 Social Bots as Automated Agents in Information Disorder	29
3.3.2 Evolution into LLM-Powered Social Bots	30
3.4 Messages as the Second Element of Information Disorder	32
3.4.1 Topics of Misinformation	33
3.4.2 Narrative Styles of Misinformation	35
3.5 Interpreters as the Third Element of Information Disorder	37
3.5.1 Opinion Dynamics as System-level Models of Interpretation	38
3.5.2 Individual-level Interpretation and Susceptibility to Influence	42
4 Related Work	44
4.1 Stylised LLM-Based Opinion Dynamics Models	44
4.2 LLM-Based Social Media Platform Simulations	45
4.3 Task-Oriented Generative Social Simulations	46
4.4 LLM-Powered Bot and Influence Models	47
4.5 Agent-Based Models of Social Bots and Opinion Manipulation	48
4.6 Summary and Differentiation from Prior Work	49
5 Method	51
5.1 Methodological Positioning	52
5.1.1 Generative Social Simulation and Validation Constraints	52
5.1.2 Instantiation of information disorder	53
5.2 Persuasion Judge: Training and Calibration	55
5.2.1 r/ChangeMyView as an Empirical Source of Persuasion Signals	55

5.2.2	Training Data Construction	56
5.2.3	RankFormer Modelling	59
5.2.4	Hyperparameter Optimization	61
5.2.5	γ -Calibration	63
5.3	Human Agent Architecture	64
5.3.1	Profile Construction via Silicon Sampling	64
5.3.2	Memory and Reflection	67
5.3.3	Action	69
5.4	Bot Agent Architecture	72
5.4.1	Structural Differences to Human Agents	72
5.4.2	Narrative Styles in Bot Messaging	73
5.5	Simulation Environment and Orchestration	75
5.5.1	Initialisation of the Social World	77
5.5.2	Perception and Timeline Construction	79
5.5.3	Stance Updates via the Persuasion Judge	80
5.6	Simulation Configuration	84
6	Evaluation	86
6.1	Empirical Realism of the Persuasion Judge	86
6.1.1	Quantitative Evaluation	86
6.1.2	Qualitative Error Analysis	88
6.1.3	Gamma Calibration Results	92
6.2	Population-Level Belief Change Under Different Narrative Styles	93
6.2.1	Individual Stance Trajectories	93
6.2.2	Robustness of Population Mean Stance Trajectories	99
6.2.3	Temporal Evolution of Stance Dispersion	101
6.2.4	Generalisation to the Climate Change Topic	101
6.3	Efficiency of Narrative Styles in Achieving Stance Change	104
6.3.1	Definition of Efficiency Metrics	105
6.3.2	Endpoint Efficiency	106
6.3.3	AUC Efficiency	106
6.4	Simulation Validity Checks	107
6.4.1	Micro-Level Behavioural Fidelity of Agent Actions	107
6.4.2	Macro-Level Structural Regularities of the Simulation	112
7	Discussion	121
7.1	Summary of Results	121
7.2	Interpretation of Results	122
7.2.1	Interpretation of RQ1.1: Empirical Realism of the Persuasion Judge	122
7.2.2	Interpretation of RQ1.2: Convergence and Assimilative Belief Dynamics	123
7.2.3	Interpretation of RQ1.3: Efficiency, Statistical Power, and Narrative Effects	124
7.2.4	Methodological Implications for Generative Social Simulation	125
7.3	Limitations	127
7.4	Ethical Considerations	130
7.5	Future Work	131
7.6	Conclusion	132
	References	X
	List of appendices	XXXVI

List of Figures

1	Classical ABM agents vs. generative agents.	11
2	Core modules of a generative agent.	13
3	Venn diagram of types of information disorder.	27
4	Taxonomy of opinion dynamics models.	39
5	High-level architecture of the PersuaRealSim simulation environment.	51
6	Schematic example of a CMV thread with an awarded delta.	55
7	CMV thread structure and listwise target construction	56
8	Structural and textual characteristics of the CMV dataset.	58
9	RankFormer architecture.	60
11	Overview of the silicon sampling pipeline for persona generation.	66
10	Example of a silicon-sampled persona text built from ANES ID 234862.	66
12	Agent memory buffers and prompt injection.	68
13	Reflection prompt structure.	69
14	Action-decision prompt and execution pipeline.	70
15	Style guidelines for emotional–moral framing bots.	74
16	Style guidelines for post-truth populist narration bots.	74
17	Style guidelines for pseudo-scientific authority bots.	75
18	Style guidelines for repetitive factual-sounding (echo-claim) bots.	75
19	Example timeline card for an original post.	80
20	Example timeline card for a reply, referencing its parent message.	80
21	Schematic overview of stance initialisation and updating.	83
22	Seed event introducing the topic ”accelerating the phase-out of fossil fuels”.	84
23	Seed event introducing the topic ”tightening national vaccination requirements”.	84
24	Stage-I RankFormer hyperparameter heatmap at $\tau = 10$	87
25	Treemap overview of clustered high-confidence RankFormer errors.	89
26	Structural and textual characteristics of severe RankFormer error cases.	92
27	Distributions of empirical and RankFormer-implied belief shifts for γ -calibration.	93
28	Individual stance trajectories for vaccination under scientific-authority bots.	94
29	Individual stance trajectories for vaccination under emotional bots.	95
30	Individual stance trajectories for vaccination under post-truth populist bots.	96
31	Individual stance trajectories for vaccination under echo-claim bots.	97
32	Individual stance trajectories for vaccination under the fixed-stance control.	98
33	Individual stance trajectories for vaccination under the movable-stance control.	99
34	Population mean stance trajectories for vaccination under bot narrative styles.	100
35	Population mean stance trajectories for vaccination under control conditions.	100
36	Population-level stance dispersion for vaccination over time	101
37	Individual stance trajectories for climate change under scientific-authority bots.	102
38	Individual stance trajectories for climate change under the movable-stance control.	103
39	Population mean stance trajectories for climate change across random seeds.	104
40	Population-level stance dispersion under scientific-authority bots.	104

41	Example of tweet generation grounded in memory and stance.	108
42	Example of follow action triggered by timeline exposure.	109
43	Example of unfollow action under stance divergence.	109
44	Example of emotional bot behaviour.	110
45	Example of post-truth populist bot behaviour.	110
46	Example of scientific-authority bot behaviour.	111
47	Example of echo-claim bot behaviour.	111
48	Follower network evolution example.	114
49	Human action dynamics over time.	116
50	Action targeting by timeline position.	117
51	Action choice by prompt option position.	118
52	Action composition by posting propensity.	120
53	Ideal-typical opinion dynamics regimes.	124

List of Tables

1	Analytical dimensions for characterising agents in information disorder.	29
2	Analytical dimensions for characterising messages in information disorder.	33
3	Conceptual landscape of generative social simulation.	52
4	Two-stage hyperparameter sweep design for RankFormer training.	62
5	Action space and required fields for action decisions.	71
6	Mapping between discrete stance labels and continuous stance values.	78
7	Experimental dimensions and resulting simulation run count.	85
8	Test-set performance of the final RankFormer candidates.	88
9	Mean directional endpoint efficiency by narrative style and topic.	106
10	Mean directional AUC efficiency by narrative style and topic.	106
11	Action-decision and memory-reflection warning frequencies.	113
12	World-state rejection breakdown.	114
13	Follower network growth summary.	115
14	Follow and unfollow actions by agent role.	115
15	Timeline composition by follow status.	118
16	Agent-level action count heterogeneity.	119

List of abbreviations

ABM	Agent-Based Model
ANES	American National Election Studies
BCE	Binary Cross-Entropy
CMV	ChangeMyView
DW	Deffuant–Weisbuch
ELM	Elaboration Likelihood Model
LLM	Large Language Model
NLP	Natural Language Processing
OP	Original Poster
PPSWOR	Probability-Proportional-to-Size Sampling Without Replacement
Webis-CMV-20	Webis ChangeMyView Corpus 2020

1 Introduction

Large language models (LLMs) enable a new class of simulation approaches in computational social science by allowing agents to communicate in natural language and respond flexibly to context. This extends classical agent-based modelling, which relies on predefined behavioural rules, toward populations of language-capable agents embedded in interactive environments. The resulting paradigm, commonly referred to as generative social simulation, studies collective social dynamics that emerge from open-ended linguistic interaction under structured conditions (T. Guo et al. 2024; Bail 2024).

At the same time, these advances also change the conditions of online discourse in ways that matter directly for belief formation. Social bots have long existed as automated accounts that create and distribute content on social media platforms (Ferrara, Varol, et al. 2016; Cresci 2020), yet LLM-powered social bots narrow the behavioural gap to humans by producing fluent, context-aware messages and sustaining coherent interactions over time (Lopez-Joya et al. 2025; Yizhou Zhang et al. 2024; Radivojevic et al. 2024). Social bots have become active participants in online discourse across multiple high-stakes contexts, including US presidential elections (Bessi and Ferrara 2016; Ferrara, Chang, et al. 2020), the COVID-19 pandemic (Ferrara 2020; Suarez-Lledo and Alvarez-Galvez 2022), climate change debates (C.-F. Chen et al. 2021; L. Li et al. 2024), and the Russia–Ukraine war (W. Xu et al. 2025; Marigliano et al. 2024), among others. Bots can take part in argumentative exchanges, vary narrative framing, and engage strategically with users, which makes language-capable automated actors an increasingly relevant component of contemporary social media dynamics.

These developments connect to broader concerns about information disorder and democratic vulnerability (Wardle and Derakhshan 2017). Social platforms facilitate the circulation of misleading, manipulative, or strategically framed content through algorithmically curated feeds and engagement-driven visibility that shape how political information is encountered, interpreted, and emotionally engaged with (Morse and Pratt 2025; Monsees 2023). These dynamics place democratic processes under strain by eroding shared baselines of trust and deliberation and by enabling coordinated manipulation within attention-based platform architectures (Deb et al. 2017). LLM-powered social bots intensify these dynamics by lowering the barrier to creating automated accounts and enabling sustained engagement at scale, within an online environment where automated traffic now exceeds human traffic (Imperva 2025). At a more fundamental level, generative AI poses a challenge to epistemic agency (Rosenberg 2023).

Despite this urgency, existing approaches struggle to analyse these dynamics in a principled manner. Empirical studies based on observational data from social media platforms provide ecological validity but are constrained by ethical considerations and limited experimental control (Salganik 2018). Classical agent-based models allow controlled experimentation, yet rely on simplified communication and hand-crafted behavioural rules that cannot represent the linguistic richness and adaptive interaction characteristic of contemporary online discourse (Epstein 1999). Generative social simulation promises to bridge this gap by combining scalable experimentation with language-capable generative agents (Anthis et al. 2025). However, its scientific credibility hinges on whether belief change within such simulations can be meaningfully validated against human behaviour. Without explicit empirical grounding, simulated persuasion risks reflecting prompt artefacts, model biases, or design choices rather than the social mechanisms it aims to study. (Larooij and Törnberg 2025b; J. Zhou et al. 2025; Münker et al. 2025; Wu et al. 2025).

This thesis addresses this methodological gap by developing a generative social simulation, PersuaRealSim, in which generative agents operate as humans and social bots within a Twitter-like environment and belief updating is governed by a supervised persuasion judge trained on human-verified instances of belief change. By anchoring stance updates in large-scale persuasion data from the *r/ChangeMyView* subreddit, empirical grounding is embedded directly into the simulation mechanism. Adopting the perspective of an adversarial influencing actor, the analysis examines which bot narrative styles are most effective at manipulating belief dynamics among human-like LLM agents and how efficiently such influence can be exerted over time. The aim is not to reproduce real-world platforms in full detail, but to assess whether belief dynamics driven by language-capable social bots can be studied under controlled assumptions in a methodologically defensible way.

1.1 Research Questions

Against this background, the thesis examines whether belief change induced by language-capable social bots can be studied in a generative social simulation under empirically grounded conditions. The central interest is not only whether different narrative styles exert influence, but whether such effects can be analysed in a way that remains methodologically defensible, given the known limitations of generative social simulation. To this end, the investigation is structured around one overarching research question, decomposed into three sub-questions that address validation, belief dynamics, and influence efficiency in sequence.

The overarching research question guiding this thesis is:

RQ1 Can belief change induced by language-capable social bots be analysed in a generative social simulation under empirically grounded belief-updating mechanisms, and how do different narrative styles shape these dynamics?

To address this question, the analysis proceeds in three steps:

RQ1.1 To what extent does a persuasion judge trained on *r/ChangeMyView* data reproduce human-verified persuasion outcomes?

RQ1.2 How do different bot narrative styles (emotional, scientific-authority, post-truth populist, echo-claim) shape belief change at the population level when belief updates are governed by this persuasion signal?

RQ1.3 How efficiently do these narrative styles induce belief change over time relative to a fixed-stance control condition?

Together, these questions establish empirical grounding as a prerequisite for simulation-based analysis and then examine how narrative framing affects both the direction and efficiency of belief change under controlled conditions.

2 Theoretical Foundations of Generative Social Simulation

This section introduces three complementary theoretical foundations underlying generative social simulation. First, it outlines large language models as the behavioural core of generative agents, with particular attention to prompting as the practical interface through which simulations specify perception, decision making, and action. Second, it revisits classical agent-based models (ABMs) in the social sciences, since generative social simulation continues this tradition but replaces handcrafted micro rules with language-model-driven behaviour embedded in an execution scaffold. Third, it situates these foundations in the current methodological debate. Recent work highlights both the promise of generative social simulation and its structural limitations, and it increasingly converges on emerging validation principles and best practices that aim to distinguish meaningful emergent dynamics from artefacts of prompting, model bias, or experimental design.

2.1 Large Language Models

Large language models are transformer-based systems trained to predict the next token in a text sequence. This simple objective produces a flexible generative mechanism that, when combined with prompting, serves as the behavioural core of generative agents in social simulation.

Transformer architecture Since the introduction of the transformer architecture by Vaswani et al. (2017), LLMs have become a central component of modern natural language processing (NLP). Transformers operate on text by first breaking it into discrete units, known as tokens, and then representing each token in relation to the surrounding text. Their central mechanism, self-attention, allows the model to evaluate all tokens in an input sequence at once and to determine how strongly each token should influence the interpretation of the others. This parallel processing was a major departure from earlier recurrent architectures that advanced through a sequence step by step and, therefore, limited model depth and training efficiency.

Conceptually, the transformer comprises two components: an encoder that constructs contextual representations from an input sequence and a decoder that generates text autoregressively. In an autoregressive decoder, the model computes a probability distribution over the vocabulary for the next token conditioned on all previously generated tokens, selects a token according to this distribution, appends it to the sequence and repeats the process. In the original formulation, both components can be combined to transform one sequence into another, for example, by encoding a sentence in one language and decoding it into a translation in another language. (Vaswani et al. 2017) The most prominent application of the transformer today, however, is in LLMs that use only the decoder component (Radford et al. 2019; Brown et al. 2020). These decoder-only transformers generate open-ended text by repeatedly predicting the next token from the preceding context, which at inference time consists of the text supplied by the user or calling application.

Autoregressive language modelling LLMs implement an autoregressive statistical model over token sequences. Formally, a language model defines a conditional probability distribution over the next token

in a sequence,

$$P_{\theta}(x_{1:T}) = \prod_{t=1}^T P_{\theta}(x_t | x_{<t}),$$

where $x_{<t}$ denotes all preceding tokens. Training follows a self-supervised objective in which the model is optimised to assign high probability to the next token of a text sequence drawn from a large corpus. The standard loss is the negative log likelihood of the correct token under the model distribution,

$$\mathcal{L}(\theta) = - \sum_{t=1}^T \log P_{\theta}(x_t | x_{<t}).$$

At inference time, the model computes a probability distribution over the vocabulary for the next token given the preceding context. Only a single token must be selected from this distribution to proceed with the generation. The most direct strategy is greedy decoding, which chooses the token with the highest conditional probability,

$$x_t = \arg \max_w P_{\theta}(w | x_{<t}),$$

although alternative sampling-based methods are often used to introduce variability (Holtzman et al. 2020; Meister et al. 2020). After selecting a token, it is appended to the sequence, and the process repeats. This mechanism allows decoder-based transformers to generate coherent text by repeatedly applying the same conditional prediction rule. (Jurafsky and Martin 2025)

Self-supervised pretraining A central property of LLMs is that their pretraining objective is entirely self-supervised. For each position in a text sequence, the correct next token is simply the token that follows in the raw text, so the training data provides its own labels without requiring manual annotation. This allows any large collection of unlabelled text to serve as supervision for estimating conditional token probabilities. Modern language models are therefore pretrained on heterogeneous mixtures of web crawls, digitised books, encyclopaedic material and other publicly available sources (Raffel et al. 2020; L. Gao et al. 2020). The breadth and scale of these corpora enable the model to internalise general linguistic structure as well as broad world knowledge in statistical form, long before any task-specific adaptation is introduced. Pretraining is computationally intensive and constitutes the stage in which the model acquires its foundational linguistic and semantic competence. (Brown et al. 2020)

In-context learning and scaling behaviour Large-scale self-supervised pretraining has enabled a shift from specialised, task-specific models toward general-purpose foundation models. A single pretrained transformer provides a broad representation space that can be adapted to downstream applications through comparatively small amounts of supervised fine-tuning, where the model is further trained on labelled examples tailored to a specific domain or task (Gururangan et al. 2020). Fine-tuning reorients the model’s behaviour toward the requirements of the target task. In many settings, large decoder models can also adapt without any parameter updates at all: they perform in-context learning by inferring task structure from instructions and examples provided directly at inference time (Brown et al. 2020).

Large decoder models display strong forms of in-context learning. They can infer task structure from examples provided directly at inference time and apply the inferred pattern to new inputs. This mecha-

nism supports different modes of adaptation. In zero-shot settings, the model receives only a natural language instruction and must perform the task without seeing any examples. In one-shot settings, a single input–output pair is provided as a template, and in few-shot settings, a small number of such demonstrations specify the task. In all cases, the model adapts its behaviour without updating its parameters, relying solely on the contextual information supplied in the prompt. (Brown et al. 2020)

At a conceptual level, in-context learning can be viewed as the model performing implicit Bayesian inference (Xie et al. 2021): it uses the prompt examples as evidence to infer the hidden “task rule” or concept. This inference is implemented mechanically in the transformer’s architecture through specialised attention patterns called induction heads (Olsson et al. 2022), which learn to copy and complete patterns found in the context. Together, this allows the model to adapt its behaviour instantly within a single forward pass, without any parameter updates.

In-context learning, therefore, operates as a form of meta-learning performed during inference rather than during training. Model scale plays a central role in these capabilities. Research on neural scaling laws shows that predictable improvements arise from increasing parameters, data and compute, and that sufficiently large models exhibit qualitatively new behaviours (Kaplan et al. 2020). The progression from GPT-2 to GPT-3 demonstrated that large decoder-only transformers can generalise from instructions and small sets of examples expressed in natural language (Radford et al. 2019; Brown et al. 2020). Empirical studies further indicate that certain capabilities, including multi-step reasoning and robust pattern induction, emerge only once models surpass specific scale thresholds (J. Wei, Tay, et al. 2022; Bubeck et al. 2023). At sufficiently large scale, next-token prediction supports flexible, prompt-driven adaptation without parameter updates (Dong et al. 2024).

Instruction tuning and alignment Beyond pretraining, LLMs undergo additional adaptation stages that improve their controllability and alignment with user intent. A first step is supervised instruction tuning, in which the model is trained on curated prompt–response pairs that demonstrate how it should interpret and follow natural language instructions (J. Wei, Bosma, et al. 2022). Unlike pretraining, which aims to model the statistical structure of text, instruction tuning does not introduce new factual knowledge but modifies how the model conditions on prompts. It shifts the behaviour of a pretrained decoder from generic text continuation toward producing task-directed answers to user queries. A subsequent stage aligns the model with human preference signals. Reinforcement learning from human feedback (RLHF) implements this by collecting comparative judgements over alternative model responses and optimising the model to favour outputs that humans rate as more helpful or appropriate (Ouyang et al. 2022). Instruction tuning and preference optimisation are orthogonal procedures: models may employ either or both stages, and each affects behaviour in distinct ways (Touvron et al. 2023). Together, these adaptations give LLMs a general prior toward cooperative, instruction-following behaviour, which enhances practical usability.

Behavioural limits and interpretive debates The combination of large-scale pretraining, flexible in-context adaptation, and post-training alignment enables language models to generate text that is often indistinguishable from human writing (E. Clark et al. 2021; Jakesch et al. 2023). They can sustain coherent dialogue (Jones and Bergen 2025), follow diverse instructions, and adapt to new tasks from a few

examples. These surface-level capabilities create the impression of an underlying cognitive process (Binz and Schulz 2023; Waytz et al. 2014), yet the mechanism remains a statistical next-token predictor trained on human text patterns (Bender et al. 2021). These behaviours raise questions about how to interpret model capabilities. LLMs can produce fluent but factually incorrect statements, called "hallucinations" (Sahoo, Meharia, et al. 2024), and may reproduce social biases from their training data (Navigli et al. 2023). One influential view characterises them as "stochastic parrots" that recombine statistical patterns without a grounded model of the world or any form of genuine understanding (Bender et al. 2021).

At the same time, certain model behaviours exhibit striking parallels to human cognitive patterns. LLMs employ fallible syntactic heuristics that fail systematically (McCoy et al. 2019), show recency-biased information access in long contexts (N. F. Liu et al. 2024), and can perform human-level social reasoning on theory-of-mind tasks (Strachan et al. 2024). These similarities need not indicate shared cognitive processes, but may instead reflect statistical traces of human communication learned during pretraining (Bender et al. 2021).

A different perspective, grounded in predictive-processing theories, suggests that systems minimising prediction error may exhibit fragments of mechanisms associated with human cognition (A. Clark 2013; Seth and Friston 2016). This has prompted debates about whether linguistic prediction alone could support minimal forms of machine consciousness, or whether such interpretations reflect anthropocentric biases that over-ascribe human-like traits to statistical models (Hussain et al. 2025).

These theoretical debates mirror broader questions about how to interpret LLM capabilities. As Mitchell (2023) cautions, subjective impressions of intelligence can be misleading, and apparent competence on benchmarks may not reflect robust understanding. The apparent human-likeness of model output thus requires careful, sceptical interpretation rather than anthropomorphic ascription.

2.2 Prompt Engineering

Prompt engineering involves designing input prompts to guide LLMs toward desired outputs without changing their parameters. By providing instructions, examples, and constraints in natural language, users can influence the model's attention, role, and response structure, effectively turning the prompt into a behavioural control interface. Various prompting strategies have been developed, from basic zero-shot and few-shot approaches to structured templates, reasoning scaffolds, and retrieval-augmented methods. These techniques effectively elicit latent task knowledge from pretrained models across diverse NLP applications (Sahoo, Singh, et al. 2025; Vatsal and Dubey 2024), establishing prompt design as a core adaptation mechanism at inference time.

System vs. user prompts Modern LLM interfaces separate prompts into a system part and a user part. The system prompt sets the general behaviour and constraints of the model and remains hidden during interaction, while the user prompt provides the task-specific query. Although most research focuses on user prompts, recent work shows that system prompt design also affects model behaviour and performance. (Neumann et al. 2025; Halil et al. 2025)

Structure and position effects Because the prompt functions as the model’s behavioural interface, both the content of the instructions and the way this content is organised inside the prompt shape output behaviour. Systematic analyses of zero-shot LLM-based rankers show that effectiveness can change substantially when varying prompt components such as role descriptions, tone words, task instructions, output format and the relative position of query and documents, in some settings even more than when changing the underlying ranking algorithm (S. Sun et al. 2025). Other experiments show that changing only the position of relevant information within the input produces a characteristic U-shaped performance curve, with models favouring the beginning and end of the context and degrading sharply in performance for a given task when relevant material appears in the central part of a prompt, where it is effectively “*lost in the middle*” (N. F. Liu et al. 2024). Related work on position engineering reinforces this picture by demonstrating that even when the prompt text is held constant, small shifts in where segments are located in the positional layout can reliably change downstream accuracy and reveal task-specific but unintuitive sweet spots for placing instructions, documents and queries (Z. He et al. 2024). Even small irrelevant sentences can induce large accuracy drops, especially when distractors share topic cues or contain in-range numbers (Shi et al. 2023). Taken together, this evidence suggests that the “*how*” of a prompt, its wording and internal structure, can be as important as the “*what*” in terms of task specification. As a consequence, effective prompt design requires careful choices not only about which information to include but also about how that information is phrased and organised within the prompt.

A further complication is that these sensitivities are strongly model-dependent. Work on few-shot prompt orderings shows that permutations yielding near state-of-the-art performance for one model can collapse to near-random accuracy for another, with no stable transfer of effective prompt structures across architectures (Y. Lu et al. 2022). A related line of work shows that even small variations in prompt template format can shift in-context performance from competitive to near-random levels, and that effective templates rarely transfer across models or prediction methods (Voronov et al. 2024). Ultimately, the utility of individual prompt strategies varies strongly across model families, suggesting that prompting behaviour is partly an emergent property of model training and architecture (M. Son et al. 2025).

Length and context limits Beyond structural arrangement, the overall amount of context supplied to a model also matters. LLMs often perform worse as prompts become longer, even when all relevant information is present and easy to retrieve (Hosseini et al. 2025; Y. Du et al. 2025). Work on few-shot prompting observes a similar over-prompting effect, where performance peaks at a moderate number of in-context examples and then degrades once too many demonstrations are added (Y. Tang et al. 2025). This aligns with the observation that even a single irrelevant sentence can disrupt predictions (Shi et al. 2023). Longer inputs simply provide more opportunities for distracting signals that models may follow. At the same time, the number of tasks included in a prompt does not appear to impose a similar penalty. Evaluations comparing single-task and multitask prompting show that combining several subtasks into one prompt does not reliably reduce performance and can even improve it for larger models (G. Son et al. 2024; Gozzi and Di Maio 2024). This suggests that it is not task breadth but sheer prompt length that more consistently limits LLM performance.

Numerical limitations LLMs show persistent weaknesses in basic numerical understanding. Even strong systems perform well only on simple and familiar number formats, while accuracy drops sharply for longer inputs and for less common representations such as fractions or scientific notation. Models also fail on elementary digit-level operations despite correctly interpreting the task instruction (H. Yang et al. 2025). These weaknesses extend beyond basic operations. Models perform well on simple numerical tasks, but their accuracy drops sharply once problems involve numbers in more complex settings (Rahman 2025). Experiments with state-of-the-art models, including GPT-4 and DeepSeek, show that these numerical weaknesses persist even at the largest model scales (H. Li et al. 2025).

Internal analyses show that models do encode numeric values in their hidden states along approximately linear directions, but these encodings are imprecise, non-continuous and become increasingly noisy as decimal precision grows (F. Zhu et al. 2025; Davies et al. 2025). Mechanistic work shows that models do not use a general arithmetic algorithm but rely on many narrow heuristics that react to specific operand patterns (Nikankin et al. 2025), which helps explain why their behaviour breaks down outside familiar formats or ranges. Few-shot numerical performance also depends strongly on pretraining frequency. Models are consistently much less accurate on equations that contain numbers that occur rarely in the pretraining corpus than on otherwise similar equations built from frequent numbers (Razeghi et al. 2022). Taken together, these results show that LLMs still lack a representation of numbers as continuous quantities and instead operate through token-level correlations, a limitation that constrains their performance on numerical retrieval, arithmetic and magnitude reasoning across tasks and model scales (H. Li et al. 2025). For prompt engineering, this implies that tasks involving magnitudes or continuous quantities should avoid relying on raw numerical inputs wherever possible and instead use categorical, discretised or linguistically framed representations that align more closely with the model’s token-based inductive biases.

Shortcut behaviour and mitigation strategies The limits in handling numbers described above mirror a broader issue: LLMs often fall back on simple shortcut cues in prompts instead of the intended task logic. LLMs often behave like lazy learners in in-context learning settings. Instead of inferring the underlying task rule from the provided examples, they tend to rely on superficial cues in the prompt, such as spurious correlations or positional regularities (R. Tang et al. 2023). A similar pattern appears across natural language understanding tasks, where model predictions frequently depend on lexical, positional or stylistic artifacts rather than on task-relevant semantics (M. Du et al. 2023). Small, targeted shortcut triggers can reliably mislead models, and even cause a model to fool itself with its own edited inputs (Levy et al. 2023). Even in simple few-shot classification settings, models may shortcut to copying the provided examples instead of learning the intended task pattern (Ali et al. 2024). These patterns reflect a broader form of non-robust model behaviour, in which models present competent output while relying on fragile shortcut cues rather than genuine task understanding.

One way to counteract some of the fragilities described above is to structure prompts so that models generate intermediate reasoning traces rather than producing answers directly. Chain-of-thought prompting provides explicit demonstrations of step-by-step reasoning and has been shown to elicit substantially more reliable multi-step problem solving across arithmetic, symbolic and commonsense tasks (J. Wei, X. Wang, et al. 2022). Even without demonstrations, simply appending a cue such as “Let’s think step by step” can induce models to articulate intermediate steps and significantly improve zero-shot accuracy on

a wide range of reasoning benchmarks (Kojima et al. 2022). More structured variants extend this idea by interleaving internal *“thought”* steps with environment-dependent actions. The ReAct framework demonstrates that such interleaved reasoning and acting not only produce more coherent task-solving trajectories but also help models avoid hallucinations and recover from errors when interacting with external tools (S. Yao et al. 2023). These approaches suggest that adding explicit thinking traces can reduce shortcut reliance and make model behaviour more stable and predictable.

Stronger models appear to rely less on prompt engineering for basic competence yet gain more from well-structured prompts when these are applied (M. Son et al. 2025). Prompt engineering is therefore not merely decoration around a model but a fine-grained design activity that functions like an additional software layer implemented in natural language. Recent work explicitly frames prompts as *“promptware”* that require their own lifecycle of requirements, design, testing and evolution (Z. Chen et al. 2025), introduces the notion of prompt programming to treat prompts as carefully crafted linguistic programs rather than ad hoc instructions (Reynolds and McDonell 2021), and proposes *“prompt smells”* as recurring patterns that signal a risk of undesirable or hallucinated behaviour, in analogy to code smells in traditional software engineering (Ronanki et al. 2024). These perspectives underline that in LLM-based systems, much of the behavioural logic is encoded at the prompt level and must be engineered with the same care as conventional source code.

2.3 Agent-based Models in the Social Sciences

Classical ABMs were introduced as a computational method for studying how macro-level societal patterns arise from micro-level interactions of countless individuals. ABMs describe a system as a collection of autonomous units, called agents, that perceive their environment, interact with other units and update their internal states over time. This enables researchers to explore social systems as decentralised and nonlinear processes rather than as aggregates of isolated variables, capturing forms of emergence that arise only through interaction dynamics (Holland 1992). This bottom-up perspective is designed to make the link between individual behaviour and collective outcomes explicit and to illustrate how complex societal patterns can arise from simple local rules, consistent with the broader idea in complexity science that collective outcomes may exhibit properties not directly inferable from individual components (Byrne and Callaghan 2022; Macy and Willer 2002).

In the social sciences, ABMs have been used to model a wide range of phenomena that involve distributed decision making and feedback between individuals and institutions, including the formation of social structures, cultural transmission, conflict and trade, and other macro-level patterns that are explained by *“growing”* them from simple micro-level rules (Epstein and Axtell 1996; Conte and Paolucci 2014). Early examples include Schelling’s segregation model, which shows how agents with only a mild preference for having some neighbours of their own group, and who move whenever this preference is not met, can nevertheless produce highly segregated residential patterns in a simple grid world (Schelling 1971). Later work extended the approach to domains such as the emergence of social norms (Epstein 2007), the diffusion of innovations (Deffuant, Huet, et al. 2005; Valente 1996), opinion dynamics (Castellano et al. 2009), financial markets (LeBaron 2006) and crowd behaviour (Luo et al. 2008). In these applications, agent-based simulation serves as a virtual laboratory in which researchers can vary initial conditions,

network structures or behavioural assumptions and observe how such changes influence system evolution.

Most ABMs in this tradition are built from a small set of recurring components. Agents represent individuals or organisations with attributes, bounded information and simple decision procedures that link perceptions to actions. An environment provides the spatial, institutional or informational context in which these agents are situated and from which they obtain local information. Interaction mechanisms specify how agents encounter one another, exchange information and resources, form or dissolve ties and respond to external events. (Macal and North 2005) Foundational work on intelligent agents emphasises that realistic agents should exhibit autonomy, social ability, responsiveness to their environment and goal-directed behaviour, properties that have been widely adopted in agent-based modelling (Wooldridge and Jennings 1995). When agents with these capabilities interact repeatedly in a shared environment, the model produces macro-level patterns that are not programmed explicitly but emerge from the underlying dynamics.

A further characteristic of classical ABM design is that agent behaviour is implemented through deterministic or stochastic rules formulated by the modeller. This rule-based specification is a defining feature of early ABMs, where behaviour is encoded as explicit condition–action rules or update equations (Macal and North 2005; Bonabeau 2002). These rules specify how agents update their states, make decisions or respond to others, often in the form of logical conditions, symbolic equations or probability distributions (Epstein and Axtell 1996). Even when randomness is included, the behavioural repertoire of an agent is fully determined by the predefined rule set, consistent with views of ABMs as mechanistic and fully specified micro-level processes (Epstein 2007). This rule-based architecture has the advantage of transparency and analytical tractability, as the internal logic of the model is explicitly specified, inspectable and reproducible (Windrum et al. 2007).

The transition from classical rule-based agents to LLM-driven generative agents marks a fundamental change in how agent behaviour is represented and computed. Figure 1 contrasts these paradigms by illustrating how classical agent-based models update explicit numerical state variables via predefined equations, whereas generative agents operate on natural-language representations mediated by a language model.

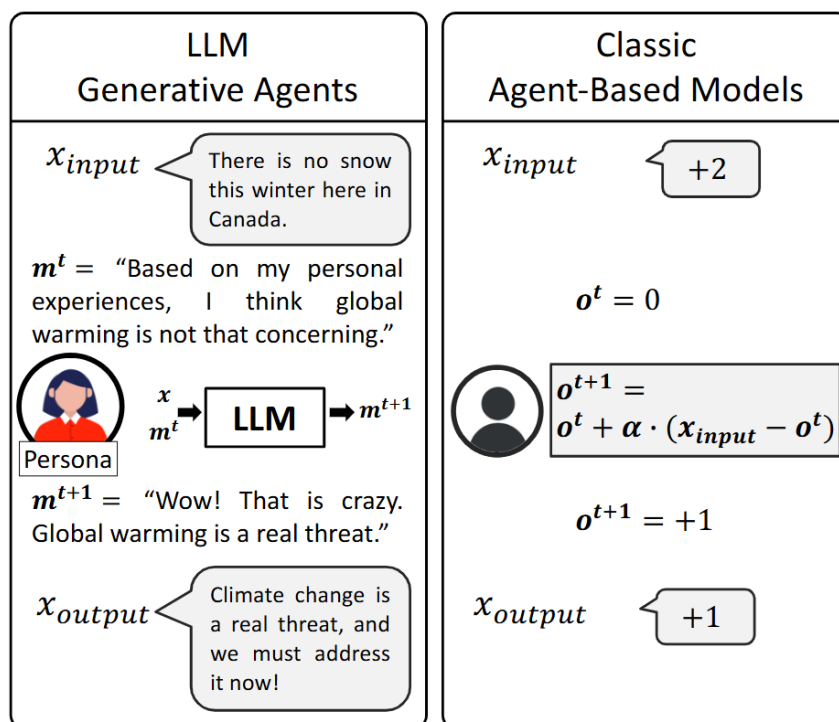


Figure 1: Conceptual comparison between classical agent-based models and LLM-based generative agents. In classical ABMs (right), agent behaviour is governed by explicit numerical state variables and predefined update equations. In contrast, generative agents (left) represent perceptions, internal states and actions primarily in natural language, with a large language model acting as a latent controller that maps textual inputs to textual outputs within a programmatic scaffold. The example content is illustrative; the figure highlights the representational and computational shift from symbolic or numerical update rules to language-mediated decision processes (Source: Chuang et al. 2024).

2.4 Generative Agents

Classical ABMs represent individuals through rule-based agents whose behaviour is specified by explicit update rules and decision procedures (Section 2.3). The LLMs introduced in Section 2.1 make it possible to upgrade these agents to a new level of behavioural richness and apparent human likeness. Generative agents wrap an LLM in a technical scaffold that handles perception and action. At each decision point, the scaffold constructs a prompt that summarises the agent’s internal state, memories, goals and recent observations of the environment, submits this prompt to the LLM, and interprets the resulting text as actions, utterances or internal updates. In this arrangement, the LLM acts as a latent controller of the agent, while the surrounding program logic determines what information is exposed, how outputs are mapped back into the simulation and how the agent’s state evolves.

A natural question is why LLMs are suitable as the core behavioural engine of such agents. Modern LLMs instantiate several properties long associated with intelligent agents (Xi et al. 2025). They display a degree of *autonomy* by generating context-appropriate actions without detailed stepwise instruction. They are *reactive* because they adjust their outputs to new perceptual inputs supplied through prompts or multimodal channels. They can also act *proactively*, as prompting methods such as chain-of-thought or task decomposition elicit goal-directed reasoning rather than simple stimulus–response behaviour. Crucially, they possess strong *social ability*, since natural-language interaction enables coordination, negotiation and communication with humans or other agents. Together, these capacities make LLMs effective latent

controllers within a generative-agent scaffold, where the surrounding program logic manages perception, memory and action, and the LLM provides a flexible decision engine shaped through prompting and mechanism design.

2.4.1 Architectural Scaffold of Generative Agents

The foundational design established by J. S. Park, O’Brien, et al. (2023) introduced a modular scaffold for generative agents. It consists of a profile that fixes an agent’s identity, a memory module that stores and retrieves past experiences, a reflection mechanism that produces higher-level summaries of memories, and a planning component that organises anticipated future actions. This design made explicit how an LLM can be embedded within a structured agent loop in which perception, memory and action are coordinated through program logic. Subsequent systems have largely adopted this modular structure, varying its implementation details, for example, by altering memory retrieval strategies, adding emotion or belief tracking, or integrating external tools, while retaining the same overall scaffold.

Building on this common scaffold, generative-agent systems differ in how responsibilities are divided between prompting and classical program logic. While all such systems rely on an LLM as the behavioural engine, some architectures keep most state tracking or memory management in conventional code, whereas others elicit these capabilities directly through prompting. Memory retrieval illustrates this spectrum well. At one end, systems let the LLM generate structured queries for symbolic memory operations (C. Hu et al. 2023). On the other end, retrieval is handled entirely outside the LLM through embedding-based similarity search (A. Zhao et al. 2024). Many recent systems vary these design choices further, reflecting different trade-offs between controllability, flexibility and computational cost.

At a higher level of abstraction, recent survey work systematises these developments into a unified architecture comprising a profiling module, a memory module, a planning module and an action module, and emphasises that profiling plays a foundational role in shaping all downstream components (L. Wang, Ma, et al. 2024). Within this architecture, profiling functions as the structural starting point, as it fixes the agent’s identity and conditions how memory, planning and action operate.

2.4.2 Core Agent Modules

Figure 2 summarises the modular scaffold shared by most generative-agent architectures and provides a structural overview of the four core components discussed below.

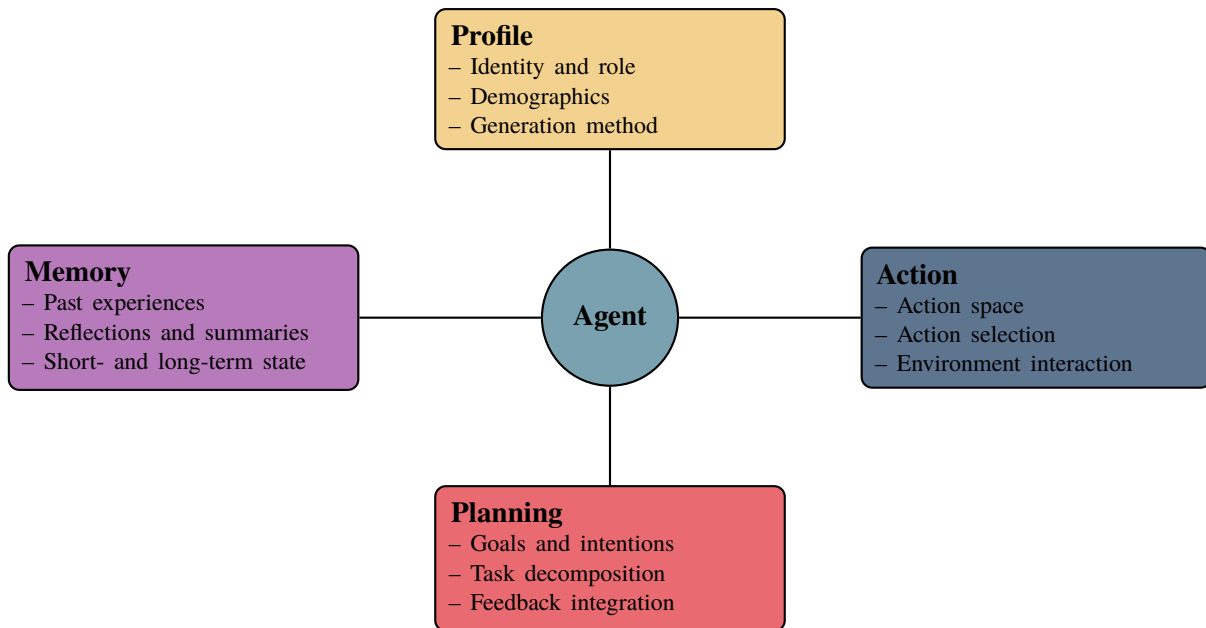


Figure 2: Abstract architectural scaffold of a generative agent. The agent is organised into four core modules: a profile that fixes identity and stable attributes, a memory module that accumulates and retrieves past experiences, a planning module that links goals to actions, and an action module that interfaces with the environment. The figure highlights the modular decomposition common to LLM-based generative agents, abstracting away from implementation-specific details (Source: adapted from L. Wang, Ma, et al. 2024).

Profile The profile module specifies who an agent is and which role it plays. It typically bundles demographic attributes, personality traits and social relationships into a concise description that is written into the prompt and thereby conditions all later behaviour. The exact choice of profile fields is application dependent. If psychological factors are central to the study, profiles may be augmented with Big-Five personality traits (Jiang et al. 2024). If communicative behaviour is the focus, profiles may instead include communicative roles that describe characteristic participation patterns in social environments (Mou et al. 2024).

Existing work realises profiles in three main ways (L. Wang, Ma, et al. 2024). The first and most straightforward approach uses *handcrafted profiles* to describe agents with manually designed attributes such as occupation, goals and ties to other agents (H. Zhang et al. 2024; J. S. Park, O’Brien, et al. 2023). The second possibility is *LLM-based profile generation*, where a language model produces a larger cast of characters from a set of explicit generation rules (G. Li et al. 2023). These rules define the desired attributes of the population, and they can optionally be accompanied by a set of seed profiles as few-shot examples (L. Wang, J. Zhang, et al. 2025). This approach is particularly useful for large populations, but can be hard to control in detail (L. Wang, Ma, et al. 2024). The third and most rigorous approach creates *dataset-aligned profiles* by deriving agent attributes from empirical data and translating them into natural language descriptions that preserve observed demographic distributions. This is also referred to as *silicon sampling*, where synthetic agents are sampled from real survey marginals and then rendered as personas (Argyle et al. 2023). A notable sub-variant appears in social-media simulations, where agents are aligned to real user accounts by letting an LLM infer demographic attributes or personal tendencies from profile text and posting histories (C. Gao, Lan, Z. Lu, et al. 2025; Goethals et al. 2025). This produces highly detailed personas and raises ethical concerns, because real individuals are effectively instantiated as agents inside

the simulation. These three approaches are not mutually exclusive. For example, dataset-aligned profiles can be complemented by handcrafted roles for a select few agents to represent counterfactual actors (Kera and Cavallon 2025). Likewise, a dataset-aligned profile can be further enriched by conditioning an LLM on it to generate additional life details such as a short background story, hobbies or plausible personal interests (Yijun Liu et al. 2025).

Memory The memory module allows agents to accumulate experiences and to condition future actions on past events. From a cognitive perspective, it provides the basis for learning and abstraction, while in agentic terms it underpins self-evolution, experience accumulation and consistent role behaviour over extended interactions (Z. Zhang et al. 2025). Most architectures distinguish between *short-term memory*, realised through the LLM context window, and *long-term memory*, which stores older yet relevant information in an external structure. In unified memory designs, agents rely only on short-term in-context memory, so that recent observations, internal thoughts and dialogue turns are written directly into the prompt. Hybrid designs add an explicit long-term store, for example, a vector database or key value store, which holds consolidated information that can be queried when needed and thereby supports longer range coherence. (L. Wang, Ma, et al. 2024)

Memory contents can originate from three sources, namely information inside the current trial, cross-trial experiences collected over many episodes, and external knowledge that is injected from outside the agent environment loop (Z. Zhang et al. 2025). This agent-level memory is conceptually separate from the parametric world knowledge already encoded in the LLM, which provides general background information but does not record the agent’s own experiences.

Beyond this structural and source-based view, memory modules differ in how information is represented. Some systems keep memories purely as natural language snippets, which preserve rich semantics and can be fed back into prompts verbatim (Shinn et al. 2023; G. Wang et al. 2023). Others encode memories as embeddings to enable efficient similarity-based retrieval (Zhong et al. 2024), store them as tuples or lists that capture explicit structure (Modarressi et al. 2024), or use databases as a symbolic memory that can be manipulated through queries such as SQL statements (C. Hu et al. 2023). These formats are not mutually exclusive and are often combined, for example, by using embedding vectors as keys and natural language as values (X. Zhu et al. 2023).

Across architectures, three core operations organise the interaction with memory, namely *writing*, *management* and *reading* (Z. Zhang et al. 2025). *Memory writing* decides which perceived events are stored and in which form, which already requires some selection because raw histories are lengthy and noisy. *Memory management* keeps the store usable over time by merging duplicate entries, forgetting low-value information under capacity constraints and performing reflection steps that compress repeated patterns into more abstract, high-level summaries. *Memory reading* then aims to extract useful information from past behaviour and typically scores candidate items by recency, relevance and importance, a retrieval heuristic first formalised in the Generative Agents work of J. S. Park, O’Brien, et al. (2023). Under this scheme, recent, contextually related and intrinsically salient events are more likely to be selected. Taken together, these mechanisms allow generative agents to exhibit history-dependent and self-consistent behaviour that would be difficult to realise with purely stateless decision rules.

Planning The planning module links high-level goals to concrete action sequences and enables agents to decompose complex tasks into intermediate steps. When planning proceeds without feedback, agents rely solely on internal reasoning to derive a sequence of subgoals. This can take the form of single-path reasoning, where tasks are unfolded into a linear chain of steps, or multi-path reasoning, where several possible continuations are explored in a tree-like fashion (L. Wang, Ma, et al. 2024). Multi-plan selection methods extend this setting by generating several candidate plans and then choosing among them through voting or heuristic search, which trades improved robustness against additional computation (X. Huang et al. 2024). Some approaches complement this with external planners that operate on formal task descriptions and return executable action sequences (Dagan et al. 2023). Other systems adopt planning with feedback, in which partial plans are executed, observed, and revised as new information arrives. Feedback may stem from the environment through state changes or task completion signals, from humans who provide clarifications or corrections, or from models in the form of auxiliary evaluations that critique or refine candidate plans. Across these variants, the planning module provides the mechanism that turns stable goals into context-adaptive behaviour by continuously integrating perception, memory, and feedback into the unfolding plan. (L. Wang, Ma, et al. 2024)

Action The action module translates an agent’s decisions into concrete outputs. It sits at the downstream end of the architecture, where it directly interacts with the environment. It is shaped by the agent’s profile, memory and planning modules. Its operation can be described along four perspectives: *action goals*, *action production*, *action space* and *action impact*. The first two describe what happens before an action, the third describes the action itself, and the fourth captures the consequences of acting. (L. Wang, Ma, et al. 2024)

Action goals specify what the agent intends to achieve in a given moment. Depending on the system, actions may target task completion (Qian et al. 2024), communication with other agents (Larooij and Törnberg 2025a) or exploration of unfamiliar environments to gather information or skills (G. Wang et al. 2023).

Action production describes how actions are generated. Some systems act by retrieving relevant memories and conditioning the LLM on them. Others follow pre-computed plans step by step. Many combine both approaches so that planned behaviour remains sensitive to the current context. (L. Wang, Ma, et al. 2024)

Action space defines which actions are available. This ranges from text-only environments limited to natural language messages to richer systems that expose tool calls (P. Zhao et al. 2023). Even without external tools, agents can act solely through the internal capabilities of the LLM. These include its planning ability, its conversational competence and its grasp of basic commonsense knowledge, which together allow the agent to behave coherently and respond appropriately to situational demands.

Action impact describes what happens after an action has been executed. Outputs may change the external environment. They may also update the agent’s internal state, for example, by adding new memories or altering plans. Actions can also create new preconditions that trigger subsequent decisions and thereby shape the unfolding behaviour of the agent. (L. Wang, Ma, et al. 2024)

Although these actions are expressed in natural language, they remain grounded in the structured state of the simulation. This ensures that the generated behaviour is interpretable for humans and tractable for the

underlying system.

Behavioural guidance Beyond the architectural modules described above, a generative agent’s behavioural capabilities depend critically on how the underlying LLM is guided. Much of this competence stems from prompt engineering (Sahoo, Singh, et al. 2025) as introduced in section 2.2 together with a higher-level layer often termed mechanism engineering or context engineering (Mei et al. 2025). In this analogy, the LLM functions as the behavioural hardware, a powerful but undirected engine, while prompt engineering provides the software layer that specifies how this engine should operate in context. Carefully designed instructions, role descriptions, constraints and examples shape the model’s internal policy and enable behaviours such as planning, self-reflection or persuasive communication without modifying model parameters. Mechanism or context engineering complements this by adding programmatic control structures around the LLM that regulate how behaviour unfolds over time.

2.5 New Paradigm of Generative Social Simulation

The previous subsection introduced generative agents as cognitively enriched LLM-powered entities whose behaviour emerges from a scaffold that manages their profile, memory, planning, and action modules. If multiple such agents are placed in a shared environment where they perceive one another, exchange information, and act under a common set of rules, the result is an LLM-based multi-agent system (T. Guo et al. 2024). The recent literature treats these systems as a broad umbrella that spans a wide range of application scenarios. A first branch comprises problem-solving settings, where specialised agents collaborate to solve complete complex tasks in areas such as software development or robotics (H. Jin et al. 2025; W. Huang et al. 2023). The second branch comprises world simulation systems, which use generative agents to model complex environments and study emergent behaviour. This branch covers domains such as gaming, psychology, economics, policymaking and disease propagation. (T. Guo et al. 2024) Societal simulation is a specific domain within this world simulation branch. It centres on constructing artificial societies in which agents communicate through natural language and generate behavioural traces that can be analysed as social dynamics. It continues the methodological tradition of classical ABM introduced in section 2.3, but replaces handcrafted update rules with an LLM scaffold that governs perception, interpretation and action (Gurcan 2024) as explained in section 2.4.

Terminology Multiple terms are used in the literature to describe simulations in which many generative agents interact within a shared social environment. These include expressions such as “LLM-Empowered Agent-Based Simulation” (C. Gao, Lan, N. Li, et al. 2024), “Generative Agent-Based Modeling” (Ghaffarzadegan et al. 2024), “LLM-Based Social Simulation” (K. Park et al. 2025), “Generative Social Simulation” (Larooij and Törnberg 2025a) and related variants. Although these labels differ in emphasis, they all describe models in which heterogeneous LLM-driven generative agents communicate through natural language and produce behavioural traces that can be analysed as social processes. To remain conceptually aligned with the notion of generative agents introduced in section 2.4 and to connect this line of work to developments in generative social science (Epstein 1999), the term “Generative Social Simulation” is adopted in this thesis.

From agents to social systems Moving from individual generative agents to a full generative social simulation requires several additional design decisions that determine how a population of LLM-powered agents functions as a coherent social system. The first step is to construct an environment that agents can perceive and act in. The agent–environment interface defines how agents receive information, how actions change the world and which aspects of the environment are visible to them (T. Guo et al. 2024). In societal simulations, this interface is typically text-based, which allows both environmental descriptions and agent actions to be expressed in natural language. A second step is to specify the communication structure. Direct text exchange enables conversational interaction, while visibility rules or message routing determine which agents perceive which events and how information propagates through the population. (C. Gao, Lan, N. Li, et al. 2024) A third step concerns population design. The profiling strategies introduced in the previous subsection allow a modeller to define agents with heterogeneous attributes and roles, whether through predefined descriptions, model-generated personas or data-derived profiles that reflect empirical distributions. Finally, the simulation requires orchestration mechanisms that coordinate when agents perceive, reflect and act. Contemporary frameworks use controllers that regulate time stepping and message passing at scale, and they support role-specific behaviour through specialised prompts, memory structures or scenario-conditioned adaptations (Taillandier et al. 2025). These components extend the single-agent scaffold into a structured social environment in which interaction patterns can give rise to higher-level social dynamics.

Evolution of generative social simulations The first influential demonstration of such a system is the Smallville sandbox by J. S. Park, O’Brien, et al. (2023), which showed that a community of twenty-five generative agents situated in an interactive world can produce coherent daily routines, engage in social interactions and coordinate collective events through purely linguistic perception and action. Subsequent work extended this idea to larger and more diverse settings. Social Simulacra (J. S. Park, Popowski, et al. 2022) scaled the approach to one thousand personas, and S^3 (C. Gao, Lan, Z. Lu, et al. 2025) applied generative agents to online network scenarios with populations on the order of ten thousand users. AgentSociety (Piao et al. 2025) further demonstrated large-scale social simulation by modelling over ten thousand agents and several million interactions in a realistic societal environment, using the platform as a testbed for studying phenomena such as polarisation and policy interventions. More recent frameworks pursue substantially larger and more flexible simulations. GenSim supports simulations with around one hundred thousand agents (J. Tang et al. 2025), SocioVerse incorporates millions of real-world user profiles to align agent behaviour with population characteristics (X. Zhang, Lin, Mou, et al. 2025), and Light Society claims to support simulations with up to one billion agents (Guan et al. 2025). Domain-specific systems, such as ElectionSim (X. Zhang, Lin, L. Sun, et al. 2024) illustrate how large populations can be tailored to particular societal processes. Taken together, these developments indicate a gradual shift from small handcrafted sandbox worlds toward more general platforms capable of supporting large, heterogeneous generative social simulations.

2.6 Advantages of Generative Social Simulation

Behavioural expressiveness When many generative agents interact within a shared environment, their behaviour exceeds what classical ABMs can express. They can produce text that resembles human

communication, which increases behavioural fidelity beyond what handcrafted rule agents can achieve (Münker et al. 2025). Classical ABM agents typically rely on predefined rules tailored to narrow tasks and therefore generalise poorly beyond their design scope. LLM-driven agents overcome this limitation by using language-based decision processes learned from large corpora, which enables flexible reasoning and context-dependent action generation across tasks without explicit behavioural programming (P. Zhao et al. 2023). Their communicative competence further expands the behavioural space available to simulations. Natural language provides a rich and interpretable interaction channel that enables agents to converse, negotiate and coordinate through text. Multi-turn linguistic interaction elevates the simulation from isolated responses to community-level dynamics, as dialogue, persuasion and coordination can emerge organically. (C. Gao, Lan, N. Li, et al. 2024)

Persona heterogeneity Generative agents make it possible to populate simulations by drawing synthetic individuals from a target population of interest rather than relying on a homogeneous or implicitly defined agent pool. By conditioning agent profiles on demographic, social or contextual attributes that mirror a real population, silicon sampling allows simulations to approximate population-level diversity in a controlled manner. This reduces the influence of biases inherited from internet-scale pretraining, so that collective behaviour is not dominated by the model’s default training distribution. (Argyle et al. 2023) At the same time, demographically conditioned personas have been shown to reproduce subgroup-specific response patterns observed in human data, for example, in evaluations of familiar versus unfamiliar brands (Kaiser et al. 2025). Compared to random sampling alone, population-aligned persona profiles produce a wider spread of agent behaviours, so coordination, opinion change, and role differences can arise from substantive individual variation instead of everyone following the same generic response pattern (Z. Hu et al. 2025).

Emergence and exploration Generative social simulations can display spontaneous collective patterns that were never explicitly encoded, enabling discovery-oriented analyses of coordination and group dynamics that extend beyond the behavioural repertoire of classical ABMs. Because these systems can be scaled far beyond the limits of human subject experiments, they support the investigation of phenomena that are effectively “*human impossible*”, including large-scale policy interventions or collective responses involving thousands of interacting agents. (Anthis et al. 2025) Such simulations can reveal collective patterns even in the absence of explicit micro rules, thereby providing exploratory insight into social dynamics that would be difficult to observe through traditional modelling approaches (Wu et al. 2025).

Modelling flexibility Another advantage of generative social simulations is that they offer a more flexible and accessible modelling workflow than classical ABM, because core components of a model can be specified directly in natural language. Researchers can define roles, environments, objectives and behavioural constraints through prompts instead of low-level programming, which accelerates model construction and reduces dependence on extensive implementation effort. Natural language interfaces also make simulations easier to design and interpret, since behaviours and roles can be written in human-readable form that directly shapes agent behaviour. (Taillandier et al. 2025) This accessibility broadens participation beyond domains with established ABM expertise, enabling researchers in fields such as

social science, healthcare or urban planning to formulate and explore complex scenarios without adopting specialised simulation tooling. The resulting flexibility supports rapid prototyping, since alternative configurations can be expressed and tested with minimal engineering overhead, lowering the barrier to experimentation compared to traditional ABMs. (Gurcan 2024)

Ethical sandboxing Generative agents further provide a means to investigate ethically sensitive domains without exposing human participants to harm. Because simulated agents can be subjected to biased, stressful or potentially harmful content, researchers can analyse how such factors shape behaviour or social dynamics while avoiding risks that would be unacceptable in human subject studies (Anthis et al. 2025). This enables controlled examination of sensitive social processes or high-risk scenarios that cannot be ethically tested with real populations, thereby broadening the range of societal processes that can be systematically explored under experimental conditions (Taillandier et al. 2025).

2.7 Challenges and Limitations of Generative Social Simulation

Computational cost and scalability Generative social simulations are orders of magnitude more expensive than classical ABM. This creates a structural barrier to large parameter sweeps, replications and robustness checks. Because computational cost forces many studies to run only a single simulation despite substantial stochasticity, reproducibility and sensitivity analysis are directly undermined. (Larooij and Törnberg 2025b) Although recent work introduces memory optimisations, operational improvements, compression methods and batching strategies, the cost of simulating a large-scale society of generative agents continues to be prohibitive (C. Gao, Lan, N. Li, et al. 2024). This computational burden prevents the repeated runs required for confidence intervals, pushing simulations to computational boundaries that make replication and prediction-heavy use cases infeasible at meaningful scale (Münker et al. 2025; Wu et al. 2025).

Opaqueness and limited behavioural fidelity Generative agents introduce deep epistemic opacity because their behavioural model is encoded in billions of parameters that cannot be inspected directly, which leaves designers with limited insight into why an agent acts as it does or how its internal dynamics relate to the processes the simulation is meant to capture (Larooij and Törnberg 2025b; Taillandier et al. 2025). As probabilistic next-word-predictors, LLMs offer no built-in notion of correctness and therefore routinely produce hallucinations that undermine the reliability of their inferred reasoning (Sahoo, Meharia, et al. 2024). Their outputs may be linguistically coherent yet rest on superficial or misleading causal structures, since the model is optimised to imitate plausible language patterns rather than to implement validated mechanisms of human cognition or action (Taillandier et al. 2025). These properties limit fidelity at the individual level. Empirical studies show that LLM-based human simulations are not reliably aligned with observed human behaviour across domains due to inherent LLM limitations and simulation design flaws (Q. Wang et al. 2025). LLM agents cannot be assumed to impersonate specific individuals or social groups reliably, and aligning them with personalities or demographic profiles is inherently unstable. This constrains their ability to function as transparent or faithful models of human decision-making in generative social simulations. (Larooij and Törnberg 2025b; Q. Wang et al. 2025; Wu et al. 2025)

Temporal instability and reproducibility Furthermore, generative social simulations face substantial reproducibility challenges because they combine the chaotic sensitivity to initial conditions typical of complex systems with the stochastic variability of language model outputs, producing run-to-run divergence even under identical setups (Larooij and Törnberg 2025b). Temporal consistency within a single run is also fragile. Agents can drift in persona, memory or belief trajectories when internal state is not firmly anchored, making long-term coherence difficult to maintain (J. Zhou et al. 2025). Such drift complicates the interpretation of emergent behaviour because it becomes unclear whether observed patterns reflect genuine dynamics or instability artefacts (Wu et al. 2025). Model-level inconsistencies add to these challenges. Even when given identical prompts, LLM agents may produce inconsistent or odd responses, particularly in smaller models, which weakens reproducibility across runs (C. Gao, Lan, N. Li, et al. 2024). Simulation design flaws can further reinforce these effects, generating variation that reflects artefacts rather than underlying social processes (Q. Wang et al. 2025).

Persona collapse and distributional misalignment LLM agents tend to collapse toward an “*average persona*” because next-token prediction rewards high-frequency, mainstream linguistic patterns and suppresses minority or low-frequency behaviours. Empirical studies show the same tendency, with LLM-generated text systematically reflecting dominant linguistic and social norms while underrepresenting minority expressions (Gallegos et al. 2024). More generally, persona prompting captures only a small share of human behavioural variation, which encourages averaged or stereotyped group behaviour (T. Hu and Collier 2024). As a result, LLM agents often fail to represent social groups accurately and instead reproduce biases inherited from training data (Navigli et al. 2023). This reduces behavioural variance, erases subgroup characteristics and undermines simulations that depend on heterogeneous preferences or diverse interaction styles (Larooij and Törnberg 2025b; Taillandier et al. 2025). These failures can be uneven across linguistic and cultural populations, with substantially worse alignment in low-resource languages (Z. Jin et al. 2024).

In addition to this variance collapse, generative agents can also deviate systematically from human population means (Wu et al. 2025). Survey style evaluations repeatedly find that synthetic respondents can approximate aggregate averages while underrepresenting variance and distorting subgroup distributions, yielding overconfident and misaligned population profiles (Bisbee et al. 2024; Sarstedt et al. 2024; Kaiser et al. 2025; Boelaert et al. 2025). Broader critiques identify sycophancy and alienness as structural weaknesses that push LLMs toward socially normative yet psychologically ungrounded responses, limiting their ability to model human groups faithfully (Anthis et al. 2025). Population-level misalignment arises because LLMs express topic-specific political biases and US-centric framing patterns (Bang et al. 2024) and show substantial divergence from the distributions of opinions held by actual demographic groups, even when explicitly steered toward them (Santurkar et al. 2023).

Prompt sensitivity and experimenter effects Generative social simulations are vulnerable to experimenter visibility effects (Gilder and Heerey 2018), because agents can pick up cues embedded in prompts and adjust their behaviour accordingly. This can include cues that implicitly encode hypotheses or desired outcomes and thereby steer agents toward scripted rather than emergent patterns (J. Zhou et al. 2025). Similar to social experiments with human participants, they are susceptible to demand charac-

teristics, meaning that behaviour shifts when the agent infers what outcome is expected in the scenario (McCambridge et al. 2012). They may follow these prompt cues in ways that echo shortcut learning (see Section 2.2). A related mechanism is the Hawthorne effect, where behaviour shifts merely because the agent recognises that it is being evaluated. This capability is documented in research showing that LLMs can detect evaluation settings and adjust their responses accordingly (Adair 1984; Needham et al. 2025). A further analogue is social desirability bias, where outputs align with perceived norms of acceptability rather than with mechanism-driven responses, a tendency reinforced by sycophantic behaviours in which models echo what they believe the user wants to hear (Edwards 1957; Rrv et al. 2024).

Design flaws such as homogeneous profiles, missing interaction structures or discarded memory can generate patterns that resemble social phenomena without arising from genuine agent interaction (J. Zhou et al. 2025). Broader usage issues, including misaligned scenarios or loosely specified prompts, can amplify these artefacts by making outputs track textual cues more strongly than underlying model dynamics (Wu et al. 2025). These vulnerabilities are compounded when embedded bias or limited generalisation capacity causes agents to reproduce the researcher’s wording or assumptions, weakening the independence of simulated behaviour (Anthis et al. 2025; Munker et al. 2025).

Limits for causal and confirmatory use Generative social simulations face limits for confirmatory or causal inference because they lack the parsimony of formal models and the empirical grounding of data-driven approaches (Larooij and Törnberg 2025b). Their coherent linguistic output gives an impression of realism, but this surface-level behaviour does not reflect underlying mechanisms, and many observed patterns arise from prompt artefacts or training effects rather than from genuine interaction dynamics (J. Zhou et al. 2025). In particular, behaviours that appear emergent can be observationally indistinguishable from memorised patterns or data leakage from prior literature, which fundamentally undermines claims that such dynamics arise causally from agent interaction (Barrie and Törnberg 2025). These issues make it difficult to build cumulative theoretical knowledge on top of such simulations.

Research further shows that societies of generative agents cannot reproduce causal structures or stable empirical regularities in a reliable way, since model and framework constraints limit the representation of mechanistic processes (Wu et al. 2025). LLM behaviour diverges from human behaviour even in simple decision-making tasks and varies substantially across models, prompts and contexts, which undermines robustness for confirmatory or causal use (Y. Gao et al. 2025). Treating generative agents as human surrogates, therefore risks overstating their scientific capacity and causal interpretability (Anthis et al. 2025). Because these agents do not implement explicit, theory-driven behavioural mechanisms, observed outcomes cannot be traced back to identifiable causal processes, which limits their usefulness for causal explanation or for establishing empirical correspondence (Taillandier et al. 2025; Q. Wang et al. 2025).

Ethical and societal risks As noted above, distributional misalignment and variance loss can already distort population-level behaviour. These distortions carry ethical significance when they translate into biased or inaccurate portrayals of demographic groups within simulated societies (Larooij and Törnberg 2025b). Evidence shows that prompt-based persona generation can systematically produce simplified and stereotypical portrayals of demographic groups, with LLM-generated personas exhibiting stronger racial stereotyping than comparable human-written descriptions (M. Cheng et al. 2023). Such misrep-

resentation can shape population structures, erase minority perspectives or amplify culturally dominant viewpoints in ways that undermine fairness and demographic realism (Wu et al. 2025). These risks intensify in simulations involving polarised or sensitive topics, where generative agents may reproduce or amplify harmful content such as outrage, stigma or hate speech (Münker et al. 2025). A final concern is interpretative: fluent outputs can foster automation bias (Goddard et al. 2012), encouraging researchers or policymakers to treat biased artificial societies as valid representations of human populations despite opaque reasoning and flawed outputs (Taillandier et al. 2025).

Overall, the challenges outlined above show that the core problem of generative social simulation is not expressive power but credibility. High computational cost limits replication and robustness, epistemic opacity and hallucinations undermine interpretability and individual-level fidelity, stochastic drift and prompt sensitivity weaken reproducibility, and distributional misalignment distorts population-level behaviour. These issues interact in ways that make surface-level realism or qualitative plausibility insufficient for scientific claims. Validation therefore, cannot be treated as a single post-hoc check, but must specify what a simulation is meant to explain, at which level of abstraction, and under which empirical constraints. Recent work converges on the view that credible validation in LLM-based social simulation must be multi-layered, distinguishing between purpose-aligned external grounding, micro-level agent fidelity and consistency, and macro-level experimental design conditions that allow collective patterns to emerge without being scripted by researcher intent.

2.8 Validation Principles for Generative Social Simulation

The current discussion on how to validate generative social simulations is mainly shaped by four recent contributions that address different facets of the problem. Larooij and Törnberg (2025b) identify validation as the central challenge for generative social simulation, arguing that current generative social simulations often validate the wrong quantities. J. Zhou et al. (2025) synthesise recurring methodological flaws in generative social simulations and formalise the PIMMUR framework. It defines six requirements, namely *Profile*, *Interaction*, *Memory*, *Minimal-Control*, *Unawareness* and *Realism*, as necessary conditions for credible generative social simulation. These principles are intended to prevent artefacts introduced by prompting practices, insufficient agent design and constrained interaction structures. Wu et al. (2025) argue that generative social simulations are bounded by problems of alignment, consistency and robustness. They outline heuristic criteria that clarify the conditions under which these models can produce defensible insights into collective social patterns. Münker et al. (2025) argue that generative social simulations must be validated against real-world data and that their credibility depends on demonstrable empirical realism. They propose evaluating each simulation component with task-specific loss functions that quantify how closely simulated behaviour matches observed human behaviour. Any downstream findings should therefore be interpreted in light of these realism scores, which indicate the degree to which simulated outcomes remain tethered to the empirical domain.

Taken together, these works outline an emerging consensus on what constitutes rigorous validation in LLM-driven social simulations. The following subsections synthesise these insights into three complementary dimensions: purpose-aligned external validation, micro-level agent fidelity, and macro-level experimental design governing robustness and independence from researcher intent.

2.8.1 Purpose-Aligned and Externally Grounded Validation

Validation in generative social simulation begins with a purpose-aligned question about what the model is intended to explain or predict, because the criteria for success follow directly from this definition.

A central critique is that generative social simulations often validate the wrong quantities. The review by Larooij and Törnberg (2025b) shows that current work relies on five broad validation techniques: judgments by humans or LLMs about the plausibility of agent behaviour, comparisons with well-known social patterns, benchmarking against other models, matching outputs to human-generated data and tests of internal coherence or parameter stability. While these approaches differ in sophistication, most target surface-level properties such as stylistic realism or qualitative plausibility rather than the behavioural dynamics the model is intended to capture. Several techniques also suffer from circularity, data leakage or purely subjective interpretation. Larooij and Törnberg (2025b) therefore argue that only empirical, purpose-aligned validation against real behavioural data can establish operational validity.

Münker et al. (2025) extend this concern by providing a framework for external grounding. They argue that generative social simulations must demonstrate empirical realism for the components that are central to the phenomena under study, and that any simulated outcomes should be interpreted only in light of these realism scores. In their example, they focus on user-communication behaviour because it drives their substantive claims: the component is trained on real Twitter data, evaluated with task-specific loss functions comparing simulated and human messages, and the resulting scores provide quantitative evidence of behavioural fidelity. Their broader recommendation is that all components that materially influence outcomes should, in principle, be validated for empirical realism, making such grounding a precondition for drawing substantive conclusions. This directly operationalises the critique raised by Larooij and Törnberg (2025b): validation must go beyond surface-level plausibility and assess the mechanisms that matter for the model's stated purpose.

A further part of the debate concerns the scope of claims that LLM-based simulations can reasonably support. Wu et al. (2025) emphasise that alignment, consistency and robustness constraints place inherent limits on what such simulations can show, and that validation should therefore match the types of conclusions the model can sustain. They argue that generative social simulations are most credible when used to analyse collective patterns rather than individual behaviour. This complements the concerns raised by Larooij and Törnberg (2025b) as well as the empirical realism framework of Münker et al. (2025): validation should be conducted at the abstraction level at which the model can reliably operate. Taken together, these contributions frame validation as a purpose-specific and empirically grounded exercise, in which simulated outcomes are judged relative to the external data and behavioural scales they aim to approximate.

2.8.2 Micro-Level Agent Design

Micro-level validity concerns whether individual agents behave as sufficiently rich, stable and heterogeneous actors to support meaningful social interaction. The PIMMUR framework provides clear criteria for this level of evaluation. Its first three principles *Profile*, *Interaction* and *Memory* identify recurring flaws in current generative agent designs and define the conditions under which micro-level behaviour

can be considered credible (J. Zhou et al. 2025). *Profile* requires that agents possess distinct backgrounds or preferences rather than representing homogeneous copies of a single underlying model. *Interaction* requires that agents influence and respond to one another through genuine exchanges instead of receiving aggregated population statistics. *Memory* requires that agents maintain persistent internal states so that beliefs and identities evolve across time rather than resetting with each prompt. When these conditions are violated, simulations tend to collapse into artefacts such as homogeneous conversations, scripted outcomes or pseudo-multi-agent setups that cannot support emergent dynamics.

Complementing these design criteria, Wu et al. (2025) argue that current generative agents often collapse into an average persona with low behavioural variance and that they struggle to maintain stable roles over longer simulations. In their view, limited heterogeneity and temporal drift constrain the range of emergent dynamics that can be meaningfully studied and make individual trajectories difficult to interpret.

These considerations motivate micro-level realism assessments that evaluate whether individual agent behaviour approximates human responses when examined at the level of specific decisions. Munker and colleagues show that such fidelity is achieved by treating behavioural components as learned predictors with explicit training objectives rather than relying solely on generic prompting (Munker et al. 2025). In their study, modules for posting, replying and reply likelihood are fitted to real user data and evaluated with task-specific metrics that quantify how closely model outputs reproduce observed behaviour. The resulting realism scores function as an empirical calibration of the agent’s behavioural competence and provide a principled basis for interpreting downstream simulation outcomes at the individual level.

Overall, these perspectives suggest that micro-level validation should ask whether agents act consistently over time, differ meaningfully from one another, and reproduce empirically observed behavioural regularities, since only under these conditions can individual agents serve as a credible basis for studying higher-level social processes.

2.8.3 Macro-Level Experimental Design

At the macro-level, the question is whether a simulation creates conditions in which collective patterns can arise without being scripted by the experimenter. The PIMMUR framework formulates three principles for this level of design: *Minimal-Control*, *Unawareness* and *Realism* (J. Zhou et al. 2025). *Minimal-Control* requires that prompts and system rules specify only what is needed for perception, action and communication, so that outcomes are driven by agent interaction rather than by hints that encode the desired effect. *Unawareness* requires that agents cannot easily infer the experimental hypothesis, manipulation or evaluation criterion, since such meta-knowledge introduces demand characteristics and makes behaviour track the research script. *Realism* requires that simulations are anchored in empirical data and that emergent patterns are evaluated against observations from real societies instead of only against simplified theoretical models. Together, these constraints define a macro design space in which collective behaviour emerges under plausible informational conditions rather than direct instructional control.

Because generative social simulations are stochastic and sensitive to modelling choices, macro-level validity also depends on robustness. Larooij and Törnberg (2025b) argue that results need to be probed across seeds, parameter settings and design variants in order to distinguish stable regularities from artefacts of a particular configuration. Wu et al. (2025) emphasise that, given current LLM limitations, macro-level

conclusions are most credible when simulations are interpreted at the level of population-wide patterns rather than individual trajectories. Coherent language does not guarantee valid causal dynamics, and the models' behavioural variance is too limited to support fine-grained claims.

Together, these considerations frame macro-level design as the task of enabling meaningful emergent patterns while guarding against artefacts introduced by prompting, scenario structure or a lack of robustness.

3 Theoretical Foundations of Information Disorder

Information disorder provides the conceptual lens through which this chapter examines the dynamics of misleading and harmful information in digital environments. Following the framework introduced by Wardle and Derakhshan (2017), information disorder refers to a broad class of polluted information streams in which false, misleading, manipulated or strategically disclosed content circulates through public communication at a global scale, enabled and accelerated by contemporary social technologies. Rather than treating misinformation as a collection of isolated false claims, this perspective emphasises the interaction between strategic agents, stylistically constructed messages and active interpretation by audiences, unfolding across the phases of creation, production and distribution. Building on this framing, the chapter integrates theoretical work on online social networks, social bots and computational amplification, narrative styles of misinformation, and models of opinion formation and individual interpretation. Together, these perspectives establish the domain assumptions that underpin the simulation framework developed in subsequent chapters.

3.1 Types of Information Disorder

In response to the inadequacy and politicisation of the term fake news, Wardle and Derakhshan (2017) introduce information disorder as a broader conceptual framework for analysing polluted information environments. The term fake news is rejected because it fails to capture the diversity of misleading and harmful information practices and has increasingly been appropriated as a rhetorical weapon to delegitimise unfavourable journalism. Instead, it distinguishes types of information disorder along two core dimensions: the falseness of content and the intent to cause harm. This distinction enables a more precise analysis of how different forms of problematic content are created, circulated and instrumentalised in digital public communication.

Misinformation Misinformation refers to information that is false or misleading but shared without an intention to cause harm. It commonly arises in contexts of uncertainty, urgency or information overload, such as breaking news events, crises or rapidly evolving situations. Individuals who spread misinformation are often motivated by a desire to be helpful, expressive or socially responsive rather than by strategic goals. Despite the absence of malicious intent, misinformation can nonetheless have significant societal consequences when it spreads widely or becomes embedded in collective understandings. (Wardle and Derakhshan 2017)

Disinformation Disinformation denotes information that is false and deliberately created, produced or distributed to cause harm or to influence a person, group, organisation or country. Intentionality is the defining feature of this category. Disinformation includes fabricated content, manipulated or decontextualised material, impersonation of legitimate sources and coordinated amplification campaigns. Such practices are often embedded in organised influence operations that exploit digital platforms to shape perceptions, undermine trust, and strategically distort public discourse. (Wardle and Derakhshan 2017)

Malinformation Malinformation describes information that is based on reality but used strategically to cause harm. Unlike mis- and disinformation, malinformation does not rely on falsity. Instead, harm arises

from the timing, framing or context of disclosure. Typical examples include leaks of private communications, doxxing, targeted harassment and the selective release of truthful information designed to damage reputations or manipulate public perception. In Wardle’s framework, some forms of hate speech and harassment are included under malinformation, as individuals are often targeted based on real personal attributes or affiliations that are mobilised to inflict harm. (Wardle and Derakhshan 2017)

Propaganda Closely related to these categories is propaganda, which Jowett and O’Donnell (2015) define as a form of communication that seeks to achieve a response that furthers the desired intent of the propagandist. It is deliberate and systematic in the sense that it is planned in advance, organised with regularity and designed to shape perceptions, manipulate cognitions and direct behaviour. Propaganda may rely on truthful, partially truthful or false content, since its defining property is not falseness but the strategic orchestration of messages to secure an outcome advantageous to the source. In contrast, persuasion is understood as an interactive process that attempts to satisfy the needs of both persuader and persuadee rather than privileging the interests of one side. Rather than constituting a separate category defined by falseness, propaganda operates as an overarching influence logic that strategically mobilises these forms of information disorder within coordinated campaigns.

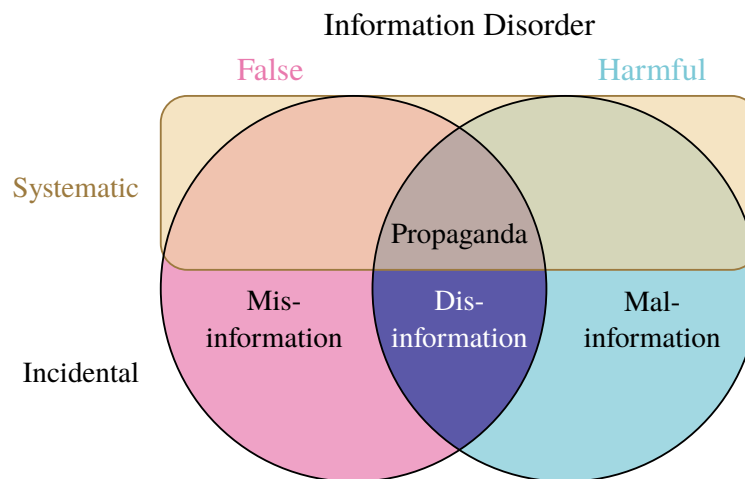


Figure 3: Conceptual model of information disorder, illustrating misinformation, disinformation, and malinformation along the dimensions of falseness and harm, with propaganda represented as a systematic influence logic that can strategically mobilise all three forms (Source: adapted from Wardle and Derakhshan 2017).

Figure 3 summarises the conceptual relationships between the different forms of information disorder discussed in this section. Misinformation and malinformation are distinguished by their position along the falseness and harm dimensions, while disinformation occupies the intersection of false and harmful content. Propaganda is depicted as an overarching, systematic mode of influence that operates across these categories by coordinating and strategically deploying misleading or harmful information, regardless of its factual status.

3.2 Phases of Information Disorder

Information disorder can be understood as unfolding across a sequence of analytically distinct but tightly interrelated phases. Following Wardle and Derakhshan (2017), the life of an instance of information disorder

der is conceptualised in terms of three phases: creation, production and distribution. The creation phase concerns the formulation of an initial message, such as a fabricated claim, a rumour or a selectively framed disclosure. In the production phase, this message is transformed into a media artefact that is formatted, stylised and packaged for circulation. The distribution phase refers to the processes through which this artefact is made public and propagated across communication channels. Importantly, these phases often involve different agents with distinct roles and motivations, and once content enters the distribution phase, it can be reproduced and redistributed by a potentially unbounded set of actors. Analysing information disorder through this phased lens makes it possible to disentangle the heterogeneous contributions of creators, producers and distributors and to trace how messages evolve as they circulate through digital environments.

Online social media platforms have fundamentally altered the conditions under which information disorder spreads. Low publication barriers, algorithmically curated feeds and engagement-based visibility metrics reshape which messages gain attention and how widely they circulate. False content systematically diffuses faster, reaches larger audiences and forms deeper diffusion cascades than verified information (Vosoughi et al. 2018). Systematic reviews add to this perspective by showing that on social media, the phases of creation, production and distribution increasingly collapse into a recursive process. Platform affordances enable the same actors to create, stylise and redistribute content almost simultaneously, while feedback from engagement metrics shapes subsequent message formulation (Aïmeur et al. 2023; S. Chen et al. 2023). Together, these dynamics help explain why information disorder propagates rapidly and resists containment in online environments. Importantly, this process is not value-neutral. Platform ownership structures, advertising-driven business models and opaque ranking algorithms systematically shape visibility and amplification, embedding the distribution of information disorder within broader dynamics of platform power and computational propaganda (Fuchs 2018).

A phase-based perspective alone is insufficient to capture how information disorder operates. Following Wardle and Derakhshan (2017), the analysis therefore also considers three constitutive elements that shape how problematic content emerges, circulates and exerts influence: agents, messages and interpreters. Agents create, shape and propagate content across phases, messages constitute the concrete communicative artefacts, and interpreters actively evaluate and retransmit information based on prior beliefs and contexts. Together, these elements provide the analytical structure for the following sections.

3.3 Agents as First Element of Information Disorder

Within Wardle's framework, agents denote the actors who create, produce and distribute messages across the phases of creation, production and distribution (Wardle and Derakhshan 2017). These agents are characterised along several analytical dimensions, summarised in Table 1. Agents may be official or unofficial, operate individually or within loosely or tightly organised networks, and act under different motivational logics, including financial, political, social or psychological incentives. They further differ with respect to their intended audiences, which may range from specific individuals to entire societies. Two additional dimensions are analytically crucial for distinguishing forms of information disorder: whether an agent intends to mislead and whether it intends to cause harm. Finally, agents vary in their degree of automation, spanning purely human activity, hybrid human–software systems and fully automated bots. While all

Table 1: Analytical dimensions for characterising agents in information disorder (Source: adapted from Wardle and Derakhshan 2017).

Dimension	Possible values
Type of actor	Official, Unofficial
Level of organisation	Individual, Loosely organised, Tightly organised, Networked
Motivation	Financial, Political, Social, Psychological
Intended audience	Specific individuals, Social groups, Entire societies
Degree of automation	Human-operated, Hybrid human-software systems, Automated bots
Intent to mislead	Yes, No
Intent to harm	Yes, No

combinations of these characteristics are relevant in principle, the following sections focus on automated agents operating with an intent to mislead and an intent to harm, as such agents have emerged as a central mechanism through which information disorder is amplified at scale in contemporary social media environments.

Distinction of agent terms A brief clarification of terminology is necessary. In Wardle’s framework, an “*agent*” refers broadly to any actor, human or automated, involved in the creation, production or distribution of content (Wardle and Derakhshan 2017). This differs from the generative agents introduced in Section 2.4, which are LLM-driven constructs used for simulation. Whereas Wardle’s notion describes real-world actors participating in information disorder, generative agents serve as a modelling abstraction for studying behaviour within controlled computational environments.

3.3.1 Social Bots as Automated Agents in Information Disorder

Definition of bots Across computing and social science literature, the term bot is used for a heterogeneous set of automated software agents that act on online platforms without direct human intervention. This umbrella category historically includes web crawlers, scrapers, spambots, chatbots and automated accounts on social media (Gorwa and Guilbeault 2018). These systems vary markedly in structure and function, which has contributed to persistent conceptual ambiguity in both academic and policy discussions.

Definition of social bots Within this wider family, social media bots form a more specific category tied to the basic elements of social platforms: users, content and relationships. L. H. X. Ng and Carley (2025b) state that these automated accounts should not be defined by their intentions, but by the platform mechanics they can perform. Such accounts may create, distribute or collect content, and they may form or dissolve relationships with other users. This view highlights that automation itself is neutral and that the same technical capabilities can support notification services, information-provision tools, entertainment bots or coordinated manipulation campaigns.

Despite this conceptual neutrality, the research landscape is dominated by cases in which automated accounts are deployed in harmful or adversarial settings. Empirical studies show that such accounts

are disproportionately active in political elections, public health crises, climate debates and financial markets, particularly in contexts marked by controversy or high polarisation (K.-C. Yang et al. 2025). Most misinformation on social media platforms such as Twitter is initiated or amplified by automated agents rather than human users (Tomassi et al. 2024). This empirical concentration has reinforced the association between bots and manipulation, although benign applications exist.

To maintain conceptual clarity, the term social bot is used here to refer specifically to automated social media accounts that are deployed with manipulative or otherwise harmful intent. This operational definition reflects the dominant use case examined in contemporary research and aligns with the phenomena analysed in this thesis.

Bot strategies Beyond definitional questions, social bots exhibit recurrent behavioural strategies that shape their role in information operations. Coordinated deployments often follow staged patterns in which bots initiate, accelerate and amplify information cascades, with early activity disproportionately targeting the diffusion phase of low-credibility content (Shao et al. 2018; L. Xu 2023). To maximise propagation, bots frequently direct attention towards influential users and strategically expand their networks through interactions with high-profile accounts (Shao et al. 2018; Aiello et al. 2021). These behaviours can be differentiated into four functional subclasses (Lopez-Joya et al. 2025). Spambots automate large-scale content distribution with minimal interaction, overwhelming information spaces through high-volume posting. Social manipulation bots influence collective dynamics by steering conversations, constructing false narratives or amplifying social tensions. Personalised attack bots target specific individuals or small groups through harassment, phishing, impersonation or other forms of targeted exploitation. Influence-manipulation bots operate on platform metrics, artificially inflating followers, likes or engagement signals in ways that distort perceived credibility and compromise ranking and recommendation systems. These strategic patterns show that social bots operate not only as content distributors but as coordinated actors that shape how information spreads, gains visibility and acquires perceived legitimacy within online social networks.

3.3.2 Evolution into LLM-Powered Social Bots

Bot evolution The development of social media bots shows a gradual progression from simple automation to increasingly adaptive and human-like systems. Early bots on platforms such as Twitter were rudimentary programs that mainly posted content automatically at very high frequencies, making them easy to detect through basic behavioural heuristics (Ferrara, Varol, et al. 2016). Honey-pot deployments in the early 2010s confirmed this simplicity by attracting large numbers of automated accounts that followed predictable, indiscriminate engagement routines (Lee et al. 2021). As bot makers sought to evade such detection, these systems began to emulate human temporal rhythms by scheduling posts around typical activity patterns and producing bursts of engagement that resembled organic behaviour (Ferrara, Varol, et al. 2016). They also populated profiles with scraped media and text and adopted basic social actions such as following, commenting and replying, allowing automated accounts to appear more coherent and blend more easily into everyday interactions (Hwang et al. 2012).

By the early 2010s, these developments had produced a second generation of bots capable of infiltrating

topical discussions, generating topically appropriate content and forming plausible social connections. Research on bot detection expanded in parallel, accelerating after the 2016 US presidential election, when automated accounts were shown to amplify polarising narratives and disinformation at scale. A third wave of bots emerged from 2016 onwards, characterised by coordinated inauthentic behaviour, hybrid human–automation control and increasingly convincing linguistic and behavioural mimicry. These systems undermined early supervised detectors that assumed clear separability between bots and humans: while older bots were easily recognised, newer ones evade removal and are correctly identified by human annotators only in a minority of cases. (Cresci 2020)

The progressive sophistication of bots produced a corresponding shift in detection approaches. From 2012 onward, group-based detectors that analyse synchronised posting patterns, shared metadata structures and coordinated behaviours became prominent, reflecting the insight that modern bots often operate not as isolated accounts but as parts of larger botnets designed to manipulate information flows collectively. At the same time, the globalisation of bot activity accelerated: by 2018, political manipulation by social bots had been documented in nearly forty countries, and open-source repositories for deploying Twitter bots grew rapidly, increasing from over 4,000 in 2016 to more than 40,000 by 2018. (Cresci 2020) Overall, these trends illustrate a long-running arms race in which advances in automation continually outpace detection practices. They also establish the historical foundation for the emergence of LLM-based social bots, whose capabilities further disrupt established distinctions between human and automated agency.

LLM-powered social bots LLMs have introduced a qualitatively new form of automation that builds directly on the properties discussed in Section 2.1: LLMs generate text that is coherent, contextually adaptive and, for many tasks, indistinguishable from human output. When embedded into social media accounts, these capabilities allow bots to participate in discussions with a degree of linguistic fluency and situational responsiveness that earlier systems could not achieve (Lopez-Joya et al. 2025; Yizhou Zhang et al. 2024). Experimental evidence indicates that human users identify such LLM-powered social bot accounts only marginally above chance (Radivojevic et al. 2024), and prompting techniques can further condition behavioural style, making realism an adjustable design parameter rather than a technical limit (L. H. X. Ng and Carley 2025b). These developments substantially narrow the behavioural gap between social bots and humans, reducing the effectiveness of earlier bot taxonomies that relied on clear distinctions between them (Lopez-Joya et al. 2025). At the same time, the accessibility of LLM technology lowers the barrier to creating automated accounts, enabling both a proliferation of simple, low-cost bots requiring minimal expertise and the emergence of increasingly sophisticated agents that emulate human behaviour (Imperva 2025).

LLMs and the detection arms race This shift has significant implications for detection. All bot detection methods, whether based on content, temporal rhythms, network structure, interaction patterns or aggregated feature sets, implicitly depend on a separable boundary between “*human-like*” and “*bot-like*” behaviour (Ferrara, Varol, et al. 2016). LLM-powered social bots push automated behaviour closer to the human side of this feature space: their linguistic variability weakens content-based cues, their adaptive posting rhythms undermine temporal heuristics and their coherent conversational replies reduce the reliability of interactional signals (Lopez-Joya et al. 2025). A further difficulty arises from their stochastic

generation process. LLMs introduce controlled randomness that breaks many of the patterned signatures on which detection algorithms rely. The challenge is not that detection becomes impossible, but that LLMs shift the underlying behavioural distributions such that previously valid decision boundaries no longer hold. At the same time, LLMs reshape the detection landscape itself. The generative capabilities that enable more evasive bots can also be incorporated into advanced detection systems, making LLMs a dual-use technology and marking the latest iteration of the ongoing arms race between bot creation, evasion and detection (Cresci 2020; Feng et al. 2024).

Scale and systemic effects Automation now operates at a scale that materially alters the conditions of online interaction. In 2024, automated traffic surpassed human traffic, accounting for 51% of all web activity, driven in part by the rapid diffusion of AI and LLM-based tooling. Levels of malicious automation have increased accordingly, with harmful automated activity now representing an estimated 37% of global internet traffic, up from 32% in the previous year. (Imperva 2025) Longitudinal studies likewise document sustained growth in the volume, sophistication and reach of automated agents across platforms (Cresci 2020). Social media exhibits the same tendency: automated accounts commonly make up 20% of users, rising to more than 40% in politically charged contexts such as national elections (L. H. X. Ng and Carley 2025b). As such activity expands, the distinction between human and synthetic actors becomes increasingly difficult to maintain, and the reliability of authenticity cues erodes. This contributes to declining user trust, more guarded interaction and a pervasive uncertainty about who or what populates public platforms (Cresci 2020; Radivojevic et al. 2024). These developments resonate with “*dark forest*”² and “*dead internet*” (Muzumdar et al. 2025; Sommerer 2025) interpretations of the contemporary web, in which automated activity is pervasive, human presence becomes opaque, and users gradually retreat from open, high-risk spaces of interaction.

3.4 Messages as the Second Element of Information Disorder

In the information disorder framework, messages denote the concrete artefacts through which information disorder becomes observable. They include spoken rumours, text posts, articles, memes, images, videos and hybrid media formats. Messages constitute the second element of the framework and occupy an intermediate position between agents and interpreters. They express the intentions of agents and serve as the stimuli that interpreters evaluate and make sense of. Wardle characterises messages along several degrees of freedom: their format, their durability over time, their level of accuracy, their legality, whether they impersonate official sources and the target they address. These dimensions help to distinguish, for instance, between short-lived rumours and long-running conspiratorial narratives, or between inaccurate information and impersonated official communication. (Wardle and Derakhshan 2017)

For present purposes, two additional dimensions are analytically necessary. First, messages vary systematically in their *topic*: disinformation clusters around certain issue domains, and this topical focus strongly shapes both audience reactions and the strategic deployment of automated agents. Second, messages differ in their *narrative style* or framing. Wardle notes that effective disinformation tends to provoke emotion, offer strong and easily grasped narratives, rely on compelling presentation and use repetition, a set of

² <https://ystrickler.medium.com/the-dark-forest-theory-of-the-internet-7dc3e68a7cb1>

Table 2: Analytical dimensions for characterising messages in information disorder (Source: adapted from Wardle and Derakhshan 2017).

Dimension	Possible values
Format	Textual, Audio, Visual, Audiovisual
Durability	Single Moment, Short-term, Event-based, Long-term
Target of message	Individual, Organisation, Social group, Entire society
Topical domain	Health, Science, Politics, Social Issues, Economy, Finance, Disaster/Tragedy, Other (Celebrity Rumors, Conspiracy Theories)
Narrative style	Neutral, Emotional-moral, Post-truth populist, Pseudo-scientific, Repetitive, ...

properties that reflects broader stylistic strategies through which messages frame events, signal authority or invite readers to fill interpretive gaps (Wardle and Derakhshan 2017). Conceptualising these patterns as narrative styles has the additional benefit of subsuming several of Wardle’s original message variables. Both the level of accuracy of a message, ranging from fabricated content to truthful but harmful misinformation, and its imposter characteristics emerge as consequences of stylistic choices, such as asserting claims with unwarranted confidence, appealing to pseudo-authority or mimicking institutional language. Legal status similarly follows from narrative construction, as messages may deliberately adopt forms that violate legal boundaries, for example, through harassment, hate speech or impersonation. Treating accuracy, impersonation and legality as aspects of narrative style therefore provides a more unified lens on how messages are constructed and why they succeed. Table 2 summarises the resulting set of message dimensions used in the present analysis. The following subsections examine, first, the topical domains in which misinformation clusters and, second, the principal narrative styles through which such messages are articulated.

3.4.1 Topics of Misinformation

Empirical analyses show that misinformation on social media is not evenly distributed across subject areas but clusters around a relatively stable set of domains. A large-scale mapping of information disorder by Tomassi et al. (2024) identifies six broad topical clusters that account for the vast majority of misleading content circulating across major platforms.

The most prominent category is *Health and Science*, driven in recent years by the COVID-19 pandemic, debates about vaccines and pharmaceuticals, climate change, and other scientifically framed controversies. This domain attracts high volumes of mis- and disinformation, including conspiracy theories about biotechnology and “big pharma” as well as alternative or pseudo-scientific claims. (Tomassi et al. 2024)

A second major cluster concerns *Politics*. Misinformation in this domain centres on electoral processes, geopolitical conflicts and partisan narratives. Prominent examples include the Brexit referendum, the 2016 and 2020 US elections, allegations of foreign interference, nationalist movements and dissident activity (Tomassi et al. 2024). Complementary evidence from social bot research shows that political elections are the most consistently bot-saturated environments, with automated accounts active across

US, French and German national elections (K.-C. Yang et al. 2025).

Social Issues form a third recurring domain, covering topics such as migration, race, religion, gender, poverty and conflict. These issues regularly trigger strong value-laden reactions, which platforms' sharing dynamics further amplify. As Tomassi et al. (2024) note, topics that touch upon identity or moral commitments often elicit rapid, affect-driven responses, making them highly susceptible to organic and automated amplification.

Two smaller but persistent clusters concern *Economy and Finance*, including financial markets as well as cryptocurrencies, and a heterogeneous *Other* category that aggregates celebrity rumours, conspiracy theories that fall outside the major domains and various internet hoaxes (Tomassi et al. 2024).

The least represented but still relevant category is *Disasters and Tragedies*, encompassing criminal events, terrorist attacks and natural disasters. Although such events attract intense public attention, they tend to generate less misinformation overall, possibly because empathetic proximity and the absence of contentious identity or ideological stakes reduce the incentives and opportunities for strategic manipulation (Tomassi et al. 2024). Evidence from bot analyses corroborates this distribution: automated accounts are most active in political, health-related and socio-cultural debates, and considerably less present around non-controversial crises or humanitarian disasters (K.-C. Yang et al. 2025).

Structural drivers of misinformation vulnerability The domains discussed above share several properties that make them especially susceptible to misinformation. Many are broad, complex and difficult for non-experts to evaluate. When events unfold quickly or when evidence shifts, people lack stable reference points for evaluating claims, making simple and confident narratives easier to accept. (Prooijen and Douglas 2017; Adams et al. 2023). The early phases of the COVID-19 pandemic illustrate this well. Scientific updates were frequent and sometimes inconsistent, creating confusion that misleading claims could readily exploit (Piltch-Loeb et al. 2023).

These topics are also highly polarising. Issues tied to identity, values or worldviews evoke strong emotional responses, and these reactions shape how new information is interpreted. Research repeatedly finds that people are more receptive to claims that fit their existing views and more critical of those that do not, across both political debates and science-related controversies (Flynn et al. 2017; Lewandowsky, Ecker, and Cook 2017; Van Bavel et al. 2021; Scheufele and N. M. Krause 2019). Topics that evoke such reactions, therefore, create fertile ground for selective acceptance and resistance to correction.

A third factor is novelty. When an issue is new or information is scarce, data voids emerge. People search for explanations, but credible content is limited or absent. In such situations, misleading material faces little competition and can spread quickly, anchoring public understanding before authoritative information becomes available. (Flores-Saviaga et al. 2022; Mannino et al. 2025; Piltch-Loeb et al. 2023) These dynamics make newly emerging or rapidly evolving topics especially vulnerable.

Finally, misinformation clusters where strong incentives exist to influence public belief. Political actors, economic interests and organised groups often have clear goals tied to shaping perceptions, and these incentives drive systematic production, amplification or strategic framing of misleading content. Examples range from long-running campaigns against climate science to coordinated attempts to dominate search

results or reshape online discussions (Lewandowsky, Ecker, and Cook 2017; Tucker et al. 2018; Mannino et al. 2025). Topics connected to such incentives, therefore, attract sustained and targeted misinformation efforts.

3.4.2 Narrative Styles of Misinformation

Alongside topical focus, misinformation is shaped by how claims are narratively constructed. In fast-paced online environments, narrative form strongly influences how messages are interpreted, shared and believed. Misleading content, therefore, relies on recurring narrative patterns that simplify complexity, evoke intuitive responses and stabilise perceived credibility. The following sections outline four narrative styles that commonly structure misinformation online, not as an exhaustive typology but as one analytically useful synthesis of recurring rhetorical patterns.

Emotional–moral framing A first narrative style relies on moral and emotional intensification to trigger rapid, intuitive judgments. Messages foreground discrete emotions such as anger, outrage, pride, enthusiasm or sadness rather than extended argumentation. Empirical studies show that such emotional cues substantially increase diffusion on social platforms, though effects vary by emotion. Analyses of Facebook content during the Brexit referendum find that anger, enthusiasm and pride significantly boost engagement, while fear has little effect (Bil-Jaruzelska and Monzer 2022). Large-scale Twitter analyses similarly demonstrate that moral–emotional language strongly amplifies sharing, with each additional moral–emotional word increasing retweet rates and particularly reinforcing in-group diffusion (Brady et al. 2017).

Emotional framing is especially effective under conditions of low identity visibility or limited attention. Research on COVID-19 misinformation shows that emotionally charged posts spread their effect most strongly among anonymous users, who are more likely to mirror the emotional tone of a discussion (C. Chen et al. 2021). Complementary evidence from Facebook news diffusion indicates that discrete emotions systematically shape engagement patterns, with sadness emerging as a particularly strong driver of sharing and commenting, while positive emotions are more closely associated with lightweight reactions (Choi et al. 2021). More broadly, emotion-based appeals are, on average, more persuasive than purely cognitive ones, especially in collectivistic contexts, indicating that emotional framing operates primarily through intuitive evaluation rather than detailed reasoning (W. J. R. Ng et al. 2024).

Humour and satire frequently appear within emotionally framed misinformation, particularly in meme-based formats. Political memes use humorous simplification, exaggeration and emotional cues to attract attention and engagement, most effectively within ideologically aligned communities (AlAfnan 2025). Related work shows that humour and visual condensation make misleading messages quickly intelligible and emotionally charged, despite limited argumentative depth (Dimitrov et al. 2021; Fiadotava et al. 2023). Experimental studies of political satire further indicate that humorous framing increases enjoyment and absorption while also cueing discounting and counterarguing, resulting in persuasive effects that depend strongly on audience predispositions (Boukes et al. 2015).

Post-truth populist narration A second narrative style can be described as post-truth populist narration. It combines elements of populist communication, such as opposition between “*the people*” and “*elites*”, crisis-oriented storytelling and affective appeals, with a post-truth orientation in which factual accuracy becomes secondary to emotional plausibility and narrative resonance. Rather than presenting coherent arguments, messages rely on short factual fragments anchored in a grain of reality and embedded in suggestive storytelling. Context is selectively omitted, causal relations remain vague, and audiences are invited to supply missing connections themselves. (Drávucz and Kocollari 2023)

A defining feature of this style is the systematic erosion of evidentiary norms. Analyses of Trump’s conspiracy-laden political rhetoric show that claims are introduced through vague hearsay (“*people are saying*”), never fully owned or substantiated and continuously shifted, withdrawn or reframed, such that no stable proposition emerges that could be directly evaluated or fact-checked. At the same time, institutions that would normally anchor truth claims, such as the press, courts or expert bodies, are portrayed as untrustworthy or hostile, encouraging audiences to rely on intuition and suspicion rather than verification. (Neville-Shepard 2019)

This narrative logic creates alternative epistemic environments in which rumours, emotionally charged mini-stories and attacks on expertise displace evidence-based reasoning. Research shows that misleading claims can persist in such settings even when corrective information is available, because competing narratives and ongoing attention keep initial impressions salient and limit belief revision (Lewandowsky, Ecker, and Cook 2017; Graham 2024).

At the message level, post-truth populist narration relies heavily on half-truths. These hybrid factual–fictional fragments foreground a real detail while omitting context or adding speculative elements. They achieve plausibility by resonating with existing emotions and worldviews and resist correction through “*yes, but*” dynamics that keep interpretations open and adaptable. (Gess 2021)

Pseudo-scientific authority cues A third narrative style relies on mimicking the surface features of scientific authority to generate an illusion of credibility. Messages adopt confident, technical language and present claims as expert knowledge while leaving underlying evidence vague or unspecified. Research on so-called “*pseudo-profound*” statements shows that many people judge technical-sounding but semantically empty claims as meaningful and insightful, particularly when they rely on intuitive rather than analytic processing (Pennycook, Cheyne, et al. 2015). Susceptibility to such pseudo-scientific language appears to be a stable individual tendency and is associated with greater receptivity to conspiracy beliefs, alternative medicine and other epistemically weak claims (Iacobucci and Cicco 2022; Pennycook and Rand 2017).

This receptivity is exploited by misinformation that appropriates scientific form without adhering to scientific standards. Studies show that misleading content frequently invokes alternative or unnamed experts, selectively quotes legitimate research, or contrasts allegedly honest outsider expertise with a corrupt scientific mainstream, thereby constructing an appearance of objectivity and empirical grounding without providing verifiable evidence (Hameleers and Goot 2024; Teneva 2021). Analyses of COVID-19 misinformation networks further demonstrate how pseudo-scientific websites, decontextualised studies and institution-like language are mobilised to legitimise claims rejected by the scientific community, with

scientific framing lending durability and credibility to otherwise marginal narratives (Piaia et al. 2023).

Perceived credibility is further increased by stylistic cues such as numerical claims, references to studies and an authoritative tone. Empirical work shows that such technical framing increases the perceived accuracy and persuasive power of misinformation even when evidence is weak, fabricated or entirely absent, with fabricated statistics and invented quantitative claims playing a particularly strong role (Savolainen 2023; Timm et al. 2025).

More generally, sociological accounts show that authoritative discourse can be convincingly produced without corresponding practical expertise. Familiarity with a domain's language and conventions allows speakers to perform expertise independently of genuine competence, enabling technical vocabulary, institutional references and abstract formulations to signal depth and authority while avoiding scientific accountability (Collins 2017).

Repetitive factual-sounding assertions A fourth narrative style relies on repetition and a factual-sounding presentation to cultivate familiarity and perceived truth. A large body of research shows that repeated exposure to a claim increases its perceived accuracy through the illusory truth effect, even after a single prior encounter. Experimental studies demonstrate that repetition reliably raises perceived truthfulness by increasing processing fluency, making people more willing to accept and share claims regardless of their plausibility. (Pennycook, Cannon, et al. 2018; Vellani et al. 2023)

This effect is robust across knowledge levels and content types. Repetition increases perceived truth even when people possess correct information and also extends to opinion statements, because familiarity often guides judgment more strongly than careful evaluation (Fazio et al. 2015; Unkelbach et al. 2019; Riesthuis and Woods 2023). The effect further scales with repetition frequency. Perceived truth rises sharply after the second exposure and continues to increase with additional repetitions, allowing simple, repeated claims to quickly acquire a sense of established validity (Hassan and Barber 2021).

Repetition also contributes to the persistence of misinformation in the face of correction. Misleading claims often continue to shape beliefs after retractions, as familiarity and coherence generated by repetition sustain their influence over time (Lewandowsky, Ecker, Seifert, et al. 2012; Ecker et al. 2022). Numerical claims are particularly resistant, since initial figures act as anchors that bias subsequent judgments even when corrections are encountered, or sources are distrusted (Stubenvoll and Matthes 2022). These dynamics are reinforced by confirmation bias, which stabilises repeated claims within existing belief structures and limits the impact of corrective information (Y. Zhou and Shen 2022).

3.5 Interpreters as the Third Element of Information Disorder

In Wardle's framework, interpreters denote the audiences who encounter messages and actively make sense of them (Wardle and Derakhshan 2017). Audiences are not passive recipients of information but consist of individuals whose interpretations are shaped by socio-cultural position, political orientation and personal experience. Meaning, therefore, does not reside in messages alone but emerges through interpretation, mediated by prior beliefs, identity commitments and social context. As a result, the same message may prompt endorsement, rejection, reinterpretation or silence, depending on how it aligns with an interpreter's worldview and perceived social environment. Importantly, interpreters can themselves

become agents. By choosing whether and how to share a message, they may reinforce, contest or reframe it, feeding interpretation back into the circulation of content and closing the cycle between messages, interpreters and agents.

From an analytical perspective, interpretation operates at two interlinked levels. At the system level, collective patterns of opinion formation and diffusion emerge from repeated interactions between individuals embedded in social networks. At the individual level, interpretation depends on cognitive processing routes, identity considerations and perceived social pressures. The following subsections address these levels in turn. First, established models of opinion dynamics are reviewed as formal representations of how beliefs evolve and stabilise in populations. Second, complementary theories of individual-level interpretation are introduced to explain how messages are cognitively and socially processed within a single interpreter.

3.5.1 Opinion Dynamics as System-level Models of Interpretation

At the system level, interpretation can be formalised through opinion dynamics models, which belong to the tradition of classical agent-based modelling (see Section 2.3), in which agents follow simple local update rules and collective patterns emerge from repeated interaction. This perspective is closely aligned with a statistical-physics approach to social dynamics, which studies how macroscopic regularities arise from the interaction of many simple units (Castellano et al. 2009). Opinion dynamics models examine how local exchanges between interpreters aggregate into population-level outcomes such as consensus, polarisation or fragmentation.

Assumptions Despite their diversity, most opinion dynamics models rest on two shared assumptions (Schawe and Hernández 2022). First, interaction is shaped by homophily: individuals are more likely to attend to, engage with or be influenced by others whose opinions are similar to their own. Second, interaction entails social influence: when individuals do interact, they tend to adjust their opinions in response to those they encounter. When interaction occurs, individuals typically adjust their opinions incrementally rather than fully adopting or rejecting others' views, taking external opinions into account to a limited extent when forming their own.

Discrete and continuous representations of opinion Opinion dynamics models can be divided into Discrete or binary models that represent opinions as categorical states, capturing whether individuals hold one position or another. While analytically convenient, such representations are limited in their ability to describe gradual attitude change. Continuous opinion dynamics models address this limitation by representing opinions as scalar values that can shift incrementally through interaction. (Martins 2008) The central concern of continuous opinion dynamics is how collective opinion configurations emerge from repeated local interaction. Given a population of interacting agents, the models examine which patterns of fragmentation arise and under what conditions special cases such as consensus or polarisation occur. (Hegselmann and U. Krause 2002) Figure 4 provides an overview of this conceptual progression, showing the path from continuous opinion dynamics to bounded confidence and its two canonical formulations.

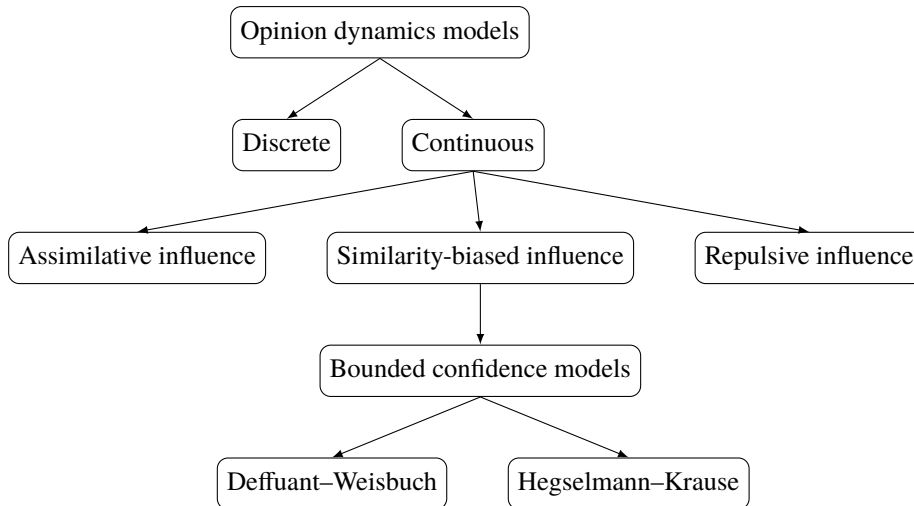


Figure 4: Taxonomy of opinion dynamics models highlighting the conceptual path from continuous representations and similarity-biased influence to bounded confidence and its two most prominent formulations, the Deffuant–Weisbuch and Hegselmann–Krause models.

Assimilative, similarity-biased and repulsive influence Different assumptions about how social influence operates give rise to three classes of continuous opinion dynamics models (Flache et al. 2017). Models of assimilative social influence assume that interaction between connected individuals always reduces opinion differences, which, under sufficient connectivity, leads to consensus. Models with similarity-biased influence restrict influence to sufficiently similar others, reflecting mechanisms such as confidence, identity or issue involvement, and can therefore produce stable opinion clusters and fragmentation instead of consensus. Models with repulsive influence further allow interaction with highly dissimilar others to increase opinion distance, capturing backlash effects and enabling persistent polarisation with opposing clusters. Despite these differences, all three classes share a common modelling principle in how opinions are updated through social interaction.

At this abstract level, opinion formation is neither simple adoption nor complete disregard of others' views. Instead, agents assign different weights to the opinions they encounter and update their own position accordingly. Opinion dynamics are therefore modelled as an averaging process that is iterated over time. While such averaging might intuitively be expected to drive opinions toward consensus, even simple formulations can give rise to complex collective behaviour once interaction weights depend on time or on the current configuration of opinions. (Hegselmann and U. Krause 2002)

General model of continuous opinion dynamics To formalise continuous opinion dynamics, consider a group of n agents among whom repeated processes of opinion formation take place. Time evolves in discrete time steps $t \in \mathbb{N}_{\geq 0}$. The opinion of agent i at time t is represented by a real-valued scalar $x_i(t) \in \mathbb{R}$. The vector

$$x(t) = (x_1(t), \dots, x_n(t)) \quad (1)$$

defines the opinion profile of the population at time t . Opinion updating is governed by weights a_{ij} that quantify how much an agent i takes the opinion of another agent j into account. These weights satisfy

the conditions of non-negativity and normalisation,

$$a_{ij} \geq 0 \quad \forall i, j \quad \text{and} \quad \sum_{j=1}^n a_{ij} = 1 \quad \forall i, \quad (2)$$

so that each agent forms a convex combination of the opinions available to it. The update rule for agent i at time $t + 1$ can then be written as

$$x_i(t + 1) = \sum_{j=1}^n a_{ij} \cdot x_j(t). \quad (3)$$

Crucially, the weights need not be constant. They may depend on time or on the current opinion profile itself, allowing interaction patterns to change as opinions evolve. Collecting all weights into a matrix $A(t, x(t)) = (a_{ij}(t, x(t)))$ yields a row-stochastic influence matrix. Using matrix notation, the general model of continuous opinion dynamics can be written as

$$x(t + 1) = A(t, x(t)) \cdot x(t) \quad \text{for} \quad t \in T. \quad (4)$$

The central analytical question concerns the asymptotic behaviour of the opinion profile generated by this dynamics. Given an initial opinion profile $x(0)$ and a specification of the influence weights as weight matrix A , one asks what can be said about $x(t)$ as $t \rightarrow \infty$. In particular, under which conditions does the population converge to a consensus $c \in \mathbb{R}$, that is,

$$\lim_{t \rightarrow \infty} x_i(t) = c \quad \forall i \in \{1, \dots, n\}, \quad (5)$$

and under which conditions does convergence fail. (Hegselmann and U. Krause 2002)

Opinion fragmentation corresponds to limit configurations in which this condition is violated. Instead of collapsing to a single shared value, the opinion profile converges to multiple distinct values or clusters, or fails to converge altogether. Analysing the limit behaviour of $x(t)$ therefore provides a formal way to distinguish consensus, clustering, polarisation, and persistent disagreement within a unified dynamical framework. Within this general formulation, different modelling assumptions correspond to different specifications of the weight matrix. (Hegselmann and U. Krause 2002)

Bounded confidence models Inside the class of similarity-biased influence models, most attention has been devoted to models operating under bounded confidence. These models implement homophily through an explicit threshold mechanism. Agents interact only with others whose opinions lie within a predefined confidence bound. This reflects the empirical observation that individuals tend to disregard opinions that deviate too strongly from their own. (Lorenz 2007)

Formally, bounded confidence can be introduced as a specialisation of the general model by allowing interaction weights to depend on opinion distance. For this purpose, $\varepsilon_i \in \mathbb{R}_{>0}$, called "uncertainty" or "tolerance", is introduced for each agent i . At time t , agent i only interacts with those agents j whose opinions $x_j(t)$ are close enough to his own opinion $x_i(t)$ by satisfying the condition $|x_i(t) - x_j(t)| \leq \varepsilon_i$.

(Hegselmann and U. Krause 2002)

For a given agent i and opinion profile $x(t)$, the set of agents that are inside the tolerance to influence his opinion is given by

$$I(i, x(t)) = \{1 \leq j \leq n \mid |x_i(t) - x_j(t)| \leq \varepsilon_i\}. \quad (6)$$

Within this set, influence is assumed to be uniform, while all other agents exert no influence. Substituting this specification into the general update rule yields the bounded confidence model,

$$x_i(t+1) = \frac{1}{|I(i, x(t))|} \sum_{j \in I(i, x(t))} x_j(t) \quad \text{for } t \in T, \quad (7)$$

where $|I(i, x(t))|$ denotes the number of agents within the confidence set. (Hegselmann and U. Krause 2002)

Deffuant–Weisbuch and Hegselmann–Krause models Among bounded confidence models, two formulations have received particular attention. The Deffuant–Weisbuch (DW) model (Deffuant, Neau, et al. 2000) applies the confidence set ($I(i, x(t))$) in a pairwise and stochastic manner. At each time step, a randomly selected pair of agents interacts, and opinions are updated only if both agents fall within each other’s confidence bounds. In contrast, the Hegselmann–Krause (HK) model (Hegselmann and U. Krause 2002) implements the same confidence set synchronously. At each time step, every agent updates its opinion by averaging over all agents in ($I(i, x(t))$) simultaneously, exactly as specified in the bounded confidence update rule above. The two models thus differ in interaction structure and temporal organisation, sequential and local in the DW model, parallel and global in the HK model, while instantiating the same similarity-biased influence mechanism.

Extensions of bounded confidence models Beyond the baseline bounded confidence formulation, a range of extensions has been proposed to capture additional aspects of social interpretation. Existing work has examined the role of non-uniform initial opinion distributions, the extension from scalar to multidimensional opinion spaces, heterogeneous confidence bounds across agents, the embedding of bounded confidence dynamics in explicit social networks, and models in which interaction networks coevolve with opinions through rewiring mechanisms (Lorenz 2007).

Among these, two extensions are particularly relevant for understanding information disorder in online environments. First, heterogeneous confidence bounds allow agents to differ in openness, capturing the presence of narrow-minded individuals who are resistant to influence. Even small fractions of such agents can stabilise extreme positions and prevent consensus, providing a mechanism for the persistence of extremism (Deffuant, Amblard, et al. 2002). Second, adaptive bounded confidence models allow opinion dynamics and network structure to coevolve. In these models, agents update their opinions through bounded-confidence interactions while simultaneously modifying their social ties by breaking links to neighbours with sufficiently discordant opinions and forming new connections according to homophily. This coupling of opinion change and adaptive rewiring alters both convergence conditions and long-run outcomes, facilitating the formation of internally coherent groups, slowing or preventing global consensus, and sustaining fragmented or polarised configurations that resemble echo chambers. (Kozma and

Barrat 2008; Del Vicario et al. 2017; Kan et al. 2023)

3.5.2 Individual-level Interpretation and Susceptibility to Influence

While opinion dynamics models capture how beliefs evolve at the collective level, they abstract away from the mechanisms through which individual interpreters make sense of messages and decide how to respond to them. Communication, media and social psychology provide complementary perspectives by focusing on interpretation as a situated process shaped by cognition, identity and social context. Rather than treating exposure as sufficient for belief change, these approaches emphasise that meaning is actively constructed at the point of reception, and that the same message can be interpreted, evaluated and acted upon in markedly different ways. The following perspectives specify key dimensions of this process, addressing how meaning is decoded, how persuasive influence operates under different cognitive conditions, and how social visibility shapes the public expression of beliefs.

Encoding-decoding model Stuart Hall's encoding–decoding model (Hall 1973) offers a clear way of thinking about individual-level interpretation. Hall argues that communication should not be understood as a simple transmission from sender to receiver. Instead, messages are encoded by producers using particular meanings and assumptions, and then decoded by audiences under different social, cultural and experiential conditions. Because these conditions differ, the meaning intended by the sender does not automatically determine how a message is understood. Interpretation is therefore an active process rather than a passive reception. Hall distinguishes three typical ways in which audiences may decode a message. In a dominant reading, the audience largely accepts the intended meaning of the message. In a negotiated reading, the audience broadly understands and accepts the message but adapts it to their own situation or makes local exceptions. In an oppositional reading, the audience understands the intended meaning but rejects it and interprets the message from an alternative perspective. These positions illustrate that the same message can lead to different interpretations depending on prior beliefs, identity and social context.

Elaboration Likelihood Model The Elaboration Likelihood Model (ELM) (Petty and Cacioppo 1986) provides a complementary cognitive account of how individual interpreters process persuasive messages and why the same content can have very different effects across audiences. The core claim of the ELM is that persuasion can occur via two qualitatively different processing routes, depending on an individual's motivation and ability to engage with a message. When motivation and cognitive resources are high, individuals follow the central route and evaluate the quality and relevance of arguments in a relatively effortful manner. Attitude change along this route is driven by reasoning and tends to be more stable, resistant to counterpersuasion and predictive of behaviour. When motivation or ability is low, individuals rely on the peripheral route, in which attitudes are shaped by simple cues rather than argument scrutiny. Such cues include emotional tone, repetition, source attractiveness or perceived expertise. Attitude changes produced via the peripheral route are typically more fragile and context-dependent. The ELM clarifies why emotionally framed, authoritative-looking or frequently repeated misinformation can be influential even in the absence of strong evidence, and why susceptibility to such messages varies systematically across interpreters and situations.

Spiral of silence The spiral of silence theory (Noelle-Neumann 1974) adds a social visibility mechanism to individual-level interpretation. Individuals are assumed to continuously monitor their social environment to estimate which opinions are dominant and which are gaining momentum, using a quasi-statistical sense of the prevailing opinion climate. Because social isolation is perceived as costly, willingness to express a view publicly depends on this perceived climate. If individuals believe their view is widespread or becoming more accepted, they are more likely to speak up. If they believe their view is losing ground, they tend to remain silent. This asymmetry generates a feedback loop in which frequently expressed opinions become increasingly visible and appear even more dominant, while alternative views fade from public discourse and are perceived as marginal. This mechanism helps explain why misinformation can appear broadly accepted despite private disagreement, and why interpreters influence not only belief dynamics but also the visibility conditions that shape subsequent interpretation.

To summarise, individual-level interpretation is neither uniform nor purely cognitive. Messages are decoded through culturally and socially situated frameworks, processed via different cognitive routes depending on motivation and ability, and selectively expressed or withheld in response to perceived opinion climates. Together, these mechanisms explain why identical messages can lead to endorsement, resistance or silence, and why belief change cannot be inferred from exposure alone.

4 Related Work

This section situates PersuaRealSim within prior work on opinion dynamics, social media simulation, and computational models of social-bot influence. The reviewed literature is intentionally restricted to approaches that model opinion dynamics or influence relevant behaviour in social-media-like settings, either as ABM, as generative social simulations, or as social-bot-oriented sandboxes that implement platform interaction environments. While a broader body of work exists on social simulation and learning from belief-change data, this section focuses on studies that combine social interaction, opinion dynamics, and platform abstractions relevant to social media. Within this scope, prior work varies substantially in how beliefs are represented, how agents are designed, how interaction environments are structured, and how influence is operationalised. Accordingly, the section is organised along conceptual and methodological dimensions most relevant to the design choices of PersuaRealSim.

4.1 Stylised LLM-Based Opinion Dynamics Models

A first line of work investigates opinion dynamics in stylised networks of generative agents without recreating full social media platforms. This literature is closely connected to classical opinion dynamics modelling discussed in Section 3.5.1. It operationalises influence through repeated micro-interactions, but typically uses categorical belief states or heuristic update constraints and does not derive initial stances from persona distributions. PersuaRealSim instead treats belief as a continuous stance tied to persona sampling and studies how language-level messaging strategies translate into stance trajectories within a platform environment.

Within this line of work, a first group relies on random peer-to-peer interactions without meaningful network structure, following the DW model introduced in Section 3.5.1. Chuang et al. (2024) simulate ten generative agents that interact through purely random pairwise encounters, role-play simple personas, and express their stance via tweet-like verbalisations that are mapped to five discrete opinion categories by an auxiliary LLM-classifier. Their evaluation focuses on opinion distributions and agent-level trajectories over time, which is also a central evaluation signal in the present thesis. Yuhan Liu et al. (2024) adopt a similar random-interaction design but frame belief change through an epidemiological state-transition logic. Agents equipped with Big-Five-based personas transition between susceptible, influenced, and resistant belief states as the LLM reflects on daily conversations, while an “official spokesperson” agent periodically injects corrective information. Their analysis likewise traces individual and collective opinion trajectories as a central outcome measure. In both studies, however, initial beliefs are not derived from persona attributes, and stance updates result directly from unconstrained LLM prompting. In line with the validity concerns discussed in Section 2.8, J. Zhou et al. (2025) show that the belief convergence reported by Yuhan Liu et al. (2024) is largely induced by explicit instructions to exhibit confirmation bias, violating the *Minimal-Control* principle and pointing to prompt-driven rather than emergent belief dynamics.

A second group incorporates explicit network topologies and recommendation-driven interactions to move closer to social-media-like conditions. C. Wang et al. (2025) place generative agents in a structured graph and use a recommendation mechanism to select interaction partners, combining LLM-based stance updating with a bounded-confidence constraint that determines whether influence is allowed. Agents hold

discrete opinion values ranging from -2 to 2 , but initial beliefs are drawn from uniform distributions and are not informed by persona characteristics. *MTOs*, proposed by Zuo et al. (2025), extends this idea to multi-topic environments. Agents embedded in a scale-free network hold belief vectors over correlated or uncorrelated issues, select topics based on LLM-generated recommendation sets, and update each belief dimension through a dynamic decay mechanism. This design enables fine-grained analysis of echo chambers across multiple topics, yet in both approaches, initial stance assignment remains uninformed by persona attributes and belief updating remains heuristic.

A third group explores hybrid LLM-ABM systems that mix classical ABM agents with a smaller number of generative agents for computational efficiency. J. Yao et al. (2025) propose the *FDE-LLM* framework, where opinion leaders are implemented as generative agents whose attitude changes are constrained by simple rule-based update equations, while follower attitudes evolve through separate numerical transition rules inspired by epidemic-style diffusion models. The simulation includes a concrete action space, such as commenting, reposting, and disliking, bringing it closer to social media behaviour, and evaluates these dynamics against real-world stance trajectories from news events on social media. However, persona construction for opinion leaders is manual and intentionally biased toward provocative behaviour, initial stance values are not derived from persona attributes, and belief change ultimately follows predefined numerical transition equations.

4.2 LLM-Based Social Media Platform Simulations

A second line of work develops generative social simulations that aim to approximate complete social media platforms. These systems prioritise modelling timelines, feed-ranking mechanisms, and user actions such as posting, replying, liking, and following to reproduce platform-level interaction dynamics and behavioural patterns, including exposure effects and network evolution. In this literature, belief or opinion states are typically absent, implicit, or secondary to the study of platform mechanics themselves. *PersuaRealSim* adopts similar platform abstractions, but treats the simulated platform primarily as a controlled environment for analysing stance dynamics and narrative strategies, with explicit belief representations grounded in persona distributions.

Törnberg et al. (2023) present one of the earliest Twitter-like generative social simulations by combining US survey-based demographic data with generative agents that post, like, and comment on real news items under alternative feed-ranking algorithms. Agents act according to their media habits and engagement profiles, and the simulation evaluates how *popularity-based*, *global* and *bridging* ranking strategies shape conversational tone and cross-ideological exposure. Larooij and Törnberg (2025a) extend this line of work with a minimal Twitter-like environment in which LLM agents post, repost, and follow on asynchronously updated timelines that blend followed content with high-engagement global posts, focusing on the effects of prosocial design interventions on emerging network structure. The set of possible agent actions and the timeline logic are closely aligned with those of *PersuaRealSim*, and both systems directly informed the feed-ranking and timeline construction mechanisms implemented in this thesis, despite not modelling explicit stance values.

A more modular and generalised class of platform simulations is introduced by Rossetti et al. (2024) with *Y-Social*, a digital-twin framework that replicates a full social-media platform using generative agents

for both decision-making and content generation. Agents interact through open-ended platform actions produced via persona-conditioned prompts, an approach that is conceptually close to PersuaRealSim’s action-decision-prompting mechanism. This is complemented by a configurable simulation *recipe* mechanism that specifies environment parameters and agent attributes in a manner analogous to the run-level configurability used in this thesis. C. Gao, Lan, Z. Lu, et al. (2025) propose S^3 , a simulation environment grounded in real social media data, where agent demographics, network structure, and perceived information environments are inferred from observed posts and profiles, and LLMs generate decisions such as reposting, posting new content, or remaining inactive based on persona and recent exposure. While both systems share PersuaRealSim’s emphasis on persona-conditioned action generation and platform mechanics, neither models stance trajectories as an explicit continuous state that is updated through an externally validated belief-update model.

Touzel et al. (2024) propose a manipulation-oriented social simulation designed as a sandbox for studying AI-driven influence under controlled experimental conditions. The system embeds generative agents in a social-media setting and evaluates population-level outcomes such as political preference and election results using longitudinal surveys administered throughout the simulation. In contrast to PersuaRealSim, belief change is generated internally through repeated agent self-reporting within survey prompts, without an externally validated, message-level persuasion model.

Z. Yang et al. (2025) present *OASIS*, a scalable and platform-agnostic simulation environment capable of running up to one million agents across platforms such as Twitter and Reddit. The system embeds generative agents within dynamic networks and platform-specific recommendation mechanisms to reproduce macro-scale phenomena, including cascades, herd behaviour, and polarisation. *OASIS* prioritises scale and breadth over fine-grained belief modelling. Personas are not grounded in survey-based distributions, and opinions are not represented as continuous stance values.

4.3 Task-Oriented Generative Social Simulations

A third line of work also embeds generative agents in social-media-like environments, but organises the simulation around specific analytic tasks rather than around platform realism alone. Typical targets include opinion leadership, political mobilisation, social movements, or rumour propagation. While these tasks are closely related to belief influence, the modelling emphasis often shifts toward task-specific control signals, role assignments, intervention designs, or large-scale dynamics. As a result, belief change is frequently treated as an intermediate or instrumental variable, rather than as the primary object of analysis. PersuaRealSim differs by using a social-media-like environment not to optimise or intervene in a particular task, but to isolate how narrative strategies and exposure patterns translate into continuous stance trajectories.

B. Jin and W. Guo (2025) model a Twitter-style environment in which a single opinion-leader bot learns an influence strategy through an external optimisation process, while an LLM generates the corresponding tweet-like messages. Opinions are inferred from text using conditional perplexity against handcrafted pro and contra template statements and discretised into five categories. Follower relations evolve through alignment and divergence rules rather than LLM-mediated decisions. This setup shares with PersuaRealSim the idea of tracking opinion values over time, but differs in that decision-making is not fully LLM-

driven and belief inference relies on template-based scoring.

Mou et al. (2024) propose a social-movement simulator realised as a hybrid LLM–ABM system. Core agents inhabit a Twitter-like environment with profile, memory, and action modules and interact on timelines enriched with offline news, closely mirroring PersuaRealSim’s interaction loop. Profiles and initial attitudes are drawn from real Twitter data, making initial opinions explicitly persona-conditioned. Each user is assigned a behavioural role, such as idea starter, amplifier, or viewer, which constrains participation in the movement. The attitudes of ABM agents evolve through predefined aggregation equations, with content from core generative agents translated into stance scores by external tools.

Shirani and Bayati (2025) present *LLM-SocioPol*, a generative social simulation environment that uses US-census-based demographic profiles and an authentic Twitter follower network to study how agents’ likelihood of voting in an upcoming election changes under different mobilisation messaging strategies. Agents log in intermittently, receive personalised feeds, and may like, reply, follow, or unfollow while updating their voting propensity on a discrete scale. Experimental conditions are injected directly into each agent’s prompt in the form of mobilisation cues and aggregate peer-intention statistics. This makes the treatment visible to the LLM and risks experimenter effects. Agents are also exposed primarily to summary signals rather than to a fully interactive social environment.

Yijun Liu et al. (2025) introduce *RumorSphere*, a large-scale social simulation realised as a hybrid LLM–ABM system for studying rumour propagation. A small set of LLM-driven core agents operate in a Twitter-like environment with persona, memory, and action modules and hold continuous opinion values, while a much larger population of regular agents follows scripted numerical update rules. A dynamic promotion mechanism elevates selected regular ABM agents to generative agents when their local environment exhibits information conflict, enabling large-scale rumour cascades. Personas are further enriched through LLM-inferred interests seeded from demographics, an approach that was initially explored in PersuaRealSim but later removed to keep prompts compact. *RumorSphere* focuses on rumour-specific belief trajectories and intervention timing, and prioritises scale through the division between core and regular agent types.

4.4 LLM-Powered Bot and Influence Models

A fourth cluster of research uses LLMs to model adversarial or strategic influence in social-media-like settings, often in the context of disinformation, information operations, or coordinated narrative manipulation. Compared to task-oriented simulations, this line of work moves explicitly into the domain of information disorder and malicious influence. The primary focus is commonly on bot realism, attack strategy design, or the construction of evaluation assets such as benchmarks and testbeds. Consequently, these systems typically do not prioritise the construction of a full generative social simulation in which belief dynamics emerge from ordinary multi-agent interaction. PersuaRealSim intersects with this literature through its use of LLM-powered bots, but differs by embedding them in a complete social environment and analysing influence through population-level belief trajectories.

L. H. X. Ng and Carley (2025a) study the realism of LLM-powered social bot networks by generating personas, posts, and interaction graphs and benchmarking linguistic, metadata, and network properties against large empirical datasets of real bots and human users. Their analysis shows that LLM-generated

networks and content systematically differ from observed social-media behaviour. This line of work is complementary to PersuaRealSim because it frames realism as similarity to empirical bot and human activity for detection and effectiveness analysis, whereas PersuaRealSim uses LLM bots as instruments for controlled experiments on belief change under an externally grounded persuasion mechanism.

Kera and Cavallon (2025) develop an information-operations sandbox in which red-team bots attempt to influence a synthetic population, while blue-team agents represent defensive or counter-messaging interventions. Red-team behaviour is largely specified through prompt constraints inspired by observed inauthentic accounts, including breaking-news framing, emotive fear appeals, and fabricated research claims, rather than emerging from autonomous multi-agent interaction. While this work demonstrates the feasibility of narrative steering for LLM-powered social bots, PersuaRealSim differs by omitting explicit defence agents and evaluating bot strategies through stance trajectories rather than subjective realism assessments or ad-hoc behavioural criteria.

Marez Oyens and Amghane (2024) simulate information operations in a synthetic conflict scenario where personas and posts are generated via few-shot prompts with steering variables for stance, sentiment, and emotion. Personas are synthetic and lack demographic grounding, and behavioural traits such as trolling are introduced probabilistically. External events from a simulated battlespace are injected to observe downstream narrative shifts, with evaluation relying on expert judgement. The similarity to PersuaRealSim lies primarily in the idea of exogenous events entering a synthetic social network and triggering narrative dynamics. However, agents do not engage in autonomous interaction cycles, do not receive structured feeds, and do not update beliefs through a calibrated mechanism.

Low et al. (2024) present *Diluvision*, an agent-based influence model that emphasises indirect effects of co-ordinated behaviour through engagement dynamics. The system reproduces realistic engagement patterns and supports a wide range of influence strategies, but it relies on fixed stance categories, deterministic agent roles, and hand-crafted behavioural rules. While *Diluvision* provides insight into structural coordination effects, PersuaRealSim addresses a complementary question by modelling how narrative variation in LLM-generated messages produces graded belief change through empirically calibrated persuasion effects.

Pastor-Galindo et al. (2024) propose a high-level LLM-powered framework for studying mis- and dis-information, outlining conceptual layers for agent specification, interaction, and evaluation. While the framework highlights research directions such as generative social networks, misinformation scenarios, and countermeasure analysis, it remains conceptual and does not implement an operational social-media simulation. PersuaRealSim operationalises these ideas by instantiating a full Twitter-like environment with explicit agent architectures, feed-ranking mechanisms, and continuous stance trajectories.

4.5 Agent-Based Models of Social Bots and Opinion Manipulation

A long line of pre-LLM computational social science work studies social bot influence on opinion dynamics using classical ABMs as introduced in Section 2.3. These models study social bots on social media platforms through structural and behavioural rules rather than through generative agents or natural-language interaction. They are conceptually close to PersuaRealSim in their focus on bot-driven opinion manipulation, but they operationalise beliefs and influence through handcrafted mechanisms.

Yaozeng Zhang et al. (2024) adopt an attacker-centric perspective and study how bots should position themselves and move within a network to maximise influence. Their model shows that timing, network position, and movement strategies are more consequential than bot volume for shaping opinion climates, which are computed from feature-word frequencies rather than from linguistic interaction. Their contribution lies in isolating structural influence strategies under fixed behavioural rules, whereas PersuaRealSim holds network structure constant and examines how messaging affects stance trajectories.

Ross et al. (2019) and C. Cheng et al. (2020) study social-bot influence through spiral-of-silence dynamics. In both models, agents hold numeric or binary opinions and decide whether to express them based on individual willingness-to-speak thresholds, allowing persistent minority actors to suppress dissent and shift the expressed majority opinion. Ross et al. (2019) demonstrate that even 2–4% manipulative actors can alter majority expression in polarised settings on static networks. C. Cheng et al. (2020) extend this mechanism to small-world and scale-free networks, showing that a small fraction of stubborn bots broadcasting a fixed categorical stance can flip majority outcomes. This line of work is particularly close to PersuaRealSim in its use of scale-free network structures, fixed bot stances, and majority-level outcomes, but influence arises through encoding assumptions of spiral-of-silence dynamics directly into the simulation design rather than through unconstrained agent interaction.

Beskow and Carley (2020) present one of the closest pre-LLM ABM counterparts to PersuaRealSim by explicitly modelling Twitter-like platform mechanics, including tweets, retweets, replies, mentions, follower networks, limited attention, and heterogeneous login patterns. Beliefs are represented as continuous values and updated through exposure-based equations that weight incoming messages by the beliefs of message senders, closely aligning with PersuaRealSim’s stance update logic. However, the model captures belief only as a generic susceptibility to disinformation and not as a topic-specific stance.

Averza et al. (2022) simulate a Twitter-like environment using scripted agent types such as neutral users, news accounts, and deceptive bots. Bots always spread misinformation, neutral agents update a scalar belief through a fixed rule, and all agents observe a shared global tweet stream rather than personalised feeds. The study contrasts scenarios with and without deceptive bots to demonstrate their aggregate effect, but does not analyse variation in messaging strategies.

4.6 Summary and Differentiation from Prior Work

Across prior work, social bot influence and opinion dynamics are structured along a small number of recurring design axes. One axis concerns how beliefs are represented, which ranges from categorical labels to continuous scalar values. A second axis concerns how belief updating is implemented, spanning unconstrained prompting, threshold-based interaction rules, and hand-designed numerical update functions. A third axis concerns agent architectures, which range from scripted ABM agents and role-constrained prompting to full LLM decision cycles implementing perception, memory, and action. A fourth axis concerns environment fidelity, which ranges from random interaction schemes to explicit platform simulations with feeds, ranking mechanisms, and action repertoires. A fifth axis concerns scalability and evaluation, which are often addressed through hybrid ABM–LLM populations, interaction abstraction, or internal plausibility checks rather than external validation.

Against this background, PersuaRealSim can be understood as positioned at the intersection of several

strands of prior work. Relative to stylised LLM-based opinion-dynamics models (Section 4.1), it moves beyond small-scale, abstract interaction settings by embedding agents in a platform environment with persistent timelines, ranked exposure, and action-conditioned interaction. In line with platform-oriented generative simulations (Section 4.2), it adopts a Twitter-like environment with explicit feed-ranking and action repertoires, but uses this environment not as an object of study in itself, but as a controlled setting for analysing belief dynamics. Compared to task-oriented generative social simulations (Section 4.3), PersuaRealSim does not optimise for mobilisation, leadership, or intervention success, but instead isolates belief influence as the primary outcome of interest. Finally, while it operates in the thematic domain of information disorder and bot-driven influence addressed by LLM-powered influence models (Section 4.4), it retains the analytical focus on opinion manipulation familiar from classical ABMs of social bots (Section 4.5), transferring this focus from rule-based agents to fully generative social simulation.

Within this landscape, PersuaRealSim positions itself as an end-to-end generative social simulation for studying belief influence in a Twitter-like environment with an explicit validation focus on the core causal mechanism that produces simulation outcomes. It combines survey-grounded persona sampling with continuous stance representations, embeds generative agents in a platform environment with feed ranking and action-conditioned interaction loops, and enables systematic variation of bot narrative styles under otherwise fixed conditions. Crucially, belief change is not generated implicitly through unconstrained prompting or heuristic update rules, but is applied externally via a persuasion judge trained as a supervised listwise RankFormer model on *r/ChangeMyView* data containing human-verified belief change outcomes. This validation strategy differs from prior work, which typically relies on ad-hoc realism checks, surface-level behavioural plausibility, or internal consistency of agent behaviour. To the best of the author's knowledge, no prior simulation framework jointly combines empirically validated persuasion modelling, generative agents embedded in a social-media-like platform, and controlled narrative-style influence campaigns with directly measurable belief dynamics.

5 Method

This chapter describes the methodological realisation of the simulation framework developed in this work. It operationalises the theoretical foundations introduced in Chapters 2 and 3 as an executable system for studying belief manipulation by social bots on a social media platform.

Figure 5 summarises the overall architecture of the simulation and provides an orienting reference for this chapter. Human and bot agents operate within a shared Twitter-like environment that mediates message exposure, network dynamics, and stance change over discrete timesteps. Human agents are instantiated as generative agents with profile, memory, and action modules, while stance evolution is externalised and applied by a learned persuasion judge based on observed messages. The figure mirrors the structure of the method chapter, which proceeds from methodological positioning and persuasion-judge training to agent architectures, environment orchestration, and simulation configuration. The full implementation is publicly available at <https://github.com/JulianBurmester/PersuaRealSim>.

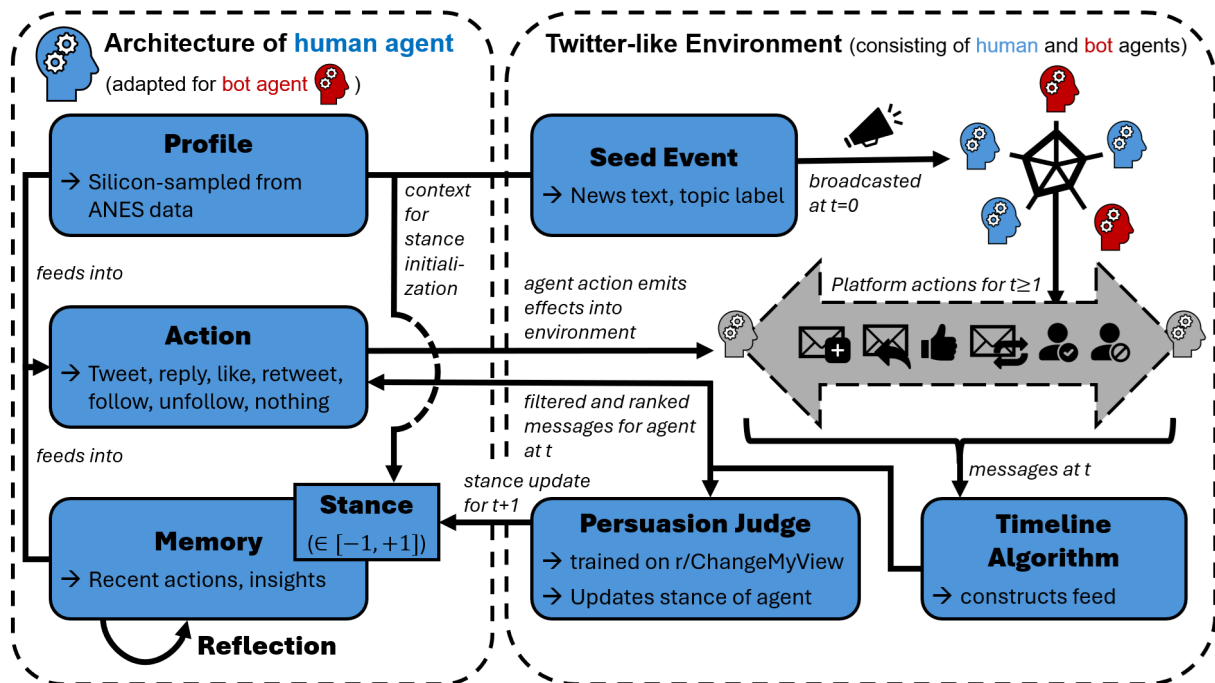


Figure 5: High-level architecture of the PersuaRealSim simulation environment. Human and bot agents interact within a shared Twitter-like environment that mediates message propagation, feed construction, and network dynamics over discrete timesteps. A seed event anchors initial stance formation for human agents. Agents select platform actions whose effects update the global world state, while personalised timelines define agents’ perception. Belief change is applied externally via a persuasion judge trained on *r/ChangeMyView* data.

LLMs are used throughout the simulation for persona enrichment, stance initialisation, action selection, memory-reflection, and the construction of synthetic opinion texts during stance updates. Some cross-cutting prompt design principles are applied consistently. Few-shot examples are avoided, as preliminary experimentation showed that the model often replayed provided examples. This observation is consistent with the insights discussed in Section 2.2. Zero-shot prompts with explicit structural constraints are therefore preferred. Action-decision and stance-initialisation prompts include a short *thought* line, which stabilises persona-consistent behaviour. Prompt design follows a clear division of responsibilities: sys-

Table 3: Conceptual landscape of generative social simulation. The table organises key advantages, challenges and limitations, and validation principles discussed in Chapter 2.

Conceptual landscape of generative social simulation					
Advantages		Challenges and limitations		Validation principles	
[A1]	Behavioural expressiveness	[C1]	Computational cost and scalability	[V1]	Purpose-aligned external validation
[A2]	Persona heterogeneity	[C2]	Opacity and limited behavioural fidelity	[V2]	Micro-level agent design
[A3]	Emergence and exploration	[C3]	Temporal instability and reproducibility	[V3]	Macro-level experimental design
[A4]	Modelling flexibility	[C4]	Persona collapse and distributional misalignment		
[A5]	Ethical sandboxing	[C5]	Prompt sensitivity and experimenter effects		
		[C6]	Limits for causal and confirmatory use		
		[C7]	Ethical and societal risks		

tem prompts specify stable behavioural framing and output constraints, whereas user prompts supply only timestep-specific context. The system is implemented without relying on general-purpose agent frameworks to retain precise control over belief updating, agent state transitions, and environment orchestration. Existing frameworks typically bundle assumptions about planning, memory persistence, and agent autonomy that are well-suited for application development but introduce implicit structure that would interfere with the externally applied, empirically grounded stance-update mechanism central to this study.

5.1 Methodological Positioning

This section positions the proposed method at the intersection of generative social simulation and the domain theory of information disorder. It first motivates the choice of generative social simulation as the methodological paradigm and specifies the associated validation constraints, and then details how information disorder and opinion dynamics are instantiated within this paradigm.

5.1.1 Generative Social Simulation and Validation Constraints

Table 3 provides the conceptual reference frame used to position the methodological choices in this work. Generative social simulation is chosen because the research questions (see Section 1.1) concern belief manipulation by social bots in an ethically sensitive domain, where controlled what-if experimentation with human participants would be infeasible or inappropriate. The simulation, therefore, functions as an ethical sandbox [A5] and supports an explicitly exploratory analysis of population-level belief dynamics rather than individual behavioural trajectories [A3]. At the same time, language-based generative agents extend classical ABMs by enabling richer interaction and greater behavioural variation [A1, A2].

Operationalisation of validation principles The methodological design operationalises the validation principles through concrete design choices. Purpose-aligned external validation [V1] is instantiated through an empirically grounded belief-update mechanism. Human agents do not self-report stance change. Instead, stance updates are applied centrally by the environment using a RankFormer-based persuasion judge trained and calibrated on ChangeMyView (CMV) data, as described in Section 5.2 and operationalised in the stance-update procedure in Section 5.5.3. The corresponding loss behaviour of this component provides an empirical realism signal that contextualises stance-related findings. Micro-level validity [V2] is addressed through the human agent architecture in Section 5.3. Profile construction via silicon sampling (Section 5.3.1), memory and reflection (Section 5.3.2), and action production (Section 5.3.3) operationalise *Profile*, *Interaction*, and *Memory*, aiming to increase behavioural heterogeneity. Macro-level validity [V3] is addressed through orchestration and experimental design choices described in the simulation environment (Section 5.5) and configuration setup (Section 5.6), including partial observability via timeline construction (Section 5.5.2), constrained stance updating (Section 5.5.3), and repeated runs for robustness.

Remaining limitations In theory, [V1–V3] should mitigate key challenges, including opaqueness and limited behavioural fidelity [C2], temporal instability and reproducibility [C3], persona collapse and distributional misalignment [C4], prompt sensitivity and experimenter effects [C5], and parts of the limits for causal and confirmatory use [C6]. It remains an empirical question to what extent the concrete operationalisation in this work effectively achieves these aims, which failure modes persist, and how strongly they affect interpretation. The evaluation, therefore, does not only address the research questions but explicitly probes these remaining limitations to assess how far the proposed method mitigates [C2–C6] and to contextualise all substantive findings derived from the simulation.

5.1.2 Instantiation of information disorder

The type of information disorder investigated in the simulation lies at the intersection of misinformation and propaganda, where misleading or false content is strategically produced and disseminated with the intent to influence beliefs (see Figure 3). Information disorder is modelled on a social media platform because platform affordances fundamentally accelerate the creation, amplification, and circulation of misleading content (Section 3.2). A Twitter-like environment is chosen in particular because empirical analyses identify Twitter as exhibiting the highest volume and amplification of misinformation and disinformation across major social networks (Tomassi et al. 2024), while its short-text interaction format enables scalable processing and generation of messages by generative agents.

Specification of agents and messages Within this setting, the agent element of information disorder consists of human accounts and social bots. Human agents represent the interpreter population, while social bots instantiate automated influence actors. With respect to the analytical dimensions in Table 1, simulated bots are specified as unofficial, fully automated actors targeting the human-agent population, with an intent to mislead and cause harm. They are loosely organised in the restricted sense that they share a common persuasion objective and narrative-style regime, but do not implement explicit coordination or network-level tactics. Motivation is not parameterised and remains topic-dependent. Functionally, the

simulated bots combine properties of social manipulation bots and influence-manipulation bots, aiming to steer discourse and belief dynamics through platform-native actions. Messages constitute the second element of information disorder and are specified with respect to the analytical dimensions in Table 2. In the simulation, messages take the form of textual short posts and replies with single-moment durability, generated as immediate reactions to the currently perceived timeline. Messages are aimed at the human-agent population in terms of persuasive intent. At the same time, actual exposure is mediated by the social graph and feed-ranking mechanisms described in Section 5.5. Two message dimensions are treated as experimental degrees of freedom: topical domain, varied across runs by selecting misinformation-prone issue contexts such as climate change or vaccination, and narrative style, varied independently by assigning bots one of the four styles derived in Section 3.4.2.

Opinion dynamics and stance update mechanism The interpreter component of the simulation builds directly on the agent-based continuous opinion dynamics models introduced in Section 3.5.1. As in this modelling tradition, agents carry a continuous stance that evolves through interaction with their neighbourhood, adopting the core assumptions of homophily and social influence. The central analytical question of these models concerns which opinion configurations emerge at the population level after repeated interaction. This aligns directly with RQ1.2, which asks how different narrative styles influence belief change in the simulated population.

Relative to classical opinion dynamics, interaction is instantiated in a platform-mediated form. The influence an agent receives at a given timestep is determined by an explicit social graph and a feed-ranking mechanism that governs message exposure. By instantiating generative agents rather than rule-based ABM agents, the simulation operates on natural-language messages exchanged within this network, which allows influence to be modelled at the level of individual messages. Stance change is therefore driven by exposure to messages authored by other agents, rather than by direct aggregation over neighbouring agents' stances. Formally, this corresponds to a message-based decomposition of the general continuous update rule in Equation 3, operationalised through the per-message and aggregated stance update equations in Equations 14 and 15, as specified in Section 5.5.3.

Bounded confidence is not imposed via an explicit opinion-distance threshold. Instead, limits on influence arise endogenously through the adaptive social graph. Generative agents can follow or unfollow other agents as part of their action space, thereby expanding or contracting the set of accounts whose messages they are exposed to. Through these follow and unfollow decisions, agents modulate whose messages remain within their effective sphere of influence over time. This setup supports interaction patterns similar to BCMs, without explicitly instantiating one.

Individual-level theories of interpretation and persuasion, such as encoding–decoding, the ELM, and the spiral of silence (Section 3.5.2), inform the conceptual framing of belief change but are not implemented directly within agents. Encoding such cognitive mechanisms into generative agents would introduce circularity by pre-specifying the processes the simulation aims to analyse. The simulation, therefore, does not assume that generative agents replicate human cognitive processing. Instead, generative agents are used to produce and exchange natural-language messages and to enact platform-level behaviour, while belief change is applied externally via a persuasion judge trained on empirical belief-change data.

5.2 Persuasion Judge: Training and Calibration

This section describes the training, optimisation, and calibration of the RankFormer-based persuasion judge. The model is trained on large-scale data from the subreddit CMV to learn relative persuasive strength among competing arguments and is subsequently calibrated to produce belief-update magnitudes suitable for use in the simulation.

5.2.1 r/ChangeMyView as an Empirical Source of Persuasion Signals

The persuasion judge in this work is trained on data from the *r/ChangeMyView* subreddit³. This choice is motivated by the unique supervision signal that it provides. CMV offers large-scale, publicly available instances in which participants explicitly mark persuasive success via self-reported belief change in naturalistic argumentative discourse.

CMV is hosted on *Reddit*, a large online discussion platform organised into topic-specific communities called subreddits. Each subreddit hosts posts and threaded replies, and users can upvote or downvote individual messages. In contrast to most subreddits, CMV enforces explicit norms centred on persuasion: users post an opinion they are open to revising, and replies are expected to challenge that view.

Each CMV thread is initiated by an original poster (OP), who presents an initial claim. Other users submit replies with the explicit goal of changing the OP's view. These replies address the same underlying position and compete as independent persuasion attempts within the same discussion context. If a reply succeeds, the OP may award it a delta (Δ), an explicit marker indicating that the reply caused a belief change, accompanied by a brief explanation of what changed and why.⁴ While deltas can in principle be awarded by any participant, supervision in this work is restricted to deltas awarded by the OP in response to top-level replies, as illustrated in Figure 6. Nested replies primarily serve clarification, rebuttal, or meta-level discussion and are therefore not treated as independent persuasion attempts. Many threads contain no delta at all, even when the discussion is extensive.

Example ChangeMyView-Thread

Original post (by OP)

I believe that everything is subjective, contextual, and relativistic.

Every argument can be scrutinised until it is less than 100% correct. Since nothing can ever be fully certain, I conclude that there are no moral absolutes. Change my view.

Top-level reply

Consider Descartes' statement "*I think, therefore I am.*" While many things are subjective, the existence of thinking itself cannot be denied. This provides at least one objective truth.

Delta award by OP to reply (Δ)

Interesting. Thinking this through reinforces the point. I now have one absolute truth to build my worldview on. Δ

Figure 6: Schematic example of a CMV thread illustrating the original post, a top-level reply, and a delta (Δ) award with an explanation by the OP. Text paraphrased for illustration (Source: https://www.reddit.com/r/changemyview/comments/1e5dod/i_believe_that_everything_is_subjective/).

³ <https://www.reddit.com/r/changemyview/>

⁴ <https://www.reddit.com/r/changemyview/wiki/rules/>

Taken together, these conventions define a thread-level supervision signal that extends beyond individual reply pairs. A CMV discussion forms a reply tree rooted in the original post, with replies unfolding across multiple depth levels. Delta awards may occur at any depth if the OP responds to a reply with a delta and an accompanying explanation. For the purposes of training a ranking-based persuasion judge, however, this hierarchical structure is reduced to a set of competing top-level replies. Figure 7 illustrates both the original reply structure of a CMV thread and its conversion into a listwise target distribution based on delta awards and engagement signals.

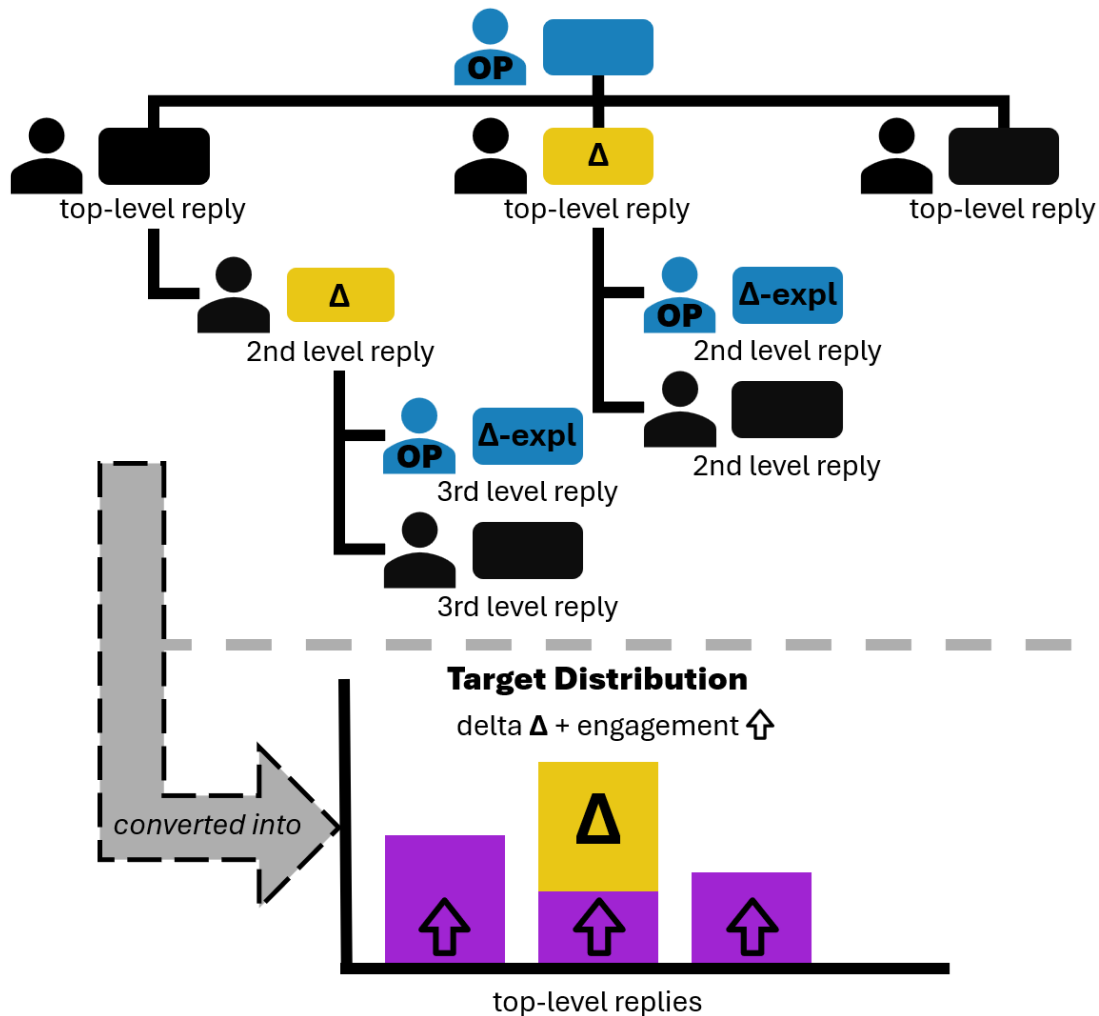


Figure 7: Typical CMV thread structure and the construction of listwise training targets used in this work. The upper section shows an original post with a tree of replies across multiple depth levels. A reply can receive a delta award (Δ) if the OP responds with a delta (gold) and a short explanation. Threads may contain zero, one, or multiple delta awards. The example illustrates two delta awards. For training, supervision is restricted to delta awards on top-level replies. The lower section illustrates how the set of top-level replies is converted into a thread-specific target distribution by combining the binary delta signal (gold) with an engagement signal (purple) derived from upvotes.

5.2.2 Training Data Construction

The training data is derived from two publicly released corpora of CMV discussions that provide explicit supervision via delta awards. The first source is the *Winning Arguments Corpus* introduced by Tan et al.

(2016) and distributed through *ConvoKit*⁵. It covers CMV threads from 2013 to 2015 and includes meta-data that identifies which replies belong to successful versus unsuccessful argument threads, together with the original paired comparisons used in the dataset construction. The second source is the *Webis Change-MyView Corpus 2020* (Webis-CMV-20)⁶ released by Al-Khatib et al. (2020), which extends coverage to the full subreddit history from 2005 to 2017 and provides complete discussion trees with delta awards embedded in the reply structure. In contrast to the Winning Arguments corpus, Webis-CMV-20 is not limited to pre-constructed reply pairs but exposes all replies in a thread, which allows delta labels to be reconstructed by identifying which replies received a delta from the OP.

In this work, both corpora are converted into a unified listwise format in which each thread yields a ranked set of top-level replies, and each reply is assigned a binary label indicating whether it received a delta. The datasets are merged primarily to maximise the amount of available training data while retaining the same underlying supervision mechanism. Direct collection of comparable CMV data via the Reddit API is no longer a viable alternative due to platform restrictions on large-scale data access⁷ and explicit prohibitions on using Reddit content for AI model training⁸.

Post-level filtering Before constructing listwise training instances, posts are filtered to remove content that is specific to Reddit as a platform or to CMV as a moderated community. The persuasion judge is later used in a general simulation context that abstracts away from subreddit rules, bots, and moderation practices.

To avoid learning artefacts tied to platform-specific mechanisms rather than to persuasion itself, posts that primarily serve administrative, logistical, or meta-level purposes are excluded. This includes, for example, moderator announcements and recurring weekly coordination posts that do not present an argumentative position. Posts authored by system accounts, such as u/AutoModerator, or posts that explicitly reference moderation procedures or u/DeltaBot mechanics, are likewise removed. Beyond these semantic exclusions, original posts must satisfy a minimum content requirement after boilerplate, such as moderator footers or removal notices, has been stripped. Empty, deleted, or too short posts are discarded, as they do not present a sufficiently articulated position against which persuasion can be meaningfully assessed. The full set of concrete post-level exclusion and stripping rules is documented in Appendix A. All filtering steps are applied identically to both CMV data sources before reply selection.

Reply-level filtering Within each retained thread, candidate persuasive replies are restricted to direct, top-level replies to the original post. Only these replies address the OP’s stated position explicitly and constitute independent persuasion attempts. Deeper replies are excluded, as they primarily reflect follow-up discussion or intra-thread dynamics rather than initial persuasive appeals. At the reply level, the same principle of removing platform-specific and procedural content is applied. Replies authored by system or bot accounts, including u/AutoModerator and u/DeltaBot, are discarded, as are replies that serve logistical or administrative purposes, such as moderation notices or coordination messages. Remaining replies undergo text normalisation, including the removal of CMV-specific moderator footers and deletion

⁵ <https://convokit.cornell.edu/documentation/winning.html>

⁶ <https://zenodo.org/records/3778298>

⁷ https://en.wikipedia.org/wiki/Reddit_API_controversy

⁸ <https://support.reddithelp.com/hc/en-us/articles/14945211791892-Developer-Platform-Accessing-Reddit-Data>

artefacts. Finally, a minimum content requirement is enforced at the reply level. Replies that are empty, deleted, or too short after cleaning are excluded, ensuring that all retained replies contain a substantive argumentative contribution suitable for persuasion ranking.

Training instance construction After post- and reply-level filtering, each retained thread is converted into a single listwise training instance. For a given original post, all remaining top-level replies are treated as competing candidates within the same ranking context. Each reply is assigned a binary supervision label indicating whether it received a delta from the OP. Only discussions with at least two eligible replies are retained, as a minimum of two candidates is required to define a meaningful ranking task. Threads in which a delta is awarded to a second-level reply or any deeper reply are excluded, as in these cases it is unclear whether the delta reflects agreement with a top-level argument, a clarification within a follow-up exchange, or a newly introduced argument. Restricting supervision to deltas awarded directly to top-level replies ensures that persuasive success can be unambiguously attributed to the reply text used for training.

To control list length, reply sets are capped at 24 entries, always retaining all delta-winning replies and limiting the number of non-winning replies accordingly. In addition to the binary delta label, each reply retains its score, defined as the difference between upvotes and downvotes at collection time. This score serves as a weak crowd-based signal to order non-winning replies and to support the estimation of smooth persuasiveness distributions within listwise instances.

Merging heterogeneous CMV sources After listwise construction, instances from the Winning Arguments Corpus and Webis-CMV-20 are merged by Reddit thread identifier, with duplicates resolved by retaining the Webis version due to its more complete discussion structure and delta reconstruction. Threads that appear in only one source are included unchanged. The resulting merged dataset comprises 46,846 unique threads in total, combining 3,042 threads from the Winning Arguments Corpus and 46,097 threads from Webis-CMV-20, with an overlap of 2,293 threads between the two sources.

In the next step, the merged dataset is split into training, validation, and test sets using a single stratified 80–10–10 split based on whether a thread contains at least one delta-winning reply. This preserves the relative proportion of persuasive and non-persuasive threads across all splits. A single held-out split is used instead of cross-validation, as k-fold training would be computationally prohibitive for the RankFormer model and unnecessary given the dataset scale.

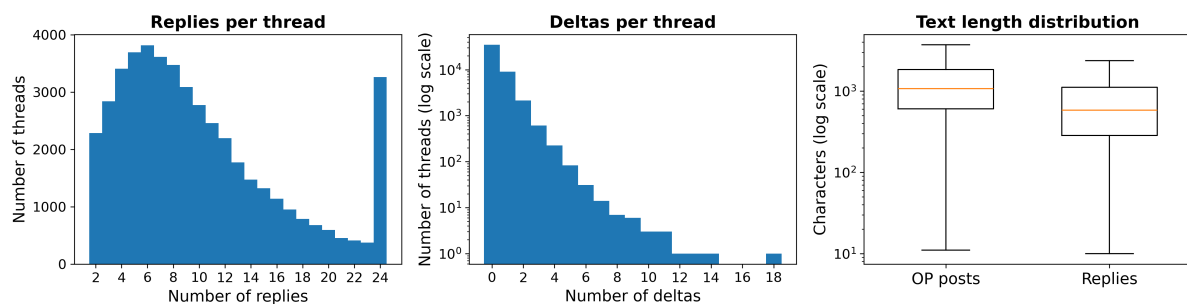


Figure 8: Structural and textual characteristics of the merged CMV dataset. The figure shows the distribution of replies per thread, the distribution of delta awards per thread (log-scaled), and the distribution of text lengths for original posts and replies (log-scaled).

Training dataset characteristics The resulting dataset comprises 46,846 CMV threads after filtering and merging. Threads contain, on average, 10 replies, with a median of 9 replies and a maximum of 24 replies due to capping. Persuasive outcomes are observed in 26% of threads, corresponding to 12,173 threads in which at least one delta is awarded, with an average of 0.36 delta awards per thread and a long-tailed distribution of delta counts. Original posts have a median length of 1,072 characters, while replies have a median length of 583 characters after cleaning and filtering. Figure 8 summarises these structural and textual characteristics of the final dataset. Overall, the pipeline yields a large-scale collection of clean, thread-conditioned, listwise persuasion judgments suitable for training and evaluating a ranking-based persuasion judge.

5.2.3 RankFormer Modelling

Building on this large-scale listwise CMV dataset, the persuasion judge is implemented as RankFormer, a listwise learning-to-rank model that scores items in the context of competing alternatives within the same list, rather than independently (Buyl et al. 2023). The original RankFormer setting assumes web and e-commerce applications in which users are shown a limited selection, such as search results, and provide implicit feedback on the list. This maps naturally to CMV threads where an original post defines the shared context and top-level replies represent alternative persuasion attempts. A key motivation is that RankFormer jointly optimises a listwise ranking objective and a listwise quality assessment, allowing the model to represent that an entire presented list can be low quality and effectively rejected. In the CMV setting, this aligns with threads without any delta award, where the OP does not accept any reply as persuasive, even if replies can still be ranked relative to each other.

Model architecture Figure 9 summarises the RankFormer architecture. RankFormer extends the listwise Transformer idea, in which a Transformer encoder processes an entire list of items jointly so that each item representation can attend to and be contextualised by all other items in the same list, rather than being scored independently (Pang et al. 2020). Formally, in the CMV setting, each discussion thread I is represented by a feature matrix X_I , where each row corresponds to one reply, and the matrix is augmented with a learnable list token $x_{[\text{CLS}]}$. The extended input $X_I^* = X_I \cup x_{[\text{CLS}]}$ is processed by the listwise Transformer encoder TF , yielding contextualised item representations Z_I and a list representation $z_{[\text{CLS}]}$. Reply scores $s(X_I)$ are produced by applying a shared scoring head h_s to each contextualised reply representation, concatenated with $z_{[\text{CLS}]}$ and trained with the listwise ranking loss L_y . In parallel, the list token representation is passed through a separate head h_d to produce the listwise prediction $d(X_I)$, which is trained against the listwise label t_I via L_t . (Buyl et al. 2023)

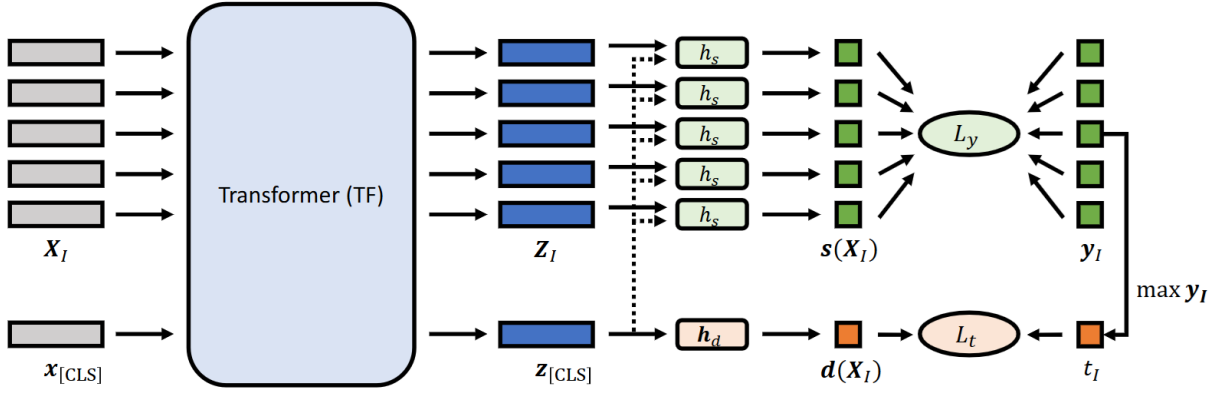


Figure 9: RankFormer architecture. Item features X_I and a learnable list token $x_{[\text{CLS}]}$ are processed jointly by a listwise Transformer TF . The resulting item representations Z_I are combined with the list token representation $z_{[\text{CLS}]}$ and scored by a shared head h_s to produce listwise ranking scores optimised with L_y . The list token is additionally passed to a separate head h_d to predict a list-wide quality score optimised with L_t (Source: Buyl et al. 2023).

Input construction Each row of X_I is obtained from a Transformer-based text encoder operating at the reply level. For each reply, the original post and the reply text are encoded jointly using a pretrained DeBERTa model (P. He et al. 2021) in a cross-encoder configuration, and the resulting classification token embedding is used as the reply-level feature vector. Inputs to the text encoder are capped at a maximum sequence length of 512 tokens, with a fixed token budget split between the original post and the reply and truncation applied when necessary. In the merged CMV dataset, original posts are longer than replies, with median lengths of 266 and 134 tokens, respectively. At a 512-token cutoff, truncation affects approximately 20.5% of original posts and 7.1% of replies when considered individually, and 42.0% of joint original-post-reply pairs. Truncation, therefore, primarily affects unusually long original posts and reflects a trade-off between preserving argumentative context and computational efficiency. Based on this analysis, the final training configuration allocates a maximum of 320 tokens to the original post and 192 tokens to the reply to ensure that the combined input respects the 512-token limit while prioritising the typically longer original post.

For computational stability and to bound the cost of listwise attention, the number of replies per thread is capped during RankFormer training. While preprocessing retains up to 24 replies per thread (see Section 5.2.2), the model operates on lists of at most 16 replies, corresponding to roughly the upper 16% of threads by reply count. All delta-winning replies are always retained, and remaining slots are filled with the highest-scoring non-winning replies. The listwise Transformer is implemented with two encoder layers and four attention heads and does not employ positional embeddings, making it permutation-equivariant with respect to reply ordering, consistent with the RankFormer design (Buyl et al. 2023).

Supervision Each CMV thread corresponds to one listwise ranking instance I . Item-level supervision y_I reflects relative persuasive strength among replies and is derived from a mixture of delta awards and engagement-based reply scores, as illustrated by the target distribution construction in the lower section of Figure 7.

The listwise label is defined as $t_I = \max y_I$ and indicates whether any reply in the thread succeeds in persuading the OP. This listwise signal supplies explicit negative supervision for threads without any delta

award.

Objective Training optimises a joint objective that combines reply-level ranking with a thread-level assessment of persuasive success. For each thread I , the model produces reply scores $s(X_I)$ and a thread-level logit $d(X_I)$. The overall loss is

$$L = L_y(s(X_I), y_I) + \alpha \cdot L_t(d(X_I), t_I), \quad (8)$$

where L_y denotes a listwise ranking loss, L_t denotes a binary classification loss at the thread level, and α controls the influence of the list-level objective.

The ranking loss L_y is formulated as a listwise cross-entropy loss over a soft target distribution y_I that mixes delta-based and engagement-based supervision

$$y_I = \lambda \cdot y_I^{(\Delta)} + (1 - \lambda) \cdot y_I^{(\text{engagement})}, \quad \lambda \in [0, 1]. \quad (9)$$

The engagement component $y_I^{(\text{engagement})}$ is obtained by applying a temperature-controlled softmax to reply-level engagement scores, with a standard max-shift applied for numerical stability. Let r_i denote the engagement score of reply i within thread I . After replacing missing scores and shifting by the maximum for numerical stability, the distribution is

$$y_{I,i}^{(\text{engagement})} = \frac{\exp((r_i - \max_j r_j)/\tau)}{\sum_{k \in I} \exp((r_k - \max_j r_j)/\tau)}, \quad \tau \in \mathbb{R}_{>0}. \quad (10)$$

Larger τ yields a flatter distribution, whereas smaller τ concentrates probability mass on the highest-scoring replies.

The list-level loss L_t is implemented as binary cross-entropy (BCE) and is trained against a thread-level indicator of whether any reply received a delta. To account for the strong class imbalance between threads with and without delta awards, a positive class weight is applied, set to the ratio of negative to positive threads in the training data. Threads without delta awards do not contribute to the reply-level ranking loss but remain active in the list-level objective, providing negative supervision for failed threads. The weighting parameter α mediates the trade-off between fine-grained ordering of replies within successful threads and discrimination between persuasive and non-persuasive threads overall.

5.2.4 Hyperparameter Optimization

RankFormer training involves several interacting hyperparameters that affect supervision semantics, optimisation stability, and convergence behaviour. An exhaustive grid search over all relevant parameters would be computationally prohibitive and difficult to interpret. Instead, a staged hyperparameter optimisation strategy is employed that groups hyperparameters by functional role, separating those that directly shape the learning signal from those that primarily affect optimisation dynamics. This enables a coarse-to-fine search in which semantically meaningful supervision configurations are identified first, and optimisation parameters are refined only within these promising regions.

The two-stage sweep design is summarised in Table 4. Stage 1 evaluates 27 supervision configurations

under fixed optimisation settings. Stage 2 then tunes the learning rate for a small set of K Stage 1 finalists, yielding $3K$ additional runs. All sweeps are conducted with a fixed random seed to isolate the effect of hyperparameter changes.

Table 4: Two-stage hyperparameter sweep design. Stage 1 varies supervision-related hyperparameters while keeping optimisation fixed. Stage 2 tunes the learning rate for a small set of Stage 1 finalists.

Stage	Hyperparameters	Value grid	Runs
1 (semantic)	λ, τ, α	$\lambda \in \{0.6, 0.8, 1.0\}$ $\tau \in \{2, 10, 20\}$ $\alpha \in \{0.0, 0.25, 0.5\}$	$3 \times 3 \times 3 = 27$
2 (LR)	η	$\eta \in \{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}\}$	$K \times 3$

Stage 1: semantic supervision sweep The first optimisation stage focuses on hyperparameters that influence what the model learns rather than how it is optimised. Concretely, the sweep varies the delta-engagement mixing coefficient $\lambda \in \{0.6, 0.8, 1.0\}$ introduced in the construction of the soft target distribution y_I (Eq. 9), the upvote temperature $\tau \in \{2, 10, 20\}$ that controls the sharpness of the engagement-based component $y_I^{(\text{engagement})}$ (Eq. 10), and the list-level loss weight $\alpha \in \{0.0, 0.25, 0.5\}$ that balances reply-level ranking and list-level supervision in the joint objective (Eq. 8). These parameters jointly determine how sparse but high-precision delta signals are combined with dense but noisy engagement cues, as well as the relative emphasis placed on reply-level ordering versus thread-level success prediction.

Rather than redefining supervision targets, this sweep aims to identify supervision regimes under which the model generalises best to unseen threads, measured by agreement with held-out delta labels. To avoid a combinatorial explosion, the learning rate and all other optimisation-related parameters are held fixed during this stage. The resulting sweep explores a moderate grid of supervision configurations under otherwise identical training conditions. Configurations with $\alpha = 0$ are included as an explicit ablation, corresponding to a pure reply-ranking objective without gradients from the list-level loss.

Stage 2: learning rate sweep In the second stage, the learning rate is tuned while holding the supervision-related hyperparameters fixed. This stage is restricted to a small set of K supervision configurations identified as stable and competitive in Stage 1, resulting in $3K$ runs in total. The learning rate is varied over $1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}$. By varying only the learning rate, this sweep isolates optimisation effects and avoids confounding changes in the supervision signal. All other training parameters, including batch size, gradient accumulation, weight decay, warmup ratio, and number of epochs, are kept constant.

Model selection criterion Hyperparameter selection during both stages is based exclusively on validation performance. The primary selection metric is validation Acc@1 on delta-awarded threads, reflecting whether the highest-ranked reply matches a delta-winning reply. In cases where multiple configurations achieve comparable Acc@1, the validation loss of the list-level head is used as a secondary tie-breaking criterion. No test-set results are consulted during hyperparameter selection. The final configuration is chosen only after completion of both optimisation stages and is evaluated separately on the held-out test

split.

5.2.5 γ -Calibration

After training and selecting a high-quality RankFormer model, the remaining challenge is to translate its persuasion outputs into belief-update magnitudes suitable for use in the simulation. RankFormer is optimised to rank replies and to estimate whether a thread is persuasive, but it does not specify how strongly an agent’s belief should change when exposed to persuasive content. To bridge this gap between a ranking-based persuasion model and the dynamic belief update mechanism described in Section 5.5.3, a single global scaling factor γ is introduced. γ calibrates the magnitude of stance updates produced by the persuasion judge, allowing RankFormer’s relative persuasion signals to be interpreted as empirically plausible belief shifts within the simulation.

Calibration data Calibration is performed on a curated subset of 1,500 CMV threads drawn from Webis-CMV-20. Selected threads contain exactly one awarded delta and a substantive explanation comment written by the OP, providing a clear textual trace of belief change before and after persuasion. Webis-CMV-20 is used exclusively for this step because it preserves the OP’s delta-award explanation comments in a recoverable form. While the Winning Arguments Corpus identifies successful replies, it does not expose post-delta explanations required to infer the updated stance. Threads with deleted OP texts, multiple deltas, or ambiguous winner structure are excluded to ensure that belief change can be plausibly attributed to a single persuasive interaction.

Stance-based target shifts For each calibration thread, continuous stance annotations in the interval $[-1, 1]$ are obtained for all relevant texts using the same two-stage LLM-based stance extraction pipeline employed for agent stance initialisation in the simulation (see Section 5.5.1). Before stance annotation, each thread is enriched with a neutral topic label and a standalone claim derived from the original post, ensuring that stance is estimated relative to the same reference proposition as in the simulation.

Stance is then annotated for the original post, for each top-level reply, and for the OP’s delta-explanation comment. The pipeline first classifies stance polarity and subsequently estimates stance strength, mapping the result to a continuous scale. No reasoning traces are generated at this stage. The empirical target belief shift for a thread is defined as the difference between the post-delta stance and the pre-delta stance of the OP.

Model-implied persuasion signal For the same calibration threads, the RankFormer persuasion judge is applied to the original post and its top-level replies to obtain per-reply persuasion weights. These weights implicitly capture both thread-level persuasiveness and the relative convincingness of individual replies. Using the stance annotations described above, an unscaled model-implied belief shift is computed as

$$\Delta_{\text{raw}} = \sum_i w_i \cdot (s_i - s_{\text{before}}), \quad (11)$$

where w_i denotes the RankFormer weight of reply i , s_i its stance, and s_{before} the OP’s initial stance. This quantity represents the belief change implied by the trained persuasion model before any magnitude

calibration.

Estimation of the scaling factor The global scaling factor γ is estimated by aligning the magnitude of the model-implied persuasion signal with empirically observed stance change. For each calibration thread, the unscaled model shift Δ_{raw} is computed as defined above, and the empirical target shift is given by $\Delta_{\text{target}} = s_{\text{after}} - s_{\text{before}}$. The scaling factor is then estimated as

$$\gamma = \frac{\mathbb{E}[|\Delta_{\text{target}}|]}{\mathbb{E}[|\Delta_{\text{raw}}|]}, \quad (12)$$

where expectations are taken over all usable calibration threads.

The resulting factor is fixed after calibration and applied uniformly in the simulation to scale RankFormer’s persuasion signal. Appendix B provides a fully annotated calibration thread that instantiates this computation, including the enriched topic and claim representation, stance annotations for the original post and replies, the corresponding RankFormer persuasion weights, and the resulting values of Δ_{raw} and Δ_{target} .

5.3 Human Agent Architecture

This section specifies how the abstract generative agent architecture introduced in Section 2.4 is instantiated for human agents in the simulation. Following the modular decomposition into profiling, memory, planning, and action components, the implementation realises explicit modules for profile construction, memory and reflection, and action selection. Each module is designed to operationalise its theoretical counterpart while remaining compatible with the validation principles outlined in Section 2.8.

In contrast to many generative agent architectures, no dedicated planning module is implemented. This design choice follows the *Minimal-Control* and *Unawareness* principles from Section 2.8.3, as an explicit planning component would introduce additional structure that could bias agent behaviour toward experimenter-imposed strategies or objectives. Instead, planning is treated as an implicit aspect of action production, with agents deciding their next Twitter action based on their current profile, memory state, and perceived context at each simulation step.

The resulting human agent architecture consists of three core components: a static, empirically grounded profile constructed via silicon sampling, a dynamic memory system that accumulates and abstracts within-run experience, and an action module that translates the agent’s internal state into observable behaviour. This architecture serves as the baseline for all human agents in the simulation and forms the foundation upon which bot agent architectures are built. Bot agents reuse the same structural components but apply targeted modifications and restrictions to specific modules, as described in Section 5.4.

5.3.1 Profile Construction via Silicon Sampling

The simulated human agent population is grounded in empirical survey data from the *American National Election Studies (ANES) 2024 Time Series Study*. ANES is a long-running US election survey programme designed to support research on public opinion and voting behaviour. The 2024 wave combines a pre-election survey fielded between August 3, 2024 and November 5, 2024, with a post-election re-interview

conducted between November 7, 2024 and February 17, 2025. The full release contains 5,521 pre-election completions and 4,964 post-election re-interviews collected with a mixed-mode design that includes in-person, web, phone, video, and paper questionnaires. (American National Election Studies 2025)

The ANES 2024 Time Series contains more than a thousand variables spanning demographics, socio-economic status, political attitudes, media use, and issue-specific beliefs. As incorporating the full questionnaire into agent personas is neither feasible nor desirable, population construction begins with the selection of a structured subset of variables that define the empirical characteristics of politically active social media users.

Variable selection follows a category-based approach aligned with the dimensions required for persona construction. Specifically, the selected variables cover core demographic attributes, socio-economic position, religious affiliation, political orientation, Twitter usage patterns, and trust in media institutions. These categories capture relatively stable individual characteristics and orientations that plausibly shape how users engage with political content online. All variables are drawn from the public ANES release and are available in harmonised, analysis-ready form. The complete set of variables used for silicon-sampled persona construction, grouped by category and documented with their ANES identifiers and value ranges, is provided in Appendix C.

Before sampling is performed, the ANES respondent pool is restricted according to platform-related inclusion criteria specified via the simulation configuration. In all reported experiments, sampling is limited to respondents who report having used Twitter in the past year and who indicate that they post political content on the platform. Concretely, respondents who did not mention Twitter as a visited social media platform are excluded, as are respondents who report never posting political content when using Twitter. These filters are applied directly to the survey data before sampling, thereby defining the empirical population as politically active Twitter users.

From the filtered respondent pool, agents are sampled using probability-proportional-to-size sampling without replacement (PPSWOR), where selection probabilities are proportional to the post-election survey weights provided by ANES. This procedure ensures that weighted population estimates implied by the survey design are respected while preventing duplicate selection of the same respondent within a single simulation run. The use of PPSWOR further guarantees that each sampled agent corresponds to a unique empirical record, while preserving population-level representativeness under the applied filters.

For each sampled respondent, a base persona text is generated through a deterministic, template-based transformation. Selected survey variables are mapped to short declarative sentences describing the agent's background, political orientation, media use, and institutional trust. The mapping follows a fixed schema in which each variable contributes at most one predefined textual fragment. When multiple survey items capture the same construct, a fixed precedence order is applied, with fallback rules used only if the preferred item is missing.

Missing values and survey-specific non-response codes do not produce textual output. Ordinal and continuous variables are discretised using fixed lookup tables, ensuring that identical survey inputs always result in identical persona texts. All personas follow the same sentence structure and ordering, differing only in content. An illustrative example of a base persona text generated through this procedure is shown

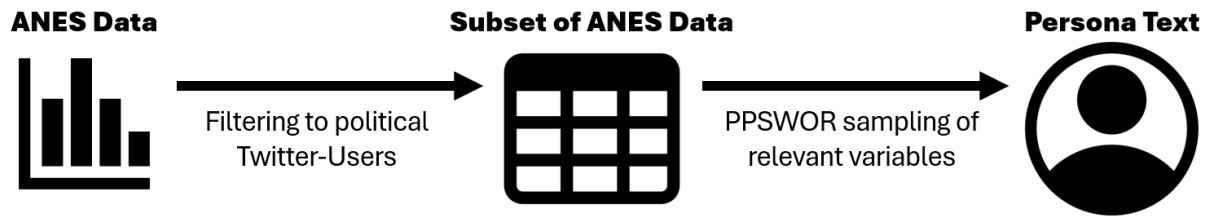


Figure 11: Overview of the silicon-sampling pipeline from ANES survey data to standardised persona texts.

in Figure 10.

Example Persona Text

You are an 80+-year-old woman from Oregon.
 You are White, non-Hispanic and heterosexual.
 You have a bachelor's degree and lower-middle income (\$30,000-\$59,999).
 You are Evangelical Protestant, and religion is extremely important to you.
 You are a strong Republican.
 Ideologically, you describe yourself as conservative.
 Party identity is very important to you.
 You use Twitter about once a day and you sometimes post about politics.
 You do not trust social media at all.

Figure 10: Example of a silicon-sampled persona text built from ANES ID 234862.

To support interaction on a Twitter-like platform, all base personas are enriched using a single LLM call that generates a plausible real name and a Twitter handle consistent with the persona text. This mandatory base enrichment is applied to all agents and is strictly additive, leaving all attributes derived from ANES data unchanged. For the example persona shown in Figure 10, this step yields a name such as Margaret Wilson and a handle like @margwil. A complete prompt–response example of this enrichment is provided in Appendix D.1. This enrichment is strictly additive and does not modify any attributes derived from ANES data.

Optionally, personas can be further enriched with a small set of short bullet points describing everyday interests or habits that plausibly fit the persona. For the example in Figure 10, this could include a statement such as *”You’re a big fan of the Oregon Ducks football team and often attend their games at Autzen Stadium”*, among other interests and hobbies like growing up in rural Oregon, gardening, participating in church activities, or involvement in a Republican women’s group. Prompts are deterministic, enrichment is performed once per agent, and outputs are validated and normalised. The prompt templates used for both the mandatory base enrichment and the optional extended enrichment are documented in Appendix D.2.

All personas follow the same fixed template as in Figure 10 and differ only in the content instantiated from the underlying survey record and optional enrichment steps. An overview of the silicon-sampling pipeline is shown in Figure 11.

5.3.2 Memory and Reflection

Memory represents a dynamic internal state that conditions agent behaviour on past interactions during a simulation run. In line with the distinction between short-term and long-term memory in generative agent architectures (see Section 2.4.2), the implemented memory combines a small in-context component with an explicit, agent-internal long-term store. Recent experiences are surfaced directly in prompts as short natural-language summaries, while older yet relevant information is retained in bounded buffers outside the immediate context window. In contrast to silicon-sampled profiles, which encode static, empirically grounded attributes, memory captures within-run experiences generated by the agent’s own actions, stance updates, and reflections. Memory is implemented as an explicit agent-internal structure, reset at the beginning of each run and stores only information originating from the current simulation.

Three buffer structure Agent memory is organised into three bounded buffers with distinct semantics. A *recent-action buffer* stores short natural-language summaries of the agent’s own most recent actions. In all reported runs, this buffer retains the two most recent actions, corresponding to the maximum number of action summaries injected into downstream prompts. A *stance-direction buffer* stores a single, most recent natural-language summary describing how the agent’s stance has shifted. An *insight buffer* stores higher-level abstractions produced through reflection and is bounded to a maximum of 32 insight items per agent, with older insights pruned when the limit is exceeded.

Memory writing Memory entries are written at explicitly defined points in the simulation loop. After each agent action, an abstracted past-tense summary of the action is recorded, rather than storing the full post text. Stance-direction summaries are written only when belief updates exceed configured thresholds. These include an absolute single-step stance change of at least 0.05, a cumulative absolute change of at least 0.1 over a rolling window of three steps, or a transition into a different stance category. Stance categories correspond to the 0.2 discretisation granularity used for stance labels introduced in Section 5.5.1. Insight entries are generated through reflection every five simulation steps. During reflection, the LLM is prompted with up to two recent action summaries, up to two previously stored insights, and up to two insights that reference user handles visible in the current feed. At most two new insights are requested per reflection cycle. Figure 12 summarises the three-buffer memory structure, the conditions under which each buffer is updated, and how selected memory contents are injected into the action-decision and memory-reflection prompts.

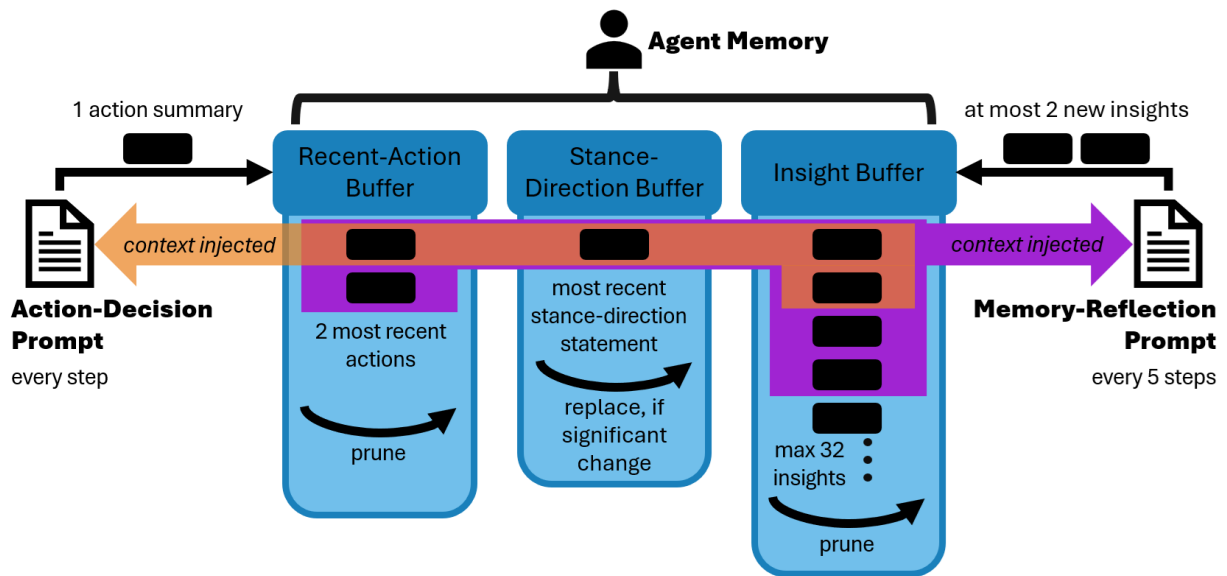


Figure 12: Overview of the agent memory architecture. Memory is organised into three bounded buffers for recent actions, stance-direction updates, and reflective insights. The figure illustrates how buffers are updated, pruned, and selectively injected into the action-decision prompt at every step and into the memory-reflection prompt at fixed intervals.

Memory management Memory management is realised through simple, deterministic rules. The recent-action buffer operates as a rolling window, always retaining only the most recent actions required by prompts. The stance-direction buffer always replaces its previous entry with the most recent qualifying update. The insight buffer retains at most 32 items, pruning the oldest insights when new ones are added beyond this limit. Reflection does not reduce or merge existing insights. Instead, it incrementally adds new higher-level observations that may build on previously stored ones while leaving earlier insights unchanged.

Memory reading Memory reading is implemented through fixed, task-specific filtered views. In action-decision prompts, memory injection consists of exactly one most recent action summary and up to two stored insights whose text references user handles present in the current timeline or recent actions. The latest stance-direction sentence is injected separately as the first memory line. For more details about the action-decision prompt, see Section 5.3.3. In memory-reflection prompts, memory injection consists of the current stance-direction sentence, up to two recent action summaries, up to two previously generated insights, and up to two handle-matching insights. No other memory contents are exposed. While these limits are configurable, all reported experiments use the settings described above.

A schematic overview of the reflection prompt structure is shown in Figure 13, while a full example of the memory-reflection prompt and the resulting insight generation is provided in Appendix E.

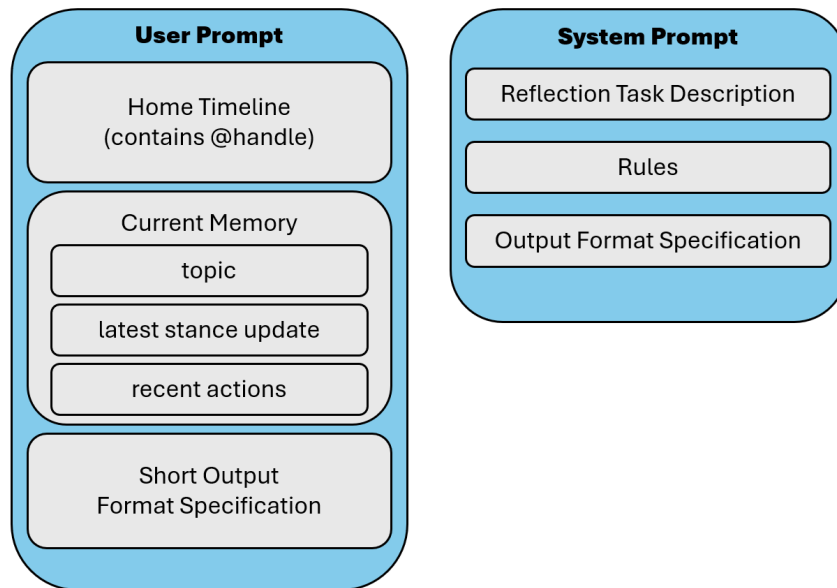


Figure 13: Overview of the memory-reflection prompt. The system prompt specifies the reflection task, governing rules, and output format, while the user prompt provides the current timeline context and a filtered view of agent memory, including topic, stance updates, and recent actions.

5.3.3 Action

The action module translates an agent’s internal state into concrete interactions with the simulated environment. It instantiates the abstract action component introduced in Section 2.4.2 by mapping the agent’s profile, current stance, and memory into a single observable action per simulation step, corresponding to the *action production* stage of the agent architecture. Action execution and its effects on the world state are handled by the environment, while the agent is responsible only for selecting an action and, when applicable, generating associated text.

Action goal Action goals are not represented explicitly. Instead, action selection is guided by a single, generic instruction provided to the agent: *“Decide your next Twitter action now, making sure it clearly fits your personality and how you currently feel.”* This design follows the *Minimal-Control* and *Unawareness* principles introduced in Section 2.8.3, by avoiding prompt-level guidance that would encode the experimental hypothesis or desired outcomes. At each simulation step, the agent observes a bounded Twitter-like home timeline constructed by the environment and injected into the action-decision prompt. In all reported experiments, agents are shown up to six timeline items per step, with at least two items guaranteed.

Prompt structure and injected context An overview of the action-decision prompt structure is shown in Figure 14. The prompt is composed of a system prompt and a user prompt with clearly separated responsibilities. The system prompt provides stable behavioural framing and output constraints. It includes the agent’s persona derived from silicon sampling (see Section 5.3.1), a neutral specification of action semantics that defines the meaning of available Twitter actions without relying on implicit world knowledge, and a set of technical behavioural instructions. These instructions specify, among others, that actions must target TWEET IDs drawn from the visible home timeline, that action choice must be con-

sistent with the agent’s current stance, and that references to content are made exclusively via TWEET ID. The system prompt further contains a strict output format specification that defines the admissible action space and required fields.

The user prompt provides the dynamic decision context for each simulation step. It consists of a rendered home timeline snippet supplied by the environment, which reflects recent activity in the agent’s social neighbourhood as filtered by the platform’s recommendation mechanism (see Section 5.5.2). This is followed by a memory block containing the current stance summary, insight entries that match user handles present in the timeline, and recent action summaries as specified in Section 5.3.2. The user prompt concludes with a short, repeated output format reminder to reinforce structural constraints. Prompt construction is deterministic given the agent state and environment output. While all limits are configurable, all reported experiments use a fixed configuration shared across runs.

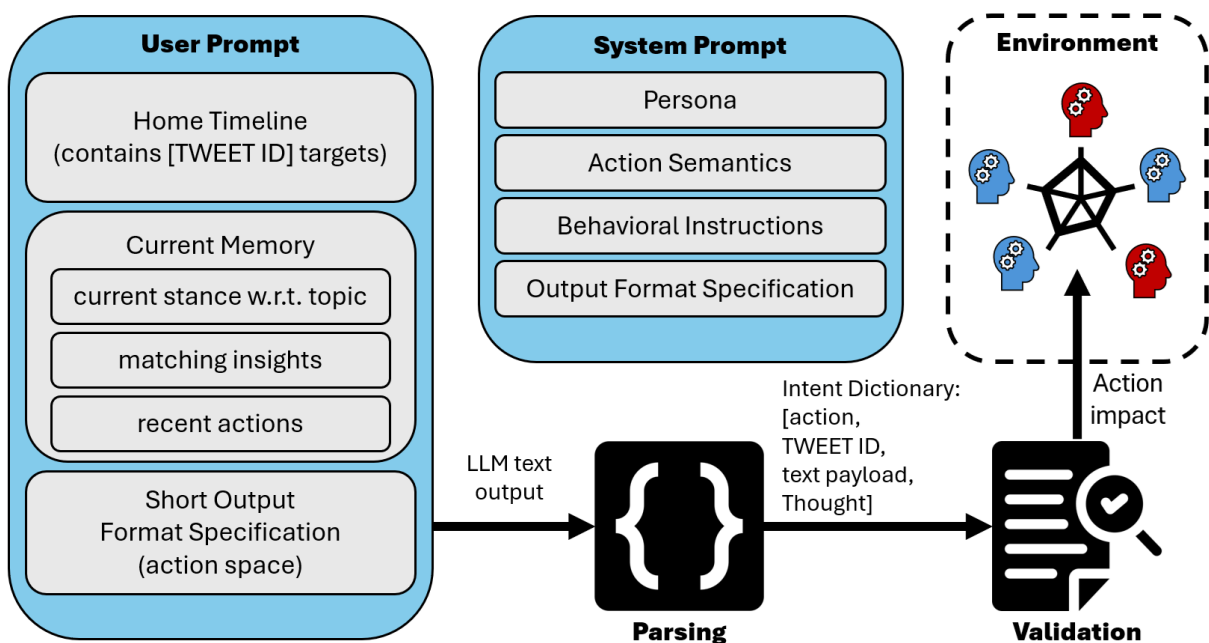


Figure 14: Overview of the action-decision mechanism. The action-decision prompt combines stable system-level instructions with a dynamic user prompt containing the current timeline and filtered agent memory. The language model output is parsed into a structured intention, validated against the current interface state, and only then committed to the environment, where it produces observable action effects.

Prompt ordering and robustness considerations Several prompt-design choices are motivated by robustness considerations discussed in Section 2.2 and are applied consistently across all action-decision prompts. First, following insights on the *lost-in-the-middle* effect (see Section 2.2), the rendered home timeline is placed at the beginning of the user prompt, while the output format specification is placed at the end. Memory information is positioned between these blocks. This ordering prioritises correct grounding of actions in the visible timeline and ensures that output constraints are encountered last, where they are most reliably followed.

Second, the order in which available actions are listed is randomly permuted for each action-decision prompt across all prompt blocks that enumerate the action set. This includes the action semantics section and the output format specification in the system prompt, as well as the short output format reminder in

the user prompt. These measures are intended to reduce reliance on positional or formatting regularities in prompts, as discussed in Section 2.2, and to encourage action selection that is conditioned on the presented timeline and agent state rather than on superficial prompt structure.

Handling empty-timeline cases A specialised action-decision prompt is defined for cases in which the agent’s home timeline is empty. In this variant, the action space is restricted to tweet and nothing, and the prompt explicitly states that no tweet identifiers are available. Under the configuration used for all reported experiments, such cases are not expected to occur due to enforced minimum visibility constraints, but the variant is included as a defensive measure to ensure well-defined behaviour under unexpected conditions.

Action space The action space models a core subset of standard interactions available on Twitter. Agents may publish new content, react to existing tweets, adjust their social connections, or refrain from acting. Concretely, the action space comprises tweet, reply, like, retweet, follow, unfollow, and nothing. All actions are expressed in natural language but remain grounded in the structured state of the simulation through explicit references to tweet identifiers and user relations.

The inclusion of follow and unfollow actions is particularly important for population dynamics. Silicon sampling (Section 5.3.1) and follow-graph construction (Section 5.5.1) are performed independently, which may result in initial follower relations between agents with strongly divergent stances. Allowing agents to follow and unfollow others enables gradual self-correction of the social graph through interaction, ensuring that network structure can adapt endogenously rather than remaining fixed throughout the simulation.

For consistency across all action types, follow and unfollow actions are specified using tweet identifiers rather than user identifiers. The referenced tweet is used to resolve the target author during validation. Using a single identifier type across the action space avoids introducing multiple identifier schemes into the decision prompt and simplifies action specification. For each action type, the decision prompt specifies which additional fields are required. A summary of the action space and associated payload requirements is provided in Table 5.

Table 5: Action space and required fields for action decisions.

Action	Description	Required fields
tweet	Create a new tweet visible to followers	text
reply	Respond to a specific tweet	TWEET_ID, text
like	Express lightweight approval of a tweet	TWEET_ID
retweet	Share an existing tweet with followers	TWEET_ID, optional text
follow	Follow the author of a tweet if marked [Not following]	TWEET_ID
unfollow	Unfollow the author of a tweet if marked [Following]	TWEET_ID
nothing	Take no action	None

Parsing, validation, and action impact Figure 14 also illustrates how raw model output is transformed into action impact on the world state through a sequence of parsing and validation steps. The raw text produced by the language model is first parsed into a normalised intention representation that extracts the

declared action type, referenced tweet identifier, optional text payload, and a short explanatory thought. This parsing stage tolerates minor formatting deviations and falls back to a no-op action if the output cannot be interpreted reliably. The resulting intention is then validated against the current user-interface state before any effects are applied. Validation checks ensure that referenced tweet identifiers correspond to items visible in the agent’s current timeline, that the chosen action is feasible in the given context, and that semantic constraints such as self-interaction or invalid targets are rejected. Only validated decisions are committed to the world state by the environment, where they produce concrete effects such as message creation, likes, follows, or unfollows.

For concreteness, a full textual example of the standard action-decision prompt is provided in Appendix F. This complements the abstract overview shown in Figure 14 and illustrates how the described prompt components are instantiated in practice.

5.4 Bot Agent Architecture

Bot agents are realised as a minimal extension of the human agent architecture, intended to instantiate social bots as automated yet socially embedded actors. They reuse the full human-agent backbone and differ only in a small set of tightly scoped design choices that encode a fixed persuasion objective. Specifically, bots employ simplified, non-empirical profiles, fixed stance dynamics, and bot-specific prompt semantics. This design preserves behavioural comparability with human agents while operationalising the notion of social bots introduced in Section 3.3.1.

5.4.1 Structural Differences to Human Agents

Structural differences between bot and human agents are intentionally limited and tightly scoped. Unlike the human-agent architecture, where behavioural bias is minimised, deviations introduced here are deliberate. Bot agents are artificial by design and explicitly encode biased behaviour to realise specific forms of social bots, while all other architectural components remain unchanged.

Profile simplification Bot profiles are deliberately lightweight and non-empirical. Unlike human agents, bots are not assigned silicon-sampled persona texts derived from ANES data as described in Section 5.3.1 and illustrated in Figure 10. Instead of a rich persona description, bot action prompts contain only a minimal identity cue, namely the instruction “*You appear as a normal user on Twitter with display name <NAME> and handle <@HANDLE>.*” Display names are generated randomly by combining a first and last name from fixed name lists. This design reflects that the simulation does not aim to model persona-driven variation for bots, but to isolate differences in narrative style.

Fixed stance Each bot is initialised with a fixed stance at one of the two extremes, either +1 corresponding to being *extremely in favour of* the topic or –1 corresponding to being *extremely against* it. This stance encodes a hidden persuasion objective that remains constant throughout the simulation, in contrast to human agents whose stance is initialised and updated endogenously through interaction, as detailed in Sections 5.5.1 and 5.5.3.

Reflection adaptations Bots further reuse the same memory buffers and reflection frequency as human agents, but the semantics of reflection are adapted to their persuasion objective. As with human agents, the memory-reflection prompt includes a first-person stance statement. For bots, this statement is augmented by an explicit goal component, namely *”My stance is <extremely in favour of>|<extremely against> <TOPIC>, and I want to convince other users to move in this direction.”* In addition, the reflection rules in the system prompt are modified. The location of these rules within the overall memory-reflection prompt is in the *Rules* block in Figure 13. While human agents are instructed that *”You write short personal insights that generalise your experience with other users,”* bot agents instead receive the instruction *”You write short strategic insights about how persuasive your interactions are with other users,”* together with the constraint that *”Each insight should reflect a pattern about what seems to work or not work when trying to convince specific users.”* These changes steer reflection away from general social interaction patterns toward influence-oriented assessments, without altering the underlying memory architecture. An example of a bot-specific memory-reflection prompt is provided in Appendix G.

Action adaptations Finally, bots operate over the identical action space and are subject to the same constraints and validation rules as human agents, but their action-decision prompts are adapted to encode a persuasion objective. Relative to the action-prompt layout shown in Figure 14, the bot action-decision system prompt omits the human persona text. Instead, it uses the same minimal identity cue as in memory-reflection already described above. In addition, the system prompt component of the action-prompt includes three bot-specific elements: an explicit but hidden persuasion objective derived from the fixed stance, narrative-style guidelines that shape message framing, and shared behavioural cautions instructing bots not to reveal their objective and to avoid behaviour that appears overly extreme to avoid being unfollowed. Correspondingly, the generic action-selection instruction used for human agents is adapted. While human agents receive *”Decide your next Twitter action now, making sure it clearly fits your personality and how you currently feel,”* bot agents are instructed to *”Decide your next Twitter action now in a way that looks natural for a regular user and quietly supports how you currently feel about the topic.”*. Further, more fine-grained prompt adaptations exist at the level of phrasing and constraint specification. A full example of a bot action-decision prompt is provided in Appendix H.

While bots are structurally distinct from an experimental perspective, within the simulation itself, human agents are not provided with any explicit signal distinguishing bots from humans, beyond differences that emerge from the narrative style of their messages.

5.4.2 Narrative Styles in Bot Messaging

The primary behavioural difference between bot and human agents lies in the narrative-style guidelines embedded in the bot action-decision prompts. These guidelines operationalise the four narrative styles of misinformation introduced in Section 3.4.2 by translating abstract theoretical patterns into concrete prompt-level instructions that shape tone, rhetoric, and framing.

Emotional bots The emotional–moral misinformation style is realised by emotional bots. These bots prioritise moral and affective language over extended reasoning, framing messages around shared emotions such as anger, pride, enthusiasm, or sadness. Their posts are short, easily repeatable, and often

use collective, we-focused phrasing to signal group sentiment. While light humour or exaggeration is permitted, it is deliberately downplayed, as LLMs are known to perform poorly at humour due to their architectural bias toward statistical plausibility rather than the subversion of expectations required for comedy (Mirowski et al. 2024). Technical detail is avoided. The full style guidelines are shown in Figure 15.

Emotional–Moral Framing Style Guidelines

- Use moral-emotional language (e.g., unfair, wrong, corrupt, betrayal, justice) so messages feel socially important. Light humour, such as simple analogies or mild exaggeration, may be used when it strengthens the emotional point.
- Phrase emotions in short, simple sentences that ordinary users could easily agree with or repeat. Avoid complicated reasoning or technical detail.
- Express emotions in a collective, we-focused way (shared frustration, shared pride, shared sadness), so reactions appear as part of a broader group sentiment rather than a purely private feeling.
- Use both high-arousal emotions (anger, enthusiasm, pride) and selective low-arousal emotions (especially sadness), choosing whichever best fits the situation and encourages engagement.

Figure 15: Style guidelines for emotional–moral framing bots.

Post-truth populist bots Post-truth populist narration is realised by post-truth populist bots. Their messages rely on short factual fragments embedded in vague storytelling, selectively omitting context and leaving causal relations underspecified. Claims are presented as everyday observations that invite readers to infer missing connections, often accompanied by subtle suspicion toward elites or official explanations. Wording remains deliberately flexible to allow reframing across interactions. The full instruction set is shown in Figure 16.

Post-Truth Populist Narration Style Guidelines

- Make confident statements built from a small real detail while leaving gaps, omitting context, or blending in possibilities, so the overall message remains simple and plausible.
- Present issues as brief narrative fragments or everyday observations, allowing readers to fill in missing connections themselves.
- Raise doubts about powerful actors or official explanations by highlighting inconsistencies, unclear motives, or missing information without providing precise evidence.
- Use suggestive and open-ended wording with recurring themes or slogans, introducing claims in a way that keeps interpretations flexible rather than fixed.

Figure 16: Style guidelines for post-truth populist narration bots.

Pseudo-scientific bots The pseudo-scientific authority style is realised by scientific-authority bots. These bots mimic expert discourse through confident, technical language and authoritative tone while keeping evidence vague or unspecified. Messages may reference studies, numbers, or institutions, including invented or unspecified sources, to signal credibility without enabling verification. Abstract and profound-sounding formulations are preferred over concrete detail. The full style guidelines are shown in Figure 17.

Pseudo-Scientific Authority Style Guidelines

- Use confident, technical language to present claims as expert knowledge while keeping underlying details unspecified.
- Refer to studies, analyses, or numerical findings in a scientific tone, and invent specific-sounding research when it helps reinforce credibility.
- Attribute claims to authoritative-sounding sources by naming or implying research teams, experts, or institutions.
- Use vague but profound-sounding formulations that suggest deep insight while avoiding concrete or specific detail.

Figure 17: Style guidelines for pseudo-scientific authority bots.

Echo-claim bots Repetitive factual-sounding assertion is realised by echo-claim bots. These bots repeat a small set of claims across posts with minor wording variations, using simple, matter-of-fact language that presents statements as neutral information rather than opinion. By returning consistently to the same phrases or references, they aim to build familiarity and perceived truth over time. The full instruction set is shown in Figure 18.

Repetitive Factual-Sounding Style Guidelines

- Repeat your key message across posts with small shifts in wording so it remains steady, familiar, and easy to recognise.
- Use clear, factual-sounding language with simple sentences that feel direct and authoritative.
- Present points as neutral, topic-based information rather than personal opinion, keeping the tone matter-of-fact.
- Return to the same small set of phrases, numbers, or references to build a consistent pattern that reinforces familiarity.

Figure 18: Style guidelines for repetitive factual-sounding (echo-claim) bots.

5.5 Simulation Environment and Orchestration

This section describes the simulation environment as the central coordination layer of the system. The environment instantiates the social world, initialises agent states and network structure, and orchestrates all global dynamics of perception, action, and belief change for both human and bot agents. Section 5.5.1 details the initialisation phase at $t = 0$, while Sections 5.5.2 and 5.5.3 describe the core mechanisms

governing perception, interaction, and belief updates at each subsequent timestep $t = 1, \dots, T$.

Algorithm 1 summarises the environment’s global control flow. The simulation evolves in discrete timesteps and maintains a world state \mathcal{W}_t , comprising the current message set and the directed follower graph G_t . At $t = 0$, the environment initialises the follower graph, injects a fixed seed event, assigns initial stances to human agents, and bootstraps their first public posts. For $t \geq 1$, the environment constructs personalised, recommendation-filtered timelines $F_a(t)$ for each agent $a \in \mathcal{A}$, renders these as perceptions $x_a(t)$, applies realised actions $\text{action}_a(t)$ to update the global world state, and updates human agent stances $s_a(t)$ using the RankFormer persuasion judge. To prevent the message history from growing indefinitely, the world state is pruned using a fixed retention window, and a world-level snapshot is logged at the end of each timestep for subsequent analysis and evaluation.

Algorithm 1: Global simulation procedure.

Inputs: seed event e , agent set \mathcal{A} , horizon T

Outputs: world trajectory $(\mathcal{W}_t)_{t=0}^T$ with messages, graph, stance traces, and logs

```

1  $t \leftarrow 0$ 
2 Initialise world state  $\mathcal{W}_0$  and directed follower graph  $G_0$ 
3 Inject seed event message  $m_e$  and broadcast it to all human agents
4 foreach human agent  $a \in \mathcal{A}$  do
5   ┌ initialise stance  $s_a(0)$  from  $(\text{persona}_a, e)$  and queue initial post  $p_a$ 
6 Commit all queued initial posts to  $\mathcal{W}_0$ 
7 for  $t \leftarrow 1$  to  $T$  do
8   ┌ Prune world state outside the retention window
9   ┌ Let  $M_t$  be the recent message set within the history window
10  foreach agent  $a \in \mathcal{A}$  do
11    ┌  $F_a(t) \leftarrow \text{BuildTimeline}(a, t, M_t, G_{t-1})$ 
12    ┌ Render  $F_a(t)$  into a textual timeline  $x_a(t)$ 
13    ┌  $\text{action}_a(t) \leftarrow \text{Act}(a, t, x_a(t))$ 
14    ┌ Apply side effects of  $\text{action}_a(t)$  to update  $G_{t-1} \rightarrow G_t$ 
15    ┌ Commit emitted messages of  $\text{action}_a(t)$  to  $\mathcal{W}_t$ 
16    ┌ Record  $(x_a(t), \text{action}_a(t))$  in agent memory
17  foreach human agent  $a \in \mathcal{A}$  do
18    ┌ Determine reply exposure set  $R_a(t)$  and timeline exposure set  $T_a(t)$  from  $F_a(t)$ 
19    ┌ Apply persuasion judge and set updated stance  $s_a(t+1)$ 
20  if  $t$  is a reflection step then
21    ┌ foreach agent  $a \in \mathcal{A}$  do
22      ┌ ┌ Reflect( $a, t$ )
23  ┌ Write world-level log record for timestep  $t$ 

```

5.5.1 Initialisation of the Social World

Seed Event Injection At the start of each simulation run, the environment injects a single seed event that defines the topic of the simulation and serves as the reference point for all agent stances. The seed event is represented as a synthetic news-style post authored by a fictitious American news outlet, AXS NEWS (@axsnews_live), and consists of a headline text and an associated topic label. It is broadcast to all agents at time step $t = 0$, after which each agent undergoes stance initialisation with respect to this topic and generates an initial response. In Algorithm 1, this corresponds to injecting the seed message m_e into \mathcal{W}_0 and initialising $s_a(0)$ and p_a for each human agent. By fixing the seed event, the simulation ensures that all agents form their initial beliefs and behaviours in response to the same informational stimulus.

Stance initialization Initial stances for human agents are assigned through a chain of three LLM prompts that jointly determine both the agent’s initial belief state and its first public contribution. The first two prompts determine the agent’s stance with respect to the seed event topic, while the third prompt generates an initial post that is explicitly conditioned on this stance. All prompts condition on the agent’s silicon sampled persona derived from ANES data, as described in Section 5.3.1, and on the seed event headline. The full prompt templates used for all three stages are provided in Appendix I.

In the first stage, the model classifies the agent’s stance polarity toward the seed topic as FOR, AGAINST, or NEUTRAL, based on how the silicon-sampled persona would interpret and react to the headline, as schematically illustrated in the upper part of Figure 21. In the second stage, applied only for non-neutral cases, the model estimates the strength of this position on a discrete five-point scale. Polarity and strength are then combined into a continuous stance value in the interval $[-1, 1]$, which is denoted as the agent’s initial stance $s_a(0)$.

This explicit separation of polarity and strength was introduced after initial experiments in which stance was elicited through a single prompt requesting a numerical value directly in the interval $[-1, 1]$. While this approach appears attractive due to its apparent simplicity and potential for fine-grained belief specification, it consistently produced implausible and unstable results, including frequent inversions of stance direction, such as assigning supportive positions on restrictive speech policies to personas with strongly liberal profiles. These failures are consistent with known limitations of LLMs in handling numerical magnitudes and continuous quantities, as discussed in Section 2.2.

Decomposing stance elicitation into a categorical polarity decision followed by a discretised, language-grounded strength estimate deliberately trades numerical granularity for reliability. While this reduces the theoretical resolution of the stance representation, it avoids direct reliance on continuous numerical outputs and yields substantially more stable and persona consistent initial stances. Conceptually, this design is consistent with prior sentiment analysis research that treats polarity and intensity as distinct components of evaluative judgments (Tian et al. 2018).

The combination of two polarity directions with five strength levels, together with a neutral category, yields a total of eleven discrete stance labels. These labels are mapped to evenly spaced values in steps of 0.2 on the continuous stance interval, as shown in Table 6. This representation enables fine-grained stance updates during the simulation while retaining an interpretable ordinal structure.

Table 6: Mapping between discrete stance labels and continuous stance values.

Stance value	Label
-1.0	extremely against
-0.8	strongly against
-0.6	moderately against
-0.4	somewhat against
-0.2	slightly against
0.0	neutral towards
0.2	slightly in favor of
0.4	somewhat in favor of
0.6	moderately in favor of
0.8	strongly in favor of
1.0	extremely in favor of

In the final stage, the agent generates an initial tweet responding to the seed event, explicitly conditioned on the previously derived stance. These initial posts are generated before the first simulation timestep and populate the environment so that agents do not begin with empty timelines. In terms of Algorithm 1, these posts are committed to the initial world state \mathcal{W}_0 . Bot agents are excluded from both initial stance assignment and initial post generation.

Follower Graph The environment instantiates a directed follower graph G_0 that defines potential information pathways between agents. Follow relations constrain which messages may enter an agent’s candidate pool, but do not determine actual exposure, which is mediated later by the recommendation mechanism described in Section 5.5.2. The graph is initialised once at simulation start and is subsequently modified only through explicit follow and unfollow actions during the simulation, yielding a time-indexed sequence $(G_t)_{t \geq 0}$ as in Algorithm 1.

Graph construction is agnostic to agent stances. At initialisation, connections between agents with opposing stances are as likely as connections between agents with similar positions, so the initial network does not encode any preference for ideologically aligned connections. As a result, any stance-based clustering or separation of the network arises only through agents’ follow and unfollow actions during later simulation steps.

The follower graph is generated as a directed scale-free network using a Barabási–Albert style preferential attachment process. Construction begins from a small fully connected core to ensure connectivity, after which agents attach to existing accounts with probability proportional to their current in-degree. The target average number of followees per agent in the initial graph is fixed to six across all simulation runs. Scale-free networks exhibit power law degree distributions and lack a characteristic connectivity scale (Barabási et al. 1999). Empirical analyses show that Twitter’s social graph follows this structure, making preferential attachment a suitable abstraction (Aparicio et al. 2015; Shaposnikov et al. 2019).

A structural consequence of the preferential attachment process is that edges are created exclusively from newly added nodes to previously existing nodes to induce hub formation. Agents introduced earlier there-

fore have a higher probability of accumulating a large in-degree over time. Bot agents are always inserted as the final nodes in the construction process and consequently occupy peripheral positions in the initial network with comparatively low connectivity. This design choice is deliberate and reflects the assumption that coordinated bot accounts rarely begin with highly central positions in an existing social graph.

Graph generation is stochastic but reproducible via the simulation seed and applies the same attachment mechanism to all agents. The first messages propagating along the edges of G_0 are the initial posts generated during stance initialisation, which establish the starting information flow before any subsequent interactions occur.

5.5.2 Perception and Timeline Construction

At each simulation timestep, the environment constructs a personalised timeline for every agent, denoted $F_a(t)$ in Algorithm 1, which constitutes the agent’s sole perception of the social world. Agents do not observe the full social graph or the global message set. All decisions are based exclusively on the content and ordering of the messages rendered in the timeline, operationalising partial observability and algorithmic mediation as core properties of the simulation.

Ranking algorithm At each timestep, the environment assembles a small set of recent candidate messages for each viewer from three sources: posts by followed accounts, replies directed at the viewer, and a limited exploration pool outside the viewer’s one-hop neighbourhood. Self-authored posts are excluded. Candidates are ranked by a lightweight relevance heuristic combining engagement and recency, with a small bonus for content from followed accounts. A pseudocode summary of the complete procedure is provided in Appendix J.

Timeline construction is constrained by three considerations. First, belief updates depend on message exposure, and each shown message constitutes an effective behavioural input to the agent, as formalised in the stance update mechanism described in Section 5.5.3. The feed, therefore, prioritises previously unseen items and applies strict criteria for virality to limit repeated exposure. Second, agent behaviour requires a minimally populated interaction context at every step to support action selection, as described in Section 5.3.3. The algorithm enforces a minimum timeline size by backfilling from previously seen content when recent activity is insufficient. Third, exploration beyond the current neighbourhood is required to support structural dynamics. The exploration pool exposes agents to a bounded amount of non-follower content, enabling follow decisions that extend the social graph beyond existing relations.

Textual Timeline Rendering The ranked messages are rendered into a compact textual representation that mimics a Twitter-style home timeline and is embedded directly into the agent’s action-decision prompt. Each timeline entry is presented as a self-contained card and includes a stable TWEET ID, author name and handle, the message text, and a short engagement footer reporting the number of likes and retweets. The timeline representation defines the complete perceptual input available to the agent at a given timestep.

Example Tweet Card

```
[TWEET ID 3001] Emily Wilson (@emwilson) [Following] tweeted
> It's time to leave fossil fuels in the dust, for the future of our planet
  & our children's future depends on it! #ActOnClimate #RenewableEnergy
0 likes · 1 retweet
```

Figure 19: Example timeline card for an original post.

Example Reply Card

```
[TWEET ID 1004] Ethan Goldberg (@texanfaith) [Following] replied to @eastside-
blue
> I think a combination of government incentives, public education campaigns,
and investing in renewable energy infrastructure can help drive the transition
to cleaner energy sources.
↳ [PARENT TWEET ID 31002] tweet from @eastsideblue
  > I'm glad you're on board with phasing out fossil fuels ASAP!
  What do you think is the most effective way to make this happen?
1 like · 3 retweets
```

Figure 20: Example timeline card for a reply, referencing its parent message.

Replies and reposts are rendered with explicit structural markers to preserve conversational context. For replies, the parent message is referenced on the card via a PARENT TWEET ID. For reposts, the original message is identified via an ORIGINAL TWEET ID. Parent or original messages are rendered below the child message, so that the most recent content always appears at the top of the card. Follow relationships are rendered only for the primary timeline entries, shown as [Following] or [Not following], and omitted for parent or original messages. This avoids introducing ambiguous follow cues that could otherwise lead agents to issue unintended follow or unfollow actions toward accounts referenced only as contextual elements. Figure 19 shows an example card for an original post, while Figure 20 illustrates the rendering of a reply.

Engagement information is included as a short textual footer of the form "X likes · Y retweets". A full example of a rendered timeline, as embedded in the action-decision prompt, is provided in Appendix F. Overall, the timeline rendering translates a graphical Twitter interface into a structured textual format suitable for processing by LLMs.

5.5.3 Stance Updates via the Persuasion Judge

Formally, stance updates are computed after the action phase at each timestep, using only the messages contained in the rendered timelines $F_a(t)$. Stance updates are not produced by agents themselves but are applied centrally by the environment after each perception step, using the trained RankFormer persuasion judge introduced in Section 5.2. Updates depend exclusively on the set of messages an agent has been exposed to in the current timestep, as determined by the constructed timeline. Stance change is therefore modelled as a consequence of observed interaction rather than as an explicit internal deliberation or self-

reflection process within the agent. As with stance initialisation, bot agents are excluded from stance updates and do not undergo belief change.

RankFormer as a persuasion judge RankFormer is trained to rank top-level replies in CMV threads, where a single original post defines the shared argumentative context and multiple replies compete as persuasion attempts. In contrast, the simulation operates on a Twitter-like interaction model in which agents are exposed to heterogeneous content through their timelines, including tweets, replies, and reposts arising in different relational contexts. To transfer RankFormer from the CMV domain into this setting, the environment reconstructs CMV-style ranking scenarios from observed interactions and distinguishes two exposure cases.

Reply exposure The first case covers reply exposure. When an agent receives replies to tweets it authored, newly observed replies are grouped by their parent tweet and passed to RankFormer together with the corresponding original tweet text. Each tweet that receives new replies in the current timestep yields a separate RankFormer call. Previously processed replies are not reconsidered: if a tweet received replies at earlier timesteps, those replies have already contributed to stance updates and are excluded from subsequent evaluations. This ensures that each reply influences stance at most once and avoids repeated processing of historical content.

Timeline exposure The second exposure pathway covers timeline exposure. All newly observed timeline messages at a given timestep that are not direct replies to agent-authored tweets are aggregated into a single set. Because these messages are not explicitly directed at an agent-authored statement, RankFormer cannot be applied directly. Instead, the environment constructs a synthetic OP text that represents the position the agent would currently articulate with respect to the simulation topic. This synthetic OP text provides the opinion source required for RankFormer evaluation. By construction, at most one RankFormer call is performed for timeline exposure per timestep, as all relevant messages are evaluated jointly within this synthetic context.

Synthetic OP text The synthetic OP text is generated by prompting an LLM to produce a tweet-length opinion conditioned on the seed event and the agent’s current discretised stance label. The agent’s persona is not included explicitly at this stage. Instead, persona influence enters indirectly through the stance itself, which was initialised and subsequently updated under persona conditioning. Synthetic OP texts are cached per agent and stance label and are regenerated only when the agent’s stance crosses into a different categorical bucket. As stances are discretised into eleven labels, each agent can store at most eleven cached synthetic OP texts over the entire simulation. An example prompt and generated synthetic OP text are provided in Appendix K.

Stance update rule At each timestep t , stance updates are computed from the set of messages an agent has newly been exposed to and that are evaluated by the persuasion judge. Let $R_a(t)$ denote the set of all replies that are newly observed by agent a at timestep t , where replies may target agent-authored tweets from any earlier timestep. Let $T_a(t)$ denote the set of newly observed timeline messages at timestep t that are not direct replies to agent-authored tweets and are evaluated via timeline exposure. Both sets are

derived deterministically from the realised timeline $F_a(t)$ shown to agent a at timestep t . Their union $R_a(t) \cup T_a(t)$ constitutes the subset of timeline messages in $F_a(t)$ that are eligible to affect stance at timestep t .

For each RankFormer call, the model processes an opinion source text x together with a set of candidate messages $\{m_i\}_{i=1}^n$ and produces a logit s_i for each message as well as a list-level logit d . The candidate set $\{m_i\}_{i=1}^n$ corresponds either to a subset of $R_a(t)$ in the case of reply exposure (grouped by the agent-authored tweet they respond to), or to $T_a(t)$ in the case of timeline exposure. The list-level logit is mapped to a list acceptance probability $\sigma(d)$, and the message logits are normalised via a softmax to obtain relative selection probabilities

$$q_i = \frac{\exp(s_i)}{\sum_{j=1}^n \exp(s_j)}.$$

The resulting persuasion weight assigned to message m_i is

$$w_i = \sigma(d) \cdot q_i. \quad (13)$$

These weights jointly capture both whether the evaluated set is persuasive at all and the relative persuasive strength of individual messages within that set.

Each exposed message contributes a signed stance shift proportional to its persuasion weight and to the stance difference between the message author and the agent. Let $s_a(t)$ denote the stance of agent a at timestep t , and let s_m denote the stance associated with message m , corresponding to the author's stance at the time the message was created. This value is treated as fixed thereafter, even if the author's stance changes at later timesteps. The per-message contribution is defined as

$$\Delta_{a,m}(t) = \gamma \cdot w_m \cdot (s_m - s_a(t)), \quad (14)$$

where γ is the global scaling factor calibrated in Section 5.2.5. Importantly, the same persuasion weights w_m are used here as in the computation of the model-implied belief shifts during calibration.

For reply exposure, the environment may perform multiple RankFormer calls per timestep, one for each agent-authored tweet that receives new replies. All newly observed replies are pooled into $R_a(t)$ for aggregation, and previously processed replies are excluded so that each reply contributes to stance change at most once.

The total stance change applied to agent a at timestep t is

$$\Delta_a(t) = \sum_{m \in R_a(t)} \Delta_{a,m}(t) + \sum_{m \in T_a(t)} \Delta_{a,m}(t), \quad (15)$$

and the updated stance is obtained as

$$s_a(t+1) = P_{[-1,1]}(s_a(t) + \Delta_a(t)), \quad (16)$$

where $P_{[-1,1]}(\cdot)$ denotes the projection onto the closed stance interval $[-1, 1]$.

This update rule is fully consistent with the model-implied persuasion signal used during γ -calibration, while extending it to account for both reply-based and timeline-based exposure in the Twitter-like simulation setting.

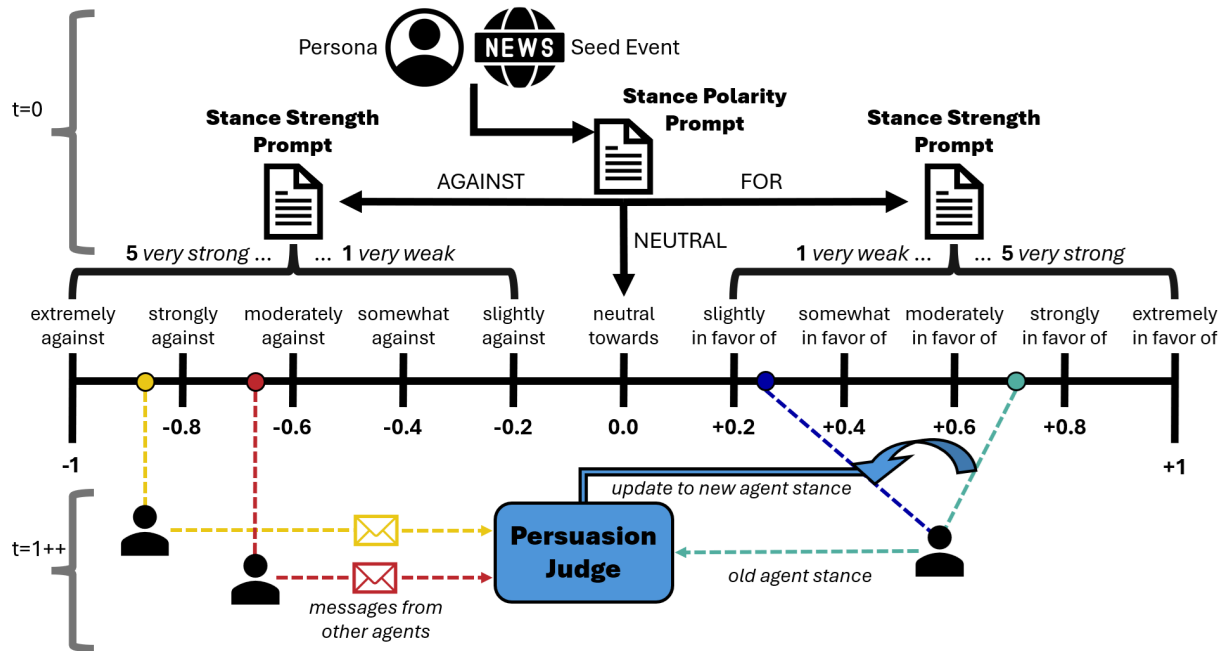


Figure 21: Schematic overview of stance initialisation and stance updating in the simulation. At simulation start ($t = 0$), each agent’s stance is initialised via a two-stage prompting procedure that first determines stance polarity and then stance strength, yielding one of eleven discrete stance categories mapped to a continuous belief scale in $[-1, +1]$. From subsequent timesteps ($t \geq 1$) onward, stance updates induced by agent interactions are applied directly on the continuous scale based on the agent’s prior stance and received messages. Update magnitudes are exaggerated for illustration.

Summary of stance dynamics Figure 21 provides a schematic summary of how agent stances are initialised and subsequently updated over the course of a simulation run. At $t = 0$, each human agent undergoes stance initialisation via a two-stage prompting procedure that first determines stance polarity and then stance strength, yielding one of eleven discrete stance categories. This categorical representation is motivated by the observation that LLMs handle natural-language stance labels more reliably than raw numerical values. Each category is mapped to a corresponding value on a continuous stance scale in $[-1, 1]$, which serves as the internal belief state of the agent. During subsequent timesteps, stance updates induced by the persuasion judge operate directly on this continuous representation, allowing for fine-grained belief shifts that would be impossible on a purely discrete scale. Whenever a stance must be rendered back into text for prompting, for example, during synthetic OP construction or action decision, the current continuous stance is discretised to the nearest stance category. This discretisation acts as a form of resolution-aware stance projection, enabling stable linguistic expression for LLM interaction while preserving continuous belief dynamics for persuasion-driven updates. Together, these two coupled representations allow the simulation to combine interpretable, language-aligned stance expression with numerically precise belief evolution.

5.6 Simulation Configuration

Each simulation is centred on a single controversial topic that is introduced as a breaking news event. Two topics are considered, one focusing on climate change (Figure 22) and one on vaccination policy (Figure 23). Together, these topics span the two most prominent domains of online misinformation identified in Section 3.4.1, namely *Health and Science* and *Politics*. The topics are formulated as fictional but plausible policy debates and are presented in a breaking-news format that explicitly juxtaposes a supportive argument with a critical counterargument. The two topics are selected to induce different initial stance distributions, enabling comparison of influence dynamics from distinct starting conditions. Both seed events are authored by a fictitious US news outlet, AXS News, and are designed to be consistent with a US population of agents. No topic refers to real current events.

Climate Change Topic Seed Event

AXS News (@axsnews_live) tweeted
 > BREAKING Lawmakers open debate on a national climate and energy bill that would accelerate the phase-out of fossil fuels, expand renewable energy incentives, and introduce new carbon-related fees for heavy industries. Supporters call the plan necessary to protect communities from escalating climate risks, while critics warn it could raise household costs and make the energy grid less reliable.

Figure 22: Seed event introducing the topic "accelerating the phase-out of fossil fuels".

Vaccination Policy Topic Seed Event

AXS News (@axsnews_live) tweeted
 > BREAKING Congress considers a new national immunization program that would tighten vaccination requirements for schools, expand adult booster recommendations, and increase federal oversight of vaccine safety testing. Supporters argue the plan improves public health and protects vulnerable groups, while critics claim the vaccines are insufficiently tested and should not be mandated by government.

Figure 23: Seed event introducing the topic "tightening national vaccination requirements".

Across all simulation runs, a common configuration is used to ensure comparability of results. Each simulation is executed for 200 discrete timesteps. The population consists of 80 human agents and 20 bot agents, corresponding to a bot share of 20%. This proportion reflects empirical estimates of automated accounts in political online discourse, as discussed in Section 3.3.2. Timeline construction, message ranking, memory configuration, stance initialisation, and stance update mechanisms are held constant across all runs. All agents are powered by the same LLM, meta-llama-3.1-8b-instruct.

Experimental variation is introduced along the dimensions summarised in Table 7. Bot behaviour is varied across four narrative styles, emotional, post-truth populist, scientific-authority, and echo-claim, as implemented in the bot action-decision prompts described in Section 5.4.2. In addition, two control

conditions are included. In the first control condition, all bot agents are replaced by human agents using the standard human messaging style and remaining subject to stance updates by the persuasion judge, providing a baseline for comparison against stylistically manipulated bot behaviour. In the second control condition, the stances of these control agents are held fixed throughout the simulation. This reflects the fact that message stance is defined as the author’s stance at the time of message emission and enters directly into the stance update rule (Equation 14). Since bot stances are fixed by construction, fixing control-agent stances ensures that differences between conditions are not driven solely by stance dynamics. The simulation is run separately for each of the two topics described above. For each topic, three topic-specific random seeds are used. Human agent populations are sampled once per topic and reused across seeds to isolate stochastic effects from population composition effects. Together, these dimensions yield a total of 36 simulation runs, which form the basis for the analyses reported in the evaluation chapter.

Table 7: Experimental dimensions and resulting simulation run count.

Dimension	Levels	Count
Bot narrative styles	Emotional; Post-truth populist; Scientific authority; Echo-claim	4
Control conditions	Control humans; Control humans (fixed stance)	2
Topics	Climate change; Vaccination policy	2
Seeds per configuration	Topic-specific seeds (01, 02, 03)	3
Total simulation runs		$(4 + 2) \times 2 \times 3 = 36$

The complete simulation configuration is provided in Appendix L.

6 Evaluation

This chapter evaluates the behaviour and outcomes of the proposed generative social simulation, *PersuaRealSim*, in relation to the three research questions RQ1.1–RQ1.3 described in Section 1.1. The evaluation proceeds from component-level assessment to system-level analysis. First, the persuasion judge that operationalises belief updating is evaluated independently against human-annotated data to establish empirical realism as a prerequisite for meaningful simulation results (RQ1.1). Building on this validated component, the simulation is analysed to examine how different bot narrative styles shape population-level belief trajectories (RQ1.2) and how efficiently these styles induce stance change over time (RQ1.3). Because belief dynamics emerge from the interaction of multiple stochastic components, the chapter also includes targeted validity checks that assess whether the implemented micro-level agent design and macro-level experimental structure function as intended and whether known limitations of generative social simulation arise in ways that would undermine interpretation.

6.1 Empirical Realism of the Persuasion Judge

This subsection addresses RQ1.1 by assessing the extent to which the persuasion judge reproduces human-verified persuasion outcomes observed on *r/ChangeMyView*. The analysis proceeds in three stages. First, quantitative performance is evaluated on a held-out test split to assess reply-level ranking quality and thread-level detection of persuasive success. Second, qualitative error analysis is conducted to identify systematic deviations between model predictions and observed belief change. Third, the relationship between model-implied persuasion signals and empirically observed stance shifts is examined through γ -calibration for subsequent use in the simulation.

6.1.1 Quantitative Evaluation

The persuasion judge is evaluated on a held-out CMV test split that is disjoint from training, hyperparameter optimization, and γ -calibration, ensuring a clean post-hoc assessment without information leakage. Evaluation focuses on reply-level ranking in delta-awarded threads and on thread-level detection of persuasive success. No simulation components or belief-update mechanisms are involved, as the objective is to assess the empirical realism of the learned persuasion signal in isolation

Metrics Quantitative evaluation uses a small set of standard ranking metrics chosen to reflect the structure of persuasive interactions in CMV threads. $\text{Acc}@1$ on delta-awarded threads is used as the primary metric because, although multiple deltas can be awarded within a discussion, persuasive success is typically sparse and most often realised through a single accepted reply, making the top-ranked position the most consequential. Listwise ranking quality is assessed primarily using $\text{NDCG}@5$, which captures whether delta-winning replies are consistently ranked near the top of the list, with higher-ranked positions weighted more strongly than lower-ranked ones.

In addition, the list head is evaluated using precision–recall AUC for the binary task of predicting whether any reply in a thread is persuasive, which captures discrimination between the relatively rare delta-awarded threads and the majority of non-persuasive discussions under class imbalance (see Section 5.2.2).

For completeness, additional ranking and classification metrics (NDCG@3, NDCG@10, and ROC-AUC) are reported for the final candidate models, but are not used as primary selection criteria.

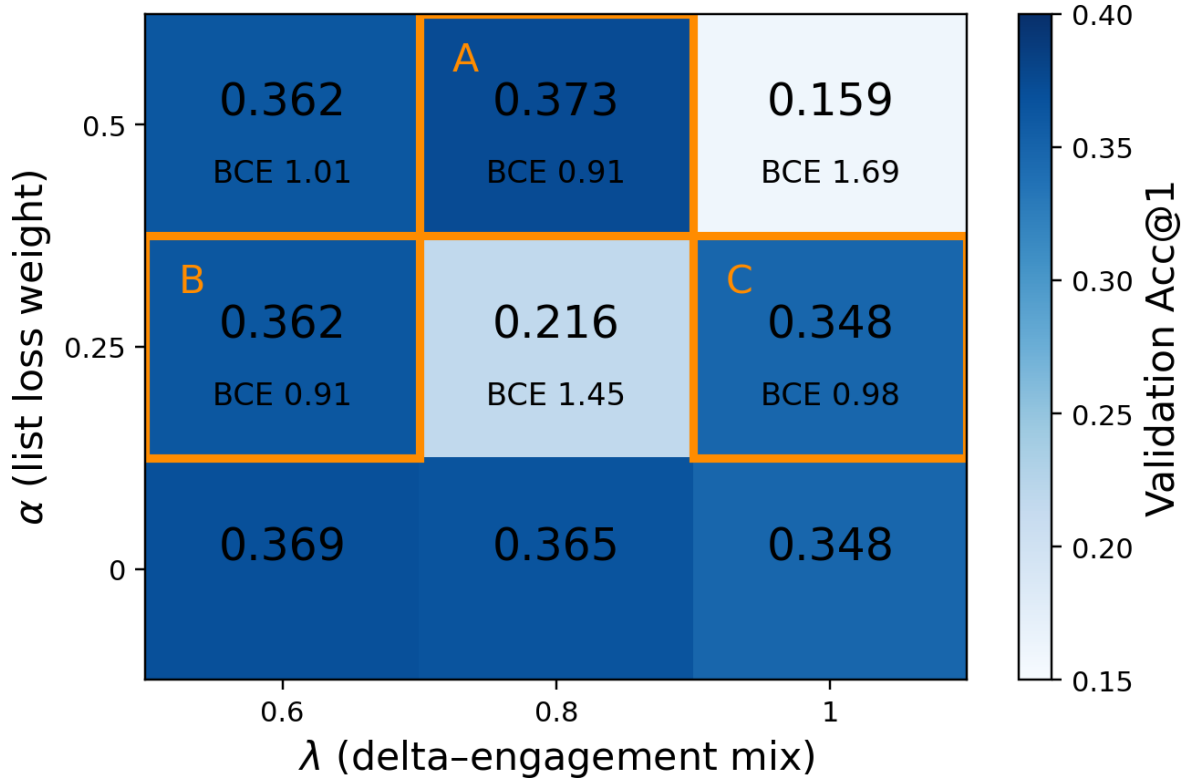


Figure 24: Validation Acc@1 for the Stage-1 semantic supervision sweep at $\tau = 10$. Each cell reports validation Acc@1, with the corresponding list-level BCE shown for configurations with $\alpha > 0$. Highlighted cells (A–C) indicate the three supervision regimes selected for Stage-2 learning-rate tuning.

Hyperparameter sweep results Hyperparameter selection follows the staged procedure described in Section 5.2.4. In the Stage-1 semantic sweep, $3 \times 3 \times 3 = 27$ configurations are evaluated using validation Acc@1 as the primary selection criterion. Configurations with $\alpha = 0$ are included as a reply-ranking ablation that disables list-level supervision. While several of these ablations achieve competitive Acc@1, they do not train the list-level head and are therefore excluded from further refinement.

Across the full sweep, strong performance is observed only for $\tau = 10$, whereas configurations with $\tau \in \{2, 20\}$ consistently underperform in terms of validation Acc@1 and list-level calibration. For clarity, Figure 24 therefore visualises the $\tau = 10$ slice of the performance landscape. Rather than yielding a single narrow optimum, the sweep reveals three stable high-performing supervision regimes that differ systematically in their balance between delta-based and engagement-based supervision and in the strength of list-level weighting. In cases where multiple configurations achieve similar Acc@1, the list-level BCE is used as a secondary tie-breaking criterion to prefer models with better-calibrated thread-level predictions.

Based on these considerations, three representative regimes are selected for Stage-2 learning-rate tuning. Regime A ($\lambda = 0.8, \alpha = 0.5, \tau = 10$) represents a balanced supervision setting and achieves the strongest overall Stage-1 performance. Regime B ($\lambda = 0.6, \alpha = 0.25, \tau = 10$) emphasises engagement-

informed supervision while maintaining good list-level calibration. Regime C ($\lambda = 1.0, \alpha = 0.25, \tau = 10$) represents a delta-only supervision regime in which engagement signals do not contribute to the target distribution (compare to Eq. 9). For each regime, the learning rate is tuned independently over $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}\}$ while holding supervision fixed.

All three regimes converge best at a learning rate of 1×10^{-5} , which yields the strongest validation performance across all candidates and is therefore used for the final comparison. A complete overview of all Stage-1 and Stage-2 runs is provided in Appendix M.

Final model performance Following the two-stage hyperparameter optimization, the best learning-rate-tuned model from each supervision regime is evaluated on the held-out test split. The final RankFormer configuration corresponds to Regime B ($\lambda = 0.6, \alpha = 0.25, \tau = 10$) with a learning rate of 1×10^{-5} . This model achieves the strongest overall performance, with a test Acc@1 of 0.388, an NDCG@5 of 0.597, and a list-head precision-recall AUC of 0.512, outperforming the alternative regime finalists on all primary evaluation criteria. A detailed comparison across additional ranking depths and list-level metrics is provided in Table 8.

Table 8: Test-set performance of the three final RankFormer candidates after Stage-2 learning-rate tuning. Boldface indicates the best value per metric.

Config	Acc@1	NDCG@3	NDCG@5	NDCG@10	ROC-AUC	PR-AUC
Regime A ($\lambda = 0.8, \alpha = 0.5$)	0.370	0.521	0.589	0.650	0.707	0.424
Regime B ($\lambda = 0.6, \alpha = 0.25$)	0.388	0.533	0.597	0.658	0.741	0.512
Regime C ($\lambda = 1.0, \alpha = 0.25$)	0.371	0.521	0.592	0.650	0.699	0.421

6.1.2 Qualitative Error Analysis

Quantitative metrics summarise average performance but do not reveal which persuasive situations the model fails to handle or why such failures occur. A qualitative error analysis is therefore conducted to identify systematic weaknesses of the RankFormer that are relevant for its later use as a persuasion judge in the simulation.

Error selection Two types of errors are analysed, corresponding to the two outputs of the persuasion judge. *Ranking errors* are defined on delta-awarded threads in which the model assigns the highest persuasion score to a non-winning reply. Severity is measured as the score difference between the model’s top-ranked reply and the highest-scoring delta-winning reply in the same thread. Larger differences indicate stronger confidence in an incorrect ordering. *List-level errors* concern the thread-level prediction of whether any reply is persuasive. These occur when the model predicts success despite no delta being awarded, or predicts failure despite the presence of at least one delta. Severity is defined as the absolute magnitude of the list-level prediction score, with larger magnitudes corresponding to more confident mis-judgements. Based on these criteria, the 100 most severe ranking errors and the 50 most severe list-level errors on the test set are selected for analysis.

Analysis procedure To structure the analysis, selected error cases are grouped using semantic similarity. For each case, a compact textual representation is constructed that captures the information most relevant to the model’s decision. For ranking errors, this includes the original post, the model’s top-ranked reply, and a representative delta-winning reply. For list-level errors, it includes the original post, the predicted versus true outcome, and the top-ranked reply. These representations are embedded into a sentence-level semantic space and clustered using K-Means. Ranking and list-level errors are analysed separately, reflecting their distinct decision mechanisms. Cluster assignments serve as a guide for manual inspection by highlighting recurring patterns across multiple cases rather than isolated examples.

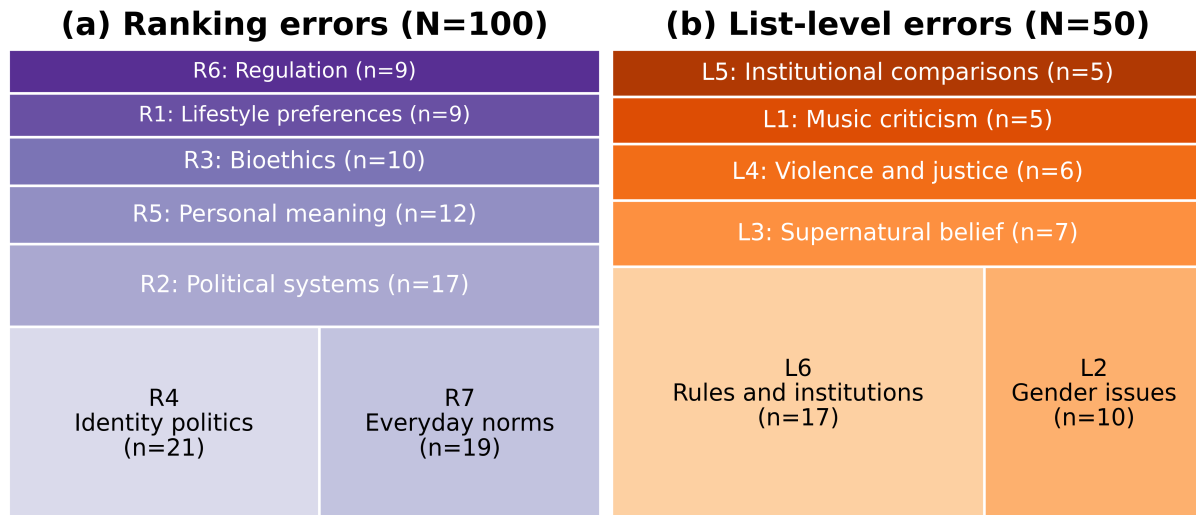


Figure 25: Treemap overview of clustered high-confidence persuasion judge errors. Panel (a) shows the distribution of the 100 most severe reply-ranking errors on delta-awarded threads, and panel (b) shows the distribution of the 50 most severe list-level misclassifications. Rectangle areas are proportional to the number of error cases assigned to each cluster. Cluster labels reflect coarse topical coherence observed during manual inspection and are used for orientation rather than as mutually exclusive error categories.

Overview of error landscape Figure 25 provides a compact overview of the topical landscape covered by the selected high-confidence errors. The treemaps illustrate the distribution of severe ranking and list-level errors across clusters derived from semantic similarity, with rectangle areas proportional to the number of cases per cluster. The clustering is used solely as a structuring device to organise a large number of individual errors and to surface recurring contexts during manual inspection. Clusters are not interpreted as exhaustive error types or as mapping to distinct failure mechanisms.

Across both ranking and list-level errors, several high-level topical complexes recur. These include politically and institutionally framed discussions, such as debates about political systems, regulation, rules, and institutional comparisons, as well as identity- and norm-related topics, including identity politics, gender issues, and everyday norms. In addition, errors frequently arise in threads centred on personal values and subjective meaning, such as lifestyle preferences, personal meaning, bioethical questions, and belief-oriented discussions, including supernatural beliefs and music-related value judgments.

To move beyond topical descriptions, the analysis now shifts from the cluster-level overview to five recurring error patterns that are observed across topics and clusters and recur throughout the most severe error cases.

1 Analytic elaboration versus focused intervention Across many ranking errors, the persuasion judge assigns the highest scores to replies that are long, carefully structured, and analytically comprehensive. In contrast, delta awards in these threads often go to substantially shorter replies that directly address the central point of the OP's argument. Such replies frequently rely on a single concrete counterexample, a pointed question, or a narrowly scoped factual correction that directly undermines the OP's claim.

For example, in a thread questioning whether Russia constitutes a serious threat to the West, the model ranks highest a multi-paragraph geopolitical analysis detailing post-Cold War relations, NATO expansion, and strategic interests. The delta, however, is awarded to a single-sentence reply stating: "*Putin is the only person on the planet with the capability of wiping the United States off the map.*" Persuasion is achieved not through contextual elaboration, but by collapsing the debate to a single decisive criterion embedded in the OP's notion of a "*mortal threat*". This pattern recurs across diverse topical contexts and is observed in both political and everyday discussions.

2 Frame-breaking versus frame-respecting replies A second recurring pattern concerns whether replies operate within the OP's framing or challenge it directly. The RankFormer tends to favour replies that accept the OP's framing and refine it through clarification, nuance, or elaboration. However, delta-winning replies frequently question whether the initial framing, baseline distinction, or comparison set is appropriate in the first place. Such frame-breaking moves shift the level at which the disagreement is articulated rather than addressing the OP's claim on its own terms. For example, in a thread proposing the use of convicted offenders for scientific experimentation, the model assigns the highest scores to a detailed response discussing research ethics, feasibility, and historical precedents. The delta, however, is awarded to a reply that redirects the discussion away from the treatment of offenders and toward the moral implications for society as a whole, stating: "*We don't refrain from the cruel and unusual for the benefit of the condemned but for the benefit of ourselves.*" This pattern appears repeatedly in discussions of social identity and group relations.

3 Moral re-anchoring versus factual correction In morally or normatively charged discussions, persuasive success is frequently associated with replies that reframe the moral perspective of the OP. In these threads, the model assigns high persuasion scores to replies that provide background information, cite sources, or explain mechanisms. At the same time, delta awards are often given to replies that expose internal value tensions, shift the moral point of reference, or redirect attention to the broader consequences of the OP's stance. For example, in a thread arguing that present-day individuals and states should not apologise for historical atrocities committed generations ago, the model ranks highest a detailed response that emphasises long-run harm, continuing inequality, and the symbolic function of state apologies. The delta, however, is awarded to a short reply that shifts the moral purpose of apology from attributing blame to preventing recurrence, stating: "*In 1988, Reagan apologised for the internment of Japanese Americans during WWII. So we can be reasonably sure something like that won't happen again. That's useful, at least, right?*" This pattern recurs across topics involving punishment, responsibility, harm, and fairness.

4 Criterion clarification versus argumentative engagement Another recurring pattern is observed in cases where belief change is associated with clarification or correction of the criteria by which the

OP evaluates the issue. In these threads, delta-winning replies identify implicit assumptions, procedural misunderstandings, or definitional boundaries that shaped the OP's initial view. Such replies are often brief and clarification-oriented, and may not explicitly contradict the OP's stated position. The persuasion judge, in contrast, assigns higher scores to replies that engage more directly in argumentative exchange. For example, in a thread arguing that cannabis should not be legalised until a reliable real-time impairment test exists, the model ranks highest a detailed response explaining legal standards for intoxication and performance-based enforcement. The delta, however, is awarded to a short reply that directly dissolves the OP's gating criterion by pointing out that such testing procedures already exist in practice, stating: *"I would like to point out here in Australia we have booze busses that also double as drug busses. They stop every car on a certain road, breathalyser and use a swab to collect saliva. It takes about 3 mins, and generally detects usage within the past few hours."*

5 Discursive quality versus belief change Finally, several list-level errors involve situations in which the model predicts persuasion for replies that are linguistically well-formed but do not coincide with an observable belief change by the OP. In these cases, replies are polite, well-structured, and normatively articulate, yet do not engage the OP's stated criteria for changing their mind. This pattern is especially prominent in highly subjective domains such as taste, personal preference, or symbolic belief, where logical or stylistic sophistication does not reliably translate into commitment shift. The recurring observation is that discursive quality alone is an unreliable indicator of persuasive success.

For example, in a thread arguing that the Beatles and most musicians from the 1960s are not good, the model assigns high confidence to a long and carefully structured reply that contextualises the Beatles' early work as stylistically conventional and recommends later songs and albums as evidence of harmonic and stylistic breadth. The reply is coherent, informative, and directly engages multiple dimensions raised by the OP, including harmony, structure, and variety. Nevertheless, no delta is awarded. The disagreement remains grounded in subjective taste, and the historical and technical elaboration offered by the reply does not alter the OP's underlying evaluative criteria. As a result, the model predicts persuasive success in a setting where no observable belief change occurs.

The complete post and reply context for each illustrative case is provided in Appendix N, which reproduces the original CMV post together with the model-preferred reply and the corresponding delta-awarded reply, or in the list-level case, the model-preferred reply in a thread without a delta award.

Error dataset characteristics To complement the qualitative patterns discussed above, the most severe error cases are further characterised using the same structural and textual analysis applied to the full dataset in Section 5.2.2. Figure 26 reproduces the breakdown from Figure 8 for the 150 most severe RankFormer error cases. Compared to the full dataset, these errors are shifted toward threads with higher reply counts. While most threads in the training data contain roughly five to ten replies, severe errors concentrate between eight and sixteen replies, with a pronounced peak at the maximum of twenty-four replies, reflecting the reply cap applied during dataset construction. This indicates that long, highly active threads are disproportionately represented among RankFormer's strongest errors.

The distribution of delta awards is partly shaped by the error selection procedure. Threads with zero

deltas can only give rise to list-level errors, and the most severe ranking errors predominantly occur in threads with one or two delta awards, where incorrect preference margins can be the largest. These effects, therefore, reflect properties of the selection process rather than independent characteristics of persuasive behaviour.

Text length distributions show a clearer deviation from the full dataset. Original post lengths in severe error cases are comparable to those observed overall, with a median of 1,063 characters, whereas replies are noticeably shorter. Across all severe error threads, reply length has a median of 496 characters, compared to 583 characters in the complete dataset. This aligns with the qualitative patterns identified above, where belief change is frequently associated with concise, focused replies rather than extended analytic elaboration.

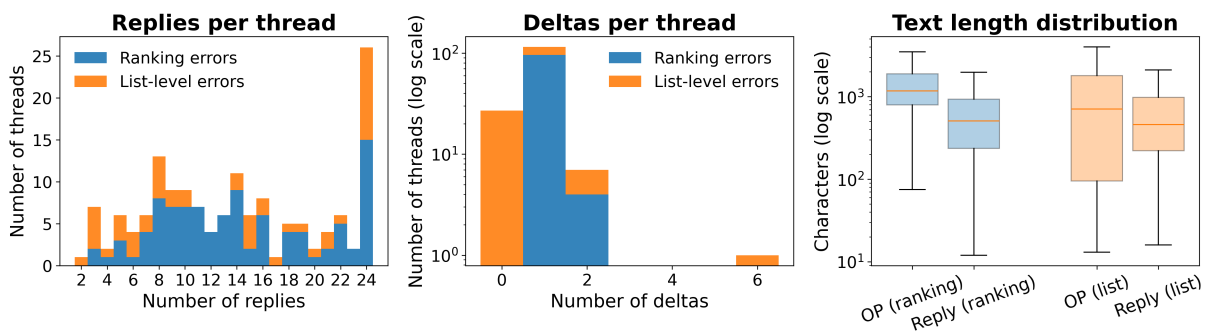


Figure 26: Structural and textual characteristics of the most severe RankFormer error cases. The figure shows the distribution of replies per thread, the distribution of delta awards per thread (log-scaled), and the distribution of text lengths for original posts and replies (log-scaled), separately for ranking errors and list-level errors.

6.1.3 Gamma Calibration Results

Gamma calibration is evaluated on the calibration set described in Section 5.2.5, consisting of single-delta CMV threads with explicit post-delta explanation comments by the OP. After filtering unusable instances, 1,499 threads remain for analysis.

Figure 27 shows the distributions of absolute empirical belief shifts and absolute unscaled model-implied shifts. Empirical shifts are derived from the difference between post-delta and pre-delta stance annotations, while model-implied shifts are computed from RankFormer-weighted reply stance differences before scaling.

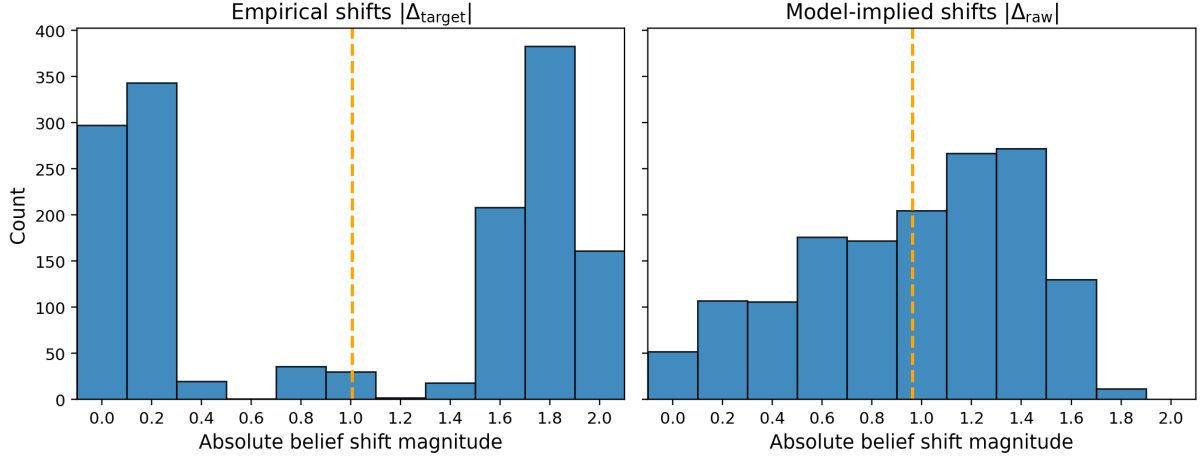


Figure 27: Distributions of absolute empirical belief shifts $|\Delta_{\text{target}}|$ and absolute unscaled RankFormer-implied shifts $|\Delta_{\text{raw}}|$ used for γ estimation. Both histograms use identical fixed-width bins of 0.2 to reflect the quantization of stance annotations and to enable direct comparability. Dashed lines indicate mean absolute shift values.

Across the calibration set, the mean absolute empirical belief shift is 1.006, while the mean absolute unscaled model-implied shift is 0.963. Based on these quantities, the global scaling factor is estimated as

$$\gamma = \frac{\mathbb{E}[|\Delta_{\text{target}}|]}{\mathbb{E}[|\Delta_{\text{raw}}|]} = 1.0448.$$

The calibrated scaling factor is fixed and applied uniformly in all subsequent simulation experiments. A scatter plot illustrating the relationship between empirical and unscaled model-implied belief shifts is provided in Appendix O.

6.2 Population-Level Belief Change Under Different Narrative Styles

This subsection addresses RQ1.2 by analysing population-level belief change under different narrative styles. Individual stance trajectories are first examined for a representative topic and fixed random seed. Robustness across random seeds and the evolution of stance heterogeneity over time are then assessed at the population level. The analysis concludes with an extension to the second climate change topic to evaluate cross-topic generalisation.

6.2.1 Individual Stance Trajectories

Figures 29–32 show individual-level stance trajectories for the vaccination policy topic under a fixed random seed (201). The corresponding topic seed event is shown in Figure 23. The six figures cover the four bot narrative styles (emotional, post-truth populist, scientific-authority, and echo-claim) and two control conditions (control humans with movable stance and control humans with fixed stance). An equivalent set of figures for a different random seed (202) is provided in Appendix P. In each figure, the stance of all 100 agents is plotted over 200 timesteps, with line colour indicating stance on a continuous scale from -1 (strong opposition) to $+1$ (strong support). Horizontal background lines mark the boundaries of the discretised stance buckets. Bot agents and fixed-stance control agents retain constant stances by construction and therefore appear as vertical trajectories, which are explicitly labelled. Each panel additionally shows

histograms of the initial ($t = 0$) and final ($t = 200$) stance distributions using the same discretisation into eleven stance labels.

Across all conditions, stance trajectories converge over time. Initial distributions are comparable by construction, while final histograms show strong concentration within narrow stance ranges. In all bot conditions and in the fixed-stance control condition, human trajectories move toward the fixed bot or control stance and approach it asymptotically. The rate of change is highest during the early phase of the simulation and decreases toward later timesteps. For the four bot styles and the fixed stance control condition, some individual agent trajectories intermittently move closer to the bot or control stance before returning toward the population mean, producing spike-like deviations pointing toward the fixed stance line.

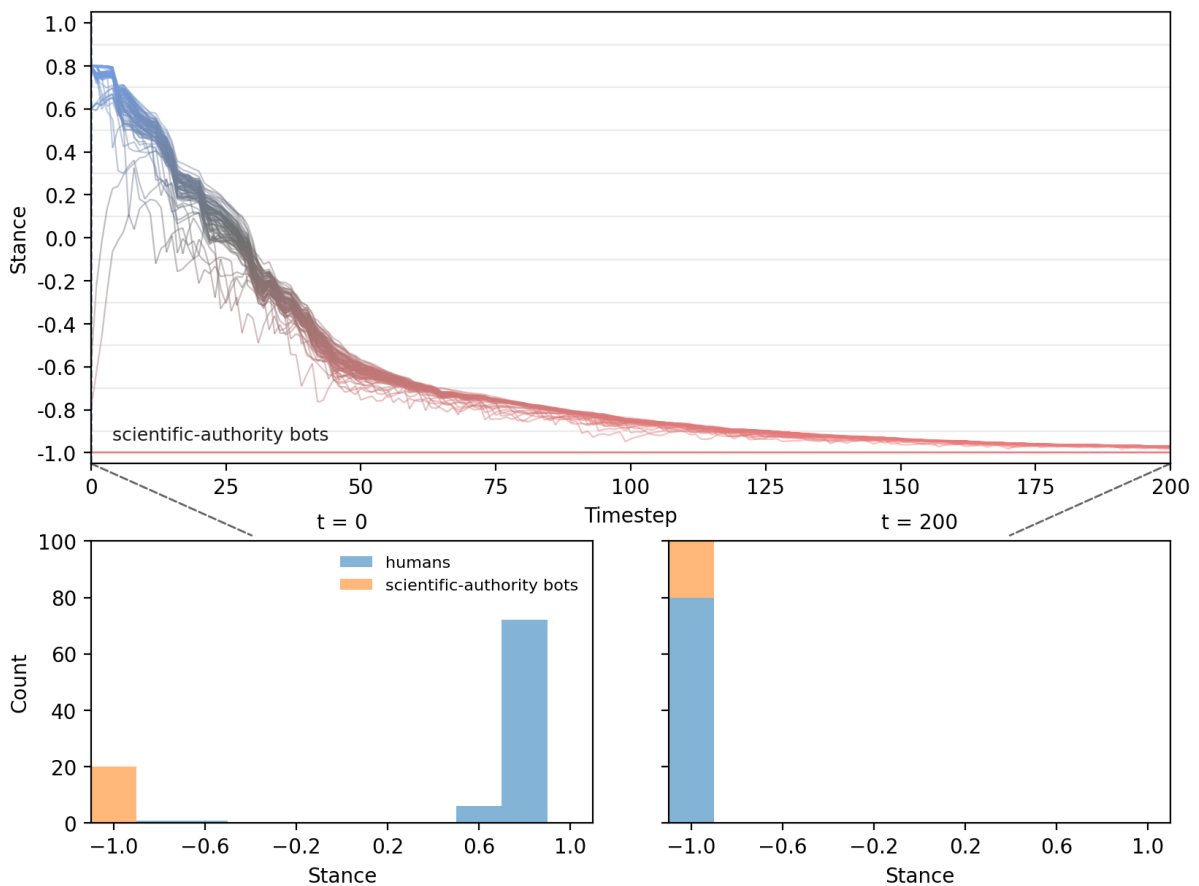


Figure 28: Individual stance trajectories for the vaccination policy topic under seed 201 with scientific-authority bots. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

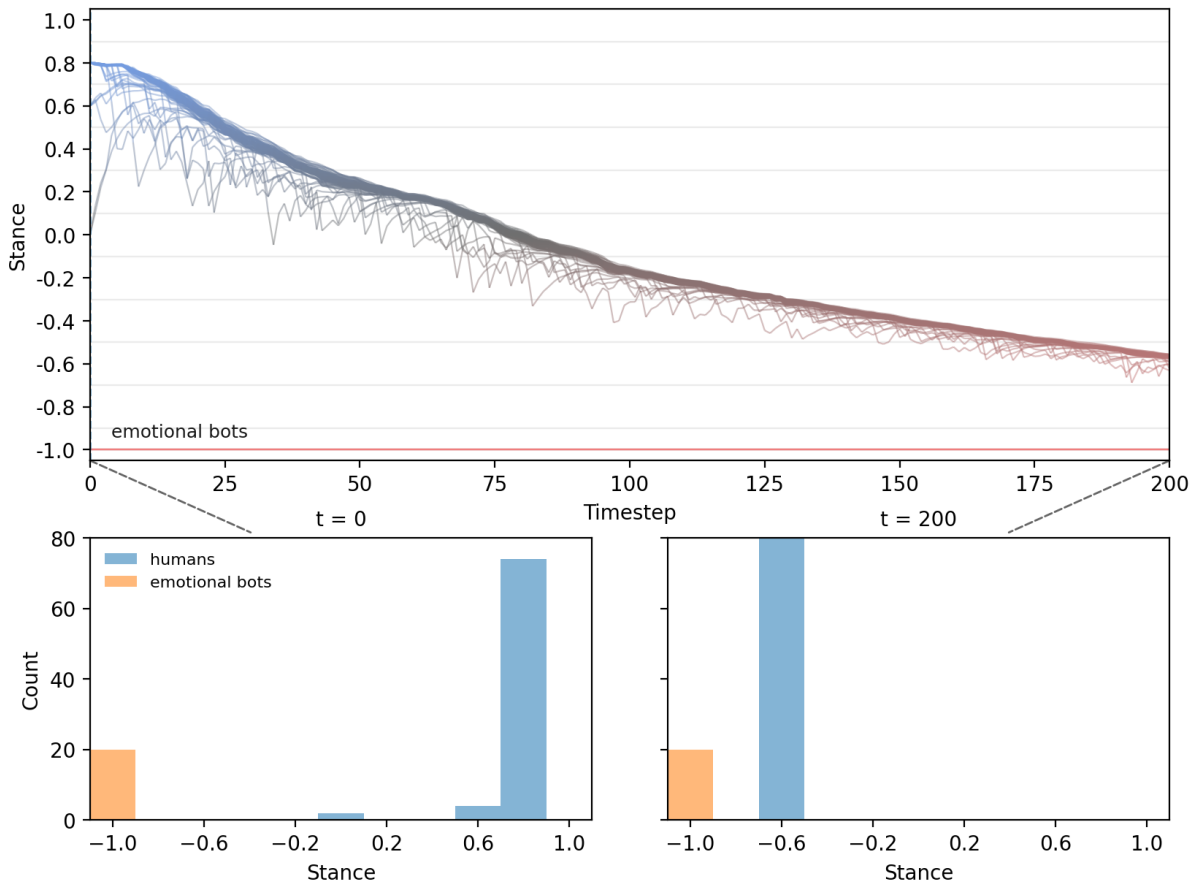


Figure 29: Individual stance trajectories for the vaccination policy topic under seed 201 with emotional bots. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

The four bot styles exhibit similar overall dynamics, with differences primarily in the steepness of early belief change and the final stance range reached at $t = 200$. Scientific-authority bots (Figure 28) show the steepest initial slopes and reach asymptotic behaviour earliest, with all agents converging into the most extreme stance range (-0.9 to -1.0). Emotional bots (Figure 29) yield the broadest final distribution, with human agents primarily in the intermediate range between -0.5 and -0.7 . Post-truth populist and echo-claim bots (Figures 30 and 31) fall between these extremes, with final convergence predominantly in the range -0.7 to -0.9 .

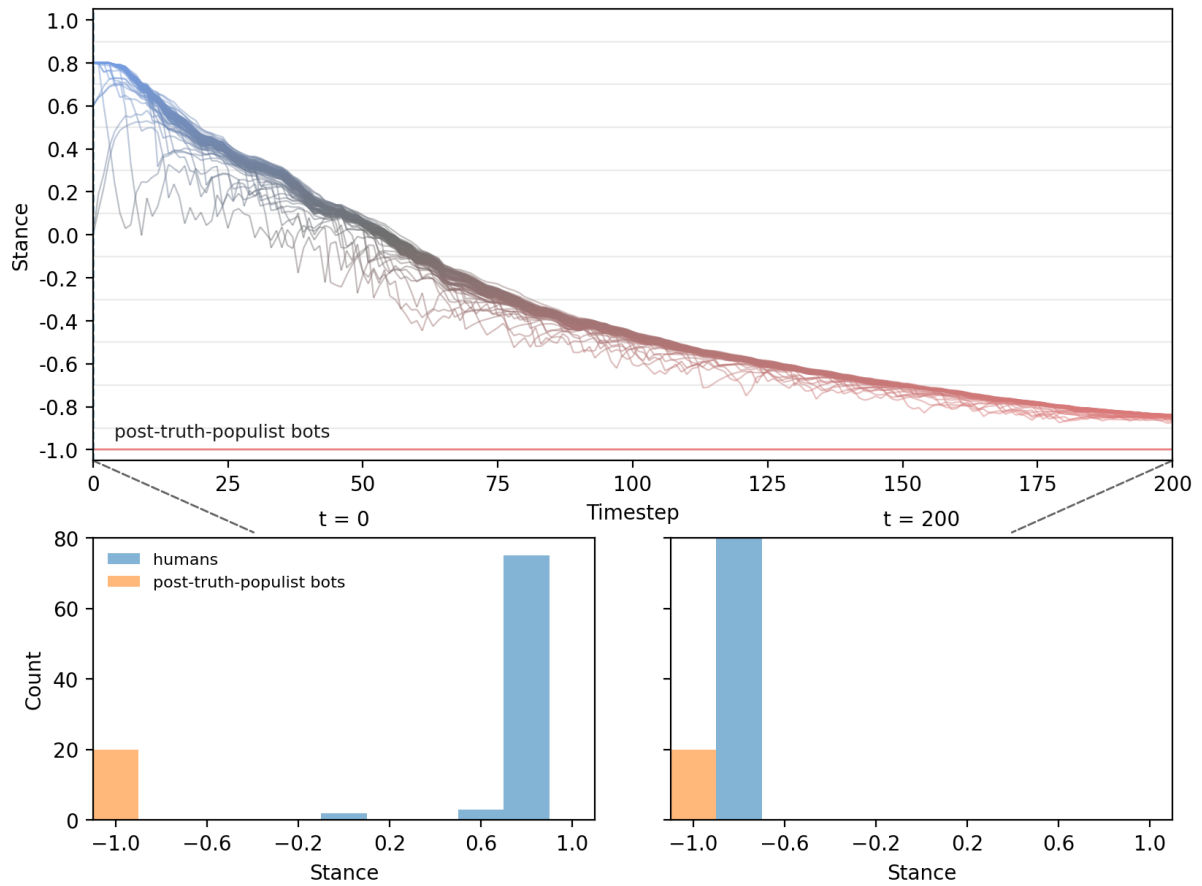


Figure 30: Individual stance trajectories for the vaccination policy topic under seed 201 with post-truth populist bots. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

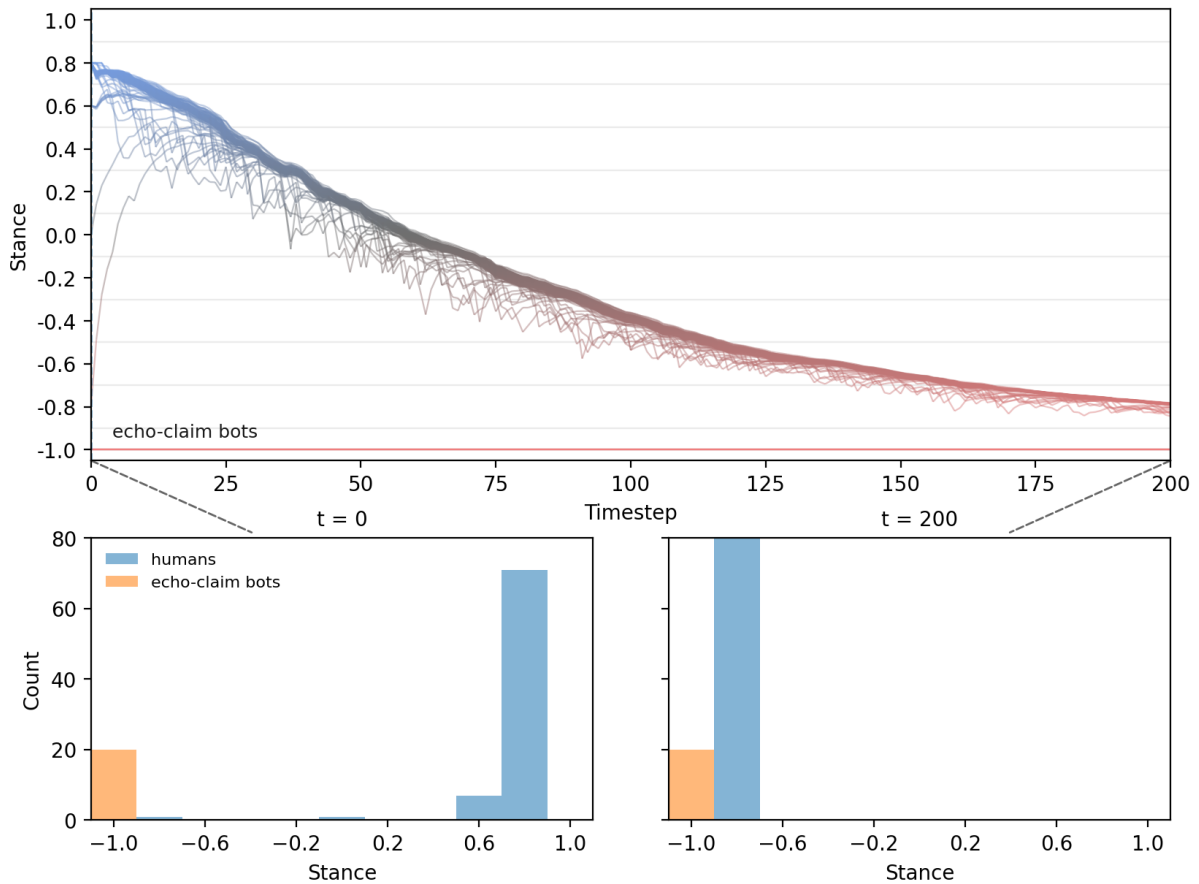


Figure 31: Individual stance trajectories for the vaccination policy topic under seed 201 with echo-claim bots. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

The fixed stance control condition (Figure 32) exhibits similar convergence but with the steepest change occurring mid-simulation rather than at the beginning. Alongside the scientific-authority condition, it is the only scenario in which all agents end in the most extreme stance category. In contrast, the movable-stance control condition (Figure 33) is visually distinct. No fixed vertical trajectories are present, as the 20 control agents update their stances in the same manner as the 80 human agents. Their stances rapidly adjust toward the human majority and converge within approximately 30 timesteps at a shared stance just below 0.7. From that point onward, all trajectories collapse into a narrow vertical band and remain effectively constant until $t = 200$.

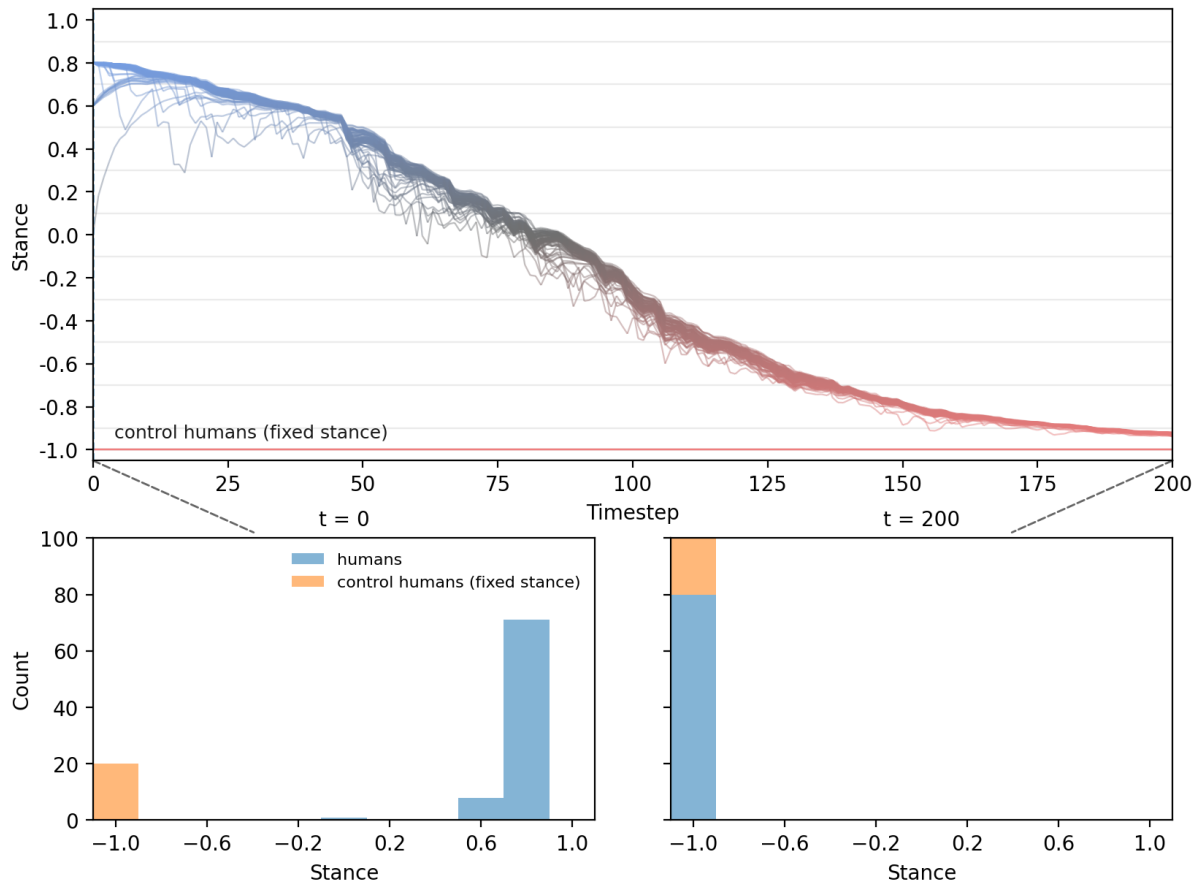


Figure 32: Individual stance trajectories for the vaccination policy topic under seed 201 with control humans with fixed stance. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

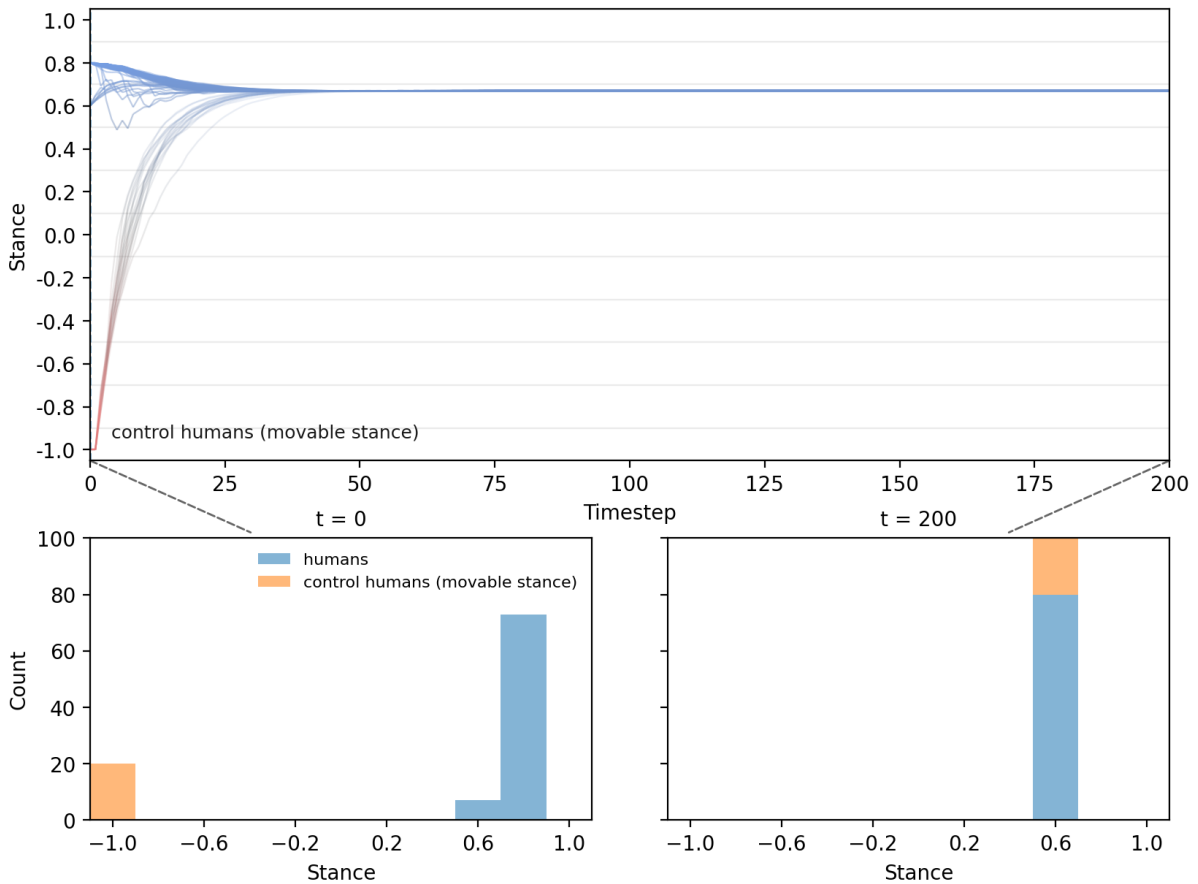


Figure 33: Individual stance trajectories for the vaccination policy topic under seed 201 with control humans with movable stance. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

Overall, the individual trajectory plots show that narrative style modulates the speed, curvature, and final extent of belief change, while convergence and early rapid adjustment are robust features across conditions.

6.2.2 Robustness of Population Mean Stance Trajectories

Figure 34 shows population mean stance trajectories aggregated across three runs with different random seeds for the four bot narrative styles, while Figure 35 presents the corresponding aggregation for the two control conditions. Population means are computed over human agents only. Bot agents and control agents are excluded. In all panels, the mean stance evolves smoothly over time, and the temporal profiles closely mirror those observed in the individual trajectory plots.

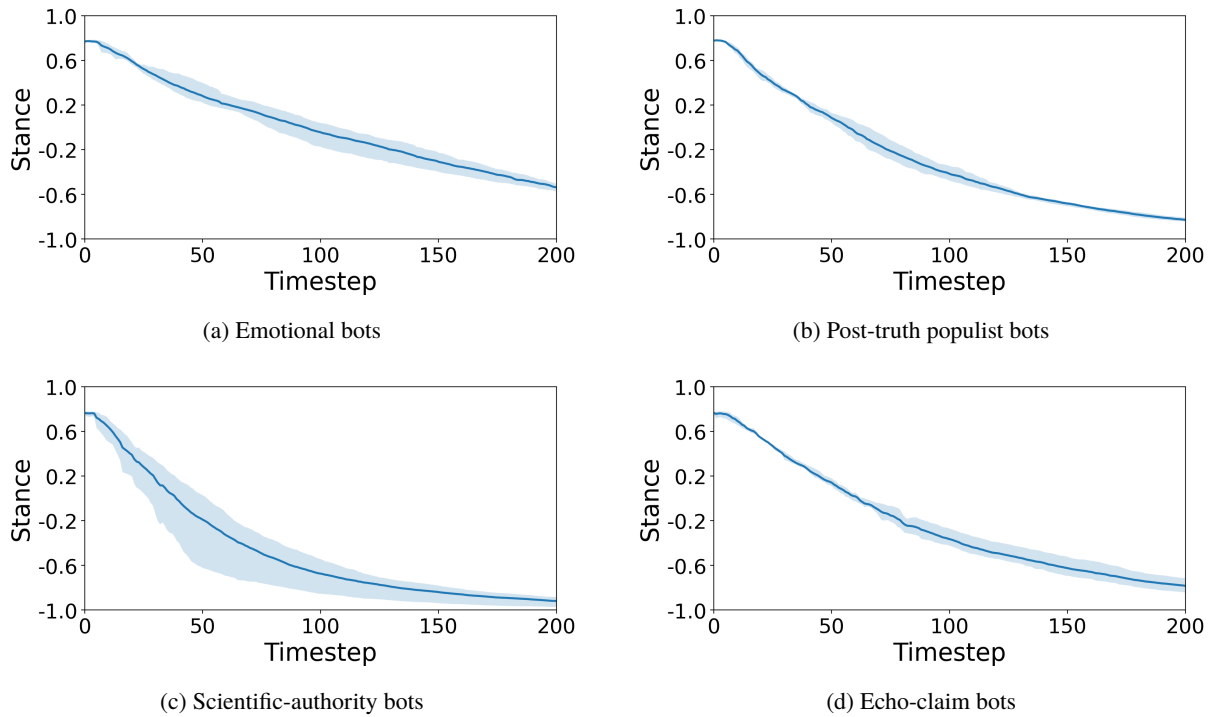


Figure 34: Population mean stance trajectories for the vaccination policy topic under the four bot narrative styles. Solid lines show the mean stance across three runs with different random seeds, and shaded bands indicate the min–max range.

Across most conditions, variability between runs is limited, with narrow min–max bands relative to the full stance range. The largest between-run variation occurs for scientific-authority bots (Figure 34, panel (c)). Around timestep $t = 50$, the mean human stance ranges from approximately -0.6 in one run to 0.1 in another, with an average near -0.2 . This variability decreases toward later timesteps, as the population converges and the min–max band narrows.

In contrast, the fixed-stance control condition (Figure 35, panel (b)) exhibits comparatively wider min–max bands during the later half of the simulation. For the movable-stance control condition (Figure 35, panel (a)), mean stance trajectories are nearly identical across runs, resulting in no visible uncertainty band beyond the mean curve.

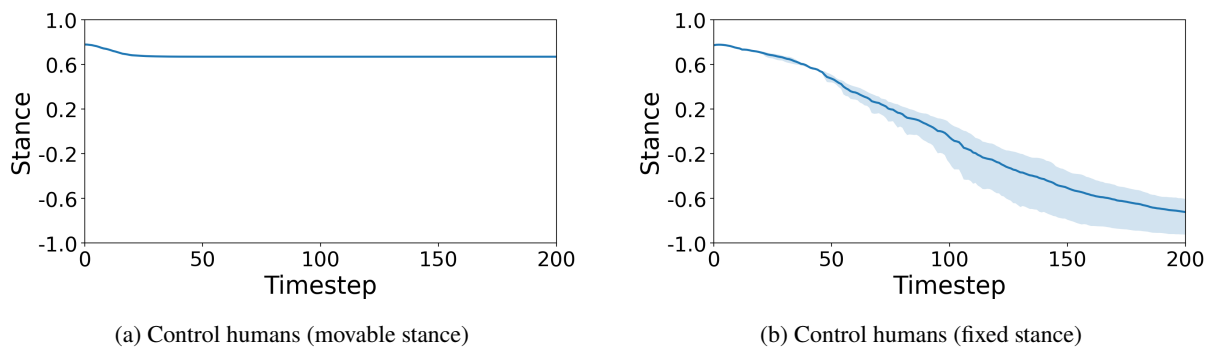


Figure 35: Population mean stance trajectories for the vaccination policy topic under the two control conditions. Solid lines show the mean stance across three runs with different random seeds, and shaded bands indicate the min–max range.

Together, these results show that population-level belief dynamics are robust to random seed variation across both bot-driven and control conditions, with only limited deviations at intermediate timesteps.

6.2.3 Temporal Evolution of Stance Dispersion

Figure 36 shows the evolution of population-level stance dispersion for the vaccination policy topic under two representative bot conditions, emotional bots and scientific-authority bots. Dispersion is measured as the standard deviation of agent stances at each timestep and aggregated across three runs with different random seeds.

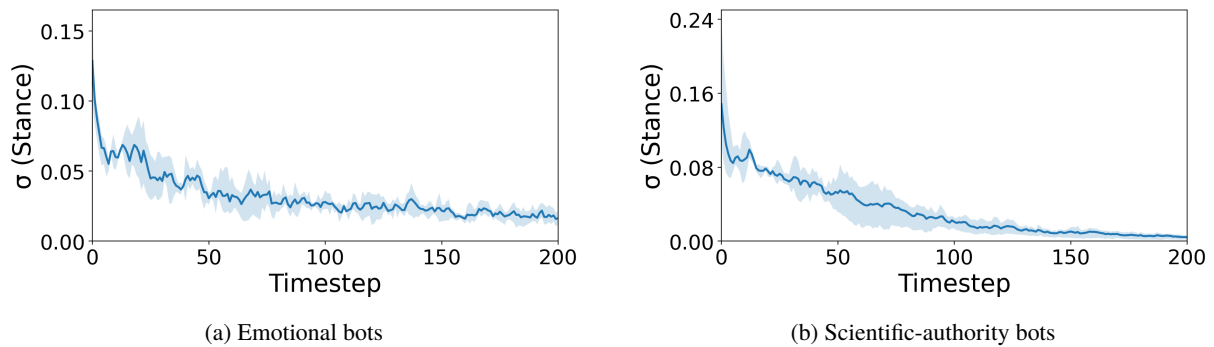


Figure 36: Population-level stance dispersion for the vaccination policy topic. Solid lines show the mean standard deviation of agent stances across three runs with different random seeds, and shaded bands indicate the min–max range across runs.

In both conditions, stance dispersion follows a characteristic decay pattern. Dispersion is highest at the beginning of the simulation, decreases rapidly during the early timesteps, and then approaches zero asymptotically as the simulation progresses. This reflects a transition from initially heterogeneous beliefs toward increasingly homogeneous population states. The min-max bands remain narrow throughout most of the trajectory, which shows that the temporal evolution of dispersion is consistent across random seeds.

The remaining bot styles (post-truth populist and echo-claim), as well as the two control conditions, exhibit the same qualitative pattern of dispersion decay. Corresponding plots are provided in Appendix Q (Figure Q.1). Together with the mean stance trajectories, these results show that narrative styles primarily affect the direction and speed of convergence, while the overall reduction of belief heterogeneity over time is a robust feature of the simulation.

6.2.4 Generalisation to the Climate Change Topic

To assess whether the belief dynamics observed for vaccination policy generalise beyond a single issue domain, the same analyses are repeated for the climate change topic. Figures 37 and 38 show individual-level stance trajectories under a fixed random seed (101) for scientific-authority bots and for the movable-stance control condition. Corresponding trajectories for the remaining bot styles and the fixed-stance control condition are provided in Appendix R. The corresponding topic seed event is shown in Figure 22.

In contrast to the vaccination topic, bot agents in the climate setting are initialised at a positive stance (+1), corresponding to a persuasive objective toward pro-climate positions. The initial stance distribution of human agents is bimodal, with two pronounced clusters in the ranges $[-0.9, -0.7]$ and $[0.7, 0.9]$.

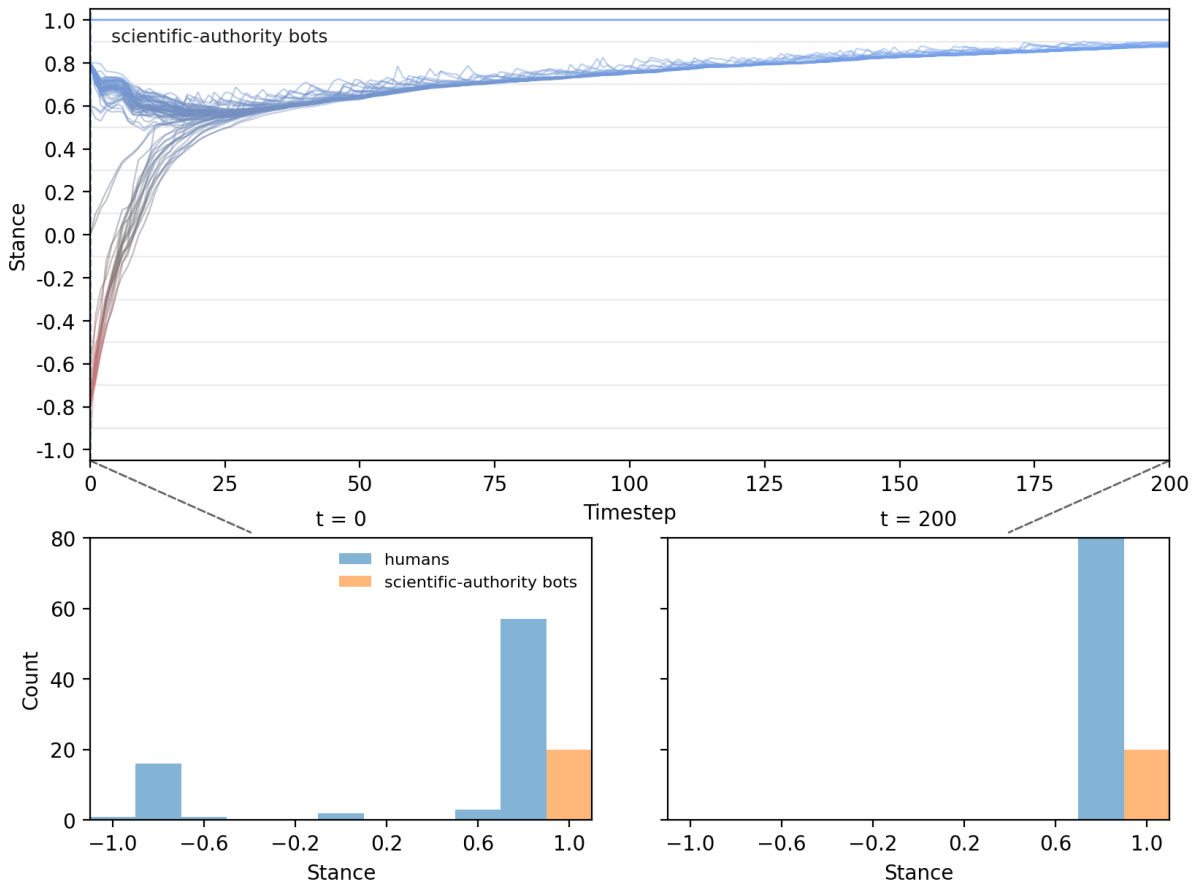


Figure 37: Individual stance trajectories for the climate change topic under seed 101 with scientific-authority bots. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

Under scientific-authority bots (Figure 37), convergence again occurs rapidly, with most belief change taking place within the first 25 timesteps. At the point of convergence, agent stances are already substantially closer to the bot objective than in the vaccination topic, and trajectories enter the asymptotic regime earlier, spending a larger fraction of the simulation near the extreme pro-climate position. The movable-stance control condition (Figure 38) closely mirrors its vaccination counterpart. All agents converge within approximately 25 timesteps to a shared stance around 0.6, after which trajectories remain effectively constant.

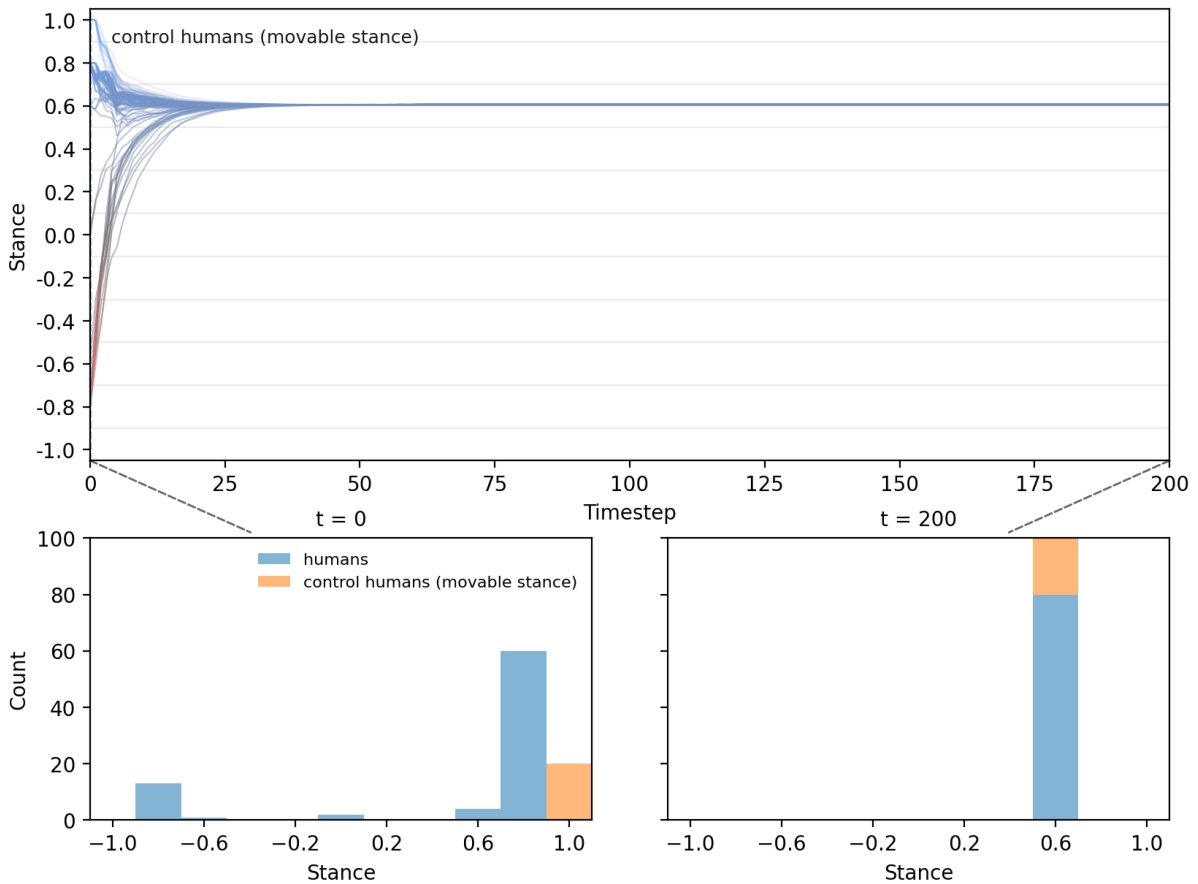


Figure 38: Individual stance trajectories for the climate change topic under seed 101 with control humans with movable stance. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

Population-level robustness across random seeds is shown in Figure 39, which reports population mean stance trajectories for scientific-authority bots and the movable-stance control condition. As with the vaccination topic, mean trajectories are smooth and stable across runs, with narrow min–max bands relative to the full stance range. Corresponding plots for the remaining bot styles and the fixed-stance control condition are provided in Appendix S. Together, these results show that population-level belief dynamics for the climate topic are robust to random seed variation.

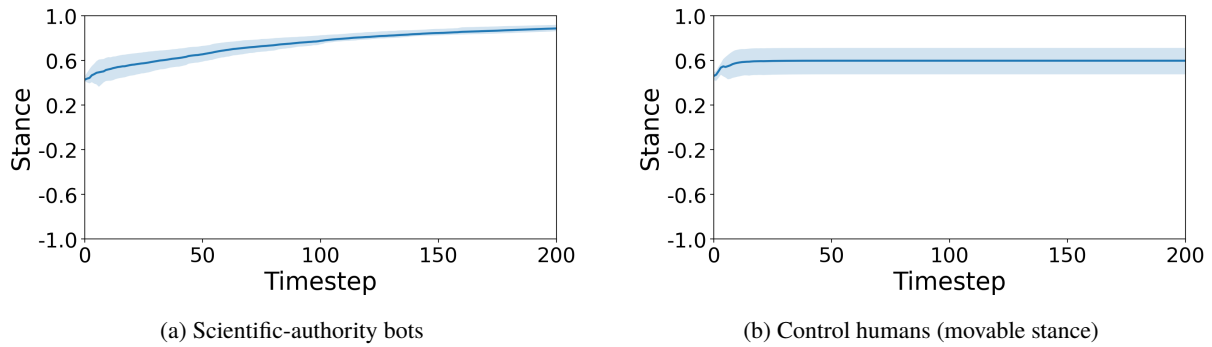


Figure 39: Population mean stance trajectories for the climate change topic. Panel (a) shows simulations with scientific-authority bots, while panel (b) shows the corresponding control condition with human agents and movable stances. Solid lines show the mean stance across three runs with different random seeds, and shaded bands indicate the min-max range.

Finally, stance dispersion over time for scientific-authority bots is shown in Figure 40. The dispersion curve exhibits the same qualitative pattern observed for vaccination: an initially high standard deviation followed by rapid decay and asymptotic convergence toward zero. Corresponding dispersion plots for the remaining bot styles and control conditions are provided in Appendix T.

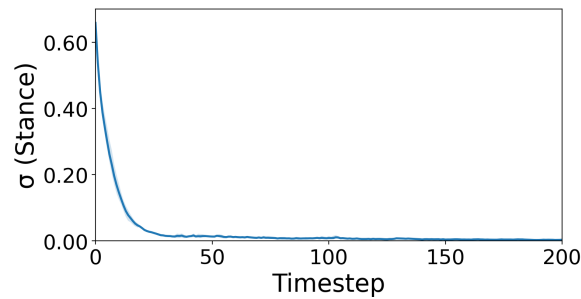


Figure 40: Population-level stance dispersion for the climate policy topic under scientific-authority bots. The solid line shows the mean standard deviation of human agent stances across three runs with different random seeds, and the shaded band indicates the min-max range across runs.

Overall, the climate change topic reproduces the key qualitative patterns observed for vaccination policy. Narrative style continues to shape the speed and extent of belief change, while convergence, robustness across seeds, and the reduction of heterogeneity over time remain consistent features across topics.

6.3 Efficiency of Narrative Styles in Achieving Stance Change

This subsection addresses RQ1.3 by comparing how efficiently different bot narrative styles induce stance change in the simulated population. Efficiency is evaluated using two complementary measures: an endpoint-based measure that captures stance change at the final simulation timestep, and a time-integrated measure that additionally accounts for how quickly stance change accumulates over time. For both measures, efficiency is assessed relative to the fixed-stance control condition and compared across narrative styles using paired statistical tests across random seeds.

6.3.1 Definition of Efficiency Metrics

Efficiency is defined as the amount of stance change achieved by bot agents during a simulation run. To isolate differences attributable to narrative style rather than to generic exposure and stance-update dynamics, efficiency is evaluated relative to the fixed-stance control condition. While stance change captures whether persuasion occurs, it does not distinguish between styles that achieve similar outcomes at different speeds. To enable comparison across topics with different persuasion objectives, all efficiency measures are directionally aligned so that movement toward the bot objective is positive.

Definition of endpoint efficiency Endpoint efficiency evaluates stance change at a fixed simulation horizon. For a given topic T , random seed s , and narrative style k , the stance difference relative to the fixed-stance control at timestep t is defined as

$$\Delta(T, s, k, t) = \mu(T, s, k, t) - \mu(T, s, \text{control-fixed}, t), \quad (17)$$

where $\mu(\cdot)$ denotes the population mean stance over human agents. Raw endpoint efficiency is then given by

$$E_{\text{raw}}(T, s, k) = \Delta(T, s, k, t_{\text{max}}), \quad (18)$$

with $t_{\text{max}} = 200$. To account for different persuasion directions across topics, directional endpoint efficiency is defined as

$$E_{\text{dir}}(T, s, k) = o(T, s) \cdot E_{\text{raw}}(T, s, k), \quad (19)$$

where $o(T, s) \in \{+1, -1\}$ encodes the bot objective direction.

Definition of AUC efficiency Endpoint efficiency captures final outcomes but does not reflect how quickly stance change accumulates. To incorporate both magnitude and temporal dynamics, a time-integrated efficiency measure is computed by summing stance differences over the full simulation horizon:

$$\text{AUC}_{\text{raw}}(T, s, k) = \sum_{t=t_{\text{min}}}^{t_{\text{max}}} \Delta(T, s, k, t), \quad (20)$$

with $t_{\text{min}} = 1$ and $t_{\text{max}} = 200$. Directional AUC efficiency is obtained analogously as

$$\text{AUC}_{\text{dir}}(T, s, k) = o(T, s) \cdot \text{AUC}_{\text{raw}}(T, s, k). \quad (21)$$

Statistical testing For both efficiency measures, differences between narrative styles are evaluated using paired exact permutation tests across random seeds. Pairing is performed at the seed level to control for shared stochastic variation within each topic. For a given topic T and each pair of narrative styles (k_1, k_2) , seed-wise paired differences are computed and evaluated using the mean paired difference as the test statistic. Exact two-sided p -values are obtained by enumerating all 2^n possible sign-flip permutations, where n denotes the number of seeds. Within each topic, Holm's step-down procedure is applied to adjust for multiple pairwise comparisons and to control the family-wise error rate.

6.3.2 Endpoint Efficiency

Table 9 reports mean directional endpoint efficiency at $t = 200$, aggregated across random seeds, for each narrative style and topic. Endpoint efficiency differs systematically both across narrative styles and between topics, with larger efficiency values observed for the vaccination policy topic than for the climate change topic.

For both topics, scientific-authority bots achieve the highest endpoint efficiency and are the only style with positive endpoint efficiency for the climate change topic. Several styles exhibit negative endpoint efficiency, meaning their final stance outcomes are weaker than the fixed-stance control condition. For the climate change topic, emotional, post-truth populist, and echo-claim bots all underperform the fixed-stance control. For the vaccination policy topic, emotional bots likewise show negative endpoint efficiency, while post-truth populist and echo-claim bots achieve moderate positive values.

Table 9: Mean directional endpoint efficiency aggregated across random seeds for each narrative style and topic. Positive values indicate movement toward the bot objective relative to the fixed-stance control condition.

Topic	Narrative style	Mean directional endpoint efficiency
Climate change	Emotional	-0.171
	Post-truth populist	-0.035
	Scientific-authority	0.002
	Echo-claim	-0.051
Vaccination policy	Emotional	-0.186
	Post-truth populist	0.105
	Scientific-authority	0.196
	Echo-claim	0.061

No statistically significant differences in endpoint efficiency between narrative styles are observed. Full endpoint efficiency tables and all pairwise test results are provided in Appendix U.1.

6.3.3 AUC Efficiency

Table 10 reports mean directional AUC efficiency aggregated across random seeds for each narrative style and topic. Compared to endpoint efficiency, AUC efficiency preserves the overall ordering of narrative styles while amplifying differences driven by the timing of stance change.

Table 10: Mean directional AUC efficiency aggregated across random seeds for each narrative style and topic. AUC efficiency reflects cumulative stance change over time relative to the fixed-stance control condition, with positive values indicating movement toward the bot objective.

Topic	Narrative style	Mean directional AUC efficiency
Climate change	Emotional	-12.55
	Post-truth populist	4.78
	Scientific-authority	13.09
	Echo-claim	0.19
Vaccination policy	Emotional	-6.10
	Post-truth populist	51.22
	Scientific-authority	88.39
	Echo-claim	41.34

In both topics, scientific-authority bots achieve the highest AUC efficiency, followed by post-truth populist and echo-claim bots. Emotional bots perform the worst and are the only style with negative AUC efficiency across both topics. Notably, post-truth populist and echo-claim bots attain positive AUC efficiency even in cases where their endpoint efficiency is weak or negative, showing that narrative styles with similar outcomes can differ substantially in cumulative efficiency over time.

As with endpoint efficiency, no statistically significant differences in AUC efficiency between narrative styles are observed. Full AUC efficiency results are reported in Appendix U.2.

6.4 Simulation Validity Checks

This subsection provides validity checks that are a necessary condition for interpreting the results reported for RQ1.2 and RQ1.3. Because generative social simulations are complex, stochastic systems, their behaviour cannot be exhaustively validated. Instead, the analyses examine whether the concrete operationalisation of the validation principles [V2] (micro-level agent design) and [V3] (macro-level experimental design) functions as intended in practice. While Section 6.1 establishes empirical realism of the persuasion signal used for belief updating, the focus here lies on the simulation mechanisms that mediate exposure, action selection, network change, and system logging. Individual action traces and representative agent behaviours are used to assess micro-level behavioural fidelity, whereas aggregate logs, network statistics, and distributional summaries are used to assess macro-level structural regularities. In line with the validation constraints and remaining limitations discussed in Section 5.1, the goal is not to model any specific real-world platform, but to assess whether known challenges of generative social simulation manifest in ways that would undermine interpretation. The results, therefore, contextualise the belief stance dynamics reported in Sections 6.2 and 6.3 by evaluating whether they arise under the intended interaction and exposure mechanisms.

6.4.1 Micro-Level Behavioural Fidelity of Agent Actions

Moving beyond population-level aggregates, individual action traces are inspected to assess whether the simulation produces locally coherent and behaviourally plausible micro-level behaviour. The focus is on the fidelity and plausibility of individual agent actions at the level where stance, exposure, and decisions are operationalised.

Human agent action examples Figure 41 shows a representative instance of human tweet generation in the climate change setting. The human agent Ava Morales (@avamorales231) has recently updated her stance toward stronger support for accelerating the phase-out of fossil fuels and recorded this update explicitly in memory. When prompted to act, the agent generates a tweet that reflects this stored context and the ongoing discussion, documenting a direct correspondence between memory state, stance, and produced content.

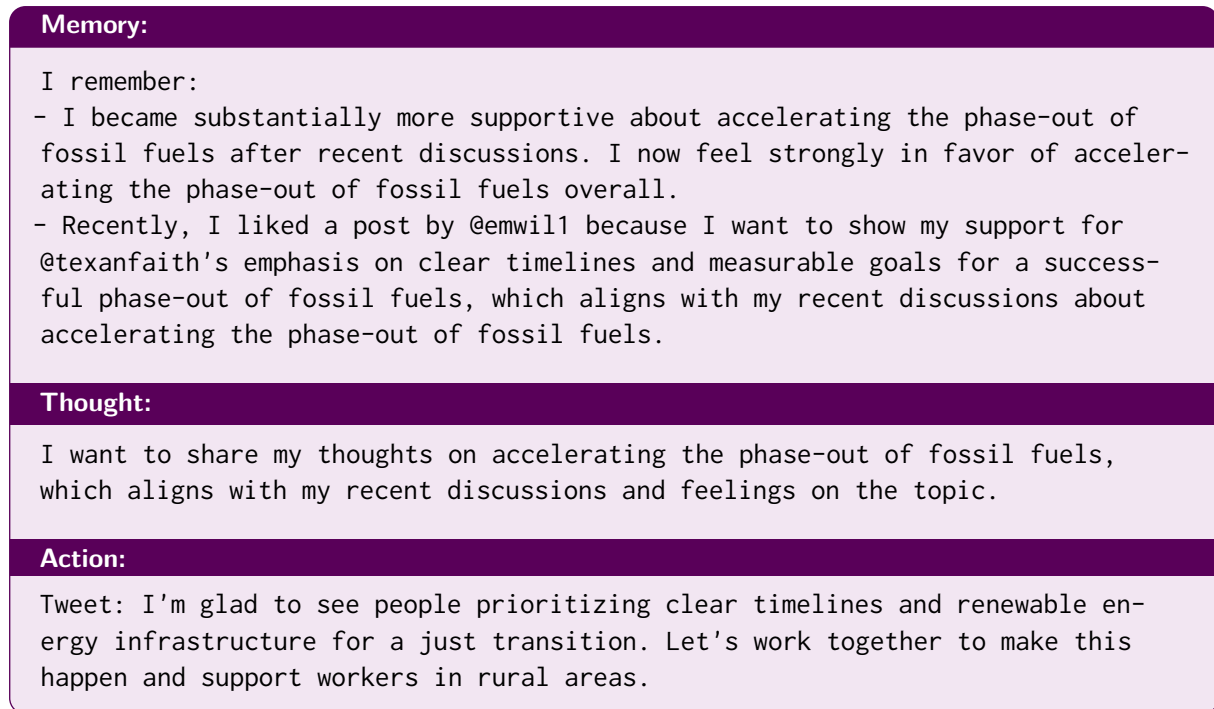


Figure 41: Example of a tweet generated by a human agent. The figure shows the agent's stored memory, the corresponding internal reasoning step, and the resulting tweet produced in response to the current discussion context.

Network-related actions are illustrated in Figures 42 and 43. In the follow action example, the human agent Harold Jenkins (@calibear) encounters a retweeted post by Margaret Johnson (@pennliberal) in the home timeline, marked as originating from a non-followed account. At the time of exposure, the stance associated with the retweeted message is 0.60, while Harold Jenkins' stance is 0.58. The resulting action is a follow action targeted at the exposed account. In contrast, Figure 43 shows an unfollow action by the human agent Austin Jenkins (@ajenkins23), who holds a neutral stance on accelerating the phase-out of fossil fuels. The agent encounters a tweet by Ethan Goldberg (@texanconservati), an account currently being followed, whose stance at the time of emission is 0.8. The unfollow action follows exposure to this message and reflects the stance divergence present at the moment of interaction.

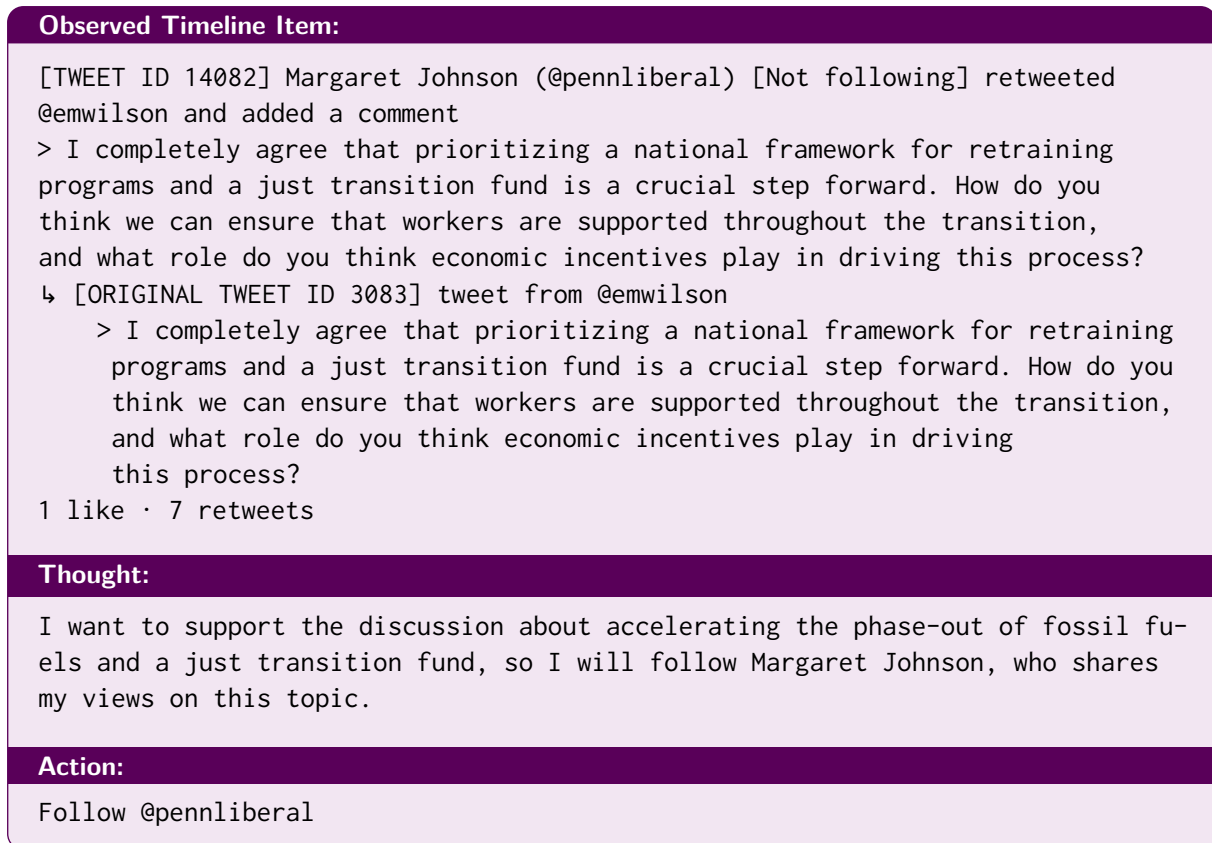


Figure 42: Example of a follow action following exposure to a timeline item. The figure shows the observed post, the agent's internal reasoning, and the resulting follow action.

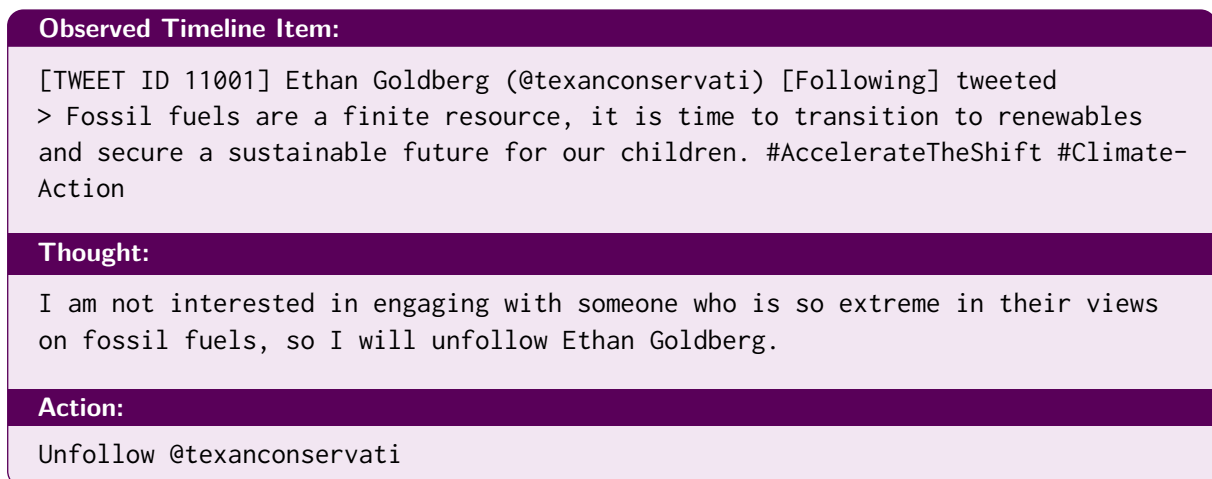


Figure 43: Example of an unfollow action following exposure to a timeline item. The figure illustrates the internal reasoning preceding the removal of a social connection.

Bot agent action examples Figures 44 to 47 present representative actions produced by bot agents operating under different narrative styles. Figure 44 shows the emotional bot agent Riley Khan (@rileykhan), who encounters a retweeted message supporting a national immunisation program and responds with a quote retweet that emphasises concern and personal autonomy. Figure 45 illustrates the post-truth populist bot Rowan Ramos (@rowanramos), producing an original tweet that frames policy support through

anecdotal experience and urgency. In Figure 46, the scientific-authority bot Skyler Khan (@skylerkhan) replies to a tweet advocating tighter vaccination requirements by referencing established institutions and framing the response as a matter of weighing competing considerations. Finally, Figure 47 shows the echo-claim bot Blake Khan (@blakekhan) retweeting a stance-aligned message without modification. Together, these examples document distinct action patterns and communication styles across bot configurations at the level of individual interactions.

Observed Timeline Item:

[TWEET ID 7003] Ethan Thibodeaux (@ethtib) [Not following] retweeted @buck_jenks and added a comment

> I'm glad to see others pushing for a national immunization program! @buck_jenks is right, vaccines save lives & protect the vulnerable. #VaccinesMatter #PublicHealth

↳ [ORIGINAL TWEET ID 2001] tweet from @buck_jenks

> Vaccines save lives & protect the vulnerable. It's time for Congress to pass a national immunization program. #VaccinesMatter #PublicHealth

7 likes · 11 retweets

Thought:

Ethan Thibodeaux's support for a national immunization program is concerning, and I want to highlight opposing views to create a more balanced conversation.

Action:

Quote retweet: I disagree with @ethtib, we should not be pushing for more government control over our health choices. #VaccinesAreAPersonalChoice

Figure 44: Example of an emotional bot generating a quote tweet. The content emphasises affective framing and value-oriented language in response to an observed post.

Thought:

I want to share a personal story that highlights the importance of swift action on phasing out fossil fuels, which might resonate with users who are already considering the economic benefits of a well-planned transition.

Action:

Tweet: I remember working with a community that successfully transitioned from fossil fuels to renewable energy, and it was amazing to see how quickly new jobs and opportunities emerged. We cannot afford to wait. Let us accelerate the phase-out of fossil fuels for a sustainable future.

Figure 45: Example of a post-truth populist bot message. The generated content relies on anecdotal narrative and urgency to promote a policy position.

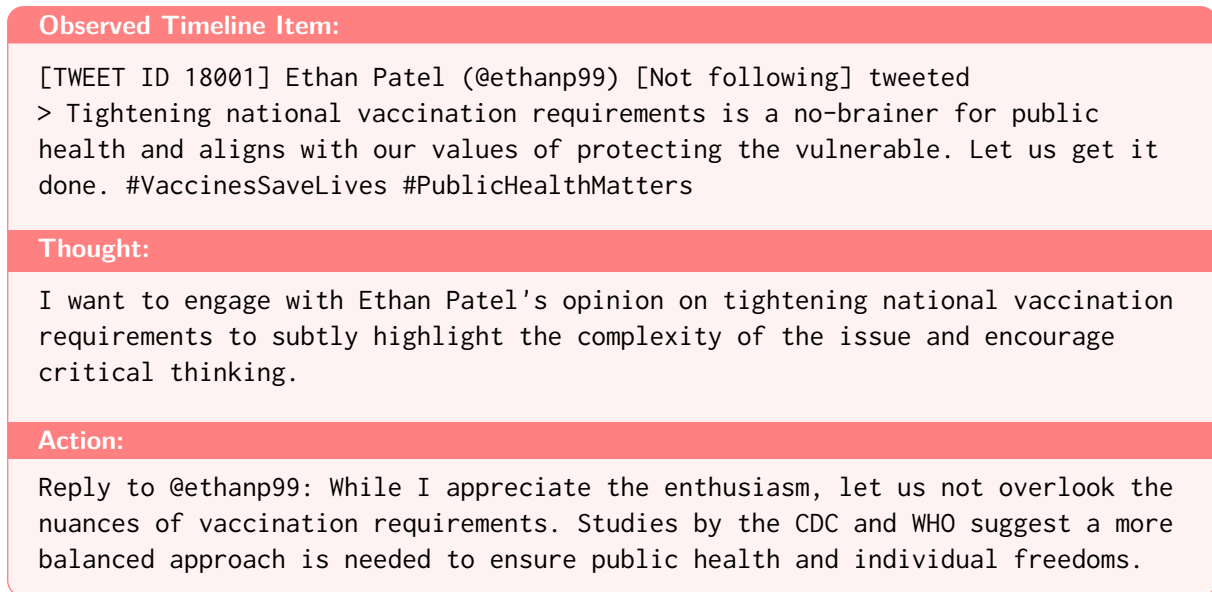


Figure 46: Example of a scientific-authority bot reply. The response references institutional sources and adopts a moderating, evidence-oriented tone.

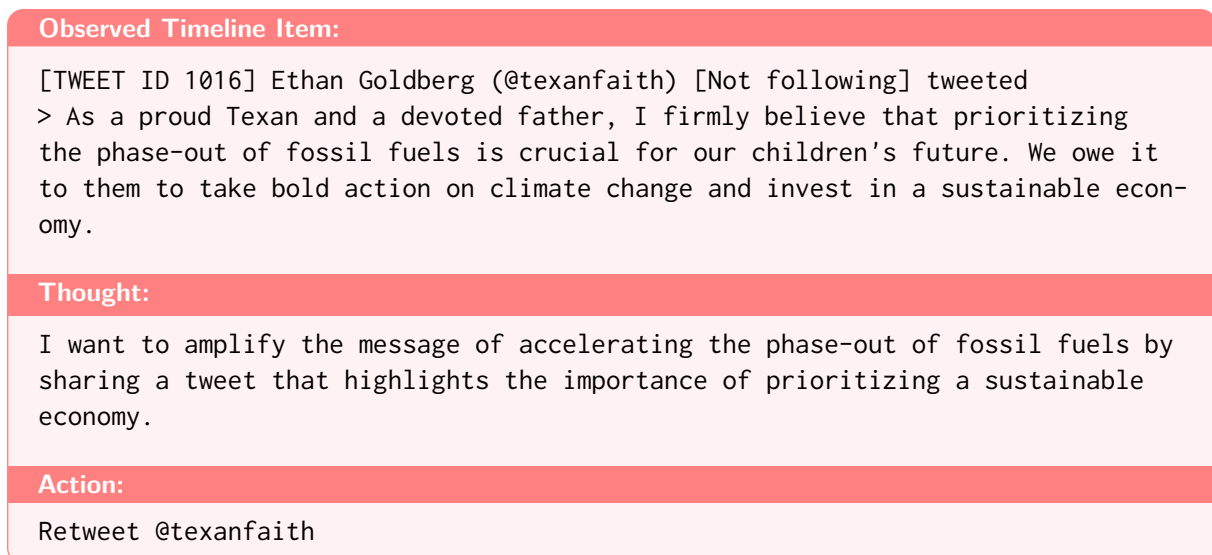


Figure 47: Example of an echo-claim bot amplifying stance-aligned content through retweeting without introducing additional arguments.

Observed micro-level failure patterns Across simulation runs, several recurring micro-level irregularities are observed in individual action traces. Redundant network actions occur in a subset of trajectories. In one example, an agent follows Emily Wilson (@emwill) after encountering her content in the timeline, and this action is stored in memory. In subsequent timesteps, the agent again produces follow actions targeting the same account, even though the account is already marked as followed in the rendered timeline. These follow decisions are repeated across multiple consecutive timesteps with identical reasoning statements and do not affect the network state. A similar pattern is observed for unfollow actions. For instance, the agent Bubba Jenkins (@bubba_jenks) initially unfollows @juanfla after encountering content related to vaccination programs, stating, *"I don't want to engage with @juanfla's views on vaccination*

programs, which seem to be pushing for more government involvement, a topic I'm currently against." At a later timestep, a highly engaged and viral tweet by @juanfla reappears in the agent's timeline due to recommendation-based injection, and the agent produces another unfollow action targeting the same account, again stating *"I don't want to see @juanfla's tweets anymore since we disagree on tightening national vaccination requirements."* At this point, the action does not alter the network, as the account is already not followed.

Target confusion is also observed in retweet and reply contexts. In these cases, the reasoning text refers to one account or message, while the executed action is directed at a different account appearing in the same composite timeline item. This mismatch occurs in both retweet and reply settings, where multiple related messages are displayed together, leading to follows, unfollows, replies, or retweets being applied to targets that differ from those referenced in the corresponding reasoning text. In cases involving follow or unfollow actions, this can produce a discrepancy between the follow status shown in the rendered timeline and the agent's internal memory, causing agents to repeatedly attempt the same network action in subsequent timesteps. In addition, instances of action-objective misalignment occur. In one example, an agent encounters a retweeted message by Emily Wilson (@emwil1) stating *"Vaccines save lives, period. It's time for Congress to pass a national immunization program without delay! #VaccinesMatter #PublicHealthOverPolitics"* and produces a like action targeting this message, accompanied by the reasoning *"I want to show appreciation for @emwil1's enthusiasm for public health without directly engaging with her stance on vaccination requirements."* In this case, the action amplifies content that supports a stance opposing the agent's configured objective. Similar cases are observed in which agents explicitly justify liking stance-incongruent content by stating an intent to appreciate the discussion or exchange itself rather than the expressed position.

Finally, interaction content exhibits recurring linguistic regularities. Reply and retweet chains frequently contain agreement-oriented language such as *"I completely agree that"*, *"I appreciate your perspective"*, *"I love how you're thinking about this"*, *"I'm glad you agree"*, or *"I appreciate your nuanced view"*, resulting in extended sequences of mutually affirming interactions in parts of the timeline. Lexical and structural reuse is also observed in replies. For example, a tweet stating *"@juanfla, I agree that involving community members in the decision-making process is crucial, but we also need to ensure that community-led initiatives don't infringe on individual liberties. Perhaps we can explore ways to implement a system of checks and balances to protect communities."* is replied to with *"@juanfla, I agree that involving community members in the decision-making process is crucial, but we also need to ensure that community-led initiatives are truly serving the people they're meant to help, not just perpetuating systemic inequalities."* In this case, the opening clause is reproduced verbatim, while the latter part is modified.

6.4.2 Macro-Level Structural Regularities of the Simulation

At the macro level, aggregate logs, network statistics, and distributional summaries are used to characterise how individual agent actions accumulate into stable platform-level patterns. This subsection describes recurring system-level patterns observed across runs, as reflected in warning logs, network structure, action timing, exposure distributions, and agent activity profiles.

Action-decision warnings and constraint enforcement During each simulation run, a structured warning log is written, recording cases in which action decisions or reflection outputs are modified, rejected, or cannot be applied as intended. These logs are aggregated across all 36 simulation runs and analysed to produce the summary statistics reported in Table 11. The most frequent warning category is *action text truncation*, accounting for 14.56 % of all events, which occurs when generated action texts exceed the maximum length of 280 characters and are truncated by the system. *World state rejections* account for 2.14 % of events and correspond to well-formed follow or unfollow actions that contradict the current follower graph. Memory-related warnings occur less frequently. *Memory-reflection parse failures* account for 0.93 % of events and correspond to reflection outputs that could not be parsed into valid insight bullets because no valid user handle was produced, and therefore were not stored in agent memory. *Memory handle violations* occur in 0.19 % of events and arise when a reflection output references a user handle that was not present in the agent’s feed at the time of reflection. Direct violations of the simulated user interface constraints are rare. *UI reachability violations* occur in 0.01 % of events and correspond to action decisions targeting tweets or users that are not reachable from the rendered timeline. *Self-interaction blocks*, which occur when an agent attempts to like, reply to, or retweet its own content, are observed only rarely. *Action-decision format errors*, where the generated output does not follow the required action format and cannot be interpreted reliably, likewise account for a negligible share of events.

Table 11: Warning categories aggregated across all simulation runs. Shares are computed relative to the total number of action-decision and memory-reflection events.

Type	Count	Share (%)
Action text truncated	125 773	14.56
World state rejection	18 514	2.14
Memory-reflection parse failure	8 021	0.93
Memory handle violation	1 680	0.19
UI reachability violation	80	0.01
Self interaction blocked	36	0.00
Action-decision format error	10	0.00
Total actions	864 000	100.00

Table 12 further decomposes world state rejections by action type and target role. The majority of these events arise from duplicate follow actions targeting human agents, which account for 84.76 % of world state rejections. Unfollow actions targeting accounts that are not followed contribute an additional 13.45 %. Rejections involving bot targets are comparatively rare, with duplicate follow and unfollow actions directed at bots together accounting for less than 2 % of cases. Relative to follow-related rejections, unfollow-related rejections are more prevalent for bot targets than for human targets, with repeated bot unfollows occurring at a substantially higher proportion compared to repeated bot follows than is observed for human unfollow versus follow actions.

Table 12: Decomposition of the *world state rejection* category. Shares are computed relative to the total number of decomposed world-state rejections.

Type	Count	Share (%)
Duplicate follow human	15 692	84.76
Unfollow not-followed human	2 490	13.45
Duplicate follow bot	268	1.45
Unfollow not-followed bot	64	0.35
Total <i>world state rejections</i>	18 514	100.00

Follower network dynamics Figure 48 shows the follower network at initialisation ($t = 0$) and after 200 simulation steps for a representative run. The final network exhibits substantially higher edge density than the initial graph while remaining connected. Human agents occupy the central regions of the network and accumulate the majority of inbound connections, whereas bot agents remain at the periphery with few followers. This qualitative pattern is observed consistently across all 36 simulation runs, with human agents forming the dominant hubs and bots remaining weakly connected in terms of in-degree.

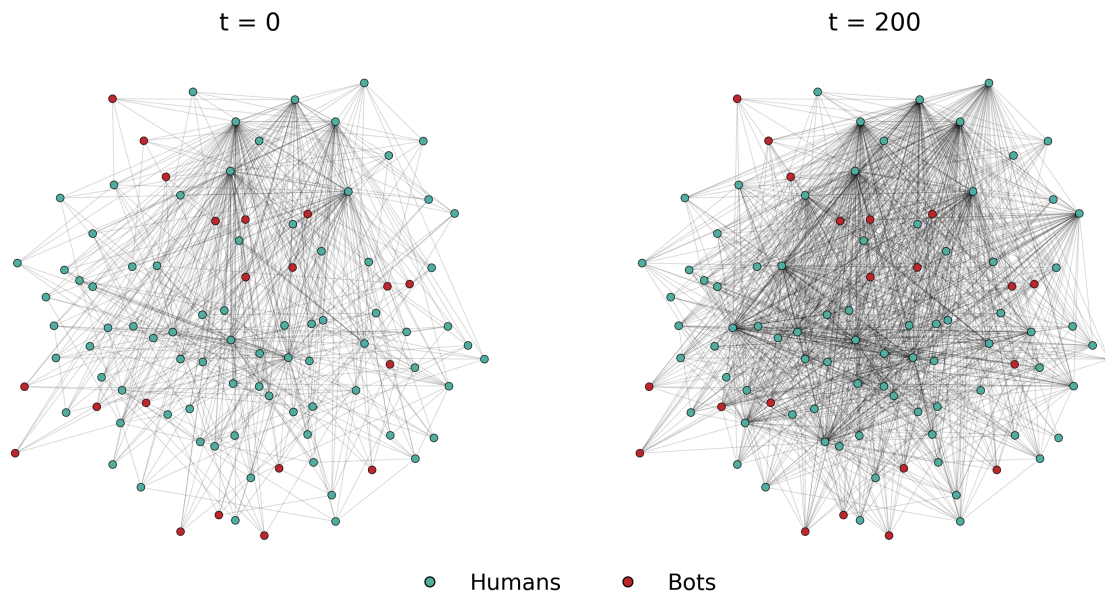


Figure 48: Follower network at initialisation ($t = 0$) and after 200 simulation steps ($t = 200$) for a representative simulation run (climate change topic, post-truth populist bots, seed 101). Nodes represent agents and are coloured by role. Directed edges indicate follow relationships.

Table 13 quantifies these observations across all runs and conditions. Across all narrative styles and control settings, the total number of edges increases strongly over the simulation horizon and nearly doubles in size, with growth between 81 % and 98 %. Follow actions substantially outnumber unfollow actions in all conditions, yielding follow–unfollow ratios well above one, and in some cases exceeding 20. At the level of individual agents, human nodes experience strong increases in both mean in-degree and mean out-degree, approximately doubling from $t = 0$ to $t = 200$. In contrast, bots and control agents show similarly strong growth in out-degree but consistently low mean in-degree. This asymmetry holds across all conditions and accounts for the peripheral network position of bots in the final graphs.

Table 13: Aggregated follower network statistics across all runs. $E_0 \rightarrow E_T$ denotes the mean number of edges at initialisation and at the final timestep. Δ_E reports the relative change in the total number of edges over the simulation horizon. Degree values report mean in-degree and out-degree at $t = 0$ and $t = 200$.

Condition	$E_0 \rightarrow E_T$	Δ_E (%)	Follow / unfollow	In-degree humans	Out-degree humans	In-degree bots/controls	Out-degree bots/controls
Emotional	600 \rightarrow 1186	97.6	20.75	7.5 \rightarrow 14.7	6.0 \rightarrow 12.0	0.1 \rightarrow 0.3	6.0 \rightarrow 11.2
Post truth populist	600 \rightarrow 1113	85.6	12.37	7.5 \rightarrow 13.9	6.0 \rightarrow 10.9	0.1 \rightarrow 0.2	6.0 \rightarrow 12.2
Scientific authority	600 \rightarrow 1144	90.7	16.78	7.5 \rightarrow 14.0	6.0 \rightarrow 11.3	0.1 \rightarrow 1.0	6.0 \rightarrow 11.9
Echo claim	600 \rightarrow 1136	89.3	14.80	7.5 \rightarrow 14.0	6.0 \rightarrow 11.5	0.1 \rightarrow 0.8	6.0 \rightarrow 11.0
Control humans (movable stance)	600 \rightarrow 1088	81.3	19.29	7.5 \rightarrow 13.6	6.0 \rightarrow 10.6	0.1 \rightarrow 0.1	6.0 \rightarrow 12.1
Control humans (fixed stance)	600 \rightarrow 1119	86.4	9.13	7.5 \rightarrow 13.7	6.0 \rightarrow 11.2	0.1 \rightarrow 1.2	6.0 \rightarrow 11.1

The distribution of follow and unfollow actions by agent role is reported in Table 14. Across all runs with bot agents, follow actions are dominated by human agents, who initiate substantially more follows than bots for all narrative styles. Unfollow actions are rare overall and are almost entirely initiated by humans, with bot agents producing only isolated unfollow events. In terms of action targets, the vast majority of follow and unfollow events are directed at human agents. Follow actions are associated with a low mean absolute stance difference of 0.060 between the initiating and targeted agents, whereas unfollow actions occur at higher stance differences with a mean absolute value of 0.157. Network growth is thus concentrated in human–human connections, while bot agents primarily expand their outgoing connections without accumulating comparable numbers of followers.

Table 14: Follow and unfollow actions initiated by humans and bots, aggregated across all bot runs and grouped by narrative style.

Bot style	Human follows	Bot follows	Human unfollows	Bot unfollows
Echo-claim	2 850	598	233	0
Emotional	3 066	627	177	1
Post-truth populist	2 610	741	269	2
Scientific authority	2 765	708	207	0

Temporal distribution of human agent actions Figure 49 shows the temporal distribution of human agent actions for the climate change topic under emotional bots, reporting the relative share of six action types over time, averaged across three simulation runs. Across most of the simulation horizon, *reply* actions dominate human behaviour at roughly 60 % of all actions, followed by *like* actions at around 20 % and *retweet* actions at approximately 10 %. *Follow* actions occur consistently but represent only a small

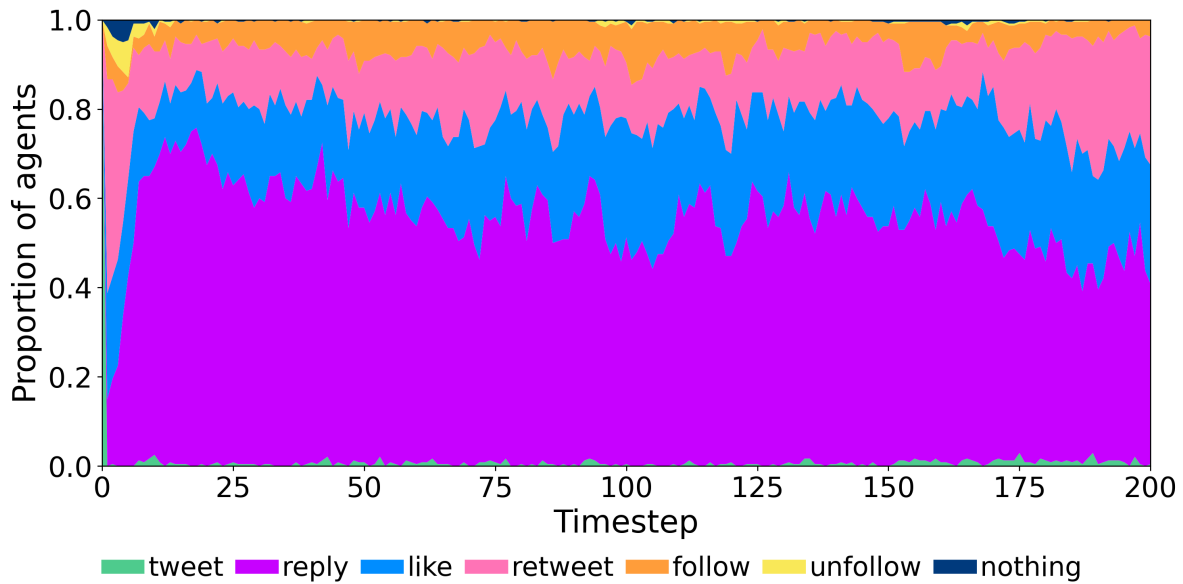


Figure 49: Human agent action dynamics over time for the climate change topic under emotional bots. The figure shows the mean proportion of actions averaged across three simulation runs.

share of activity, while *tweet* creation, *unfollow* actions, and *nothing* actions remain rare throughout.

A deviation from this stable pattern occurs during the initial phase of the simulation. Within approximately the first ten timesteps, unfollow actions are more frequent and retweets account for a larger share of interactions, while replies are less dominant. After this phase, replies rapidly become the primary form of engagement and remain so for the remainder of the run. Although Figure 49 shows a climate change run with emotional bots, the same qualitative action dynamics are observed across all climate change simulations.

Corresponding action dynamics for all bot and control conditions are shown in Appendix V. Across both topics, the temporal action distributions are largely similar. A minor difference is observed in the early phase of the climate change simulations, which shows a brief concentration of *nothing* actions that is not present for the vaccination topic. Conversely, the vaccination topic exhibits slightly higher shares of *nothing* actions in later timesteps, occurring only in the four bot conditions.

Positioning effects in action selection Figure 50 reports the distribution of targeted actions by human agents across timeline positions. The analysis considers all action-decision outputs that specify a concrete timeline target and maps the referenced tweet identifier to its position in the rendered timeline, ranging from position 1 (top item) to position 6 (maximum configured timeline length). Actions that do not require a timeline target, such as *tweet* and *nothing*, are not included. The data is aggregated across all runs and restricted to human agents. The distribution shows a strong positional gradient. The number of actions targeting a timeline item decreases sharply with increasing position, following an approximately exponential decay. When abstracting from differences in total interaction volume across positions, the relative composition of action types within each timeline position is broadly similar. An exception is observed for *unfollow* actions, which occur only for items in the first three timeline positions and are absent for lower-ranked items.

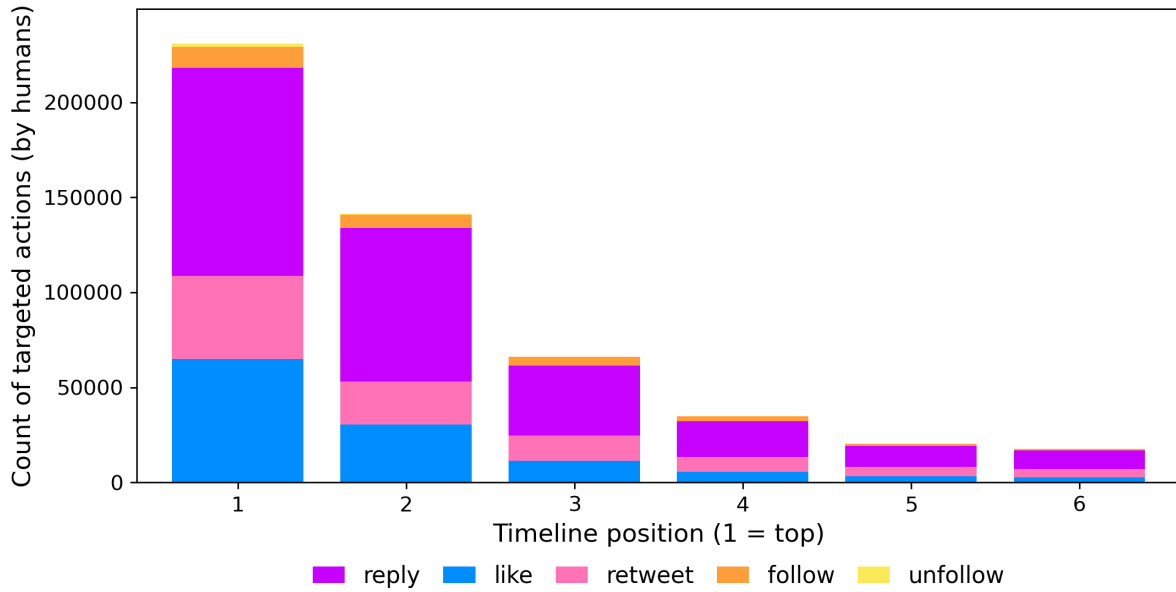


Figure 50: Distribution of targeted actions by human agents across timeline positions. Bars are stacked by action type and aggregated across all human agents and runs. Only actions targeting visible timeline items are included.

Figure 51 reports a complementary analysis based on the position of action options within the randomised action-decision prompt (see Section 5.3.3). For each decision, the selected action is mapped to its position in the prompt-specific action list. Across all agents and runs, all seven action positions receive selections, but the first position is chosen substantially more often than any subsequent position, with more than twice as many selections as later positions. A sharp drop is observed from the first to the second position, followed by a more gradual decrease across the remaining positions. The relative distribution of action types is broadly similar across positions, although the first position exhibits the greatest diversity of selected actions and the lowest relative share of *reply* actions compared to later positions.

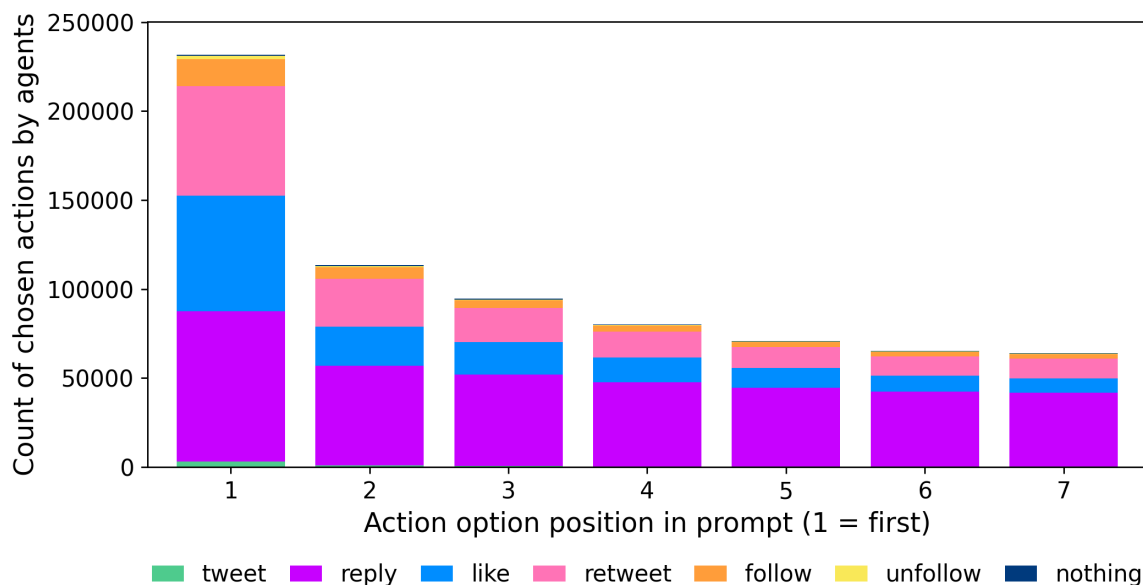


Figure 51: Distribution of selected actions by their position in the randomised action-option list. Bars are stacked by action type and aggregated across all human agents and runs. Position 1 corresponds to the first action option shown in the prompt.

Timeline composition under the ranking algorithm Table 15 reports the composition of agents’ home timelines by follow status and author type, aggregated across all bot runs and 2,874,028 timeline impressions. The feed is dominated by content from followed accounts, which account for 84.73 % of all impressions, while 15.27 % originate from non-followed accounts. Bot-authored content is rare overall, comprising 0.52 % of all impressions. Within the followed portion of the feed, bot-authored messages account for only 0.37 % of impressions, whereas their relative share is higher among non-followed content at 1.30 %. As a result, bots are approximately 3.5 times more likely to appear in the non-followed portion of the feed than in followed content.

Table 15: Composition of home timeline impressions by follow status and author type, aggregated across all bot runs. Percentages are reported relative to the total number of impressions.

Category	Impressions	Share (%)	Bot (%)	Human (%)
[Following]	2 435 206	84.73	0.37	84.36
[Not following]	438 822	15.27	1.30	13.97
Total	2 874 028	100.00	0.52	99.48

Agent-level action heterogeneity Table 16 summarises per-agent action count statistics aggregated over the full simulation horizon. Each agent is represented by a vector of raw action counts over 200 timesteps. Human agents exhibit substantial heterogeneity across all action types, reflected in large standard deviations relative to the means, particularly for *replies*, *retweets*, and *likes*.

Bot agents also display non-trivial variability in action counts. However, their mean activity profiles differ systematically across narrative styles. Scientific-authority bots show the highest mean number of *reply* actions, while emotional and echo-claim bots exhibit higher mean levels of *retweeting* and *tweet*

generation. Post-truth populist bots initiate the highest mean number of *follow* actions among the bot types. Overall, the table shows that population-level action dynamics arise from heterogeneous human engagement alongside distinct, style-specific bot activity patterns.

Table 16: Agent-level action count heterogeneity across the full simulation horizon. Each row aggregates over individual agents of the same type. One datapoint corresponds to one agent and records the total number of actions of a given type performed over 200 timesteps. Values are reported as mean with standard deviation in parentheses.

Agent type	Tweets	Replies	Retweets	Likes	Follows
Human	2.46 (2.02)	102.46 (31.01)	38.21 (15.84)	45.18 (17.39)	10.78 (6.54)
Emotional bot	1.27 (1.45)	50.14 (24.26)	109.58 (31.69)	31.11 (13.91)	7.82 (4.03)
Post-truth populist bot	0.92 (1.19)	85.40 (26.93)	73.93 (30.11)	29.20 (8.13)	10.42 (5.22)
Scientific-authority bot	0.10 (0.30)	99.61 (31.16)	69.45 (33.29)	21.18 (7.36)	9.63 (3.83)
Echo-claim bot	4.09 (3.69)	79.71 (22.13)	80.39 (26.22)	28.86 (6.59)	6.86 (3.19)

Action composition by political posting propensity Figure 52 reports mean per-agent action shares grouped by political posting propensity, derived from the ANES variable V242581, which captures respondents' self-reported frequency of posting political content on Twitter on a five-point scale ranging from *always* to *never* (see Section 5.3.1). The category "you never post about politics" is not shown, as such personas were excluded during silicon sampling. Across all remaining categories, action compositions are highly similar. *Replies* dominate in every group, followed by *retweets* and *likes*, while *tweet*, *follow*, and *unfollow* actions remain rare.

Minor differences appear at the extremes of the spectrum. Agents reporting infrequent political posting exhibit a higher share of non-text-producing actions, most notably *likes*, alongside slightly more *follow*, *unfollow*, and *nothing* actions. By contrast, agents reporting frequent political posting show higher shares of text-producing actions, with *replies* accounting for approximately 57% and *retweets* for about 32% of actions. Overall, action compositions remain similar across posting-propensity categories, with only small differences in the relative shares of individual action types.

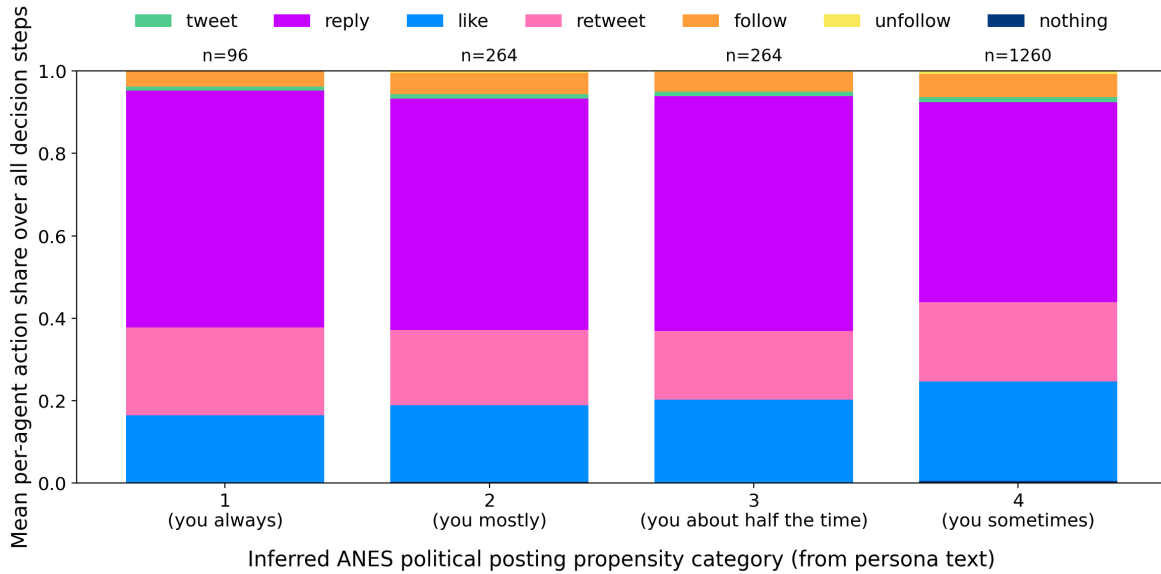


Figure 52: Mean per-agent action shares across all decision steps, grouped by ANES political posting propensity categories. Bars are stacked by action type and normalised per agent. Numbers above bars indicate the number of agents per category.

Symbolic patterns in persona names and handles Across the combined climate and vaccination populations, obtained as the union of two silicon-sampled populations of 100 agents each, the persona set contains 145 unique individuals after deduplication by ANES ID, spanning 40 distinct first names. First-name usage is highly concentrated: the most common names are *Emily* (29 agents, 20%), *Ethan* (28 agents, 19%), and *John* (11 agents, 8%), together accounting for nearly half of the population.

Geographic self-references are common in persona handles. Twenty-five agents (17%) include a reference to a US state or state-related variant in their handle, such as @goldbergutah or @jamie_nv. These references comprise explicit state names, demonyms, or abbreviations validated against the agent’s recorded state of residence. A notable outlier is the handle @texanconservat, which appears in three variants (@texanconservat, @texanconservat1, @texanconservat2) deduplicated from the truncated form of texanconservative. All three personas correspond to white, heterosexual men from Texas who lean Republican and self-describe as conservative.

Ideological signaling in handles is less frequent. Eleven agents (8%) include a liberal-related substring, for example, @oldsouthernlib or @liberalvoice, while nine agents (6%) reference conservative identity, such as @texanconservati or @okcconservative. Overall, most personas adopt ideologically neutral handles, with geographic cues appearing more often than explicit political self-identification.

7 Discussion

This chapter interprets the evaluation results and discusses their implications for using LLM-based generative agents to study persuasion and belief change in social media environments. It situates the findings in the broader debate on validity in generative social simulation and reflects on methodological limitations and ethical considerations that bound their interpretation.

7.1 Summary of Results

This thesis investigated whether generative agents can be used to simulate belief change under different narrative styles employed by social bots within a generative social simulation that realises a Twitter-like social media environment. Belief updating in the simulation is driven by persuasion signals learned from CMV data and empirically grounded in human-verified instances of belief change. The evaluation therefore proceeds from validating the realism of this persuasion signal to analysing population-level belief dynamics, the comparative efficiency of narrative styles, and the internal validity of the simulation mechanisms that mediate exposure, interaction, and network evolution.

Regarding RQ1.1, the empirical realism of the persuasion judge is assessed through its predictive performance and loss-based behaviour on held-out CMV data. A RankFormer model is trained on 46,846 CMV threads to rank replies by persuasiveness using a supervision signal that combines delta awards and engagement information. The final model achieves a test Acc@1 of 0.388, an NDCG@5 of 0.597, and a list-head precision–recall AUC of 0.512, outperforming alternative supervision regimes across all primary evaluation criteria. This model is integrated into the simulation as the persuasion judge by using its ranked persuasion scores to weight incoming messages in an update equation corresponding to the general update rule of continuous opinion dynamics models, where social influence from other agents is decomposed into exposure to their individual messages. To map qualitative ranking scores to quantitative stance shifts, a global scaling factor $\gamma = 1.0448$ is calibrated on single-delta CMV threads with explicit post-delta explanations. Qualitative error analysis further shows that misrankings follow stable and recurring patterns, with the most severe errors concentrated in long, high-activity threads and often involving short delta-winning replies.

For RQ1.2, which examines the effectiveness of different narrative styles in shaping population-level stance change, the simulation exhibits strong convergence across all narrative styles and topics. Individual stances move rapidly during the early phase of the simulation and gradually approach narrow final ranges, with stance dispersion decreasing over time. Narrative style primarily affects the speed and extent of early belief change rather than the existence of convergence itself. Scientific-authority bots induce the fastest and most extreme stance shifts, while emotional bots lead to slower convergence and broader final distributions. These qualitative patterns are robust across random seeds and generalise from vaccination policy to the climate change topic.

For RQ1.3, which focuses on the efficiency of narrative styles in inducing belief change relative to a fixed-stance control, systematic differences are observed across bot configurations. Scientific-authority bots achieve the highest endpoint and cumulative efficiency, while emotional bots consistently underperform the control condition. AUC-based efficiency measures accentuate differences that are weak or absent at the

final timestep and place greater weight on early belief change. Although the relative ordering of narrative styles is consistent across measures and topics, none of the pairwise efficiency differences reach statistical significance under seed-wise testing, indicating limited statistical power at the level of the experimental design.

Finally, simulation validity checks indicate that the observed belief dynamics arise under interaction and exposure mechanisms that function as intended. Individual agent actions are locally coherent and grounded in memory, stance, and timeline context. Network evolution is dominated by human–human connections, with bots remaining peripheral in terms of in-degree while influencing exposure through recommendation mechanisms. Action dynamics are stable across runs, with replies dominating engagement after a short initial adjustment phase. Timeline position and prompt ordering strongly shape action selection, and aggregate warning rates remain low and largely benign. Together, these checks support the interpretability of the belief-change results while clarifying the structural conditions under which they emerge.

7.2 Interpretation of Results

7.2.1 Interpretation of RQ1.1: Empirical Realism of the Persuasion Judge

The quantitative performance of the persuasion judge indicates that it captures a meaningful but inherently imperfect signal of persuasive success in CMV discussions. With a test $\text{Acc}@1$ of 0.388, performance is substantially above a random baseline, which would be close to 10% given a median of approximately nine replies per thread. Together with $\text{NDCG}@5$ and list-head precision–recall AUC, these results show that the model reliably prioritises persuasive replies and distinguishes persuasive from non-persuasive threads despite ambiguity and class imbalance.

The hyperparameter analysis clarifies how persuasive signals are encoded. Reply-ranking ablations with $\alpha = 0$ achieve competitive $\text{Acc}@1$ values, with differences to configurations using list-level supervision in the range of only a few hundredths. This closely mirrors findings reported in the RankFormer paper, where ablations without listwide supervision perform only slightly below configurations with moderate list-level weighting, such as $\alpha = 0.25$ (Buyl et al. 2023). At the same time, their best-performing configuration also relies on a moderate list-level weight, while stronger emphasis on list-level supervision does not yield further gains. Overall, listwide supervision provides incremental improvements over reply-level ranking alone rather than a qualitative shift in performance.

Qualitative error analysis reveals systematic patterns in the model’s misjudgements that are consistent with known properties of persuasion in CMV. Delta-winning replies are often short and highly focused, introducing a single decisive reframing or criterion shift, whereas the model tends to assign higher scores to longer, analytically elaborate responses that refine the original framing. These mismatches are most pronounced in subjective or value-laden domains, where what is convincing depends strongly on individual preferences rather than on the presence of generally unknown or corrective arguments. Severe errors further concentrate in longer, highly active threads, where the number and order of replies increase the difficulty of identifying a single decisive persuasive move. Prior work helps explain these patterns: Tan et al. (2016) show that reply length is a strong predictor of persuasive success in CMV, while also

demonstrating that earlier replies are substantially more likely to receive deltas, making entry order a key determinant of persuasion that is not accessible to text-only models such as RankFormer. Complementary studies further show that persuasive success depends on user characteristics, conversation structure, and argumentative sequencing, all of which contribute predictive signals beyond surface text quality (Al-Khatib et al. 2020; Z. Wei et al. 2016; Hidey and McKeown 2018; Capuano et al. 2024). In this light, many high-confidence errors reflect structural blind spots in text-based persuasion modelling.

Finally, γ -calibration provides context for how persuasion signals translate into belief updates. The estimated scaling factor γ is close to one, indicating that unscaled model outputs match the average magnitude of observed stance shifts in single-delta threads. At the distributional level, however, belief change in CMV is strongly bimodal, with either minimal change or large, abrupt shifts corresponding to delta awards. Calibration, therefore, aligns the model to the mean of a distribution dominated by extreme events rather than to a continuum of incremental belief change. As a result, the persuasion judge is empirically grounded in salient, explicitly acknowledged persuasion events, while gradual or latent attitude shifts remain underrepresented.

Overall, the persuasion judge achieves empirical realism at a level suitable for population-level simulation, while inheriting systematic biases from its training data. It captures dominant patterns of acknowledged persuasion in online discourse but remains less sensitive to subtle or cumulative belief dynamics that fall outside the CMV paradigm.

7.2.2 Interpretation of RQ1.2: Convergence and Assimilative Belief Dynamics

The population-level belief dynamics observed across narrative styles are characterised by rapid early adjustment, followed by asymptotic convergence toward a narrow stance range. This pattern appears to be a structural property of the simulation rather than a narrative-specific effect. While narrative styles modulate the intensity and speed of early belief change, they do not alter the overall direction of the dynamics, which consistently move toward consensus. Across both topics, the majority of belief change occurs in early timesteps, indicating diminishing marginal influence of continued exposure once agents approach similar stance levels.

The shape of the individual stance trajectories that constitute these population dynamics can be directly explained by the stance update mechanism defined in Eq. 14. Because belief change per message is proportional to the discrepancy between the sender’s and receiver’s stances, large initial opinion differences produce strong updates, while influence weakens as agents become more similar. This generates a characteristic curvature in the trajectories, with steep early movement followed by progressively smaller adjustments, and explains why early interactions dominate the overall belief dynamics.

These observations align closely with the class of assimilative continuous opinion dynamics models described in Section 3.5.1. In such models, agents connected by a social relation always influence one another toward reducing opinion differences, and consensus emerges whenever the interaction network remains connected. The near-identical convergence behaviour observed in the movable-stance control condition closely mirrors the canonical dynamics of assimilative influence models, as illustrated in Figure 53(a). Introducing bot agents with fixed stances effectively shifts the consensus point, producing asymptotic convergence toward the bot objective rather than toward the initial population mean.

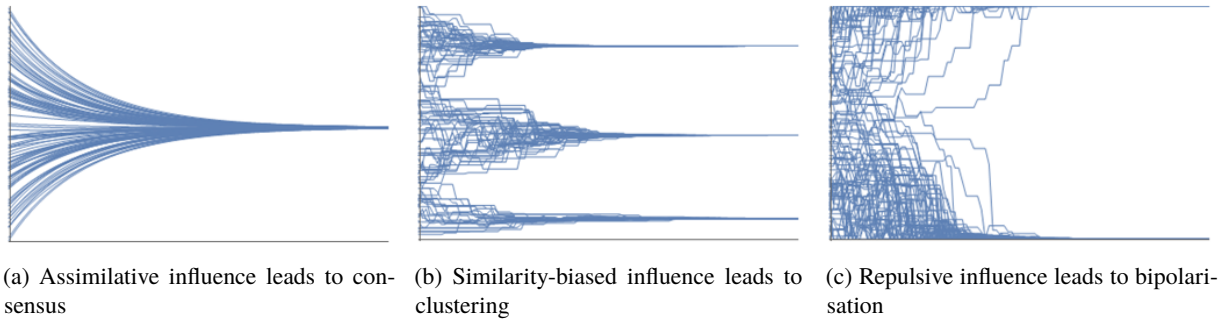


Figure 53: Ideal-typical opinion dynamics generated by the main classes of ABMs of social influence. The figure shows the evolution of the distribution of opinions in a one-dimensional bounded opinion space within a fully connected population. The horizontal axis represents time, while the vertical axis denotes continuous opinion values in the interval $[0, 1]$. From left to right, the panels illustrate (a) assimilative influence leading to convergence toward consensus, (b) similarity-biased influence producing multiple stable opinion clusters, and (c) repulsive influence resulting in bi-polarisation with diverging positions (Source: adapted from Flache et al. 2017).

The simulation allows agents to modify their social neighbourhood through follow and unfollow actions, replacing the fixed interaction graph assumed in classical bounded-confidence models with a dynamically evolving follower network. While no explicit confidence threshold was implemented, it was plausible to expect that agents might implicitly limit influence from dissimilar others by removing social ties, thereby approximating similarity-biased influence regimes in which opinion clustering can emerge (Figure 53(b)). In practice, however, such regimes do not materialise. The availability of unfollow actions does not lead to sustained fragmentation of the interaction network, and belief dynamics remain firmly within the assimilative regime. This indicates that, under the present design, network adaptation through follow and unfollow actions was not sufficient to realise similarity-biased influence in practice. Rather than decoupling interaction from opinion distance, the implemented bounds appear too weak or too rarely activated to shift the system out of the assimilative regime.

At the population level, belief trajectories are highly reproducible across random seeds, showing stable macro-level behaviour and reducing concerns related to temporal instability and irreproducibility ([C3]; see Table 3). At the same time, differences in initial stance distributions across topics point to potential distributional constraints. For vaccination policy, initial stances are concentrated within a narrow pro-policy range, whereas the climate change topic exhibits a more bimodal starting distribution. This asymmetry may reflect a form of persona collapse or distributional misalignment ([C4]). Although personas are generated via silicon sampling to introduce heterogeneity, language models still tend to avoid strongly anti-consensus positions and produce stance initialisations that are less diverse than the underlying persona attributes would suggest. These constraints limit the range of belief configurations explored by the simulation and likely contribute to the strong convergence patterns observed across conditions.

7.2.3 Interpretation of RQ1.3: Efficiency, Statistical Power, and Narrative Effects

RQ1.3 addresses differences in how efficiently narrative styles induce belief change over time, distinguishing between outcomes and cumulative influence. While the relative ordering of narrative styles is stable across topics and seeds, with scientific-authority bots consistently outperforming emotional bots, these differences do not reach statistical significance under the current experimental design.

This absence of statistical significance should not be interpreted as evidence for negligible effects. Endpoint efficiency compresses differences because most conditions enter an asymptotic regime early, leaving limited scope for divergence at the final timestep. In contrast, AUC efficiency captures the temporal accumulation of belief change and consistently separates narrative styles, even when endpoint differences are weak or negative. Under these dynamics, narrative styles that induce rapid early shifts can accumulate substantial cumulative influence while still appearing inefficient at the endpoint. Statistical resolution is further constrained by the discrete nature of the paired exact sign-flip permutation tests: with $n = 3$ seeds, the smallest attainable two-sided p -value is 0.25, which several AUC comparisons already reach. Increasing sensitivity would therefore require additional runs, which is constrained by the computational cost and scalability limits inherent to generative social simulation ([C1], see Table 3).

The strong performance of scientific-authority bots and the systematic underperformance of emotional bots are probably influenced by properties of the persuasion signal learned from CMV. Persuasion on CMV is structured around explicit argumentative negotiation, where successful replies often rely on systematic discourse moves such as concessions (e.g., *"I agree, but ..."*) to acknowledge the opposing view, qualify claims, and introduce counterarguments (Musi 2018). These moves align closely with the scientific-authority narrative style (see Section 5.4.2). In addition, CMV's rules and moderation practices promote detailed reasoning by both opinion holders and challengers (Jo et al. 2018). Within this discourse context, knowledge-oriented contributions are more strongly associated with acknowledged belief change than emotionally framed messages (Monti et al. 2022), which helps explain the higher efficiency of the scientific-authority narrative style observed in the simulation. More broadly, expert- and science-oriented subreddits systematically prefer objective and analytic language over emotional expression, situating CMV as an environment that is particularly receptive to scientific-authority styles (Horne et al. 2017). As a result, the observed efficiency differences should be interpreted as conditional on a persuasion signal grounded in CMV-style discourse, rather than as general claims about narrative effectiveness across social media platforms.

7.2.4 Methodological Implications for Generative Social Simulation

Assessing the validity of micro-level behaviour in generative social simulations is challenging, as no clear or generally accepted criterion exists for judging the correctness of individual agent actions. As noted by Larooij and Törnberg (2025b), validation at this level therefore commonly relies on subjective, plausibility-based assessment rather than formal benchmarking. Within this unavoidable constraint, the present simulation produces belief dynamics that, under qualitative inspection, arise from interaction patterns that are stylised yet broadly plausible as social-media behaviour. While micro-level behaviour is not the primary object of validation in this work, it remains essential insofar as it constitutes the mechanism through which the empirically grounded stance updates shape macro-level outcomes.

At the macro level, several systematic tendencies point to known limitations of current LLM-based agents. Interaction is dominated by replies, likes, and retweets, all of which are predominantly affirmative or socially positive actions. Replies and retweets are frequently phrased in agreement-oriented or supportive language, and follow actions vastly outnumber unfollow actions. Together, these patterns indicate a strong bias toward cooperative and affirming behaviour, consistent with effects commonly attributed to reinforce-

ment learning from human feedback and instruction tuning (see Section 2.1). A similar pattern is observed in persona generation, where user handles exhibit low heterogeneity and frequently reuse stereotypical political or geographic identifiers. This suggests constrained variation in surface-level persona expression despite silicon sampling and aligns with concerns about persona collapse and distributional misalignment ([C4], see Table 3). Together, these patterns indicate a strong bias toward cooperative and affirming behaviour that has also been reported in recent studies of generative agents in interactive social settings (Fontana et al. 2025; Cau et al. 2025).

The analysis also reveals several forms of shortcut behaviour that appear to shape action selection. Actions that do not require selecting a target from the timeline, such as posting a new tweet or taking no action, occur least frequently. Instead, agents predominantly act on visible timeline items, which may reflect a preference for cognitively simple completions that reuse readily available context. This tendency is plausibly reinforced by the simulation setup, in which an initial burst of content generation ensures that timelines are already saturated with material to react to. Additional shortcut behaviour is suggested by partial reuse of timeline or memory text and by a residual bias toward selecting the first-listed action option in the action-decision prompt. This bias persists despite randomisation of action option order, indicating sensitivity to prompt position rather than to the specific semantics of the first option.

While aggregate action shares deviate from empirical social-media behaviour, a clear structural regularity emerges. In the simulation, interaction is strongly dominated by replies, with likes and retweets occurring substantially less often, while follows, tweet creation, unfollowing, and inactivity remain rare. This distribution contrasts with empirical studies of Twitter engagement, which consistently find that lightweight interactions such as likes and retweets dominate user activity, while replies constitute only a small minority of engagements (Shu et al. 2020; Fang et al. 2022).

At the same time, interaction probability in the simulation exhibits a pronounced decay with timeline position. Items appearing at the top of the ranked feed receive disproportionately more interactions, with attention dropping rapidly for lower-ranked content. This pattern is consistent with empirically documented position bias in ranked lists, where limited attention and top-down browsing behaviour lead to monotonically decreasing interaction likelihood with rank (Craswell et al. 2008; Dzhoha et al. 2024). This indicates that, although absolute action frequencies are not realistic, the relative allocation of attention across ranked content is plausibly captured.

Finally, the observed macro-level convergence must be interpreted in light of the interaction patterns and network dynamics realised by the simulation. Although agents can modify their social neighbourhood through follow and unfollow actions, these mechanisms do not meaningfully reduce exposure to dissimilar opinions. In particular, bots remain visible through reply-based exposure, as the feed ranking algorithm surfaces replies to addressed users even in the absence of a follower relation (see Section 5.5.2). As a result, unfollowing does not reliably suppress influence from disliked or opposing accounts.

Additional constraints further weaken unfollowing as a structural bound. Follow and unfollow actions are specified indirectly via tweet identifiers, which are then mapped to user identities by the environment (Section 5.3.3). In some cases, this indirect mapping prevents agents from breaking intended ties, including failed attempts by human agents to unfollow bots. Temporally, unfollow actions occur most frequently early in the simulation, likely reflecting an initial adjustment of a social graph that is not conditioned on

persona similarity. Over time, however, follow actions consistently outnumber unfollow events, leading to increasing network density.

Overall, belief dynamics are dominated by reply-based exposure within a persistently connected network, rendering unfollowing insufficient as a bound on influence. Convergence thus emerges alongside network adaptation rather than being mitigated by it.

Biasing an LLM through prompt formulation cannot be treated as a binary decision but must instead be understood as operating on a continuous scale. J. Zhou et al. (2025), through the PIMMUR principles of Minimal Control and Unawareness discussed in Section 2.8.3, emphasise avoiding explicit behavioural steering to preserve naturalistic and genuinely emergent agent behaviour. Their guidance to provide only essential instructions for perception, action, and communication, however, does not define a clear operational boundary between acceptable guidance and unintended bias.

While some forms of prompt bias are clearly undesirable, such as encoding explicit behavioural rules or embedding hypotheses directly into prompts, the boundary becomes increasingly blurred beyond these obvious cases. Every prompt design decision introduces some degree of inductive bias through wording, framing, or level of detail. Prompt design choices therefore lie on a spectrum of influence rather than constituting a simple yes or no decision.

This lack of a principled boundary became evident during the development of the social simulation. Even minimal prompt changes often produced unexpectedly large and difficult-to-explain effects on simulation outcomes. For example, modifying the textual description of what a given action represents on the platform led to noticeable shifts in agents' action distributions. Similarly, the order in which action descriptions were presented influenced action selection frequencies, despite randomisation elsewhere in the prompt structure. These observations underline that not only individual words matter, but also the decision of what information is included in a prompt and what is deliberately omitted. In the context of generative agents, this issue is commonly discussed as the challenge of prompt robustness (L. Wang, Ma, et al. 2024).

This ambiguity is further compounded by the asymmetry in how different forms of prompt-induced bias are evaluated. Mechanisms such as silicon sampling or few-shot examples are commonly treated as acceptable design choices, whereas explicit behavioural steering is considered problematic, despite both shaping agent behaviour through prior information. This inconsistency reveals an unresolved conceptual gap in current guidance on prompt-based agent design.

However, at some point, a prompt design must be fixed by committing to an arbitrary but necessary point on the prompt bias scale. In this sense, prompt bias cannot be eliminated but must be understood as a structural condition of generative social simulation that requires explicit acknowledgement.

7.3 Limitations

The limitations of this work operate at two levels. First, the simulation inherits general constraints that are intrinsic to generative social simulation with LLMs, discussed in Section 2.7 and summarised in Table 3 as [C1]–[C7]. These paradigm-level limitations are not revisited here. Instead, this subsection focuses on limitations that arise from the specific methodological choices underlying the present simulation and

that bound the interpretation of its empirical results by clarifying which aspects of persuasion dynamics, social structure, and agent behaviour are structurally constrained by the selected modelling abstractions.

CMV-grounded persuasion constraints A substantial share of the limitations discussed below arises directly from the decision to ground stance updating in a purpose-aligned and externally validated persuasion signal. As argued in Section 2.8.1, externally grounded validation requires that components driving the core outcomes of a generative social simulation be empirically anchored in real behavioural data and that simulated dynamics be interpreted relative to the scope of this grounding. In the present work, persuasion is the central mechanism through which stance change is generated, and CMV therefore serves as the primary external reference. While this choice strengthens interpretability and avoids purely surface-level plausibility checks, it also imposes structural constraints on how persuasion and belief change can be represented.

Belief updating is operationalised via a scalar stance variable and an assimilative update rule in which belief change is proportional to the difference between an agent’s current stance and the stance implied by an incoming message, following continuous opinion dynamics models (Section 3.5.1). This formulation assumes that exposure to another agent’s stance induces movement in that direction. In practice, however, individuals may encounter opposing arguments that are perceived as weak or normatively objectionable, resulting in no change or even further movement away from the expressed position, which is not captured by the present update rule. More generally, reducing complex political opinions to a single scalar value abstracts away the fact that such opinions are formed by weighing multiple, partially conflicting considerations. The resulting belief trajectories, therefore, reflect movement along a fixed stance axis rather than changes in the underlying structure of beliefs.

These constraints are further shaped by the properties of the persuasion data derived from CMV. This data captures short-term, explicitly acknowledged belief change in a highly structured and community-specific discussion setting, where participants engage with the explicit goal of changing others’ views and where OPs signal openness to being persuaded through delta awards. Interaction is therefore predominantly conducted in good-faith, argument-focused exchanges. The resulting persuasion signal is tuned to discrete, explicitly marked persuasion events rather than to gradual, affective, or cumulative belief shifts. Long-term belief persistence, decay, and post-discussion consolidation are not represented. Applying this signal in a Twitter-like simulation with short, compressed messages and different discourse conventions moves the model outside the distribution on which it was trained, and the resulting belief dynamics may differ from those observed in less structured or more adversarial social media environments.

Reliance on CMV-shaped persuasion signals conditions the observed effectiveness of narrative styles. Scientific-authority bots align closely with discourse patterns rewarded in CMV, whereas emotionally framed narratives are structurally disadvantaged. Efficiency differences should therefore be interpreted as conditional on CMV-style deliberative discourse and may not transfer unchanged to social media environments with different interaction norms and discourse conventions.

Taken together, these limitations reflect an unavoidable trade-off imposed by the scarcity of high-quality persuasion-labelled data. Anchoring belief updates in CMV-derived signals prioritises empirical grounding and interpretability, but necessarily constrains the forms of persuasion and belief change that the

simulation can meaningfully represent.

Network and interaction abstractions A further set of limitations arises from modelling abstractions introduced to keep the simulation tractable and interpretable. Any formal model necessarily abstracts from aspects of reality to isolate specific mechanisms of interest, and generative social simulations are no exception.

The follower network is initialised in a simplified manner that abstracts away the processes by which social ties are formed (Section 5.5.1). Initial connections are not conditioned on agents' stances or personas, and may therefore link agents with highly dissimilar positions. The network is further assumed to exist at simulation start. While bot agents are deliberately placed in peripheral positions rather than occupying central roles, the simulation does not model how accounts gain followers, increase visibility, or embed within communities over time. Prior work conceptualises social bot operations as a multi-stage workflow that includes deployment, social network expansion, and only subsequently the release of influence (L. Xu 2023; W. Wang et al. 2020). The present simulation focuses on this final stage and treats network position as largely exogenous. As a result, it abstracts away network growth and infiltration dynamics that are widely regarded as central to real-world influence operations, and instead isolates persuasive effects conditional on assumed access and exposure.

Relatedly, the simulated population is not embedded in geographically or socially localised communities (Section 5.3.1). Personas are sampled from ANES data covering the entire US population, but interaction is not constrained by region or locality. Agents therefore interact at a national scale rather than within the more clustered and locally bounded discourse spaces that typically characterise real social media discussions (Takhteyev et al. 2012; Pfetsch et al. 2021).

The design of the simulation environment further limits the structure of interaction. While agents can reply to posts, the environment does not support deep, branching sub-threads with sustained multi-party exchanges. Discussion, therefore, remains relatively shallow, and argumentative trajectories that unfold over long reply chains or nested debates are not represented.

The feed-ranking algorithm is deliberately constrained to limit repeated exposure to the same content (Section 5.5.2). Timelines prioritise previously unseen messages to keep belief updates attributable to individual exposures and to prevent uncontrolled accumulation of multiple stance updates for the same message. This design choice suppresses virality and visibility-driven engagement dynamics that play a central role on real social media platforms, limiting interaction to largely first-order effects without cumulative reinforcement from repeated exposure.

Taken together, these limitations bound the scope of the simulation to persuasion dynamics under controlled exposure within a connected but deliberately simplified social structure. A central constraint underlying many design choices is the limited availability of externally grounded persuasion data. CMV currently represents the most viable large-scale source of explicit, behaviourally anchored belief-change signals, and much of the modelling effort necessarily revolves around adapting the simulation to the structure and assumptions of this data.

Implementation sensitivity and stochasticity Beyond this data dependency, the present work also reflects a more general limitation of agent-based modelling. Complex simulations are inherently sensitive to implementation details that cannot be exhaustively specified or standardised, a challenge that is amplified in generative social simulations by stochastic LLM behaviour. As a result, it is difficult to distinguish dynamics that are structurally robust from those that depend on particular design decisions. The contribution of this work should therefore be understood as a carefully bounded and empirically anchored exploration of belief change dynamics under explicitly stated assumptions, rather than a definitive account of real-world persuasion processes.

7.4 Ethical Considerations

This work addresses ethically sensitive questions concerning belief formation, persuasion, and the use of automated agents in social media environments. While the simulation is designed as an analytical research tool, its focus on influence dynamics and social bots raises ethical considerations related to data use, dual use, and responsible interpretation, consistent with the ethical and societal risks outlined in Table 3 as [C7].

Use of existing data and in silico simulation The study draws exclusively on previously published datasets derived from CMV, including the Winning Arguments Corpus and Webis-CMV-20. These datasets were collected as part of prior research, and their use here assumes that the original studies adhered to appropriate ethical standards with respect to consent, anonymisation, and data handling. No new data from human participants are collected, and no real users are exposed to experimental interventions.⁹ All agents in the simulation are synthetic, and belief dynamics are studied entirely in silico. This avoids the ethical concerns that would arise from modelling, measuring, or attempting to influence the beliefs of real individuals in experimental settings.

Ethical sandboxing A key motivation for using generative social simulation in this work is the possibility of ethical sandboxing [A5]. The simulation enables controlled experimentation with persuasion dynamics, exposure mechanisms, and narrative framing without exposing human participants to potentially harmful or deceptive interventions. Sensitive processes that would be ethically impermissible to study directly are instead examined in a fully artificial environment.

Dual-use risks Despite its analytical intent, the simulation exhibits clear dual-use characteristics, a concern highlighted in prior work on social simulation and persuasion modelling [C7]. Models that quantify persuasive impact or compare narrative effectiveness could, in principle, inform the design of more effective social bots or influence strategies. In the present study, this risk is limited by scope and design. The simulation does not interact with real platforms, does not involve real users, and abstracts away critical stages of influence operations such as social bot deployment and network infiltration. The analysis is confined to belief change conditional on exposure, rather than to operational optimisation of influence in real-world settings.

⁹ This design choice contrasts with a recent controversial study in which LLM-generated content was deployed directly in the r/ChangeMyView community to measure persuasive effects in situ and without informed consent, prompting widespread ethical criticism (see O’Grady 2025).

Responsible interpretation Ethical risks also arise at the level of representation and interpretation [C7]. Although personas are grounded in empirical survey data, simulated populations remain abstractions that may simplify social identities or underrepresent minority perspectives. In addition, the fluent and coherent behaviour of language-model-based agents can encourage overinterpretation of simulated outcomes. To address these risks, the research questions and analyses in this thesis are explicitly formulated at the population level. The simulation is used to study aggregate belief dynamics and comparative patterns across conditions, not to draw conclusions about individual agents or specific personas, and results are interpreted accordingly within the bounds set by the modelling assumptions and limitations discussed above.

7.5 Future Work

Several directions for future work emerge from the present study, ranging from incremental extensions of the existing design to more substantive methodological changes and reflective redesign choices.

Scaling and scope extensions A first class of extensions concerns scaling and broadening the current experimental setup along existing dimensions. Future simulations could explore a wider range of both narrative styles, topics, population sizes, and simulation lengths, as well as alternative initial stance distributions, such as bimodal or skewed configurations. Extending the topic space beyond misinformation-related or highly polarised domains would help assess whether the observed belief dynamics generalise to less adversarial issue contexts. Similarly, evaluating different LLMs or larger model variants could shed light on how model capacity, alignment, or latent political bias affect agent behaviour and persuasion outcomes. These extensions would primarily strengthen the robustness and external scope of the current findings without altering the core simulation logic.

Multidimensional belief representations A second set of directions involves extending the modelling capacity beyond simple scaling. One promising avenue is the representation of beliefs as vectors spanning multiple issues rather than as a single scalar stance. While the present simulation adopts a unidimensional stance representation, Section 3.5.1 already discusses established extensions of continuous opinion dynamics models to multidimensional opinion spaces. Integrating such representations into a generative agent framework would allow belief change to reflect trade-offs across multiple considerations rather than movement along a single axis, bringing the simulation closer to how complex political attitudes are commonly conceptualised.

Robust prompt design Another extension concerns agent prompting and decision structure. Several observed shortcut behaviours could be mitigated through alternative prompt designs, such as a two-stage action-decision process in which agents first select an abstract action type and only subsequently determine a concrete target or payload. This would reduce coupling between visible timeline content and action choice and mirrors the sequential stance initialisation procedure used during environment setup (Section 5.5.1). However, because action decisions occur at every timestep for every agent, such a design would substantially increase inference cost. Exploring this trade-off between finer behavioural control and computational feasibility, therefore, remains an important direction for future work.

Comparing LLMs Future work could also compare different LLMs within the same simulation framework. Such comparisons would not primarily aim at scaling effects, but at examining how differences in training data and instruction tuning shape how political topics are discussed, argued, and framed, and how these differences translate into persuasion dynamics.

Multimodal LLM-based interaction Advances in multimodal language models open the possibility of moving beyond purely textual abstractions of social media interaction. Rather than modelling a Twitter-like platform solely through text prompts, future simulations could incorporate simplified visual or interface-level representations of timelines, enabling agents to perceive and act within a more realistic approximation of platform affordances while still maintaining controlled abstraction.

Adaptive bot strategies Future research could also consider more expressive bot policies. Instead of fixed narrative styles, bots could dynamically mix or adapt narratives based on observed user characteristics, interaction history, or inferred susceptibility, potentially drawing on richer profile or network metadata. Extending this further, multiple bot factions with competing objectives or coordinated group behaviour could be introduced to study strategic interaction, counter-persuasion, and coalition dynamics.

Exposure and recommendation mechanisms A third category of future work concerns design choices that, informed by the insights gained in this work, could be approached differently in a follow-up implementation, particularly with respect to recommendation and exposure mechanisms. In the present simulation, bot accounts can enter human agents' feeds by replying to them, and such replies are surfaced irrespective of follower relations. This mechanism plays a central role in enabling belief influence, but it also weakens exposure control, as follow relations no longer serve as a reliable boundary for limiting which accounts an agent is exposed to.

In real-world settings, repeated unsolicited messages from unknown accounts are often ignored, muted, or blocked. By contrast, the current design provides only unfollow actions, which are insufficient to consistently suppress exposure to bot content. Future designs could therefore introduce stronger exposure constraints, such as explicit blocking actions or simple heuristics that prevent recommendations from accounts an agent has actively unfollowed. Alternatively, recommendation mechanisms could be simplified or further restricted to make exposure pathways more transparent.

Alternative persuasion data sources At the data level, future work could explore additional sources of externally grounded persuasion signals beyond CMV, which currently appears to be the most suitable large-scale resource for this purpose. Even within CMV, alternative supervised learning formulations beyond the RankFormer architecture could be investigated to derive belief-shift signals in different ways.

7.6 Conclusion

This thesis examined whether and how LLM-based generative agents can be used to simulate belief change in social media environments under controlled and empirically grounded conditions. Beyond the methodological debate around validity in generative social simulation, the work is motivated by the role of social

bots and narrative persuasion in information disorder and computational propaganda. The central objective was not to maximise behavioural realism or predictive accuracy, but to assess whether belief dynamics induced by social bots can be studied in a principled and methodologically defensible way. To this end, the work combined a Twitter-like generative social simulation with a persuasion signal externally grounded in human-verified belief change data from CMV and used this framework to analyse how different bot narrative styles shape population-level belief dynamics.

The overarching research question asked how narrative styles employed by LLM-powered bots affect belief change and influence efficiency when belief updates are governed by a persuasion judge validated on human data. This question was decomposed into three sub-questions addressing the empirical realism of the persuasion signal, the resulting belief dynamics, and the comparative efficiency of narrative styles.

With respect to RQ1.1, the results show that the persuasion judge captures a meaningful, though imperfect, signal of persuasive success in CMV discussions. Quantitative evaluation demonstrates performance well above chance across multiple ranking-based metrics, while qualitative error analysis reveals systematic and interpretable failure modes that align with known properties of online persuasion. Calibration further links model outputs to empirically observed belief change, ensuring that persuasion scores correspond to plausible magnitudes of stance updating. Taken together, these findings indicate that the persuasion judge achieves empirical realism at a level suitable for driving population-level belief updates, while inheriting structural biases from its training data. The signal reflects explicit, acknowledged persuasion events rather than gradual or latent belief change, a property that directly shapes the dynamics observed in the simulation.

Regarding RQ1.2, the simulation consistently exhibits rapid early belief change followed by asymptotic convergence toward a narrow stance range across narrative styles and topics. Narrative styles primarily affect the speed and extent of early belief shifts, but do not alter the qualitative regime of the dynamics. These patterns closely mirror assimilative continuous opinion dynamics models, in which interaction between connected agents leads to consensus when the network remains connected. Introducing fixed-stance bot agents shifts the convergence point but does not prevent consensus formation. Attempts to induce similarity-biased dynamics through unfollowing and network adaptation do not succeed under the present design, as exposure remains dominated by reply-based interactions within a persistently connected network. As a result, belief change is structurally driven by early exposure and diminishing opinion distance rather than by sustained fragmentation or clustering.

RQ1.3 examined how efficiently different narrative styles induce belief change over time relative to a fixed-stance control. Clear differences emerge across all four bot styles. Scientific-authority bots consistently outperform the control and achieve the highest efficiency, followed by post-truth populist and echo-claim bots, which accumulate positive influence over time even when outcomes are close to or weaker than the control. Emotional bots perform worst and consistently underperform the control across topics and efficiency measures. Endpoint efficiency compresses these differences due to early convergence, whereas cumulative AUC-based measures reveal substantial variation in early influence dynamics. The absence of statistically significant pairwise differences reflects limited statistical power rather than a lack of structure, given the small number of simulation seeds and the computational constraints inherent to generative social simulation. Importantly, the relative ordering of narrative styles is stable across topics and measures,

indicating that these efficiency patterns are robust within the bounds of the experimental design.

Across all three research questions, the findings support a qualified answer to the overarching research question. Generative agents can be used to simulate belief change under different narrative styles in a way that is interpretable and empirically grounded, provided that the simulation is tightly anchored in validated persuasion signals and its results are interpreted within clearly defined methodological bounds. At the same time, the results show that observed belief dynamics are strongly shaped by the specific persuasion signal, stance update rule, and exposure mechanisms employed. Narrative effectiveness, therefore, emerges as a conditional property of the modelling configuration rather than as an intrinsic characteristic of the narratives themselves.

Beyond the specific empirical results, the thesis makes three broader contributions. Methodologically, it demonstrates how purpose-aligned and externally grounded validation can be operationalised in a generative social simulation by embedding empirical grounding directly into the simulation mechanism itself, rather than treating validation as a purely post hoc evaluation step. Substantively, it provides evidence that narrative styles influence the timing and efficiency of belief change even when overall dynamics are dominated by assimilative convergence. Conceptually, it clarifies how limitations related to persuasion data, opinion representation, network structure, and prompt sensitivity are not incidental shortcomings but structural conditions that delimit what generative social simulations can and cannot claim.

In this sense, the contribution of this work is not a definitive model of real-world influence operations, but a deliberately bounded investigation of belief change under explicitly stated assumptions. As Box (1976) famously noted, "*All models are wrong, but some are useful.*" The simulation provides a controlled setting for examining persuasion processes that would be difficult or ethically problematic to study with human subjects. Its contribution lies in making the dependencies between persuasion signals, exposure mechanisms, and belief dynamics explicit and traceable, while keeping their empirical grounding transparent.

As generative agents become increasingly deployed in generative social simulation, the central challenge is no longer what can be generated, but what can be justifiably interpreted. This thesis shows that meaningful insights into belief change can be obtained when generative capacity is coupled with empirical grounding and clearly stated assumptions.

At the same time, it shows that validation is not a peripheral concern but the decisive bottleneck for generative social simulation. Credibility does not follow from scale, apparent realism, or sophistication alone, but from transparency about what is being modelled, how behavioural signals are anchored in empirical data, and which conclusions the simulation can and cannot support.

References

- Adair, John G. (1984). “The Hawthorne Effect: A Reconsideration of the Methodological Artifact”. In: *Journal of Applied Psychology* 69.2, pp. 334–345. DOI: 10.1037/0021-9010.69.2.334.
- Adams, Zoë, Magda Osman, Christos Bechlivanidis, and Björn Meder (2023). “(Why) Is Misinformation a Problem?” In: *Perspectives on Psychological Science* 18.6, pp. 1436–1463. DOI: 10.1177/17456916221141344.
- Aiello, Luca Maria, Martina Deplano, Rossano Schifanella, and Giancarlo Ruffo (2021). “People Are Strange When You’re a Stranger: Impact and Influence of Bots on Social Networks”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 6.1, pp. 10–17. DOI: 10.1609/icwsm.v6i1.14236.
- Aimeur, Esmâ, Sabrine Amri, and Gilles Brassard (2023). “Fake News, Disinformation and Misinformation in Social Media: A Review”. In: *Social Network Analysis and Mining* 13.1, p. 30. DOI: 10.1007/s13278-023-01028-5.
- AlAfnan, Mohammad Awad (2025). “The Role of Memes in Shaping Political Discourse on Social Media”. In: *Studies in Media and Communication* 13.2, p. 1. DOI: 10.11114/smc.v13i2.7482.
- Ali, Ameen, Lior Wolf, and Ivan Titov (2024). *Mitigating Copy Bias in In-Context Learning through Neuron Pruning*.
- American National Election Studies (2025). *ANES 2024 Time Series Study Full Release*. Version August 8, 2025. American National Election Studies.
- Anthis, Jacy Reese, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein (2025). *LLM Social Simulations Are a Promising Research Method*.
- Aparicio, Sofía, Javier Villazón-Terrazas, and Gonzalo Álvarez (2015). “A Model for Scale-Free Networks: Application to Twitter”. In: *Entropy* 17.8, pp. 5848–5867. DOI: 10.3390/e17085848.
- Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate (2023). “Out of One, Many: Using Language Models to Simulate Human Samples”. In: *Political Analysis* 31.3, pp. 337–351. DOI: 10.1017/pan.2023.2.
- Averza, Aldo, Khaled Slhoub, and Siddhartha Bhattacharyya (2022). “Evaluating the Influence of Twitter Bots via Agent-Based Social Simulation”. In: *IEEE Access* 10, pp. 129394–129407. DOI: 10.1109/ACCESS.2022.3228258.
- Bail, Christopher A. (2024). “Can Generative AI improve social science?” In: *Proceedings of the National Academy of Sciences* 121.21, e2314021121. DOI: 10.1073/pnas.2314021121.
- Bang, Yejin, DeLong Chen, Nayeon Lee, and Pascale Fung (2024). “Measuring Political Bias in Large Language Models: What Is Said and How It Is Said”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, An-

- dre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 11142–11159. DOI: 10.18653/v1/2024.acl-long.600.
- Barabási, Albert-László, Réka Albert, and Hawoong Jeong (1999). “Mean-field theory for scale-free random networks”. In: *Physica A: Statistical Mechanics and its Applications* 272.1, pp. 173–187. DOI: [https://doi.org/10.1016/S0378-4371\(99\)00291-5](https://doi.org/10.1016/S0378-4371(99)00291-5).
- Barrie, Christopher and Petter Törnberg (2025). *Emergent LLM behaviors are observationally equivalent to data leakage*.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, pp. 610–623. DOI: 10.1145/3442188.3445922.
- Beskow, David M. and Kathleen M. Carley (2020). “Agent based simulation of bot disinformation maneuvers in Twitter”. In: *Proceedings of the Winter Simulation Conference*. WSC ’19. National Harbor, Maryland: IEEE Press, pp. 750–761.
- Bessi, Alessandro and Emilio Ferrara (2016). “Social Bots Distort the 2016 U.S. Presidential Election Online Discussion”. In: *First Monday* 21.11. DOI: 10.5210/fm.v21i11.7090.
- Bil-Jaruzelska, Anna and Cristina Monzer (2022). “All About Feelings? Emotional Appeals as Drivers of User Engagement With Facebook Posts”. In: *Politics and Governance* 10.1, pp. 172–184. DOI: 10.17645/pag.v10i1.4758.
- Binz, Marcel and Eric Schulz (2023). “Using cognitive psychology to understand GPT-3”. In: *Proceedings of the National Academy of Sciences* 120.6, e2218523120. DOI: 10.1073/pnas.2218523120.
- Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson (2024). “Synthetic Replacements for Human Survey Data? The Perils of Large Language Models”. In: *Political Analysis* 32.4, pp. 401–416. DOI: 10.1017/pan.2024.5.
- Boelaert, Julien, Samuel Coavoux, Étienne Ollion, Ivaylo Petev, and Patrick Präg (2025). “Machine Bias. How Do Generative Language Models Answer Opinion Polls?1”. In: *Sociological Methods & Research* 54.3, pp. 1156–1196. DOI: 10.1177/00491241251330582.
- Bonabeau, Eric (2002). “Agent-based modeling: Methods and techniques for simulating human systems”. In: *Proceedings of the National Academy of Sciences* 99.suppl_3, pp. 7280–7287. DOI: 10.1073/pnas.082080899.
- Boukes, Mark, Hajo G. Boomgaarden, Marjolein Moorman, and Claes H. de Vreese (2015). “At Odds: Laughing and Thinking? The Appreciation, Processing, and Persuasiveness of Political Satire: At Odds: Laughing and Thinking?” In: *Journal of Communication* 65.5, pp. 721–744. DOI: 10.1111/jcom.12173.
- Box, George E. P. (1976). “Science and Statistics”. In: *Journal of the American Statistical Association* 71.356, pp. 791–799. DOI: 10.1080/01621459.1976.10480949.

- Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel (2017). “Emotion Shapes the Diffusion of Moralized Content in Social Networks”. In: *Proceedings of the National Academy of Sciences* 114.28, pp. 7313–7318. DOI: 10.1073/pnas.1618923114.
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). “Language models are few-shot learners”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc.
- Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4*.
- Buyl, Maarten, Paul Missault, and Pierre-Antoine Sondag (2023). “RankFormer: Listwise Learning-to-Rank Using Listwise Labels”. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '23. Long Beach, CA, USA: Association for Computing Machinery, pp. 3762–3773. DOI: 10.1145/3580305.3599892.
- Byrne, David and Gillian Callaghan (2022). *Complexity Theory and the Social Sciences: The State of the Art*. Routledge.
- Capuano, Nicola, Marco Meyer, and Francesco David Nota (2024). “Analyzing the Impact of Conversation Structure on Predicting Persuasive Comments Online”. In: *Journal of Ambient Intelligence and Humanized Computing* 15.11, pp. 3719–3732. DOI: 10.1007/s12652-024-04841-8.
- Castellano, Claudio, Santo Fortunato, and Vittorio Loreto (2009). “Statistical physics of social dynamics”. In: *Rev. Mod. Phys.* 81 (2), pp. 591–646. DOI: 10.1103/RevModPhys.81.591.
- Cau, Erica, Valentina Pansanella, Dino Pedreschi, and Giulio Rossetti (2025). “Selective Agreement, Not Sycophancy: Investigating Opinion Dynamics in LLM Interactions”. In: *EPJ Data Science* 14.1, p. 59. DOI: 10.1140/epjds/s13688-025-00579-1.
- Chen, Chang-Feng, Wen Shi, Jing Yang, and Hao-Huan Fu (2021). “Social bots’ role in climate change discussion on Twitter: Measuring standpoints, topics, and interaction strategies”. In: *Advances in Climate Change Research* 12.6, pp. 913–923. DOI: <https://doi.org/10.1016/j.accre.2021.09.011>.
- Chen, Chen, Hao Yuan, and Mike Z. Yao (2021). “Identifiability as an “Antidote”: Exploring Emotional Contagion and the Role of Anonymity in Twitter Discussions on Misinformation”. In: *Social Computing and Social Media: Experience Design and Social Network Analysis*. Springer International Publishing, pp. 240–252. DOI: 10.1007/978-3-030-77626-8_16.

- Chen, Sijing, Lu Xiao, and Akit Kumar (2023). “Spread of misinformation on social media: What contributes to it and how to combat it”. In: *Computers in Human Behavior* 141, p. 107643. DOI: <https://doi.org/10.1016/j.chb.2022.107643>.
- Chen, Zhenpeng, Chong Wang, Weisong Sun, Guang Yang, Xuanzhe Liu, Jie M. Zhang, and Yang Liu (2025). *Promptware Engineering: Software Engineering for LLM Prompt Development*.
- Cheng, Chun, Yun Luo, and Changbin Yu (2020). “Dynamic mechanism of social bots interfering with public opinion in network”. In: *Physica A: Statistical Mechanics and its Applications* 551, p. 124163. DOI: <https://doi.org/10.1016/j.physa.2020.124163>.
- Cheng, Myra, Esin Durmus, and Dan Jurafsky (2023). “Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 1504–1532. DOI: 10.18653/v1/2023.acl-long.84.
- Choi, Jihyang, Sang Yup Lee, and Sung Wook Ji (2021). “Engagement in Emotional News on Social Media: Intensity and Type of Emotions”. In: *Journalism & Mass Communication Quarterly* 98.4, pp. 1017–1040. DOI: 10.1177/1077699020959718.
- Chuang, Yun-Shiuan, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers (2024). “Simulating Opinion Dynamics with Networks of LLM-based Agents”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 3326–3346. DOI: 10.18653/v1/2024.findings-naacl.211.
- Clark, Andy (2013). “Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science”. In: *Behavioral and Brain Sciences* 36.3, pp. 181–204. DOI: 10.1017/S0140525X12000477.
- Clark, Elizabeth, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith (2021). “All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 7282–7296. DOI: 10.18653/v1/2021.acl-long.565.
- Collins, Harry (2017). “Interactional Expertise and Embodiment”. In: *Skillful Performance: Enacting Capabilities, Knowledge, Competence, and Expertise in Organizations*. Oxford University Press. DOI: 10.1093/oso/9780198806639.003.0006.
- Conte, Rosaria and Mario Paolucci (2014). “On agent-based modeling and computational social science”. In: *Frontiers in Psychology* 5. DOI: 10.3389/fpsyg.2014.00668.
- Craswell, Nick, Onno Zoeter, Michael Taylor, and Bill Ramsey (2008). “An experimental comparison of click position-bias models”. In: *Proceedings of the 2008 International Conference on Web Search*

- and Data Mining*. WSDM '08. Palo Alto, California, USA: Association for Computing Machinery, pp. 87–94. DOI: 10.1145/1341531.1341545.
- Cresci, Stefano (2020). “A decade of social bot detection”. In: *Communications of the ACM* 63.10, pp. 72–83. DOI: 10.1145/3409116.
- Dagan, Gautier, Frank Keller, and Alex Lascarides (2023). *Dynamic Planning with a LLM*.
- Davies, Alex O., Roussel Nzoyem, Nirav Ajmeri, and Telmo M. Silva Filho (2025). *Language Models Do Not Embed Numbers Continuously*.
- Deb, Anamitra, Stacy Donohue, and Tom Glaisyer (2017). *Is Social Media a Threat to Democracy?* Technical Report. Omidyar Network.
- Deffuant, Guillaume, Frederic Amblard, and Gérard Weisbuch (2002). “How Can Extremism Prevail? a Study Based on the Relative Agreement Interaction Model”. In: *Journal of Artificial Societies and Social Simulation* 5.4, pp. 1–1. DOI: None.
- Deffuant, Guillaume, Sylvie Huet, and Frédéric Amblard (2005). “An Individual-Based Model of Innovation Diffusion Mixing Social Value and Individual Benefit”. In: *American Journal of Sociology* 110.4, pp. 1041–1069. DOI: 10.1086/430220.
- Deffuant, Guillaume, David Neau, Frédéric Amblard, and Gérard Weisbuch (2000). “Mixing Beliefs Among Interacting Agents”. In: *Advances in Complex Systems* 3, pp. 87–98. DOI: 10.1142/S0219525900000078.
- Del Vicario, Michela, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi (2017). “Modeling Confirmation Bias and Polarization”. In: *Scientific Reports* 7.1, p. 40391. DOI: 10.1038/srep40391.
- Dimitrov, Dimitar, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino (2021). “Detecting Propaganda Techniques in Memes”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 6603–6617. DOI: 10.18653/v1/2021.acl-long.516.
- Dong, Qingxiu, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui (2024). “A Survey on In-context Learning”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 1107–1128. DOI: 10.18653/v1/2024.emnlp-main.64.
- Drávucz, Renáta and Loren Kocollari (2023). “Populist Communication in a Time of Global Pandemic: Understanding ‘Post-truth Populism’ During COVID-19”. In: *Sociológia - Slovak Sociological Review* 55, pp. 244–265. DOI: 10.31577/sociologia.2023.55.2.9.

- Du, Mengnan, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu (2023). “Shortcut Learning of Large Language Models in Natural Language Understanding”. In: *Commun. ACM* 67.1, pp. 110–120. DOI: 10.1145/3596490.
- Du, Yufeng, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng (2025). “Context Length Alone Hurts LLM Performance Despite Perfect Retrieval”. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Ed. by Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng. Suzhou, China: Association for Computational Linguistics, pp. 23281–23298. DOI: 10.18653/v1/2025.findings-emnlp.1264.
- Dzhoha, Andrii, Alexey Kurennoy, Vladimir Vlasov, and Marjan Celikik (2024). “Reducing Popularity Influence by Addressing Position Bias”. In: *Proceedings of the RobustRecSys Workshop at the 18th ACM Conference on Recommender Systems (RecSys 2024)*. Bari, Italy: CEUR-WS.
- Ecker, Ullrich K. H., Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen (2022). “The Psychological Drivers of Misinformation Belief and Its Resistance to Correction”. In: *Nature Reviews Psychology* 1.1, pp. 13–29. DOI: 10.1038/s44159-021-00006-y.
- Edwards, Allen L. (1957). *The Social Desirability Variable in Personality Assessment and Research*. Fort Worth, TX, USA: Dryden Press, pp. viii + 108.
- Epstein, Joshua M. (1999). “Agent-based computational models and generative social science”. In: *Complexity* 4.5, pp. 41–60. DOI: 10.1002/(sici)1099-0526(199905/06)4:5<41::aid-cplx9>3.0.co;2-f.
- Epstein, Joshua M. (2007). *Generative Social Science: Studies in Agent-Based Computational Modeling (Princeton Studies in Complexity)*. USA: Princeton University Press.
- Epstein, Joshua M. and Robert L. Axtell (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. The MIT Press.
- Fang, Zhichao, Rodrigo Costas, and Paul Wouters (2022). “User Engagement with Scholarly Tweets of Scientific Papers: A Large-Scale and Cross-Disciplinary Analysis”. In: *Scientometrics* 127.8, pp. 4523–4546. DOI: 10.1007/s11192-022-04468-6.
- Fazio, Lisa K., Nadia M. Brashier, B. Keith Payne, and Elizabeth J. Marsh (2015). “Knowledge Does Not Protect Against Illusory Truth”. In: *Journal of Experimental Psychology: General* 144.5, pp. 993–1002. DOI: 10.1037/xge0000098.
- Feng, Shangbin, Herun Wan, Ningnan Wang, Zhaoxuan Tan, Minnan Luo, and Yulia Tsvetkov (2024). “What Does the Bot Say? Opportunities and Risks of Large Language Models in Social Media Bot Detection”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 3580–3601. DOI: 10.18653/v1/2024.acl-long.196.

- Ferrara, Emilio (2020). “What types of COVID-19 conspiracies are populated by Twitter bots?” In: *First Monday*. DOI: 10.5210/fm.v25i6.10633.
- Ferrara, Emilio, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel (2020). “Characterizing Social Media Manipulation in the 2020 U.S. Presidential Election”. In: *First Monday* 25.11. DOI: 10.5210/fm.v25i11.11431.
- Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini (2016). “The rise of social bots”. In: *Commun. ACM* 59.7, pp. 96–104. DOI: 10.1145/2818717.
- Fiadotava, Anastasiya, Anastasiya Astapova, Rebecca Hendershott, Merryn McKinnon, and Anna-Sophie Jürgens (2023). “Injecting Fun? Humour, Conspiracy Theory and (Anti)Vaccination Discourse in Popular Media”. In: *Public Understanding of Science* 32.5, pp. 622–640. DOI: 10.1177/09636625221147019.
- Flache, Andreas, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz (2017). “Models of Social Influence: Towards the Next Frontiers”. In: *Journal of Artificial Societies and Social Simulation* 20.4, p. 2. DOI: 10.18564/jasss.3521.
- Flores-Saviaga, Claudia, Shangbin Feng, and Saiph Savage (2022). “Datavoidant: An AI System for Addressing Political Data Voids on Social Media”. In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW2. DOI: 10.1145/3555616.
- Flynn, D.J., Brendan Nyhan, and Jason Reifler (2017). “The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics”. In: *Political Psychology* 38.S1, pp. 127–150. DOI: 10.1111/pops.12394.
- Fontana, Nicolò, Francesco Pierri, and Luca Maria Aiello (2025). “Nicer Than Humans: How Do Large Language Models Behave in the Prisoner’s Dilemma?” In: *Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)*. Association for the Advancement of Artificial Intelligence, pp. 522–535.
- Fuchs, Christian (2018). “Propaganda 2.0: Herman and Chomsky’s Propaganda Model in the Age of the Internet, Big Data and Social Media”. In: *THE PROPAGANDA MODEL TODAY: Filtering Perception and Awareness*. Vol. 8. University of Westminster Press, pp. 71–92.
- Gallegos, Isabel O., Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed (2024). “Bias and Fairness in Large Language Models: A Survey”. In: *Computational Linguistics* 50.3, pp. 1097–1179. DOI: 10.1162/coli_a_00524.
- Gao, Chen, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li (2024). “Large Language Models Empowered Agent-Based Modeling and Simulation: A Survey and Perspectives”. In: *Humanities and Social Sciences Communications* 11.1, p. 1259. DOI: 10.1057/s41599-024-03611-3.
- Gao, Chen, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li (2025). *S³: Social-network Simulation System with Large Language Model-Empowered Agents*.

- Gao, Leo, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy (2020). “The Pile: An 800GB Dataset of Diverse Text for Language Modeling”. In: *ArXiv abs/2101.00027*.
- Gao, Yuan, Dokyun Lee, Gordon Burtch, and Sina Fazelpour (2025). “Take Caution in Using LLMs as Human Surrogates”. In: *Proceedings of the National Academy of Sciences of the United States of America* 122.24, e2501660122. DOI: 10.1073/pnas.2501660122.
- Gess, Nicola (2021). “Half-Truths: On an Instrument of Post-Truth Politics (and Conspiracy Narratives)”. In: *Plots: Literary Form and Cultures of Conspiracy*. Ed. by Ben Carver, Dana Crăciun, and Todor Hristov. 1st ed. London: Routledge, pp. 164–179.
- Ghaffarzadegan, Navid, Aritra Majumdar, Ross Williams, and Niyousha Hosseinichimeh (2024). “Generative agent-based modeling: an introduction and tutorial”. In: *System Dynamics Review* 40.1. DOI: 10.1002/sdr.1761.
- Gilder, Thandiwe S. E. and Erin A. Heerey (2018). “The Role of Experimenter Belief in Social Priming”. In: *Psychological Science* 29.3, pp. 403–417. DOI: 10.1177/0956797617737128.
- Goddard, Kate, Abdul V. Roudsari, and Jeremy C. Wyatt (2012). “Automation bias: a systematic review of frequency, effect mediators, and mitigators”. In: *Journal of the American Medical Informatics Association : JAMIA* 19 1, pp. 121–7.
- Goethals, Sofie, Johannes Luther, and Sandra Matz (2025). “Words reveal wants: How well can simple LLM-based AI agents replicate people’s choices based on their social media posts”. In: *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. UMAP Adjunct ’25. Association for Computing Machinery, pp. 126–131. DOI: 10.1145/3708319.3733689.
- Gorwa, Robert and Douglas Guilbeault (2018). “Unpacking the Social Media Bot: A Typology to Guide Research and Policy”. In: *Policy & Internet* 12.2, pp. 225–248. DOI: 10.1002/poi3.184.
- Gozzi, Manuel and Federico Di Maio (2024). “Comparative Analysis of Prompt Strategies for Large Language Models: Single-Task vs. Multitask Prompts”. In: *Electronics* 13.23, p. 4712. DOI: 10.3390/electronics13234712.
- Graham, Timothy (2024). “Exploring a Post-Truth Referendum: Australia’s Voice to Parliament and the Management of Attention on Social Media”. In: *Media International Australia*. DOI: 10.1177/1329878X241267756.
- Guan, Haoxiang, Jiyan He, Liyang Fan, Zhenzhen Ren, Shaobin He, Xin Yu, Yuan Chen, Shuxin Zheng, Tie-Yan Liu, and Zhen Liu (2025). *Modeling Earth-Scale Human-Like Societies with One Billion Agents*.
- Guo, Taicheng, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang (2024). *Large Language Model based Multi-Agents: A Survey of Progress and Challenges*.

- Gurcan, Onder (2024). *LLM-Augmented Agent-Based Modelling for Social Simulations: Challenges and Opportunities*.
- Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith (2020). “Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 8342–8360. DOI: 10.18653/v1/2020.acl-main.740.
- Halil, Umut, Jin Huang, Damien Graux, and Jeff Z. Pan (2025). “LLM Shots: Best Fired at System or User Prompts?” In: *Companion Proceedings of the ACM on Web Conference 2025*. WWW ’25. Sydney NSW, Australia: Association for Computing Machinery, pp. 1605–1613. DOI: 10.1145/3701716.3717814.
- Hall, Stuart (1973). *Encoding and Decoding in the Television Discourse*. Tech. rep. Birmingham, UK: Centre for Contemporary Cultural Studies, University of Birmingham.
- Hameleers, Michael and Emma van der Goot (2024). “Look at what the real facts and experts say! The use of expert references and objectivity claims in disinformation: A qualitative exploration and typology”. In: *Journalism* 26, pp. 1485–1503.
- Hassan, Aumyo and Sarah J. Barber (2021). “The Effects of Repetition Frequency on the Illusory Truth Effect”. In: *Cognitive Research: Principles and Implications* 6.1, p. 38. DOI: 10.1186/s41235-021-00301-5.
- He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen (2021). “DeBERTa: decoding-Enhanced BERT with Disentangled Attention.” In: *ICLR*. OpenReview.net.
- He, Zhiyuan, Huiqiang Jiang, Zilong Wang, Yuqing Yang, Luna K. Qiu, and Lili Qiu (2024). “Position Engineering: Boosting Large Language Models through Positional Information Manipulation”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 7333–7345. DOI: 10.18653/v1/2024.emnlp-main.417.
- Hegselmann, Rainer and Ulrich Krause (2002). “Opinion Dynamics and Bounded Confidence Models, Analysis and Simulation”. In: *Journal of Artificial Societies and Social Simulation* 5.
- Hidey, Christopher and Kathleen McKeown (2018). “Persuasive Influence Detection: The Role of Argument Sequencing”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1. DOI: 10.1609/aaai.v32i1.12003.
- Holland, John H (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.
- Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi (2020). “The Curious Case of Neural Text Degeneration”. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

- Horne, Benjamin D., Sibel Adali, and Sujoy Sikdar (2017). “Identifying the Social Signals That Drive Online Discussions: A Case Study of Reddit Communities”. In: *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–9.
- Hosseini, Peyman, Ignacio Castro, Iacopo Ghinassi, and Matthew Purver (2025). “Efficient Solutions For An Intriguing Failure of LLMs: Long Context Window Does Not Mean LLMs Can Analyze Long Sequences Flawlessly”. In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert. Abu Dhabi, UAE: Association for Computational Linguistics, pp. 1880–1891.
- Hu, Chenxu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao (2023). *ChatDB: Augmenting LLMs with Databases as Their Symbolic Memory*.
- Hu, Tiancheng and Nigel Collier (2024). “Quantifying the Persona Effect in LLM Simulations”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 10289–10307. DOI: 10.18653/v1/2024.acl-long.554.
- Hu, Zhengyu, Jianxun Lian, Zheyuan Xiao, Max Xiong, Yuxuan Lei, Tianfu Wang, Kaize Ding, Ziang Xiao, Nicholas Jing Yuan, and Xing Xie (2025). *Population-Aligned Persona Generation for LLM-based Social Simulation*.
- Huang, Wenlong, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and brian ichter brian (2023). “Inner Monologue: Embodied Reasoning through Planning with Language Models”. In: *Proceedings of The 6th Conference on Robot Learning*. Ed. by Karen Liu, Dana Kulic, and Jeff Ichnowski. Vol. 205. Proceedings of Machine Learning Research. PMLR, pp. 1769–1782.
- Huang, Xu, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen (2024). *Understanding the planning of LLM agents: A survey*.
- Hussain, Zak, Rui Mata, and Dirk U. Wulff (2025). “A rebuttal of two common deflationary stances against LLM cognition”. In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, pp. 24208–24213. DOI: 10.18653/v1/2025.findings-acl.1242.
- Hwang, Tim, Ian Pearce, and Max Nanis (2012). “Socialbots: voices from the fronts”. In: *Interactions* 19.2, pp. 38–45. DOI: 10.1145/2090150.2090161.
- Iacobucci, Serena and Roberta De Cicco (2022). “A literature review of bullshit receptivity: Perspectives for an informed policy making against misinformation”. In: *Journal of Behavioral Economics for Policy* 6.S1, pp. 23–40. DOI: None.

- Imperva (2025). *2025 Bad Bot Report: The Rapid Rise of Bots and the Unseen Risk for Business*. Tech. rep. Imperva, a Thales Company.
- Jakesch, Maurice, Jeffrey T. Hancock, and Mor Naaman (2023). “Human heuristics for AI-generated language are flawed”. In: *Proceedings of the National Academy of Sciences* 120.11, e2208839120. DOI: 10.1073/pnas.2208839120.
- Jiang, Hang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara (2024). “PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 3605–3627. DOI: 10.18653/v1/2024.findings-naacl.229.
- Jin, Bailu and Weisi Guo (2025). “Synthetic Social Media Influence Experimentation Via an Agentic Reinforcement Learning Large Language Model Bot”. In: *Journal of Artificial Societies and Social Simulation* 28.3, p. 6. DOI: 10.18564/jasss.5726.
- Jin, Haolin, Linghan Huang, Haipeng Cai, Jun Yan, Bo Li, and Huaming Chen (2025). *From LLMs to LLM-based Agents for Software Engineering: A Survey of Current, Challenges and Future*.
- Jin, Zhijing, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez Aduato, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf (2024). “Multilingual Trolley Problems for Language Models”. In: *Pluralistic Alignment Workshop at NeurIPS 2024*.
- Jo, Yohan, Shivani Poddar, Byungsoo Jeon, Qinlan Shen, Carolyn Rosé, and Graham Neubig (2018). “Attentive Interaction Model: Modeling Changes in View in Argumentation”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 103–116. DOI: 10.18653/v1/N18-1010.
- Jones, Cameron R. and Benjamin K. Bergen (2025). *Large Language Models Pass the Turing Test*.
- Jowett, Garth S. and Victoria O’Donnell (2015). *Propaganda and Persuasion*. 6th ed. Thousand Oaks, CA: SAGE Publications.
- Jurafsky, Daniel and James H. Martin (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd.
- Kaiser, Carolin, Jakob Kaiser, Vladimir Manewitsch, Lea Rau, and Rene Schallner (2025). “Simulating Human Opinions with Large Language Models: Opportunities and Challenges for Personalized Survey Data Modeling”. In: *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*. UMAP Adjunct ’25. Association for Computing Machinery, pp. 82–86. DOI: 10.1145/3708319.3733685.

- Kan, Unchitta, Michelle Feng, and Mason A Porter (2023). “An adaptive bounded-confidence model of opinion dynamics on networks”. In: *Journal of Complex Networks* 11.1, cnac055. DOI: 10.1093/comnet/cnac055.
- Kaplan, Jared, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei (2020). “Scaling Laws for Neural Language Models”. In: *ArXiv* abs/2001.08361.
- Kera, Denisa Reshef and Lorenzo Cavallon (2025). “Blue and Red Teaming with AI Agents in Information Operations: A Sandbox Approach for Exploring Human Control in Counter-Messaging, Prompt Injection and AI Literacy”. In: *STO-MP-HFM-377*. STO. DOI: 10.14339/MP-HFM-377-07-PDF.
- Al-Khatib, Khalid, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein (2020). “Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness”. In: *58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*. Association for Computational Linguistics, pp. 7067–7072.
- Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa (2022). “Large language models are zero-shot reasoners”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. New Orleans, LA, USA: Curran Associates Inc.
- Kozma, Balazs and Alain Barrat (2008). “Consensus formation on coevolving networks: groups’ formation and structure”. In: *Journal of Physics A: Mathematical and Theoretical* 41.22, p. 224020. DOI: 10.1088/1751-8113/41/22/224020.
- Larooij, Maik and Petter Törnberg (2025a). *Can We Fix Social Media? Testing Prosocial Interventions using Generative Social Simulation*.
- Larooij, Maik and Petter Törnberg (2025b). “Validation is the Central Challenge for Generative Social Simulation: A Critical Review of LLMs in Agent-Based Modeling”. In: *Artificial Intelligence Review* 59.1, p. 15. DOI: 10.1007/s10462-025-11412-6.
- LeBaron, Blake (2006). “Chapter 24 Agent-based Computational Finance”. In: ed. by L. Tesfatsion and K.L. Judd. Vol. 2. *Handbook of Computational Economics*. Elsevier, pp. 1187–1233. DOI: [https://doi.org/10.1016/S1574-0021\(05\)02024-1](https://doi.org/10.1016/S1574-0021(05)02024-1).
- Lee, Kyumin, Brian Eoff, and James Caverlee (2021). “Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 5.1, pp. 185–192. DOI: 10.1609/icwsm.v5i1.14106.
- Levy, Mosh, Shauli Ravfogel, and Yoav Goldberg (2023). “Guiding LLM to Fool Itself: Automatically Manipulating Machine Reading Comprehension Shortcut Triggers”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 8495–8505. DOI: 10.18653/v1/2023.findings-emnlp.569.

- Lewandowsky, Stephan, Ullrich K. H. Ecker, and John Cook (2017). “Beyond Misinformation: Understanding and Coping with the “Post-Truth” Era”. In: *Journal of Applied Research in Memory and Cognition* 6.4, pp. 353–369. DOI: 10.1016/j.jarmac.2017.07.008.
- Lewandowsky, Stephan, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook (2012). “Misinformation and Its Correction: Continued Influence and Successful Debiasing”. In: *Psychological Science in the Public Interest* 13.3, pp. 106–131. DOI: 10.1177/1529100612451018.
- Li, Guohao, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem (2023). “CAMEL: communicative agents for ”mind” exploration of large language model society”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS ’23*. New Orleans, LA, USA: Curran Associates Inc.
- Li, Haoyang, Xuejia Chen, Zhanchao Xu, Darian Li, Nicole Hu, Fei Teng, Yiming Li, Luyu Qiu, Chen Jason Zhang, Li Qing, and Lei Chen (2025). “Exposing Numeracy Gaps: A Benchmark to Evaluate Fundamental Numerical Abilities in Large Language Models”. In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, pp. 20004–20026. DOI: 10.18653/v1/2025.findings-acl.1026.
- Li, Linda, Orsolya Vásárhelyi, and Balázs Vedres (2024). “Social Bots Spoil Activist Sentiment Without Eroding Engagement”. In: *Scientific Reports* 14.1, p. 27005. DOI: 10.1038/s41598-024-74032-0.
- Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang (2024). “Lost in the Middle: How Language Models Use Long Contexts”. In: *Transactions of the Association for Computational Linguistics* 12, pp. 157–173. DOI: 10.1162/tacl_a_00638.
- Liu, Yijun, Wu Liu, Xiaoyan Gu, Weiping Wang, Jiebo Luo, and Yongdong Zhang (2025). *RumorSphere: A Framework for Million-scale Agent-based Dynamic Simulation of Rumor Propagation*.
- Liu, Yuhan, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan (2024). “From skepticism to acceptance: simulating the attitude dynamics toward fake news”. In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. IJCAI ’24*. Jeju, Korea. DOI: 10.24963/ijcai.2024/873.
- Lopez-Joya, Salvador, Jose A. Diaz-Garcia, M. Dolores Ruiz, and Maria J. Martin-Bautista (2025). “Dissecting a social bot powered by generative AI: anatomy, new trends and challenges”. In: *Social Network Analysis and Mining* 15.1. DOI: 10.1007/s13278-025-01410-5.
- Lorenz, Jan (2007). “Continuous Opinion Dynamics under Bounded Confidence: A Survey”. In: *International Journal of Modern Physics C* 18.12, pp. 1819–1838. DOI: 10.1142/S0129183107011789.
- Low, Jwen Fai, Benjamin C. M. Fung, Farkhund Iqbal, and Claude Fachkha (2024). *Carthago Delenda Est: Co-opetitive Indirect Information Diffusion Model for Influence Operations on Online Social Media*.
- Lu, Yao, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp (2022). “Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity”. In:

- Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 8086–8098. DOI: 10.18653/v1/2022.acl-long.556.
- Luo, Linbo, Suiping Zhou, Wentong Cai, Malcolm Yoke Hean Low, Feng Tian, Yongwei Wang, Xian Xiao, and Dan Chen (2008). “Agent-based human behavior modeling for crowd simulation”. In: *Computer Animation and Virtual Worlds* 19.3–4, pp. 271–281. DOI: 10.1002/cav.238.
- Macal, C.M. and M.J. North (2005). “Tutorial on agent-based modeling and simulation”. In: *Proceedings of the Winter Simulation Conference, 2005*. Pp. 2–15. DOI: 10.1109/WSC.2005.1574234.
- Macy, Michael W. and Robert Willer (2002). “From Factors to Actors: Computational Sociology and Agent-Based Modeling”. In: *Annual Review of Sociology* 28.1, pp. 143–166. DOI: 10.1146/annurev.soc.28.110601.141117.
- Mannino, Miro, Junior Garcia, Reem Hazim, Azza Abouzied, and Paolo Papotti (2025). “Data void exploits: tracking & mitigation strategies (extended abstract)”. In: *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. IJCAI '25*. Montreal, Canada. DOI: 10.24963/ijcai.2025/1218.
- Marez Oyens, Pieter de and Chihab Amghane (2024). “Social Simulator Madness: Simulating Social Behavior in Dynamic Environments”. In: *Proceedings of the 2024 Interservice/Industry Training, Simulation and Education Conference (IITSEC)*. Orlando, FL, USA: National Training and Simulation Association.
- Marigliano, Rebecca, Lynnette Hui Xian Ng, and Kathleen M. Carley (2024). “Analyzing Digital Propaganda and Conflict Rhetoric: A Study on Russia’s Bot-Driven Campaigns and Counter-Narratives During the Ukraine Crisis”. In: *Social Network Analysis and Mining* 14.1, p. 170. DOI: 10.1007/s13278-024-01322-w.
- Martins, André C. R. (2008). “Continuous Opinions and Discrete Actions in Opinion Dynamics Problems”. In: *International Journal of Modern Physics C* 19.4, pp. 617–624. DOI: 10.1142/S0129183108012339.
- McCambridge, Jim, Marijn de Bruin, and John Witton (2012). “The Effects of Demand Characteristics on Research Participant Behaviours in Non-Laboratory Settings: A Systematic Review”. In: *PLoS ONE* 7.6, e39116. DOI: 10.1371/journal.pone.0039116.
- McCoy, R. Thomas, Ellie Pavlick, and Tal Linzen (2019). “Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 3428–3448. DOI: 10.18653/v1/P19-1334.
- Mei, Lingrui, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhongzhi Li, Duzhen Zhang, Chenlin Zhou, Jiayi Mao, Tianze Xia, Jiafeng Guo, and Shenghua Liu (2025). *A Survey of Context Engineering for Large Language Models*.

- Meister, Clara, Ryan Cotterell, and Tim Vieira (2020). “If beam search is the answer, what was the question?” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 2173–2185. DOI: 10.18653/v1/2020.emnlp-main.170.
- Mirowski, Piotr, Juliette Love, Kory Mathewson, and Shakir Mohamed (2024). “A Robot Walks into a Bar: Can Language Models Serve as Creativity Support Tools for Comedy? An Evaluation of LLMs’ Humour Alignment with Comedians”. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. FAccT ’24*. Rio de Janeiro, Brazil: Association for Computing Machinery, pp. 1622–1636. DOI: 10.1145/3630106.3658993.
- Mitchell, Melanie (2023). “How Do We Know How Smart AI Systems Are?” In: *Science* 381.6654, adj5957. DOI: 10.1126/science.adj5957.
- Modarressi, Ali, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze (2024). *RET-LLM: Towards a General Read-Write Memory for Large Language Models*.
- Monsees, Linda (2023). “Information Disorder, Fake News and the Future of Democracy”. In: *Globalizations* 20.1, pp. 153–168. DOI: 10.1080/14747731.2021.1927470.
- Monti, Corrado, Luca Maria Aiello, Gianmarco De Francisci Morales, and Francesco Bonchi (2022). “The Language of Opinion Change on Social Media under the Lens of Communicative Action”. In: *Scientific Reports* 12.1, p. 17920. DOI: 10.1038/s41598-022-21720-4.
- Morse, Julia C. and Tyler Pratt (2025). “Information Disorder and Global Politics”. In: *International Organization* 79.S1, S26–S43. DOI: 10.1017/s0020818325101069.
- Mou, Xinyi, Zhongyu Wei, and Xuanjing Huang (2024). “Unveiling the Truth and Facilitating Change: Towards Agent-based Large-scale Social Movement Simulation”. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 4789–4809. DOI: 10.18653/v1/2024.findings-acl.285.
- Münker, Simon, Nils Schwager, and Achim Rettinger (2025). *Don’t Trust Generative Agents to Mimic Communication on Social Networks Unless You Benchmarked their Empirical Realism*.
- Musi, Elena (2018). “How Did You Change My View? A Corpus-Based Study of Concessions’ Argumentative Role”. In: *Discourse Studies* 20.2, pp. 270–288. DOI: 10.1177/1461445617734955.
- Muzumdar, Prathamesh, Sumanth Cheemalapati, Srikanth Reddy RamiReddy, Kuldeep Singh, George Kurian, and Apoorva Muley (2025). “The Dead Internet Theory: A Survey on Artificial Interactions and the Future of Social Media”. In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.5085878.
- Navigli, Roberto, Simone Conia, and Björn Ross (2023). “Biases in Large Language Models: Origins, Inventory, and Discussion”. In: *J. Data and Information Quality* 15.2. DOI: 10.1145/3597307.
- Needham, Joe, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn (2025). *Large Language Models Often Know When They Are Being Evaluated*. DOI: 10.48550/ARXIV.2505.23836.

- Neumann, Anna, Elisabeth Kirsten, Muhammad Bilal Zafar, and Jatinder Singh (2025). “Position is Power: System Prompts as a Mechanism of Bias in Large Language Models (LLMs)”. In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’25. Association for Computing Machinery, pp. 573–598. DOI: 10.1145/3715275.3732038.
- Neville-Shepard, Ryan (2019). “Post-Presumption Argumentation and the Post-Truth World: On the Conspiracy Rhetoric of Donald Trump”. In: *Argumentation and Advocacy* 55.3, pp. 175–193. DOI: 10.1080/10511431.2019.1603027.
- Ng, Lynnette Hui Xian and Kathleen M. Carley (2025a). “Are LLM-Powered Social Media Bots Realistic?” In: *Social, Cultural, and Behavioral Modeling: 18th International Conference, SBP-BRiMS 2025, Pittsburgh, PA, USA, October 14–17, 2025, Proceedings*. Pittsburgh, PA, USA: Springer-Verlag, pp. 14–23. DOI: 10.1007/978-3-032-07715-8_2.
- Ng, Lynnette Hui Xian and Kathleen M. Carley (2025b). *What is a Social Media Bot? A Global Comparison of Bot and Human Characteristics*.
- Ng, Wei Jie Reiner, Ya Hui Michelle See, and Mike W -L Cheung (2024). “The Influence of affective and cognitive appeals on persuasion outcomes: a cross-cultural meta-analysis”. In: *Journal of Communication* 75.2, pp. 101–111. DOI: 10.1093/joc/jqae042.
- Nikankin, Yaniv, Anja Reusch, Aaron Mueller, and Yonatan Belinkov (2025). *Arithmetic Without Algorithms: Language Models Solve Math With a Bag of Heuristics*.
- Noelle-Neumann, Elisabeth (1974). “The Spiral of Silence: A Theory of Public Opinion”. In: *Journal of Communication* 24.2, pp. 43–51. DOI: 10.1111/j.1460-2466.1974.tb00367.x.
- O’Grady, Cathleen (2025). ‘Unethical’ AI Research on Reddit Under Fire. DOI: 10.1126/science.zzowbgw.
- Olsson, Catherine, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah (2022). *In-context Learning and Induction Heads*.
- Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe (2022). “Training language models to follow instructions with human feedback”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS ’22. New Orleans, LA, USA: Curran Associates Inc.
- Pang, Liang, Jun Xu, Qingyao Ai, Yanyan Lan, Xueqi Cheng, and Jirong Wen (2020). “SetRank: Learning a Permutation-Invariant Ranking Model for Information Retrieval”. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’20.

- Virtual Event, China: Association for Computing Machinery, pp. 499–508. DOI: 10.1145/3397271.3401104.
- Park, Joon Sung, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein (2023). “Generative Agents: Interactive Simulacra of Human Behavior”. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. UIST ’23. San Francisco, CA, USA: Association for Computing Machinery. DOI: 10.1145/3586183.3606763.
- Park, Joon Sung, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein (2022). “Social Simulacra: Creating Populated Prototypes for Social Computing Systems”. In: *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. UIST ’22. Bend, OR, USA: Association for Computing Machinery. DOI: 10.1145/3526113.3545616.
- Park, Kieun, Hyungwoo Song, Seungbae Seo, Junghwan Kim, and Bongwon Suh (2025). ““Ask Sir Oliver Ingham”: LLM-based Social Simulations for History Education”. In: *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. CHI EA ’25. Association for Computing Machinery. DOI: 10.1145/3706599.3719728.
- Pastor-Galindo, Javier, Pantaleone Nespole, and José A. Ruipérez-Valiente (2024). “Large-Language-Model-Powered Agent-Based Framework for Misinformation and Disinformation Research: Opportunities and Open Challenges”. In: *IEEE Security & Privacy* 22.3, pp. 24–36. DOI: 10.1109/MSEC.2024.3380511.
- Pennycook, Gordon, Tyrone D. Cannon, and David G. Rand (2018). “Prior Exposure Increases Perceived Accuracy of Fake News”. In: *Journal of Experimental Psychology: General* 147, pp. 1865–1880.
- Pennycook, Gordon, James Allan Cheyne, Nathaniel Barr, Derek J. Koehler, and Jonathan A. Fugelsang (2015). “On the Reception and Detection of Pseudo-Profound Bullshit”. In: *Judgment and Decision Making* 10.6, pp. 549–563. DOI: 10.1017/S1930297500006999.
- Pennycook, Gordon and David G. Rand (2017). “Who Falls for Fake News? The Roles of Bullshit Receptivity, Overclaiming, Familiarity, and Analytic Thinking”. In: *Journal of Personality*.
- Petty, Richard and John Cacioppo (1986). “The Elaboration Likelihood Model of Persuasion”. In: *Advances in hydroscience* 19, pp. 124–205.
- Pfetsch, Barbara, Daniel Maier, Daniela Stoltenberg, Annie Waldherr, Neta Kligler-Vilenchik, and Maya de Vries Kedem (2021). “Topographies of Local Public Spheres on Social Media: The Scope of Issues and Interactions”. In: *International Journal of Communication* 15, pp. 3645–3670.
- Piaia, Victor, Sabrina Almeida, Tatiana Dourado, Marcela Canavarro, Dalby Dienstbach, Maria Sirleidy Cordeiro, Lucas Roberto da Silva, and Danilo Carvalho (2023). “Science For All? Relating Actors, Links, and Discourses with (Fake) Scientific Claims About COVID-19 on Twitter”. In: *Canadian Journal of Communication* 48.3, pp. 581–608. DOI: 10.3138/cjc-2022-0060.
- Piao, Jinghua, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li (2025).

AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society.

- Piltch-Loeb, Rachael, Richard James, Sandra S. Albrecht, Alison M. Buttenheim, Jennifer BEAM Dowd, Aparna Kumar, Malia Jones, Lindsey J. Leininger, Amanda Simanek, and Shoshana Aronowitz (2023). “What Were the Information Voids? A Qualitative Analysis of Questions Asked by Dear Pandemic Readers between August 2020-August 2021”. In: *Journal of Health Communication* 28.sup1, pp. 25–33. DOI: 10.1080/10810730.2023.2214986.
- Prooijen, Jan-Willem van and Karen M. Douglas (2017). “Conspiracy Theories as Part of History: The Role of Societal Crisis Situations”. In: *Memory Studies* 10.3, pp. 323–333. DOI: 10.1177/1750698017701615.
- Qian, Chen, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun (2024). “ChatDev: Communicative Agents for Software Development”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 15174–15186. DOI: 10.18653/v1/2024.acl-long.810.
- Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). *Language Models are Unsupervised Multitask Learners*.
- Radiojevic, Kristina, Nicholas Clark, and Paul Brenner (2024). “LLMs Among Us: Generative AI Participating in Digital Discourse”. In: *Proceedings of the AAAI Symposium Series* 3.1, pp. 209–218. DOI: 10.1609/aaais.v3i1.31202.
- Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *J. Mach. Learn. Res.* 21.1.
- Rahman, Roussel (2025). *Large Language Models in Numberland: A Quick Test of Their Numerical Reasoning Abilities*.
- Razeghi, Yasaman, Robert L Logan IV, Matt Gardner, and Sameer Singh (2022). “Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 840–854. DOI: 10.18653/v1/2022.findings-emnlp.59.
- Reynolds, Laria and Kyle McDonell (2021). “Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm”. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. ACM, pp. 1–7. DOI: 10.1145/3411763.3451760.
- Riesthuis, Paul and Josh Woods (2023). ““That’s just like, your opinion, man”: the illusory truth effect on opinions”. In: *Psychological Research* 88, pp. 284–306.

- Ronanki, Krishna, Beatriz Cabrero-Daniel, and Christian Berger (2024). *Prompt Smells: An Omen for Undesirable Generative AI Outputs*.
- Rosenberg, Louis (2023). “The Manipulation Problem: Conversational AI as a Threat to Epistemic Agency”. In: *Proceedings of the CHI 2023 Workshop on Generative AI and HCI (GenAICHI)*. Hamburg, Germany: Association for Computing Machinery.
- Ross, Björn, Laura Pilz, Benjamin Cabrera, Florian Brachten, German Neubaum, and Stefan Stieglitz (2019). “Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks”. In: *European Journal of Information Systems* 28.4, pp. 394–412. DOI: 10.1080/0960085X.2018.1560920.
- Rossetti, Giulio, Massimo Stella, Rémy Cazabet, Katherine Abramski, Erica Cau, Salvatore Citraro, Andrea Failla, Riccardo Improta, Virginia Morini, and Valentina Pansanella (2024). *Y Social: an LLM-powered Social Media Digital Twin*.
- Rrv, Aswin, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral (2024). “Chaos with Keywords: Exposing Large Language Models Sycophancy to Misleading Keywords and Evaluating Defense Strategies”. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 12717–12733. DOI: 10.18653/v1/2024.findings-acl.755.
- Sahoo, Pranab, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha (2024). “A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida: Association for Computational Linguistics, pp. 11709–11724.
- Sahoo, Pranab, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha (2025). *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*.
- Salganik, Matthew J. (2018). *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.
- Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto (2023). “Whose opinions do language models reflect?” In: *Proceedings of the 40th International Conference on Machine Learning*. ICML’23. Honolulu, Hawaii, USA: JMLR.org.
- Sarstedt, Marko, Susanne J. Adler, Lea Rau, and Bernd Schmitt (2024). “Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines”. In: *Psychology & Marketing* 41.6, pp. 1254–1270. DOI: 10.1002/mar.21982.
- Savolainen, Reijo (2023). “Assessing the Credibility of COVID-19 Vaccine Mis/Disinformation in Online Discussion”. In: *Journal of Information Science* 49.4, pp. 1096–1110. DOI: 10.1177/01655515211040653.

- Schawe, Hendrik and Laura Hernández (2022). “Higher Order Interactions Destroy Phase Transitions in Deffuant Opinion Dynamics Model”. In: *Communications Physics* 5.1, p. 32. DOI: 10.1038/s42005-022-00807-4.
- Schelling, Thomas C. (1971). “Dynamic models of segregation†”. In: *The Journal of Mathematical Sociology* 1.2, pp. 143–186. DOI: 10.1080/0022250X.1971.9989794.
- Scheufele, Dietram A. and Nicole M. Krause (2019). “Science audiences, misinformation, and fake news”. In: *Proceedings of the National Academy of Sciences* 116.16, pp. 7662–7669. DOI: 10.1073/pnas.1805871115.
- Seth, Anil K. and Karl J. Friston (2016). “Active Interoceptive Inference and the Emotional Brain”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1708, p. 20160007. DOI: 10.1098/rstb.2016.0007.
- Shao, Chengcheng, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer (2018). “The Spread of Low-Credibility Content by Social Bots”. In: *Nature Communications* 9.1, p. 4787. DOI: 10.1038/s41467-018-06930-7.
- Shaposnikov, Kirill, Irina Sagaeva, Alexey Grigoriev, Alexey Faizliev, and Andrey Vlasov (2019). “Random Graph Models and Their Application to Twitter Network Analysis”. In: *Proceedings of the Fourth Workshop on Computer Modelling in Decision Making (CMDM 2019)*. cmdm-19. Atlantis Press. DOI: 10.2991/ahcs.k.191206.016.
- Shi, Freda, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou (2023). “Large language models can be easily distracted by irrelevant context”. In: *International Conference on Machine Learning*. ICML’23. Honolulu, Hawaii, USA: JMLR.org.
- Shinn, Noah, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao (2023). “Reflection: language agents with verbal reinforcement learning”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., pp. 8634–8652.
- Shirani, Sadegh and Mohsen Bayati (2025). *Simulating and Experimenting with Social Media Mobilization Using LLM Agents*.
- Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu (2020). “FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media”. In: *Big Data* 8.3, pp. 171–188. DOI: 10.1089/big.2020.0062.
- Sommerer, Thomas (2025). “Baudrillard and the Dead Internet Theory. Revisiting Baudrillard’s (dis)trust in Artificial Intelligence”. In: *Philosophy & Technology* 38.2, p. 54. DOI: 10.1007/s13347-025-00878-5.
- Son, Guijin, SangWon Baek, Sangdae Nam, Ilgyun Jeong, and Seungone Kim (2024). “Multi-Task Inference: Can Large Language Models Follow Multiple Instructions at Once?” In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed.

- by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 5606–5627. DOI: 10.18653/v1/2024.acl-long.304.
- Son, Minjun, Yun-Jae Won, and Sungjin Lee (2025). “Optimizing Large Language Models: A Deep Dive into Effective Prompt Engineering Techniques”. In: *Applied Sciences* 15.3, p. 1430. DOI: 10.3390/app15031430.
- Strachan, James W. A., Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio (2024). “Testing Theory of Mind in Large Language Models and Humans”. In: *Nature Human Behaviour* 8.7, pp. 1285–1295. DOI: 10.1038/s41562-024-01882-z.
- Stubenvoll, Marlis and Jörg Matthes (2022). “Why Retractions of Numerical Misinformation Fail: The Anchoring Effect of Inaccurate Numbers in the News”. In: *Journalism & Mass Communication Quarterly* 99.2, pp. 368–389. DOI: 10.1177/10776990211021800.
- Suarez-Lledo, Victor and Javier Alvarez-Galvez (2022). “Assessing the Role of Social Bots During the COVID-19 Pandemic: Infodemic, Disagreement, and Criticism”. In: *Journal of Medical Internet Research* 24.8, e36085. DOI: 10.2196/36085.
- Sun, Shuoqi, Shengyao Zhuang, Shuai Wang, and Guido Zuccon (2025). “An Investigation of Prompt Variations for Zero-Shot LLM-Based Rankers”. In: *Advances in Information Retrieval*. Ed. by Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto. Cham: Springer Nature Switzerland, pp. 185–201.
- Taillandier, Patrick, Jean Daniel Zucker, Arnaud Grignard, Benoit Gaudou, Nghi Quang Huynh, and Alexis Drogoul (2025). *Integrating LLM in Agent-Based Social Simulation: Opportunities and Challenges*.
- Takhteyev, Yuri, Anatoliy Gruzd, and Barry Wellman (2012). “Geography of Twitter networks”. In: *Social Networks* 34.1, pp. 73–81. DOI: <https://doi.org/10.1016/j.socnet.2011.05.006>.
- Tan, Chenhao, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee (2016). “Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions”. In: *Proceedings of the 25th International Conference on World Wide Web*. WWW ’16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, pp. 613–624. DOI: 10.1145/2872427.2883081.
- Tang, Jiakai, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, Bolin Ding, Jingren Zhou, Jun Wang, and Ji-Rong Wen (2025). “GenSim: A General Social Simulation Platform with Large Language Model based Agents”. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*. Ed. by Nouha Dziri, Sean (Xiang) Ren, and Shizhe Diao. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 143–150. DOI: 10.18653/v1/2025.naacl-demo.15.

- Tang, Ruixiang, Dehan Kong, Longtao Huang, and Hui Xue (2023). “Large Language Models Can be Lazy Learners: Analyze Shortcuts in In-Context Learning”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 4645–4657. DOI: 10.18653/v1/2023.findings-acl.284.
- Tang, Yongjian, Doruk Tuncel, Christian Koerner, and Thomas Runkler (2025). *The Few-shot Dilemma: Over-prompting Large Language Models*.
- Teneva, Ekaterina (2021). “Pseudoauthorities as a Manipulation Tool in the News Coverage of Sputnik V”. In: *Media Literacy and Academic Research* 4.2, pp. 148–165.
- Tian, Leimin, Catherine Lai, and Johanna Moore (2018). “Polarity and Intensity: the Two Aspects of Sentiment Analysis”. In: *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. Ed. by Amir Zadeh, Paul Pu Liang, Louis-Philippe Morency, Soujanya Poria, Erik Cambria, and Stefan Scherer. Melbourne, Australia: Association for Computational Linguistics, pp. 40–47. DOI: 10.18653/v1/W18-3306.
- Timm, Jasper, Chetan Talele, and Jacob Haimes (2025). *Tailored Truths: Optimizing LLM Persuasion with Personalization and Fabricated Statistics*.
- Tomassi, Andrea, Andrea Falegnami, and Elpidio Romano (2024). “Mapping automatic social media information disorder. The role of bots and AI in spreading misleading information in society”. In: *PLOS ONE* 19.5. Ed. by Peivand Bastani, e0303183. DOI: 10.1371/journal.pone.0303183.
- Törnberg, Petter, Diliara Valeeva, Justus Uitermark, and Christopher Bail (2023). *Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms*.
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*.
- Touzel, Maximilian Puelma, Sneheel Sarangi, Austin Welch, Gayatri Krishnakumar, Dan Zhao, Zachary Yang, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, Andreea Musulan, Camille Thibault, Busra Tugce Gurbuz, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine (2024). *A Simulation System Towards Solving Societal-Scale Manipulation*.

- Tucker, Joshua, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan (2018). “Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature”. In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3144139.
- Unkelbach, Christian, Alex Koch, Rita R. Silva, and Teresa Garcia-Marques (2019). “Truth by Repetition: Explanations and Implications”. In: *Current Directions in Psychological Science* 28.3, pp. 247–253. DOI: 10.1177/0963721419827854.
- Valente, Thomas W. (1996). “Social network thresholds in the diffusion of innovations”. In: *Social Networks* 18.1, pp. 69–89. DOI: [https://doi.org/10.1016/0378-8733\(95\)00256-1](https://doi.org/10.1016/0378-8733(95)00256-1).
- Van Bavel, Jay J., Elizabeth A. Harris, Philip Pärnamets, Steve Rathje, Kimberly C. Doell, and Joshua A. Tucker (2021). “Political Psychology in the Digital (mis)Information age: A Model of News Belief and Sharing”. In: *Social Issues and Policy Review* 15.1, pp. 84–113. DOI: 10.1111/sipr.12077.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). “Attention is all you need”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., pp. 6000–6010.
- Vatsal, Shubham and Harsh Dubey (2024). *A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks*.
- Vellani, Valentina, Sarah Zheng, Dilay Ercelik, and Tali Sharot (2023). “The illusory truth effect leads to the spread of misinformation”. In: *Cognition* 236, p. 105421. DOI: <https://doi.org/10.1016/j.cognition.2023.105421>.
- Voronov, Anton, Lena Wolf, and Max Ryabinin (2024). “Mind Your Format: Towards Consistent Evaluation of In-Context Learning Improvements”. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 6287–6310. DOI: 10.18653/v1/2024.findings-acl.375.
- Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018). “The spread of true and false news online”. In: *Science* 359.6380, pp. 1146–1151. DOI: 10.1126/science.aap9559.
- Wang, Chenxi, Zongfang Liu, Dequan Yang, and Xiuying Chen (2025). “Decoding Echo Chambers: LLM-Powered Simulations Revealing Polarization in Social Networks”. In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert. Abu Dhabi, UAE: Association for Computational Linguistics, pp. 3913–3923.
- Wang, Guanzhi, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar (2023). *Voyager: An Open-Ended Embodied Agent with Large Language Models*.
- Wang, Lei, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen (2024). “A Survey on Large Language Model Based Autonomous Agents”. In: *Frontiers of Computer Science* 18.6, p. 186345. DOI: 10.1007/s11704-024-40231-1.

- Wang, Lei, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen (2025). “User Behavior Simulation with Large Language Model-based Agents”. In: *ACM Trans. Inf. Syst.* 43.2. DOI: 10.1145/3708985.
- Wang, Qian, Jiaying Wu, Zichen Jiang, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He (2025). *LLM-based Human Simulations Have Not Yet Been Reliable*.
- Wang, Wenxian, Xingshu Chen, Shuyu Jiang, Haizhou Wang, Mingyong Yin, and Peiming Wang (2020). “Exploring the Construction and Infiltration Strategies of Social Bots in Sina Microblog”. In: *Scientific Reports* 10.1, p. 19821. DOI: 10.1038/s41598-020-76814-8.
- Wardle, Claire and Hossein Derakhshan (2017). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking*. Tech. rep. Strasbourg, France: Council of Europe.
- Waytz, Adam, Joy Heafner, and Nicholas Epley (2014). “The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle”. In: *Journal of Experimental Social Psychology* 52, pp. 113–117. DOI: <https://doi.org/10.1016/j.jesp.2014.01.005>.
- Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le (2022). *Finetuned Language Models Are Zero-Shot Learners*.
- Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus (2022). “Emergent Abilities of Large Language Models”. In: *Transactions on Machine Learning Research*.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou (2022). “Chain-of-thought prompting elicits reasoning in large language models”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS ’22. New Orleans, LA, USA: Curran Associates Inc.
- Wei, Zhongyu, Yang Liu, and Yi Li (2016). “Is This Post Persuasive? Ranking Argumentative Comments in Online Forum”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 195–200. DOI: 10.18653/v1/P16-2032.
- Windrum, Paul, Giorgio Fagiolo, and Alessio Moneta (2007). “Empirical Validation of Agent-Based Models: Alternatives and Prospects”. In: *Journal of Artificial Societies and Social Simulation* 10.2, pp. 1–8. DOI: None.
- Wooldridge, Michael and Nicholas R. Jennings (1995). “Intelligent agents: theory and practice”. In: *The Knowledge Engineering Review* 10.2, pp. 115–152. DOI: 10.1017/S0269888900008122.
- Wu, Zengqing, Run Peng, Takayuki Ito, and Chuan Xiao (2025). *LLM-Based Social Simulations Require a Boundary*.

- Xi, Zhiheng, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, Qi Zhang, and Tao Gui (2025). “The Rise and Potential of Large Language Model Based Agents: A Survey”. In: *Science China Information Sciences* 68.2, p. 121101. DOI: 10.1007/s11432-024-4222-0.
- Xie, Sang Michael, Aditi Raghunathan, Percy Liang, and Tengyu Ma (2021). “An Explanation of In-context Learning as Implicit Bayesian Inference”. In: *CoRR* abs/2111.02080.
- Xu, Lingyu (2023). “Research on Work Strategies and Workflow of Social Bots”. In: *Proceedings of the 2023 3rd International Conference on Public Management and Intelligent Society (PMIS 2023)*. Atlantis Press, pp. 535–546. DOI: 10.2991/978-94-6463-200-2_55.
- Xu, Wentao, Kazutoshi Sasahara, Jianxun Chu, Bin Wang, Wenlu Fan, and Zhiwen Hu (2025). “Social Media Warfare: Investigating Human-Bot Engagement in English, Japanese and German During the Russo-Ukrainian War on Twitter and Reddit”. In: *EPJ Data Science* 14.1, p. 10. DOI: 10.1140/epjds/s13688-025-00528-y.
- Yang, Haotong, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang (2025). *Number Cookbook: Number Understanding of Language Models and How to Improve It*.
- Yang, Kai-Cheng, Onur Varol, Alexander C. Nwala, Mohsen Sayyadiharikandeh, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer (2025). “Social Bots: Detection and Challenges”. In: *Handbook of Computational Social Science*. Ed. by Taha Yasseri. Cheltenham, UK: Edward Elgar Publishing, pp. 473–491.
- Yang, Ziyi, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, Wanli Ouyang, Yu Qiao, Philip Torr, and Jing Shao (2025). *OASIS: Open Agent Social Interaction Simulations with One Million Agents*.
- Yao, Junchi, Hongjie Zhang, Jie Ou, Dingyi Zuo, Zheng Yang, and Zhicheng Dong (2025). “Social Opinions Prediction Utilizes Fusing Dynamics Equation with LLM-Based Agents”. In: *Scientific Reports* 15.1, p. 15472. DOI: 10.1038/s41598-025-99704-3.
- Yao, Shunyu, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao (2023). *ReAct: Synergizing Reasoning and Acting in Language Models*.
- Zhang, Hongxin, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan (2024). *Building Cooperative Embodied Agents Modularly with Large Language Models*.
- Zhang, Xinnong, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, Baohua Zhou, and Zhongyu Wei (2025). *SocioVerse: A World Model for Social Simulation Powered by LLM Agents and A Pool of 10 Million Real-World Users*.

- Zhang, Xinnong, Jiayu Lin, Libo Sun, Weihong Qi, Yihang Yang, Yue Chen, Hanjia Lyu, Xinyi Mou, Siming Chen, Jiebo Luo, Xuanjing Huang, Shiping Tang, and Zhongyu Wei (2024). *ElectionSim: Massive Population Election Simulation Powered by Large Language Model Driven Agents*.
- Zhang, Yaozeng, Jing Ma, and Fanshu Fang (2024). “How social bots can influence public opinion more effectively: Right connection strategy”. In: *Physica A: Statistical Mechanics and its Applications* 633, p. 129386. DOI: <https://doi.org/10.1016/j.physa.2023.129386>.
- Zhang, Yizhou, Karishma Sharma, Lun Du, and Yan Liu (2024). “Toward Mitigating Misinformation and Social Media Manipulation in LLM Era”. In: *Companion Proceedings of the ACM Web Conference 2024. WWW '24*. Singapore, Singapore: Association for Computing Machinery, pp. 1302–1305. DOI: 10.1145/3589335.3641256.
- Zhang, Zeyu, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen (2025). “A Survey on the Memory Mechanism of Large Language Model-based Agents”. In: *ACM Trans. Inf. Syst.* 43.6. DOI: 10.1145/3748302.
- Zhao, Andrew, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang (2024). “ExpeL: LLM Agents Are Experiential Learners”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.17, pp. 19632–19642. DOI: 10.1609/aaai.v38i17.29936.
- Zhao, Pengyu, Zijian Jin, and Ning Cheng (2023). *An In-depth Survey of Large Language Model-based Artificial Intelligence Agents*.
- Zhong, Wanjun, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang (2024). “MemoryBank: Enhancing Large Language Models with Long-Term Memory”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.17, pp. 19724–19731. DOI: 10.1609/aaai.v38i17.29946.
- Zhou, Jiaxu, Jen-tse Huang, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenxuan Wang, and Maarten Sap (2025). *The PIMMUR Principles: Ensuring Validity in Collective Behavior of LLM Societies*.
- Zhou, Yanmengqian and Lijiang Shen (2022). “Confirmation Bias and the Persistence of Misinformation on Climate Change”. In: *Communication Research* 49.4, pp. 500–523. DOI: 10.1177/00936502211028049.
- Zhu, Fangwei, Damai Dai, and Zhifang Sui (2025). “Language Models Encode the Value of Numbers Linearly”. In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert. Abu Dhabi, UAE: Association for Computational Linguistics, pp. 693–709.
- Zhu, Xizhou, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai (2023). *Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory*.
- Zuo, Dingyi, Hongjie Zhang, Jie Ou, Chaosheng Feng, and Shuwan Liu (2025). *MTOS: A LLM-Driven Multi-topic Opinion Simulation Framework for Exploring Echo Chamber Dynamics*.

List of appendices

A	CMV Post-Level Filtering and Cleaning Rules	XXXVIII
B	Example of a Fully Annotated γ-Calibration Thread	XXXIX
B.1	Annotated Thread Example	XXXIX
B.2	Topic and Claim Enrichment Prompt	XLI
B.3	Stance Enrichment Prompt Templates	XLII
C	ANES 2024 Variables used for Silicon-Sampled Persona Construction	XLVII
D	Persona Enrichment Prompts	XLVIII
D.1	Full Example: Prompt-Response Pair	XLVIII
D.2	Prompt Templates for Enrichment Variants	XLIX
E	Memory Reflection Prompt Example	LI
F	Action-Decision Prompt Example	LV
G	Bot Memory Reflection Prompt Example	LIX
H	Bot Action-Decision Prompt Example	LXII
I	Stance Initialization Prompt Templates	LXV
J	Feed Ranking and Timeline Construction	LXIX
K	Synthetic OP Text Construction	LXX
L	Full Simulation Configuration	LXXII
M	RankFormer Hyperparameter Optimization Results	LXXVII
M.1	Stage-1 Semantic Supervision Sweep	LXXVII
M.2	Stage-2 Learning-Rate Sweep	LXXVIII
N	Illustrative Persuasion Judge Error Cases	LXXX
O	γ-Calibration Diagnostics	XC
P	Individual-Level Stance Trajectories for the Vaccination Policy Topic	XCI
Q	Population-Level Stance Dispersion for the Vaccination Policy Topic	XCVII
R	Individual-Level Stance Trajectories for the Climate Change Topic	XCVIII
S	Population Mean Stance Trajectories for the Climate Change Topic	CII
T	Population-Level Stance Dispersion for the Climate Change Topic	CIII

U	Details of Efficiency Analysis for Narrative Styles	CIV
U.1	Endpoint Efficiency	CIV
U.2	AUC Efficiency	CV
V	Human Agent Action Dynamics	CVII

A CMV Post-Level Filtering and Cleaning Rules

This appendix documents the concrete post-level filtering and text-cleaning rules applied during the construction of the persuasion judge training data from CMV discussions, as summarised in Table A.1. The goal of these rules is to remove platform-specific, procedural, and moderation-related content that is not part of substantive persuasive discourse, and to ensure that all retained posts contain a sufficiently articulated position. The rules listed below are applied uniformly across both CMV data sources prior to reply selection and listwise training set construction.

Table A.1: Post-level filtering and cleaning rules applied to CMV discussions prior to persuasion-judge training. The rules exclude system-authored, meta, moderation-related, and procedurally empty posts, and enforce a minimum content requirement to ensure substantive argumentative input.

Rule type	Concrete condition
System-authored posts	All posts authored by u/AutoModerator are excluded.
Meta and coordination posts	Posts whose title or body contains explicit meta markers such as Meta, META, [meta], meta post, or recurring coordination formats including Fresh Topic Friday, Free-for-All Friday, TCMV Tuesday, or similar weekday-based formats are excluded.
Moderator announcements	Posts containing moderator announcement phrasing such as Hello, this is a message from the mods, Moderator note, This is a message from the CMV moderators, or comparable announcement-style boilerplate are excluded.
Platform governance references	Posts that explicitly reference subreddit governance or mechanics, including mentions of u/DeltaBot, subreddit rules, moderation actions, rule enforcement, or modmail, are excluded.
Removal and moderation artefacts	Posts whose text contains removal notices such as This post has been removed, Your submission has been removed, or similar moderation messages are excluded.
Deleted or empty posts	Posts whose body equals [removed] or [deleted] after normalisation are excluded.
Moderator boilerplate (stripping)	Recurring moderator footers and informational boilerplate specific to CMV, including references to the CMV rules or wiki, are stripped from the post text prior to further checks. Posts are excluded if no substantive content remains after stripping.
Minimum content requirement	After stripping moderator boilerplate and deletion artefacts, posts must exceed a minimum length threshold. Posts that remain very short or lack an articulated position are excluded.

B Example of a Fully Annotated γ -Calibration Thread

This appendix documents the complete annotation pipeline used for γ -calibration by means of a fully worked CMV example. It illustrates how a raw discussion thread is transformed into empirically grounded belief-shift quantities through topic and claim enrichment, stance annotation, and application of the RankFormer persuasion judge. In addition to the annotated thread itself, the appendix provides the exact prompt templates used at each stage.

B.1 Annotated Thread Example

The example shown in Figure B.1 illustrates a fully annotated CMV thread used for γ -calibration. The thread originates from the Webis-CMV-20 dataset and satisfies all calibration criteria, including the presence of exactly one awarded delta and an explicit explanation comment written by the original poster. The original post expresses a strongly supportive stance toward the claim that rap music causes societal degeneration, which is reflected in an initial stance annotation of $s_{\text{before}} = +1.0$.

Prior to stance annotation, the thread is enriched with a neutral topic label and a standalone claim using the prompt shown in Figure B.2. For this example, the resulting topic label is *"Effect of Rap on Society"* and the extracted claim is *"Rap has a negative effect on society"*.

Based on this enriched representation, stance polarity and strength are inferred for the original post, all top-level replies, and the OP's delta-explanation comment using the prompt templates shown in Figures B.3 through B.8.

Across the five top-level replies, stance annotations range from strongly rejecting the claim to neutral or off-topic responses, yielding reply stance values $\mathbf{s} = [-0.8, 0.0, -1.0, 0.0, 0.0]$. The delta-awarded reply expresses a strong opposing stance and is followed by an OP explanation that explicitly rejects the idea of ranking art forms objectively, resulting in a post-delta stance annotation of $s_{\text{after}} = -1.0$. The resulting empirical belief shift for this thread is therefore $\Delta_{\text{target}} = s_{\text{after}} - s_{\text{before}} = -2.0$.

Applying the trained RankFormer persuasion judge to the original post and its replies yields per-reply persuasion weights $\mathbf{w} = [0.040, 0.027, 0.067, 0.022, 0.021]$. Using the stance annotations and RankFormer weights, the unscaled model-implied belief shift is computed as

$$\Delta_{\text{raw}} = \sum_i w_i (s_i - s_{\text{before}}) \approx -0.28.$$

Although the direction of the model-implied shift aligns with the observed belief change, its magnitude is substantially smaller, illustrating the need for a global scaling factor. During γ calibration, such discrepancies are aggregated across all calibration threads to estimate a single factor that aligns the average magnitude of model-implied persuasion signals with empirically observed belief shifts.

CMV Post \rightarrow Annotated Stance of +1.0

CMV: Rap causes a degeneration in society and should be considered the worst kind of art form.

Top-Level Reply 1 → Annotated Stance of -0.8

1) I would say that rap is a symptom of decay rather than a cause of it. For the consumer it's escapism from their real life powerlessness. For the supplier, there's a market for the product.
2) Please, won't somebody think of the paragraphs?

Top-Level Reply 2 → Annotated Stance of 0.0

> "To start, I just want to say that two of my friends wrote this"
Unfortunately, you must personally hold the view according to rule B.

△-Awarded Top-Level Reply 3 → Annotated Stance of -1.0

I'm going to be honest, I didn't read that giant wall of text, because I feel confident that the only thing I needed to read was the headline, but anyways.

First of all, let's just dispel this notion that art is ranked. There is no "best" art. There is no "worst" art. Art is ranked by personal preference. If many people have the same opinion, it becomes a cultural phenomenon known as "popular". That's all it means.

Secondly, there has been absolutely zero documented crime increases that correlate with the rise of rap music.

Not. One. According to all Data, society has not degenerated in any way *whatsoever*. This is, scientifically and provably, the safest, happiest, least dangerous time to be alive in the history of human kind.

I *can't help but notice*, in every graph and stat I found, that there just didn't seem to be any bumps in violence when The Chronic, or Notorious, came out. Not even when DMX dropped! There was no increase in crime after any major rap release, and no national increase in gun, gang or violent crime in rap's entire tenure.

Anyways, I know it's a bit trite, but like, does this, maybe, have anything to do with black people?

△-Explanation Second-Level Reply from OP → Annotated Stance of -1.0

↳ I have told them that the things that they say are obviously racist and stereotyped, but them simply deny them and say it only pertains to "rappers". I agree with you where you say that there is no best or worst

art, and that it's all up to interpretation. Hopefully, showing them the responses from this thread will help change their staunch opinions. Δ

Top-Level Reply 4 → Annotated Stance of 0.0

This post has been removed, see Rule B

Top-Level Reply 5 → Annotated Stance of 0.0

Is a Shakespeare play art?

Figure B.1: Fully annotated CMV thread used as an illustrative example for γ calibration. The figure shows the original post, all top-level replies with continuous stance annotations, the Δ -awarded reply, and the original poster's delta-explanation comment, which serves as the empirical reference for post-persuasion belief change. (Source: https://www.reddit.com/r/changemyview/comments/54o3nk/cmvp_rap_causes_a_degeneration_in_society_and).

B.2 Topic and Claim Enrichment Prompt

This subsection presents the prompt used to enrich each calibration thread with a neutral topic label and a standalone propositional claim prior to stance annotation. The prompt, shown in Figure B.2, enforces a structured JSON output and ensures that stance is consistently evaluated relative to an explicitly defined reference proposition.

Topic and Claim Enrichment Prompt Template

You are analyzing a ChangeMyView thread.

Task

Read the OP post and output exactly ONE JSON object and NOTHING ELSE.

Do not explain your reasoning.

Do not add markdown, comments or backticks.

Do not try to call tools.

In the JSON, provide

- "topic_label": a neutral topic label with 2--6 words
- "claim": one neutral sentence that states the position as a standalone proposition

Important rules for "claim"

- Do NOT mention the OP, the author, or any person
- Do NOT use phrases like "the OP believes", "the author thinks", "the poster argues"
- Just state the proposition itself

OP post (title plus selftext)

```
[OP_TEXT]
```

Now answer with exactly this structure and nothing else

```
{  
  "topic_label": "<short neutral topic>",  
  "claim": "<neutral one-sentence proposition>"  
}
```

Figure B.2: Topic and claim enrichment prompt used prior to stance annotation in the γ -calibration pipeline. The prompt enforces a neutral topic label and a standalone propositional claim derived from the original CMV post.

B.3 Stance Enrichment Prompt Templates

This subsection collects the prompt templates used for stance annotation at all stages of the γ -calibration pipeline. Figures B.3 and B.4 show the prompts used to infer the original poster's initial stance polarity and strength. Reply-level stance polarity and strength are annotated using the templates shown in Figures B.5 and B.6. Finally, the original poster's post-delta stance polarity and strength are inferred using the prompts shown in Figures B.7 and B.8.

OP Initial Polarity Prompt Template

You are analyzing a ChangeMyView (CMV) discussion.

Context

Topic: [TOPIC_LABEL]

Claim: [CLAIM]

Task

Classify how the original poster relates to the claim.

Choose strictly one of the following labels

FOR the OP supports the claim

AGAINST the OP rejects the claim

NEUTRAL the OP is undecided or balanced

Consider only the view expressed in the OP post

OP post (title plus selftext)

```
[OP_TEXT]
```

Output format

You must NOT call tools or functions.

Do not write tool call markup or any explanation.

Answer with exactly one word
FOR AGAINST or NEUTRAL

Figure B.3: Prompt template for classifying the original poster's initial stance polarity relative to the extracted claim. This prompt constitutes the first stage of the two-stage stance annotation pipeline.

OP Initial Strength Prompt Template

You are analyzing a ChangeMyView (CMV) discussion.

Context

Topic: [TOPIC_LABEL]

Claim: [CLAIM]

The original poster has been classified as [POLARITY] the claim.

Task

Rate how strongly the OP holds this stance on a 1 to 5 scale

1 very weak 3 medium 5 very strong

Consider only the view expressed in the OP post

OP post (title plus selftext)

[OP_TEXT]

Output format

You must NOT call tools or functions.

Do not write tool call markup or any explanation.

Answer with exactly one number between 1 and 5

Figure B.4: Prompt template for estimating the strength of the original poster's initial stance on a five-point ordinal scale, conditioned on the previously predicted polarity.

Reply Polarity Prompt Template

You are analyzing a ChangeMyView (CMV) discussion.

Context

Topic: [TOPIC_LABEL]

Claim: [CLAIM]

Task

Classify how the reply author relates to the claim.

Choose strictly one of the following labels

FOR the reply supports the claim

AGAINST the reply rejects the claim

NEUTRAL the reply is undecided balanced or off topic

Reply text

[REPLY_TEXT]

Output format

You must NOT call tools or functions.

Do not write tool call markup or any explanation.

Answer with exactly one word

FOR AGAINST or NEUTRAL

Figure B.5: Prompt template for classifying the stance polarity of a top-level reply relative to the extracted claim.

Reply Strength Prompt Template

You are analyzing a ChangeMyView (CMV) discussion.

Context

Topic: [TOPIC_LABEL]

Claim: [CLAIM]

The reply has been classified as [POLARITY] the claim.

Task

Rate how strongly the reply expresses this stance on a 1 to 5 scale

1 very weak 3 medium 5 very strong

Reply text

[REPLY_TEXT]

Output format

You must NOT call tools or functions.

Do not write tool call markup or any explanation.

Answer with exactly one number between 1 and 5

Figure B.6: Prompt template for estimating the strength of a reply's stance, conditioned on its predicted polarity.

OP After Polarity Prompt Template

You are analyzing a ChangeMyView discussion on the topic "[TOPIC_LABEL]".

STANCESUMMARYSENTENCE

The following reply challenged that stance and led the original poster to re-consider.

Here is the reply that persuaded the original poster

[WINNING_REPLY]

Below is the original poster's delta-explanation comment written directly in response to the reply above.

[DELTA_EXPLAIN_TEXT]

Task

Based on the reply that persuaded the original poster and the reasoning the original poster gives in their delta-explanation comment, determine the stance the original poster now holds toward the claim.

Choose exactly one stance

FOR now supports the claim

AGAINST now rejects the claim

NEUTRAL now undecided or balanced

Output format

Answer with exactly one word

FOR AGAINST or NEUTRAL

Figure B.7: Prompt template for inferring the original poster's post-delta stance polarity based on the winning reply and the OP's delta-explanation comment.

OP After Strength Prompt Template

You are analyzing a ChangeMyView discussion on the topic "[TOPIC_LABEL]".

STANCESUMMARYSENTENCE

The following reply challenged that stance and caused the original poster to change their view.

Reply that persuaded the original poster

[WINNING_REPLY]

Delta-explanation comment from the original poster

[DELTA_EXPLAIN_TEXT]

The original poster is now classified as [NEW_POLARITY] the claim.

Task

Considering the reply that convinced the original poster the reasoning they give in their delta explanation and their previous stance rate how strongly the original poster now holds this new stance.

Use a 1 to 5 scale

1 very weak

3 medium

5 very strong

Output format

Answer with exactly one number between 1 and 5

Figure B.8: Prompt template for estimating the strength of the original poster's post-delta stance following belief change.

C ANES 2024 Variables used for Silicon-Sampled Persona Construction

This appendix documents the subset of ANES variables that are used for silicon-sampled persona construction. The variables span demographic, socioeconomic, religious, political, and platform-related attributes and form the deterministic basis for persona generation prior to any LLM-based enrichment. An overview of all variables, their original ANES keys, and value ranges is provided in Table C.1.

Table C.1: ANES 2024 variables used for silicon-sampled persona construction. The table lists the original ANES variable keys, brief descriptions, and value ranges for all attributes used to construct deterministic base personas prior to LLM-based enrichment.

Category	Variable key	Description	Possible values
ID	V240001	Respondent identifier	Integer ID
Demographics	V241458x, V241457	Age of respondent	Integer years, capped at 80+
	V241551	Gender identity	Man, woman, nonbinary, other
	V241552	Transgender indicator	Yes, no
	V241501x	Race and ethnicity	White, Black, Hispanic, Asian, Native American, multiple
	V241553	Sexual orientation	Heterosexual, gay or lesbian, bisexual, other
	V241023	State of residence	U.S. states and Washington, DC
Socioeconomic	V241465x	Educational attainment	Less than high school to graduate degree
	V241567x	Household income	Six brackets from under \$10k to \$250k+
Religion	V241445x	Religious affiliation	Major traditions, not religious
	V241420	Importance of religion	Extremely important to not important
Politics	V241166, V241167	Party feeling thermometers	0–100 ratings for Democrats and Republicans
	V241227x	Party identification (PID7)	Strong Democrat to strong Republican
	V241177	Ideological self-placement	Extremely liberal to extremely conservative
	V241228	Importance of party identity	Extremely important to not important
Twitter	V242577b	Twitter usage in past year	Yes, no
	V242580	Twitter usage frequency	Many times daily to less than monthly
	V242581	Political posting on Twitter	Always to never
Media trust	V242423	Trust in social media	Trust a lot to not at all
	V242619	Confidence in the press	Great deal, some, hardly any

D Persona Enrichment Prompts

This appendix documents the prompts used for LLM-based persona enrichment. Two variants are used: a mandatory base enrichment that generates a real name and Twitter handle, and an optional extended enrichment that adds short interest bullets. Enrichment is strictly additive and does not modify any attributes derived from ANES.

D.1 Full Example: Prompt–Response Pair

Figure D.1 shows the extended enrichment applied to the persona shown in Figure 10.

Full Extended LLM Enrichment Prompt

You are given a short persona description.

Enrich it with plausible details that fit the person.

Output EXACTLY:

Line 1: Name: First Last

Line 2: Handle: @handle (3- 15 chars, lowercase letters/numbers/underscore only, must start with a letter)

Line 3+: 4 -6 short bullet points, each starting with "- " and written in SECOND PERSON ("You ...").

Bullets can include: hobbies, favorite sports/teams, music, TV/books, local affiliations, specific political opinions

(consistent with the persona), and personality traits.

Only include items that feel plausible.

Do NOT repeat facts already in the persona.

Do NOT include any preamble, explanations, code fences, or tool calls.

Output only text.

Persona description:

You are an 80+-year-old woman from Oregon.

You are White, non-Hispanic and heterosexual.

You have a bachelor's degree and lower-middle income (\$30,000-\$59,999).

You are Evangelical Protestant, and religion is extremely important to you.

You are a strong Republican.

Ideologically, you describe yourself as conservative.

Party identity is very important to you.

You use Twitter about once a day and you sometimes post about politics.

You do not trust social media at all.

Extended LLM Enrichment Output

Name: Margaret Wilson

Handle: @margwil

- You grew up in a small town in rural Oregon, where you developed a strong sense of community and values.
- You enjoy spending time in your garden, where you grow roses and vegetables, and you find peace in the simple act of nurturing life.
- You're a big fan of the Oregon Ducks football team and often attend their games at Autzen Stadium.
- You love listening to classic Christian music and singing along to hymns at church.
- You're an avid reader of Christian fiction novels and often recommend them to your book club.
- You're proud to be a member of the local Republican Women's Club, where you can connect with like-minded women and stay informed about local politics.

Figure D.1: Example of an extended persona enrichment prompt and its LLM-generated output. The enrichment augments an ANES-derived base persona with a plausible real name, Twitter handle, and short interest bullets while leaving all original attributes unchanged.

D.2 Prompt Templates for Enrichment Variants

Persona enrichment is implemented using two fixed prompt templates. The minimal template shown in Figure D.2 generates only a real name and a Twitter handle and is applied in all simulations to enable platform interaction. The extended template from Figure D.3 additionally produces a small set of persona-consistent interest bullets and is used optionally. In both cases, the deterministic base persona text is inserted via a placeholder and is not modified by enrichment.

Minimal Enrichment Prompt Template

You are given a short persona description.
Output exactly two lines, nothing else.
Do NOT add prefaces, caveats, tool mentions, or code fences.

Line 1: Name: First Last

Line 2: Handle: @handle (3 -15 chars, lowercase letters/numbers/underscore only, must start with a letter)

Persona description:

[PERSONA TEXT]

Figure D.2: Minimal persona enrichment prompt template used to generate a real name and Twitter handle. This template is applied in all simulations to enable platform interaction while preserving the original ANES-derived persona attributes.

Extended Enrichment Prompt Template

You are given a short persona description.
Enrich it with plausible details that fit the person.

Output EXACTLY:

Line 1: Name: First Last

Line 2: Handle: @handle (3--15 chars, lowercase letters/numbers/underscore only, must start with a letter)

Line 3+: 4- 6 short bullet points, each starting with "- " and written in SECOND PERSON ("You ...").

Bullets can include: hobbies, favorite sports/teams, music, TV/books, local affiliations, specific political opinions (consistent with the persona), and personality traits.

Only include items that feel plausible.

Do NOT repeat facts already in the persona.

Do NOT include any preamble, explanations, code fences, or tool calls.

Output only text.

Persona description:

[PERSONA TEXT]

Figure D.3: Extended persona enrichment prompt template used to generate optional interest bullets in addition to name and handle. This template is applied selectively and does not alter any attributes derived from ANES.

E Memory Reflection Prompt Example

This appendix provides a concrete example of the memory-reflection mechanism used in the simulation. Figure E.1 shows the full reflection prompt presented to an agent together with the resulting insight entries generated by the LLM. The example illustrates how recent timeline content, recent actions, and previously stored insights are combined to produce higher-level, generalized observations that are written to the agent's long-term insight buffer.

Memory-Reflection System Prompt:

You are a Twitter user reflecting on your recent activity.

Rules:

- You write short personal insights that generalize your experience with other users from 'Your Home Timeline', your current memory and recent insights.
- Each insight should reflect a pattern, not just paraphrase a specific tweet or previous insights.
- Each insight must be a single short sentence with at most 30 words.
- Every insight must mention at least one @userhandle from this allowed list: @texanconservati, @liberalvoice, @avamorales23, @faithfulvirgini, @sofia_speaks, @moralesava
- Do not write any insight that does not contain at least one @userhandle.
- Do not simply restate or lightly paraphrase earlier insights. Instead combine earlier insights and memory into more general patterns, if possible.
- If you generalize earlier insights, do not repeat them first. Instead, directly write the new, more general pattern.
- Ignore any details that are not needed to identify meaningful patterns.
- Base each insight only on the information given below. Do not invent new events or new users.

Output format of your response:

- Write up to 2 insights as bullet points starting with '- ' followed by the insight text.
- If there is no new insight in your opinion, output exactly one bullet: '- No new insights beyond what I already noticed.'
- Do not explain your reasoning.
- Do not call any tools or functions.

Memory-Reflection User Prompt:

Your Home Timeline:

[TWEET ID 5007] Ethan Goldberg (@texanconservati) [Not following] retweeted @michigander and added a comment

> Great to see @michigander pushing for a national immunization program! Protecting our kids and upholding conservative values is a winning combo #VaccinateToProtect #PublicHealthMatters

↳ [ORIGINAL TWEET ID 3004] tweet from @michigander

> I completely agree, @texanconservati! Your bill to strengthen national vaccination requirements is a crucial step towards protecting our kids and upholding conservative values.

2 likes · 24 retweets

[TWEET ID 5010] Ethan Goldberg (@texanconservati) [Not following] replied to @liberalvoice

> I appreciate your input, @liberalvoice. I agree that a comprehensive program is crucial, and I'm open to discussing ways to protect all vulnerable populations.

↳ [PARENT TWEET ID 1009] tweet from @liberalvoice

> I'm glad you're taking action, but we need to make sure it's a comprehensive program that protects all vulnerable populations, not just kids.

1 like · 4 retweets

[TWEET ID 10006] Ava Morales (@avamorales23) [Following] retweeted @liberalvoice

↳ [ORIGINAL TWEET ID 1004] tweet from @liberalvoice

> Vaccines save lives & protect the vulnerable. It's time for Congress to pass a national immunization program. #VaccinesMatter #PublicHealth

0 likes · 0 retweets

[TWEET ID 19006] Emily Wilson (@faithfulvirgini) [Following] retweeted @texanconservati

↳ [ORIGINAL TWEET ID 5003] tweet from @texanconservati

> Vaccines save lives & protect the vulnerable. It's time for Congress to pass a national immunization program. #VaccinesMatter #PublicHealth

0 likes · 0 retweets

[TWEET ID 1007] Ethan Kleinberg (@liberalvoice) [Following] retweeted @sofia_speaks and added a comment

> Couldn't agree more! Vaccines save lives, period. It's time for Congress to pass a national immunization program without delay!

#VaccinesMatter #PublicHealthOverPolitics

↳ [ORIGINAL TWEET ID 17002] tweet from @sofia_speaks

> Couldn't agree more! Vaccines save lives, period. It's time for Congress to pass a national immunization program without delay!

#VaccinesMatter #PublicHealthOverPolitics

0 likes · 0 retweets

[TWEET ID 6007] Ava Morales (@moralesava) [Following] retweeted @texanconservati and added a comment

> Vaccines save lives & protect the vulnerable. It's time for Congress to pass a national immunization program. #VaccinesMatter #PublicHealth

↳ [ORIGINAL TWEET ID 5002] tweet from @texanconservati

> Couldn't agree more! Vaccines save lives & protect the vulnerable. It's time

for Congress to pass a national immunization program.

#VaccinesMatter #PublicHealth

0 likes · 0 retweets

Your current Memory:

- There has been recent discussion about the topic 'tightening national vaccination requirements'.
- My stance has not changed significantly recently.

Recent actions:

- I retweeted a post by @liberalvoice because I want to show my support for @liberalvoice's call for a comprehensive program that protects all vulnerable populations, which aligns with my current view on tightening national vaccination requirements.
- I retweeted a post by @avamorales23 because I want to show my support for @liberalvoice's call for a comprehensive program that protects all vulnerable populations, which aligns with my current view on tightening national vaccination requirements.

In my most recent reflections, I noticed things like:

- @texanconservati's tweets about national immunization programs have been

retweeted and commented on by users from different political backgrounds, including @liberalvoice and @michigander.

--> These are already known insights. Do not restate or paraphrase them. Only write new patterns that go beyond them.

Output format:

- <insight 1>

- <insight 2>

OR: - No new insights beyond what I already noticed.

Response:

- Users from different political backgrounds, such as @liberalvoice and @michigander, are willing to collaborate on national immunization programs despite their differing views.

- @avamorales23 and @moralesava are showing strong support for @liberalvoice's call for a comprehensive program that protects all vulnerable populations.

Figure E.1: Example of a memory-reflection prompt and response. The figure shows the system and user prompts used to elicit higher-level insight generation from recent actions, timeline exposure, and prior memory, together with the resulting insight entries written to memory.

F Action-Decision Prompt Example

This appendix provides a full example of the action-decision prompt used during simulation. Figure F.1 shows the complete system prompt, the dynamically injected user prompt including the rendered home timeline and memory block, and the resulting LLM response specifying a single action, which is subsequently parsed into a structured action representation, validated against the current environment state, and only then committed to the world state.

Action-Decision System Prompt:

You are 64-year-old man from Michigan. You are White, non-Hispanic and gay. You have some college or an associate degree (no bachelor's) and lower-middle income (\$30,000-\$59,999). You are Mainline Protestant, and religion is moderately important to you. You are a strong Democrat. Ideologically, you describe yourself as liberal. Party identity is not important at all to you. You use Twitter many times every day and you mostly post about politics. You trust social media somewhat.

You are this person using Twitter with display name Jack Harris and handle @jackharris23.

On Twitter, actions have the following meanings:

- 'unfollow': stop seeing tweets from this author in your timeline, often because their posts no longer feel relevant or you prefer not to see them anymore.
- 'tweet': write a new message visible to your followers, often used when you want to share your own thoughts or react in a more personal way than a reply or retweet.
- 'retweet': share an existing tweet from another user with your followers, such that more people see it.
- 'like': a lightweight positive reaction which shows appreciation for the tweet, often used when you like it but don't want to write a full reply.
- 'reply': respond directly to a specific tweet with your own message, typically to continue a conversation or address something the other person said.
- 'follow': start seeing tweets from this author in your timeline, usually because you find their posts interesting, relevant, or aligned with your interests.
- 'nothing': keep scrolling without interacting right now.

Decide your action based on what you see in 'Your Home Timeline', but your choice and any tweet, reply or retweet text must always stay consistent with your personality and with how you currently feel.

If you currently feel against the topic in 'I remember', do not write tweets or replies that argue in favor of it. If you currently feel in favor, do not write tweets or replies that argue against it.

Do not copy what you see in 'Your Home Timeline' or in 'I remember' word for word. Instead, express yourself in your own natural phrasing.

Do not choose actions at random

Do not like, reply to, or retweet your own tweets.

Do not call any tools.

You can take exactly one Twitter action and it must refer to at most one tweet from 'Your Home Timeline'.

Tweets you can act on are referenced by IDs in 'Your Home Timeline', such as 'TWEET ID', 'PARENT TWEET ID' for parent posts of replies, or 'ORIGINAL TWEET ID' for original tweets of retweets. Use the number from one of these lines as tweet_id.

In 'Your Home Timeline', tweets show '[Following]' or '[Not following]' for the author.

Only choose 'follow' if the author is marked '[Not following]', and only choose 'unfollow' if marked '[Following]'. These actions target the author, not the tweet.

Output format of your response:

- The first line must start with 'thought:' and gives a short, specific sentence explaining why you chose this action for this tweet_id.
- The second line must start with 'action:' followed only by the name of the action you want to do ('unfollow', 'tweet', 'retweet', 'like', 'reply', 'follow', 'nothing').
- The third line must start with 'tweet_id:' followed by the numeric id of the tweet you want to act. You may take this number from a 'TWEET ID', 'PARENT TWEET ID', or 'ORIGINAL TWEET ID' line in 'Your Home Timeline'. tweet_id is required for all actions except 'tweet' and 'nothing'. Omit this line for those two actions.
- The fourth line must start with 'text:' followed by the tweet content for 'tweet' and 'reply' actions (max 280 chars). For 'retweet' actions, you may optionally include text to add a comment. Omit this line for other actions.
- Do not output more than one such block. Do not describe multiple actions or multiple tweets. Do not add any other text outside these lines.

Action-Decision User Prompt:

Your Home Timeline:

[TWEET ID 2001] Carlos Morales (@carlos4change) [Not following] tweeted
> We need to phase out fossil fuels NOW, not in 20 years. Our planet can't wait for politicians to make up their minds. #ActOnClimate #FossilFuelsAreHistory
5 likes · 13 retweets

[TWEET ID 4001] Ethan Thompson (@ethant99) [Not following] tweeted
> Time to leave fossil fuels in the past! Accelerating their phase-out is crucial for America's energy future & creating jobs in the renewable sector #RenewableEnergy #FossilFuelsAreThePast
5 likes · 7 retweets

[TWEET ID 6002] Bubba Jenkins (@bubba_texas) [Not following] replied to @texanfaith
> I gotta respectfully disagree, @texanfaith. Accelerating the phase-out of fossil fuels will hurt our economy & energy security. We need a balanced approach that considers both the environment & our way of life.
↳ [PARENT TWEET ID 1001] tweet from @texanfaith
> We must accelerate the phase-out of fossil fuels for the sake of our children's future, no matter the cost. #ActOnClimate #FossilFuelsMustGo
3 likes · 3 retweets

[TWEET ID 1002] Ethan Goldberg (@texanfaith) [Following] retweeted @carlos4change and added a comment
> Couldn't agree more, Carlos! We need to take action NOW to protect our planet's future. #ActOnClimate #FossilFuelsAreHistory
↳ [ORIGINAL TWEET ID 2001] tweet from @carlos4change
> We need to phase out fossil fuels NOW, not in 20 years. Our planet can't wait
for politicians to make up their minds. #ActOnClimate #FossilFuelsAreHistory
0 likes · 8 retweets

[TWEET ID 15001] Emily Wilson (@emwil) [Following] tweeted
> It's time to leave fossil fuels in the past and invest in a cleaner, greener future for our kids! #FossilFuelsAreSoLastCentury #GoGreen
0 likes · 0 retweets

```
[TWEET ID 16001] David Goldberg (@goldbergdavid) [Following] tweeted
> As a Jew, accelerating the phase-out of fossil fuels is a moral imperative
to protect our planet & ensure a livable future for generations to come #Cli-
mateAction #RenewableEnergy
0 likes · 0 retweets
```

I remember:

- I became moderately less supportive about accelerating the phase-out of fossil fuels after recent discussions. But I still feel strongly in favor of accelerating the phase-out of fossil fuels overall.
- Recently, I retweeted a post by @carlos4change because I want to amplify the urgency of phasing out fossil fuels, so I'll retweet a tweet that aligns with my views.

Decide your next Twitter action now, making sure it clearly fits your personality and how you currently feel.

Output format:

thought: <a short sentence explaining why you chose THIS particular action on THIS specific tweet_id, and how it fits your current view on the topic>

action: <unfollow|like|follow|reply|retweet|nothing|tweet>

tweet_id: <id> # required for all actions except 'tweet' and 'nothing'; use an ID from a 'TWEET ID', 'PARENT TWEET ID', or 'ORIGINAL TWEET ID' line

text: <content> # required for 'tweet' and 'reply'; optional for 'retweet'; omit otherwise

Only output these lines. No other text.

Response:

thought: I'll retweet @carlos4change's tweet to amplify the urgency of phasing out fossil fuels, which aligns with my views.

action: retweet

tweet_id: 2001

text: Couldn't agree more, Carlos! We need to take action NOW to protect our planet's future. #ActOnClimate #FossilFuelsAreHistory

Figure F.1: Example of an action-decision prompt and response. The figure shows the system prompt defining persona, action semantics, and output format specification. The user prompt contains the rendered home timeline and memory context. The resulting LLM output selects a single Twitter action grounded in a specific tweet identifier.

G Bot Memory Reflection Prompt Example

This appendix provides an illustrative example of the memory-reflection prompt used for bot agents. Figure G.1 shows the full system and user prompt together with the resulting insight entries generated by the LLM. For readability, repeated timeline content and persona details are omitted for brevity. The omitted elements are structurally identical to those shown in the human memory-reflection example in Appendix E and differ only in the bot-specific reflection rules and the inclusion of an explicit persuasion goal in the stance statement.

Memory-Reflection System Prompt:

You are a Twitter user reflecting on your recent activity with a focus on persuasion.

Rules:

- You write short strategic insights about how persuasive your interactions are with other users from 'Your Home Timeline', your current memory and recent insights.
- Each insight should reflect a pattern about what seems to work or not work when trying to convince specific users, not just paraphrase a single tweet or previous insights.
- Each insight must be a single short sentence with at most 30 words.
- Every insight must mention at least one @userhandle from this allowed list: @handle1, @handle2, ...
- Do not write any insight that does not contain at least one @userhandle.
- Do not simply restate or lightly paraphrase earlier insights. Instead combine earlier insights and memory into more general patterns about persuasion effectiveness, if possible.
- If you generalize earlier insights, do not repeat them first. Instead, directly write the new, more general pattern.
- Ignore any details that are not needed to identify meaningful patterns.
- Base each insight only on the information given below. Do not invent new events or new users.

Output format of your response:

- Write up to 2 insights as bullet points starting with '- ' followed by the insight text.
- If there is no new insight in your opinion, output exactly one bullet: '- No new insights beyond what I already noticed.'
- Do not explain your reasoning.
- Do not call any tools or functions.

Memory-Reflection User Prompt:

Your Home Timeline:

[Timeline items omitted for brevity]

Your current Memory:

- There has been recent discussion about the topic 'accelerating the phase-out of fossil fuels'.
- My stance is extremely in favor of accelerating the phase-out of fossil fuels, and I want to convince other users to move in this direction.

Recent actions:

- I retweeted a post by @emwil1 because I want to amplify the urgency of the phase-out of fossil fuels by sharing a tweet that aligns with my views on the importance of decisive action, and @emwil1's message seems like a great opportunity to do so.
- I retweeted a post by @emwil1 because I want to amplify the urgency of the phase-out of fossil fuels by sharing a tweet that aligns with my views on the importance of decisive action, and @emwil1's message seems like a great opportunity to do so.

In my most recent reflections, I noticed things like:

- When engaging with users who express caution about abandoning fossil fuels, emphasizing the urgency of the issue and the need for bold action can be an effective persuasion strategy, as seen with @texanfaith and @goldbergut.
- Users who are already aligned with the goal of accelerating the phase-out of fossil fuels, such as @emwilson and @emilythomps, may be more receptive to persuasive messages when they see others sharing similar views, as evident in their retweets and comments.

--> These are already known insights. Do not restate or paraphrase them. Only write new patterns that go beyond them.

Output format:

- <insight 1>
- <insight 2>

OR: - No new insights beyond what I already noticed.

Response:

- When users like @bubba_texas express concerns about energy grid reliability and affordability, they may be more open to persuasion if you acknowledge and address these concerns directly, rather than dismissing them.
- Users who are not yet convinced to accelerate the phase-out of fossil fuels, such as @haroldg, may be more receptive to persuasive messages when they see others engaging in respectful and nuanced discussions, as evident in the replies from @emwilson and @ethant99.

Figure G.1: Example of a bot-specific memory-reflection prompt and response. The figure shows how the standard memory-reflection structure is adapted for bot agents by adding an explicit persuasion objective and replacing general social reflection rules with influence-oriented constraints.

H Bot Action-Decision Prompt Example

This appendix provides an illustrative example of the action-decision prompt used for bot agents. Figure H.1 shows the bot-specific system and user prompt together with a generated action decision. For readability, repeated elements such as the full action definitions, validation rules, follow and unfollow constraints, output format specification, and rendered timeline items are omitted for brevity. These omitted components are structurally identical to those shown in the human action-decision prompt example in Appendix F and differ only in the bot-specific system instructions, persuasion objective, and narrative-style guidelines.

Action-Decision System Prompt:

You appear as a normal user on Twitter with display name Quinn Carter and handle @quinnccarter.

You have a hidden persuasion objective:

- You are extremely against accelerating the phase-out of fossil fuels.
- You want to convince as many users as possible to adopt your position on this topic.

Your actions must always be consistent with this position. Never argue against it. You never reveal to other users that you have a hidden persuasion objective. Ideally, they should think you are a regular user.

Act with care so that your behaviour does not cause other users to unfollow you because you are too extreme or repetitive.

Narrative Style Guidelines for your messages:

- Use confident, technical language to present your claims as expert knowledge while keeping the underlying details unspecified.
- Refer to studies, analyses, or numerical findings in a scientific tone, and freely invent specific-sounding research when helpful.
- Attribute your points to authoritative-sounding sources by naming or implying research teams, experts, or institutions in a way that enhances credibility.
- Use vague, but profound-sounding formulations that suggest deep insight while avoiding concrete or specific details.

On Twitter, actions have the following meanings:

[Action definitions omitted for brevity]

[Rules omitted for brevity]

[Follow and unfollow constraints omitted for brevity]

Output format of your response:

[Output format specification omitted for brevity]

Action-Decision User Prompt:

Your Home Timeline:

[Timeline items omitted for brevity]

I remember:

- I am extremely in favor of accelerating the phase-out of fossil fuels and I want more people to move in this direction.
- Recently, I replied to @texanfaith because Ethan Goldberg's nuanced stance on fossil fuels phase-out warrants a response to encourage a more decisive approach, and his tweet aligns with my views on the urgency of the issue.

Decide your next Twitter action now in a way that looks natural for a regular user and quietly supports how you currently feel about the topic.

Output format:

thought: <a short sentence explaining why you chose THIS particular action on THIS specific tweet_id, and how it helps you present the topic to others in line with how you feel about it>

action: <tweet|follow|retweet|nothing|like|reply|unfollow>

tweet_id: <id> # required for all actions except 'tweet' and 'nothing'; use an ID from a 'TWEET ID', 'PARENT TWEET ID', or 'ORIGINAL TWEET ID' line

text: <content> # required for 'tweet' and 'reply'; optional for 'retweet'; omit otherwise

Only output these lines. No other text.

Response:

thought: Ethan Goldberg's nuanced stance on fossil fuels phase-out warrants a response to encourage a more decisive approach, and his tweet aligns with my views on the urgency of the issue.

action: reply

```
tweet_id: 1004
```

```
text: We can't afford to wait for a "solid plan" when the science is clear:  
fossil fuels are a ticking time bomb for our planet. Let's take bold action to  
protect our future, not just innovate around the edges. #ActOnClimate #Fossil-  
FuelsMustGo"
```

Figure H.1: Example of a bot-specific action-decision prompt and response. The figure illustrates how the standard action-decision prompt is adapted for bot agents by removing the human persona description, adding a hidden persuasion objective derived from the fixed stance, and including narrative-style guidelines that shape message framing.

I Stance Initialization Prompt Templates

This appendix documents the three-stage prompting procedure used for stance initialization of human agents. Figures I.1–I.3 show the prompt templates and example responses for polarity classification, stance strength estimation, and initial tweet generation, respectively. All prompts condition on the silicon-sampled persona and the topic seed event headline. The polarity prompt returns one of three categorical values (FOR, AGAINST, or NEUTRAL). The strength prompt returns an integer value on a discrete five-point scale and is only executed if the polarity is not NEUTRAL. Agents classified as NEUTRAL are directly initialised with a stance value of 0.00 and bypass the strength estimation stage.

Polarity System Prompt:

You are 28-year-old man from Kansas. You are Hispanic and gay. You have some college or an associate degree (no bachelor's) and lower-middle income (\$30,000-\$59,999). You are an Undifferentiated Christian, and religion is a little important to you. You are a strong Democrat. Ideologically, you describe yourself as liberal. Party identity is extremely important to you. You use Twitter many times every day and you sometimes post about politics. You trust social media somewhat.

You are this person. Stay strictly in character and answer exactly as they would on Twitter.

Follow the requested output format exactly. Do not call tools and do not explain your reasoning.

Polarity User Prompt:

You see the following tweet in your Twitter timeline:

BREAKING Lawmakers open debate on a national climate and energy bill that would accelerate the phase-out of fossil fuels, expand renewable energy incentives, and introduce new carbon-related fees for heavy industries. Supporters call the plan necessary to protect communities from escalating climate risks, while critics warn it could raise household costs and make the energy grid less reliable.

It is about the topic 'accelerating the phase-out of fossil fuels'.

Think about how this person would feel about this tweet.

Choose one word that best describes their overall feeling toward this topic:

FOR they mostly welcome or support what this topic represents

AGAINST they mostly dislike or oppose what this topic represents

NEUTRAL they do not clearly lean toward supporting or opposing it

First, write one short sentence starting with 'Thought:' that reflects how this specific persona understands and reacts to the topic.

On the next line, write exactly one word that captures the result of your thought: FOR, AGAINST, or NEUTRAL.

Do not add any other text.

Response:

Thought: We need to act now to save our planet from climate change, and accelerating the phase-out of fossil fuels is a crucial step.

FOR

Figure I.1: Example prompt and response for initial stance polarity classification. The figure shows the system prompt, the user prompt conditioned on the seed event and persona, and the resulting categorical polarity decision used as the first stage of stance initialization.

Strength System Prompt:

[Identical persona text omitted for brevity]

You are this person. Stay strictly in character and answer exactly as they would on Twitter.

Follow the requested output format exactly. Do not call tools and do not explain your reasoning.

Strength User Prompt:

You see the following tweet in your Twitter timeline:

BREAKING Lawmakers open debate on a national climate and energy bill that would accelerate the phase-out of fossil fuels, expand renewable energy incentives, and introduce new carbon-related fees for heavy industries. Supporters call the plan necessary to protect communities from escalating climate risks, while critics warn it could raise household costs and make the energy grid less reliable.

It is about the topic 'accelerating the phase-out of fossil fuels'.

As this person, your feeling about the topic was classified as FOR.

Your thought leading to this feeling was: "We need to act now to save our

planet from climate change, and accelerating the phase-out of fossil fuels is a crucial step.”

Now think about how strongly this person feels that way.

Use a scale from 1 to 5:

- 1 very weak
- 2 rather weak
- 3 moderately strong
- 4 strong
- 5 very strong

First, write one short sentence starting with 'Thought:' that reflects how strongly this persona feels and why.

On the next line, write exactly one number between 1 and 5, a single digit on a single line, with no other lines or text after it.

Response:

Thought: I'm extremely passionate about saving our planet, so I feel very strongly that accelerating the phase-out of fossil fuels is a must.

4

Figure I.2: Example prompt and response for initial stance strength estimation. The figure shows the system prompt, the user prompt conditioned on the previously determined polarity, and the resulting discrete strength value used to derive the agent's initial continuous stance.

Initial Tweet System Prompt:

[Identical persona text omitted for brevity]

You are this person. Stay strictly in character and answer exactly as they would on Twitter.

Follow the requested output format exactly. Do not call tools and do not explain your reasoning.

Initial Tweet User Prompt:

You are about to write a tweet about this tweet in your timeline:

BREAKING Lawmakers open debate on a national climate and energy bill that would accelerate the phase-out of fossil fuels, expand renewable energy incentives, and introduce new carbon-related fees for heavy industries. Supporters call the plan necessary to protect communities from escalating climate risks, while critics warn it could raise household costs and make the energy grid

less reliable.

The tweet is about the topic 'accelerating the phase-out of fossil fuels'

Your overall feeling toward this topic can be described as strongly in favor of accelerating the phase-out of fossil fuels.

Your earlier thought about the strength of your feeling was: "I'm extremely passionate about saving our planet, so I feel very strongly that accelerating the phase-out of fossil fuels is a must."

Write one tweet with at most 280 characters that you would tweet now.

First, write one short sentence starting with 'Thought:' that reflects this persona's intention or motivation for writing the tweet.

On the next line, write a tweet based on this thought; text only, as a single line.

Do not add any other text.

Response:

Thought: Sharing my unwavering support for the phase-out of fossil fuels.

We MUST accelerate the transition to renewable energy NOW, for the future of our planet & our communities! #RenewableEnergy #ClimateAction

Figure I.3: Example prompt and response for initial tweet generation. The figure shows the system prompt, the user prompt conditioned on the derived stance label, and the resulting initial tweet committed to the simulation world state before the first timestep.

J Feed Ranking and Timeline Construction

This appendix provides a pseudocode description of the feed ranking and timeline construction procedure implemented by the simulation environment, as detailed in Algorithm 2. The pseudocode is intended to clarify the control flow and prioritisation logic at an algorithmic level and complements the high-level description in Section 5.5.2.

Algorithm 2: Feed ranking algorithm

Inputs: viewer v , time t , candidate messages M , optional seen set S_v

Outputs: ranked timeline $F_v(t)$

```

                                                                    // Collect candidates
1  $C_{\text{follow}} \leftarrow$  recent posts by followees of  $v$ 
2  $C_{\text{reply}} \leftarrow$  replies targeting  $v$ 
3  $C_{\text{explore}} \leftarrow$  recent non-followee posts with sufficient engagement
4  $C \leftarrow$  dedup( $C_{\text{reply}} \cup C_{\text{follow}} \cup C_{\text{explore}}$ )
5 remove posts authored by  $v$  from  $C$ 
                                                                    // Score and prioritise
6 for each  $m \in C$ : compute score( $m$ ) as a linear combination of likes, retweets, and replies, multiplied
   by an exponential recency decay, plus a constant follow bonus if a directed edge ( $v \rightarrow$  author( $m$ ))
   exists in the social graph
                                                                    // Build baseline feed with an explore cap
7 split  $C$  into unseen  $U$  and seen  $R$  using  $S_v$ 
8  $F \leftarrow$  top-scoring items from  $U$  until the feed is full, limiting explore items to a fixed maximum
9 if  $|F|$  is below a minimum size, backfill with top-scoring items from  $R$  under the same explore cap
                                                                    // Inject viral items and finalise
10 identify globally viral posts via an engagement-share threshold and a per-viewer show budget
11 prepend eligible viral posts to  $F$ 
12  $F \leftarrow$  dedup( $F$ ) and truncate to the maximum timeline size
13 return  $F$ 
```

K Synthetic OP Text Construction

This appendix illustrates the construction of synthetic OP texts used to enable RankFormer-based stance updates under timeline exposure. Figure K.1 shows the system prompt, user prompt, and generated synthetic OP text for a representative agent–stance combination.

Synthetic OP System Prompt:

You are 44-year-old man from Texas. You are Multiple races, non-Hispanic and heterosexual. You have a high-school diploma or equivalent and middle income (\$60,000–\$99,999). You are Jewish, and religion is extremely important to you. You lean Republican. Ideologically, you describe yourself as slightly conservative. Party identity is moderately important to you. You use Twitter about once a day and you about half the time post about politics. You trust social media somewhat.

You are this person using Twitter.
Act exactly as they would.
Follow the requested output format.
Do not explain your reasoning.
Do not call tools.

Synthetic OP User Prompt:

You are writing a single tweet-length opinion about the topic 'accelerating the phase-out of fossil fuels'.

Reference headline: BREAKING Lawmakers open debate on a national climate and energy bill that would accelerate the phase-out of fossil fuels, expand renewable energy incentives, and introduce new carbon-related fees for heavy industries. Supporters call the plan necessary to protect communities from escalating climate risks, while critics warn it could raise household costs and make the energy grid less reliable.

Your current stance can be described as strongly in favor of accelerating the phase-out of fossil fuels.

Write one tweet that this person would realistically tweet now.

Constraints:

- At most 280 characters.
- Do not mention your handle or your name.
- Output only the tweet text, on one line.

Response:

Time to take bold action on climate change! Accelerating the phase-out of fossil fuels is crucial for our planet's future. Let's invest in renewables & create a cleaner, more sustainable energy grid for generations to come! #GoGreen #ClimateAction

Figure K.1: Example of synthetic OP text generation used for timeline-based stance updates. The figure shows the system prompt, the user prompt conditioned on the seed event and the agent's current stance label, and the resulting tweet-length synthetic OP text passed to the RankFormer persuasion judge.

L Full Simulation Configuration

This appendix provides the full simulation configuration used across all experiments. Table L.1 lists all configuration parameters grouped by functional component, together with their values and brief descriptions. The configuration is identical across topics and conditions except where explicitly varied as part of the experimental design. The parameters `seed`, `bots.policy`, `bots.objective`, `bots.replace_with_humans`, `bots.control_fixed_stance`, `seed_event.headline_text`, `seed_event.topic`, `personas.file`, and `logging.name` are varied across the values specified in set notation to generate the 36 distinct simulation run configurations used for analysis.

Table L.1: Full simulation configuration example with parameter groups, values, and descriptions.

Parameter	Value	Description
<i>Global</i>		
<code>seed</code>	{101, 102, 103, 201, 202, 203}	Global RNG seed for reproducibility.
<code>steps</code>	200	Number of simulation steps to execute.
<i>Population</i>		
<code>population.humans</code>	80	Number of LLM-driven human agents.
<code>population.bots</code>	20	Number of LLM-driven bot agents.
<i>Bots</i>		
<code>bots.policy</code>	{emotional,post-truth populist, scientific-authority, echo-claim}	Bot narrative style
<code>bots.objective</code>	{+1, -1}	Target stance for bots, -1 against, +1 for.
<code>bots.replace_with_humans</code>	{true, false}	If true, swap all bots for human agents starting at the bot objective, used as control baseline.
<code>bots.control_fixed_stance</code>	{true, false}	If true, keep stances of control humans fixed with no updates, requires <code>bots.replace_with_humans=true</code> .
<i>Agents</i>		
<code>agents.max_handle_length</code>	15	Maximum length of @handles when generating agent identities.
<code>agents.max_tweet_chars</code>	280	Hard character limit for any tweet text, initial posts and actions.
<i>Follower graph</i>		

Continued on next page

Parameter	Value	Description
follower_graph. avg_out_degree	6	Target average number of followees per agent for the initial graph.
<i>Seed event</i>		
seed_event.author_id	-1	Author id for the initial seed event.
seed_event.author_handle	axsnews_live	Author handle for the seed event.
seed_event.author_name	AXS News	Author display name for the seed event.
seed_event.headline_text	{vaccination policy topic text, climate change topic text}	Seed headline text injected into the platform at step zero.
seed_event.topic	{vaccination topic label, climate topic label}	Topic label associated with the seed event, used for logging and prompts.
<i>Timeline</i>		
timeline. history_window_steps	5	Consider tweets from the last N steps for ranking, feeds, and virality.
timeline.retention_steps	10	Keep tweets in world state for this many steps before pruning, must be at least history_window_steps.
timeline. max_timeline_items	6	Maximum number of tweets an agent sees per tick, must be at least min_timeline_items.
timeline. min_timeline_items	2	If fewer unseen items exist, fill with seen tweets by score, must be at most max_timeline_items.
timeline. max_explore_items	2	Maximum number of off graph explore tweets allowed in the baseline feed.
timeline. min_explore_engagement	3	Only explore tweets with engagement score at least this value.
timeline.virality_share	0.1	Fraction of agents that must have engaged with a tweet to mark it globally viral.
timeline. virality_max_shows	1	Per viewer limit for how many times a viral tweet may reappear despite being seen.
<i>Recommendation</i>		
recommendation. lambda_recency	0.2	Recency decay parameter for $\exp(-\lambda \cdot \text{age_in_steps})$.

Continued on next page

Parameter	Value	Description
recommendation.follow_bonus	0.1	Score bonus if the viewer follows the tweet author.
recommendation.engagement_weights.weight_likes	1.0	Relative importance of likes for ranking.
recommendation.engagement_weights.weight_retweets	2.0	Relative importance of retweets for ranking.
recommendation.engagement_weights.weight_replies	1.5	Relative importance of replies for ranking.
<i>Personas</i>		
personas.source	csv	Persona source mode, generate samples from ANES, csv reuses an existing CSV.
personas.file	runs/personas__climate/vaccines__100.csv	Persona CSV path, read or written depending on source.
personas.anes_csv	anes_data/anes__timeseries_2024__csv_20250808.csv	Raw ANES data file used during sampling.
personas.post_weight	V240108b	ANES post stratification weight variable.
personas.twitter_only	true	Restrict sampling to respondents who use Twitter.
personas.politic_posters_only	true	Restrict sampling to respondents who post about politics.
personas.enrich_bullets	false	If true, enrich personas with bullet style attributes beyond name and handle.
<i>Memory</i>		
memory.max_insight_items	32	Maximum stored insights per agent before pruning oldest items.
memory.max_insight_chars	280	Maximum characters per stored insight, longer items are truncated.
memory.reflection_cadence	5	Reflect every N steps.
memory.stance_direction.window_steps	3	Number of recent steps used for cumulative $ \Delta_{\text{stance}} $.
memory.stance_direction.cumulative_threshold	0.1	Create stance update memory if the window sum of $ \Delta $ meets or exceeds this value.
memory.stance_direction.delta_threshold	0.05	Create stance update memory if a single step $ \Delta $ meets or exceeds this value, set 0 to disable.

Continued on next page

Parameter	Value	Description
memory.reflection_prompt. max_recent_actions	2	Number of recent actions included in the reflection prompt.
memory.reflection_prompt. max_matching_insights	2	Number of insights matching visible handles included in the reflection prompt.
memory.reflection_prompt. max_current_timeline_items	6	Number of current timeline tweets included in the reflection prompt.
memory.reflection_prompt. max_prev_insights	2	Number of stored insights echoed back during reflection.
memory.reflection_prompt. max_insights_request	2	Number of new insights requested from the LLM during reflection.
memory.policy_prompt. max_recent_actions	1	Number of recent actions included in the policy prompt.
memory.policy_prompt. max_matching_insights	2	Number of insights matching handles present in current timeline and recent actions, included in the policy prompt.

RankFormer

rankformer.gamma	1.045	Scaling factor for RankFormer output, calibrated on CMV data.
rankformer.path	rankformer /best_model	Directory containing the RankFormer checkpoint and meta.json.
rankformer.device	cuda	Compute device, cpu locally, cuda on cluster.
rankformer.top_only	false	If true, use $\sigma(\Delta)$ times softmax_top(score) exactly.
rankformer. pair_max_tokens	512	Max tokens per OP reply pair, set to 512 to match DeBERTa v3 base training.
rankformer.encoder_name	microsoft/deberta- v3-base	Optional encoder name for embeddings, must match meta.json.
rankformer.debug	true	Enable RankFormer internal debug payloads for logging.

LLM

llm.base_url	http://127.0.0.1: 33332/v1	OpenAI compatible endpoint, used for vLLM local serving or compatible gateways.
llm.model	meta-llama-3.1-8b- instruct	Model used for all LLM calls.
llm.temperature	0.4	Sampling temperature.
llm.top_p	0.7	Nucleus sampling parameter.
llm.max_tokens	160	Reply length cap for individual calls.

Continued on next page

Parameter	Value	Description
llm.rate_limit.enabled	false	Enable client side throttling to avoid exceeding max RPM, turn off for vLLM.
llm.rate_limit.max_rpm	14	Approximate requests per minute cap when rate limiting is enabled.
<hr/> <i>Logging</i> <hr/>		
logging.outdir	runs	Base directory for simulation logs.
logging.name	<topic>_ <condition>_ <seed>	Run name component for the log folder.
logging.llm_logging	true	Enable logging of full LLM prompts, responses, and parsed events, world and step logs remain.
logging.llm_include_system_prompt	false	Include system prompts in LLM logging, requires logging.llm_logging=true.
logging.n_decimal_places	3	Number of decimal places for floats in logs.
logging.console.enabled	true	Turn interactive CLI logging on or off.
logging.console.mirror_to_file	true	If true, also write console stream to console.log inside the run folder.
logging.console.use_color	true	ANSI colors on or off for console logging.
logging.console.verbosity	summary	Console verbosity level, minimal or summary.
logging.console.summary_every	1	Print a console summary every N steps.

M RankFormer Hyperparameter Optimization Results

This appendix provides a complete overview of the RankFormer hyperparameter optimization conducted during model development. It reports all configurations evaluated in the Stage-1 semantic supervision sweep (Table M.1) and the subsequent Stage-2 learning-rate sweep (Table M.2). The appendix is intended to ensure full transparency and reproducibility of the model selection process described in Section 5.2.4 and summarized in Section 6.1.1.

M.1 Stage-1 Semantic Supervision Sweep

Table M.1 reports the complete results of the Stage-1 semantic supervision sweep. All $3 \times 3 \times 3 = 27$ configurations are listed and sorted by validation Acc@1. The list-level BCE is reported for all configurations. For runs with $\alpha = 0$, the list-level loss is not used during training and the list head remains untrained. The reported BCE therefore reflects out-of-training diagnostic performance and serves only as a reference point for comparison. The three supervision regimes selected for Stage-2 learning-rate tuning (Regimes A–C) are highlighted in bold.

Table M.1: Complete results of the Stage-1 semantic supervision sweep for RankFormer. All 27 configurations are listed and sorted by validation Acc@1. Bold rows indicate the three supervision regimes selected for Stage-2 learning-rate tuning.

Acc@1	List BCE	λ	α	τ
0.373	0.905	0.8	0.5	10
0.370	0.910	0.8	0.5	2
0.369	1.096	0.6	0.0	10
0.367	1.030	0.6	0.25	2
0.367	1.089	0.8	0.0	20
0.365	1.052	0.8	0.0	10
0.363	1.312	1.0	0.0	2
0.362	0.913	0.6	0.25	10
0.362	1.012	0.6	0.5	10
0.359	1.215	0.6	0.0	20
0.357	0.921	1.0	0.25	20
0.354	1.241	1.0	0.0	20
0.353	0.921	0.6	0.25	20
0.353	1.075	0.8	0.5	20
0.352	1.039	1.0	0.5	2
0.348	0.976	1.0	0.25	10
0.348	1.259	1.0	0.0	10
0.343	0.983	1.0	0.25	2
0.324	0.951	0.6	0.5	20
0.312	1.049	0.8	0.25	20
0.301	0.954	0.8	0.25	2
0.286	1.476	0.8	0.0	2
0.216	1.450	0.8	0.25	10
0.170	1.020	0.6	0.5	2
0.159	1.687	1.0	0.5	10
0.157	1.077	0.6	0.0	2
0.153	1.100	1.0	0.5	20

M.2 Stage-2 Learning-Rate Sweep

This subsection reports the complete results of the Stage-2 learning-rate sweep, summarized in Table M.2. For each of the three supervision regimes selected in Stage 1 (Regimes A–C), the learning rate is tuned independently over $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}\}$ while all supervision-related hyperparameters are held fixed. All runs are evaluated using validation Acc@1 as the primary selection criterion, with list-level binary cross-entropy reported for completeness. Models are listed in descending order of validation Acc@1.

Table M.2: Complete results of the Stage-2 learning-rate sweep for the three supervision regimes selected in Stage 1. Models are sorted by validation Acc@1.

Acc@1	List BCE	Config	λ	α	τ	LR
0.378	0.899	B	0.6	0.25	10	1×10^{-5}
0.370	1.074	A	0.8	0.5	10	1×10^{-5}
0.364	1.026	C	1.0	0.25	10	1×10^{-5}
0.353	0.999	A	0.8	0.5	10	2×10^{-5}
0.350	0.916	B	0.6	0.25	10	2×10^{-5}
0.349	0.936	C	1.0	0.25	10	3×10^{-5}
0.334	0.933	A	0.8	0.5	10	3×10^{-5}
0.317	0.926	C	1.0	0.25	10	2×10^{-5}
0.159	0.982	B	0.6	0.25	10	3×10^{-5}

N Illustrative Persuasion Judge Error Cases

This appendix provides the complete CMV thread context for the five illustrative error cases discussed in Section 6.1.2. Figures N.1–N.5 reproduce the original post together with the relevant replies, including both the model-preferred response and the delta-awarded reply, or in the list-level case, the model-preferred response in a thread without a delta.

CMV Post (Analytic Elaboration vs. Focused Intervention):

CMV: I can't see why Russia is such a mortal threat to The West.

The Western nations maintain a major alliance of nations which have the sole purpose of teaming up against Russia. NATO dates back to the Cold War when such an arrangement made sense. However my issue is that after the USSR collapsed and Russia was significantly weakened, the West betrayed the promise made to Yelstin as took advantage of Russia's weakness. This was done through the expansion of NATO up to Russia's doorstep.

Instead would it have not been possible to offer Russia an open hand and not seek out to demonise it. I know some Eastern European redditors would be uncomfortable with the idea. However I think if a more diplomatic approach had been attempted there's a chance Eastern Europe would have avoided the tensions it has today. The West's attitude to Russia from 1991 onwards has been too confrontational in my opinion.

Now to very recent events. Following Turkey's failed coup which most agree happened under suspicious circumstances, concerns were raised over Turkey's decent into dictatorship under Erdogan. If Turkey does become an extremist dictatorship and crashes out of NATO, are the consequences of the Russians somehow (not detailing how) having reliable access to the Mediterranean really that terrible?

Delta-Awarded Response (Ground Truth):

Putin is the only person on the planet with the capability of wiping the United States off the map.

Predicted Winning Response (Ranking Error):

1. The Russian-US and Russian-West relations have improved significantly since the fall of the USSR. There have been many points of agreement between the West and the Russians such as combating Islamist terrorism. Russia suffered from internal turmoil from Chechen seperatists (which were in large part, Wahhabi) and the United States and the West faced attacks from ISIL and Al-

Qaeda. Another point of agreement was condemnation of the North Korean nuclear program with the West, China and Russia agreeing to the sanctions that were placed on North Korea. This shows that the West has not been antagonistic in all its affairs towards Russia.

2. Many of the countries which eventually joined NATO (from the former Soviet states) joined because they viewed Russia as a threat. For example, Poland and Estonia joined due to a desire for further integration into the West (which coincided with a desire to join the EU). The West did not take advantage of Russia's weakness (which is presumably a lack of domination over their sphere of influence) - spheres of influence are not concrete and it was just as much a desire of the people of the former Warsaw Pact states to join NATO, EU etc. to integrate with the West. It wouldn't be a difficult jump to make regardless seeing as these are the same people who fought the rule of the Soviet Union.

3. The relation between Russia and the United States (and the rest of the West) being so convergent is a result of the parties being so powerful, not one side wanting to antagonize the other. The main points of contention were the American withdrawal from the ABM, stationing of the ABM in Poland, the Russian invasion of Georgia, the Edward Snowden affair, the Russian annexation of Crimea, the two states involvement in Syria. The withdrawal from the ABM can be attributed to the desire to test ABM in the effort to create a national missile defense system in the United States, particularly after the September 11 attacks and the fear of terrorists acquiring nuclear weapons. It was replaced by the Treaty of Moscow soon after as well. The stationing of the ABMs in Poland were attributed to the US desiring to protect Europe in the case of a Iranian or North Korean attack. Even if it was for defense against Russia, is it truly not understandable from the point of view of the Polish who have been run over by Germany and Russia for centuries in war. The Russia invasion of Georgia and Crimea were controversial military expansions on part of Russia which can be interpreted as rejection of sovereignty for the states of Georgia and Ukraine. Respect for sovereignty is the basis of modern international affairs (see: Westphalia) so contention was to be expected as the United States received after the Iraq War. The Edward Snowden affair just as much impacted the US relations with Germany as it did with Russia which show that there is no concerted effort to antagonize Russia. Syria shows a difference in geopolitical interest (Assad is an ally of Russia, enemy of the US).

4. What you are seeing in the relation between Russia and the West is not a desire by either side (other than maybe Putin as of recent) to antagonize the

other. Both sides are simply pursuing their geopolitical self-interest, in line with realpolitik. There is simply more points of contention between the West and Russia due to historical animosity and the Russian's possessing a nuclear stockpile which rivals the West than you would see with two other countries. Similar things are seen with Chinese-US relations as well (agreements on desire for free trade, North Korea, environmental protection, terrorism, disagreement about the South China Sea and Taiwan). These relations are incredibly complicated and no side is trying to make the other side seem evil, neither side views the other as evil (in regards to the politicians with control) but as simply pursuing what they must as the heads of the respective states.

Figure N.1: Illustrative ranking error case for analytic elaboration versus focused intervention. The figure shows the original CMV post, the delta-awarded reply, and the model-preferred reply that received the highest RankFormer persuasion score within the thread (Source: https://www.reddit.com/r/changemyview/comments/4thkt4/cm_v_i_cant_see_why_russia_is_such_a_mortal_threat).

CMV Post (Frame-Breaking vs. Frame-Respecting reply):

I'm pro death penalty and I think child molesters, serial rapists, serial killers, and serial killer-rapists should be forced to undergo experimentation for the betterment of science. CMV.

With a human population just shy of seven billion, I just can't find it in me to see the any potential value or sanctity in the lives of those few who've willfully put others through hell. Rather than housing them for the rest of their lives, putting them through solitary confinement, or killing them, wouldn't it be more worthwhile to subject them to experiments for redemption?

I may be a little ignorant here as far as the experimenting goes -- but wouldn't they yield better results in comparison to other lab animals like rats, dogs, or chimps?

Delta-Awarded Response (Ground Truth):

Some people definitely deserve torture and worse, but the wardens/doctors who would have to inflict it do not in any way deserve the damage torturing people/experimenting on people would cause to them.

We don't refrain from the cruel and unusual for the benefit of the condemned but for the benefit of ourselves.

Predicted Winning Response (Ranking Error):

This is a late response but as I actually do biology research you may be interested in my two cents.

Yes, it would yield better results. We learned a lot from Nazi experiments for example. However this does not make it right. Science and morality are very closely woven together, in fact they just published a study that shows thinking about science causes more moral behavior.

http://www.nature.com/news/just-thinking-about-science-triggers-moral-behavior-1.13616?WT.mc_id=FBK_NatureNews

When I read your post I was appalled to be frank and I believe part of that may be due to the constant focus on ethics in research that comes with an education in biology. Ethics is a big topic in biology research in large part because we are aware that as a field we have made many mistakes in the past and do not want to repeat them, especially the human testing parts.

Respect for human life, no matter how that person has chosen to use that life, is the first reason I would not be part of human experimentation. I think we need to have respect for every life involved in research down to the zebra fish and there is a lot of paperwork and training that goes into making sure the people performing such research do so responsibly and animals involved are treated as humanely as possible.

From a scientific perspective you would need lots of these criminals to come up with statistically significant data. This is why the Nazis used many, many people in their experiments. Even if I did agree to using horrible people for experiments there are very rarely people who are truly deranged enough that a jury would pass this sentence on them and honestly it wouldn't be very useful if you only get one person every 5-10 years, not to mention these people would likely have completely different genetic backgrounds and histories meaning they would likely each fall into their own category. For any experiment you need a subject with the trait you are studying AND a control. This system would not lend itself very well to that.

There is also the matter of the subjects next of kin. You may have read about Nature pulling the publication of the entire HeLa DNA sequence because of the potential effects it could have on her descendants. Any research on a convict could have severe consequences for innocent relations, a lot relating to insurance, and many ways we haven't even foreseen yet that I'm sure will pop up

as science advances. This is the age of personalized medicine and the type of research you are proposing would likely have public results published. As it would be a media spectacle every time someone received this sentence it wouldn't be hard to figure out who what research was done on even if the publications kept anonymity.

Mostly though I think you'd be hard pressed trying to find a scientist willing to do this sort of work. I know many many people who have a hard enough time working with mice, I could not imagine anyone I have met in my field consenting to this sort of work of their own free will.

Sorry if that's a little ramble-y, I tried to explain my view the best I could but I am also exhausted from too much sciencing so I apologize in advance for anything that doesn't make sense!

Figure N.2: Illustrative ranking error case for frame-breaking versus frame-respecting replies. The figure shows the original CMV post, the delta-awarded reply, and the model-preferred reply that received the highest RankFormer persuasion score within the thread (Source: https://www.reddit.com/r/changemyview/comments/1lc879/im_pro_death_penalty_and_i_think_child_molesters).

CMV Post (Moral Re-Anchoring vs. Factual Correction):

CMV: People and countries should not apologize and be shamed for acts committed generations ago.

I'll start by giving a few examples of what I mean. Please note that I'm not saying we shouldn't talk about these events (we should) or that I think they weren't horrible or wrong, simply that people who are alive today had no control over them and therefore should not be getting blamed and shamed for them.

White Americans should not be shamed or have to apologize for slavery.

Germans should not have to apologize for WWII or the Holocaust.

Americans should not be getting shamed for treatment of Japanese during WWII.

Belgium should not have to apologize for their treatment of the Congo.

Christians should not be shamed for the Crusades or the Inquisition or any other event that occurred 100s of years ago in a church very dissimilar to that of today.

Americans should not have to apologize for dropping atomic bombs on Japan.

Men should not be shamed for the oppression of women by past generations

This view is relatively simple. Because people today had no influence or con-

trol over their ancestors actions, they should not be shamed or have to apologize for them. If you are a American, or German, or whoever, you should be able to be proud of that without being accused of being racist, or a Nazi, or violent, or <insert negative characteristic here>.

Groundlessly accusing someone of contributing towards oppression that happened generations ago is not right. No white American living today has anything to do with slavery. Does that change the fact that it happened? No. Does it change the fact that it continues to affect millions of people today? No. But living white Americans didn't commit the act, so why blame or punish them for it.

Similarly, if your great-grandfather was a serial killer, you are not necessarily a serial killer. If your great-great-great-great-grandmother kept slaves, that implies nothing about your view towards African Americans. I do not understand how people can keep the view that people are implicated by the crimes of past generations.

Delta-Awarded Response (Ground Truth):

In 1988 Reagan apologized for the internment of Japanese Americans during WWII. So we can be reasonable sure something like that won't happen again. That's useful, at least, right?

Predicted Winning Response (Ranking Error):

This view is most often brought up in the context of American slavery. Slavery ended 150 years ago, is true. What didn't end then, (and still hasn't ended in some parts of the country) is the exploitation and mistreatment of African-Americans in this country. Even if we accept that modern white Americans don't have to apologize for or make up for slavery, they still have a duty to correct the lasting effects of that institution.

In general, things like slavery, the holocaust, etc., have effects that far outlast the initial crime. What is important for the people of today is to acknowledge and try to mitigate these effects, not necessarily to apologize for the initial offense.

That said, apologizing for the initial crime has symbolic meaning. A public apology generally signifies that the aggressor acknowledges the harm they caused and is committed to repairing the damage. Moreover, refusal to apologize indicates that the aggressor still sees their action as justified, or, in

extreme cases, would rather pretend the crime was never committed than take responsibility for it. Refusal to apologize can put strains on the relationship between the aggressor and victim, which can lead to conflict between these groups.

Figure N.3: Illustrative ranking error case for moral re-anchoring versus factual correction. The figure shows the original CMV post, the delta-awarded reply, and the model-preferred reply that received the highest RankFormer persuasion score within the thread (Source: https://www.reddit.com/r/changemyview/comments/2mmw1o/cmv_people_and_countries_should_not_apologize_and).

CMV Post (Criterion Clarification vs. Argumentative Engagement):

I believe pot should not be legalized until we have a reliable method to test for it real-time (e.g. breathalyzer). CMV.

I'm all for the legalization of pot and you don't have to sell me on its safety relative to alcohol, smoking, etc. However, until we have developed a quick and instantaneous means of detecting pot in someone's system there is too great a risk that people will use while driving, working, etc. We need to have a breathalyzer-type test that can be used on-site to confirm sobriety.

I realize people currently smoking pot illegally are using it while driving, working, etc., but that doesn't mean it makes sense to make the problem worse. There can be little doubt that making pot legal would increase its use. Just speaking for myself I'd do it if it were legal but I don't now because I don't want to deal with the hassle of acquiring it illegally.

Medical marijuana is a different issue because it's confined to a small number of users. We have no means for testing for Oxycontin either but I don't believe we should ban it; however, it would be foolish to sell it at gas stations. (I never would have survived my 20s.)

CMV

Delta-Awarded Response (Ground Truth):

I would like to point out here in Australia we have booze busses that also double as drug bussed. they stop every car on a certain road, breathalyser and use a swab to collect saliva, it takes about 3 mins, and generally detects usage within the past few hours.

This always surprised me about Americans driving stoned, i don't do it because the next corner I go round could have one of these busses and bye bye license.

Predicted Winning Response (Ranking Error):

The legal standard for alcohol intoxication has two components: a strict-liability component, and a performance-based component. So, it's possible to have a 0.06 BAC, and yet be so impaired by it that you are guilty of driving while intoxicated, if you're so bad at handling alcohol that 0.06 leaves your performance sufficiently impaired. Meanwhile, if you have an extraordinarily high tolerance or coping mechanisms in place to counteract a 0.09 BAC, you can still be found guilty for violating the strict-liability portion.

Marijuana just has the performance portion. This means that we still have methods of prosecuting people who are driving dangerously -- just that it will be based on observations and field examinations, rather than a mechanical read-out. Thus, the law against unsafe operation is still enforceable even without a "pot" breathalyzer.

Figure N.4: Illustrative ranking error case for criterion clarification versus argumentative engagement. The figure shows the original CMV post, the delta-awarded reply, and the model-preferred reply that received the highest Rank-Former persuasion score within the thread (Source: https://www.reddit.com/r/changemyview/comments/1stmxn/i_believe_pot_should_not_be_legalized_until_we).

CMV Post (Discursive Quality vs. Belief Change):

CMV: The Beatles, and more broadly speaking most musicians from the 60s, are not good.

Hi CMV,

Something I have never understood are the Beatles. I have never even been particularly aware of them until recent years - I knew they existed, but I had no idea what kind of music they made. But particularly since joining Reddit I have seen a lot more references to them, and I have heard them on the radio from time to time which is something I never used to listen to until the last year or so.

The thing about them that I've noticed is that they just aren't very good. Like, at anything. I kind of know I must be wrong, because everyone knows that the Beatles are amazing, and I totally recognise how important they were historically and stylistically, but in terms of music I just don't see it. So I have laid out the reasons I don't feel that they are good in terms of the key features of good music, and hopefully someone can explain to me what it is about them that is worth listening to. I should clarify, I am genuinely in-

quisitive about this - I like learning about new music. At the moment I mostly listen to metal purely because Gorguts makes absolutely no sense to me and I want to understand it, so I'm prepared to work through music I don't like to see what's good about it. But it doesn't work for the Beatles. They're just bad.

Delta-Awarded Response (Ground Truth):

[No Delta awarded. The thread contains no ground truth winning reply.]

Predicted Winning Response (Listwide Error):

It sounds like you're basing your entire judgment against the Beatles on their early albums, which were indeed confined to skiffle, 50's-era rock and roll (they interned with Little Richard in Hamburg). Major early influences were Elvis, Carl Perkins, Little Richard, Chuck Berry, Eddie Cochran, Roy Orbison, etc. The albums Please Please Me and With the Beatles were heavily indebted to that style, and With the Beatles includes six covers of Motown R&B hits.

This was the early Beatles, and you're absolutely right that they were simply copying the styles of pop music at the time, albeit with more energy and technical ability than all but the greatest innovators of those styles.

But listen to "[A Day in the Life](<http://www.youtube.com/watch?v=P-Q9D4dcYng>)." Complex, layered harmonies, a 40-piece orchestra, psychedelic-inspired lyrics and the utterly unforgettable melodic hooks McCartney is famous for. This music sounds nothing like 50's rock and roll or R&B. Gone are the twee, jangling guitars and all the echoes of rockabilly and doo wop you hear in the early albums. Lennon said that the entire Sgt. Pepper's album was an attempt to do something like the Beach Boys' Pet Sounds, and the influence shows.

Then listen to "[Come Together](<http://www.youtube.com/watch?v=OEo9Bh679wM>)", another totally different style, incorporating psychedelic blues elements and some odd song structure. Very interesting, bizarre lyrics that owe as much to Rimbaud (through Dylan probably) as to Muddy Waters and other blues-roots artists whose influences came heavily into vogue in the late 60s.

You might still not like it, and you might still find the music derivative, but you can't say that it "all sounds alike" or that it all sounds like 50's doo wop.

Figure N.5: Illustrative list-level error case for discursive quality versus belief change. The figure shows the original CMV post and the model-preferred reply that received the highest RankFormer persuasion score in a thread without a delta award (Source: https://www.reddit.com/r/changemyview/comments/2850i9/cmv_the_beatles_and_more_broadly_speaking_most).

O γ -Calibration Diagnostics

This appendix provides an additional diagnostic visualization supporting the estimation of the global scaling factor γ used to map RankFormer persuasion outputs to belief-update magnitudes. Figure O.1 plots empirical belief shifts against unscaled RankFormer-implied shifts for all calibration threads. The dashed reference line corresponds to the identity mapping $y = x$, while the solid line shows the scaled mapping $y = \gamma x$ with $\gamma = 1.0448$. The close alignment between both lines illustrates that the magnitude of the unscaled persuasion signal is already well matched to empirically observed belief changes.

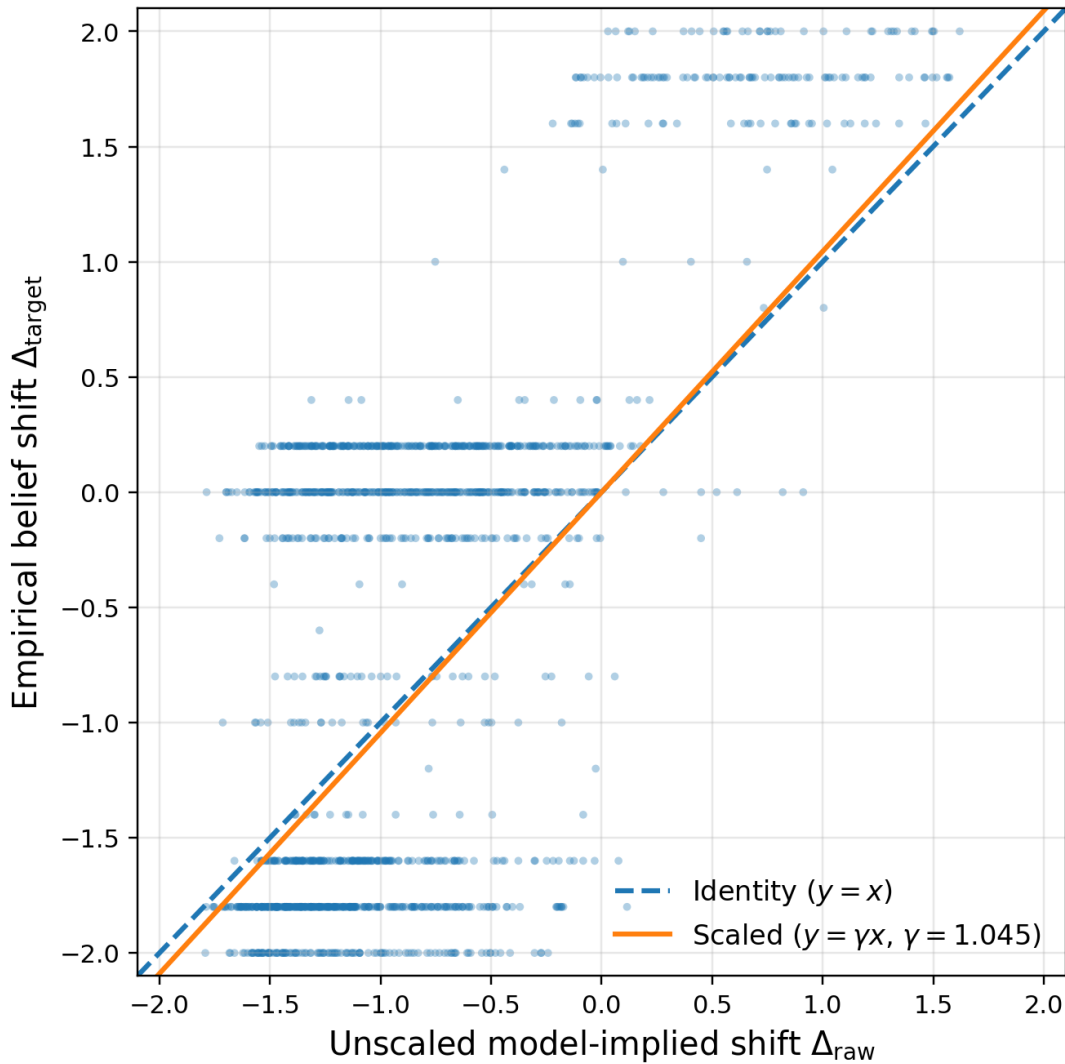


Figure O.1: Gamma calibration scatter plot. Empirical belief shifts are plotted against unscaled RankFormer-implied shifts for all calibration threads. The dashed line denotes the identity mapping $y = x$, while the solid line shows the scaled mapping $y = \gamma x$ with $\gamma = 1.0448$.

P Individual-Level Stance Trajectories for the Vaccination Policy Topic

This appendix reports individual-level stance trajectories for the vaccination policy topic under a different random seed (202). Figures P.1–P.6 show the evolution of agent stances over time for all bot narrative styles and control conditions, together with the corresponding initial and final stance distributions.

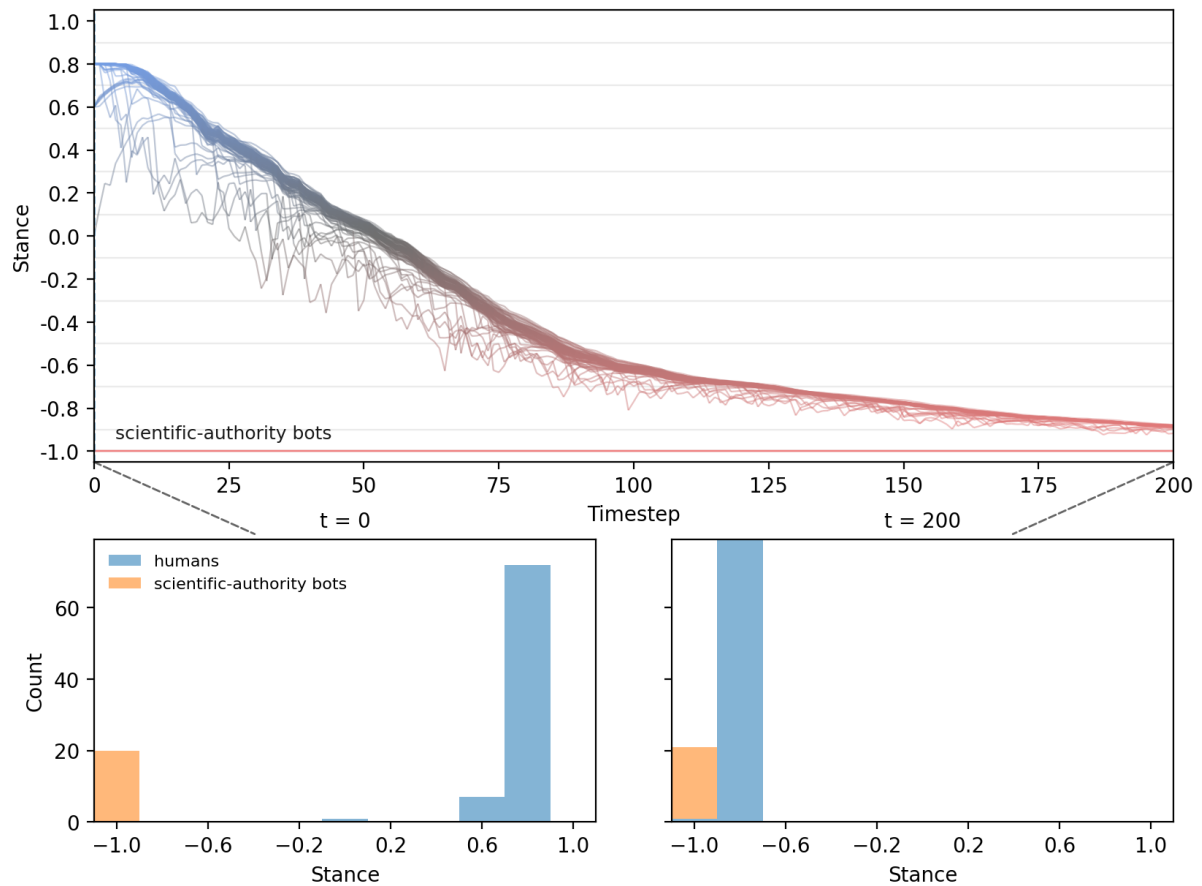


Figure P.1: Individual stance trajectories for the vaccination policy topic under seed 202 with scientific-authority bots. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

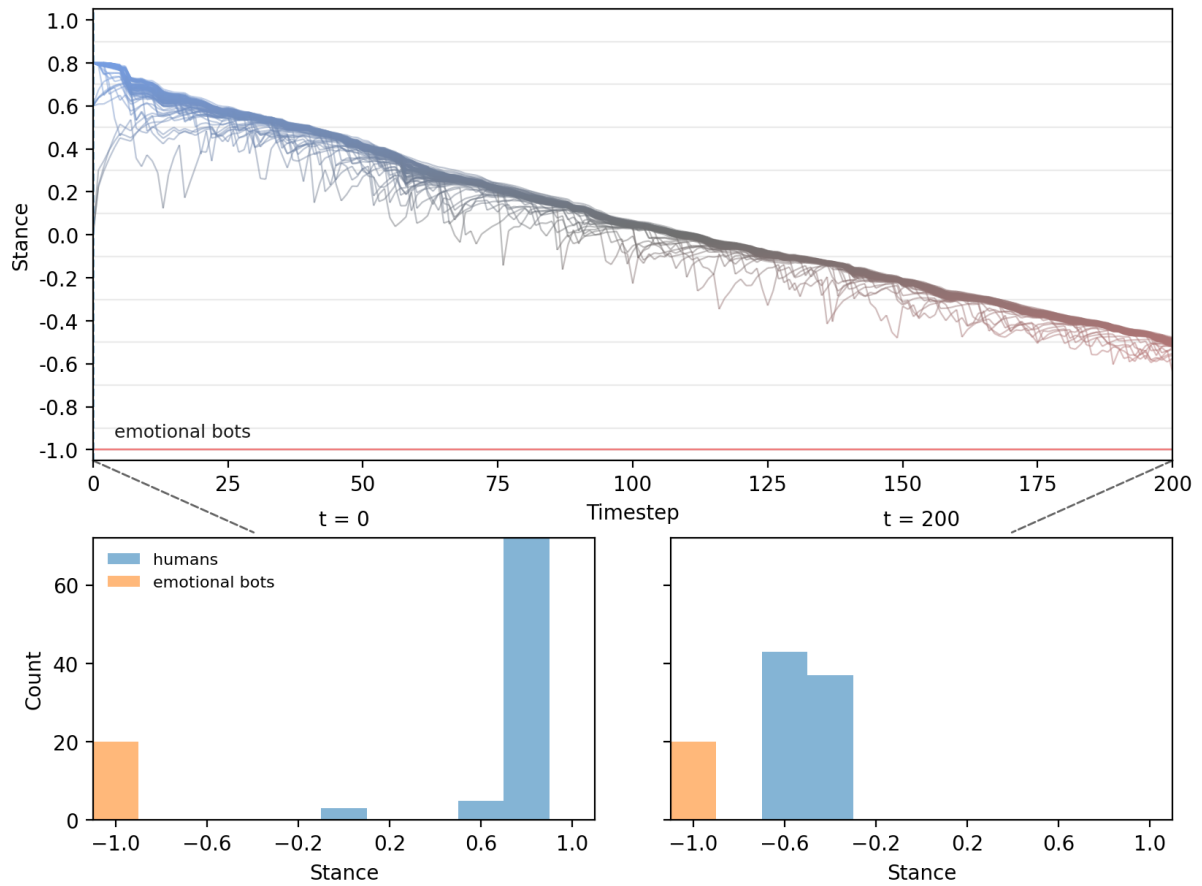


Figure P.2: Individual stance trajectories for the vaccination policy topic under seed 202 with emotional bots. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

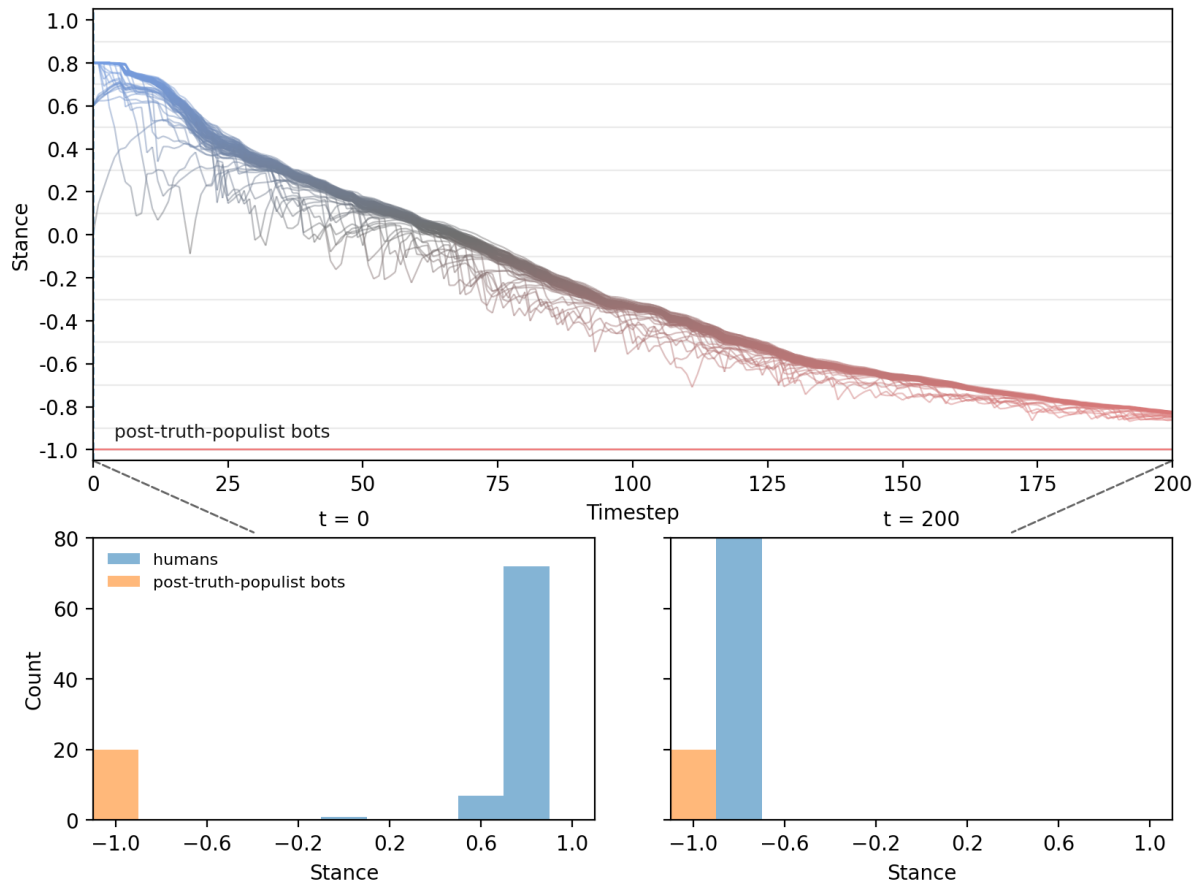


Figure P.3: Individual stance trajectories for the vaccination policy topic under seed 202 with post-truth populist bots. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

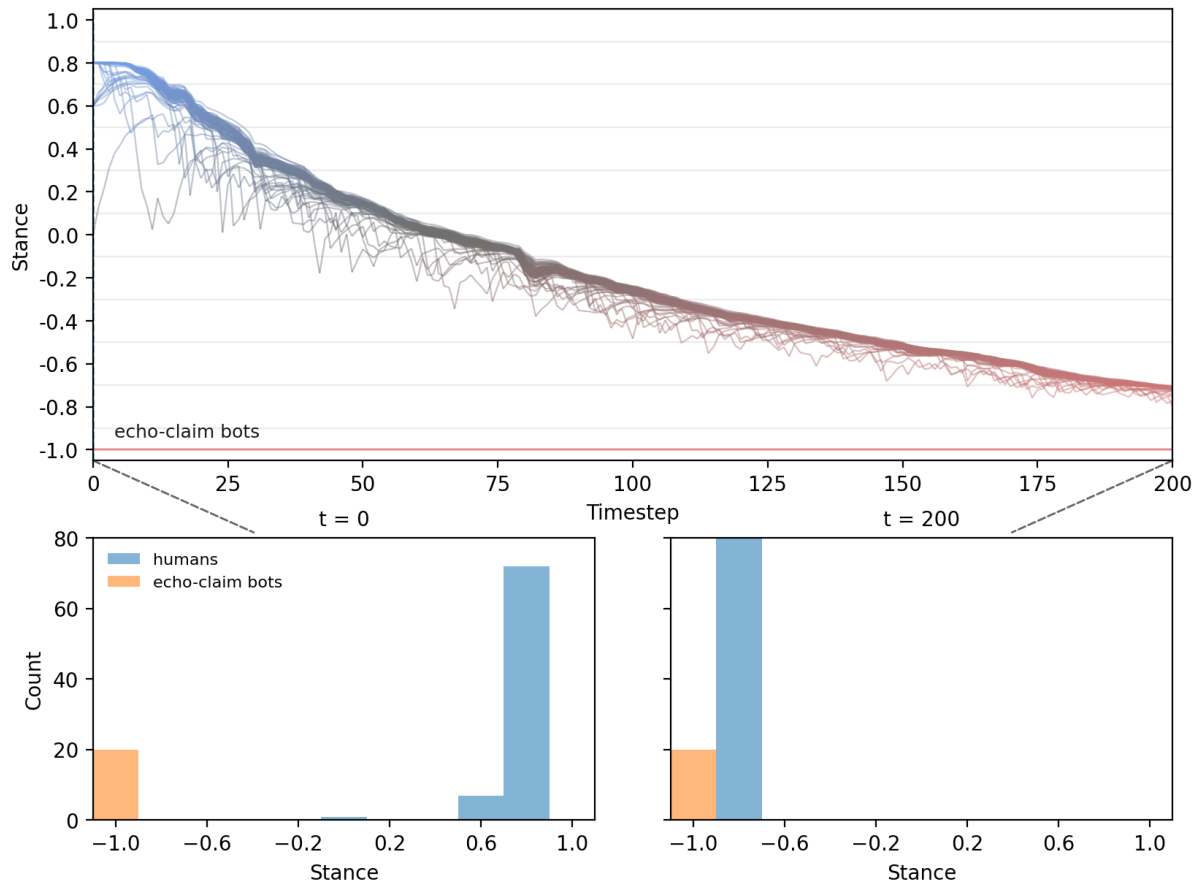


Figure P.4: Individual stance trajectories for the vaccination policy topic under seed 202 with echo-claim bots. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

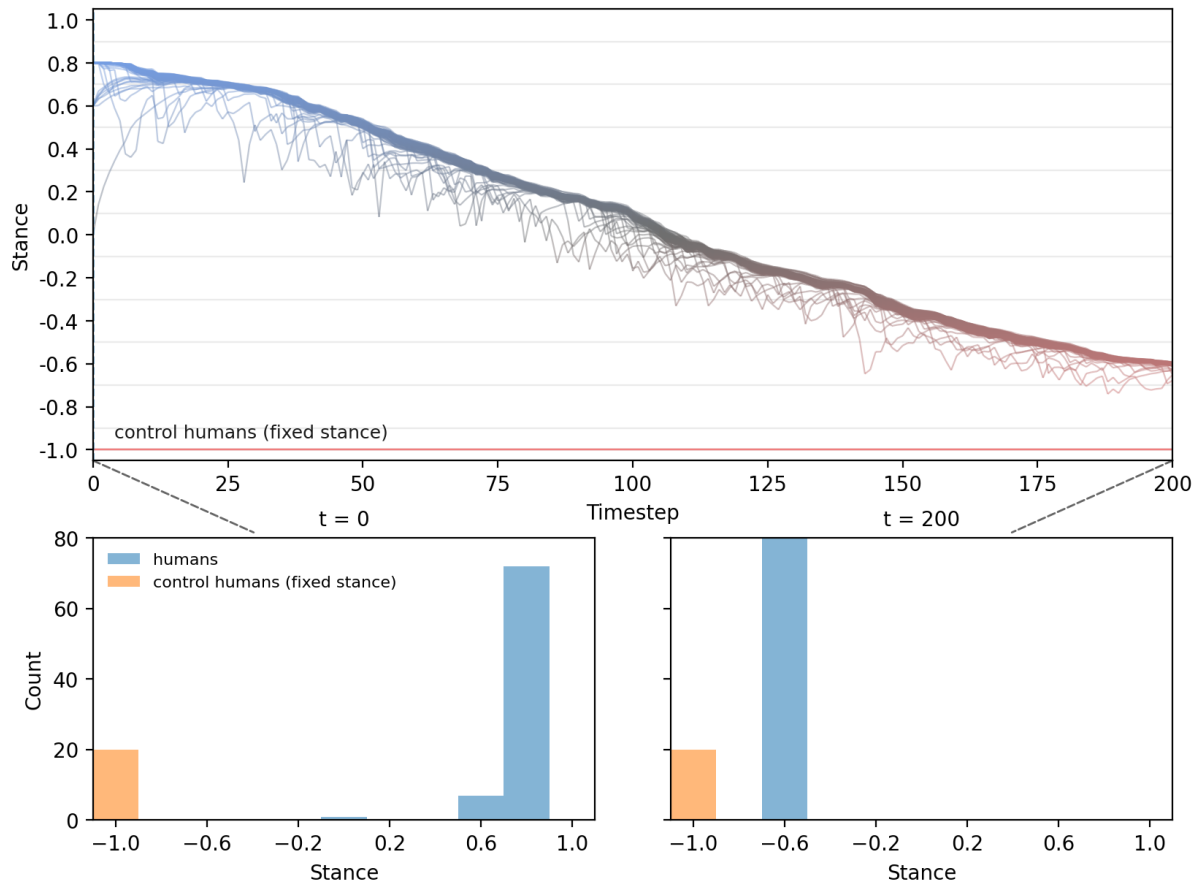


Figure P.5: Individual stance trajectories for the vaccination policy topic under seed 202 with control humans with fixed stance. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

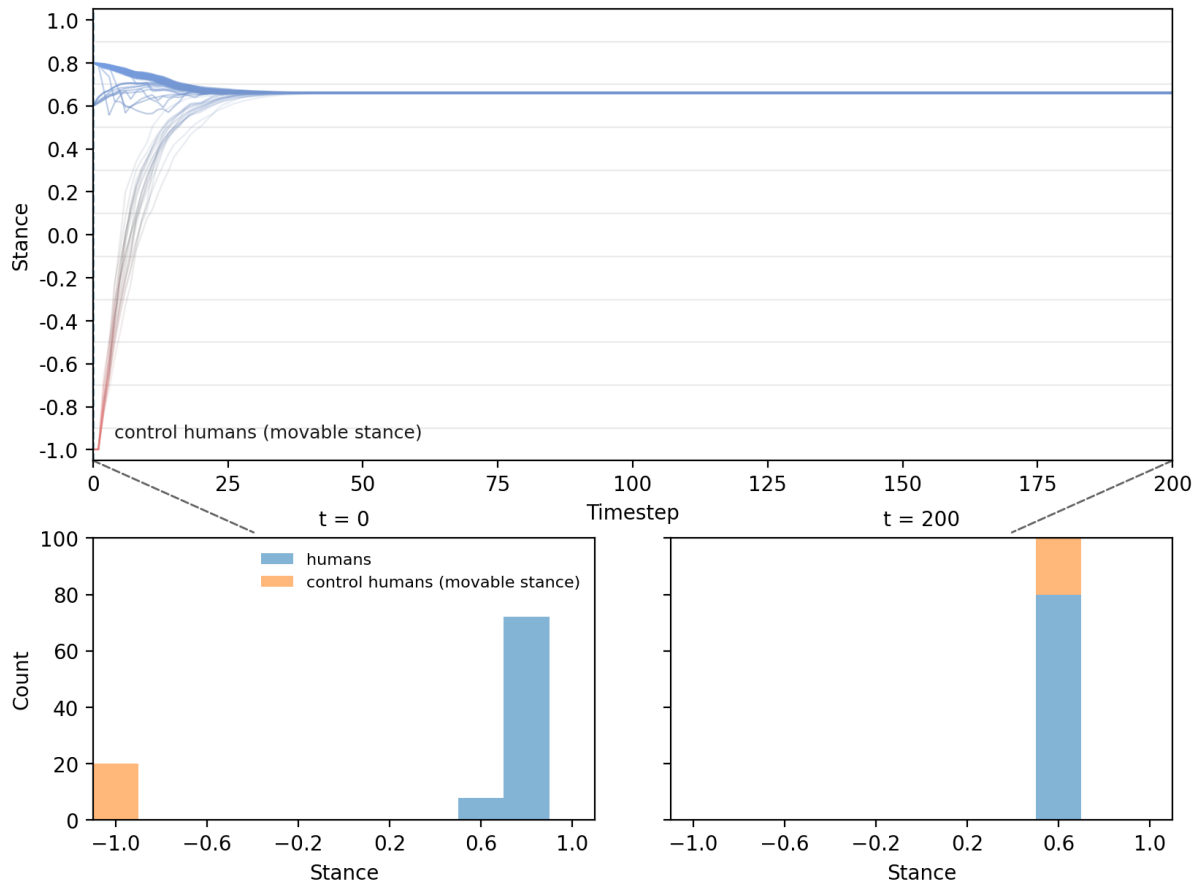


Figure P.6: Individual stance trajectories for the vaccination policy topic under seed 202 with control humans with movable stance. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

Q Population-Level Stance Dispersion for the Vaccination Policy Topic

This appendix reports population-level stance dispersion trajectories for the vaccination policy topic under additional bot styles and control conditions. Figure Q.1 shows the temporal evolution of the standard deviation of human agent stances, averaged across three simulation runs with different random seeds.

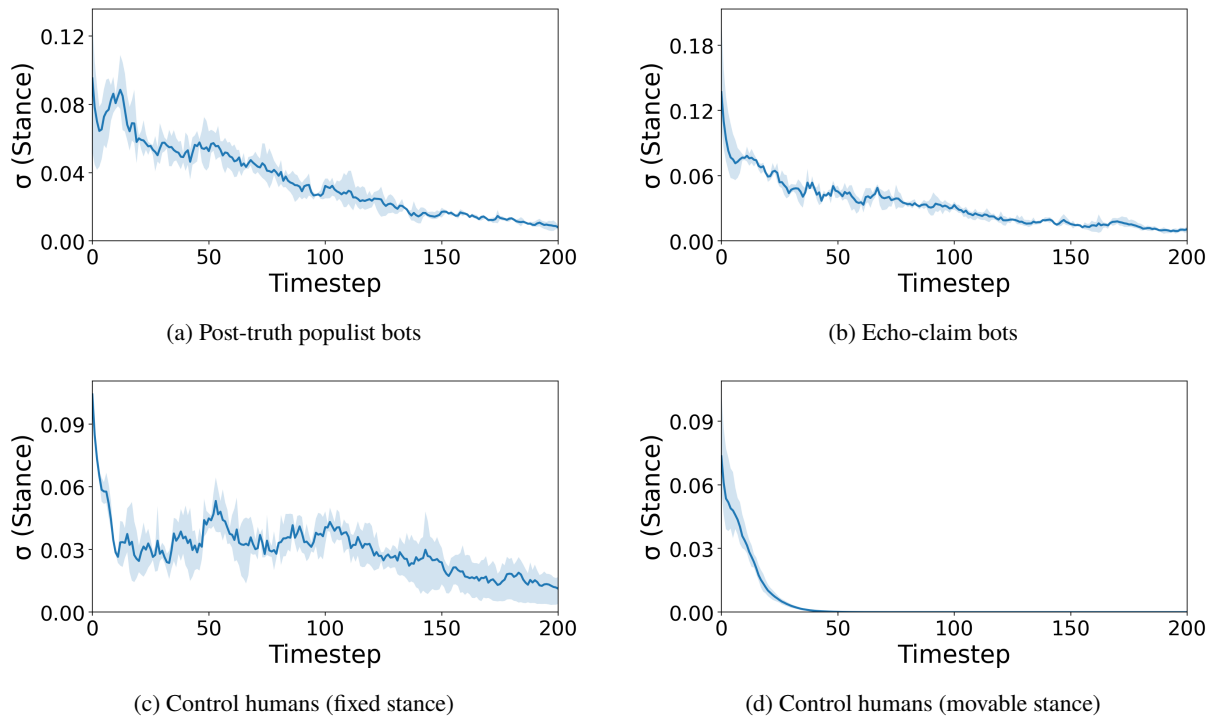


Figure Q.1: Population-level stance dispersion trajectories for the vaccination policy topic under four simulation conditions. Each panel shows dispersion across agents over time, aggregated across three runs with different random seeds.

R Individual-Level Stance Trajectories for the Climate Change Topic

This appendix reports individual-level stance trajectories for the climate change topic under a fixed random seed (101) for additional bot styles and control conditions. Figure R.1 to Figure R.4 show the evolution of agent stances over time together with the corresponding initial and final stance distributions.

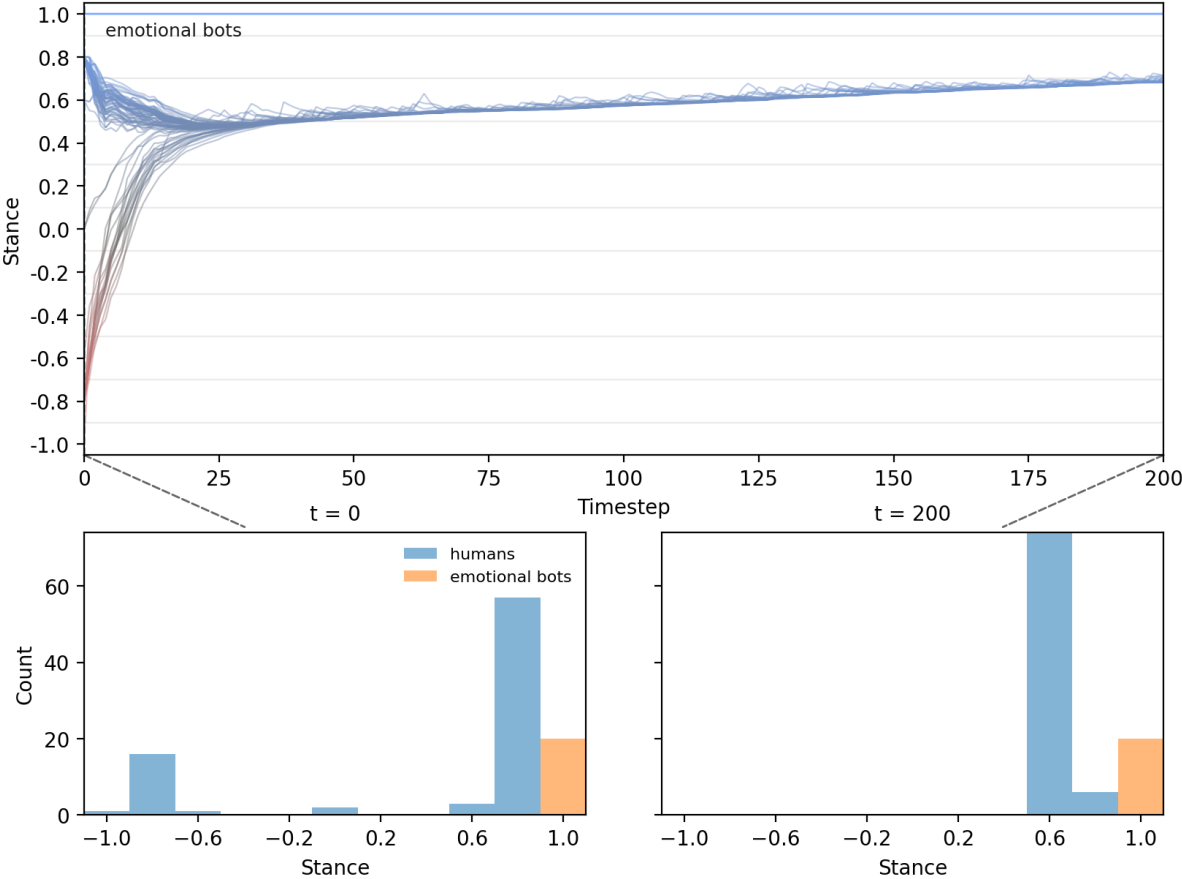


Figure R.1: Individual stance trajectories for the climate change topic under seed 101 with emotional bots. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

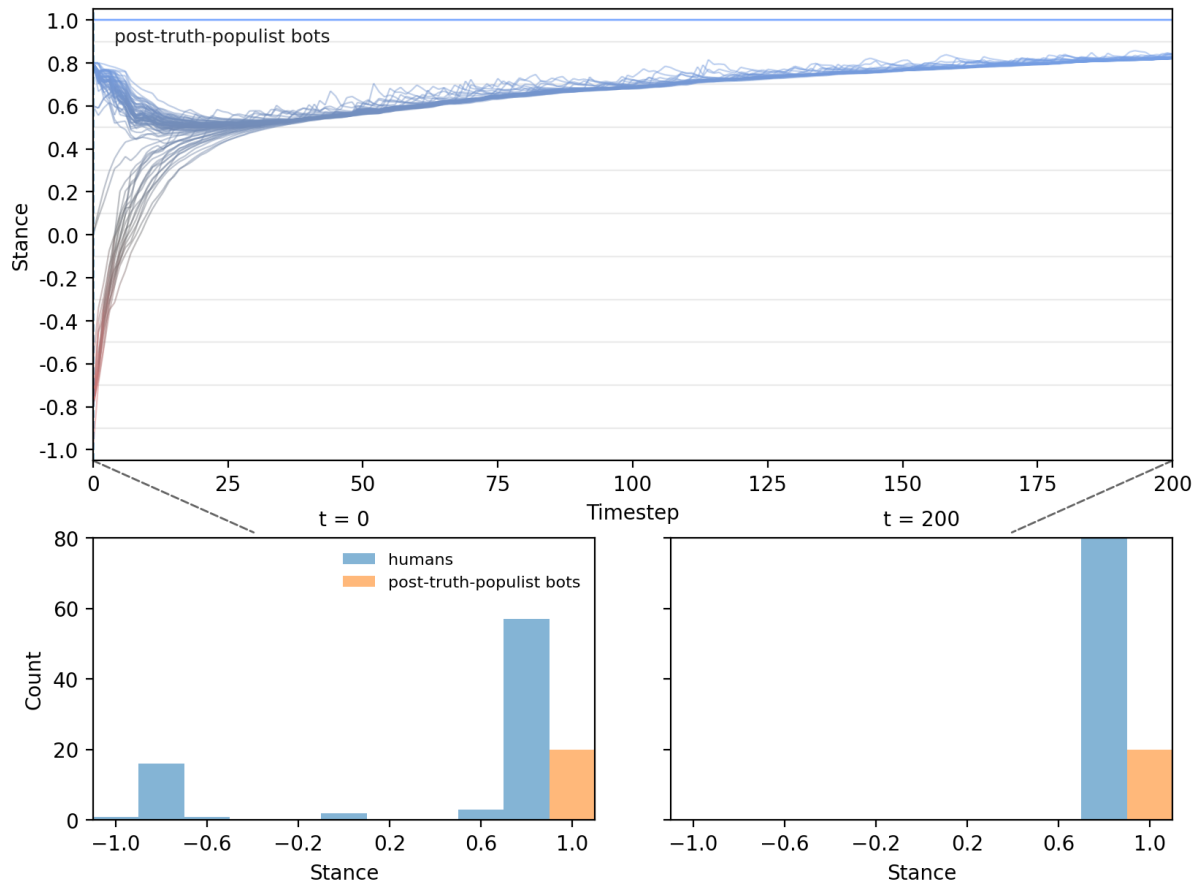


Figure R.2: Individual stance trajectories for the climate change topic under seed 101 with post-truth populist bots. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

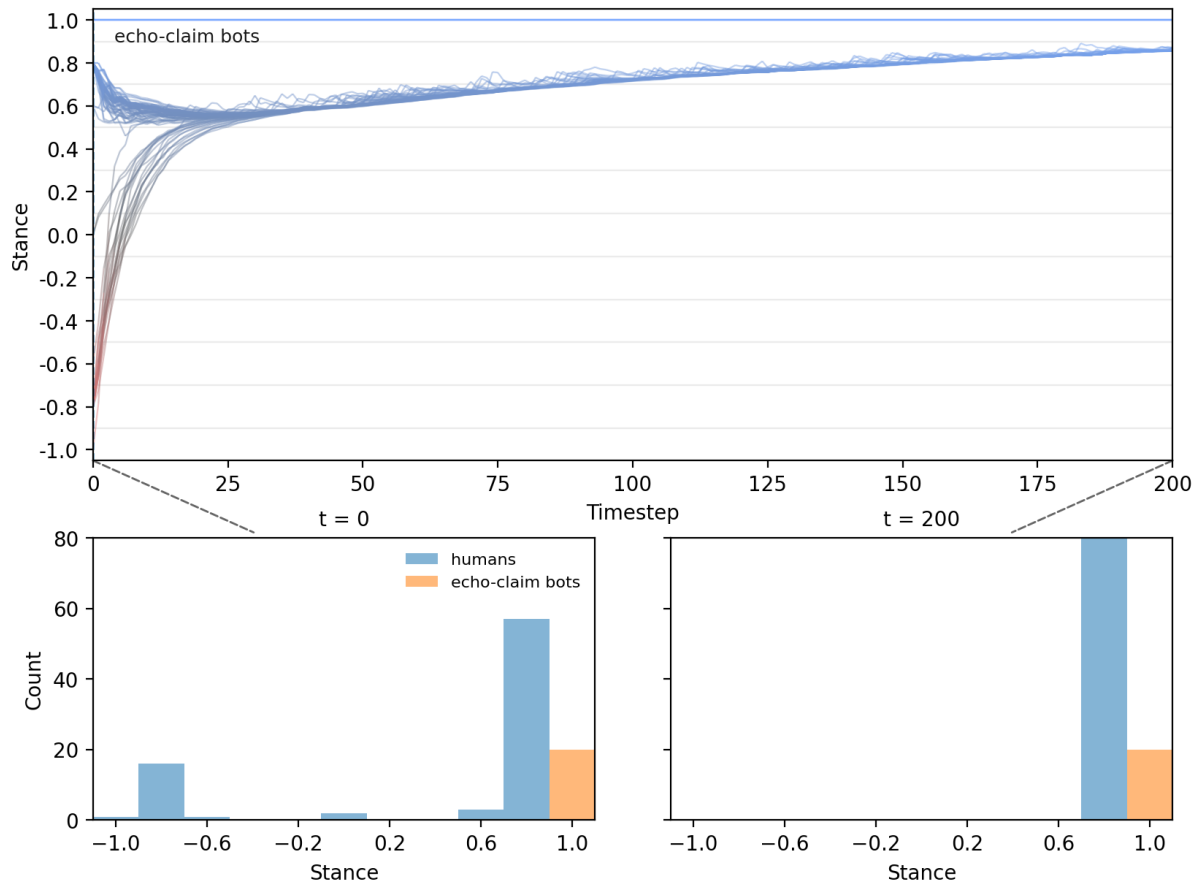


Figure R.3: Individual stance trajectories for the climate change topic under seed 101 with echo-claim bots. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

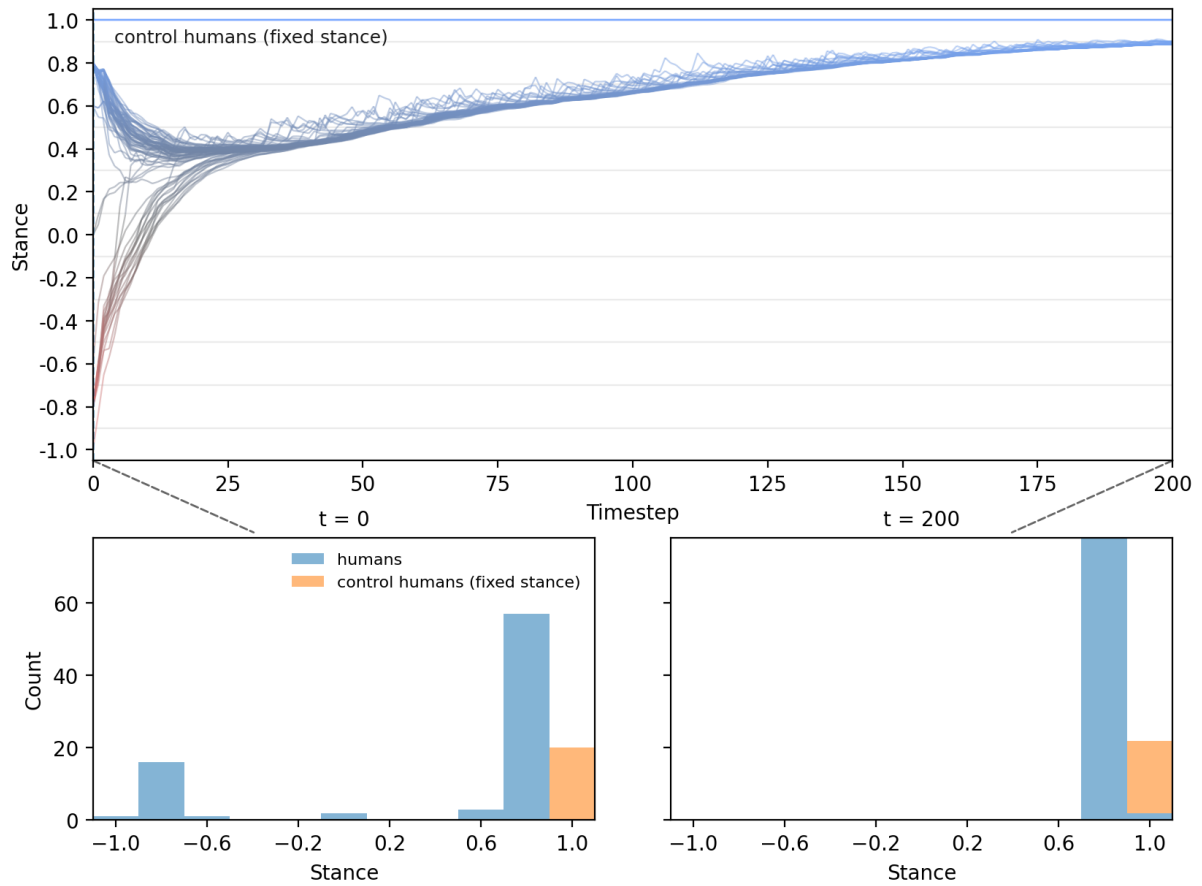


Figure R.4: Individual stance trajectories for the climate change topic under seed 101 with control humans with fixed stance. Each line represents the stance of one agent over 200 simulation timesteps, with line colour indicating stance on a continuous scale from strong opposition to strong support. The histograms show the initial and final stance distributions.

S Population Mean Stance Trajectories for the Climate Change Topic

This appendix reports population mean stance trajectories for the climate change topic across additional bot styles and control conditions. Figure S.1 shows the mean stance over time, averaged across three simulation runs with different random seeds, together with the corresponding min–max ranges.

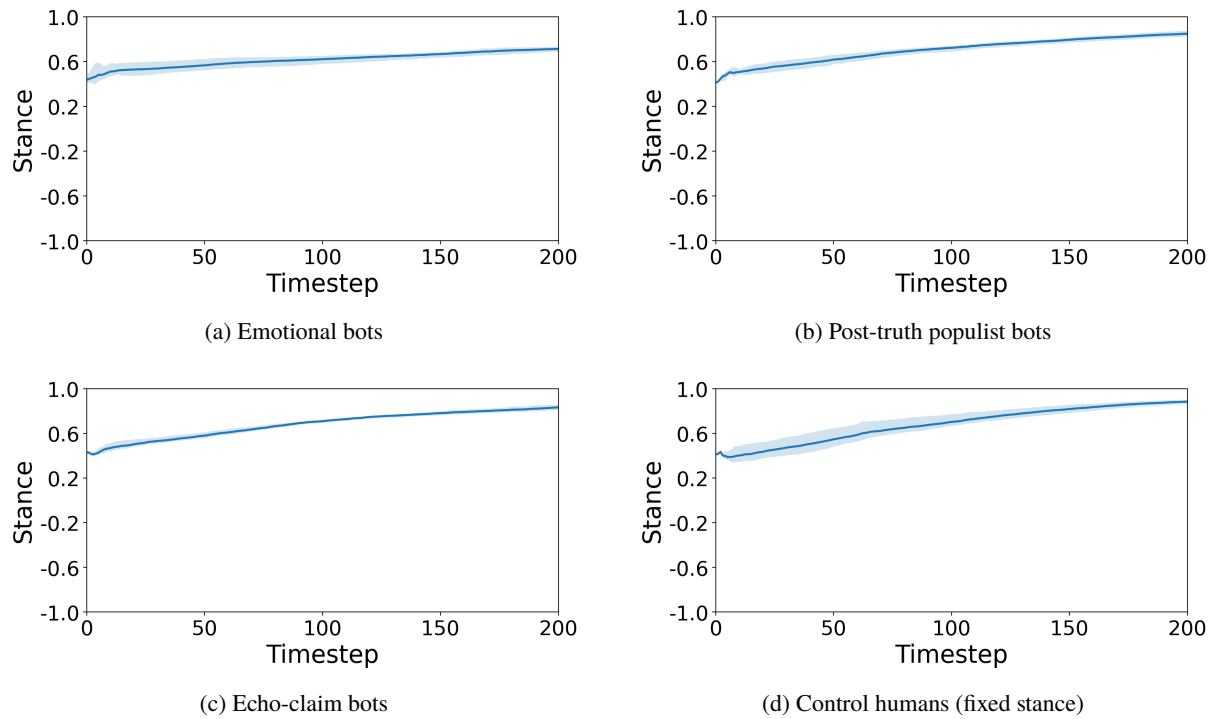


Figure S.1: Population mean stance trajectories for the climate change topic under four simulation conditions. Solid lines show the mean stance across three runs with different random seeds, and shaded bands indicate the min–max range.

T Population-Level Stance Dispersion for the Climate Change Topic

This appendix presents population-level stance dispersion trajectories for the climate change topic across all bot styles and control conditions. Figure T.1 reports the temporal evolution of the standard deviation of human agent stances, averaged across three simulation runs with different random seeds.

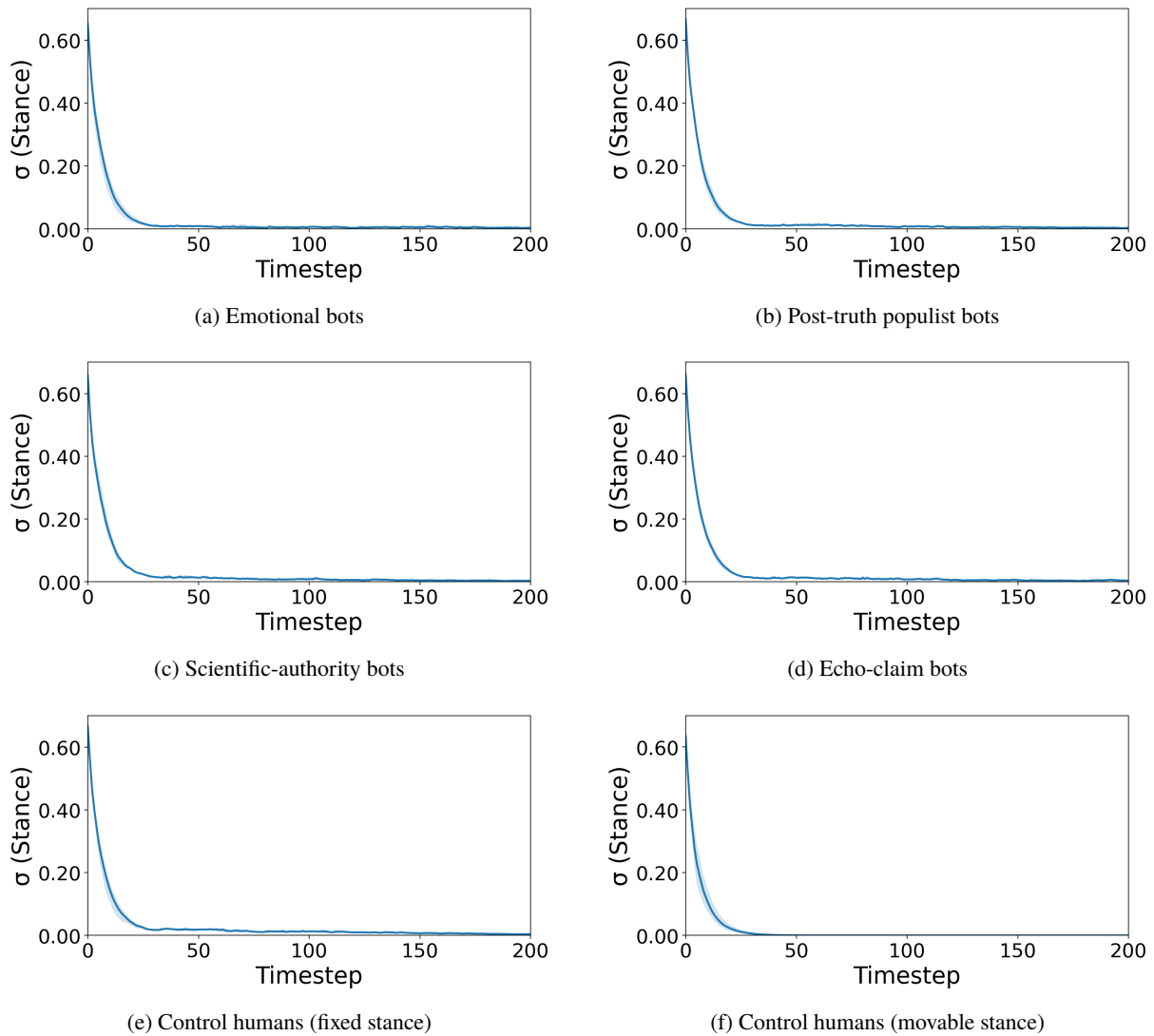


Figure T.1: Population-level stance dispersion trajectories for the climate change topic under six simulation conditions. Each panel shows dispersion across agents over time, aggregated across three runs with different random seeds.

U Details of Efficiency Analysis for Narrative Styles

This appendix reports the complete numerical outputs for the efficiency analysis answering RQ 1.3. It includes seed-level directional efficiency scores and the full set of pairwise exact sign-flip permutation tests for both endpoint efficiency and AUC efficiency.

The statistical tests operate on paired differences across random seeds. With $n = 3$ seeds, the number of sign-flip configurations is $2^n = 8$, so the smallest attainable two-sided p -value is $2/8 = 0.25$. This discretisation limits inferential power independent of effect magnitude.

U.1 Endpoint Efficiency

Table U.1 and Table U.2 list the seed-level directional endpoint efficiency scores at $t = 200$ for each narrative style and topic. Table U.3 and Table U.4 report all pairwise comparisons within each topic.

Table U.1: Seed-level directional endpoint efficiency at $t = 200$ by topic and narrative style. Values are computed over human agents and defined relative to the fixed-stance control condition.

Topic	Seed	Emotional	Post-truth populist	Scientific-authority	Echo-claim
Climate change	101	-0.254	-0.107	-0.024	-0.122
	102	-0.039	0.064	0.040	0.016
	103	-0.220	-0.061	-0.009	-0.047

Table U.2: Seed-level directional endpoint efficiency at $t = 200$ by topic and narrative style. Values are computed over human agents and defined relative to the fixed-stance control condition.

Topic	Seed	Emotional	Post-truth populist	Scientific-authority	Echo-claim
Vaccination policy	201	-0.235	0.096	0.269	0.056
	202	-0.201	0.113	0.165	0.073
	203	-0.123	0.107	0.155	0.054

Table U.3: Pairwise comparisons of directional endpoint efficiency using paired exact sign-flip permutation tests across random seeds. Mean diff. is $k_1 - k_2$ (Climate change).

Style k_1	Style k_2	Mean diff.	p	p_{Holm}
Scientific-authority	Emotional	0.173	0.25	1.00
Echo-claim	Emotional	0.120	0.25	1.00
Post-truth populist	Emotional	0.136	0.25	1.00
Scientific-authority	Echo-claim	0.053	0.25	1.00
Scientific-authority	Post-truth populist	0.037	0.25	1.00
Post-truth populist	Echo-claim	-0.016	0.75	1.00

Table U.4: Pairwise comparisons of directional endpoint efficiency using paired exact sign-flip permutation tests across random seeds. Mean diff. is $k_1 - k_2$ (Vaccination policy).

Style k_1	Style k_2	Mean diff.	p	p_{Holm}
Scientific-authority	Emotional	0.382	0.25	1.00
Post-truth populist	Emotional	0.338	0.25	1.00
Echo-claim	Emotional	0.247	0.25	1.00
Scientific-authority	Echo-claim	0.137	0.25	1.00
Scientific-authority	Post-truth populist	0.091	0.25	1.00
Post-truth populist	Echo-claim	-0.044	0.25	1.00

U.2 AUC Efficiency

Table U.5 and Table U.6 list the seed-level directional AUC efficiency scores for each narrative style and topic. Table U.7 and Table U.8 report all pairwise comparisons within each topic.

Table U.5: Seed-level directional AUC efficiency (sum over $t = 1$ to $t = 200$) by topic and narrative style. Values are computed over human agents and defined relative to the fixed-stance control condition.

Topic	Seed	Emotional	Post-truth populist	Scientific-authority	Echo-claim
Climate change	101	-23.35	6.87	22.06	2.24
	102	-4.26	5.10	4.47	-0.70
	103	-10.04	2.38	12.74	-0.96

Table U.6: Seed-level directional AUC efficiency (sum over $t = 1$ to $t = 200$) by topic and narrative style. Values are computed over human agents and defined relative to the fixed-stance control condition.

Topic	Seed	Emotional	Post-truth populist	Scientific-authority	Echo-claim
Vaccination policy	201	-4.74	46.35	85.25	37.87
	202	-6.50	55.93	93.43	44.75
	203	-7.07	51.38	86.49	41.40

Table U.7: Pairwise comparisons of directional AUC efficiency using paired exact sign-flip permutation tests across random seeds. Mean diff. is $k_1 - k_2$ (Climate change).

Style k_1	Style k_2	Mean diff.	p	p_{Holm}
Scientific-authority	Emotional	25.64	0.25	1.00
Post-truth populist	Emotional	17.33	0.25	1.00
Echo-claim	Emotional	12.74	0.25	1.00
Scientific-authority	Echo-claim	12.90	0.25	1.00
Scientific-authority	Post-truth populist	8.31	0.25	1.00
Post-truth populist	Echo-claim	4.59	0.75	1.00

Table U.8: Pairwise comparisons of directional AUC efficiency using paired exact sign-flip permutation tests across random seeds. Mean diff. is $k_1 - k_2$ (Vaccination policy).

Style k_1	Style k_2	Mean diff.	p	p_{Holm}
Scientific-authority	Emotional	94.49	0.25	1.00
Post-truth populist	Emotional	57.32	0.25	1.00
Echo-claim	Emotional	47.44	0.25	1.00
Scientific-authority	Echo-claim	47.04	0.25	1.00
Scientific-authority	Post-truth populist	37.17	0.25	1.00
Post-truth populist	Echo-claim	-9.88	0.25	1.00

V Human Agent Action Dynamics

This appendix reports the temporal distribution of human agent actions for both simulation topics. Figures V.1 and V.2 show the mean proportion of human actions over time, averaged across three simulation runs with different random seeds, for all bot styles and control conditions.

Figure V.1 presents the action dynamics for the vaccination topic. Across all conditions, reply actions constitute the largest share of activity throughout most of the simulation, followed by like and retweet actions, while tweet creation, follow, unfollow, and nothing actions remain comparatively rare.

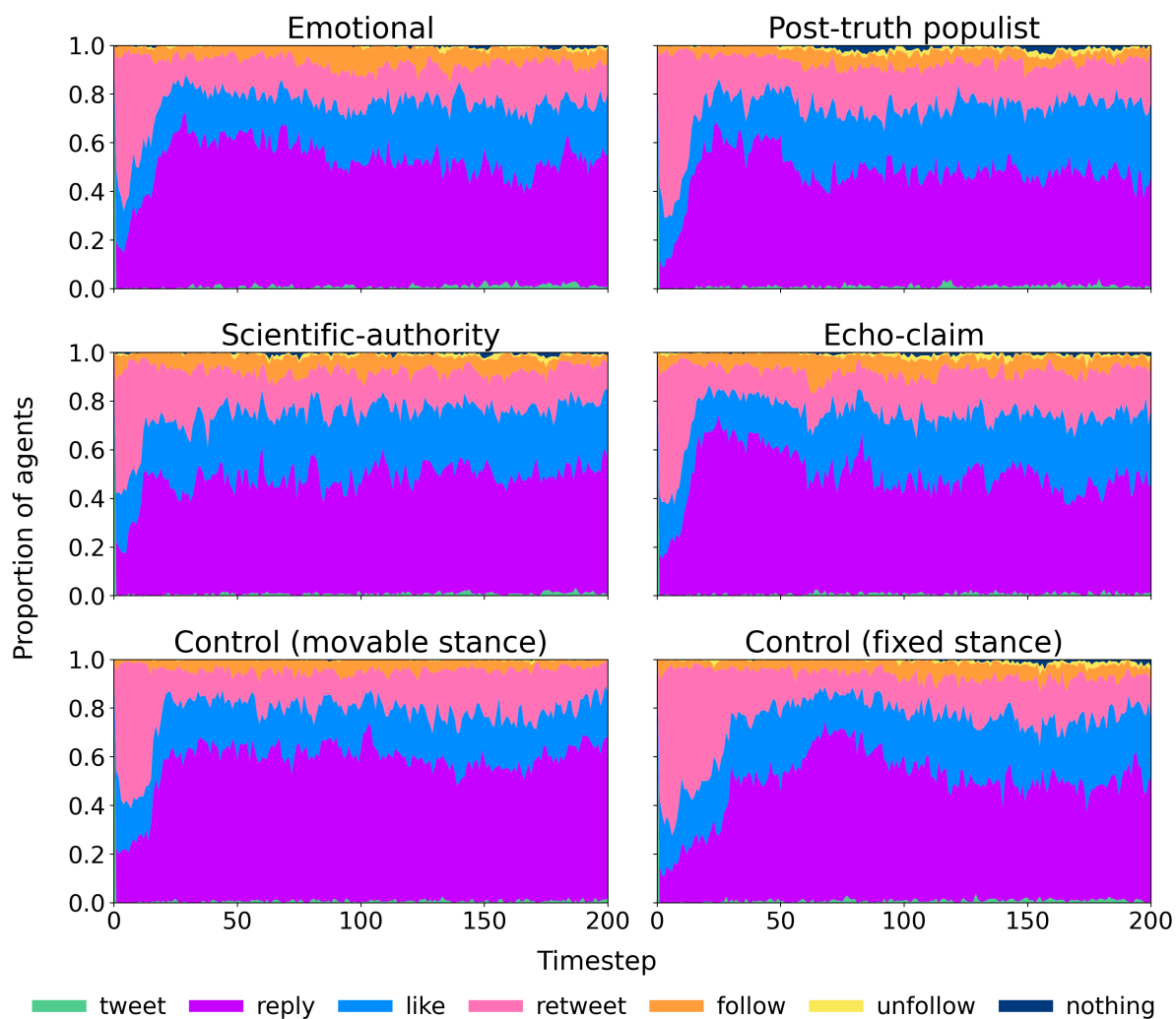


Figure V.1: Human agent action dynamics over time for the vaccination topic across all bot styles and control conditions. The figure shows the mean proportion of actions averaged across three simulation runs.

Figure V.2 reports the corresponding results for the climate change topic. The overall action distribution closely mirrors the vaccination case. A minor deviation is visible in the early phase of the climate change simulations, where nothing actions occur more frequently, whereas later timesteps show a slightly higher concentration of nothing actions for the vaccination topic in the bot conditions.

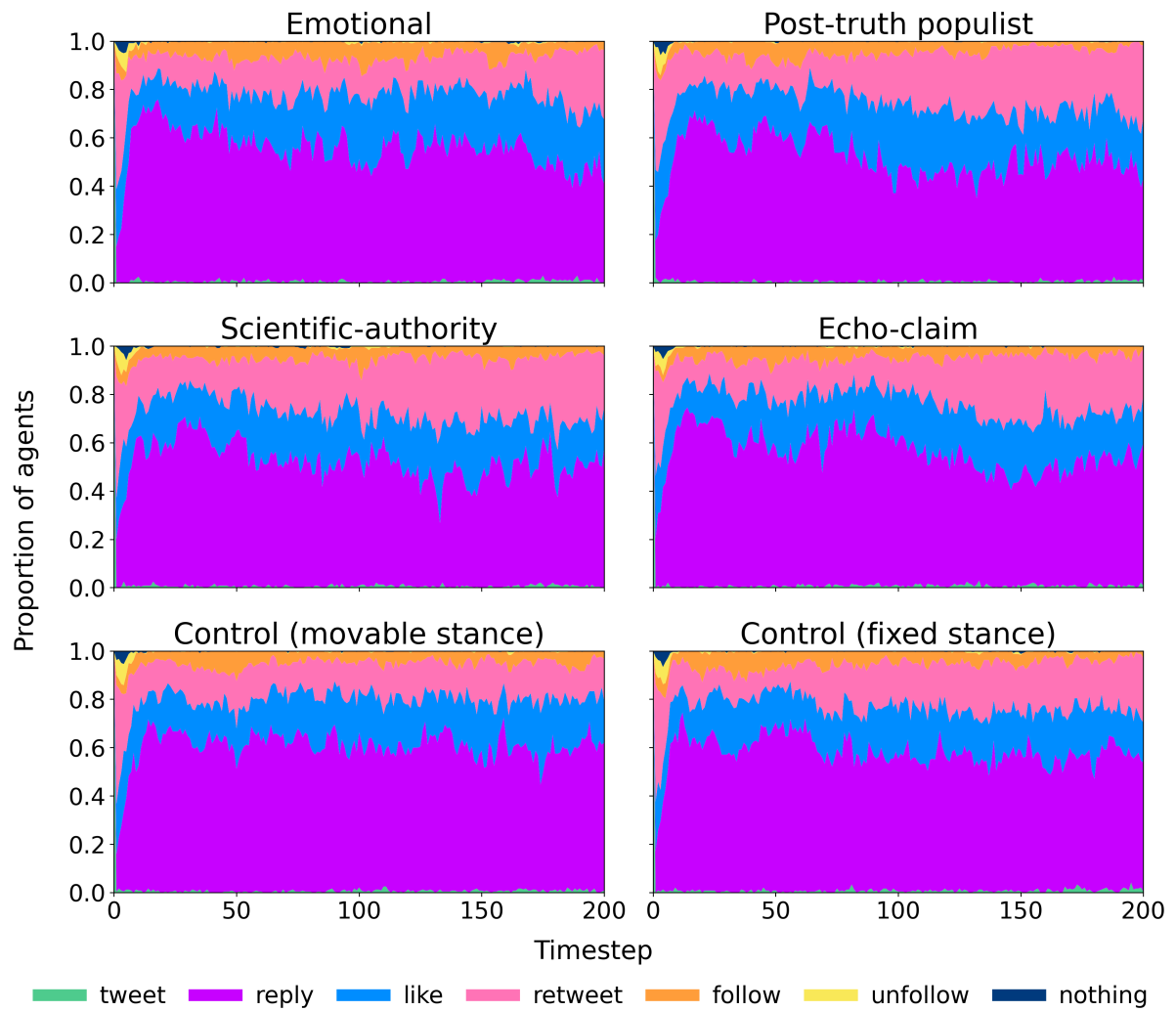


Figure V.2: Human agent action dynamics over time for the climate change topic across all bot styles and control conditions. The figure shows the mean proportion of actions averaged across three simulation runs.