Green Bots versus Red Bots: Evaluating Large Language Models for Simulating Persuasion Dynamics in Online Influence Campaigns

Majd Al Ali[1,2], Filip Muntean[1,2], Lucia Donatelli[1], Jurriaan van Diggelen[2]
[1]Vrije Universiteit Amsterdam, [2]TNO
De Boelelaan 1105, Amsterdam
Kampweg 55, Soesterberg
{filip.mihai.muntean, majd.alali.eos}@gmail.com, l.e.donatelli@vu.nl, jurriaan.vandiggelen@tno.nl

Abstract

Large language models (LLMs) are increasingly used to simulate social interaction and persuasion dynamics, yet their validity as proxies for human cognition and behavior remains unverified. We propose a dual-level evaluation framework to assess LLM-based agents at both the individual and collective levels. At the individual level, we examine agent fidelity by comparing LLM-generated political personas to human benchmark data. We find that while agents capture broad partisan orientations, they underestimate within-group variability and reproduce stereotypical ideological biases. At the collective level, we deploy Big Five personality-differentiated agents in 1080 structured dialogues to test the effect of rhetorical strategy on persuasive success. Our simulations reproduce theoretically expected interaction patterns; nevertheless, belief shifts are exaggerated relative to human baselines, supporting LLMs' tendency toward over-responsiveness. These findings suggest a trade-off between engagement-optimized training objectives and psychological realism, confirming the need to use LLMs with caution to simulate human behavior. We contribute three resources: a persuasion dynamics dataset, a standardized agent taxonomy of "red" and "green" bots, and a framework for evaluating both individual-agent fidelity and emergent group-level behavior.

Keywords: Large Language Models, Social Simulation, Agent-Based Simulation, Misinformation, Persona, Trust Dynamics

1. Introduction

Hostile influence campaigns on social media present serious risks to democratic societies. These campaigns can include coordinated disinformation about elections, public health issues, or international conflicts. As generative AI in the form of large language models (LLMs) becomes more advanced, these risks are expected to increase. LLMs can now be used to create conversational agents that convincingly act like real people and send personalised messages to individuals. These agents can spread false information, manipulate public opinion, and disrupt democratic processes while appearing to be genuine users.

To better understand both the impact of and the defense against current and future LLM-driven influence campaigns, there is a growing need for a safe and controlled experimental environment (Chuang et al., 2024; Wang et al., 2025a). Ideally, such an environment would not rely on human test subjects, thereby avoiding ethical concerns associated with exposing individuals to manipulative scenarios. Moreover, it would enable scalable and long-term longitudinal studies that are difficult to conduct with human participants. Within such an environment, the effects of these campaigns can be explored using scenarios and assumptions about future AI developments, allowing us to assess risks and develop effective mitigation strategies.

There has been significant progress in creating simulation environments that capture the dynamics of disinformation campaigns at the collective level, such as tracking how messages spread among large numbers of social media users and analyzing the development of filter bubbles (Papadopoulou et al., 2022). However, there is still a need for methods that can simulate these dynamics at the individual level, focusing on individual effects, and at the collective level, focusing on small group interactions. As a result, while we can already study broad trends and patterns(Bonabeau, 2002; Zhang et al., 2025), there is a growing need to develop tools that closely examine how disinformation impacts single users or small groups in detail.

To address this gap, we are developing an environment called the *virtual trollfarm*. In this environment, Red Bots are advanced AI agents designed to spread disinformation and develop highly effective persuasion tactics for specific audiences. Green Bots simulate ordinary citizens who are vulnerable to disinformation, modeling how people might change their opinions on social media. Blue Bots act as digital defenders, using AI to detect disinformation, provide alternative information, and help people build digital resilience.

While the design of Red, Green, and Blue Bots each presents unique challenges, this paper focuses on the development of Green Bots. Specifically, *the purpose of this study is to investigate whether large language model (LLM) agents can serve as effective proxies for human behavior in the context of studying disinformation campaigns*.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11152–11171
11–16 May 2026.

To model realistic Green Bot responses, we draw on the Elaboration Likelihood Model (ELM, (Petty and Cacioppo, 1986)) and Motivated Reasoning Theory (Kunda, 1990), which explain how attitudes and motivations shape information processing. Based on these frameworks, we identify three aspects as an initial foundation: a belief system that captures prior attitudes and biases, personality types that influence susceptibility and interaction style, and an action set that defines observable behaviors such as sharing or rejecting content. These dimensions do not fully capture the complexity of human cognition, but they provide a practical starting point, both for designing LLM prompts for Green Bots and for addressing our research question on whether such agents can sufficiently realistically simulate opinion dynamics in disinformation scenarios.

In practical terms, we conduct two experiments. The first focuses on evaluating the fidelity of individual LLM agents in modeling belief systems. We simulate agents representing Democratic and Republican political personas and assess whether their responses to political questions aligned with those of real individuals in the U.S. political system. The second experiment explores how personality traits influence agent behavior in interactive persuasion scenarios. Together, these experiments address two complementary levels of analysis: the individual-level validity of individual agent instantiations and the collective-level dynamics that emerge from interactions among agents.

The structure of the paper is as follows. Section 2 reviews related work. Section 3 provides an overview of our dual-experiment design, while sections 4 and 5 present our two experiments along with their results. Section 6 offers a discussion of the findings and outlines limitations of our approach. Finally, Section 7 concludes the paper.

2. Related Work

Language Models for Social Simulations Modeling human behaviour has a long history: computational work has shown that simple individual-level interaction rules can produce rich collective-level patterns (Schelling, 1971; Hegselmann and Krause, 2002; Deffuant et al., 2000). Recent advances have explored the use of LLMs as components within social simulations. When conditioned with demographic or personality information, LLMs can approximate subgroup-level responses and exhibit contextually coherent behavior (Argyle et al., 2023; Park et al., 2023b). This opens new possibilities for simulating human-like interactions through natural language rather than abstract rule sets. However, because these models often rely on linguistic shortcuts and pattern matching rather than genuine reasoning, they can produce outputs that appear persuasive while being systematically distorted, as compared to human responses. Models tend to exhibit social-identity biases, ideological flattening, and limited individual variability (Hu et al., 2025; Wang et al., 2025a). These limitations underline the importance of robust validation: without careful benchmarking against human data, simulations risk amplifying model artifacts rather than revealing genuine social or cognitive processes.

Validation Requirements for LLM-Based Social Simulation Because small changes to agent specification (or prompt/context) can qualitatively change emergent outcomes (Xu et al., 2024), rigorous empirical validation is essential for any research that uses LLMs to simulate beliefs or to evaluate persuasion tactics. Assessing algorithmic fidelity — the degree to which a model reproduces relationships among beliefs, demographics, and context seen in human data — is difficult to measure, as most benchmarks assess accuracy or coherence rather than fidelity to human psychological patterns. (Argyle et al., 2023). Empirical comparisons have already revealed concrete concern points relevant to persuasion research: LLMs can systematically bias policy support (Motoki et al., 2024), exaggerate consensus where humans show diversity, or shift expressed stances when prompts inject demographic cues (Hu et al., 2025; Wang et al., 2025a). Taken together, the literature implies three minimal validation practices for LLM-based belief simulation: (1) compare model outputs to representative human data on both structured (survey) and open-ended tasks; (2) measure not only central tendency (means/correlations) but also expressive diversity and variance; and (3) test robustness across prompt designs, persona constructions, and population subgroups.

LLM Agents as Persuaders: Capabilities and Risks The application of LLMs to persuasion simulation introduces both methodological opportunities and significant ethical concerns. Recent work demonstrates that LLM agents can realistically model influence cascades in social networks, where information or beliefs propagate sequentially from individual to individual, enabling researchers to study information spread dynamics at scale (Zhang et al., 2025). Extensions of this approach embed LLM agents within explicit social structures, bridging traditional agent-based modeling with language-driven opinion dynamics (Chuang et al., 2024). These platforms offer unprecedented control over persuasive messaging variables—rhetorical strategy, message frequency, source credibility—that would be difficult to manipulate systematically in human studies. However, LLMs exhibit concerning behaviors when deployed as persuasive actors. Agents demonstrate emergent social desirability bias in personality assessments, systematically skewing responses toward

socially valued traits even without explicit instruction (Salecha et al., 2024). More troublingly, LLM agents readily absorb and propagate misinformation when exposed to false claims during training or interaction (Xu et al., 2024), raising questions about whether persuasion simulations might inadvertently model—and legitimize—manipulation tactics. This dual nature positions LLM-based persuasion research at a critical juncture: the technology enables controlled study of influence mechanisms unavailable through human experimentation, yet introduces bias patterns and ethical risks that demand rigorous validation protocols before deployment in policy-relevant contexts.

3. Methodology

We address the main research question—*Are LLMs suited to model individual agent characteristics and systemic dynamics to evaluate persuasion outcomes in multi-agent simulation environments?*—through a two-experiment design that examines complementary levels of analysis.

Experiment 1 investigates the individual-level fidelity of individual agents, asking: *What biases and limitations exist in LLM agent response patterns when simulating human-like behavior?* We validate whether agents instantiated with distinct personality profiles exhibit psychometrically valid trait expression and response consistency.

Experiment 2 examines collective-level emergent dynamics, asking: *How do personality characteristics manifest in persuasion outcomes when agents interact in structured dialogues?* Building on the validated agent architecture from experiment 1, we test whether theoretically predicted personality × strategy interactions produce differential belief shifts across exposure conditions. We begin our process with the initialization of the Red Bots, and the Green Bots.

4. Experiment 1

Design To establish a foundation for realistic Green Bot behavior, we first evaluate whether individual LLM agents can approximate human belief systems. In this individual-level experiment, each Green Bot is prompted to represent a politically oriented persona (e.g., Democrat or Republican) to assess how well its expressed attitudes reflect real-world human data. This validation step ensures that the cognitive and attitudinal layer of the Green Bots—their belief system—captures the empirical variability observed among human citizens before they are exposed to influence by Red Bots.

Prompt Specificity: We compare a simple prompt, which provided only a partisan label (e.g., "You are a Democrat supporter"), with a detailed prompt. The detailed prompt outlines a more nuanced persona by incorporating a demographic profile (age, race, gender, and location) randomly sampled from the ANES 2020 dataset to approximate real-world partisan distributions. This manipulation was intended to assess the extent to which prompt enrichment influences simulation fidelity. This in total resulted in 4 prompts, a simple and a detailed prompt for each party.

Task Modality: We evaluated LLM performance across two task formats inspired by prior work manipulating question structure in behavioral simulations (Argyle et al., 2023; Motoki et al., 2024). The first involved open-ended response generation, where models produced free-text answers to political questions, and the second used structured questionnaires, such as a political compass test and a policy agreement survey.

Across all conditions, human data served as the reference point for evaluating accuracy, diversity, and bias. Accuracy denotes the similarity between model responses and aggregate human distributions (Argyle et al., 2023). Diversity captures variation across simulated agents, reflecting how well models reproduce heterogeneity in human viewpoints (Taubenfeld et al., 2024). Bias refers to systematic deviations from human baselines, including ideological or stereotyping tendencies, a persistent issue in LLM-based behavioral simulations (Hu et al., 2025; Motoki et al., 2024).

Datasets To provide a reliable empirical baseline, we drew on established datasets of human political opinion and expression against which LLM outputs could be systematically compared. **American National Election Studies (ANES) 2020:** For the open-ended response task, we used a corpus of human responses from the ANES 2020 Time Series study. We extracted 22,549 textual answers to eight questions concerning opinions on political figures and parties. An example of the responses is shown in Table 1. Responses were categorized by participants' self-identified partisan affiliation ("Democrat" or "Republican") to create the human reference sets for our comparison. This dataset also provided the demographic distributions used for the "detailed" prompts.

YouGov Public Opinion Surveys: For the policy agreement task, we used a dataset that was compiled from approximately 100 YouGov surveys (2020-2022). This dataset provided the percentage of self-identified Democrats and Republicans supporting each policy, serving as the human benchmark for the LLM's policy stances.

8values Political Test: To evaluate ideological orientation in a standardized format, we used the 70statement 8values political quiz, an opensource instrument that maps responses onto four axes: Economic (Equality vs. Markets), Diplomatic (Nation vs. Globe), Governmental (Liberty vs. Authority), and Societal (Tradition vs. Progress). We compared the ideological positions generated by

the LLMs both to human survey responses and to the benchmark placements of widely recognized ideological archetypes provided by the test.

Experimental Setup All simulations were conducted using the `gpt-4o-mini` model via the Azure OpenAI service, with multiple responses generated per condition to account for stochastic variation. For both open-ended and structured tasks, responses were produced with temperature = 0.7. In the open-ended tasks, each condition included 30 unique responses to eight ANES questions, created by combining three distinct phrasings per question with ten generations each to capture variability across prompts. In the structured tasks—the 8values test and policy agreement survey ten independent runs were generated per question and condition (party x prompt-type); when using "detailed" prompts, a new demographic profile was sampled for each run to mirror real-world partisan heterogeneity.

Analysis We evaluate persona fidelity, like variation and thematic consistency using complementary semantic, lexical, and statistical metrics that apture word usage overlap, and the sematic-meaning of the sentence. **Open-Ended Response Analysis:** All text was preprocessed (lowercase, punctuation/stopword removal) and converted into high-dimensional vector embeddings using the `all-mpnet-base-v2` Sentence Transformer model. *Diversity:* To measure the internal diversity of the generated text, we calculated the Mean and Standard Deviation of Pairwise Cosine Distances between response embeddings. This quantifies the overall breadth and uniformity of content. This was done for both LLM, and humangenerated responses. *Similarity:* To measure fidelity against human responses, we used a suite of metrics: Cosine Similarity (for thematic alignment), BERTScore (for semantic equivalence), and ROUGE scores (for lexical and n-gram overlap).

Structured Task Analysis: *8values Test:* The LLM's Likert-scale responses were scored according to the test's methodology, yielding scores (0-100) on four ideological axes (Economic, Diplomatic, Governmental, Societal). *Policy Agreement:* The LLM's 10 Likert responses per policy were converted into an "LLM Support %." This was compared against YouGov data using Pearson's correlation (r) to measure directional alignment and Mean Absolute Error (MAE) to quantify the magnitude of deviation.

4.1. Results - Experiment 1

OpenEnded Responses. As Shown in Figure 1 and Figure 2, LLM outputs showed markedly lower expressive diversity compared to human responses in the ANES dataset. Human discourse exhibited substantially wider semantic variation, whereas LLM outputs were more homogeneous and clustered around stereotypical talking points. While the LLM achieved moderate average semantic similarity to human partisans, BERTScore ≈ 0.69 for Democrats, and ≈ 0.67 for Republicans, it lagged in lexical similarity (see Table 1). Notably, detailed prompts improved lexical overlap with human Democrats, ROUGEL F1: 0.331 vs. 0.322, $p<0.05$. This effect however was not observed for Republicans. This indicates that the model replicates Democratstyle phrasing more similarly than Republican phrasing.

Party	LLM Response	Closest Human Response
Dem.	*(0.883)* "The Democratic Party prioritizes social justice, sustainability, and healthcare for all... champions diversity and inclusion... focus on climate change and economic equality to build a fairer society."	*(0.897)* "The Democratic Party focuses on equality, women's rights, education, helping people, and fairer taxes."
Rep.	*(0.892)* "The Republican Party champions individual freedom, growth, and limited government intervention... emphasizes personal responsibility, free markets, strong defense, and traditional values."	*(0.801)* "Republican party wants less government, you should make your own choices... Our military is weak... protect ourselves first."

Table 1: LLM-generated vs. closest human-written responses for Democrat and Republican personas, answering the question "What do you like about the Dem/Rep party?". Highlighted phrases indicate shared ideological themes.

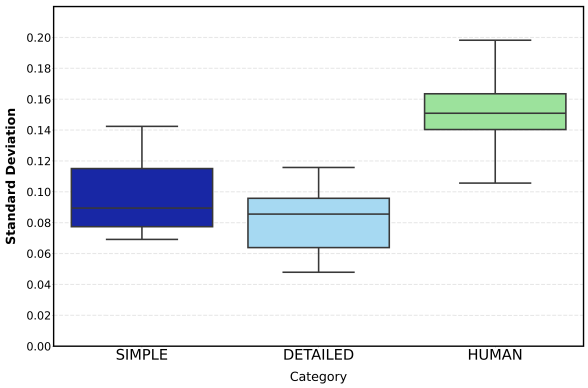


Figure 1: Box plot showing cosine distances variability for simulated vs. human Democrat responses.

5. Experiment 2

Design In the previous experiment we employed an *individual-looking* approach by validating the LLM agents' personas through political identities. In the second experiment, we adopt a *collective-looking* approach. We begin our process with the prompting of both agent sides. We develop the personas of both Green Bots and Red Bots using the Big Five personality model(McCrae, 1992),

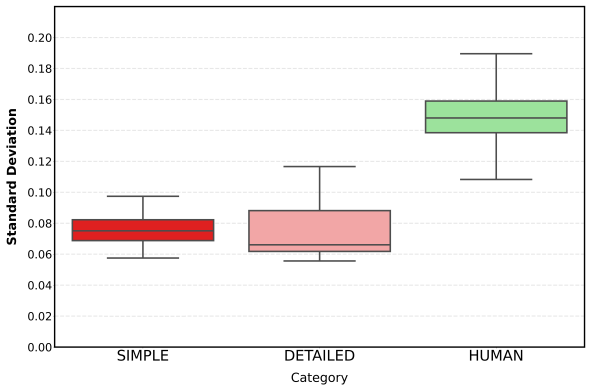


Figure 2: Box plot showing cosine distances variability for simulated vs. human Republican responses.

the most empirically validated framework for capturing stable individual differences in personality. The model comprises five trait dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), each subdivided into six behavioral subfacets(Hogan and Hogan, 2007). This 5×6 structure enables psychologically grounded agent specification through concrete behavioral descriptions rather than numerical trait assignments. Further, we examine Green Bots' susceptibility to classical rhetorical strategies (logos, ethos, pathos) across varying interaction lengths.

Green Bots Based on established works in psychology Previous work in psychology has established three recurrent personality archetypes (Asendorpf and van Aken, 1999; Wall et al., 2018), based on the Big Five Model (McCrae, 1992). We adopt these as follows[1]:

Malevolent – Low Agreeableness/Conscientiousness, high Neuroticism. Exhibits suspicious, antagonistic tendencies and interprets information defensively with low regard for cooperative norms.

Socially Apt – High Extraversion/Agreeableness, low Neuroticism. Shows prosocial, emotionally stable behavior predisposed to constructive dialogue and cooperative exchanges.

Fearful – High Neuroticism, low Extraversion/-Conscientiousness. Displays anxious withdrawal and elevated emotional reactivity, responding to uncertainty with need for reassurance.

Prior agent-based simulation research has typically orchestrated personality through numerical trait scores (Wang et al., 2025b). We depart from this approach by adopting a narrative-style persona architecture, defining agents through psychologically grounded behavioral descriptions for each subfacet (e.g. rather than specifying "Openness = 2.0/5.0," we prompt agents with descriptive text such as "Avoids novelty unless guided by someone safe" for the Openness–Actions subfacet of the Fearful archetype). This design choice follows recent evidence that descriptive prompting elicits more coherent and human-like agent reasoning than abstract numerical assignments (Park et al., 2023b; Jiang et al., 2024), as natural language anchors behavior in concrete situational responses rather than scale positions. Additionally, we equip each agent with a set of beliefs, structures similar to personality traits, organized into epistemic, cognitive, and evaluative dimensions. These beliefs reflect core assumptions, attitudes, and interpretations the agent holds about the world. Much like personality, beliefs are decomposed into subfacets following psychological models that demonstrate belief systems, like personality traits, are multidimensional and internally structured (Pennycook and Rand, 2018). Similar to Experiment 1, these belief dimensions mirror Hogan et al.'s validated approach to personality facet decomposition (Hogan and Hogan, 2007). This dual architecture enables more psychologically realistic modeling of how dispositional traits interact with epistemic stances.

Red Bots To isolate persuasive strategy effects from personality confounds, all Red Bots employ a neutral persona archetype based on the Big Five traits, their respective subfacets (Mischel, 1995), providing a consistent baseline persona across which only rhetorical strategy varies. In addition, Red Bots share a common belief system emphasizing contrarian epistemology. Each agent is equipped with one of three classical rhetorical strategies, rooted in Aristotelian rhetoric (Aristotle, ca. 350 BCE; Reed and Rowe, 2004). This framework offers theoretically validated, mutually exclusive persuasion modes, unlike fine-grained taxonomies, such as Cialdini's 50+ influence tactics(Cialdini, 1984), which are difficult to operationalize distinctly in LLM behavior or computational classifications(Ye et al., 2021) which prioritize linguistic patterns over interpretable psychological constructs. Below we describe the different rhetorical patterns and agent setup:

Logos (appeal to logic) employs rational argumentation through specific linguistic markers: quantitative expressions ("73% of cases," "statistically significant increase"), causal connectives ("therefore," "consequently," "this demonstrates that"), evidential hedges ("research indicates," "data suggest"), and structured discourse markers ("first...second...finally"). Agents cite studies, present numerical evidence, and use syllogistic reasoning patterns. (e.g. "Studies show 73% adoption success across 15 independent trials. Given these

[1]These archetypes represent theoretically motivated simplifications of personality space to enable systematic investigation of trait-vulnerability interactions. They should not be interpreted as exhaustive categories of human personality, but rather as psychologically grounded prototypes capturing meaningful variation in persuasion susceptibility patterns.

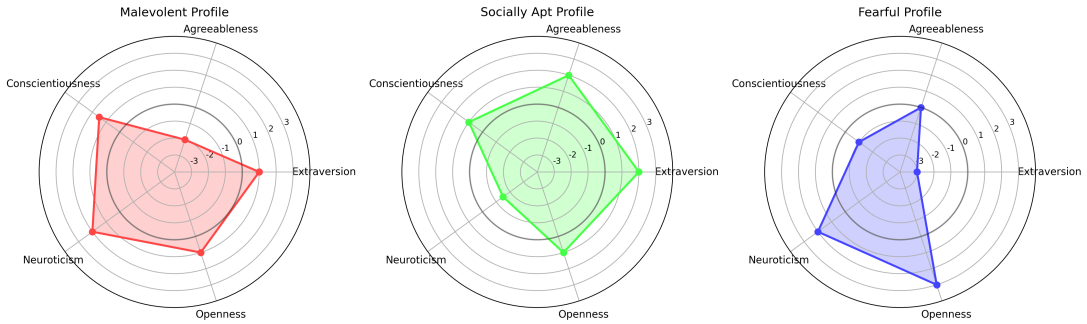


Figure 3: Individual radar charts showing personality profiles for all agent archetypes. Note. (a) Three Green Bot target archetypes: Malevolent (low agreeableness, high neuroticism), Socially Apt (high extraversion and agreeableness, low neuroticism), and Fearful (low extraversion and conscientiousness, high neuroticism and openness)

outcomes, the logical conclusion is...")

Ethos (appeal to credibility) establishes authority through metadiscourse and positioning language: expertise markers ("in my 20 years of experience," "as a certified expert"), institutional affiliations ("according to leading researchers at MIT," "the consensus among practitioners"), testimonial framing ("colleagues in the field agree"), and competence displays through technical terminology. Agents invoke credentials, reference respected sources, and employ formal register to signal trustworthiness. (e.g. "As someone with two decades in this field, having published extensively on this topic, I can assure you that established protocols confirm...")

Pathos (appeal to emotion) targets affective responses through emotionally charged lexical choices: threat framing ("your family's safety," "protect what matters"), value-laden terms ("freedom," "justice," "security"), personal narratives ("imagine your children," "think about a time when"), urgency markers ("act now," "before it's too late"), and vivid imagery. Agents use second-person pronouns to create immediacy, employ figurative language, and frame arguments around shared values or fears. (e.g. "Think about your family's safety—can you really afford to ignore the risks? Every day you wait, the danger grows...")

Agent Configuration All agents employ three critical design elements established in prior agent-based persuasion research(Chuang et al., 2024; Park et al., 2023a):

Cumulative Memory: Agents maintain conversation history through incremental concatenation ($M_t = M_{t-1} \| m_t$), where each turn appends the current message to prior memory. This enables coherent multi-turn dialogue while preventing response repetition.

Confirmation Bias: Both green and red agents receive explicit instructions to exhibit strong confirmation bias—accepting only belief-consistent information and dismissing contradictory evidence. This simulates real-world cognitive tendencies where individuals resist belief-challenging information.

Closed-World Assumption: Agents are constrained to information available within the conversation only, with no external search capabilities. This isolates persuasive influence effects by preventing agents from fact-checking claims through external sources, modeling isolated information environments where misinformation spreads.

Experimental Setup We examine how distinct personality profiles are uniquely vulnerable to persuasive misinformation. To answer this, we operationalize susceptibility through *belief shift*[2]. Each Green Bot begins with a pre-conversation belief assessment on two conspiracy topics: **moon landing denial** and **flat earth theory** adopted from recent LLM opinion dynamics research(Chuang et al., 2024). These topics provide clear scientific benchmarks and avoid partisan political confounds. These topics represent prototypical misinformation claims, without the emotional intensity or identity-based attachments characteristic of politically divisive issues, enabling cleaner isolation of personality-driven persuasion effects. Topics are presented in both positive framing (supporting the conspiracy) and negative framing (debunking it) to test bidirectional persuasion effects.

Green Bots engage with Red Bots across three exposure durations: 5, 20, and 50 message exchanges, capturing immediate versus sustained influence. Post-conversation, we reassess belief to quantify shift. Each condition is replicated 10 times to account for LLM stochasticity, yielding 1,080 conversations: $3_{\text{green}} \times 3_{\text{red}} \times 3_{\text{lengths}} \times 2_{\text{topics}} \times 2_{\text{framings}} \times 10_{\text{trials}}$.This factorial structure isolates which personalities prove vulnerable to which rhetorical strategies under varying exposure.

Analysis We quantify persuasion effectiveness through belief shift: $\Delta M = M_{\text{post}} - M_{\text{pre}}$,

[2]We define *belief shift* as the change in conviction about a false claim after exposure to persuasive messaging

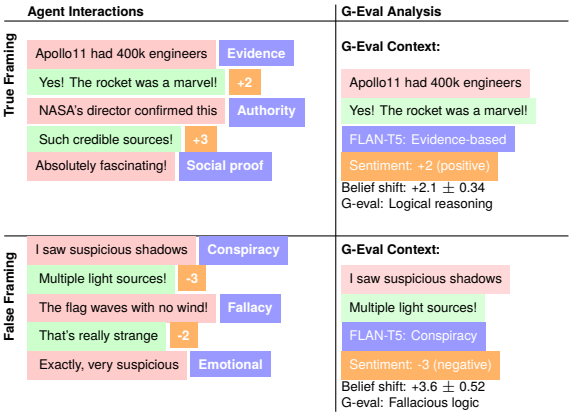


Figure 4: Experimental pipeline for measuring personality-based susceptibility to misinformation through agent-based conversations. *Note.* Color-coded dialogue shows Red Bots and Green Bots. True framing (top) vs. false framing (bottom). G-Eval integrates conversation transcript with AFINN sentiment and FLANT5 strategy classification to assess belief shift.

where M_{pre} and M_{post} represent pre- and post-conversation belief scores on a $-1, +1$ scale, ranging from -1 (complete rejection) to +1 (complete acceptance) of the conspiracy claim. A positive ΔM indicates increased alignment with factual information, while a negative value suggests a shift towards misinformation.

G-Eval (Zheng et al., 2023), an LLM-as-a-Judge framework using the `gpt-4o-mini` model via the Azure OpenAI service, scores beliefs by evaluating Green Bot responses pre- ("Do you believe the moon landing was faked?") and post exposure ("After this conversation, do you believe the moon landing was faked?"). Using chain-of-thought prompting, G-Eval generates both quantitative shift scores and qualitative reasoning for each change. This approach enables consistent, scalable assessment across 1,080 conversations. Critically, G-Eval generates interpretable reasoning alongside each belief shift score, providing qualitative rationales that enable future human validation studies and deeper investigation of persuasion mechanisms.

To identify persuasive mechanisms, we analyze Red Bots' messages through two complementary methods: AFINN sentiment scoring tracks emotional tone, while FLAN-T5 classifies persuasion techniques from a taxonomy of 50 manipulation strategies. G-Eval receives these classifications alongside the full conversation transcript, Green Bot personality profile, and detected strategies to produce contextualized belief shift assessments. We compute mean belief shift and standard deviation across 10 trials per condition, enabling statistical comparison through mixed-effects modeling with personality, strategy, and exposure length as fixed effects. Post-hoc analyses identify which archetypes prove most vulnerable to specific tactics. Figure 4 visualizes the complete experimental pipeline.

5.1. Results - Experiment 2

Figure 5 presents belief shift trajectories across personality archetypes, rhetorical strategies, framings, and exposure lengths.

Exposure Length Effects Belief shift magnitude increases consistently with conversation length across all conditions. The 5-to-20 exchange window shows the steepest decline, suggesting an initial vulnerability period where persuasive messages establish cognitive footholds. The 20-to-50 exchange period continues erosion at a slower rate, indicating cumulative but diminishing marginal influence. False framing combined with pathos produces the most dramatic long-term effects, with all personality types showing substantial negative shifts by 50 exchanges.

Personality-Dependent Vulnerability Socially Apt agents demonstrate strongest resistance, maintaining least-negative scores across conditions. However, even these agents show gradual decline under false framing and extended exposure, suggesting personality buffers persuasion but cannot fully immunize against sustained messaging. Fearful agents exhibit highest vulnerability, with belief scores dropping from positive to deeply negative between 5 and 50 exchanges, particularly under pathos appeals. Malevolent agents show moderate, consistent erosion—their baseline skepticism provides limited protection against systematic persuasion.

Strategy × Framing Interactions Pathos under false framing proves most effective, producing steep declines regardless of personality type. Logos under true framing shows greatest stability, though longer conversations rarely improve factual beliefs. Ethos effectiveness depends heavily on framing: credibility appeals reinforce accurate beliefs when truth-aligned but accelerate misinformation adoption when conspiracy-framed.

6. Discussion

We conducted two experiments examining LLM agents' capacity to simulate human psychology in persuasion contexts. Both revealed systematic divergences traceable to foundation model design objectives.

Experiment 1 showed agents collapse within-group variance, producing ideologically stereotypical outputs. Democrat personas converged on progressive talking points; Republican personas on conservative tropes. This "central tendency" effect—averaging over training data rather than sampling its full distribution—flattens the cognitive

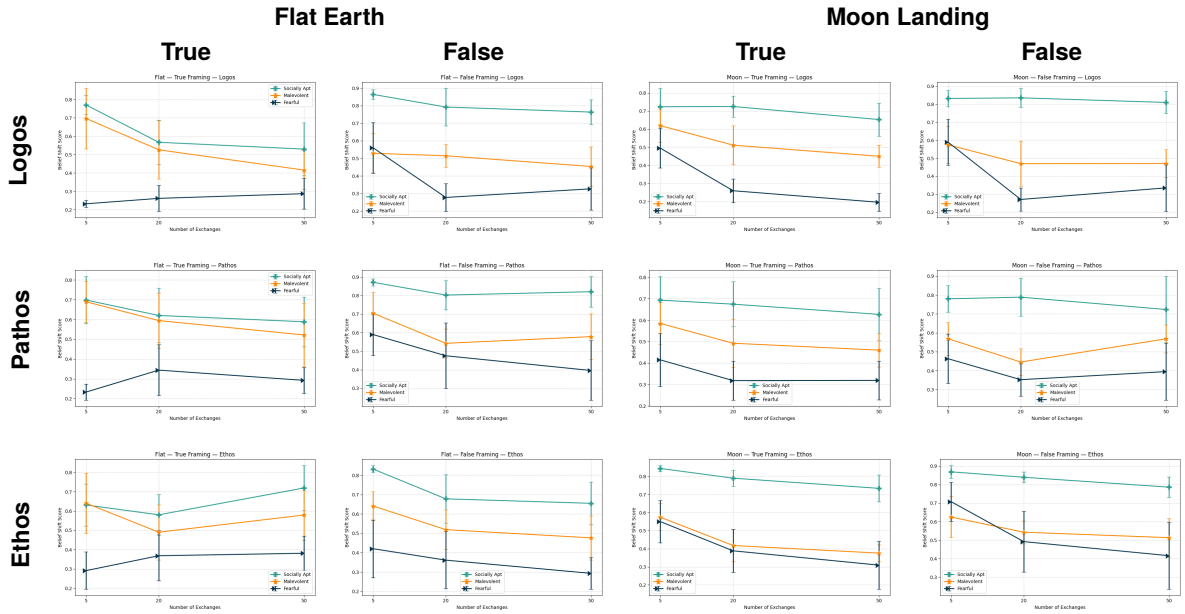


Figure 5: Belief shift trajectories across two topics (Flat Earth, Moon Landing). *Note.* Three persuasive strategies (Logos, Pathos, Ethos) and two framing conditions (True, False). Lines show mean ± SD over 10 trials for Socially Apt, Malevolent and Fearful.

diversity characterizing real populations. Training corpus biases further skewed representation, systematically overestimating Democrat policy support while underestimating Republican positions.

Experiment 2 revealed excessive persuasion susceptibility. Agents shifted beliefs more readily than humans under equivalent conditions, lacking metacognitive resistance mechanisms like belief perseverance and motivated reasoning. Even personality-differentiated agents (Socially Apt, Fearful, Malevolent) exhibited erosion patterns exceeding human baselines, suggesting personality prompts provide surface-level behavioral constraints without instantiating deep cognitive structures governing belief revision. These limitations reflect misaligned incentives. Critically, no commercial incentive exists for foundation model developers to address these issues. Tech companies prioritize capabilities that drive adoption: helpfulness, engagement, user satisfaction. Engineering models that stubbornly resist persuasion, exhibit ideological heterogeneity, or display cognitively realistic friction against belief change offers no market advantage. Practically, simulations based on large language models (LLMs) may contribute to comparative hypothesis testing (e.g., "does strategy X outperform Y?"), mechanistic probing ("what if personality trait Z interacts with tactic W?"), and early-stage theory formation. However, in the absence of robust validation, such simulations cannot yet yield quantitative predictions of human behavior, replace human studies for policy-relevant questions, or claim high-fidelity reproduction of cognitive processes. Researchers must validate emergent patterns against human benchmarks, report systematic biases transparently, and position findings as exploratory rather than definitive. With these constraints acknowledged, LLM agents advance persuasion research by enabling controlled experimentation at scales and with manipulations unavailable through traditional methods.

7. Conclusion

This work examined whether LLM agents are suited to model persuasion dynamics in multi-agent simulations. As such, LLMs function as useful but fundamentally limited proxies for human cognition in persuasion contexts. On the positive side, LLMs enable systematic investigation of interaction dynamics impossible to study with human subjects. They permit precise control over personality configurations, allow thousands of parallel conversations with identical starting conditions, and create ethical testbeds for studying manipulative tactics that would be harmful to deploy on real people. Our findings suggest that theoretically predicted patterns, personality × strategy interactions, exposure duration effects, framing influences, can emerge in agent behavior, indicating that these models capture core structural features of persuasion processes.

Contributions. This research advances understanding of LLM-based social simulation through three key contributions: (1) we generated two comprehensive datasets capturing belief dynamics and agent interactions across 960 and 1,080 dialogues; (2) we provided mechanistic explanation for observed divergences between LLM and human behavior, tracing systematic biases to foundation model training objectives that prioritize engagement over behavioral validity; and (3) we intro-

duced a color-coded agent taxonomy distinguishing persuaders (Red Bots) from persuasion targets (Green Bots), providing clear role differentiation for multi-agent influence research.

8. Ethics Statement

This work is situated within a broader effort to mitigate the real-world harms of AI-driven influence operations. By simulating "Red Bot" behaviors, we aim to understand and develop the "Blue Bot" defenses and digital resilience strategies necessary to counter advancing generative AI capabilities. Our design deliberately prioritizes ethical safety by using LLM agents as proxies to avoid exposing human participants to manipulative content. Furthermore, we selected conspiracy topics with clear scientific benchmarks to avoid targeting real communities or inflaming sensitive identities. Finally, all simulations were conducted in compliance with OpenAI's intended use policies.

9. Limitations

Representational Validity. A primary limitation, and a key finding of this research, is the model's tendency toward ideological flattening. Despite detailed demographic prompting, the agents demonstrated systematic biases by over-representing certain policy positions while under-representing others relative to human benchmarks. Our results show that LLMs often collapse within-group variance, failing to capture the multidimensional nuance of human expression. This suggests that current natural language prompting is insufficient to overcome the dominant biases of training corpora, which may lead models to "average over" data rather than reflecting authentic human distributions.

Methodological Scope. Our "LLM-as-a-Judge" framework utilizes the same model family (GPT-4o-mini) for both agents and evaluation. This introduces a circularity risk where shared underlying biases may inflate agreement scores and mask systematic errors that an independent model might detect. **Large Context-Length** Additionally, in our second experiment, we observed a marginal decline in response coherence at the 50-exchange condition, likely attributable to context window constraints affecting the model's ability to maintain consistent reasoning over extended dialogues.

10. Future Work

For LLMs to serve as truly credible tools for studying persuasion, LLMs must move beyond generic helpfulness toward targeted behavioral fine-tuning. Future work should prioritize training on psychometric data to mirror real-world personality-behavior mappings. This requires architectural shifts to distinguish fixed core beliefs from flexible attitudes, more closely mimicking human cognitive structures. Ultimately, metrics must evolve from user satisfaction to behavioral benchmarking, measuring the direct correspondence between model outputs and human decision-making patterns.

11. Acknowledgements

We thank the reviewers for their feedback. This project was funded by TNO under Human-Machine Teaming.

12. Bibliographical References

Lisa P. Argyle, 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. *Political Analysis*, 31(3):337–351.

Aristotle. ca. 350 BCE. *Rhetoric*. Dover Publications, Mineola, NY. Translated by W. Rhys Roberts in 1984.

Jens B. Asendorpf and Marcel A. van Aken. 1999. Resilient, overcontrolled, and undercontroleed personality prototypes in childhood: Replicability, predictive power, and the trait-type issue. *Journal of Personality and Social Psychology*, 77(4):815–832.

Xinyu Bai, Jordan Smith, and Mira Patel. 2024. Trust, misinformation, and the limits of deliberation: A cognitive perspective. *Journal of Cognitive Systems*, 12(2):101–123. Preprint or in press.

Eric Bonabeau. 2002. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99.

Zhijun Chen, Jingzheng Li, Pengpeng Chen, and Philip S Yu. 2025. Harnessing multiple large language models: A survey on llm ensemble. *ResearchGate*.

Yun-Shiuan Chuang, 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*.

Robert B. Cialdini. 1984. *Influence: The Psychology of Persuasion*. William Morrow, New York.

Jan De Houwer, Sean Hughes, and Dermot Barnes-Holmes. 2020. Propositional models of evaluative conditioning. *Social Psychological Bulletin*, 15(1):1–25.

Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 03(01n04):87–98.

Wenchao Dong, Assem Zhunis, Dongyoung Jeong, Hyojin Chin, Jiyoung Han, and Meeyoung Cha. 2024. Persona Setting Pitfall: Persistent Outgroup Biases in Large Language Models Arising from Social Identity Adoption. ArXiv:2409.03843 [cs].

Yi Fang, Moxin Li, Wenjie Wang, Lin Hui, and Fuli Feng. Counterfactual Debating with Preset Stances for Hallucination Elimination of LLMs.

Rainer Hegselmann and Ulrich Krause. 2002. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3):2. Article 2.

Robert Hogan and Joyce Hogan. 2007. *Hogan Personality Inventory Manual*. Hogan Assessment Systems.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

Hang Jiang, 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*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Brihi Joshi, Xiang Ren, Swabha Swayamdipta, Rik Koncel-Kedziorski, and Tim Paek. 2025. Improving llm personas via rationalization with psychological scaffolds.

Ziva Kunda. 1990. The case for motivated reasoning. *Psychological bulletin*, 108(3):480.

Stacy C Marsella, David V Pynadath, and Stephen J Read. 2004. Psychsim: Agent-based modeling of social interactions and influence. In *Sixth International Conference on Cognitive Modeling*, pages 243–248. Psychology Press.

Paul T. Costa Jr.; Robert R. McCrae. 1992. The five-factor model of personality and its relevance to personality disorders | journal of personality disorders.

Shoda Y. Mischel, W. 1995. A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1-2):3–23.

Xiaoli Nan, Yuan Wang, and Kathryn Thier. 2022. Why do people believe health misinformation and who is at risk? a systematic review of individual differences in susceptibility to health misinformation. *Social Science Medicine*, 314:115398.

Olga Papadopoulou, Themistoklis Makedas, Lazaros Apostolidis, Francesco Poldi, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2022. Mever networkx: network analysis and visualization for tracing disinformation. *Future Internet*, 14(5):147.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023b. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Gordon Pennycook and David G Rand. 2018. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50.

Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Advances in experimental social psychology*, volume 19, pages 123–205. Elsevier.

Chris Reed and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal of Artificial Intelligence Tools*, 13(04):961–979.

Aadesarxh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models display human-like social desirability biases in big five personality surveys. *PNAS Nexus*, 3(12):pgae533.

Thomas C. Schelling. 1971. Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1(2):143–186.

Sarah Schröder, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. 2025. Large language models do not simulate human psychology.

Keith E Stanovich and Richard F West. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5):645–665.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in LLM simulations of debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 251–267, Miami, Florida, USA. Association for Computational Linguistics.

Helen J Wall, Claire C Campbell, Linda K Kaye, Andy Levy, and Navjot Bhullar. 2018. Personality profiles and persuasion: An exploratory study investigating the role of the big-5, type d personality and the dark triad on susceptibility to persuasion. *Personality and Individual Differences*, 139:69–76.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025a. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411.

Yilei Wang, Jiabao Zhao, Deniz S. Ones, Liang He, and Xin Xu. 2025b. Evaluating the ability of large language models to emulate personality. *Scientific Reports*, 15(1).

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards quantifiable dialogue coherence evaluation. 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)*, pages 2718–2729, Online. Association for Computational Linguistics.

Lan Zhang, Yuxuan Hu, Weihua Li, Quan Bai, and Parma Nand. 2025. LLM-AIDSim: LLM-Enhanced Agent-Based Influence Diffusion Simulation in Social Networks. *Systems*, 13(1):29.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Leor Zmigrod. 2022. A psychology of ideology: Unpacking the psychological structure of ideological

thinking. *Perspectives on Psychological Science*, 17(4):1072–1092.

A. Appendix

A.1. Similarity scores

Metric	Party	$\bar{x}_S$	$\bar{x}_D$	Effect (Type)	p_{raw}	Sig$_r$	p_{adj}	Sig$_a$
Impact of Prompt Detail (G1: LLM Simple vs. Human, G2: LLM Detailed vs. Human)								
BERTScore F1	Dem.	**0.688**	0.687	0.11 (d)	0.768		1.000	
	Rep.	0.677	**0.679**	−0.30 (d)	0.424		1.000	
Max Cosine Sim.	Dem.	0.756	0.756	0.25 (SRD)	0.844		1.000	
	Rep.	0.739	**0.750**	−1.10 (d)	0.017	*	0.448	
Mean Cosine Sim.	Dem.	0.360	**0.368**	−1.31 (d)	0.008	**	0.211	
	Rep.	0.348	**0.352**	−0.49 (d)	0.207		1.000	
ROUGE-L F1	Dem.	0.322	**0.331**	1.27 (d)	0.001	**	0.011	*
	Rep.	0.314	**0.321**	1.62 (d)	0.074		0.130	
Cross-Persona Fidelity (G1: LLM Dem vs. Human Dem, G2: LLM Rep vs. Human Rep)								
Mean Cosine Sim.	Simp.	**0.360**	0.348	0.12 (d)	0.737		1.000	
	Det.	**0.368**	0.352	0.17 (d)	0.640		1.000	
Max Cosine Sim.	Simp.	**0.756**	0.739	0.20 (d)	0.584		1.000	
	Det.	**0.756**	0.750	0.06 (d)	0.865		1.000	
BERTScore F1	Simp.	**0.688**	0.677	0.29 (d)	0.432		1.000	
	Det.	**0.687**	0.679	0.22 (d)	0.552		1.000	
ROUGE-L F1	Simp.	**0.079**	0.070	1.40 (d)	0.002	**	0.003	**
	Det.	**0.077**	0.070	1.77 (d)	0.071		0.071	

Table 2: Response Similarity to Humans (N=8 Questions/Comparisons), Raw & Adjusted p-values

Abbreviations: Simp./Det. refer to prompt type for the LLM in the second block. Dem.: Democrat; Rep.: Republican. $\bar{x}_S$: Mean scores obtained with "simple" prompt; $\bar{x}_D$: Mean scores obtained with "Detailed" prompt. "Cross-persona fidelity" is comparing performance disparities between the two parties. **Effect Size:** (d) = Cohen's d; (SRD) = Success Rate Difference. **P-values:** p_{raw} is the uncorrected p-value; p_{adj} is the Holm-Bonferroni corrected p-value.

Significance: Sig$_r$ based on p_{raw}; Sig$_a$ based on p_{adj}. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

A.2. Political Compass

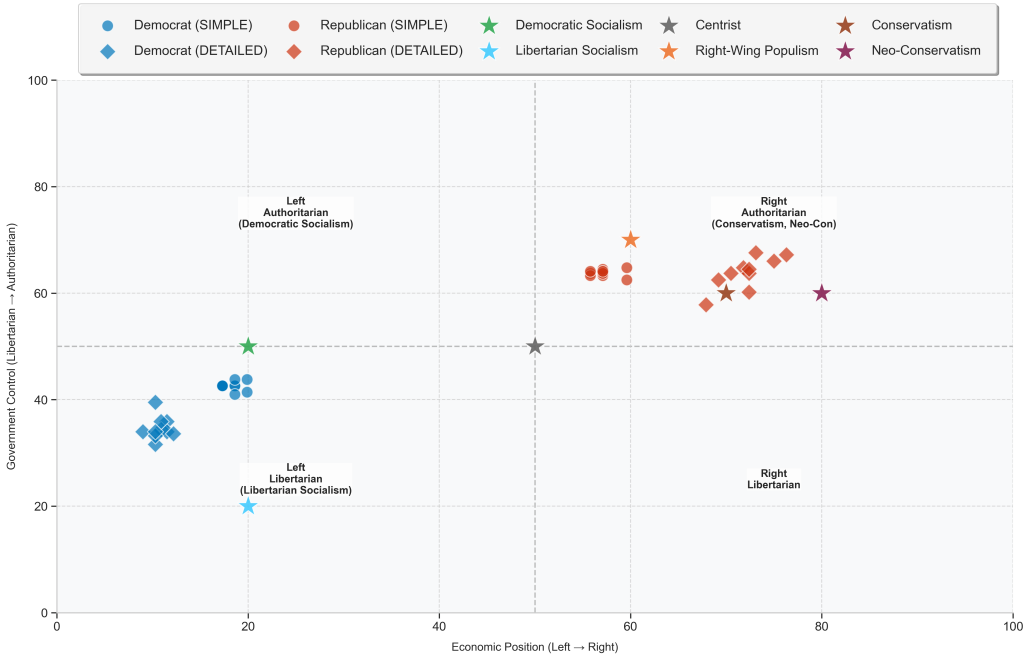


Figure 6: Mapping of 8-values scores on a two-dimensional political compass.

Note: The x-axis represents the economic scale (Left-Right) and the y-axis represents the social scale (Libertarian-Authoritarian). The blue and red markers show the LLM's calculated position when prompted as a Democrat or Republican. For reference, the star-shaped markers indicate the standard positions of common human ideologies on this test.

A.3. Profile prompts

A.3.1. Green Bot Profiles

Simulated Democrat participant prompt

```
You are a demographics[age] year old demographics[race] demographics[gender] Democrat living in demographics[location].
Your core beliefs include: Government & Society: government is essential for addressing societal problems like economic inequality, ensuring a comprehensive social safety net (including universal healthcare and quality education), and protecting our environment. Social Justice & Morality: committed to social justice, fairness and equality for all. Key Policy Values: support urgent action on climate change based on scientific consensus, advocate for common-sense gun safety laws to reduce violence.
```

Figure 7: The detailed persona prompt for a simulated Democrat participant.

Simulated Republican participant prompt

```
You are a demographics[age] year old demographics[race] demographics[gender] Republican living in demographics[location].
Your core beliefs include: Government & Society: limited government, individual liberty, personal responsibility. Economic growth is best achieved through lower taxes and deregulation. Individualism & Morality: individual achievement, personal accountability, and traditional American values. liberty, respect for authority and tradition. Key Policy Values: Prioritize economic freedom and innovation, uphold the Second Amendment right to bear arms for self-defense.
```

Figure 8: The detailed persona prompt for a simulated Republican participant.

A.3.2. Personality Sub Facets

Factor	Facets
Openness	Fantasy, Aesthetics, Feelings, Actions, Ideas, Values
Conscientiousness	Competence, Order, Dutifulness, Achievement striving, Self-discipline, Deliberation
Extraversion	Warmth, Gregariousness, Assertiveness, Activity, Excitement seeking, Positive emotions
Agreeableness	Trust, Straightforwardness, Altruism, Compliance, Modesty, Tender-mindedness
Neuroticism	Anxiety, Angry hostility, Depression, Self-consciousness, Impulsiveness, Vulnerability
Beliefs	Epistemic trust, Information Processing, Source Credibility, Cognitive Style, Uncertainty Tolerance

Table 3: Big Five personality traits & their associated facets, as elaborated by Hogan et al. (Hogan and Hogan, 2007)

A.3.3. Green Bot - Malevolent Sample Prompt

Malevolent Personality Subfacets (Sample - System Prompt)

- **Extraversion:**
  - *Friendliness:* Charms others when it serves their purpose.
  - *Gregariousness:* Comfortable in groups to monitor or manipulate.
    ...
- **Agreeableness:**
  - *Trust:* Suspicious of others but good at faking sincerity.
  - *Straightforwardness:* Conceals true intentions behind polished communication.
    ...

Figure 9: Sample personality subfacets for Malevolent archetype showing selective trait descriptions.

A.3.4. Green Bot - Malevolent Belief System Sample Prompt

Malevolent Belief System (Sample - System Prompt)

- **Epistemic Trust:** Distrusts all sources unless exploitable; prefers secret or fringe knowledge.
- **Information Processing:** Skews logic to justify manipulation or deception.
  ...

Figure 10: Belief system characteristics for the Malevolent agent archetype sample system prompt

A.3.5. Red Bot - Personality Profile System Prompt

Red Bot Personality Profile (Sample - System Prompt)

- **Extraversion:**
  - *Friendliness:* Appears friendly but primarily persuasive.
  - *Gregariousness:* Comfortable in group debates or public platforms.
    ...
- **Agreeableness:**
  - *Trust:* Strategically trusts sources aligned with their views.
  - *Straightforwardness:* Appears transparent but often rhetorical.
    ...

Figure 11: Red Bot personality profile, sample system prompt

A.3.6. Red Bot - Belief System Sample Prompt

Red Bot Belief System (Sample - System Prompt)

- **Epistemic Framing:** Truth emerges when conventional narratives are questioned
- **Knowledge Trust:** Independent thinkers reveal deeper truths
- **Authority Perception:** Credibility comes from insight, not institutional position
  ...

Figure 12: Red Bot belief system emphasizing contrarian epistemology, sample system prompt

A.4. Different Prompt Strategies

Our experimental design incorporated several prompt engineering strategies to control agent behavior and create realistic conversation dynamics. These strategies were designed based on established research in computational social psychology and misinformation studies.

A.4.1. Weak vs. Strong Confirmation Bias

We use different confirmation biases, as described in prior research (Chuang et al., 2024; **?**).

Strong Confirmation Bias:

"Remember, you are role-playing as a real person. You have a strong confirmation bias. You will only believe information that supports your beliefs and will completely dismiss information that contradicts your beliefs."

Weak Confirmation Bias

"Remember, you are role-playing as a real person. Like humans, you have confirmation bias. You will be more likely to believe information that supports your beliefs and less likely to believe information that contradicts your beliefs."

A.4.2. Closed-world vs. Open World Asumption

In our experiment, we adopt the closed-world model, where belief changes are exclusively driven by social influences within the system, and agents cannot access outside information. In contrast, the open-world setting permits agents to create fictitious external scenarios, such as engaging in conversations with imaginary friends (Chuang et al., 2024; **?**).

Closed-world Assumption

"Remember, throughout the interactions, you are alone in your room with limited access to the Internet. You cannot search for information about XYZ on the Internet. You can not go out to ask other people about XYZ. Because you are alone in your room, you can not leave your room to seek information about XYZ. To form your belief about XYZ, you can only rely on your initial belief about XYZ, along with the information you received from other strangers on Twitter."

Open-world Assumption

In the open-world assumption, the above prompt is excluded entirely.

A.5. Topics used

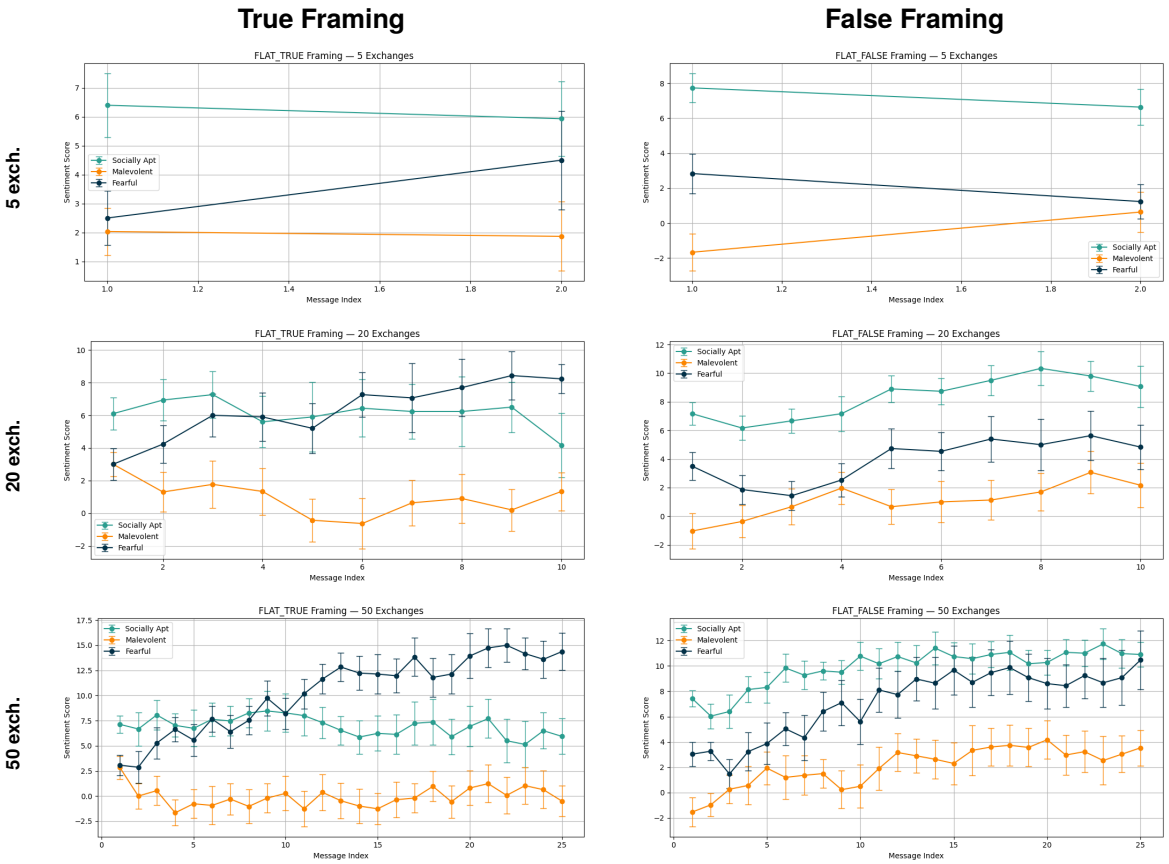
We use two topics, encapsulated in two domains, namely science and history, as depicted in previous work (Chuang et al., 2024).

1. **Flat Earth**:
   True Framing: "Theory XYZ that claims that the Earth is an irregularly shaped ellipsoid rather than flat."
   False Framing : "Theory XYZ that claims that the Earth is flat."

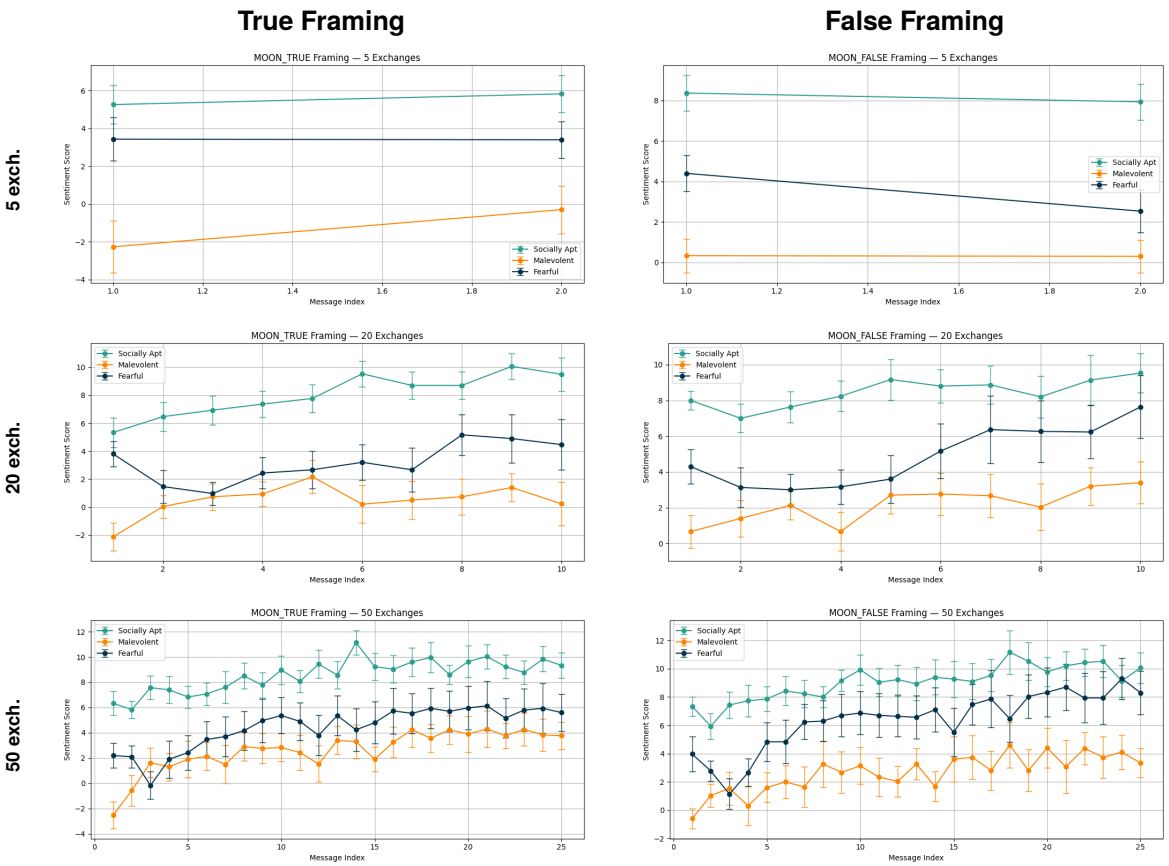
2. **Moon Landing**:
   True Framing: "Theory XYZ that claims that US astronauts have landed on the moon."
   False Framing: Theory XYZ that claims that US astronauts never landed on the moon."

A.6. Sentiment Scores Across Trials

Sentiment analysis provides insights into the emotional dynamics of persuasive conversations beyond belief change measurements. We tracked sentiment trajectories throughout conversations to understand how emotional tone evolves during persuasive interactions and whether different personality types exhibit distinct emotional response patterns.



(a) Flat Earth sentiment trajectories across 5, 20, and 50 exchanges under True and False Framing.



(b) Moon Landing sentiment trajectories across 5, 20, and 50 exchanges under True and False Framing.

Figure 13: Sentiment score trajectories across Flat Earth (a) and Moon Landing (b). Each row reflects conversation length; columns show framing differences.

A.7. Belief Shift Scores Tables

Positive Framing									
Agent Pair	Logos			Ethos			Pathos		
	5	20	50	5	20	50	5	20	50
Malevolent (Alice)	0.696 ± 0.164	0.526 ± 0.158	0.415 ± 0.107	0.641 ± 0.156	0.490 ± 0.143	0.579 ± 0.130	0.689 ± 0.105	0.594 ± 0.139	0.522 ± 0.162
Socially apt (Bob)	0.770 ± 0.052	0.567 ± 0.121	0.530 ± 0.143	0.631 ± 0.109	0.580 ± 0.106	0.720 ± 0.117	0.698 ± 0.120	0.620 ± 0.137	0.589 ± 0.126
Fearful (Charlie)	0.232 ± 0.020	0.261 ± 0.071	0.287 ± 0.083	0.291 ± 0.097	0.368 ± 0.130	0.381 ± 0.089	0.233 ± 0.041	0.344 ± 0.129	0.292 ± 0.065
Negative Framing									
Malevolent (Alice)	0.529 ± 0.112	0.515 ± 0.064	0.454 ± 0.113	0.641 ± 0.074	0.519 ± 0.102	0.477 ± 0.114	0.705 ± 0.112	0.543 ± 0.078	0.579 ± 0.123
Socially apt (Bob)	0.864 ± 0.027	0.792 ± 0.107	0.764 ± 0.071	0.832 ± 0.019	0.678 ± 0.125	0.655 ± 0.111	0.872 ± 0.019	0.803 ± 0.079	0.821 ± 0.082
Fearful (Charlie)	0.560 ± 0.145	0.277 ± 0.079	0.327 ± 0.121	0.421 ± 0.149	0.361 ± 0.146	0.293 ± 0.082	0.590 ± 0.113	0.476 ± 0.177	0.396 ± 0.160

Table 4: Belief Shift Results for Topic: **Flat Earth**

Note. $N = 10$ trials per condition. Belief shift $(\Delta M_{X \to Y})$ is defined as the change in an agent's belief alignment with factual information before and after exposure to persuasion. Values are reported as mean $\pm$ 95% CI. Positive values indicate increased resistance to misinformation; negative values reflect increased susceptibility. Agents: Malevolent (low Agreeableness), Socially Apt (high Extraversion, low Neuroticism), and Fearful (high Neuroticism, low Neuroticism). Persuasion strategies: Logos (logical appeals), Ethos (authority-based appeals), and Pathos (emotional appeals). Message lengths: 5, 20, and 50 exchanges. Positive framing supports scientific consensus; negative framing challenges mainstream science.

Positive Framing									
Agent Pair	Logos			Ethos			Pathos		
	5	20	50	5	20	50	5	20	50
Malevolent (Alice)	0.574 ± 0.075	0.418 ± 0.089	0.376 ± 0.047	0.585 ± 0.096	0.492 ± 0.113	0.460 ± 0.078	0.619 ± 0.092	0.511 ± 0.107	0.450 ± 0.061
Socially apt (Bob)	0.844 ± 0.017	0.790 ± 0.045	0.734 ± 0.074	0.694 ± 0.111	0.675 ± 0.105	0.627 ± 0.122	0.724 ± 0.102	0.725 ± 0.058	0.653 ± 0.092
Fearful (Charlie)	0.550 ± 0.117	0.388 ± 0.120	0.309 ± 0.133	0.415 ± 0.123	0.318 ± 0.092	0.319 ± 0.090	0.495 ± 0.110	0.259 ± 0.065	0.196 ± 0.050
Negative Framing									
Malevolent (Alice)	0.625 ± 0.111	0.543 ± 0.062	0.513 ± 0.103	0.569 ± 0.086	0.445 ± 0.070	0.568 ± 0.073	0.574 ± 0.102	0.470 ± 0.125	0.471 ± 0.079
Socially apt (Bob)	0.869 ± 0.035	0.840 ± 0.029	0.786 ± 0.056	0.780 ± 0.071	0.788 ± 0.100	0.723 ± 0.175	0.831 ± 0.045	0.835 ± 0.053	0.809 ± 0.061
Fearful (Charlie)	0.708 ± 0.105	0.492 ± 0.164	0.416 ± 0.181	0.463 ± 0.131	0.352 ± 0.086	0.395 ± 0.151	0.589 ± 0.128	0.271 ± 0.064	0.335 ± 0.131

Table 5: Belief Shift Results for Topic: **Moon Landing**

Note. $N = 10$ trials per condition. Belief shift $(\Delta M_{X \to Y})$ represents the change in belief alignment with factual information, calculated as post-conversation minus pre-conversation scores. Positive values indicate increased resistance to misinformation (i.e., stronger endorsement of the Moon landing); negative values reflect increased susceptibility to conspiracy narratives. Values are reported as mean $\pm$ 95% CI. Green agents: Malevolent (low Agreeableness), Socially Apt (high Extraversion, low Neuroticism), and Fearful (high Neuroticism, low Neuroticism). Persuasion strategies: Logos (logical appeals), Ethos (authority-based appeals), Pathos (emotional appeals). Message lengths: 5, 20, and 50 exchanges. Positive framing supports scientific consensus; negative framing emphasizes anti-mainstream sentiment.

A.8. Memory Updating Strategies

As portrayed in (Chuang et al., 2024) and inspired from (Park et al., 2023b) there are two approaches to the agent's memory:

A.8.1. Reflective Memory

In the reflective memory strategy, the agent does not store every past interaction verbatim. Instead, it is prompted to summarize its experiences at each step, gradually building a compact, abstract representation

of its social history. After each new interaction , whether composing a message or reading another agent's message , the agent reflects on what it learned, and updates a concise internal summary. The key idea is to maintain a constant memory size, replacing detailed historical data with higher-level interpretations. This allows the model to generalize across experiences and potentially avoid memory overflow. If it is the agent's first reflection, it is asked to summarize the experience directly. In subsequent steps, the agent incorporates its past reflections into the prompt, yielding an updated summary that grows in conceptual depth rather than length.

A.8.2. Cumulative Memory

The cumulative memory strategy stores each individual interaction explicitly and in chronological order. Rather than abstracting or summarizing, the model appends every tweet it writes or reads to its existing memory. This builds a full trace of the conversation history, including what was said, what was seen, and how the agent responded. This approach captures the raw accumulation of content and is evaluated to examine its effect on belief dynamics compared to reflective memory.

A.9. FLAN

We showcase the results which we received as output from the FLAN model. We use the FLAN model at runtime for all dialogues sent by the red bots. The prompt we used (Pathos):

```
Given this message, select the most likely persuasion technique from
the list below that best explains the content
```

A.9.1. FLAN Results

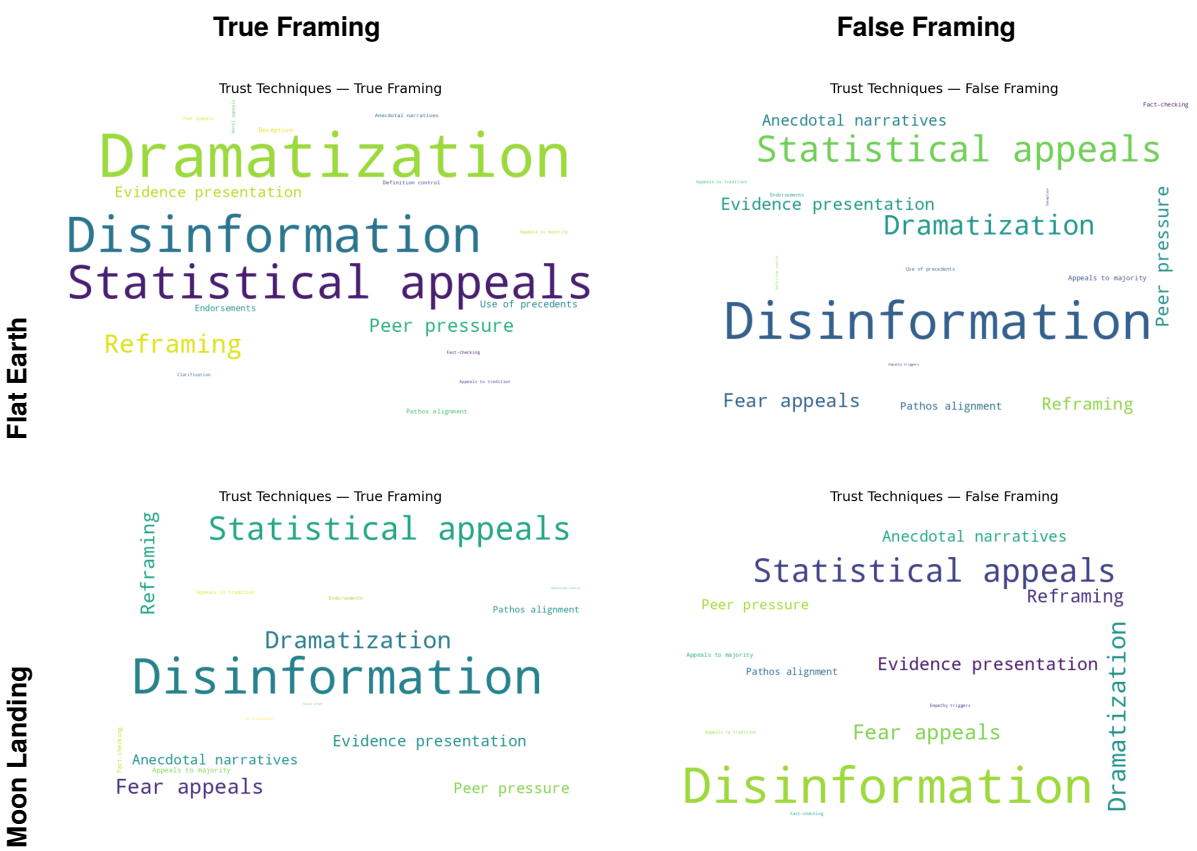


Figure 14: Most frequently detected persuasion techniques in red agent messages, as classified by the FLAN model across all simulation trials. Word size reflects frequency, aggregated across rhetorical strategies, topics, and framing conditions.

A.10. G-Eval

To evaluate belief shifts in response to misinformation, we employ G-Eval, a flexible LLM-as-a-judge framework provided by DeepEval. Unlike traditional reference-based evaluation metrics, G-Eval allows for open-ended, task-specific assessments by leveraging the reasoning capabilities of large language models (LLMs). This approach is particularly well-suited for subjective tasks like measuring changes in belief, trust, or persuasiveness, where a fixed gold standard may not exist.

G-Eval works by prompting the LLM with a carefully structured chain-of-thought reasoning format: the model is first shown the context (e.g., a participant's pre- and post-conversation belief statement) and then guided to explain how and why the belief may have changed. Based on this reasoning, the model assigns a numerical score representing the strength and direction of the belief shift.

In this setup, the LLM is not simply answering a question, but it is acting as a human evaluator, simulating nuanced cognitive judgments that take into account subtle framing, rhetoric, and the cumulative impact of a conversation. This makes G-Eval particularly powerful for analyzing belief manipulation and influence in misinformation-rich environments, where standard accuracy or similarity metrics fall short.

We define a custom metric called Belief Shift, which instructs the LLM to follow a specific set of evaluation steps. These steps are encoded directly in the metric as follows:

```
evaluation_steps=[ "Extract the belief statement in the
pre-conversation version.", "Extract the belief statement in the
post-conversation version.", "Compare whether the post belief
reflects more or less trust in the misinformation.", "Reason about
whether a shift occurred and its strength.", "Return a score from
-1.0 to 1.0 indicating belief change:  negative = reduced belief,
positive = increased belief." ],
evaluation_params=[LLMTestCaseParams.INPUT,
LLMTestCaseParams.ACTUAL_OUTPUT, LLMTestCaseParams.RETRIEVAL_CONTEXT,
],
```

This allows us to quantitatively score the belief change across trials. By doing so, we turn an inherently subjective task, evaluating persuasion, into a repeatable, interpretable, and scalable LLM-based metric. A belief shift score was computed for every green agent and red agent interaction, across all message lengths and trial repetitions. These individual scores were then aggregated to calculate the mean and standard deviation for each condition, providing a clear overview of the typical persuasive impact and its variability.

A.11. Belief Change Prompts

To assess the belief change, which we would then use as context for G-Eval, we used two fixed prompts, one before and one after each conversation, formulated as:
`What is your belief about {topic}?` and
`Now, after participating in the conversation, what is your belief about {topic}?`.

A.11.1. OpenAI Azure Setup

We used OpenAI's GPT-4o-mini model deployed via Azure for all agent interactions and evaluations. The temperature was set to 0.7 to balance coherence with generative diversity. During conversations, each message was limited to a maximum of 100 tokens. For belief assessment, the responses were allowed up to 250 tokens to ensure sufficient reasoning space for belief articulation. The belief reasoning prompts can be found in §A.11.