Overstating Attitudes, Ignoring Networks: LLM Biases in Simulating Misinformation Susceptibility

Eun Cheol Choi[1,2], **Lindsay E. Young**[1], **Emilio Ferrara**[1,2]
[1]Annenberg School for Communication and Journalism, University of Southern California
[2]Thomas Lord Department of Computer Science, University of Southern California
{euncheol, lindsay.young, emiliofe}@usc.edu

Abstract

Large language models (LLMs) are increasingly used as proxies for human judgment in computational social science, yet their ability to reproduce patterns of misinformation susceptibility remains unclear. We evaluate whether LLM-simulated survey respondents replicate human patterns of misinformation belief and sharing. Using participant profiles from three online surveys that include network, demographic, attitudinal, and behavioral features, we prompt LLMs to simulate survey responses to misinformation items and compare the results to human data on distributions and associations. LLM simulations capture broad distributional tendencies and show modest correlation with human responses. However, they systematically overstate the association between belief and sharing. Linear models fitted to simulated responses show substantially inflated explained variance compared to those fitted to human data. They also place disproportionate weight on attitudinal and behavioral features while largely ignoring personal network characteristics, a pattern not observed in human data. Analyses of LLM training data and model-generated reasoning paths suggest that these distortions reflect systematic biases in how misinformation-related concepts are represented. Our findings indicate that LLM-based survey simulations are better suited for diagnosing systematic deviations from human judgment than for substituting for it.

Code — https://github.com/EunCheolChoi0123/llm-simulating-misinformation-susceptibility

Introduction

With the rapid advancement of large language models (LLMs), scholars have increasingly explored their use as simulated agents in social network settings, including scenarios of misinformation spread (Pastor-Galindo, Nespoli, and Ruipérez-Valiente 2024; Yang et al. 2024). Historically, research focused on simple automated scripts or social bots (Ferrara et al. 2016). Earlier misinformation research has often turned to agent-based models or epidemic-style diffusion frameworks to simulate how falsehoods spread across populations (Tambuscio and Ruffo 2019; Gausen, Luk, and Guo 2021). While powerful, these approaches have been criticized for limited ecological validity: agents typically follow simplified heuristics, and their behaviors are modeled in stylized rather than psychologically grounded ways (Chuang et al. 2024). The appeal of LLMs is that, when conditioned on prompts describing personal traits and contexts, they may generate responses that approximate human judgments, incorporating demographic characteristics, attitudes, and situational cues (Törnberg et al. 2023; Mou, Wei, and Huang 2024; Chang et al. 2025).

Yet, it remains unclear whether such LLM-based agents recover the empirical relationships between misinformation susceptibility (belief in false claims and willingness to share them) and key social predictors. Prior research with human subjects shows that susceptibility is shaped not only by demographic, cognitive, and attitudinal attributes (Nan, Wang, and Thier 2022; Ecker et al. 2022), but also by the composition and structure of *personal networks* (Facciani and Steenbuch-Traberg 2024). If LLM agents fail to reproduce these relationships, simulations of misinformation dynamics may overlook or mischaracterize key drivers. However, only a handful of studies have conducted LLM simulations of misinformation susceptibility (Ma et al. 2024; Pratelli and Petrocchi 2025; Dash et al. 2025; Plebe et al. 2025), and these have focused primarily on sociodemographic predictors while largely neglecting network features.

Contributions of This Work

This study addresses a key gap in LLM-based social simulations by systematically evaluating whether LLMs reproduce patterns associated with both network-based and individual-level factors related to misinformation susceptibility. Using three distinct online survey datasets as ground truth, we directly compare LLM-based survey simulations against human responses to address the following research questions:

RQ1: *To what extent do LLMs approximate the empirical distributions of human misinformation belief and sharing?*

RQ2: *To what extent do LLMs recover the empirically observed structure of associations among social predictors and misinformation susceptibility?*

RQ3: *Do LLMs systematically differ from human data in the estimated effects of specific features on misinformation susceptibility?*

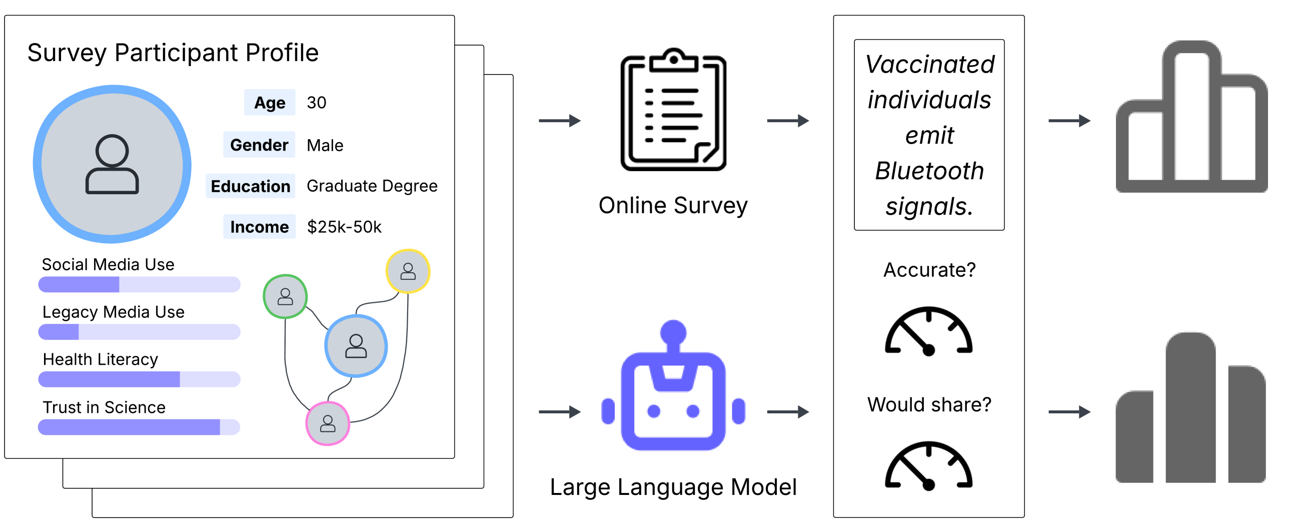

Figure 1: Study pipeline. Human survey participants provided demographic, attitudinal, behavioral, and personal network information through online surveys. Structured profiles containing the same information were provided to large language models (LLMs). Both human participants and LLMs evaluated the same set of false claims across domains and contexts, reporting their beliefs and sharing intentions. We directly compared the resulting outcome distributions and estimated predictor–outcome relationships from both sources, aiming to assess how closely LLM-generated responses mirror human response patterns and to identify systematic differences between simulated and empirical results.

Related Works

Misinformation Susceptibility

A large body of survey research has examined individual differences in misinformation susceptibility. Demographic factors such as age, gender, race/ethnicity, political identification, and education have all been linked to variation in susceptibility (Nan, Wang, and Thier 2022; Ecker et al. 2022). Cognitive and attitudinal factors, including trust in science, literacy, and trust in media, also influence whether individuals accept or reject misinformation. Network dynamics are also related to susceptibility: discussions with peers can reinforce misinformation through interpersonal validation and repeated exposure (Jia and Zhang 2025). Thus, it is important to move beyond examining susceptibility in isolation and consider the broader social environment. Network composition has been shown to be a key predictor of both belief in and sharing of false headlines (Facciani and Steenbuch-Traberg 2024).

Misinformation susceptibility is multi-dimensional: not everyone exposed to misinformation believes or shares it (Altay, Berriche, and Acerbi 2023; Luceri et al. 2025); belief and sharing are tightly correlated, yet they are conceptually and empirically distinct (Nan, Wang, and Thier 2022; Ecker et al. 2022; Sirlin et al. 2021). Notably, sharing does not always signal belief: one study found that 16% of headlines were shared despite being judged inaccurate (Pennycook and Rand 2021). Recent work shows that sharing is often driven less by mistaken beliefs than by social and emotional motives—for example, identity expression, entertainment, or strategic signaling) (Paletz, Auxier, and Golonka 2018; Altay, De Araujo, and Mercier 2022; Lee and Jang 2023). Accordingly, treating belief and sharing as separate targets and evaluating both is essential to better capture the nuances of misinformation susceptibility.

Social Simulation with LLMs

Research on social simulation with LLMs spans two broad directions. A first line of work focuses on *multi-agent interactions*, where agents (i) freely interact with one another (Park et al. 2022, 2023), (ii) interact within social networks under constraints or restricted action spaces (Ashery, Aiello, and Baronchelli 2025; Chang et al. 2025; Chuang et al. 2024), or (iii) specifically operate in social media environments (Yang et al. 2024; Törnberg et al. 2023; Mou, Wei, and Huang 2024; Gao et al. 2023; Orlando et al. 2025). These studies typically explore emergent behaviors (Park et al. 2023; Orlando et al. 2025) or evaluate how global and temporal distributions of simulated actions compare with empirical traces from social media platforms (Yang et al. 2024; Chang et al. 2025).

A second line of work embeds LLMs in *lab-style experimental or social survey settings*. Here, models are prompted to act as participants in controlled experiments (Gui and Toubia 2023), respond to survey instruments (Yu et al. 2024), or combine both approaches (Park et al. 2024). Evaluation metrics include alignment with real-life outcome distributions (Wang et al. 2024), correlations among core variables (Park et al. 2024; Salecha et al. 2024), recovery of detailed associations using regression or structural equation models (Ma et al. 2024; Kim et al. 2024), and tasks framed as causal inference problems, such as treatment effects and structural recovery (Hewitt et al. 2024; Gui and Toubia 2023; Manning, Zhu, and Horton 2024). Other studies extend this paradigm to forecasting study results (Lippert et al. 2024) or automating annotation tasks in NLP research (Hu and Collier 2024).

Despite this promise, several recurring challenges have been identified. One study identified five recurring issues in LLM social simulations: lack of human *diversity*; systematic *biases* in simulated outcomes; *sycophancy*, or user-pleasing outputs; *alienated* outputs that appear accurate but fail to reflect underlying social processes; and limited *generalizability* (Anthis et al. 2025). Additional concerns include assigning disproportionate weight to predictors that are overrepresented in pretraining corpora (Chang et al. 2025). Our study contributes by incorporating *diverse* profiles derived from real-world surveys and by systematically examining deviations and potential *biases* in simulated outcomes.

LLM-Simulated Misinformation Susceptibility

A subset of recent research has examined how sociodemographic and psychological factors shape LLM-simulated misinformation susceptibility. For example, one study tested whether "life-story" profiles encoding demographic features (e.g., age, gender, background traits) influenced simulated judgments of misinformation accuracy and sharing (Ma et al. 2024). Another incorporated Big Five personality traits into prompts to evaluate their effects on simulated news discernment (Pratelli and Petrocchi 2025). Other studies have incorporated psychological mechanisms and modalities, such as motivated reasoning (Dash et al. 2025) and susceptibility to visual misinformation (Plebe et al. 2025). One study compared the persuasive impact of misinformation in human–LLM dyads, pairing human participants from diverse demographic backgrounds with LLMs prompted using equally diverse demographic personas (Borah, Mihalcea, and Pérez-Rosas 2025).

These studies suggest that LLM simulations can reproduce certain psychological regularities, but they also raise concerns: models may place disproportionate emphasis on well-documented features while giving less weight to less codified factors. As a result, LLM simulations may be more effective at reproducing established patterns than at revealing novel drivers of susceptibility. Our study extends this line of work by introducing *personal network features*, offering a comprehensive test of whether LLM simulations capture the social bases of misinformation susceptibility.

Methodology

Dataset

Domain	Country	Platform	N	Year	Misinfo
Public Health	US	Prolific	486	2023	5 items
Climate Change	US	Prolific	377	2025	3 items
Pandemic Politics	KR	Embrain	708	2020	5 items

Table 1: Overview of analytic samples from social surveys.

We draw on three survey datasets designed to measure *misinformation susceptibility* alongside a range of *individual- and network-level features* across three domains: public health, climate change, and politically charged issues. Table 1 summarizes the key characteristics of each dataset.

The **public health** dataset (Choi and Young 2026) was collected online in 2023 from a U.S.-based sample and focuses on five misinformation items related to public health topics, including vaccines and chronic diseases. The **climate change** dataset (Huang et al. 2026; Kim et al. 2026) was collected in 2025 from a U.S.-based online sample as a part of a larger project, and we use the relevant subset of survey modules. This survey focused on misinformation related to climate change and included three image-based stimuli presented as memes. For the purposes of this study, these meme stimuli were transformed into text descriptions prior to analysis. The original images are in Appendix A. The **pandemic politics** dataset (Ihm and Lee 2023; Lee, Lee, and Hwang 2023) was collected in South Korea in 2020 using an online panel as a part of a larger project. In this paper, we focus our analysis on five politically charged misinformation items circulating during the COVID-19 pandemic, including claims involving anti-establishment narratives, nationalist or Sinophobic framing, and election-related denial narratives. Across the three studies, estimated hourly compensation was approximately \$10 to \$30, reflecting differences in survey length and payment structures. Average survey completion time ranged from 15 to 20 minutes. These surveys resulted in a total participant compensation cost of approximately \$15,000.

Because the three surveys were designed for different research purposes, the sets of measured variables and the scales used to operationalize them differ across datasets. Nevertheless, the variables examined in this study map onto a shared conceptual framework and can be meaningfully grouped into four broad categories. See Appendix A for the detailed breakdown of the variables.

- **Misinformation susceptibility:** Participants reported their belief in and their willingness to share each of the false claims with others.
- **Personal network measures:** Name-generator and name-interpreter modules captured participants' discussion partners, alter attributes, and/or alter–alter ties, allowing construction of personal network measures.
- **Demographic measures:** Standard demographic information including age, gender, race/ethnicity, education, income, and/or region.
- **Attitudinal and behavioral measures:** Self-reported attitudes and behaviors, including political leaning, trust in science (Sturgis, Brunton-Smith, and Jackson 2021), trust in social media (Obadă and Dabija 2022), traditional media use (Lee 2009), social media use, and/or health literacy (Chew et al. 2008).

Together, these measures provide candidates for *multi-dimensional correlates of susceptibility*, incorporating demographic, attitudinal/behavioral, and personal network factors. The dataset serves as an *empirical reference* against which we compare LLM-simulated responses.

SYSTEM	This survey was conducted in the United States in Oct 2023. Suppose you are a survey participant described below, adopting their demographics, attitudes, behaviors, and personal network. Based ONLY on the profile, answer the given question. If the profile lacks information, make your best estimate from available signals; never reply with "unknown" or similar. Output format (JSON): {"response": "integer"}
INPUT	Participant profile: {*personal network information*} {*demographic information*} {*attitudes and behaviors*} Consider the following claim: "Vaccinated individuals emit Bluetooth signals." Question: How would you describe the above claim? (1=Inaccurate, 7=Accurate) *OR* Question: I am willing to share this information with my friends or family. (1=Strongly disagree; 7=Strongly agree)

Figure 2: Example prompt format used to simulate public health survey participants' misinformation belief or sharing intention. Belief and sharing questions were queried separately using identical system prompts, participant profiles, and claims.

Experiment Procedure

We prompted large language models (LLMs) to impersonate as *synthetic survey respondents* (Figure 1; see Appendix C for prompts used for other survey domains). Each prompt contained a structured profile describing a survey respondent's individual attributes and egocentric network features (Figure 2; see Appendix D for full profile templates). The profile information was organized into three conceptual blocks: personal network attributes, demographic characteristics, and self-reported attitudes/behaviors. Missing values in survey variables were imputed using median imputation. Models were instructed to answer misinformation belief and sharing intention items based solely on the provided profile, following the same response format as the original human survey.

To account for the stochastic nature of LLM generation and potential sensitivity to prompt design, we implemented two robustness variations: *alternative variable ordering* and *prompting with composite scores*. First, we varied the ordering of profile blocks within the prompt to assess whether response patterns varied with presentation order. Specifically, while the network feature block was presented first in the original prompt format, the alternative ordering placed it last. Second, instead of providing raw item-level responses for each construct, we experimented with supplying composite scores in the profile that summarized each variable, allowing us to assess the robustness of model responses to alternative representations of the same underlying constructs (see Appendix E for composite score profile templates). Using composite scores in prompting introduces a cost trade-off: while additional preprocessing by researchers shortens prompts and lowers inference costs, it may obscure or omit important contextual information present in the original inputs.

We evaluated a diverse set of LLMs spanning multiple model sizes, architectural families, provider nationalities, levels of content moderation, and degrees of openness. This set included both *reasoning models* and *chat/instruction-tuned models*, reflecting different training objectives and usage paradigms in modern LLM systems. In addition, for chat models, we experimented with explicit chain-of-thought elicitation by adding a dedicated reasoning field to the JSON output to examine whether encouraging intermediate reasoning altered response distributions or the reliance on variables. Models were accessed through official provider APIs or standardized inference services, depending on availability.

Analytic Procedure

For each respondent profile, we generated simulated survey responses using LLMs and directly compared them to the human survey responses conditioned on the same profile information, enabling profile-level comparisons between simulated and empirical outcomes. We assessed whether LLM outputs (i) reproduced the distribution of misinformation belief and sharing, (ii) approximated the observed relationships between profile features and misinformation susceptibility measures, and (iii) exhibited systematic bias patterns across subgroups.

Our analysis proceeds in six steps to evaluate how closely LLM-simulated survey responses align with human data and identify potential sources of systematic divergence.

1. **Distributional Similarities (RQ1).** We compare the distributions of misinformation belief and sharing between ground truth (human responses) and LLM-simulated results using Jensen–Shannon divergence (JSD) and earth mover's distance (EMD).
2. **Correlation between Ground Truth and Simulated Responses (RQ1).** We compute ranked correlations (Spearman's ρ) between ground truth and simulated responses to quantify the extent to which LLM-generated outputs align with human responses.
3. **Correlation between Simulated Belief and Sharing. (RQ2)** We assess whether LLM simulations exhibit stronger or weaker associations between misinformation belief and sharing than in human data. Although these

constructs are strongly related, they remain theoretically and empirically distinct.

4. **Predictive Modeling of Simulated Susceptibility. (RQ2)** We fit Elastic Net regression models to predict simulated misinformation belief and sharing outputs from demographic, cognitive, behavioral, and network features. Elastic Net regularization allows us to account for multicollinearity across correlated predictors while retaining interpretability at the block and variable levels. This analysis enables us to (i) assess the relative predictive contribution of predictor blocks, (ii) evaluate whether simulated outputs place disproportionate predictive weight on certain types of features, and (iii) compare the magnitude and direction of effects against ground-truth models.
5. **Simulation Bias in Predictor Effects (RQ3).** To test whether LLM outputs systematically differ in the estimated relationships among profile features and misinformation outputs, we estimate pooled Elastic Net models that include an indicator for simulated responses and their interactions with key predictors. Interaction terms capture whether the association between a given predictor and misinformation outcomes differs in magnitude or direction between LLM-simulated and human responses.
6. **Model Reasoning and Training Corpus Analysis (RQ3).** Finally, we conduct exploratory analyses of LLM chain-of-thought processes and training data to explore possible contributors to systematic patterns in simulated responses. For chat-oriented models, we analyze chain-of-thought (CoT) rationales generated alongside survey responses to examine how demographic, attitudinal, behavioral, and network variables are invoked and linked to misinformation belief and sharing decisions. In parallel, we analyze an open-source training corpus using OLMo-Trace (Liu et al. 2025), a tool developed by the Allen Institute for AI (AI2) that traces relevant spans in the pre- and post-training corpus. This allows us to examine how variables of interest co-occur and are framed in the training data. Together, these analyses provide preliminary descriptive evidence regarding whether observed simulation patterns co-occur with model reasoning traces or patterns present in pre- and post-training data.

Results

Descriptive statistics for variables derived from human participants' responses are reported in Appendix A. The rank correlations (Spearman's ρ) between misinformation belief and sharing were .606 for public health, .418 for climate change, and .342 for political misinformation (Table 2), consistent with the prior works that while the two susceptibility measures are closely related, they remain empirically distinct (Pennycook and Rand 2021). Elastic Net models fitted to profile features as predictors and misinformation belief and sharing as outcomes explained modest variance (cross-validated R^2 ranging from .042 to .228; Table 3), suggesting that demographic, attitudinal, behavioral, and network characteristics are associated with misinformation belief and sharing, albeit with substantial noise, as is typical in social science data.

RQ1: *To what extent do LLMs approximate the empirical distributions of human misinformation belief and sharing?*

LLM outputs exhibit partial alignment with broad distributional patterns of human misinformation susceptibility and exhibit moderate correlation with human responses. However, distributional similarity and correlation strength vary across models, prompting formats, and survey contexts, with no single configuration consistently performing best.

To answer RQ1, we computed distributional divergence and rank-order agreement between ground-truth and LLM-simulated misinformation susceptibility across survey domains, model architectures, and prompting formats (Appendix B). Across most of the settings, LLM-simulated outputs are moderately correlated with human responses, though the degree of divergence and correlation varies by domain, model, and prompt configuration, and no single model or prompting strategy consistently outperforms others across metrics.

Given the absence of a dominant configuration, and to reduce model-specific noise while preserving consistent differences observed across prompting formats, we aggregate LLM-simulated outputs at the prompting-format level in subsequent analyses, rather than identifying a best-performing system. For each survey domain, we average misinformation measures across all models within each prompting format, yielding three aggregated outputs per belief and sharing outcome.

RQ2: *To what extent do LLMs recover the empirically observed structure of associations among social predictors and misinformation susceptibility?*

Although LLM outputs capture broad associations between profile features and misinformation outcomes, they exhibit consistent deviations from empirical patterns, including (i) an exaggerated coupling between belief and sharing, (ii) inflated explained variance, and (iii) disproportionate emphasis on attitudinal and behavioral predictors while underweighting network-based factors.

First, we focus on the correlation between two forms of misinformation susceptibility, belief and sharing. As shown in Table 2, LLM-simulated responses exhibit near-monotonic correlations between belief and sharing across

	Health	Climate	Political
Ground Truth	.606	.418	.342
LLM-Simulated			
Original Prompt	.925	.956	.749
Alternative Ordering	.930	.957	.773
Composite-Score	.898	.953	.742

Table 2: Rank correlation (ρ) between misinformation belief and sharing across domains.

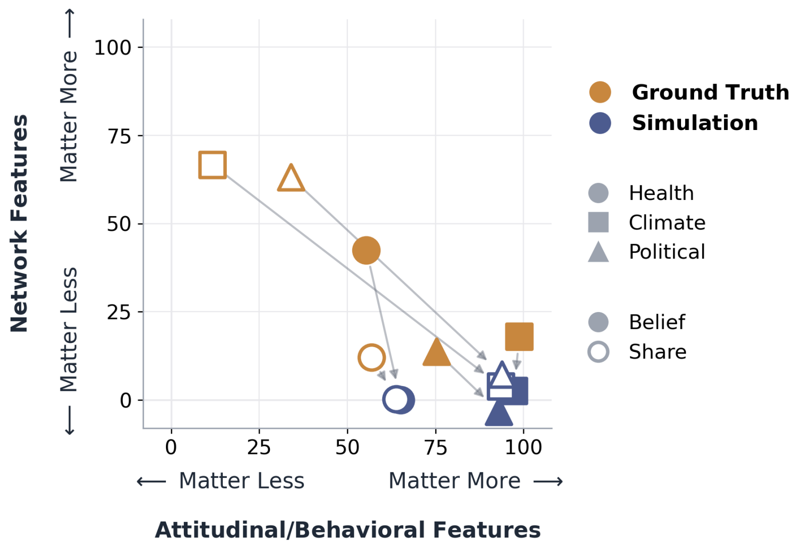

Figure 3: Simulation distorts the structure of feature importance relative to ground truth. Each point shows the proportion of predictive power lost when a feature block is removed: $1 - R^2_{\text{block removed}}/R^2_{\text{full model}}$ (in %); higher values indicate that a block matters more. The x-axis shows the importance of attitudinal/behavioral feature blocks and the y-axis the importance of network feature blocks, computed separately for ground-truth and simulated outcomes. Arrows connect each ground-truth point to its simulated counterpart. Values are averaged across three prompt specifications, and all R^2 values are five-fold cross-validated.

domains, substantially exceeding those observed in human data. Thus, LLM outputs treat belief and sharing as much more tightly coupled than human responses do, attenuating the empirical dissociation between the two measures.

Next, we examine the extent to which LLM outputs reflect the empirical structure of associations linking profile features to misinformation susceptibility. To this end, we fit Elastic Net regression models predicting both misinformation belief and sharing from demographic, attitudinal/behavioral, and network variables, and compare models trained on LLM-simulated outputs with those trained on human ground-truth responses. Across all domains, models fitted on LLM-simulated responses achieve substantially higher R^2 values than those fitted on human data, often by an order of magnitude (Table 3). This pattern indicates that simulated responses exhibit much stronger alignment with the provided profile features than is observed in human data, yielding deterministic mappings uncharacteristic of human belief or behavior.

Figure 3 examines the contribution of each feature block by removing blocks from the full model and measuring the resulting drop in explained variance. This procedure serves as an ablation-like check, isolating the extent to which each feature block contributes to the predictive signal. In human data, both attitudinal/behavioral and network features carry substantial predictive signal for belief and sharing outcomes across all domains, indicating that misinformation susceptibility reflects both individual dispositions and network context.

LLM-simulated data, however, provide a systematically distorted picture. Attitudinal/behavioral features dominate prediction in every condition, while removing network features produces little or no loss relative to the full model; a pattern absent in human data. Simulated responses thus exhibit an individual-centric structure: the simulation *overstates* the role of attitudes and behaviors in shaping susceptibility while largely *ignoring* the social network. Full per-prompt results are reported in Appendix F.

	Health		Climate		Political	
	Belief	Sharing	Belief	Sharing	Belief	Sharing
Ground-Truth	.069	.191	.102	.205	.228	.042
LLM-Simulated						
Original	.831	.862	.759	.752	.740	.580
Alt. Order	.810	.874	.741	.735	.751	.592
Composite	.687	.808	.722	.734	.736	.703

Table 3: Five-fold cross-validated out-of-sample explained variance (R^2) for ElasticNet regression models across domains and outcomes. Models predict misinformation beliefs or sharing intentions from human survey data or LLM-simulated responses, using demographic, attitudinal/behavioral, and network features.

RQ3: *Do LLMs systematically differ from human data in the estimated effects of specific features on misinformation susceptibility?*

LLM-simulated responses exhibit structured differences in feature effect estimates, with consistently larger deviations for salient attitudinal and behavioral predictors such as trust in science, political leaning, and social media use, relative to human data. Exploratory analyses of model reasoning traces and open-source pre-/post-training corpus indicate that these predictors are frequently empha-

sized in both reasoning outputs and training data contexts, which may help contextualize the observed differences.

To address RQ3, we directly test whether LLM responses systematically over- or under-estimate susceptibility associated with specific profile features. We concatenate human and LLM-simulated responses into a pooled dataset and include an indicator variable denoting simulated observations. For each domain and outcome, we fit Elastic Net regression models that include main effects for all predictors, the simulation indicator, and interaction terms between the simulation indicator and each predictor. These interaction terms capture whether the estimated association between a given feature and misinformation belief or sharing differs between simulated and human data, holding all other features constant.

Figure 4 reports standardized coefficients for the simulation-by-feature interaction terms. Across domains and outcomes, the largest interaction effects in absolute magnitude are concentrated among attitudinal and behavioral features, including trust in science, political leaning, and social media use, which are commonly cited predictors of misinformation susceptibility in prior literature (Nan, Wang, and Thier 2022). Overall, these patterns suggest that deviations between simulated and human responses are not evenly distributed across predictors, but are more pronounced for a subset of individual-level attitudinal and behavioral features. Relative to human data, simulated responses tend to place greater weight on these salient predictors, while differences associated with network features are generally smaller in magnitude.

To contextualize the patterns observed in RQ3, we conducted post hoc exploratory analyses of the model's reasoning and training data. We first examine model reasoning processes by analyzing chain-of-thought (CoT) rationales generated during misinformation judgments. Our dataset includes 42,583 unique reasoning chains generated by both open-weight reasoning models and chat models prompted with explicit CoT instructions. We then examine whether related regularities are observable in LLM training data. While the training corpora of major proprietary models are not publicly accessible, we analyze open-source pre- and post-training data using OLMoTrace (Liu et al. 2025), a tool that retrieves up to ten training-data spans most relevant to a given query. Using this approach, we retrieve 187 unique text spans that explicitly reference relationships between misinformation and profile features. The full analytical procedures are documented in Appendix G.

Descriptive analyses of reasoning traces and open-source training corpus point to similar patterns. LLMs frequently invoke attitudinal and behavioral predictors in their reasoning traces, most notably trust in science, political leaning, social media use, and health literacy, while network-related attributes appear less frequently (Figure 5). Furthermore, analysis of the training corpus shows that features such as trust in science, social media use, and health literacy frequently co-occur with misinformation-related concepts in non-uniform, directional patterns (Figure 6).

Discussion

This study evaluates how well large language models (LLMs) approximate patterns of human susceptibility to misinformation when prompted with rich survey profiles that include demographic, attitudinal, behavioral, and personal network information. Across three domains and multiple prompting formats, we find that LLMs reproduce coarse-grained distributional patterns in misinformation belief and sharing and generate responses that are moderately correlated with human judgments. At the same time, substantial and systematic discrepancies emerge beyond this surface alignment. Taken together, these results indicate that LLM-based simulations may be informative for exploratory or diagnostic purposes (Dillion et al. 2023; Wu et al. 2025), but they are limited as stand-ins for human data, especially when the relational structure of social behavior is of central interest.

Our findings reveal systematic divergences that limit the validity of LLMs as substitutes for human data. First, LLM responses substantially exaggerate the association between misinformation belief and sharing. As a result, these two theoretically and empirically distinct dimensions of susceptibility appear far more tightly coupled in simulated data than in human data. In human data, belief and sharing are correlated but separable, reflecting the fact that individuals may share content for social, emotional, or strategic reasons even when they do not fully believe it (Ecker et al. 2022). By reducing this differentiation, LLM simulations may obscure distinctions that are empirically present in human behavior, with implications for how misinformation dynamics are modeled in downstream analyses.

Second, regression models fitted on LLM-simulated outcomes exhibit dramatically inflated explained variance compared to models fitted on human responses. This pattern indicates that LLM-generated responses are more tightly aligned with the provided profile features than is observed in human data. Rather than reflecting the substantial variability and residual noise that characterize human beliefs and behaviors, simulated responses appear comparatively more regular and predictable given the same inputs (Kaiser et al. 2025; Anthis et al. 2025). As a result, the relationships estimated from simulated data may overstate the degree to which observed features account for misinformation belief and sharing in real populations.

Crucially, this over-structuring is not uniform across predictors. In human data, both attitudinal/behavioral and personal network features contribute meaningfully to explaining misinformation susceptibility. In contrast, LLM simulations place substantially greater weight on attitudinal and behavioral variables, while assigning comparatively less explanatory importance to network features. As a result, LLM outputs tend to yield larger estimated associations for well-documented attitudinal and behavioral predictors, accentuating familiar individual-level explanations, while producing weaker or less stable associations for less codified network-related factors.

Our exploratory analyses of model reasoning and training data provide suggestive evidence regarding patterns that co-occur with the observed distortions. LLMs more frequently

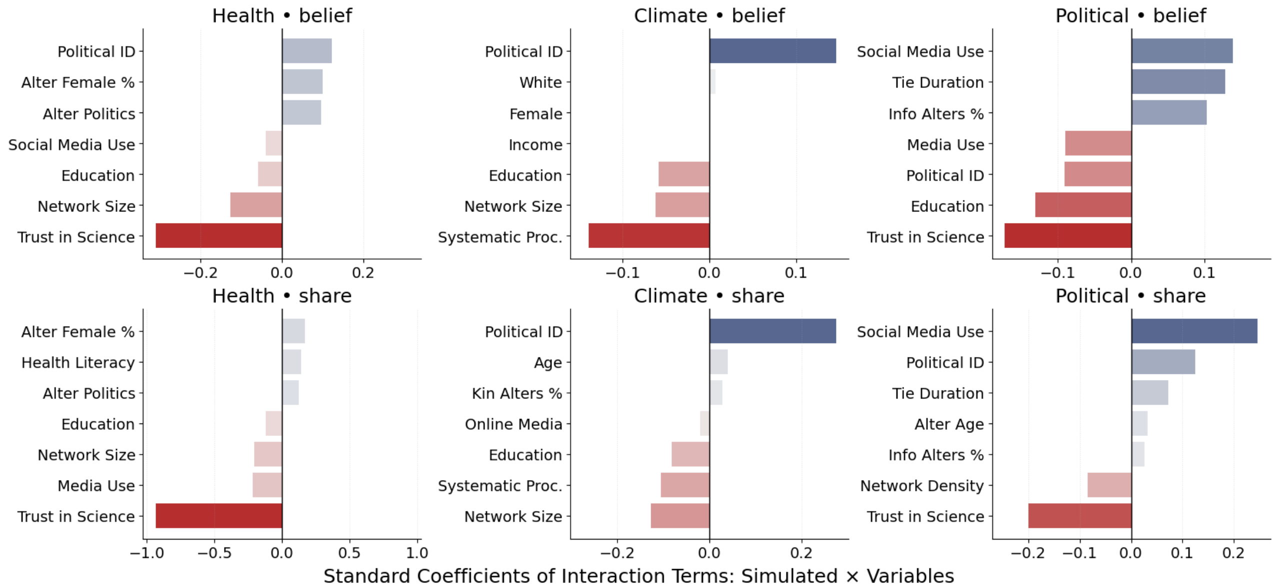

Figure 4: Standardized coefficients for simulation-by-feature interaction terms (Simulated $\times$ Variables) across domains and outcomes. Shown are the seven variables with the largest absolute interaction coefficients. Each coefficient captures the difference in the estimated feature–outcome association between the simulated and ground-truth data (i.e., the change in slope from the simulated to the ground-truth data). The vertical line at zero indicates that the simulated and ground-truth estimates are identical.

reference attitudinal and behavioral variables when generating reasoning traces for susceptibility judgments, with political leaning and trust in science referenced most frequently across domains. By comparison, network-related attributes are referenced relatively infrequently in these reasoning traces. Taken descriptively, this pattern indicates that LLM-generated responses are more often accompanied by references to salient individual-level predictors that are prominent in prior misinformation research (Nan, Wang, and Thier 2022).

Complementing this observation, our analysis of an open-source training corpus shows that several of these same attitudinal and behavioral variables frequently co-occur with misinformation-related concepts in the retrieved text spans. While these analyses do not establish a causal link, the convergence between reasoning traces and training data resonates with concerns about LLMs' reliance on statistical stereotypes overrepresented in their training data (Kotek, Dockum, and Sun 2023; Anthis et al. 2025). More broadly, these findings align with ongoing discussions about fairness, representation, and the potential for generative AI systems to reproduce societal stereotypes (Ferrara 2024).

Taken together, our study suggests that LLM-based survey simulations do not simply differ from human responses due to random variation. Instead, they exhibit patterned deviations that are consistent with differences observed in both reasoning traces and training data associations. In particular, LLM outputs tend to be more strongly associated with features that are well documented and frequently discussed in existing literature, while placing comparatively less emphasis on relational and contextual factors that are less explicitly represented in a text-based corpus. This imbalance is consistent with prior observations that deep neural network models may rely more heavily on readily available or salient cues (Geirhos et al. 2020).

Unlike attitudes or demographics, personal network features are inherently relational, higher-order, and often non-narrative. They require combining information across multiple entities and ties rather than mapping a single attribute to an outcome, a representational challenge that has been shown to benefit from explicit relational inductive biases (Battaglia et al. 2018). This representational mismatch may help explain why network features appear less influential in simulated outputs, even when such information is available in the prompt. Consistent with this interpretation, our robustness checks indicate that surface-level prompt modifications alone are unlikely to substantially alter these patterns. Addressing such limitations may therefore require architectural or representational approaches that more directly encode relational structure, rather than relying exclusively on text-based descriptions (Tang et al. 2024; Chen et al. 2024).

Our results have important implications for the use of LLMs in social simulation. While LLMs may be useful for examining how well-established predictors are reflected in model outputs, they should not be treated as interchangeable proxies for human respondents, particularly in studies where social context and network structure play a central role. In such settings, reliance on LLM-based simulations may yield overly regularized relationships that place disproportionate emphasis on individual-level factors while underrepresenting social interaction, diffusion, and peer influence. More broadly, our work highlights the need for evaluation frame-

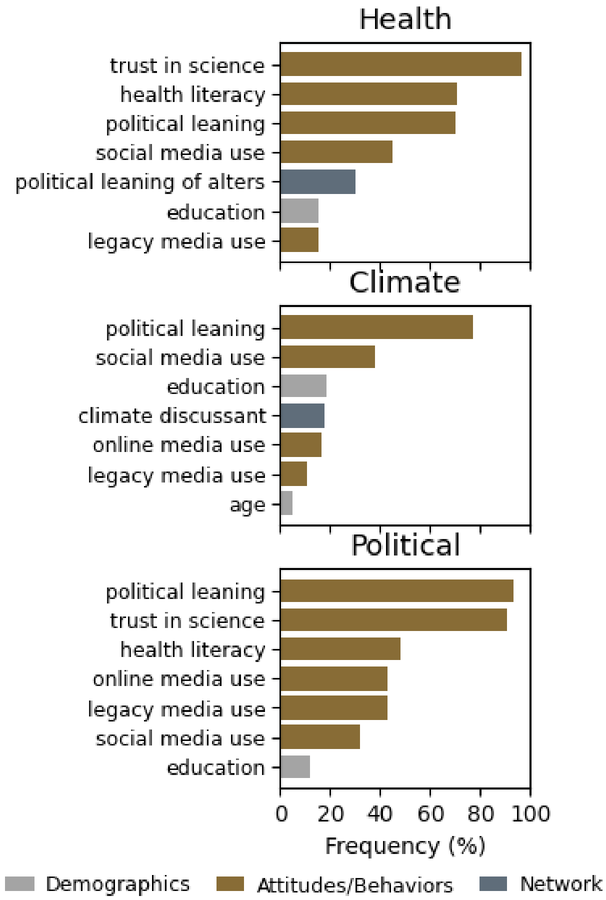

Figure 5: Frequency of the seven most frequently referred variables across reasoning chains by domain.

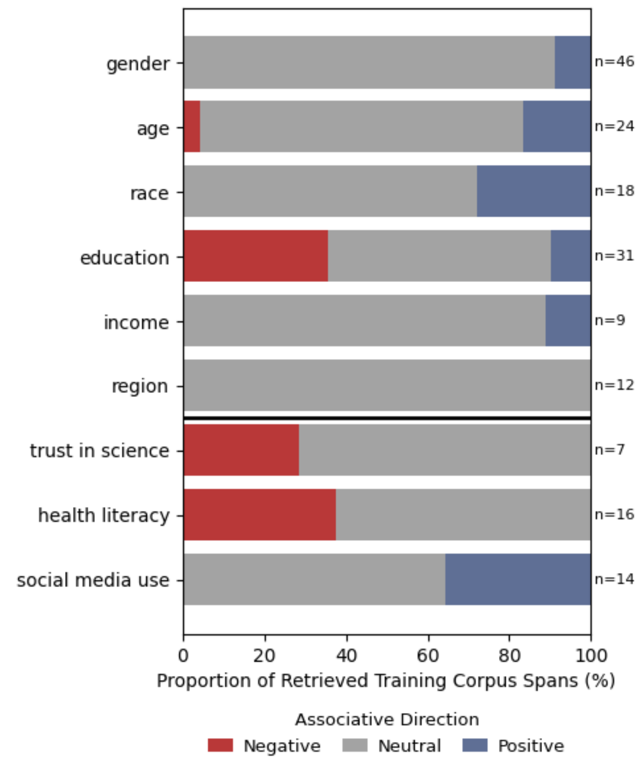

Figure 6: Distribution of inferred association directions in retrieved OLMo pre- and post-training corpus spans. Stacked bars show the proportion of spans classified as indicating negative, neutral, or positive relationships between each variable and misinformation belief or sharing; n indicates the number of retained spans per variable. Variables with fewer than five retained spans are omitted.

works that move beyond surface-level distributional alignment and explicitly test whether simulated relationships preserve key structural properties of human social processes. Without such evaluation, LLM simulations risk implicitly encoding simplified or normative assumptions about how social behavior works.

Methodologically, this study provides a generalizable pipeline for diagnosing bias and distortion in LLM-based social surveys, combining distributional metrics, predictive modeling, interaction analysis, reasoning inspection, and training-data tracing. Future work should extend this approach to additional domains, alternative network representations, and emerging model architectures, and examine whether targeted training, prompting, or architectural interventions can reduce the systematic patterns documented here.

Limitations

This study has several limitations that warrant caution in interpreting the results. Although we draw on three survey datasets spanning different domains and national contexts, they rely on online panels, which may limit representativeness and introduce sampling biases. While we aimed to cover multiple domains and items, the scope of misinformation topics remains necessarily constrained, and results may not extend to other issue areas, content formats, or claim types. The climate survey included image-based misinformation stimuli that were converted to text for this study, and the political misinformation survey was administered in Korean and translated for analysis, raising the possibility that modality- or language-specific cues relevant to human judgment were attenuated or altered. Future work should extend LLM-based simulations to multimodal settings (Plebe et al. 2025) and multilingual contexts to more closely approximate how misinformation is encountered and evaluated in real-world settings.

In addition, our interpretation of model distortions, such as the inflated influence of trust in science or political leaning, is primarily descriptive. While analyses of reasoning traces and training corpus provide convergent but exploratory evidence for potential sources of these patterns, they do not establish a causal relationship between training data, reasoning processes, and simulated response behavior. Stronger theoretical and empirical frameworks will be nec-

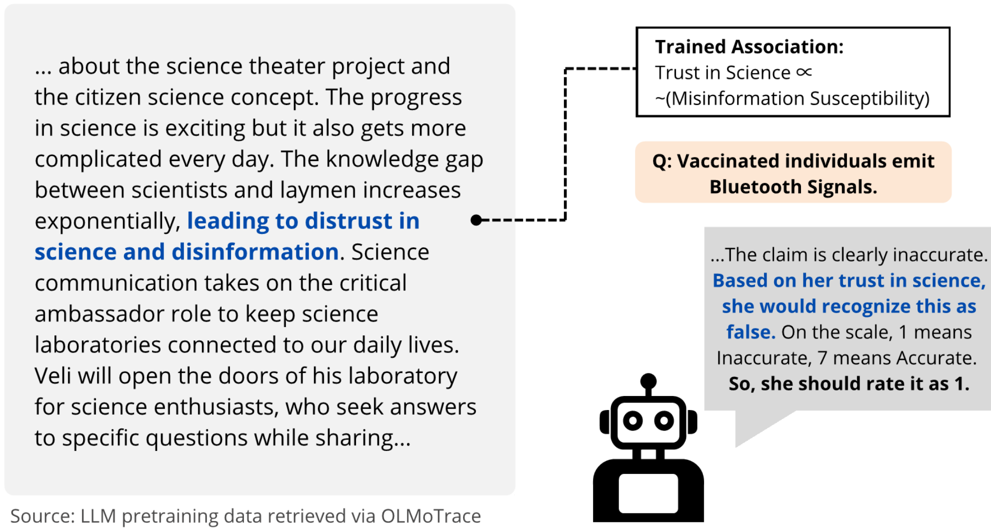

Figure 7: Conceptual illustration of a possible pathway linking training data associations, model reasoning traces, and simulated outputs. The figure illustrates how associations observed in pre- and post-training corpus spans may align with patterns in model-generated reasoning traces, helping to contextualize the exaggerated feature effects and structured distortions observed in LLM-simulated misinformation susceptibility across domains and prompting formats.

essary to explain why such biases arise and under what conditions they may be mitigated. Taken together, these limitations underscore that our findings should be viewed as an initial diagnostic rather than a comprehensive or definitive account, and that LLM-based survey simulations should be interpreted cautiously and used in conjunction with diverse, representative, and theoretically grounded human data.

Ethical Statement

All human data were collected through an online survey of U.S. adults, with approval and oversight from the university's institutional review board, in accordance with standard ethical guidelines for human-subjects research. Participation was voluntary, responses were anonymized, and no personally identifiable information was retained; all analyses were conducted at the aggregate level to further minimize risk. Thus, we consider the likelihood of harmful consequences from using these data to be low.

We acknowledge the possibility of mischaracterizing certain groups' susceptibility to misinformation and, if such patterns were taken at face value, of reinforcing bias in downstream simulation studies. A central motivation of this work is to make such risks visible by demonstrating that surface-level alignment between real and simulated responses can mask important discrepancies in relational structure and feature effects. This is a challenge that future studies in this area will also need to confront.

Acknowledgments

This project was in part supported by the NSF (Award Number 2331722).

AI Tools Usage Disclosure. During the preparation of this work, the authors used ChatGPT for text refinement and code review. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

Altay, S.; Berriche, M.; and Acerbi, A. 2023. Misinformation on misinformation: Conceptual and methodological challenges. *Social Media + Society*, 9(1): 1–13.

Altay, S.; De Araujo, E.; and Mercier, H. 2022. "If this account is true, it is most enormously wonderful": Interestingness-if-true and the sharing of true and false news. *Digital Journalism*, 10(3): 373–394.

Anthis, J. R.; Liu, R.; Richardson, S. M.; Kozlowski, A. C.; Koch, B.; Brynjolfsson, E.; Evans, J.; and Bernstein, M. S. 2025. Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Ashery, A. F.; Aiello, L. M.; and Baronchelli, A. 2025. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20): 1–10.

Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; Gulcehre, C.; Song, F.; Ballard, A.; Gilmer, J.; Dahl, G.; Vaswani, A.; Allen, K.; Nash, C.; Langston, V.; Dyer, C.; Heess, N.; Wierstra, D.; Kohli, P.; Botvinick, M.; Vinyals, O.; Li, Y.; and Pascanu, R. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Borah, A.; Mihalcea, R.; and Pérez-Rosas, V. 2025. Persuasion at play: Understanding misinformation dynamics in demographic-aware human-LLM interactions. *arXiv preprint arXiv:2503.02038*.

Chang, S.; Chaszczewicz, A.; Wang, E.; Josifovska, M.; Pierson, E.; and Leskovec, J. 2025. LLMs generate structurally realistic social networks but overestimate political homophily. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 341–371.

Chen, R.; Zhao, T.; Jaiswal, A.; Shah, N.; and Wang, Z. 2024. LLaGA: Large language and graph assistant. In *Proceedings of the 41st International Conference on Machine Learning*, 7809–7823.

Chew, L. D.; Griffin, J. M.; Partin, M. R.; Noorbaloochi, S.; Grill, J. P.; Snyder, A.; Bradley, K. A.; Nugent, S. M.; Baines, A. D.; and VanRyn, M. 2008. Validation of screening questions for limited health literacy in a large VA outpatient population. *Journal of General Internal Medicine*, 23(5): 561–566.

Choi, E. C.; and Young, L. E. 2026. Who you know, what you believe, what you share: Personal networks and health misinformation susceptibility. Manuscript under review.

Chuang, Y.-S.; Goyal, A.; Harlalka, N.; Suresh, S.; Hawkins, R.; Yang, S.; Shah, D.; Hu, J.; and Rogers, T. 2024. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 3326–3346.

Dash, S.; Reymond, A.; Spiro, E.; and Caliskan, A. 2025. Persona-assigned large language models exhibit human-like motivated reasoning. In *First Workshop on Social Simulation with LLMs*.

Dillion, D.; Tandon, N.; Gu, Y.; and Gray, K. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7): 597–600.

Ecker, U. K.; Lewandowsky, S.; Cook, J.; Schmid, P.; Fazio, L. K.; Brashier, N.; Kendeou, P.; Vraga, E. K.; and Amazeen, M. A. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1): 13–29.

Facciani, M.; and Steenbuch-Traberg, C. 2024. Personal network composition and cognitive reflection predict susceptibility to different types of misinformation. *Connections*, 44(1): 165–180.

Ferrara, E. 2024. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1): 3.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM*, 59(7): 96–104.

Gao, C.; Lan, X.; Lu, Z.; Mao, J.; Piao, J.; Wang, H.; Jin, D.; and Li, Y. 2023. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*.

Gausen, A.; Luk, W.; and Guo, C. 2021. Can we stop fake news? Using agent-based modelling to evaluate countermeasures for misinformation on social media. In *Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media*, 1–5.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Gui, G.; and Toubia, O. 2023. The challenge of using LLMs to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*.

Hewitt, L.; Ashokkumar, A.; Ghezae, I.; and Willer, R. 2024. Predicting results of social science experiments using large language models. Preprint.

Hu, T.; and Collier, N. 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)*, 10289–10307.

Huang, Y.; Mao, S.; Wang, Z.; and Young, L. E. 2026. Where personal talk meets public concern: Mapping multiplex ties in climate-change discussions. Manuscript under review.

Ihm, J.; and Lee, C.-J. 2023. Communication networks and individual resilience for individual well-being during a time of Crisis. *Health Communication*, 38(14): 3178–3192.

Jia, P.; and Zhang, Z. 2025. Misinformation dissemination and correction in online informal groups: The role of social distance and group dynamics. *New Media & Society*.

Kaiser, C.; Kaiser, J.; Manewitsch, V.; Rau, L.; and Schallner, R. 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*, 82–86.

Kim, J. Y.; Young, L. E.; Kim, D.; Curry, M.; Wang, Z.; Twyman, M.; and Yang, A. 2026. Interpersonal communication environments and counterarguing intentions toward climate misinformation: The role of personal climate discussion networks, norms, and attitudes. Manuscript under review.

Kim, S.; Jeong, J.; Han, J. S.; and Shin, D. 2024. LLM-mirror: A generated-persona approach for survey pretesting. *arXiv preprint arXiv:2412.03162*.

Kotek, H.; Dockum, R.; and Sun, D. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM Collective Intelligence Conference*, 12–24.

Lee, C.-J. 2009. The role of Internet engagement in the health-knowledge gap. *Journal of Broadcasting & Electronic Media*, 53(3): 365–382.

Lee, E.-J.; and Jang, J.-w. 2023. How political identity and misinformation priming affect truth judgments and sharing intention of partisan news. *Digital Journalism*, 11(1): 226–245.

Lee, S. J.; Lee, C.-J.; and Hwang, H. 2023. The role of deliberative cognitive styles in preventing belief in politicized COVID-19 misinformation. *Health Communication*, 38(13): 2904–2914.

Lippert, S.; Dreber, A.; Johannesson, M.; Tierney, W.; Cyrus-Lai, W.; Uhlmann, E. L.; Collaboration, E. E.; and Pfeiffer, T. 2024. Can large language models help predict results from a complex behavioural science study? *Royal Society Open Science*, 11(9): 240682.

Liu, J.; Blanton, T.; Elazar, Y.; Min, S.; Chen, Y.-S.; Chheda-Kothary, A.; Tran, H.; Bischoff, B.; Marsh, E.; Schmitz, M.; et al. 2025. OLMoTrace: Tracing language model outputs back to trillions of training tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 178–188.

Luceri, L.; Ye, J.; Jiang, J.; and Ferrara, E. 2025. The susceptibility paradox in online social influence. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 1122–1138.

Ma, W.; Deng, C.; Moossavi, A.; Wang, L.; Vosoughi, S.; and Yang, D. 2024. Simulated misinformation susceptibility (SMISTS): Enhancing misinformation research with large language model simulations. In *Findings of the Association for Computational Linguistics ACL 2024*, 2774–2788.

Manning, B. S.; Zhu, K.; and Horton, J. J. 2024. Automated social science: Language models as scientist and subjects. *arXiv preprint arXiv:2404.11794*.

Mou, X.; Wei, Z.; and Huang, X.-J. 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*, 4789–4809.

Nan, X.; Wang, Y.; and Thier, K. 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: 1–17.

Obadă, D.-R.; and Dabija, D.-C. 2022. "In flow"! Why do users share fake news about environmentally friendly brands on social media? *International Journal of Environmental Research and Public Health*, 19(8): 4861.

Orlando, G. M.; Ye, J.; La Gatta, V.; Saeedi, M.; Moscato, V.; Ferrara, E.; and Luceri, L. 2025. Emergent coordinated behaviors in networked LLM agents: Modeling the strategic dynamics of information operations. *arXiv preprint arXiv:2510.25003*.

Paletz, S. B.; Auxier, B. E.; and Golonka, E. M. 2018. A multidisciplinary, theoretical model of information propagation: Why do people share information and narratives on social media. Technical report, Defense Technical Information Centre.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.

Park, J. S.; Popowski, L.; Cai, C.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–18.

Park, J. S.; Zou, C. Q.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Willer, R.; Liang, P.; and Bernstein, M. S. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Pastor-Galindo, J.; Nespoli, P.; and Ruipérez-Valiente, J. A. 2024. Large-language-model-powered agent-based framework for misinformation and disinformation research: Opportunities and open challenges. *IEEE Security & Privacy*, 22(3): 24–36.

Pennycook, G.; and Rand, D. G. 2021. The psychology of fake news. *Trends in Cognitive Sciences*, 25(5): 388–402.

Plebe, A.; Douglas, T.; Riazi, D.; and del Rio-Chanona, R. M. 2025. I'll believe it when I see it: Images increase misinformation sharing in Vision-Language Models. *arXiv preprint arXiv:2505.13302*.

Pratelli, M.; and Petrocchi, M. 2025. Evaluating the simulation of human personality-driven susceptibility to misinformation with LLMs. *arXiv preprint arXiv:2506.23610*.

Salecha, A.; Ireland, M. E.; Subrahmanya, S.; Sedoc, J.; Ungar, L. H.; and Eichstaedt, J. C. 2024. Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, 3(12): 1–4.

Sirlin, N.; Epstein, Z.; Arechar, A. A.; and Rand, D. G. 2021. Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. *Harvard Kennedy School Misinformation Review*.

Sturgis, P.; Brunton-Smith, I.; and Jackson, J. 2021. Trust in science, social consensus and vaccine confidence. *Nature Human Behaviour*, 5(11): 1528–1534.

Tambuscio, M.; and Ruffo, G. 2019. Fact-checking strategies to limit urban legends spreading in a segregated society. *Applied Network Science*, 4(1): 116.

Tang, J.; Yang, Y.; Wei, W.; Shi, L.; Su, L.; Cheng, S.; Yin, D.; and Huang, C. 2024. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 491–500.

Törnberg, P.; Valeeva, D.; Uitermark, J.; and Bail, C. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

Wang, P.; Zou, H.; Yan, Z.; Guo, F.; Sun, T.; Xiao, Z.; and Zhang, B. 2024. Not yet: Large language models cannot replace human respondents for psychometric research. OSF Preprint.

Wu, Z.; Peng, R.; Ito, T.; and Xiao, C. 2025. LLM-based social simulations require a boundary. *arXiv preprint arXiv:2506.19806*.

Yang, Z.; Zhang, Z.; Zheng, Z.; Jiang, Y.; Gan, Z.; Wang, Z.; Ling, Z.; Chen, J.; Ma, M.; Dong, B.; et al. 2024. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*.

Yu, C.; Ye, J.; Li, Y.; Li, Z.; Ferrara, E.; Hu, X.; and Zhao, Y. 2024. A large-scale simulation on large language models for decision-making in political science. *arXiv preprint arXiv:2412.15291*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, we discuss risks and mitigation, and do not propose any deployment that could exacerbate inequities. See Ethical Statement
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Methodology
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Limitations
   (e) Did you describe the limitations of your work? Yes, see Limitations
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Ethical Statement
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Experiment Procedure and Ethical Statement
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Analytic Procedure and Results
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see Ethical Statement
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see Ethical Statement
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see Ethical Statement
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes, see Appendix
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, see Ethical Statement
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes
   (d) Did you discuss how data is stored, shared, and deidentified? Yes, see Ethical Statement

Appendix A. Survey Materials

A.1. Misinformation Items

Survey	Claim
Public Health	1. Ivermectin pills, known as the antiparasitic drug, have been approved by the FDA to treat COVID-19 2. Vaccinated individuals emit Bluetooth signals 3. Most generic sunscreens on the market contain benzenes which are a cancer-causing agent 4. Diabetes can be treated by wearing a copper bracelet 5. WHO has said smoking prevents people from getting infected with the novel coronavirus
Climate Change	1. The increase in the global polar bear population from about 5,000 in the 1960s to over 25,000 today proves that global warming is exaggerated or a hoax. 2. The Paris Climate Treaty hurts the U.S. while letting China and India pollute more, making it useless for protecting the environment. 3. A single eruption of Mount Etna releases more carbon dioxide than all human activity combined.
Pandemic Politics	1. The government is deliberately minimizing the group of people eligible for COVID-19 diagnostic test to reduce the number of confirmed cases before the election. 2. The government has exclusive control over COVID-19 clinical information, refusing to share them with the experts. 3. Purchasing public masks at a pharmacy will lead to leakage of personal information which will be used in election fraud. 4. There was a shortage in mask at hospitals in Korea because the government sent masks to China. 5. The government is providing masks purchased with tax money to China.

Table 4: Misinformation items used in the study.

Figure 8: Three meme stimuli used in climate change survey, namely (i) stop global warming hype (top left), (ii) Paris Climate Treaty (right), and (iii) Mount Ena CO2 (bottom left).

A.2. Descriptive Statistics

Variables	M **(SD) or %**
Demographics	
Female	50.82%
White	76.95%
Age	45.45 (16.01)
Education	4.24 (1.31)
Income	3.23 (1.53)
Attitudinal/Behavioral	
Political identification	3.29 (1.78)
Trust in science	3.22 (0.65)
Social media use	2.79 (1.74)
Health literacy	4.39 (0.68)
Health media exposure	1.75 (0.78)
Personal Network	
Network size	5.61 (1.73)
Density	0.67 (0.29)
Prop. of male alters	0.54 (0.23)
Prop. of white alters	0.71 (0.35)
Mean alter age	45.66 (11.98)
Mean alter education	4.07 (1.03)
Mean alter political leaning	3.58 (1.28)
Misinformation Susceptibility	
Belief	2.14 (0.85)
1. Ivermectin cures covid	2.55 (1.80)
2. Vaccinated emit signals	1.33 (1.08)
3. Sunscreen and cancer	3.75 (1.56)
4. Bracelet cures diabetes	1.49 (1.17)
5. Smoking prevents covid	1.59 (1.33)
Sharing Intention	2.18 (1.25)
1. Ivermectin cures covid	2.44 (2.06)
2. Vaccinated emit signals	1.57 (1.51)
3. Sunscreen and cancer	3.58 (2.25)
4. Bracelet cures diabetes	1.68 (1.60)
5. Smoking prevents covid	1.63 (1.52)

Table 5: Descriptive statistics of public health survey.

Variables	M (SD) or %
Demographics	
Female	59.95%
White	77.72%
Age	45.88 (13.52)
Education	4.22 (1.51)
Income	3.48 (1.59)
Attitudinal/Behavioral	
Political identification	3.49 (1.27)
Systemic Processing	3.94 (0.47)
Primary source: legacy media	2.79 (1.74)
Primary source: social media	4.39 (0.68)
No. of social media	5.56 (2.42)
Personal Network	
Climate network size	5.80 (3.46)
Climate alter prop.	0.78 (0.23)
Prop of kin alters	0.49 (0.28)
Mean tie strength	4.03 (0.51)
Mutual awareness	2.90 (0.84)
Misinformation Susceptibility	
Belief	0.50 (0.36)
1. Stop climate change hype	40.32%
2. Paris Climate Treaty	58.62%
3. Mount Ena CO2	50.40%
Sharing Intention	0.23 (0.35)
1. Stop climate change hype	23.61%
2. Paris Climate Treaty	24.14%
3. Mount Ena CO2	20.42%

Table 6: Descriptive statistics of climate change survey.

Variables	M (SD) or %
Demographics	
Female	46.75%
Seoul metropolitan area	83.42%
Age	47.21 (13.03)
Education	5.60 (1.02)
Income	4.97 (1.98)
Attitudinal/Behavioral	
Political identification	4.80 (1.90)
Trust in science	6.32 (1.57)
Social media use	4.11 (1.16)
Health literacy	3.26 (0.66)
Health media exposure	2.81 (0.50)
Personal Network	
Network size	4.86 (1.49)
Density	0.72 (0.29)
Prop. of male alters	0.24 (0.21)
Prop. of info alters	0.50 (0.37)
Mean alter age	46.32 (10.65)
Mean tie duration	19.92 (11.92)
Misinformation Susceptibility	
Belief	1.97 (0.80)
1. Gov't minimizes cases	1.80 (0.87)
2. Gov't controls clinic info	1.81 (0.83)
3. Masks and election fraud	1.71 (0.88)
4. Mask shortage due to China	2.29 (1.06)
5. Give masks away to China	2.15 (1.00)
Sharing Intention	1.27 (0.44)
1. Gov't minimizes cases	1.19 (0.47)
2. Gov't controls clinic info	1.30 (0.56)
3. Masks and election fraud	1.43 (0.65)
4. Mask shortage due to China	1.21 (0.49)
5. Give masks away to China	1.21 (0.49)

Table 7: Descriptive statistics of pandemic politics survey.

Appendix B. Divergence and Correlation between Human and Simulated Susceptibility

	Climate Change (US, 2025)						Public Health (US, 2023)						Pandemic Politics (KR, 2020)					
	Misinfo Belief			Misinfo Sharing			Misinfo Belief			Misinfo Sharing			Misinfo Belief			Misinfo Sharing		
	JS ↓	**EMD ↓**	$\rho\uparrow$	**JS ↓**	**EMD ↓**	$\rho\uparrow$	**JS ↓**	**EMD ↓**	$\rho\uparrow$	**JS ↓**	**EMD ↓**	$\rho\uparrow$	**JS ↓**	**EMD ↓**	$\rho\uparrow$	**JS ↓**	**EMD ↓**	$\rho\uparrow$
Original Format																		
Reasoning Models																		
GPT-5.1	.034	.122	.260	.011	**.039**	.170	.171	.119	.189	.023	.060	.249	.027	.093	**.434**	.021	.041	**.102**
GPT-5-mini	.074	.208	.240	**.006**	.051	.136	.194	.103	.185	.035	.051	.247	.089	.181	.325	.102	.107	-.021
Deepseek-V3	.023	**.056**	.283	.035	.169	.135	.047	.053	.172	.051	.063	.216	.133	.120	.186	.195	.112	.001
Grok-4.1-fast	.054	.087	**.301**	.064	.222	.102	**.041**	**.048**	.239	**.010**	**.016**	.282	**.021**	**.085**	.359	.026	.052	.083
Olmo-3.1-32b-think	.042	.159	.279	.009	.068	.145	.277	.148	.045	.040	.067	.171	.241	.198	-.019	.219	.139	-.042
Chat Models (CoT)																		
GPT-4.1-mini	.122	.249	.219	.034	.083	.162	.206	.126	.183	.128	.103	.248	.150	.151	.134	.281	.148	-.072
Deepseek-V3 (nr)	**.021**	.104	.289	.034	.123	.138	.056	.055	.230	.016	.023	.219	.095	.114	.242	.032	.037	.002
Grok-4.1-fast (nr)	.027	**.056**	.282	.054	.182	**.171**	.087	.104	.243	.017	.051	.225	.026	.087	.235	.045	.070	.002
Olmo-3.1-32b-inst	.125	.256	.158	.053	.133	.136	.220	.136	.187	.080	.072	.254	.195	.157	.114	.272	.154	-.024
Chat Models (ZS)																		
GPT-4.1-mini	.076	.176	.261	.049	.104	.169	.421	.160	.084	.203	.150	.178	.239	.210	.101	.143	.069	-.050
Deepseek-V3 (nr)	.031	.134	.286	.034	.083	.131	.088	.086	.176	.031	.038	.215	.137	.128	.286	.025	.054	.008
Grok-4.1-fast (nr)	.034	.060	.296	.058	.203	.143	.240	.132	.195	.060	.075	**.283**	.077	.151	.277	**.020**	**.021**	.016
Olmo-3.1-32b-inst	.437	.435	.115	.130	.193	.119	.513	.182	.082	.198	.137	.180	.331	.232	.028	.380	.166	.072
All Outputs Averaged	.037	.114	.293	.032	.085	.154	.113	.100	**.244**	.048	.056	.279	.146	.153	.316	.194	.079	.020
Alternative Ordering																		
Reasoning Models																		
GPT-5.1	.032	.124	.256	.009	**.031**	.167	.179	.121	.199	.038	.077	.228	.032	.097	**.441**	.014	.033	**.109**
GPT-5-mini	.084	.223	.244	**.005**	.041	.181	.197	.103	.210	.034	.052	.203	.092	.187	.293	.088	.101	.011
Deepseek-V3	.021	.067	.280	.034	.163	.158	.051	.059	.238	.048	.065	.233	.125	.114	.193	.169	.104	-.041
Grok-4.1-fast	.031	.065	.300	.051	.202	.147	**.043**	**.048**	.242	**.011**	**.017**	.256	**.024**	**.096**	.360	.021	.046	.070
Olmo-3.1-32b-think	.031	.125	.283	.014	.090	.157	.309	.152	.128	.034	.059	.150	.234	.192	-.023	.204	.130	-.020
Chat Models (CoT)																		
GPT-4.1-mini	.129	.259	.227	.024	.069	.160	.177	.121	.194	.090	.083	.219	.141	.161	.120	.242	.127	-.059
Deepseek-V3 (nr)	.023	.107	.249	.021	.124	.129	.074	.062	.215	.019	.023	.226	.097	.111	.238	.050	.037	-.006
Grok-4.1-fast (nr)	.017	**.040**	.262	.047	.179	.146	.082	.097	.222	.019	.045	.236	.055	.140	.246	.046	.053	.022
Olmo-3.1-32b-inst	.130	.266	.213	.063	.144	.165	.206	.132	.139	.031	.045	.255	.189	.167	.084	.221	.135	.061
Chat Models (ZS)																		
GPT-4.1-mini	.095	.209	.231	.040	.093	**.224**	.392	.156	.041	.147	.116	.245	.290	.226	.084	.072	.052	.008
Deepseek-V3 (nr)	.024	.122	.295	.028	.100	.149	.064	.074	.218	.026	.043	.228	.121	.111	.306	.027	.058	.040
Grok-4.1-fast (nr)	**.012**	.065	**.302**	.027	.147	.162	.334	.149	.161	.049	.069	.242	.113	.186	.256	**.012**	**.018**	-.003
Olmo-3.1-32b-inst	.475	.444	-.023	.154	.214	.186	.601	.192	.086	.310	.154	.100	.370	.271	.002	.522	.170	-.005
All Outputs Averaged	.037	.127	.293	.027	.083	.168	.117	.101	**.245**	.036	.052	**.260**	.158	.163	.314	.174	.071	.029
Composite Score Format																		
Reasoning Models																		
GPT-5.1	.054	.135	.244	.011	**.039**	.161	.334	.152	.176	.107	.131	.225	.033	.104	**.394**	.012	**.028**	**.083**
GPT-5-mini	.039	.139	.225	.012	.088	.089	.303	.141	.047	.105	.115	.224	.064	.162	.176	.053	.089	.008
Deepseek-V3	.020	.077	.276	.025	.138	.124	.101	.071	.203	.028	.032	.238	.091	**.090**	.196	.084	.065	.022
Grok-4.1-fast	.064	.165	.231	**.008**	.055	.126	.122	.094	.222	.045	.086	.269	.040	.128	.297	**.011**	.032	.050
Olmo-3.1-32b-think	.017	.088	.289	.018	.100	.115	.326	.156	.140	**.017**	.025	.221	.170	.169	-.029	.194	.155	-.015
Chat Models (CoT)																		
GPT-4.1-mini	.060	.156	.261	.020	.053	.154	.242	.133	.182	.127	.086	.286	.109	.138	-.037	.253	.157	-.011
Deepseek-V3 (nr)	.019	.094	.264	.042	.164	.106	.109	.077	.192	.019	**.020**	.235	.103	.118	.094	.045	.050	-.055
Grok-4.1-fast (nr)	.024	.083	.281	.013	.099	.122	.085	**.067**	.169	.023	.053	.281	**.028**	.103	.182	.055	.078	-.005
Olmo-3.1-32b-inst	.039	.144	**.307**	.009	.048	.169	.452	.173	.057	.030	.052	.221	.177	.164	-.004	.211	.132	-.046
Chat Models (ZS)																		
GPT-4.1-mini	.053	.127	.293	.041	.070	.134	.328	.141	.116	.144	.088	.288	.254	.206	.029	.188	.112	-.026
Deepseek-V3 (nr)	**.009**	.057	.256	.075	.181	**.669**	**.064**	.067	.167	.019	.026	.216	.155	.130	.170	.055	.066	-.009
Grok-4.1-fast (nr)	.025	**.055**	.299	.060	.210	.120	.416	.157	.134	.054	.072	**.323**	.101	.187	.170	.052	.038	-.009
Olmo-3.1-32b-inst	.297	.373	.149	.144	.208	.148	.612	.193	.064	.292	.172	.114	.323	.268	-.020	.420	.135	-.038
All Outputs Averaged	.030	.109	.290	.020	.086	.115	.210	.130	**.228**	.059	.077	.289	.143	.161	.200	.165	.076	.010

Table 8: Distributional divergence and rank-order agreement between ground-truth and LLM-simulated misinformation susceptibility across survey domains, models, and prompting formats. Misinformation measures are min–max normalized.

Appendix C. Simulation Prompt Format

C.1. Climate Change Simulation Prompt Format

SYSTEM This survey was conducted in the United States in Feb-Apr 2025.
Suppose you are a survey participant described below, adopting their demographics, attitudes, behaviors, and personal network.
Based ONLY on the profile, answer the given question.
If the profile lacks information, make your best estimate from available signals; never reply with "unknown" or similar.
Output format (JSON):
{"response": "integer"}

INPUT Participant profile:
{*personal network information*}
{*demographic information*}
{*attitudes and behaviors*}

Consider the following claim:
"A single eruption of Mount Etna releases more carbon dioxide than all human activity combined."

Question:
Please choose the option below that best describes your agreement with the above message.
1. I agree with the information above
2. I disagree with the information above.
OR
Question:
Please choose the option below that best describes your intention to share the above message with other people.
1. I am likely to share it with others.
2. I am unlikely to share it with others.

Figure 9: Example prompt format used to simulate climate change survey participants' misinformation belief or sharing intention. Belief and sharing questions were queried separately using identical system prompts, participant profiles, and claims.

C.2. Pandemic Politics Simulation Prompt Format

SYSTEM This survey was conducted in South Korea in May 2020.
Suppose you are a survey participant described below, adopting their demographics, attitudes, behaviors, and personal network.
Based ONLY on the profile, answer the given question.
If the profile lacks information, make your best estimate from available signals; never reply with "unknown" or similar.
Output format (JSON):
{"response": "integer"}

INPUT Participant profile:
{*personal network information*}
{*demographic information*}
{*attitudes and behaviors*}

Consider the following claim:
"Purchasing public masks at a pharmacy will lead to leakage of personal information which will be used in election fraud."

Question:
Above is information related to the coronavirus and coronavirus disease (COVID-19). To the best of your knowledge, how accurate do you think the following statements are? (1=Not at all accurate; 4=Very accurate)
OR
Question:
Would you be likely to share the following information related to the coronavirus and coronavirus disease (COVID-19) with others via social media such as Facebook or Twitter, or via messaging apps such as KakaoTalk, LINE, or WhatsApp? Please indicate how likely you are to share it. (1=No, I would not share it; 2=I would probably share it; 3=Yes, I would share it)

Figure 10: Example prompt format used to simulate pandemic politics survey participants' misinformation belief or sharing intention. Belief and sharing questions were queried separately using identical system prompts, participant profiles, and claims.

Appendix D. Original Profile Format

D.1. Public Health Survey Participant Profile

Demographics

- Gender: Female
- Age: 42
- Race/Ethnicity: White
- Education: Bachelor's degree
- Household income: $25,000 to $49,999

Attitudes and Behaviors

Political Leaning

- Here is a 7-point scale on which the political views that people might hold are arranged from extremely liberal (1) to extremely conservative (7). Where would you place yourself on this scale?: 5

Trust in Science

Please state how much you trust:

- Scientists in the US: A lot
- Science: A lot
- Scientists to find out accurate information about the world: A lot
- Scientists working in colleges or universities benefiting the public: A lot
- Scientists working in colleges or universities being honest about who is paying for their work: A lot
- Scientists working for companies making medicines or agricultural products benefiting the public: A lot
- Scientists working for companies making medicines or agricultural products being honest about who is paying for their work: A lot

Health Literacy

- How often do you have someone help you read hospital materials: Never
- How confident are you filling out medical forms by yourself: Extremely
- How often do you have problems learning about your medical condition because of difficulty understanding written information: Never
- How often do you have a problem understanding what is told to you about your medical condition: Never

Social Media Use

- How often do you use social media platform (e.g., Facebook, Instagram, Twitter) to get health-related news and information?: Once a day

Health Information Use (past 30 days)

How often have you done each of the following in the past 30 days?

- Read health information on the Internet: Once per week
- Read about health issues in newspapers or general magazines: Once per week
- Watched special health segments of television newscasts: Everyday
- Watched television programs (other than news) which address health issues or focus on doctors or hospitals: Everyday

Figure 11: Illustrative example of demographic and attitudinal/behavioral feature blocks included in public health survey participant profiles. Responses are replaced with fictitious examples to maintain anonymity.

Personal Network

Contacts

- You listed 3 regular contact(s): pp, cc, sh
- You discuss health issues with: pp

Mutual Awareness

- pp knows cc
- cc knows pp, sh
- sh knows cc

Contact Profiles

pp

- Gender: Other
- Age: 26
- Race: Asian
- Education: Bachelor's Degree
- Political leaning: Liberal

cc

- Gender: Male
- Age: 30
- Race: Asian
- Education: Master's Degree
- Political leaning: Slightly liberal

sh

- Gender: Female
- Age: 15
- Race: Asian
- Education: Less than high school
- Political leaning: Moderate

Figure 12: Illustrative example of a network feature block included in public health survey participant profiles. Responses are replaced with fictitious examples to maintain anonymity.

D.2. Climate Change Survey Participant Profile

Demographics

- Gender: Female
- Age: 42
- Race: White
- Education: High school diploma or GED
- Household income: Less than $30,000

Attitudes and Behaviors

Political Leaning

- In general, you consider yourself to be: Conservative

Systematic Processing

- After I encounter information about complex societal issues, I am likely to stop and think about it.: Agree
- If I need to act on complex societal issues, the more viewpoints I get the better.: Strongly agree
- It is important for me to interpret information about complex societal issues in a way that applies directly to my life.: Neither agree nor disagree
- After thinking about complex societal issues, I have a broader understanding.: Agree
- When I encounter information about complex societal issues, I read or listen to most of it, even though I may not agree with its perspective.: Agree

Primary Information Source

- How do you primarily access information about climate change?: Social media platforms

Social Media Use

- Which social media platforms do you have accounts on and log on at least past month? Please select all that apply.: Facebook, Instagram, Pinterest, Reddit, X/Twitter, YouTube

Figure 13: Illustrative example of demographic and attitudinal/behavioral feature blocks included in climate change survey participant profiles. Responses are replaced with fictitious examples to maintain anonymity.

Personal Network

Contacts

- You listed 3 contact(s): pp, cc, sh
- You discuss climate change with 2 contact(s): pp, sh

Perceived Mutual Awareness

- To the best of your knowledge, please indicate the degree to which you think they know one another: Most of them know each other

Climate Discussant Profiles

pp

- Relationship: Acquaintance
- Closeness (1–5 scale): 4

sh

- Relationship: Family
- Closeness (1–5 scale): 5

Figure 14: Illustrative example of a network feature block included in climate change survey participant profiles. Responses are replaced with fictitious examples to maintain anonymity.

D.3. Pandemic Politics Survey Participant Profile

Demographics

- Gender: Male
- Age: 35
- Region: Capital Region
- Education: Associate degree
- Monthly household income: 1–2 million KRW

Attitudes and Behaviors

Political Leaning

- On an 11-point scale of political (ideological) orientation, where would you place yourself? (0 = Extremely liberal, 5 = Moderate, 10 = Extremely conservative): 5

Trust in Science

- When seeking information about controversial science-related issues, how much do you trust scientists? (0 = Not at all, 10 = Completely): 7

Health Literacy

- How often do you have someone help you read hospital materials?: Sometimes
- How often do you have problems learning about your medical condition because of difficulty understanding written information?: Occasionally
- How often do you have a problem understanding what is told to you about your medical condition?: Sometimes
- How confident are you filling out medical forms by yourself?: Quite a bit

Social Media Use

- How often do you usually use social networking services such as Facebook, KakaoStory, Instagram, Naver Band, or Twitter (X)?: Several times a day
- How often do you usually use instant mobile messaging services such as KakaoTalk, LINE, or Telegram?: Once a day

Health Information Use

In the past month, how often have you encountered news reports or information about health and medical issues through the following sources?

- Daily newspapers or magazines: Not at all
- Television: Rarely
- Online news: Frequently
- Health- or medical-specialty websites: Rarely

Figure 15: Illustrative example of demographic and attitudinal/behavioral feature blocks included in political survey participant profiles. Responses are replaced with fictitious examples to maintain anonymity.

Personal Network

Contacts

- You listed 3 contact(s): ㅍㅍ, ㅊㅊ, ㅅㅎ

Mutual Awareness

- ㅍㅍ knows ㅊㅊ
- ㅊㅊ knows ㅍㅍ, ㅅㅎ
- ㅅㅎ knows ㅊㅊ

Contact Profiles

ㅍㅍ

- Gender: Female
- Age: 26
- Relationship duration (years): 2.1
- Mainly provides information support: Yes

ㅊㅊ

- Gender: Male
- Age: 30
- Relationship duration (years): 3.1
- Mainly provides information support: Yes

ㅅㅎ

- Gender: Female
- Age: 15
- Relationship duration (years): 15
- Mainly provides information support: No

Figure 16: Illustrative example of a network feature block included in political survey participant profiles. Responses are replaced with fictitious examples to maintain anonymity.

Appendix E. Composite Scores Profile Format

E.1. Public Health Survey Participant Profile

Personal Network

- Total number of alters: 7
- Network density: 0.48
- Alters by gender: Female 43%, Male 57%
- Alters by race/ethnicity: White 43%, Black 29%, Latino 14%, Native 14%
- Average age of alters: 29
- Average education of alters: Associate degree
- Average political leaning of alters: Liberal

Demographics

- Gender: Female
- Age: 27
- Race: White
- Education: Bachelor's degree
- Yearly household income: $25,000 to $49,999

Attitudes and Behaviors

- Political Leaning (1–7 scale): 1
- Trust in Science (1–4 scale): 3.0
- Health Literacy (1–5 scale): 5.0
- Social Media Use (1–6 scale): 4.0
- Health Information Use (1–4 scale): 3.2

Figure 17: Illustrative example of demographic, network, and attitudinal features included in public health survey participant profiles. Responses are replaced with fictitious examples to maintain anonymity.

E.2. Climate Change Survey Participant Profile

Personal Network

- Number of climate alters: 3
- Share of all named alters: 100%
- Kinship share of climate alters: 33%
- Mean tie strength with climate alters (1–5 scale): 3.8
- Mutual awareness among all named alters (1–4 scale): 3

Demographics

- Gender: Male
- Age: 54
- Race: White
- Education: Bachelor's Degree
- Household income: Less than $30,000

Attitudes and Behaviors

- Political Leaning: Somewhat conservative
- Systematic processing (1–5 scale): 3.8

Primary Climate Change Information Source

- Legacy media: No
- Online media: Yes

Social Media Use

- Number of social media platforms used (min 0; max 11): 2

Figure 18: Illustrative example of demographic, network, and attitudinal features included in climate change survey participant profiles. Responses are replaced with fictitious examples to maintain anonymity.

E.3. Pandemic Politics Survey Participant Profile

Personal Network (the people you discuss important matters)

- Total number of alters: 2
- Network density: 0.0
- Portion of female alters: 100.0%
- Portion of male alters: 0.0%
- Portion of alters providing information support: 100.0%
- Average age of alters: 39.0
- Average duration of relationship with alters (years): 15.0

Demographics

- Gender: Male
- Age: 43
- Region: Non-capital Region
- Education: Bachelor's degree
- Monthly household income: 5–6 million KRW

Attitudes and Behaviors

- Political Leaning (0–10 scale): 5.0
- Trust in Science (0–10 scale): 6.0
- Health Literacy (1–5 scale): 4.0
- Social Media Use (1–6 scale): 3.0
- Health Information Use (1–4 scale): 3.1

Figure 19: Illustrative example of demographic, network, and attitudinal features included in political survey participant profiles. Responses are replaced with fictitious examples to maintain anonymity.

Appendix F. Results from Alternative Prompts

F.1. Block Removal Analysis

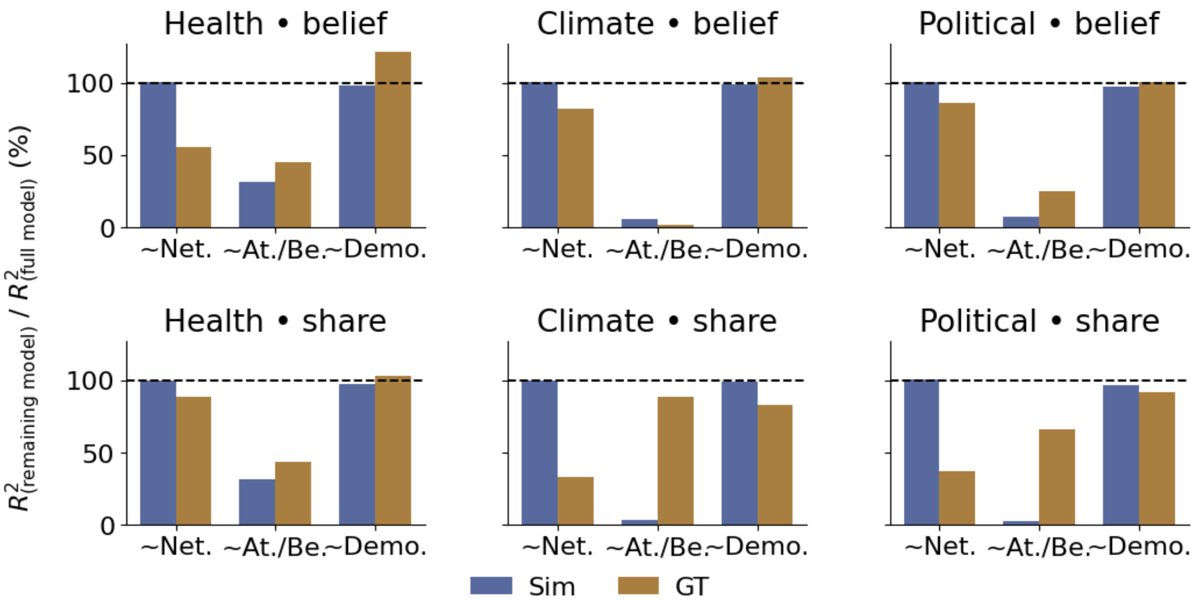

Figure 20: Relative explained variance after removing each feature block. This serves as an ablation-style analysis of feature contributions. Bars show the proportion of predictive power retained after block removal, $R^2_{\text{block removed}}/R^2_{\text{full model}}$ (in %), for simulated (Blue) and ground-truth (Orange) outcomes. All R^2 values are 5-fold cross-validated. $\sim$Net.: network block removed; $\sim$At./Be.: attitudinal/behavioral block removed; $\sim$Demo.: demographics block removed.

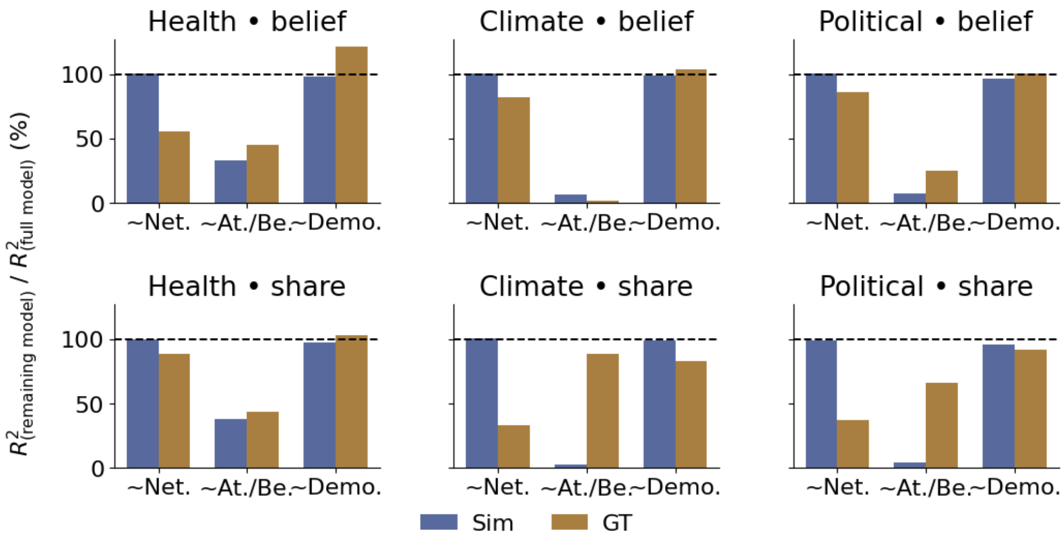

Figure 21: Relative explained variance after removing each feature block. Alternative profile ordering.

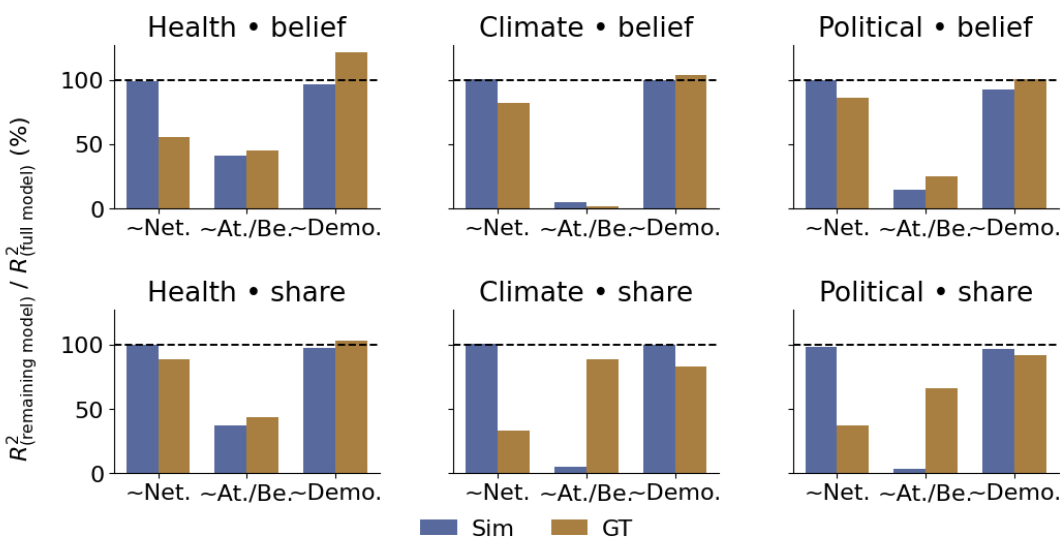

Figure 22: Relative explained variance after removing each feature block. Composite score profile format.

F.2. Interaction Analysis

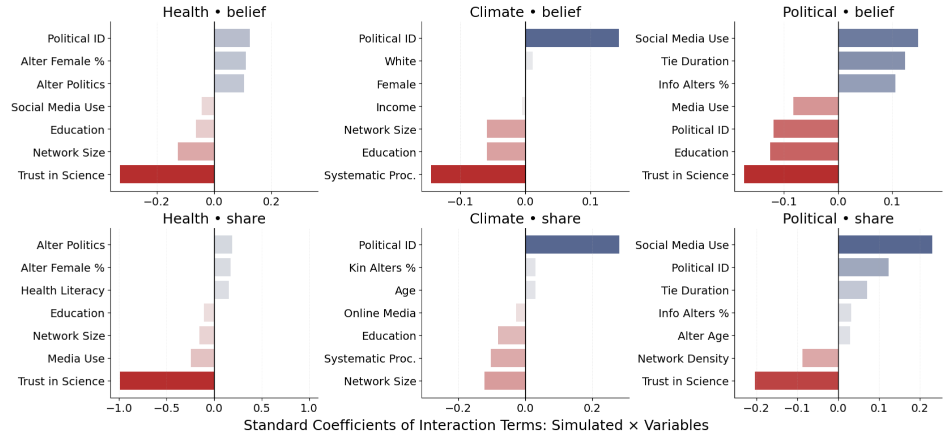

Figure 23: Standardized coefficients for simulation-by-feature interaction terms (Simulated × Variables) across domains and outcomes. Alternative profile ordering.

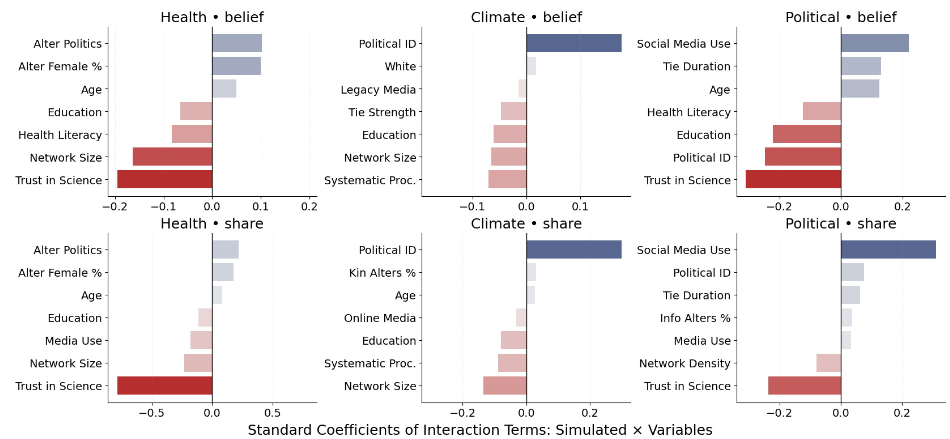

Figure 24: Standardized coefficients for simulation-by-feature interaction terms (Simulated × Variables) across domains and outcomes. Composite score profile format.

Appendix G. Procedure of Model Reasoning and Training Corpus Analysis

We examine two plausible, non-exclusive mechanisms that may help contextualize the distorted simulation results: (i) differential salience of predictors in model-generated reasoning traces and (ii) uneven representation of predictor–misinformation associations in training data, as visualized in Figure 7.

We first examine model reasoning processes. Our dataset includes reasoning chains produced by open-weight reasoning models (DeepSeek-V3 and OLMo-3.1-32B-Think) and chat models prompted with explicit chain-of-thought (CoT) instructions (GPT-4.1-mini, DeepSeek-V3, Grok-4.1-fast in non-reasoning mode, and OLMo-3.1-32B-Instruct). Across belief/sharing intention items and survey domains, this yields 42,584 unique reasoning chains.

To identify which profile features were invoked during model reasoning, we conducted a post-hoc extraction analysis using three independent LLMs (GPT-4.1-mini, DeepSeek-V3, and Grok-4.1-fast). Each model was prompted to identify variables that appeared most salient for supporting the reasoning, restricted to features explicitly mentioned or directly implied in the reasoning chain, and to return the result as a Python list (prompt shown in Figure 25). Agreement across the three model-based extractions returned an average pairwise Jaccard similarity of .750.

To reduce the influence of ambiguous or weakly implied variables, we adopted a conservative inclusion rule: a variable was retained as important for a given reasoning chain only when all three extraction models independently selected it. Under this unanimity setting, each reasoning chain yielded an average of 4.07 retained variables (SD = 1.90).

As a robustness check, two independent human raters, one of whom was an author, manually annotated 100 randomly sampled reasoning chains using the same decision rules as the prompt (Figure 25). Agreement between model-retained and human annotations was moderate (average pairwise Jaccard similarity = .628), indicating moderate alignment. Accordingly, we interpret the extracted variables as coarse indicators of reasoning salience rather than precise representations of underlying cognitive processes.

Next, we examine whether similar patterns emerge from LLM training data. While the training corpora of major proprietary models are not publicly accessible, we analyze open-source training data made available by the Allen Institute for AI (AI2) using OLMoTrace, a tool that retrieves up to 10 training-data spans most relevant to a given query. For each profile feature, we construct four query templates combining the variable name with misinformation-related concepts: (i) {*variable*} and misinformation, (ii) misinformation and {*variable*}, (iii) {*variable*} and disinformation, and (iv) disinformation and {*variable*}. To improve coverage, each variable is queried using three paraphrased name variants; the complete list of variable names and aliases is provided in Table 9. Retrieved spans are deduplicated and post-processed by retaining only those in which all query terms explicitly appear in the text. This procedure yields 188 unique spans.

To characterize the direction of association implied by each retrieved span, we further processed the text using three independent LLMs (GPT-4.1-mini, DeepSeek-V3, and Grok-4.1-fast). Each model was prompted to classify whether the association between the queried variable and misinformation susceptibility reflected in the span was positive, negative, or neutral (prompt shown in Figure 26). For example, given a span mentioning trust in science, models assessed whether the text implied a higher, lower level of susceptibility, or no clear directional association.

Agreement across the three model-based annotations was moderate (Fleiss' $\kappa = .580$). To minimize the influence of ambiguous or weakly specified associations, we adopted a strict coding rule: a span was labeled as positive or negative only when all three models reached unanimous agreement; spans lacking unanimity were treated as neutral and excluded from directional interpretation.

As an additional robustness check, two independent human raters, one of whom was an author, manually annotated the retrieved training corpus spans using the same labeling scheme. Agreement between model-consensus labels and human annotations was high (Fleiss' $\kappa = .634$), suggesting a fair amount of overlap in inferred associations. We treat these annotations as indicative rather than definitive and use them to support comparisons of associative patterns rather than causal claims.

You are given a model's reasoning and a list of candidate variables that may explain the conclusion of the reasoning.

Model reasoning:
{model reasoning}

Candidate variables:
(1) Demographics
gender, age, race, education, income, region

(2) Attitudinal
political leaning, trust in science, health literacy

(3) Behavioral
social media use, legacy media use, online media use

(4) Network Characteristics
network size, relationship with alters, issue-specific discussants, tie strength, tie duration, education level of alters, political leaning of alters, age of alters, gender of alters, race of alters, information support from alters, network density, mutual awareness

Task:
Select the set of variables that DIRECTLY SUPPORTS the conclusion according to the reasoning above.

Rules:
Only include variables that are explicitly implied or directly used in the reasoning.
Do not include redundant or weakly related variables.

Output:
Return ONLY valid JSON with exactly these keys:
{ "reasoning": "Brief evidence-based rationale (1–3 sentences).", "label": [List of selected variable names from the candidate variables] }

Figure 25: Prompt used to identify key explanatory variables from model-generated reasoning.

You are an expert researcher analyzing text data.

Text to Analyze:
"{text}"

Task:
Based ONLY on the text above, determine the DIRECTION OF ASSOCIATION between the variable "{variable_name}" and the concept/outcome "{target}" as described in the text, in a statistical, correlational, or causal sense.

Interpretation:
"Positive" means the text implies that higher/more of "{variable_name}" is associated with higher/more of "{target}" in belief or sharing.
"Negative" means the text implies that higher/more of "{variable_name}" is associated with lower/less of "{target}" in belief or sharing.
"Neutral" means the text does not clearly specify a direction (only mentions both, is descriptive without linking them, is ambiguous, mixed/conditional with no net direction, or no association is stated).

For the following variables, always interpret the association relative to the specified reference category:
Gender: Interpret the variable as a binary indicator of being female (female = 1, not female = 0).
Race: Interpret the variable as a binary indicator of being White (White = 1, non-White = 0).
Region: Interpret the variable as a binary indicator of living in a Metropolitan Area (metro = 1, non-metro = 0).
Political leaning: Interpret the variable as a binary indicator of being conservative (conservative = 1, not conservative = 0).

Rules:
Only consider belief in "{target}" or sharing tendency. If the text only states that they are related/associated without direction, return "Neutral". If the text states both directions in different contexts without a clear overall direction, return "Neutral". Do NOT use sentiment or moral approval/disapproval to decide the label.

Output:
Return ONLY valid JSON with exactly these keys: { "reasoning": "Brief evidence-based rationale (1–3 sentences).", "label": "Positive" | "Neutral" | "Negative" }

Figure 26: Prompt used for direction-of-association annotation.

Variable	Alias	Variable	Alias
gender	gender sex gender identity	network size	network size group size social reach
age	age years respondent age	relationship with alters	relationship with alters ties with contacts connections with others
race	race ethnicity ancestry	climate discussant	climate discussant climate interlocutor climate facilitator
education	education education level educational attainment	tie strength	tie strength relationship closeness connection intensity
income	income earnings household income	tie duration	tie duration connection length bond longevity
region	region area locale	education level of alters	education level of alters alters education alters educational attainment
political leaning	political leaning political orientation political stance	political leaning of alters	political leaning of alters alters political orientation alters political ideology
trust in science	trust in science faith in science confidence in science	age of alters	age of alters alters' ages network member age
health literacy	health literacy health knowledge health understanding	gender of alters	gender of alters alter gender alters' sex
social media use	social media use social media engagement online platform usage	race of alters	race of alters alter race alters race
legacy media use	legacy media use traditional media consumption mainstream media exposure	info. support from alters	information support from alters information support from network members informational assistance from close contacts
online media use	online media use digital media consumption internet media engagement	network density	network density connection density tie density
		mutual awareness	mutual awareness shared awareness reciprocal awareness

Table 9: Variable names and aliases used to query OLMoTrace.