BELIEFSIM: Towards Belief-Driven Simulation of Demographic Misinformation Susceptibility

Anonymous ACL submission

Abstract

Misinformation is a growing societal threat, and susceptibility to misinformative claims varies across demographic groups due to differences in underlying beliefs. As Large Language Models (LLMs) are increasingly used to simulate human behaviors, we investigate whether they can simulate demographic misinformation susceptibility, treating *beliefs* as a primary driving factor. We introduce BELIEFSIM, a simulation framework that constructs demographic belief profiles using psychology-informed misinformation taxonomies and survey priors. We study prompt-based conditioning and post-training adaptation, and conduct a multi-fold evaluation using: (i) susceptibility alignment and (ii) counterfactual demographic sensitivity. Across both datasets and modeling strategies, we show that beliefs provide a strong prior for simulating misinformation susceptibility, with alignment up to 92%.

1 Introduction

Misinformation is a critical challenge in today's information ecosystem, with impacts that vary substantially across demographic groups (Khachaturov et al., 2025; Timm et al., 2025). These differences reflect not only differential targeting (Ribeiro et al., 2019), but also variation in how content is perceived (Guess et al., 2019a). Prior work illustrates these complexities: younger adults were more susceptible to believing COVID-related misinformation in the UK and Brazil (Vijaykumar et al., 2021), while older adults were more likely to share false articles on social media (Guess et al., 2019b; Brashier and Schacter, 2020). Conversely, prior studies also show that beliefs play a strong role (Sultan et al., 2024): people are more likely to accept false information when it aligns or coincides with beliefs they already hold, a phenomenon known as "belief consistency" (Flynn et al., 2017; Taber and Lodge, 2006; Roozenbeek et al., 2020). Therefore, understanding how demographic differences relate to underlying beliefs is essential to explain *misinformation susceptibility*, i.e., how likely someone is to believe a misinformative claim.

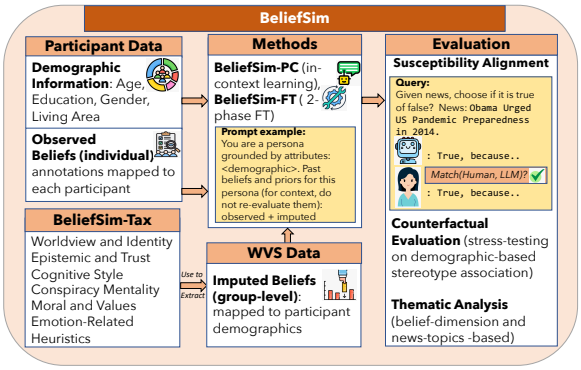

Figure 1: **BeliefSim Framework.** (1) Participant Data, Observed and Imputed Beliefs (mapped to BeliefSim-Tax) are collected from surveys, (2) Methods consist of prompt-conditioning and post-training adaptation and (3) Evaluation is performed using Susceptibility Alignment, Counterfactual and Thematic Analysis.

LLMs are increasingly used to simulate social processes (Park et al., 2023; Zhou et al., 2023; Borah et al., 2025a), including behaviors such as biased endorsement and echo chambers (Acerbi and Stubbersfield, 2023; Nehring et al., 2024; Borah et al., 2025b; Sharma et al., 2024). Closest prior work on demographic misinformation susceptibility simulation relies on coarse personas (Borah et al., 2025b), without modeling belief structures. Conversely, most belief-based studies primarily align LLMs with broad political ideologies or moral foundations (Santurkar et al., 2023; Argyle et al., 2023), or simulate general personality traits (Pratelli and Petrocchi, 2025). We believe that LLMs can serve as scalable testbeds for studying demographic susceptibility, but only if simulations move beyond surface-level personas toward belief-grounded representations.

Therefore, we ask three research questions: (1) Do beliefs improve demographic-aware simulations of misinformation susceptibility? (2) How

Dimension	Description	Example from WVS
Worldview & Identity	Group-based and political identities shape how individuals interpret and endorse information.	How proud are you to be from your country?
Epistemic Trust	Beliefs about what sources are credible and how knowledge should be evaluated (e.g., trust in science, media, institutions).	Would you say that the world is better off, or worse off, because of science?
Cognitive Style	Preferred thinking approach, ranging from analytical, accuracy-driven reasoning to intuitive impression-based judgments.	Nowadays, does one often have trouble deciding which moral rules to follow?
Conspiracy Mentality	General tendency to perceive hidden plots or malevolent intent behind significant social or political events.	How many people who are in the media do you believe are involved in corruption?
Morals & Values	Core moral values and cultural worldviews influence judgments of right, wrong, fairness, and societal norms.	Is it justifiable - claiming government benefits to which you are not entitled?
Emotion-Related	Emotion-driven beliefs related to fear, perceived risk, or threat sensitivity that heighten susceptibility to alarming misinformation.	How satisfied are you with your life as a whole these days?
Heuristics	Reliance on mental shortcuts (e.g., familiarity implies truth) that increase vulnerability to repeated or simplified claims.	Please indicate how much you use it: Daily newspaper, TV News, Radio News.

Table 1: BeliefSim-Tax for modeling demographic-aware misinformation susceptibility in LLMs. We provide representative WVS items as examples of how survey questions align with each belief dimension.

can such simulations be rigorously evaluated using utility and counterfactual demographic analyses? and (3) What modeling strategies are best suited for LLM-based simulation?

To answer these questions, we propose BELIEFSIM, a belief-driven simulation framework for demographic misinformation susceptibility, illustrated in Fig 1. We summarize our contributions as follows: **(1)** We construct a belief taxonomy (BELIESIM-TAX) and build a simulation dataset by aggregating data from existing surveys and studies; **(2)** We conduct empirical analyses to identify factors across belief and demographic dimensions that are most important for simulation; **(3)** We explore simulation techniques including prompt-based (BELIEFSIM-PC) and post-training (BELIEFSIM-FT) approaches using several LLMs; and **(4)** We perform a counterfactual study to understand when and how demographic information may provide useful priors vs stereotype-like sensitivity. Finally, we outline actionable steps that can help design future intervention methods.

2 Related Work

Psychological research has established that susceptibility to misinformation is heavily driven by *belief consistency*. While demographic factors like age, gender, and education provide significant predictive signals, recent large-scale meta-analyses suggest these effects can be context- and belief-dependent Sultan et al. (2024). However, most studies that establish benchmarks with large-scale human annotations face an inherent scalability bottleneck (Maertens et al., 2024; Borah et al., 2025b). High-quality human data collection is resource-intensive and slow, often limiting research to static snapshots (e.g. COVID-19) or narrow domains.

The emerging capability of LLMs to simulate human behavior offers a potential solution. Early work showed that LLMs could act as "silicon subjects" via demographic prompting (Argyle et al., 2023). Studies also investigated if LLMs reproduce classic social science findings and moral values (Park et al., 2023; Borah et al., 2025a; Nair and Wang, 2025). However, fidelity often degrades when simulations depend only on demographic labels (Giorgi et al., 2024), leading to stereotypical associations (Borah and Mihalcea, 2024).

To move beyond demographics, recent research has focused on conditioning models with richer context such as beliefs (Moon et al., 2024; Namikoshi et al., 2024). Relevant relevant to our work, Chuang et al. (2024) introduced "Human Belief Networks", showing that seeding an agent with a single belief (e.g., a stance on welfare) was more predictive of downstream responses than an entire demographic profile. However, most belief-focused studies have largely not examined their use in misinformation susceptibility. Our work bridges this gap by focusing on this domain, where the interplay between demographics and beliefs is highly complex.

3 BeliefSim-Tax

Beliefs are an important tool for modeling demographic simulations for misinformation. Prior work

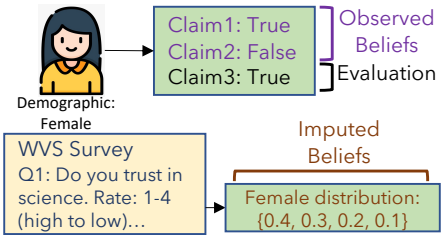

Figure 2: **Data Example.** We map participant demographics to WVS responses to derive group-level imputed beliefs, and use claim-level evaluations as participant observed beliefs These are then used to predict observed beliefs.

has shown that belief dimensions, such as conspiracy beliefs or trust in science, can help predict susceptibility to misinformation (Ecker et al., 2022; Munusamy et al., 2024).

Grounded in psychological and cognitive science research, we compiled BELIEFSIM-TAX (BeliefSim-Taxonomy) shown in Table 1, that includes seven core dimensions most associated with misinformation susceptibility: (1) Worldview and Identity Beliefs (Kahan, 2017; Van Bavel et al., 2021), (2) Epistemic Trust Beliefs (De Coninck et al., 2021; Lewandowsky et al., 2023), (3) Cognitive style (Pennycook and Rand, 2019; Ecker et al., 2022), (4) Conspiracy mentality (Douglas et al., 2019; De Coninck et al., 2021), (5) Moral and Value Beliefs (D'Errico et al., 2022; Yang et al., 2024), (6) Emotion-Related Beliefs (Brady et al., 2017; McLoughlin et al., 2024) and (7) Heuristic Beliefs (Lin et al., 2016; Fazio, 2020). While the above dimensions have traditionally been studied in isolation and within human populations, we unify them into a structured taxonomy designed for computational modeling (Further details are provided in Appendix A. This enables systematic belief simulation by providing a framework for organizing belief data and evaluating it.

4 Data for Susceptibility Simulation

For our study, we consider four demographic axes: Gender (female/male), Age (younger: <=35 years/older: >=60 years), Living Area (Rural/Urban), and Education (completed high school/not completed high school).[1] These demographics are commonly associated with systematic differences in media exposure, institutional trust, etc., all of which influence vulnerability to misinformation (Allcott and Gentzkow, 2017; Guess et al., 2019b; Allcott et al., 2019; Anspach and Carlson, 2024). Focusing on these axes enables controlled analysis of demographic context for simulating susceptibility signals in LLMs.

4.1 Evaluation Data

For susceptibility evaluation, we use two datasets containing human judgments of whether they believe a given claim: (1) PANDORA Dataset from Borah et al. (2025b), containing annotations from 318 participants. Each participant provided judgments on 3 distinct claims, along with demographic information: age, gender, living area, and education. The claims are collected from RumorEval (Gorrell et al., 2019), which consists of true/false rumors covering eight news and natural disaster events; (2) MIST dataset from Maertens et al. (2024). From MIST, we use only Study 1 (MIST-1) which includes 409 participants, each providing judgment on the same 100 claims, and demographic information including age, gender, and education. From the two datasets combined, we obtain 13.8K claims for evaluation.

4.2 Belief Data

We collect two complementary belief signals, consisting of individually observed belief judgments and group-level (demographic) belief distributions:
(1) Observed Data - claims that were directly judged by participants in the PANDORA and MIST-1 datasets. These responses capture individual-level belief judgments, reflecting each participant's personal stance rather than demographic group averages. For each evaluation instance, we include two held-out claim judgments from the same participant as observed beliefs, keeping them separate from the target evaluation claim. Across 13.8K evaluation instances, this yields 27.6K observed belief-judgment instances.
(2) Imputed Data - inferred from the World Values Survey Wave 7[2] distributions. Imputed data represent demographic belief priors (group–level), inferred from WVS, conditioned solely on demographic attributes. We map these imputed belief items to our belief taxonomy using exploratory factor analysis. This yields 126 imputed belief questions and corresponding demographic distributions. Table 1 shows examples, and Appendix A.2 contains mapping details and all questions.

[1] For ease of analysis, we focus on binary groupings, leaving finer-grained analysis to future work.

[2] https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp

Fig. 2 provides an example of simulation data. We restrict all our analysis to U.S.-based participants and English headlines, consistent with the scope of the available datasets, while still leveraging variation in living area, age, gender, and education.

5 BeliefSim-PC: Prompt-based Conditioning

We propose BELIEFSIM-PC (BeliefSim-Prompt-Cond.), which consists of prompts that condition the model on participant demographics and belief signals (both observed and imputed) from the surveys. Note that we simulate each demographic axis separately, as it enables controlled analysis of demographic-level differences.

Method. For WVS questions, answers are either Likert scales (Bertram, 2007) or Yes/No questions. Therefore, we input modal (most frequent) responses in the prompt as beliefs. For example, consider the WVS question "How important is religion in your life?", rated on a 1–10 scale. If a demographic group most frequently responds with a value of 3, we use this modal response in the prompt. Given demographic group <d>, and belief annotations <b>, we use: `You are a persona grounded by attributes: <d>. Past beliefs and priors for this persona (for context, do not re-evaluate them): <b>. When judging a claim, stay consistent with this persona's prior beliefs where reasonable.` We focus on modal responses in the main experiments for prompt stability, and include a robustness check with full belief-distribution prompting in Appendix C.5.

We perform experiments under four primary input conditions for prompt-based conditioning: (1) zero-shot (without any demographic or belief information), (2) demographic information only, (3) belief information only, and (4) both demographics and beliefs. We additionally ablate over belief dimensions (the seven in our taxonomy) and belief sources (observed vs. imputed), as well as over demographic attributes (analyzing groups separately). These experiments yield several insights across 12 total settings. We perform these experiments using several instruction-tuned LLMs: `Llama-3-8B-Instr`, `Llama-3.2-3B-Instr`, `Qwen2.5-14B-Instr`, `Mistral-7B-Instr-v0.2`, `OLMo-2-7B-Instr`, `DeepSeek-LLM-7B-Chat`, and `Grok-4-Fast`. We report averaged results across three separate runs.

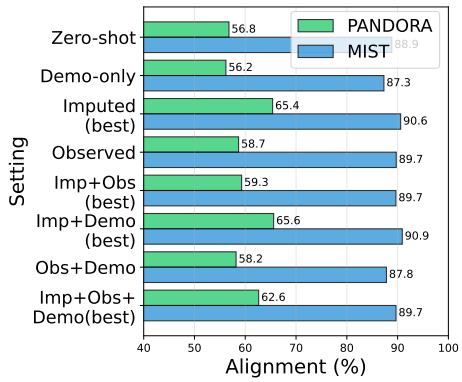

Figure 3: **Susceptibility Alignment**, averaged across demographic groups and models. Imputed + Demo(graphic) (best) performs the best. All belief-based settings do better than zero-shot and demo-only. (Fig 10 shows results of all settings.)

Evaluation. We measure susceptibility alignment: whether the LLM correctly predicts if an individual would judge a claim as true or false. We also discuss demographic divergence and the role of human judgment variability in Appendices C.6 and C.7.

5.1 Results

Fig 3 shows the susceptibility alignment across the belief and demographic configurations for both datasets. For cases that include `Imputed` beliefs, we only report `best` results by selecting the belief dimension that achieves the highest accuracy. Findings show that incorporating belief information consistently improves performance over the zero-shot and demographic-only baselines. We next analyze the impact of specific belief types, demographic ablations, datasets, and model performances. Appendix C contains detailed results for all settings. Furthermore, Appendix C.5 shows that full belief-distribution prompts do not provide stable results across models. So, we use modal responses as a more stable and compact representation in the main combined settings.

Demographic vs Belief information. Overall, beliefs drive most of the positive gains. Demo-only degrades zero-shot performances for both datasets, while beliefs-only (imputed) show larger gains (+1-16) points over demo-only and zero-shot settings. Imputed + Demo achieves the highest performance. This shows that what people *believe* may be more predictive than *who they are*. This is also in line with findings from Borah et al. (2025a), which show that belief-based signals are stronger than demographic signals in LLMs.

Imputed vs Observed Beliefs. Across settings, imputed beliefs consistently outperform observed beliefs, more so for PANDORA. A likely reason is

MODEL	PANDORA	MIST
Qwen2.5-14B	62.28	95.97
Llama-3-8B	57.49	90.53
Llama-3.2-3B	53.89	85.14
Grok-4-Fast	57.07	87.20
Mistral-7B-v0.2	55.99	70.07
OLMo-2-7B	50.60	76.95
DeepSeek-7B	52.40	59.04

Table 2: **Model performance.** Qwen performs the best followed by Llama models while Deepseek has the worst alignment.

that observed beliefs, being tied to specific participants, can be sparse, or mismatched to the new claim, while imputed demographic beliefs may act as *smoother population priors*. Furthermore, combining imputed and observed beliefs helps, but remains below imputed-only, likely because sparse individual signals can conflict with or dilute the smoother demographic prior. Thus, accuracy trends are: $imputed > imputed + observed > observed$.

Imputed Belief Ablations. Adding one belief dimension at a time to the prompt consistently yields higher performance than including all belief dimensions together. Across datasets, emotion-related and moral-values are the strongest. These also vary by demographics: emotion-related is the strongest in Gender; moral-values is the strongest for Rural/Urban and Education; heuristics/cognitive beliefs are the strongest in Education (Detailed results are in Appendix C.2).

Demographic Ablations. Adding demographic information on top of beliefs yields small gains (in the range 0.2-1.2%), especially for imputed beliefs. Thus suggesting that demographics serve as an additional context signal, helping with relevant belief priors. On the effect of belief across demographic groups, accuracies for age, education and living area are improved with belief addition, specifically using the imputed beliefs. Gender shows the largest sensitivity to which beliefs are included; prior work also show that gender effects are often smaller/mixed in misinformation contexts (Sultan et al., 2024). This underscores the need for caution in selecting belief evidence for gender-specific cases.

Model and Dataset comparison. Across models, Qwen consistently shows higher accuracies, followed by other models, as shown in Table 2.

Across datasets, MIST yields a high zero-shot baseline of 88.9% and gains the most from belief integration in the best imputed + demographic setting, reaching 90.9% (+2.1). In contrast, PANDORA has a lower zero-shot baseline of 56.8%, but shows a larger gain from belief integration, reaching 65.6% in the best imputed + demographic setting (+8.8). Therefore, the impact of adding belief priors varies across datasets: PANDORA contains fewer examples than MIST (318 vs. 13.5K), and benefits more from belief conditioning. Nevertheless, adding belief priors is beneficial across both datasets after screening outlier modal runs.

We further compare veracity/factuality accuracy with susceptibility simulation in Appendix C.8, showing that factual correctness and participant-judgment prediction are distinct objectives: models with stronger veracity accuracy do not always better predict human judgments. We also test whether the remaining gap is due to pretraining knowledge overriding persona conditioning (Appendix C.10). We find that factual confidence biases zero-shot predictions, but is not the dominant limiting factor.

Thematic Analysis. We cluster news claims into latent topics and test whether LLM susceptibility alignment differs across demographics by topic. We conduct our analysis on the MIST dataset, as it is much larger than PANDORA. We find that demographic differences in susceptibility are topic-dependent: clusters corresponding to lexically obvious or opinion-based claims exhibit minimal demographic variation, whereas more ambiguous or diverse topics like science/health claims and government show larger differences across age (older groups scoring higher) and education groups (higher-educated groups scoring higher). Gender-based differences are consistently small across all topics. These findings suggest that demographic susceptibility is also contextual and topic-sensitive, motivating further evaluation in future studies (More details are in Appendix C.4).

5.2 Demographic-based Counterfactual Evaluation

Results show that demographics can be a helpful context signal in addition to belief information and legitimately correlate with misinformation susceptibility. But they can also induce spurious signals that appear accurate but could be stereotypical (Wan et al., 2025; Geirhos et al., 2020). To investigate this, we conduct counterfactual evaluations to assess whether demographic attributes provide informative priors or introduce stereotype-sensitive associations.

Method. We conduct three complementary analy-

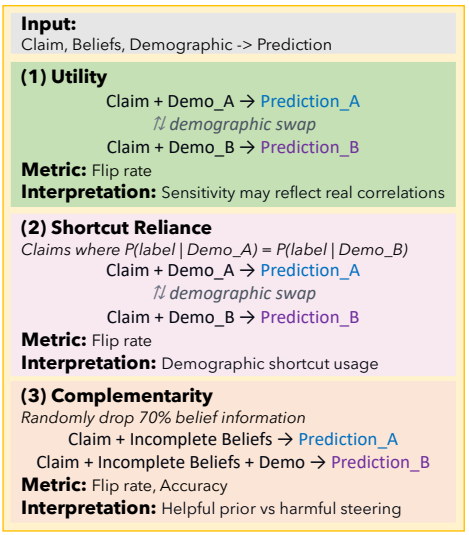

Figure 4: **Demographic-based Counterfactual Evaluation.** Note that we perform a demographic swap within the same group, i.e., swap male with female, rural with urban, etc. keeping the claim constant.

ses (Fig 4):

(1) Utility: Do demographics improve susceptibility prediction beyond claim content? We test utility first with demographic-only prompts, then swap the demographic attribute while keeping the claim fixed and measure prediction flips. Higher flip rates indicate stronger demographic sensitivity, which may reflect real demographic-label correlations rather than only spurious shortcuts.

(2) Shortcut reliance: Does the model rely on demographic cues even when they are non-informative by construction? We create a controlled subset where each claim has matched human-label distributions across demographic groups, making demographics non-predictive. High flip rates here indicate short-cutting and possible stereotyping.

(3) Complementarity: When belief evidence is incomplete, do demographics help or hurt? We simulate missing evidence by dropping 70% of total WVS belief statements, then compare belief-only vs. belief+demographic predictions. Flip rates show how much demographics steer predictions under uncertainty, while accuracy indicates whether this steering improves performance.

Results. Fig 5 shows the flip-rates per model averaged across demographic groups and datasets. Overall, `Mistral` shows the lowest flip rates, suggesting limited dependence on demographic cues. `Qwen` is also comparatively stable and low. In contrast, `Llama` and `Deepseek` models show substantially higher demographic sensitivity. Across models, the largest flip rates concentrate in Education

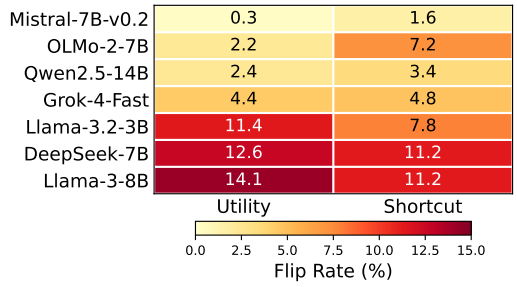

Figure 5: **Flip-Rates for Counterfactual Evaluation**: Mistral and Qwn models have lower flip-rates whereas Llama and Deepseek have higher rates. Darker (red) colors mean higher flip rates while lighter (yellow) colors mean lower flips.

and Living Area, especially in shortcut reliance, which could relate to stereotype-driven reliance (details and qualitative evidence in Appendix C.9).

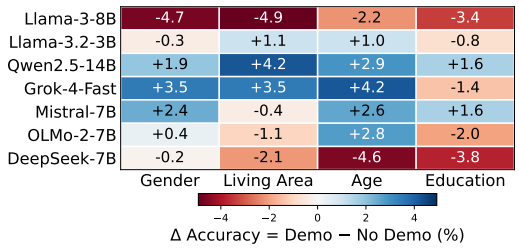

Figure 6: **Complementarity Analysis.** Adding demographics has mixed effects (mostly negative) when partial belief information is added. Positive values mean that adding demographics increases accuracy, while negative values mean that it decreases accuracy.

For complementarity analysis, Fig 6 shows that demographics show a mixed effect. `Qwen` improves for all axes while `Mistral` and `Grok` also show mostly positive gains (except one axis). However, other models show mixed or negative shifts, especially for Living Area and Education.

Overall, these findings suggest that demographic sensitivity is highly model-dependent and demographics provide limited practical benefit when partial beliefs are present, showing they are a weak prior at best. This motivates moving beyond prompt-only conditioning, as demographic cues in the prompt can introduce unstable shortcuts. Therefore, belief signals must be integrated in a way that is robust to conflicting sources.

6 BeliefSim-FT: Post-training Adaptation

Prompt-based conditioning experiments show the benefits of using beliefs for demographic misinformation susceptibility simulation. However, mixing imputed and observed beliefs in the prompt is associated with lower performance. This is intuitive, as observed (ground-truth) beliefs are tied to individual participants and imputed beliefs are only

linked to demographic groups and not individuals, creating inconsistencies. Furthermore, training directly on demographics, modal responses and susceptibility labels risks leakage, where the model may learn label shortcuts as observed in counterfactual evaluation (Wan et al., 2025; Geirhos et al., 2020). To address this, we propose BELIEFSIM-FT (BeliefSim-Fine-Tuning), which decouples imputed beliefs from observed data, while still learning transferable belief representations through a separate adapter.

Method. To decouple group-based imputed beliefs from individualistic observed beliefs, BeliefSim-FT consists of a two-phase design: (1) belief modeling, and (2) susceptibility fine-tuning.

Phase 1: Belief Modeling. We train a belief adapter by freezing the base LLM and training a belief head, implemented as a linear projection followed by softmax, to predict demographic-conditioned belief distributions for WVS questions. Concretely, given the encoder representation $h_\phi(q, d)$ for question q and demographic context d, we model $p_\theta(y \mid q, d) = \mathrm{softmax}(W h_\phi(q, d) + b)$. This captures population-level variability in belief responses. Importantly, this stage uses only survey supervision (no misinformation labels), reducing the risk of label leakage or reward hacking. Because BeliefSim-FT separates belief modeling from susceptibility prediction, extending to additional personas primarily requires collecting additional demographic-belief distributions, while the same adapter-based architecture can be reused.

Phase 2: Susceptibility Fine-tuning. We then freeze the belief adapter and train a lightweight susceptibility head that combines the base LLM's semantic representation of the prompt with the belief representation produced by the adapter. Given a persona-conditioned input (demographics and observed beliefs) and target claim x, the frozen adapter outputs a belief embedding z_{bel}, which we concatenate with the claim representation $h_\phi(x)$ and predict susceptibility via a binary classifier:

$$p_\psi(y \mid x, d) = \mathrm{softmax}(U\,[\,h_\phi(x)\,;\, z_{\text{bel}}(d)\,] + c)\,, \quad (1)$$

where $y \in \{\text{true}, \text{misinformation}\}$. We train only the susceptibility head parameters $\psi = \{U, c\}$ with cross-entropy loss, keeping the base model and belief adapter frozen. This is to ensure susceptibility learning relies on transferable belief structure rather than encoding demographic shortcuts.

Data. For Phase 1, we construct demographic belief priors from empirical WVS response distributions. For each question q and demographic group d, we estimate a categorical distribution $P(r \mid q, d)$ over response options $r \in \{1, \ldots, K_q\}$. These distributions capture meaningful demographic variability across WVS belief questions. We restrict our experiments to six demographic groups, as MIST does not include living-area annotations, and the PANDORA data is too small to support standalone data for fine-tuning. This results in 126 WVS response distributions per demographic group, yielding $6 \times 125 = 750$ examples in total. We apply textual augmentation via back-translation using Google Translate,[3] producing 1250 examples. This is sufficient as the belief adapter is a lightweight linear head requiring fewer examples to reliably capture demographic belief patterns. Training is performed using an 80/20 train/val split.

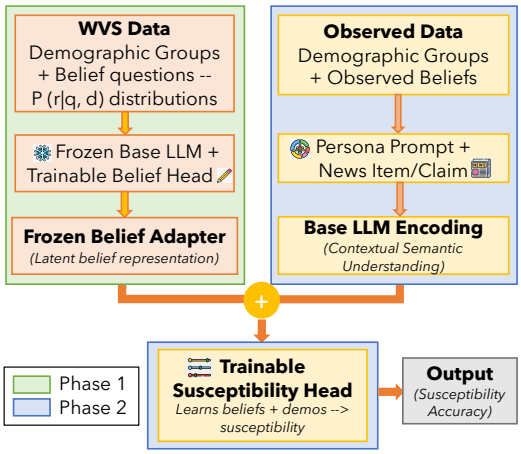

Figure 7: **BeliefSim-FT framework.** Green-shaded components correspond to Phase 1 (Belief Modeling), and blue-shaded components are Phase 2 (Susceptibility Fine-Tuning).

For Phase 2, we use (PANDORA+MIST-1) data, keeping each demographic group separate per data point (as in prompt conditioning). The target is the prediction on a claim, giving susceptibility scores. We use a 80/20 train/val split, yielding approximately 33.1k training examples and 5.1k validation examples (after removal of overlapping examples). We also use MIST-2b as a cross-study, cross-participant evaluation under the standardized MIST instrument consisting of 7k evaluation points. Appendix D.1 discusses the distinction between participant leakage and claim overlap.

We compare BeliefSim-FT against two baselines: standard one-phase LoRA fine-tuning on PANDORA+MIST-1, and a two-phase LoRA variant that replaces our head-based adaptation. We evaluate three representative models: `Qwen` (high

[3] https://translate.google.com/

Model	PANDORA+MIST-1		MIST-2	
	Acc	F1	Acc	F1
Baseline (1-phase)				
Llama-3-8b	0.750	0.751	0.680	0.673
Qwen-2.5-14b	0.760	0.761	0.720	0.712
Mistral-7B	0.740	0.761	0.650	0.653
LoRA-FT (2-phase)				
Llama-3-8b	0.809	0.809	0.876	0.861
Qwen-2.5-14b	0.800	0.800	0.893	0.892
Mistral-7B	0.800	0.790	0.887	0.867
BeliefSim-FT (2-phase)				
Llama-3-8b	0.793	0.793	0.884	0.896
Qwen-2.5-14b	0.792	0.791	0.924	0.906
Mistral-7B	0.792	0.792	0.884	0.823

Table 3: **Fine-tuning performances.** 2-phase belief training improves performance across both datasets.

alignment, low shortcut reliance), `Mistral` (low alignment, low shortcut reliance), and `Llama` (high alignment, high shortcut reliance).

Evaluation Metrics. Phase 1 reflects population-level belief fidelity, so we use distribution-level metric over Likert-scale responses, Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951). This measures how well predicted belief distributions align with each demographic (lower is better for both). Phase 2 reflects task utility. So, we compute susceptibility alignment for this phase.

6.1 Results

Phase 1. `Qwen` has the lowest KL-divergence (0.051), followed by `Mistral` (0.087) and `Llama` (0.287). With LoRA-ft, the trends remain similar (complete results are in Appendix D).

Phase 2. Table 3 shows that cross-study transfer to MIST-2 improves substantially when moving from 1-phase to 2-phase training. The 1-phase baseline performs well on PANDORA+MIST-1 but drops on MIST-2, while BeliefSim-FT achieves the strongest MIST-2 performance (up to 92.4%). The higher MIST-2 scores may partly reflect that it uses the standardized MIST instrument with reliability-selected items, whereas PANDORA+MIST-1 contains broader, more heterogeneous claims. LoRA-FT slightly improves on PANDORA+MIST-1 but transfers less strongly, possibly due to mild overfitting from updating more parameters. Across models, `Qwen` performs best in the 2-phase settings. Shortcut-reliance experiments further show that BeliefSim-FT reduces flip rates to nearly 0%, even for `Llama`, which had the highest flip rate in the prompt-based setting (Appendix D.4).

7 Lessons Learned and Actionable Steps

Our findings show that belief priors are central to simulating demographic misinformation susceptibility with LLMs. We introduce BeliefSim, and show that decoupling belief modeling from susceptibility prediction enables higher accuracy while reducing spurious demographic sensitivity. These findings offer actionable insights to design targeted demographic interventions.

Beliefs are crucial to demographic susceptibility simulation. Adding belief priors (especially imputed) leads to high accuracy gains across. Future studies can investigate additional survey sources to broaden coverage and focus on diverse demographic contexts. Furthermore, studying the effective decoupling of belief types is important.

Simple head tuning is effective for simulation alignment in smaller models. Our two-phase BeliefSim-FT approach generalizes better than full fine-tuning, suggesting that lightweight, modular adaptation can be effective while being cheaper. Future work can improve robustness by exploring alternative adapter designs and fusion mechanisms for combining belief and text representations.

Counterfactual evaluation is important for simulation studies. Counterfactual experiments quantify each model's reliance on demographic shortcuts. Such measures are necessary as demographic correlation can be a real/spurious cue. Future work can build on this by developing fine-grained stress tests to understand what drives model flips.

8 Conclusion

This paper introduced BeliefSim, a belief-driven framework for simulating demographic misinformation susceptibility using LLMs. Across two datasets and through prompt-based conditioning and post-training adaptation, we show that belief priors are a key driver for demographic-level susceptibility patterns. In contrast, demographics alone are an unreliable signal that can induce shortcut reliances. Our results highlight the importance of decoupling imputed and observed belief sources and we provide practical evaluation tools for it, such as prediction accuracy and counterfactual sensitivity. Based on these findings, we outline directions for future work and release our open-source framework, BeliefSim.[4]

[4] https://anonymous.4open.science/r/belief-sim

Limitations

Coverage and Intersectionality

Our demographic modeling is intentionally largely single-axis: we evaluate on a small set of demographic attributes (8 groups) and treat them independently. Our goal was to isolate the effects of simulation across demographic groups and simplify counterfactual analysis as well as balance labels across groups. However, intersectional groups (e.g., older x low-education, female x rural) can have interesting implications for simulation purposes, and future work can further investigate BeliefSim for intersectional belief profiles. In addition, while we mostly adhere to binary demographic axes, each axis can be further analyzed in a finer-grained approach. Finally, our study is based on US participants only, due to the availability of current data sources. We acknowledge that these may have WEIRD implications (Mihalcea et al., 2025), and future work should focus on investigating wider, cross-cultural misinformation susceptibility.

Counterfactual Flips do not always mean stereotyping

Our flip-rate analyses quantify sensitivity to demographic perturbations, but flips do not uniquely measure stereotype-driven shortcutting. Furthermore, low flip rates do not guarantee fairness - models may still encode demographic effects indirectly through correlations or beliefs. Although our experiments help isolate spurious dependence, our experiments may not capture real-world causal pathways as they require much extensive analysis. Future work can investigate extending our counterfactual flips experiments with stronger causal/robustness evaluations to better understand causal demographic effects and/or measure stereotype-driven shortcutting.

Experimental Limitations

Access to closed-source LLMs is constrained by transparency and cost. Therefore, we limit our experiments to three open-source models (to support controlled counterfactual swaps, ablations, and multiple runs). As a result, our conclusions may not fully generalize to the strongest closed-source models. Future work can extend our framework by benchmarking a broader set of closed-source models under matched prompting and budgeted sampling.

References

Alberto Acerbi and Joseph M Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & politics*, 6(2):2053168019848554.

Nicolas M Anspach and Taylor N Carlson. 2024. Not who you think? exposure and vulnerability to misinformation. *New Media & Society*, 26(8):4847–4866.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Dane Bertram. 2007. Likert scales. *Retrieved November*, 2(10):1–10.

Angana Borah, Marwa Houalla, and Rada Mihalcea. 2025a. Mind the (belief) gap: Group identity in the world of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18441–18463, Vienna, Austria. Association for Computational Linguistics.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent LLM interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9306–9326, Miami, Florida, USA. Association for Computational Linguistics.

Angana Borah, Rada Mihalcea, and Verónica Pérez-Rosas. 2025b. Persuasion at play: Understanding misinformation dynamics in demographic-aware human-llm interactions. *arXiv preprint arXiv:2503.02038*.

William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Nadia M Brashier and Daniel L Schacter. 2020. Aging in an era of fake news. *Current directions in psychological science*, 29(3):316–323.

Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. 2024. Beyond demographics: Aligning role-playing llm-based agents using human belief networks. *Preprint*, arXiv:2406.17232.

David De Coninck, Thomas Frissen, Koen Matthijs, Leen d'Haenens, Grégoire Lits, Olivier Champagne-Poirier, Marie-Eve Carignan, Marc D David, Nathalie Pignard-Cheynel, Sébastien Salerno, et al. 2021. Beliefs in conspiracy theories and misinformation about covid-19: Comparative perspectives on the role of anxiety, depression and exposure to and trust in information sources. *Frontiers in psychology*, 12:646394.

Karen M Douglas, Joseph E Uscinski, Robbie M Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding conspiracy theories. *Political psychology*, 40:3–35.

Francesca D'Errico, Giuseppe Corbelli, Concetta Papapicco, and Marinella Paciello. 2022. How personal values count in misleading news sharing with moral content. *Behavioral Sciences*, 12(9):302.

Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Jonathan St BT Evans and Keith E Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241.

Lisa K Fazio. 2020. Repetition increases perceived truth even for known falsehoods. *Collabra: Psychology*, 6(1).

Daniel J Flynn, Brendan Nyhan, and Jason Reifler. 2017. The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political psychology*, 38:127–150.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Jane Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024. Modeling human subjectivity in LLMs using explicit and implicit human factors in personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7174–7188, Miami, Florida, USA. Association for Computational Linguistics.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019a. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances*, 5(1):eaau4586.

Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019b. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586.

Dan M Kahan. 2017. Misconceptions, misinformation, and the logic of identity-protective cognition.

Carolin Kaiser, Jakob Kaiser, Vladimir Manewitsch, Lea Rau, and Rene Schallner. 2025. Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 82–86.

David Khachaturov, Roxanne Schnyder, and Robert Mullins. 2025. Governments should mandate tiered anonymity on social-media platforms to counter deepfakes and LLM-driven mass misinformation. *arXiv preprint arXiv:2506.12814*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Stephan Lewandowsky, Ullrich KH Ecker, John Cook, Sander Van Der Linden, Jon Roozenbeek, and Naomi Oreskes. 2023. Misinformation and the epistemic integrity of democracy. *Current opinion in psychology*, 54:101711.

Xialing Lin, Patric R Spence, and Kenneth A Lachlan. 2016. Social media and credibility indicators: The effect of influence cues. *Computers in human behavior*, 63:264–271.

Rakoen Maertens, Friedrich M Götz, Hudson F Golino, Jon Roozenbeek, Claudia R Schneider, Yara Kyrychenko, John R Kerr, Stefan Stieger, William P McClanahan, Karly Drabot, et al. 2024. The misinformation susceptibility test (mist): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods*, 56(3):1863–1899.

Ioana E Marinescu, Patrick N Lawlor, and Konrad P Kording. 2018. Quasi-experimental causality in neuroscience and behavioural research. *Nature human behaviour*, 2(12):891–898.

Killian L McLoughlin, William J Brady, Aden Goolsbee, Ben Kaiser, Kate Klonick, and MJ Crockett. 2024. Misinformation exploits outrage to spread online. *Science*, 386(6725):991–996.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why ai is weird and shouldn't be this way: Towards ai for everyone, with everyone, by everyone. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28657–28670.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. 2024. Virtual personas for language models via an anthology of backstories. *Preprint*, arXiv:2407.06576.

Shalini Munusamy, Kalaivanan Syasyila, Azahah Abu Hassan Shaari, Muhammad Adnan Pitchan, Mohammad Rahim Kamaluddin, and Ratna Jatnika. 2024. Psychological factors contributing to the creation and dissemination of fake news among social media users: a systematic review. *BMC psychology*, 12(1):673.

Inderjeet Nair and Lu Wang. 2025. Do language models think consistently? a study of value preferences across varying response lengths. *arXiv preprint arXiv:2506.02481*.

Keiichi Namikoshi, Alex Filipowicz, David A Shamma, Rumen Iliev, Candice L Hogan, and Nikos Arechiga. 2024. Using llms to model the beliefs and preferences of targeted populations. *arXiv preprint arXiv:2403.20252*.

Jan Nehring, Aleksandra Gabryszak, Pascal Jürgens, Aljoscha Burchardt, Stefan Schaffer, Matthias Spielkamp, and Birgit Stark. 2024. Large language models are echo chambers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10117–10123, Torino, Italia. ELRA and ICCL.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50.

Manuel Pratelli and Marinella Petrocchi. 2025. *Evaluating the Simulation of Human Personality-Driven Susceptibility to Misinformation with LLMs*. IOS Press.

Filipe N. Ribeiro, Koustuv Saha, Mahmoudreza Babaei, Lucas Henrique, Johnnatan Messias, Fabricio Benevenuto, Oana Goga, Krishna P. Gummadi, and Elissa M. Redmiles. 2019. On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 140–149, New York, NY, USA. Association for Computing Machinery.

Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *Preprint*, arXiv:2303.17548.

Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Mubashir Sultan, Alan N Tump, Nina Ehmann, Philipp Lorenz-Spreen, Ralph Hertwig, Anton Gollwitzer, and Ralf HJM Kurvers. 2024. Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proceedings of the National Academy of Sciences*, 121(47):e2409329121.

Charles S Taber and Milton Lodge. 2006. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769.

Jasper Timm, Chetan Talele, and Jacob Haimes. 2025. Tailored truths: Optimizing llm persuasion with personalization and fabricated statistics. *arXiv preprint arXiv:2501.17273*.

Jay J Van Bavel, Diego A Reinero, Victoria Spring, Elizabeth A Harris, and Annie Duke. 2021. Speaking my truth: Why personal experiences can bridge divides but mislead. *Proceedings of the National Academy of Sciences*, 118(8):e2100280118.

Santosh Vijaykumar, Yan Jin, Daniel Rogerson, Xuerong Lu, Swati Sharma, Anna Maughan, Bianca Fadel, Mariella Silva de Oliveira Costa, Claudia Pagliari, and Daniel Morris. 2021. How shades of truth and age affect responses to covid-19 (mis) information: randomized survey experiment among whatsapp users in uk and brazil. *Humanities and Social Sciences Communications*, 8(1).

Herun Wan, Jiaying Wu, Minnan Luo, Zhi Zeng, and Zhixiong Su. 2025. Truth over tricks: Measuring and mitigating shortcut learning in misinformation detection. *Preprint*, arXiv:2506.02350.

Aimei Yang, Alvin Zhou, Jieun Shin, Ke Huang-Isherwood, Wenlin Liu, Chuqing Dong, Eugene Lee, and Jingyi Sun. 2024. Sharing is caring? how moral foundation frames drive the sharing of corrective messages and misinformation about covid-19 vaccines. *Journal of Computational Social Science*, 7(3):2701–2733.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Taxonomy Dimensions and Imputed Beliefs

Beliefs are an important tool for modeling demographic simulations in the context of misinformation. Psychological work shows that several belief dimensions, such as conspiracy beliefs, political ideology, trust in science, etc. can help us predict who is most susceptible to misinformation (Ecker et al., 2022; Munusamy et al., 2024). Rather than using only demographic information as a flat category, adding belief profiles provides a more fine-grained representation of how different groups may perceive information. Recent LLM studies also argue for modeling population-level beliefs and preferences to simulate targeted groups, using belief-like representations to approximate human responses at scale (Namikoshi et al., 2024; Kaiser et al., 2025).

A.1 Taxonomy formation

We create a belief taxonomy of beliefs consisting of seven core dimensions that are associated with misinformation susceptibility grounded in psychological and cognitive science research:
(1) Worldview and Identity Beliefs: individuals interpret information through the lens of social identity and worldview (Kahan, 2017; Van Bavel et al., 2021), **(2) Epistemic Trust Beliefs**: individuals differ systematically in epistemic trust toward institutions and experts (De Coninck et al., 2021; Lewandowsky et al., 2023), **(3) Cognitive style**: individuals vary in cognition, such as reliance on analytic versus intuitive reasoning (Pennycook and Rand, 2019; Ecker et al., 2022), **(4) Conspiracy mentality**: individuals differ in a generalized predisposition to conspiracies (Douglas et al., 2019; De Coninck et al., 2021), **(5) Moral and Value Beliefs**: individuals prioritize different moral values (D'Errico et al., 2022; Yang et al., 2024), **(6) Emotion-Related Beliefs**: individuals vary in emotional responsiveness (heightened emotional arousal, such as anger or fear may amplify belief in misinformation) (Brady et al., 2017; McLoughlin et al., 2024) and **(7) Heuristic Beliefs**: individuals rely to different degrees on shortcuts such as repetition, familiarity, and social endorsement (Lin et al., 2016; Fazio, 2020).

While the above dimensions have traditionally been studied in isolation and within human populations, we unify them into a structured taxonomy designed for computational modeling. This taxonomy enables systematic belief simulation in models by providing a framework for collecting and organizing belief data and providing interpretable axes along which belief priors can be instantiated in models.

A.2 Mapping to Belief Taxonomy.

We map imputed data to the above belief dimensions. We apply exploratory factor analysis on WVS responses to identify latent belief groupings, examining which items naturally cluster together (e.g., trust-related items to Epistemic & Trust Beliefs). Post that, we conduct a manual review to ensure proper alignment with our taxonomy and validate the final question-to-dimension mapping. Overall, we obtain 126 imputed belief questions across dimensions. Table 16 contains the 126 WVS questions mapped to the 7 belief taxonomy dimensions.

B Datasets

B.1 Evaluation Data

For susceptibility evaluation, we use two ground-truth datasets containing human judgments of whether they believe a given claim: (1) PANDORA Dataset from Borah et al. (2025b), containing annotations from 318 participants. Each participant provided judgments on 3 distinct claims, along with demographic information including age, gender, living area, and education. These claims are collected from RumorEval (Gorrell et al., 2019), which consists of true or false rumors covering eight major news events and natural disaster events; (2) MIST dataset from Maertens et al. (2024). From MIST, we use only Study 1 (MIST-1) which includes 409 participants, each providing judgment on the same 100 claims, and demographic information including age, gender, and education. From the two datasets combined, we obtain 13.8K claims for evaluation.

B.2 Belief Data

We collect two complementary belief signals, consisting of individually observed belief judgments and group-level (demographic) belief distributions: **(1) Observed Data** - claims that were directly judged by participants in PANDORA and MIST-1 datasets. These responses capture individual-level belief judgments, reflecting each participant's personal stance rather than demographic group averages. We use two claim judgments as observed be-

Dataset Examples

MIST:
Participant ID: 1
Participant Demographic Group: d1
Claim: The Government Is Knowingly Spreading Disease Through the Airwaves and Food Supply
Participant Choice: fake
Gold Label: fake

PANDORA:
Participant ID: 1
Participant Demographic Group: d1
Claim: News: Updated CDC Covid Numbers Show Only Six Percent of Total US Deaths Actually Dies Solely From Virus.
Participant Choice: true
Gold Label: fake

Figure 8: Dataset Examples - MIST and PANDORA

liefs for each participant (keeping it separate from the evaluation data). This leads to 27.6 claim judgments as observed beliefs.
(2) Imputed Data - inferred from the World Values Survey Wave 7[5] distributions. Imputed data represent demographic belief priors (group–level), inferred from WVS, conditioned solely on demographic attributes. We map these imputed belief items to our belief taxonomy using exploratory factor analysis. This yields 126 imputed belief questions and corresponding demographic distributions. Table 1 shows examples, and Appendix A.2 contains mapping details and all questions.

Fig 8 shows an example from each MIST and PANDORA datasets. Both datasets are similar, except for the size. Furthermore, MIST does not contain the living area information of the participants, therefore, we do not use it for fine-tuning purposes.

C Prompt-Based Conditioning

C.1 Prompt Details

Fig 10 shows the susceptibility alignment across the belief and demographic configurations for both datasets. For `Imputed` cases, we average across all belief dimensions, and `Imputed (best)` cases, we select the belief dimension that achieves the highest susceptibility alignment. Findings show that incorporating belief information consistently improves performance over the zero-shot and demographic-only baselines. We next analyze the impact of specific belief types, demographic ablations, datasets, and model performances.

[5] https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp

Prompt

You are a persona grounded by attributes: <demographic group>.

Past beliefs and priors for this persona (for context, do not re-evaluate them): <beliefs>.

When judging a claim, stay consistent with this persona's prior beliefs where reasonable.

Figure 9: Prompt Conditioning Experiments

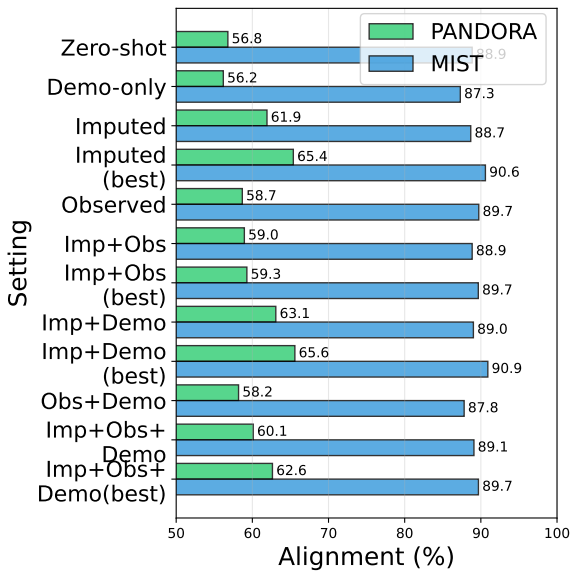

Figure 10: Susceptibility Alignment across settings and datasets

C.2 Belief-Dimension Analysis

Table 4 shows the belief-dimension-wise scores for each dataset and demographic dimension.

Across both datasets, single dimensions consistently outperform the combined prior, and this is especially stark on PANDORA, where all-dimensions collapses to much lower accuracy while the best

single dimensions remain high.

For Prolific, the strongest modal scores are generally obtained from conspiracy mindset and epistemic/trust beliefs, especially for gender, age, and education. For MIST, cognitive-style beliefs and conspiracy mindset are the strongest individual dimensions, with cognitive-style beliefs reaching the highest education-based score and conspiracy mindset performing best for age. The all-dimensions setting is not uniformly strongest, suggesting that compact dimension-specific belief priors can be more informative than combining all belief dimensions at once. Rural/urban results are omitted in the updated modal table because the corresponding screened runs were unavailable or excluded.

C.3 Dataset differences - PANDORA vs MIST

Lower accuracies on PANDORA are expected because Prolific is currently much smaller than MIST, even though both evaluate individual-level beliefs. The smaller Prolific set has less coverage across claims, demographics, and belief patterns, making estimates more sensitive to participant-level variation and noise. In contrast, the larger MIST dataset provides broader coverage and more stable training/evaluation signals, which can lead to higher accuracies.

C.4 Thematic Analysis

We perform topic clustering on the news claims to identify latent topical grouping and investigate whether LLM susceptibility alignment vary across demographics within each topic. We do this for three models – Llama-3-8b, Mistral-7b and Qwen-2.5-14b. We perform this on Dataset 2, as it is much larger and diverse than Dataset 1. Using non-negative matrix factorization on claim texts, we identify five dominant topics: (1) global power, population, and future threats; (2) political leaders and historical rankings; (3) science, health, and technology claims; (4) politics, government, and information-control narratives; and (5) public opinion, moral values, and social belief statements. Topic-level analysis shows that susceptibility alignment is not uniform across demographics. Topics related to political rankings and public opinion (Topics 2 and 5) achieve uniformly high accuracy across all demographic groups, resulting in negligible age, gender, or education gaps. Contrastingly, science/health claims and government-related narratives (Topics 3 and 4) exhibit the largest demographic variation, with older and higher-education groups generally achieving higher accuracy. Gender-based differences are minimal across all topics. Table 5 shows results across some topics and models showing differences across demographics.

These findings suggest that demographic susceptibility in LLMs is not just a broad, identity-driven phenomenon but also contextual and topic-sensitive. This also aligns with interdisciplinary work in psychology and communication suggesting that demographic susceptibility to misinformation is highly context-dependent. Prior studies show that gender alone is a weak predictor of misinformation belief, with analytic thinking and reasoning style playing a much larger role (Marinescu et al., 2018). In contrast, age and education demographics differ more for complex or ambiguous claims, particularly in science and policy domains, consistent with work on cognitive processing, and motivated reasoning (Kahan, 2017). From a cognitive perspective, this pattern supports dual-process theories of reasoning, whereby strong lexical reduce individual differences, while inference-heavy claims amplify them (Evans and Stanovich, 2013).

C.5 Experiments using belief-distributions

In our main experiments, we utilize modal responses for prompt conditioning. Here, we perform additional experiments that use belief distributions instead of modal responses in prompts. Table 6 compares modal and distribution-based prompting across models on PANDORA and MIST.

Overall, distribution-based conditioning is competitive with modal prompting, but does not consistently outperform it. Across paired comparisons, distributional prompts sometimes improve imputed-only or imputed+observed settings, suggesting that population-level uncertainty can provide useful signal. However, these gains are model- and dataset-dependent. In particular, modal prompting is more stable when imputed beliefs are combined with demographics, which is our strongest setting in the main experiments. This pattern suggests that full distributions may introduce additional ambiguity: they encode group-level uncertainty, however, the prediction target is a binary individual judgment. As a result, models may struggle to decide whether to follow the majority response, sample from the distribution, or treat minority probabilities as competing evidence.

We therefore use modal belief responses in the

Dataset	Dimension	Gender	Age	Education	Rural/Urban
Prolific-dataset	worldview_and_identity	55.69	55.88	55.51	55.80
	epistmetic_and_trust_beliefs	58.23	58.82	58.90	58.65
	cognitive_style_beliefs	55.24	57.35	55.72	56.10
	conspiracy_mindset	59.13	58.24	59.53	58.90
	moral_and_value_beliefs	55.99	55.00	54.66	55.20
	emotion_related	55.39	58.24	54.66	56.10
	heurisitics	54.94	59.41	54.24	56.53
	alldimensions	57.12	58.24	56.67	57.50
MIST-dataset	worldview_and_identity	75.99	76.17	76.23	
	epistmetic_and_trust_beliefs	74.84	73.61	76.10	
	cognitive_style_beliefs	79.48	78.81	81.17	
	conspiracy_mindset	77.85	80.39	79.17	
	moral_and_value_beliefs	74.05	75.29	73.91	
	emotion_related	75.14	72.00	72.82	
	heurisitics	75.89	76.29	75.93	
	alldimensions	86.00	74.50	73.86	

Table 4: Belief Dimension Analysis using modal runs.

main experiments for three reasons: (1) they provide a simpler and more interpretable representation of demographic belief priors. (2) they reduce prompt length and avoid requiring the model to perform implicit probabilistic reasoning over Likert distributions. (3) they lead to more stable performance in the main combined settings, especially when beliefs are paired with demographic information. Thus, while distributional prompts are a useful robustness check, modal responses provide a cleaner and more reliable prompt representation for our primary experiments given the models.

C.6 Distributional alignment within groups

We use Jensen–Shannon (JS) divergence across demographic groups to investigate how models can capture the distributions in demographic groups. Distributional representations indicate how much uncertainty remains within each group, while modal representations show what happens when this uncertainty is collapsed to a single dominant response. We also report modal disagreement, the percentage of WVS items for which demographic groups have different modal responses.

Table 7 shows that full distributional representations retain substantially more uncertainty, producing lower JS divergence across demographic groups. In contrast, modal prompts collapse uncertainty and produce sharper demographic separation. This does not mean modal prompts are always more faithful to the underlying population; rather, they provide a compact approximation of the dominant belief signal. The results therefore highlight a trade-off: modal prompts are simpler and more stable for prompt-conditioning, while distributional representations better preserve within-group uncertainty. This motivates our additional distributional prompting experiments and our use of belief distributions in BeliefSim-FT, where uncertainty can be represented more directly through distribution-level training.

C.7 Evaluation Under Human Judgment Variability

Our main evaluation metric treats each participant judgment as the prediction target, reflecting our goal of simulating individual-level susceptibility rather than estimating a single claim-level consensus label. However, susceptibility judgments are inherently noisy: participants may disagree on the same claim, and some claims may have higher response entropy than others. We therefore interpret susceptibility alignment as agreement with observed individual responses, not as recovery of an objective or consensus truth label.

To better contextualize model performance, we examine whether accuracy varies with claim-level entropy, estimated from the variability of human responses to each claim (Table 8). Higher entropy indicates greater participant disagreement and therefore a noisier prediction target. For PANDORA, accuracy decreases from 58.2% on low-

Demographic	Model	Topic/Theme	Effect and Notes
Age	Llama-3-8B	New study & ideology (T1)	+2.8 ppts (60+ <30). Older: 0.984 vs younger: 0.956.
	Llama-3-8B	Global threats & population (T2)	+2.0 ppts. Older: 0.961 vs younger: 0.941.
	Llama-3-8B	Marijuana/new-study (T4)	+1.7 ppts. 60+ group: 0.947 vs <30: 0.930.
	Qwen-2.5-14B	Marijuana/new-study (T4)	+2.5 ppts. 60+ group: 0.917 vs <30: 0.892.
	Mistral-7B	All topics	<0.5 ppts. Age differences negligible.
Education	Llama-3-8B	Sleep/blue-light health (T2)	+8.4 ppts. Completed: 0.918 vs not: 0.833 (six cases).
	Qwen-2.5-14B	Politics/elections (T3)	+16.7 ppts. Completed perfect; not-completed: 0.833 (six cases).
	Llama-3-8B	Marijuana/new-study (T4)	+0.9 ppts. Difference: 0.962 vs 0.953.
	Mistral-7B	Marijuana/new-study (T4)	+0.2 ppts. Effect negligible.
	All models	Other topics	0.0–0.2 ppts. No meaningful education effect.
Gender	Llama-3-8B	Global influence/future threats (T0)	≤ 2 ppts ($\lvert F - M \rvert$). Nearly identical accuracies.
	Llama-3-8B	Science & health claims (T2)	≤ 2 ppts. No consistent female–male separation.
	Llama-3-8B	Public opinion/social beliefs (T4)	≈ 0 ppts. Near-ceiling for both genders.
	Mistral-7B	Global influence/future threats (T0)	≈ 0 ppts. Accuracies overlap almost exactly.
	Mistral-7B	Marijuana/public opinion (T4)	≤ 1 ppt. No directional trend.
	Qwen-2.5-14B	All topics	0 ppts. Identical accuracies across all topics.

Table 5: Thematic Analysis across Demographics.

entropy claims to 42.9% on high-entropy claims, suggesting that ambiguity in participant judgments partly explains model errors. MIST remains comparatively stable across entropy bins, with only a mid-entropy dip. This distinction is important because it separates model limitations from intrinsically noisy human targets. The results suggest that ambiguity partly explains PANDORA errors, whereas screened MIST performance remains comparatively stable across agreement levels.

C.8 Veracity Prediction vs. Susceptibility Prediction

We additionally compare zero-shot veracity accuracy with belief-conditioned susceptibility alignment (best setting). Veracity accuracy measures whether the model prediction matches the gold truth label of a claim, while susceptibility alignment measures whether the model prediction matches a participant's observed judgment. These are related but distinct tasks: a model can be factually correct while failing to simulate whether a participant believes the claim.

Table 9 compares two distinct evaluation targets: veracity prediction, which measures agreement with the gold truth label, and susceptibility prediction, which measures agreement with participants' observed judgments. The results show that these objectives do not always move together. For several open-source models, belief-conditioned susceptibility alignment is higher than zero-shot veracity accuracy, especially on PANDORA, suggesting that belief conditioning can help approximate human judgments beyond factual correctness alone. However, the pattern is not uniform: `Grok-4-Fast` has strong veracity accuracy but lower susceptibility alignment, while some MIST models show only small differences between the two metrics.

Overall, this comparison highlights that factual veracity correctness and human-belief simulation capture different aspects of misinformation response. Veracity prediction evaluates whether a model identifies the correct truth status of a claim, while susceptibility prediction evaluates whether it can approximate how a specific participant judges that claim. Thus, high performance on one objective does not necessarily imply high performance on the other.

C.9 Demographic Counterfactual Analysis

Here, we provide flip rates across the three metrics in Table 10. Overall, `Mistral` shows the lowest flip rates, suggesting limited dependence on demographic cues. `Qwen` is also comparatively stable and low. In contrast, `Llama` and `Deepseek` models show substantially higher demographic sensitivity. Across models, the largest flip rates concentrate in Education and Living Area, especially in the Shortcut Reliance setting, where demographics are designed to be non-predictive. This could also relate to stereotype-driven reliance.

Qualitative evidence of shortcut reasoning.. To complement the flip-rate analysis, Table 12 shows examples where model outputs explicitly refer-

Data	Model	Belief Type	Zero-shot	Demo-only	Imputed	Observed	Imp+Obs	Imp+Demo
PANDORA	Qwen2.5-14B	modal	62.28	62.28	64.51	62.08	63.49	**64.81**
PANDORA	Qwen2.5-14B	dist	62.28	62.28	63.51	62.08	62.79	**63.91**
PANDORA	Llama-3-8B	modal	57.49	56.89	60.63	58.98	60.01	**60.93**
PANDORA	Llama-3-8B	dist	57.49	56.89	59.63	58.98	59.31	**60.03**
PANDORA	Llama-3.2-3B	modal	53.89	57.49	56.76	**60.78**	58.81	55.96
PANDORA	Llama-3.2-3B	dist	53.89	57.49	60.43	**60.78**	59.77	58.74
PANDORA	Grok-4-Fast	modal	57.07	56.60	58.20	57.00	57.85	**58.55**
PANDORA	Grok-4-Fast	dist	57.07	56.60	57.65	57.00	57.40	**57.95**
PANDORA	Mistral-7B-v0.2	modal	55.99	54.19	55.82	55.79	55.82	**56.12**
PANDORA	Mistral-7B-v0.2	dist	**55.99**	54.19	54.82	55.79	55.30	55.22
PANDORA	OLMo-2-7B	modal	50.60	49.40	49.43	**58.38**	50.51	50.23
PANDORA	OLMo-2-7B	dist	50.60	49.40	52.14	**58.38**	52.57	52.54
PANDORA	DeepSeek-7B	modal	52.40	50.00	52.53	41.32	48.96	**53.33**
PANDORA	DeepSeek-7B	dist	52.40	50.00	52.85	41.32	48.04	**53.25**
MIST	Qwen2.5-14B	modal	95.97	95.97	97.41	95.77	96.59	**97.72**
MIST	Qwen2.5-14B	dist	95.97	95.97	96.98	95.77	96.37	**97.18**
MIST	Llama-3-8B	modal	**93.32**	91.70	92.51	86.56	92.51	93.00
MIST	Llama-3-8B	dist	93.32	91.70	91.51	86.56	**93.59**	92.10
MIST	Llama-3.2-3B	modal	91.08	88.96	87.13	76.82	**94.13**	87.49
MIST	Llama-3.2-3B	dist	91.08	88.96	88.69	76.82	**93.87**	88.29
MIST	Grok-4-Fast	modal	87.20	86.70	88.30	87.05	87.95	**88.60**
MIST	Mistral-7B-v0.2	modal	71.91	68.23	73.67	71.71	72.89	**73.97**
MIST	Mistral-7B-v0.2	dist	71.91	68.23	72.67	71.71	72.19	**73.07**
MIST	OLMo-2-7B	modal	79.83	81.49	77.35	80.83	80.15	**82.29**
MIST	OLMo-2-7B	dist	79.83	81.49	80.38	80.83	**82.23**	81.89
MIST	DeepSeek-7B	modal	63.97	64.64	63.61	56.46	62.16	**65.44**
MIST	DeepSeek-7B	dist	63.97	64.64	62.55	56.46	64.43	**65.04**

Table 6: Ablations comparing modal belief conditioning and belief-distribution conditioning across datasets and models. Scores are averaged over demographic axes. Best value per row is in bold.

Axis	Dist. JS	Modal JS	Modal Disagree.
Age	0.0421	0.2511	25.53
Education	0.0498	0.2454	24.96
Gender	0.0083	0.1175	11.95
Rural/Urban	0.0072	0.0958	9.76

Table 7: Average demographic divergence for distribution-based and modal belief prompt representations. JS divergence is reported in bits; modal disagreement reports the percentage of WVS items for which demographic groups have different modal responses.

Dataset	Low	Mid	High
PANDORA	58.24	51.03	42.95
MIST	86.13	78.13	87.77

Table 8: Modal prompt-conditioning accuracy by claim-level entropy.

ence demographic attributes in Panel B, where demographics are non-predictive by construction. These examples illustrate that some models not only change predictions under demographic swaps, but also justify their decisions using demographic assumptions about education, gender, age, or living area. This supports our interpretation that high flip rates in the shortcut-reliance setting can reflect demographic shortcutting or stereotype-like reasoning. However, these excerpts are used only as qualitative evidence.

C.10 Factual Confidence and Susceptibility Simulation

A potential explanation for the limited gains from belief conditioning is that LLMs' pretraining knowledge about claim veracity acts as a compounding bias: when a model is confident a claim is false, it may resist simulating a persona that would nonetheless believe it. We analyze this hypothesis in this section.

Method. For each claim in MIST-1, we extract a zero-shot factual confidence score using a logit-based probe: we compute the renormalized probability mass assigned to `true` vs. `false` at the first generated token position, yielding a scalar in $[0.5, 1.0]$ indicating confidence in the predicted label. We then compute the Spearman rank correlation (ρ) between this confidence score and whether the model's predicted label matches the human's

Dataset	Model	Veracity	Belief-Susc.	Δ
PANDORA	Qwen2.5-14B	62.28	64.81	+2.53
PANDORA	Llama-3-8B	57.49	60.93	+3.44
PANDORA	Llama-3.2-3B	53.89	60.78	+6.89
PANDORA	Grok-4-Fast	66.77	58.55	-8.22
PANDORA	Mistral-7B-v0.2	55.99	56.12	+0.13
PANDORA	OLMo-2-7B	50.60	58.38	+7.78
PANDORA	DeepSeek-7B	52.40	53.33	+0.93
MIST	Qwen2.5-14B	96.97	97.72	+0.75
MIST	Llama-3-8B	93.32	93.00	-0.32
MIST	Llama-3.2-3B	91.08	94.13	+3.05
MIST	Grok-4-Fast	97.70	88.60	-9.10
MIST	Mistral-7B-v0.2	71.91	73.97	+2.06
MIST	OLMo-2-7B	79.83	82.29	+2.46
MIST	DeepSeek-7B	63.97	65.44	+1.47

Table 9: Comparison between zero-shot veracity accuracy and belief-conditioned susceptibility alignment. Veracity measures agreement with the gold truth label, while belief-susceptibility measures agreement with human judgments under belief- conditioned prompting.

Model	Panel A: Utility				Panel B: Shortcut				Panel C: Complementarity			
	Gender	Area	Age	Edu.	Gender	Area	Age	Edu.	Gender	Area	Age	Edu.
Mistral-7B-v0.2	0.13	0.67	0.07	0.64	2.38	0.00	2.78	0.00	3.11	0.35	3.25	2.39
Qwen3-4B	0.40	0.67	0.59	2.41	0.60	1.25	2.78	0.00	1.56	12.32	7.02	5.38
OLMo-2-7B	2.00	7.67	1.45	2.20	7.74	12.66	0.00	0.00	5.20	9.51	11.34	10.76
Qwen2.5-14B	4.41	2.33	0.72	1.99	2.98	7.50	0.00	0.00	5.02	4.93	2.91	3.59
Llama-3.2-3B	10.28	6.00	12.15	12.78	8.33	5.00	2.78	16.67	2.42	9.64	3.81	5.62
DeepSeek-7B	5.87	13.00	9.65	22.87	7.14	17.50	2.78	25.00	2.60	8.45	10.79	11.75
Llama-3-8B	1.80	8.00	10.51	32.46	8.33	6.25	13.89	33.33	6.06	8.45	3.25	5.38

Table 10: Counterfactual flip rates (%) across demographic groups. Panel A measures demographic-swap sensitivity on the original distribution, Panel B measures shortcut reliance on balanced slices, and Panel C measures complementarity under degraded beliefs.

Condition	Llama	Qwen	Mistral
Zero-shot	0.115	0.107	0.081
Demo	0.042	0.107	0.028
Imputed	−0.023	0.107	0.028
Imputed + Demo	−0.023	0.107	0.028
Observed	−0.025	0.031	0.034
Observed + Demo	−0.023	0.036	0.034
Imp. + Obs.	−0.020	0.032	0.034
Imp. + Obs. + Demo	−0.025	0.058	0.024

Table 11: Spearman ρ between zero-shot factual confidence and human-label match rate on MIST-1, restricted to false claims. Positive values indicate that higher model confidence in a claim's falsity correlates with better matching of human judgments overall; near-zero or negative values indicate decoupling from the factual prior.

belief judgment, for each (model, condition) pair. We restrict the analysis to *false claims only* (gold label: misinformation), since factual override is operative precisely when the model's factual prior conflicts with the susceptible human's belief.

Results. Table 11 reports Spearman ρ across all prompting conditions. On MIST-1, zero-shot correlations are small but consistently positive across all models (Llama: 0.115, Qwen: 0.107, Mistral: 0.081), confirming that factual confidence *somewhat* biases models toward predicting *fake*—which coincidentally aligns with the skeptical majority of annotators but at the cost of failing susceptible individuals. As belief conditioning is introduced, Llama and Mistral shift toward zero and slightly negative values under imputed conditions, suggesting partial decoupling from the factual prior. Qwen, however, remains at $\rho = 0.107$ across all modal conditions, consistent with its lower sensitivity to prompt-based conditioning observed in the counterfactual analysis (§5.2). Observed belief conditions bring all models near zero, indicating that individual-level belief signal is most effective at overriding the factual prior.

Taken together, factual confidence introduces a modest positive bias in zero-shot settings but is not the dominant factor limiting susceptibility simulation. Belief conditioning—particularly with

Dataset	Model	Axis	Hit rows	Representative raw-output evidence
MIST-1	Mistral-7B-Instruct-v0.2	Education	36/36	"A person with a **completed education is less likely to believe in such a sensational and unsubstantiated claim**."
MIST-1	Mistral-7B-Instruct-v0.2	Education	36/36	"This person, with **not completed education**, is **less likely to have the knowledge or ability to make an informed judgment**."
MIST-1	Mistral-7B-Instruct-v0.2	Gender	30/54	"This person, being **male**, is likely to believe headlines that discuss **international politics and economic transactions.**"
MIST-1	Mistral-7B-Instruct-v0.2	Gender	29/54	"This person, a **female**, is likely to be informed about **global news and economics.**"
PANDORA	OLMo-2-1124-7B-Instruct	Living Area	2/80	"The person living in a **rural area** might be more inclined to trust the news due to their **exposure to fewer alternative sources**."
PANDORA	Llama-3.2-3B-Instruct	Living Area	10/80	"The person's perspective is influenced by their **living area, which is rural,** and may **not be familiar with the specifics of the Ferguson incident**."

Table 12: Qualitative examples of explicit demographic references in raw outputs for Panel B shortcut-reliance cases. Counts indicate rows whose raw output mentions the corresponding demographic axis; excerpts are illustrative and are not used as the primary metric.

observed beliefs—largely neutralizes it, suggesting the broader simulation gap is driven by other mechanisms.

D Fine-tuning Analysis

D.1 MIST-1/MIST-2b overlap discussion

Our fine-tuning experiments use MIST-1 for training and both MIST-1 and MIST-2b for evaluation. Since the MIST benchmark is a standardized psychometric instrument, MIST-2b contains items derived from the MIST-1 item-development pool. Therefore, the MIST-2b evaluation is not a claim-disjoint test of generalization to entirely unseen misinformation claims. Instead, it evaluates cross-study and cross-participant generalization under a shared measurement instrument. This distinction is important because, for our susceptibility simulation setting, the primary leakage risk is participant-level leakage: if the same participant appeared in both training and evaluation, the model could learn participant-specific response tendencies rather than generalizable belief-susceptibility mappings. We avoid this by using independent participant samples across MIST-1 and MIST-2b. Claim overlap is less central to this particular evaluation because the goal is to predict individual susceptibility responses under a validated fixed instrument, rather than to test open-domain claim generalization. We therefore report MIST-2b results as shared-instrument cross-participant transfer, and leave fully claim-disjoint evaluation as a stricter future setting.

D.2 Phase 1 results

For Phase 1, Table 13 shows that Qwen has the lowest KL divergence, followed by Mistral and Llama for both settings - head tuning and full-fine-tuning. With LoRA-ft, the trends remain similar but slightly lower than head-ft. This indicates that Qwen more faithfully captures population-level belief distributions, a critical property for reliable downstream simulation of demographic misinformation susceptibility. Other models also perform competitively, suggesting that most models LLMs can capture belief distributions to a reasonable extent.

Model	Head Tuning	Full Fine-tuning
Qwen	0.0511	0.0732
Mistral	0.087	0.107
Llama	0.287	0.487

Table 13: KL divergence comparison across different models and tuning methods.

D.3 Phase 2 results

Table 14 shows susceptibility alignment and F1 scores across different setups and models. Across methods, the key finding is that cross-study (MIST-2) generalization improves drastically once we move from 1-phase to 2-phase pipelines. The 1-phase baseline model perform decent, however, the performance drops sharply on MIST-2. For MIST-2, our BeliefSim-FT approach performs the best, showing that belief adapter + lightweight sus-

Model	PANDORA+MIST-1		MIST-2	
	Acc	F1	Acc	F1
Baseline (1-phase)				
Llama	0.750	0.751	0.680	0.673
Qwen	0.760	0.761	0.720	0.712
Mistral	0.740	0.761	0.650	0.653
LoRA-FT (2-phase)				
Llama	0.809	0.809	0.876	0.861
Qwen	0.800	0.800	0.893	0.892
Mistral	0.800	0.790	0.887	0.867
BeliefSim-FT				
Llama	0.793	0.793	0.884	0.896
Qwen	0.792	0.791	0.924	0.906
Mistral	0.792	0.792	0.884	0.823

Table 14: Fine-tuning performances

ceptibility head captures more transferable signals aligned with out-of-domain data.

Comparing the 2-phase variants, we find that BeliefSim-FT remains the strongest on MIST-2, however, LoRA-ft slightly improves on PANDORA+MIST-1 but gives weaker transfer. This may be due to mild overfitting when more parameters are updated. Across models, Qwen consistently performs the best for 2-phase settings. Overall, the 2-phase training approach is helpful improving robustness and also adapting to out-of-domain data.

D.4 Shortcut Reliance Experiments

Table 15 shows that shortcut reliance is low, especially under head tuning. We measure flip-rate: the fraction of evaluation examples for which the model's predicted label changes when a shortcut feature is removed or perturbed, and Probability Delta: the average absolute change in the model's predicted probability for the original class after the shortcut is removed. On PROLIFIC+MIST-1, head-tuned models (Llama, Qwen and Mistral) show near-zero flip rates, indicating that removing the shortcut signal has little effect on predicted labels. Full fine-tuning also tends to have small flip rates and probability deltas in-domain, but they are more frequently non-zero (e.g., Llama/Mistral flip more than head tuning), consistent with the idea that updating more parameters can slightly increase sensitivity to spurious cues.

On MIST-2, the most notable shortcut sensitivity appears for Qwen under full fine-tuning (flip rate 0.1098, larger prob delta), whereas the corresponding head-tuned Qwen model shows 0 flip rate. This pattern supports the our previous findings that two-phase-head-only training is more robust: it preserves performance while reducing reliance on shortcut features that do not transfer.

Belief Modeling Prompt

You are modeling belief distributions for a single demographic group.

Demographic : d1

Question: q1

There are k possible response options, numbered 1 through k. Predict the probability for each option.

Figure 11: Prompt for Phase-1 Belief Modeling

D.5 Training Details

D.5.1 2-phase Head-FT details

Phase 1 trains a belief adapter to predict demographic-conditioned WVS response distributions. The model is a frozen encoder with a trainable linear head mapping the last-token hidden state to 10 logits, followed by softmax. Training uses AdamW (lr 5×10^{-4}, batch size 16, 2 epochs) and optimizes a scale-aware KL divergence loss, computed only over the valid bins 1..K for each example. For evaluation, we test generalization to unseen belief items, reporting KL distributional fit and majority-category accuracy.

Phase 2 trains a susceptibility prediction head on top of a frozen belief adapter learned in Phase 1. The model combines a frozen base encoder and frozen belief head to produce both a semantic representation of the input and a demographic-conditioned belief probability vector. These two are concatenated and fed into a lightweight, trainable classification head. We format the inputs as instruction-following prompts that include a demographic persona and up to two belief examples. During training, only the susceptibility head parameters are updated using cross-entropy loss, while the base model and belief adapter remain frozen. The model is optimized with AdamW (learning rate (5×10^{-4}), batch size 8) for 2 epochs, with reproducible initialization via fixed random seeds.

Method	Model	PANDORA+MIST-1			MIST-2		
		Flip Rate	Prob Delta	Acc Drop	Flip Rate	Prob Delta	Acc Drop
Full-FT	Llama	0.0239	0.0148	0.0049	0.0000	0.0000	0.0000
	Qwen	0.0173	0.0476	0.0012	0.1098	0.0700	0.1027
	Mistral	0.0323	0.0024	0.0000	0.0000	0.0129	0.0000
Head Tuning	Llama	0.0000	0.0125	0.0000	0.0000	0.0129	0.0000
	Qwen	0.0080	0.0261	0.0012	0.0000	0.0427	0.0000
	Mistral	0.0000	0.0071	0.0000	0.0000	0.0069	0.0000

Table 15: Shortcut reliance comparison between Full Fine-tuning and Head Tuning across different models on PANDORA+MIST-1 and MIST-2 datasets.

We finally evaluate at each epoch using accuracy and macro-F1.

D.5.2 2-phase LoRA-FT details

Here, we use Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning in both phases. In Phase 1, LoRA enables the model to adapt its representations to capture demographic belief distributions. In Phase 2, LoRA is again applied to the base model during susceptibility training, allowing the model to adjust task-relevant representations while keeping the number of trainable parameters small.

E Significance Tests Across Experiments

Prompt-based Conditioning (susceptibility alignment): Conditioning LLMs with belief information improves demographic susceptibility simulation compared to zero-shot and demo-only prompts. We averaged our findings across 3 individual runs, and improvements of belief settings are statistically significant (paired t-test, $p < 0.05$ for both PANDORA and MIST-1).

Belief Dimension Analysis. Certain belief dimensions are more predictive of susceptibility than others. We use repeated-measures ANOVA and performance variation vary significantly by belief dimensions, more for PANDORA, with a few non-significant variation for MIST (congitive, and alldimensions).

Post-Training Adaptation In PANDORA+MIST-1, accuracy differences between baseline and the 2-phase methods are modest (e.g., 0.75 vs. 0.793 for Llama), and z-tests show that both BeliefSim-FT and LoRA fine-tuning significantly outperforms the baseline (p=0.02–0.03). On MIST-2 the gains are pronounced: for Llama the baseline accuracy is 0.680 while BeliefSim-FT reaches 0.884 and LoRA-FT 0.876. Two-proportion z-tests show that both BeliefSim-FT and LoRA-FT significantly outperform the baseline ($p<0.001$).

F Model Choices, Implementation Details and Computational Resources

We evaluate a diverse set of instruction-tuned models to cover variation in model family, scale, and alignment behavior. Specifically, we include strong open-source chat models from different families (`Qwen`, `Llama`, `Mistral`, `OLMo`, `Grok`) and `DeepSeek`) to test whether belief-based conditioning is robust beyond a single architecture or training recipe. Due to resource constraints, we do not include substantially larger open-source models; instead, we focus on models in the 3B–14B range that support repeated prompting, ablations, and counterfactual evaluations at scale. Finally, `Grok-4-Fast` provides a stronger closed-source comparison point. This model set therefore lets us compare both performance and demographic sensitivity across heterogeneous instruction-following systems.

We conduct experiments using instruction-tuned LLMs including Llama, Qwen, and Mistral, implemented with the Hugging Face Transformers and PEFT libraries. Our framework adopts a modular design with explicit belief heads and susceptibility heads, and employs LoRA-based fine-tuning where base-model adaptation is required, while relying on lightweight head training in BeliefSim-FT to decouple belief modeling from susceptibility prediction. All experiments are run on NVIDIA A40 GPUs with a maximum sequence length of 1024 tokens and fixed random seeds to ensure reproducibility and computational efficiency.

G Reproducibility

We open-source our codes and data, which are uploaded to the submission system. This would help future work to reproduce our results and explore BeliefSim: demographic-aware misinformation susceptibility simulation in LLMs using belief-priors, and BeliefSim-FT.

Table 16: World Values Survey Questions mapped to Misinformation Taxonomy Dimensions

Dimension	Q ID	Question
WorldView and Identity Beliefs		
	Q6	Importance of religion (1=Very important to 4=Not at all important)
	Q4	Importance of politics (1=Very important to 4=Not at all important)
	Q171	Frequency of religious service attendance (1=More than once/week to 7=Never)
	Q173	Religious self-identification (1=Religious person, 2=Not religious, 3=Atheist)
	Q240	Political left-right scale (1-10)
	Q254	National pride (1=Very proud to 4=Not at all proud, 5=Not from US)
	Q255-Q259	Closeness to village/region/country/continent/world (1=Very close to 4=Not close at all)
	Q19, Q23	Unwanted neighbors: different race, different religion (1=Don't want, 2=Want)
	Q235-Q239	Views on government styles: strong leader, technocracy, army rule, democracy, religious law (1=Very good to 4=Very bad)
	Q241-Q249	Essential characteristics of democracy: tax/subsidize, religious law interpretation, free elections, unemployment aid, army takeover, civil rights, income equality, obedience, women's rights (1-10 scale)
Epistemic and Trust Beliefs		
	Q57	General trust in people (1=Most can be trusted, 2=Need to be careful)
	Q58-Q63	Trust in family, neighborhood, people you know, first-time meetings, other religions, other nationalities (1=Trust completely to 4=Don't trust at all)
	Q69, Q71, Q75	Confidence in police, government, universities (1=A great deal to 4=None at all)
	Q158-Q163	Views on science and technology: health/comfort, opportunities, faith vs science, moral breakdown, daily relevance, world better/worse (1-10 scale)
Cognitive Style Beliefs		
	Q152-Q153	Priority ranking: economic growth, defense forces, participation in decisions, beautification (select most and next most important)
	Q154-Q155	Priority ranking: order, participation, fighting prices, free speech (select most and next most important)
	Q156-Q157	Priority ranking: stable economy, humane society, ideas over money, fight crime (select most and next most important)
	Q176	Moral uncertainty: trouble deciding which moral rules are right (1-10 scale)
Conspiracy Mindset		
	Q112	Perceived corruption level in country (1=No corruption to 10=Abundant corruption)
	Q113-Q117	Corruption beliefs about state authorities, business executives, local authorities, civil service providers, journalists/media (1=None to 4=All of them)
	Q118	Frequency of bribery needed for services (1=Never to 4=Always)
Moral and Value Beliefs		
	Q176	Moral uncertainty (1-10 scale)
	Q177-Q195	Justifiability of: claiming unentitled benefits, fare evasion, stealing, tax cheating, bribery, homosexuality, prostitution, abortion, divorce, premarital sex, suicide, euthanasia, wife beating, child beating, violence, terrorism, casual sex, political violence, death penalty (1=Never justifiable to 10=Always justifiable)
Emotion Related		
	Q44-Q45	Views on technology development and respect for authority (1=Good, 2=Don't mind, 3=Bad)
	Q46	Happiness level (1=Very happy to 4=Not at all happy)
	Q47	Health status (1=Very good to 5=Very poor)
	Q48	Freedom of choice and control over life (1=No choice to 10=Great deal of choice)
	Q49-Q50	Life satisfaction and financial satisfaction (1=Completely dissatisfied to 10=Completely satisfied)
	Q52	Felt unsafe from crime in last 12 months (1=Often to 4=Never)
	Q131	General security feeling (1=Very secure to 4=Not at all secure)
	Q146-Q148	Worry about war, terrorist attack, civil war (1=Very much to 4=Not at all)
Heuristic		
	Q94-Q104	Active membership in: church/religious, sports/recreation, arts/music/education, labor union, political party, environmental, professional, humanitarian/charity, consumer, self-help/mutual aid, women's groups (1=Inactive, 2=Active, 3=Don't belong)
	Q201-Q208	Media/communication usage: newspaper, TV news, radio news, mobile phone, email, internet, social media, talking with friends (1=Daily to 5=Never)
	Q18-Q26	Unwanted neighbors: drug addicts, different race, AIDS, immigrants, homosexuals, different religion, heavy drinkers, unmarried couples, different language (1=Don't want, 2=Want)
	Q29-Q31	Gender attitudes: men as better political leaders, university more important for boys, men as better business executives (1=Strongly agree to 4=Strongly disagree)
	Q33-Q35	Job/gender attitudes: men's right to jobs when scarce, natives over immigrants, woman earning more causes problems (1=Strongly agree to 5=Strongly disagree)