vulnerable to prompt-driven epistemic failures, even when they internally encode correct information.

The data indicate that these vulnerabilities resemble patterns of naive human reasoning rather than expert cognition. Like laypeople, LLMs are highly sensitive to social and linguistic cues that systematically shift responses toward weaker correction independent of evidential quality, mirroring well-documented framing and source-cue effects in lay human judgment[40,41]. Experts, by contrast, maintain epistemic vigilance regardless of such contextual signals[42,43]. Evidently, current architectures reproduce human-like heuristics rather than expert-level reasoning, reinforcing fragility rather than correcting it. This is concerning in cases where, for example, individuals with polarized political ideologies who are more susceptible to certain types of misinformation[44] often communicate with lower epistemic openness[45,46]. Our results suggest that such assertive, closed prompting may reduce the likelihood of correction, inadvertently reinforcing misinformation among high-risk groups. This underscores the need for LLMs to be calibrated not only to factual accuracy but also to the epistemic and communicative context of the user.

Models differed in both correction strength and strategy use. Claude Sonnet 4.5 consistently delivered strong refutations, while Gemini 2.5 Pro was more hesitant, often signaling mild skepticism. Although all models drew from similar strategies, their frequency varied: Grok-4 and ChatGPT-5 favored “call-to-verify” tactics, whereas Claude Sonnet 4.5 and Gemini 2.5 Pro leaned on empathetic framing. Notably, ChatGPT-5 employed less effective strategies than all others, suggesting that differences in stance and strategy reflect design priorities, such as balancing persuasion with user rapport.

These variations underscore a broader challenge: epistemic fragility in LLMs. A model that hesitates to correct misinformation even when evidence is clear risks amplifying uncertainty rather than reducing it. Conversely, overly assertive models may alienate users or fail to accommodate nuanced contexts. Standardizing approaches to epistemic fragility could involve clearer calibration of stance confidence, which may include integrating confidence thresholds tied to evidence quality or harmonizing strategy use so that corrective efforts remain consistent across systems. Our stance-and-strategy framework offers one candidate set of metrics that could be integrated into such calibration pipelines. Ultimately, these findings highlight the need for guidelines that balance assertiveness with adaptability.

Differences in stance strength and strategy use across models pose safety risks: when some hedge and others overstate certainty, users receive inconsistent signals about truth, undermining trust and complicating misinformation management. Unlike traditional media, LLMs present as neutral and factual while offering personalized responses, making confidence calibration critical[47,48]. Alignment should prioritize truthfulness over compliance, and strategy use needs standardization. Research shows that combining logical appeals with relational strategies like empathy or verification improves correction effectiveness[49,50], yet current models apply these inconsistently. Governance should set clear rules for strategy selection based on context and user intent, alongside transparency in stance scoring, auditable logic, and cross-model benchmarks. Users must also be informed about these differences and the impact of prompting, reducing epistemic fragility in high-stakes advice-seeking contexts.

This study has several limitations. First, we assessed stance and strategies in isolated responses rather than multi-turn conversations, so we did not capture how correction dynamics might evolve over time. Second, our analysis focused on four major LLMs; findings may not generalize to smaller or domain-specific models, and models evolve rapidly, meaning results may not hold for future versions or alignment regimes. Third, strategies were not mutually exclusive, so observed associations may reflect underlying response types rather than independent strategy choices. Finally, strategy effectiveness was estimated using empirical data and meta-analyses. However, because no single study has systematically compared all strategies, this measure may vary depending on context. These constraints should be considered when interpreting our findings and point to opportunities for future work on conversational dynamics, broader model coverage, and causal modeling of strategy effectiveness.

Our results leave several important gaps to address. Real-world robustness remains untested, particularly in dynamic environments like social media where misinformation spreads quickly and user behavior is unpredictable. Future work should also examine multi-turn conversations to understand how correction strategies evolve over time and whether persistence improves outcomes. Another challenge is reducing epistemic homogenization, the tendency of models to suppress minority or alternative viewpoints during alignment[51]. Future work could examine how differences in user bases and training data shape model strategies, potentially optimizing certain corrective approaches for specific audience profiles. Addressing these issues will require testing LLMs in diverse, high-noise contexts, developing adaptive strategies for extended dialogues, and designing alignment frameworks that preserve epistemic diversity while maintaining factual integrity.

Methods

Study Design

To evaluate the tendencies of LLMs to correct misinformation under varying conditions, we varied four prompt characteristics and generated prompts with different combinations of each characteristic. The four characteristics included, open-mindedness (open versus closed), user intent (information-seeking versus opinion-sharing versus task-oriented output versus creative writing), user role (naive inquirer versus assertive expert), and prompt complexity (simple versus complex phrasing). This resulted in a 2 × 4 × 2 × 2 factorial design, totaling 32 unique prompt conditions (see supplementary materials for examples). Prompt generation, response generation, stance coding, and strategy tagging was conducted on October 18th, 2025.

Misinformation Domains

Ten misinformation domains were selected based on their relevance to public health, civic engagement, and epistemic risk, guided by prior work in misinformation benchmarking[52]. Topics were chosen to represent a mix of high-stakes, well-documented, and diverse epistemic contexts, while minimizing overlap. The selected misinformation domains included: evolution, vaccines and

autism, flat earth, climate change, moon landing, election fraud, alternative medicine, 5G technology, GMO foods, and COVID-19 origin.

Prompt analysis

ChatGPT-5 was used to generate 320 (2 x 4 x 2 x 2 x 10) prompts (Figure 1A), which were then sent to four different models (ChatGPT-5, Gemini 2.5 Pro, Claude Sonnet 4.5, and Grok-4), twice each, generating a total of 2,560 responses across the models (Figure 1B). They were then coded using ChatGPT-5-mini to evaluate the epistemic stance and correction strategies used (Figure 1C).

Misinformation correction coding (stance score)

ChatGPT-5-mini was used to code the stance of LLM rebuttals to the 2,560 misinformation prompts in terms of level of endorsement (see supplementary materials for full coding scheme). Responses were scored from 0 to 11, with 0 representing full endorsement (≥95% implied probability claim is true), 4 representing a neutral stance (≈50% probability), 7 representing explicit refutation (10-25% probability) and 11 representing absolute refutation with high certainty (<5% probability). Additional descriptors of what each level represented were provided to each model for improved consistency in coding.

We employed cumulative logit (proportional odds) ordered logistic regression using Python 3.10 to model the relationship between experimental factors and stance score. The dependent variable was stance score and independent variables included prompt complexity, user role, user intent, open-mindedness, topic, and model. Ordered logistic regression was selected over binary or linear models to preserve the ordinal structure of the stance score and avoid loss of information from dichotomization.

No correction for multiple comparisons was applied because all predictors were specified a priori based on a fully balanced factorial design. In such designs, testing main effects without correction is standard practice, and the risk of Type I error inflation is minimal due to the orthogonality of predictors. For conservative interpretation, a Bonferroni-adjusted threshold ($p < .0028$) was considered, under which 10 of the 13 significant effects remained statistically significant. Throughout, we focus on effect sizes and confidence intervals rather than dichotomous significance, interpreting coefficients as directional shifts in stance rather than precise point estimates.

Strategy use

In addition to epistemic stance, the types of strategies used to correct misinformation were also coded. The models were told to “Indicate every rhetorical, cognitive, or affective tactic the model employs in its response. Tags are non-exclusive - a single reply may use multiple strategies.” There were 19 strategies included, based on the misinformation literature[11,44,49,50,53-58] (see supplementary materials for descriptions): citing of evidence, appeals to authority, consensus appeals, empathetic tone, alternative explanations, socratic questioning, policy refusal, analytical reasoning, inoculation, accuracy nudges, calls to verify, redirection, social norm appeals,