Epistemic Fragility in Large Language Models: Prompt Framing Systematically Modulates Misinformation Correction

Sekoul Krastev[1], Hilary Sweatman[2], Anni Sternisko[3], Steve Rathje[3]

[1]*The Decision Lab, Montreal, Canada*
[2]*Montreal Neurological Institute, McGill University, Montreal, Canada*
[3]*New York University, New York, United States*

Corresponding author: Sekoul Krastev
sekoul@thedecisionlab.com

Abstract

As large language models (LLMs) rapidly displace traditional expertise, their capacity to correct misinformation has become a core concern. We investigate the idea that prompt framing systematically modulates misinformation correction - something we term 'epistemic fragility'. We manipulated prompts by open-mindedness, user intent, user role, and complexity. Across ten misinformation domains, we generated 320 prompts and elicited 2,560 responses from four frontier LLMs, which were coded for strength of misinformation correction and rectification strategy use. Analyses showed that creative intent, expert role, and closed framing led to a significant reduction in correction likelihood and effectiveness of used strategy. We also found striking model differences: Gemini 2.5 Pro had 74% lower odds of strong correction than Claude Sonnet 4.5. These findings highlight epistemic fragility as an important structural property of LLMs, challenging current guardrails and underscoring the need for alignment strategies that prioritize epistemic integrity over conversational compliance.

Keywords: large language models, misinformation, epistemic fragility, prompt engineering, AI alignment, sycophancy

Introduction

Large language models (LLMs) are rapidly becoming a primary knowledge source. In a sample of 1.1 million conversations, 24.4% used ChatGPT for information-seeking and 72% for non-work purposes[1]. Among executives, 92% plan to increase AI investments within three years[2]. Reliance on traditional sources is declining: Google's share of general searches fell from 73% to 67% between February and August 2025[3], Wikipedia page views dropped 8%[4], and only 5 of the top 50 U.S. news sites saw traffic growth, with declines partly linked to users relying on AI-generated summaries instead of clicking through to source articles[5]. These trends position LLMs as a first-line epistemic interface between lay users and complex scientific or political information.

LLMs are not just replacing information sources; they are displacing human expertise. Previous work found that 28.8% of ChatGPT interactions sought practical guidance[1], while 48.7% of Americans with mental health diagnoses reported using LLMs for psychological support[6]. Their appeal lies in accessibility, low cost, and anonymity, which reduces judgment – a known barrier to care[7-9]. Beyond mental health, users increasingly consult LLMs on vaccines, nutrition, and climate change, positioning these systems to assume a role traditionally held by experts: correcting misinformation.

Exposure to misinformation shapes attitudes, risk perceptions, and real-world behaviors[10,11]. Correction has thus become a central focus in misinformation research, with decades of research supporting its effectiveness, especially when it offers clear fact replacements, causal explanations, and comes from trustworthy sources[12-14].

LLMs' wide usage, personalized content, and increasingly high levels of user trust mean that, given the correct strategies, they have great potential to effectively correct misinformation at scale[15,16]. Indeed, LLMs are trained to correct misinformation in user prompts. All frontier models, when prompted with the assumption that the moon landing was faked, among other conspiracies, are likely to refute it and can even reduce the user's belief in that conspiracy[15,17].

However, the same properties that make LLMs appealing make their corrections precarious. First, LLM outputs are highly sensitive to prompt framing: small differences in wording, stance (e.g., "be sympathetic coach" vs. "be strict fact-checker"), or conversational context can systematically shift responses. Second, users often approach LLMs with advice-seeking, high-trust mindsets that amplify automation bias and source-credibility heuristics, increasing the risk that confident but flawed answers go unchallenged[18,19]. Third, models can be overcalibrated or undercalibrated about uncertainty, providing corrections that appear definitive even when evidence is mixed. Finally, LLM sycophancy (i.e., indiscriminately reinforcing users' attitudes and beliefs and avoiding direct disagreement) can cause models to prioritize user satisfaction over epistemic integrity[20].

We use the term 'epistemic fragility' to describe these vulnerabilities; instances where an LLM's correction of misinformation is influenced by features of the interaction rather than solely by the truth value of the claim. Indeed, if LLMs are to function as responsible "new experts," we must understand when prompt framing helps or hinders correction, quantify the size of these effects,

and design guardrails that make correction robust to conversational variance. The speed and depth of LLM implementation across domains creates an urgency around understanding this epistemic fragility and integrating it into broad implementation, model alignment, and AI literacy efforts.

Because LLMs optimize for word patterns that satisfy users, their responses are highly personalized but lack an internal representation of truth states, unlike expert reasoning[21]. Research shows that prompt framing strongly shapes outputs, often reflecting human biases rather than expert-level reasoning[22-24]. Even arbitrary features such as information order or categorization can introduce statistical biases[25], while format, structure, and specificity affect hallucination risk[26]. Subtle cues like political leaning or moral framing further shift responses toward user expectations rather than objective truth[27].

This sensitivity to user input has an effect on misinformation generation as well. Studies show that the way a prompt is worded, whether emotionally charged, polite, vague, or directive, can significantly influence the likelihood that a model will produce inaccurate or misleading information. For example emotional prompts, particularly polite ones, may increase disinformation generation[28], while factual accuracy varies unpredictably with changes in prompt phrasing[29]. Previous work has shown that asking models to identify credible sources before generating a correction did not reliably help models ground political misinformation responses in real news sources[30]. Others found that citation-based rebuttals led to the highest rates of regressive sycophancy[31], where models switched from correct to incorrect responses, likely because the presence of perceived authority in the prompt increased the model's tendency to defer to the user. These findings suggest that prompt sensitivity is not just a performance issue, but has direct implications for the epistemic integrity of model outputs.

While some past research suggests that prompt framing plays a critical role in LLMs generation of misinformation, significant gaps exist in understanding misinformation correction. For example, the specific conditions under which LLMs resist or succumb to misinformation, such as variations in user tone, epistemic stance, or implied intent, have not been systematically mapped. Moreover, there is limited understanding of how different prompt framings influence the types of correction strategies LLMs deploy (e.g., calls to authority, citation of evidence, or deflection). These gaps make it difficult to assess the robustness of alignment mechanisms in real-world misinformation scenarios.

To address these gaps, we ask: *How does prompt framing affect the strength of misinformation correction and the strategies LLMs use?* Methodologically, we introduce a reusable evaluation framework combining factorial prompt design, cross-model comparison, and literature-based scoring of correction strategies. Empirically, we quantify prompt-induced shifts in correction strength across ten misinformation domains and four frontier LLMs. Conceptually, we position epistemic fragility as a structural property of LLM–user interactions, linking alignment challenges to cognitive models of misinformation correction.