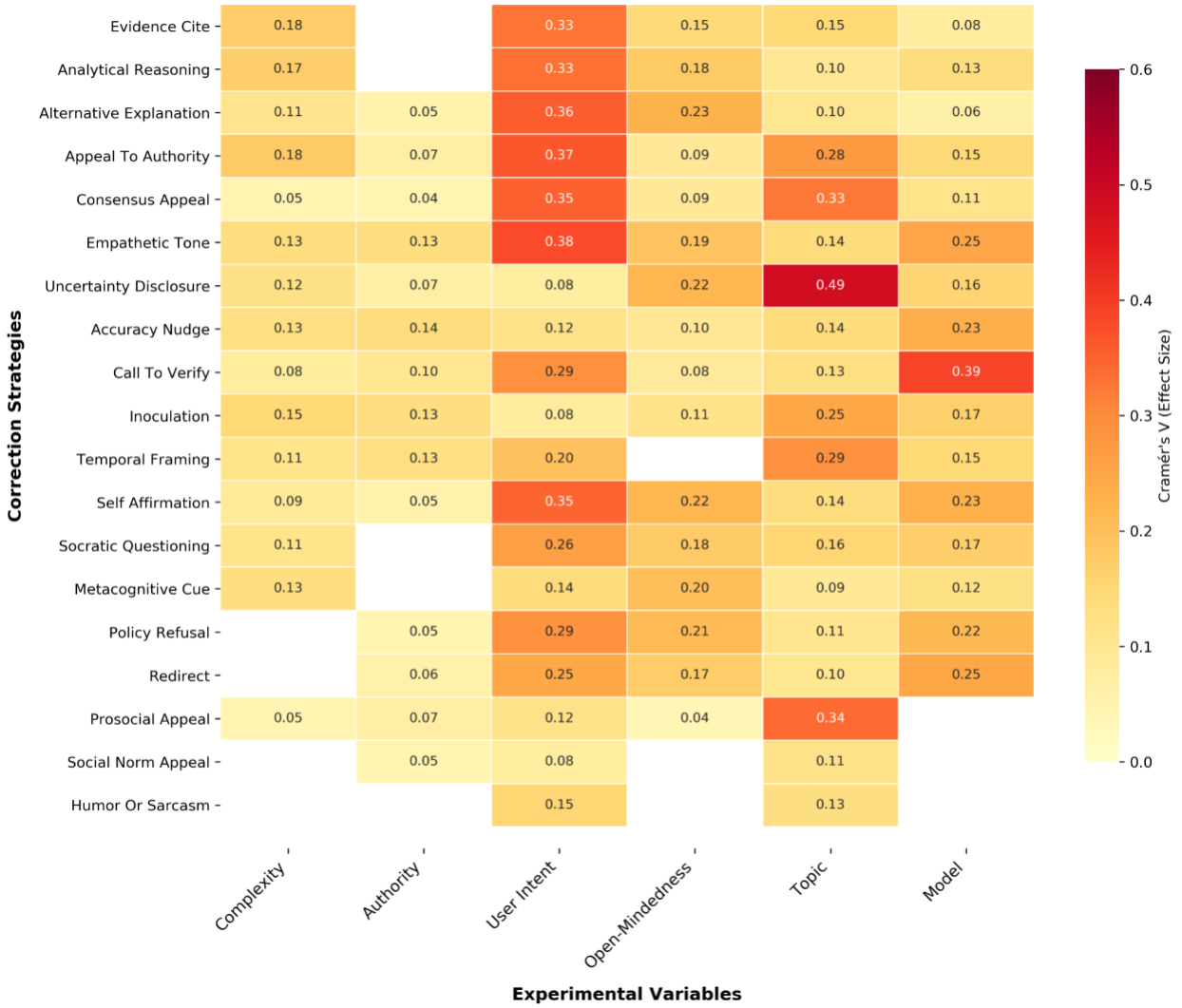

Figure 5. Significant relationships between prompt characteristics and strategy use in correction of misinformation. Only significant relationships (pFDR < .05) are shown.

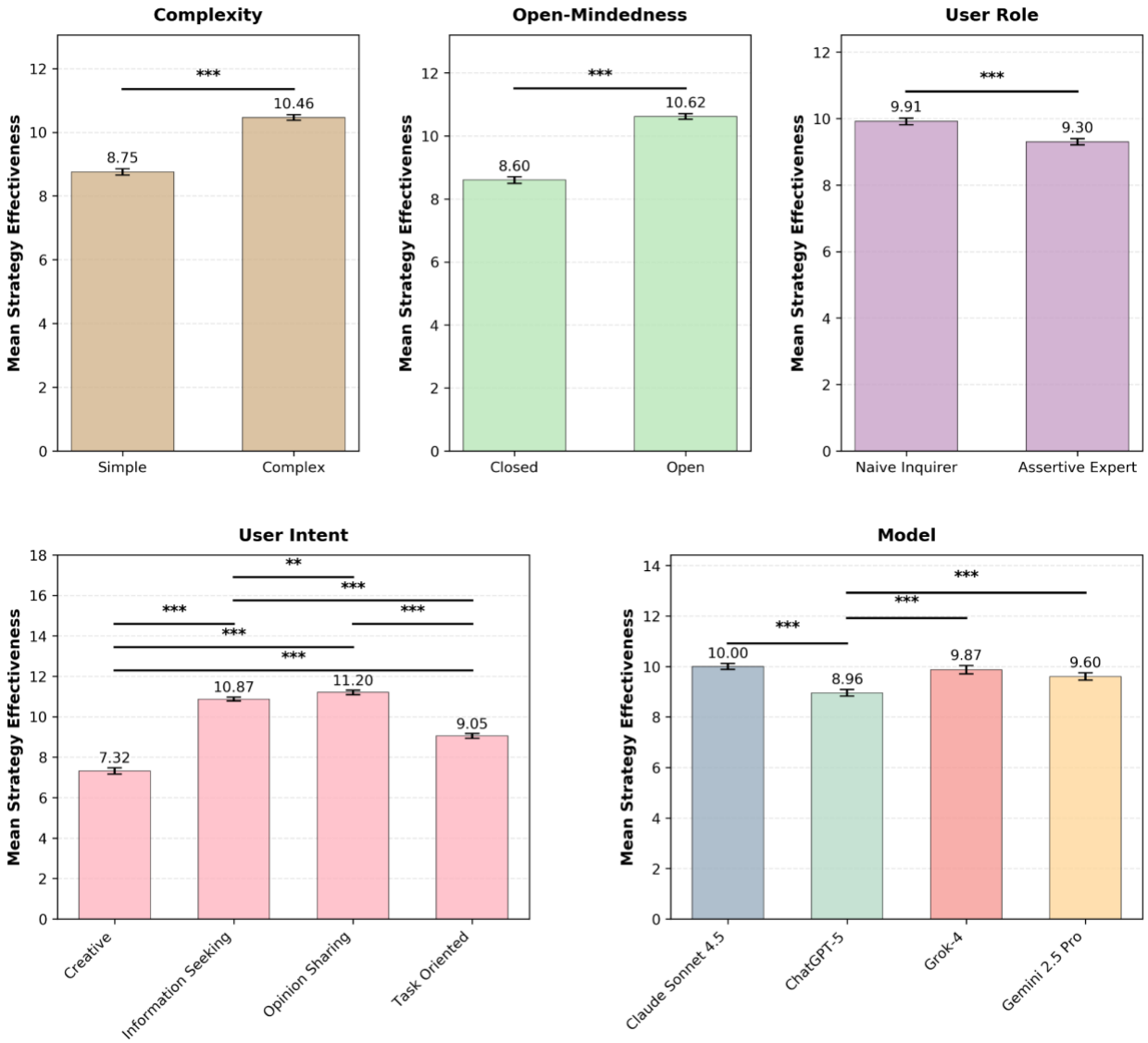

Figure 6. Mean effectiveness of strategies used for different prompt characteristics. Strategy effectiveness scores were calculated based on a literature-derived effectiveness scheme (see Supplementary Table 4): responses received 2 points for moderate–high effective size strategies, 1 point for small–moderate effective size strategies, and 0 points for unsupported or negligible effect size strategies. Error bars show standard error of the mean. **pFDR < .01, ***pFDR < .001.

Strong agreement between GPT-5-Mini and human coders

Human verification showed high consistency in stance coding across 128 pseudo-randomly selected prompt and response pairs. Inter-rater reliability was substantial between the two human coders (Cohen's κ = 0.761), and between each human coder and ChatGPT (human coder 1: κ = 0.847; human coder 2: κ = 0.795). These results suggest that the coding scheme was applied reliably and that ChatGPT's stance outputs were consistent with human judgement.

Discussion

In this paper, we identified key factors that raise the likelihood of epistemic fragility. While there was a strong general tendency across all LLMs to correct misinformation, this robustness was fragile to interaction context: corrections weakened significantly when prompts signaled the user's expertise assertively, requested creativity, or framed information in an epistemically closed manner. We also found significant effects of misinformation topics on correction, possibly reflecting the level of consensus in the real world surrounding these topics. Interestingly, we found significant model differences, with Gemini 2.5 Pro having 74% lower odds of providing stronger correction than Claude Sonnet 4.5. Finally, we found that the overall correction strategies used by models were broadly aligned with best practices from the misinformation literature, with citing evidence, analytical reasoning, alternative explanations, appeal to authority, and consensus appeal being the most common. However, we once again found that prompt framing had an effect on the effectiveness of strategies used.

Our results suggest that LLMs become less effective in correcting misinformation, both through weaker corrections and less effective corrective strategies, when prompts increase in assertive expertise, creativity, and epistemic closedness. Interestingly, this epistemic fragility mirrors findings in human misinformation research. Individuals projecting assertive expertise often display overconfidence and are perceived as authoritative, reducing both their openness to correction and others' willingness to correct them[32-34]. Likewise, low epistemic openness has been identified as a strong predictor of misinformation persistence and ideological polarization[35]. These parallels suggest that the same traits that hinder truth acceptance in humans may also influence how LLMs respond to misinformation based on certain prompt styles.

Building on these findings, creative prompts consistently produced the least effective corrections, indicating that stylistic goals can override epistemic priorities. In contrast, open framing and complex phrasing were associated with substantially more effective responses, suggesting that prompts signaling flexibility and depth encourage more rigorous engagement with misinformation. These results highlight the sensitivity of LLM corrective performance to prompt design, where shifts in tone or intent can markedly alter the effectiveness of epistemic output.

Epistemic fragility likely reflects current training and alignment methods that mimic human behavior and prioritize user satisfaction over truthfulness. Our findings align with evidence that LLMs struggle to distinguish belief from fact, often failing to acknowledge first-person false beliefs while succeeding in third-person contexts[36]. These limitations underscore the need for alignment strategies that reward epistemic integrity such as uncertainty markers, policy refusal, and verifiable sourcing, rather than merely reinforcing "helpful" or persuasive outputs. Recent work supports this direction: An *Epistemic Alignment Framework* has been proposed operationalizing uncertainty disclosure and evidence quality as alignment objectives[37], while others argue that existing fine-tuning methods fail to address epistemic accountability and call for new evaluation paradigms[38]. Similarly, previous work demonstrates that epistemic markers like uncertainty expressions significantly improve robustness, underscoring the need for alignment strategies that incentivize honesty over stylistic compliance[39]. Without such changes, models are likely to remain