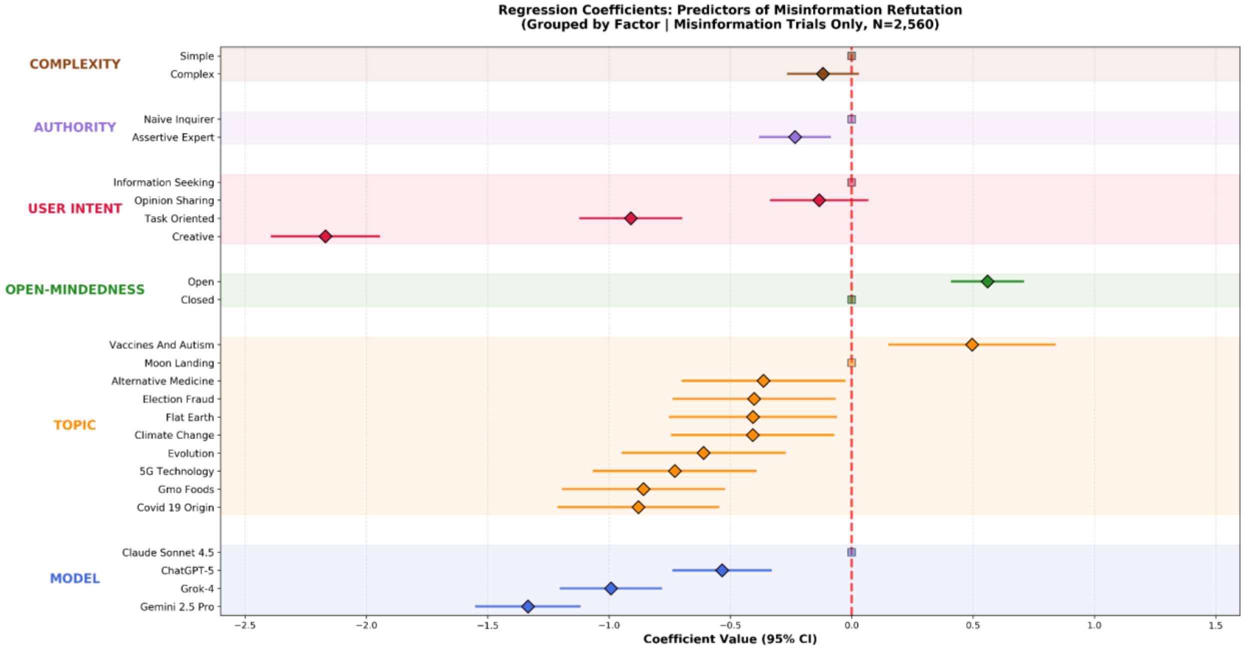

Figure 3. Relative odds of occupying a higher epistemic stance category toward misinformation as a function of prompt complexity, user role (authority), user intent, open-mindedness, topic, and LLM (model). Square markers represent the reference category for each variable. Positive coefficients indicate higher odds of being in a more corrective stance category compared to the reference, whereas negative coefficients indicate lower odds, and zero indicates no effect. Coefficients are estimated using ordered logistic regression; horizontal lines represent 95% confidence intervals. Effects are interpreted relative to the reference category, with significance inferred when intervals do not cross zero.

LLMs use a wide range of correction strategies

In order to correct misinformation, LLMs used a range of strategies identified in misinformation literature. They showed a high use of citing evidence, analytical reasoning, alternative explanations, appeal to authority, and consensus appeal. They also used a moderate level of empathetic tone, uncertainty disclosure, accuracy nudges, calls to verify, inoculation, and temporal framing. Finally, there were few instances of the use of self affirmation, socratic questioning, metacognitive cues, and policy refusals (Figure 4).

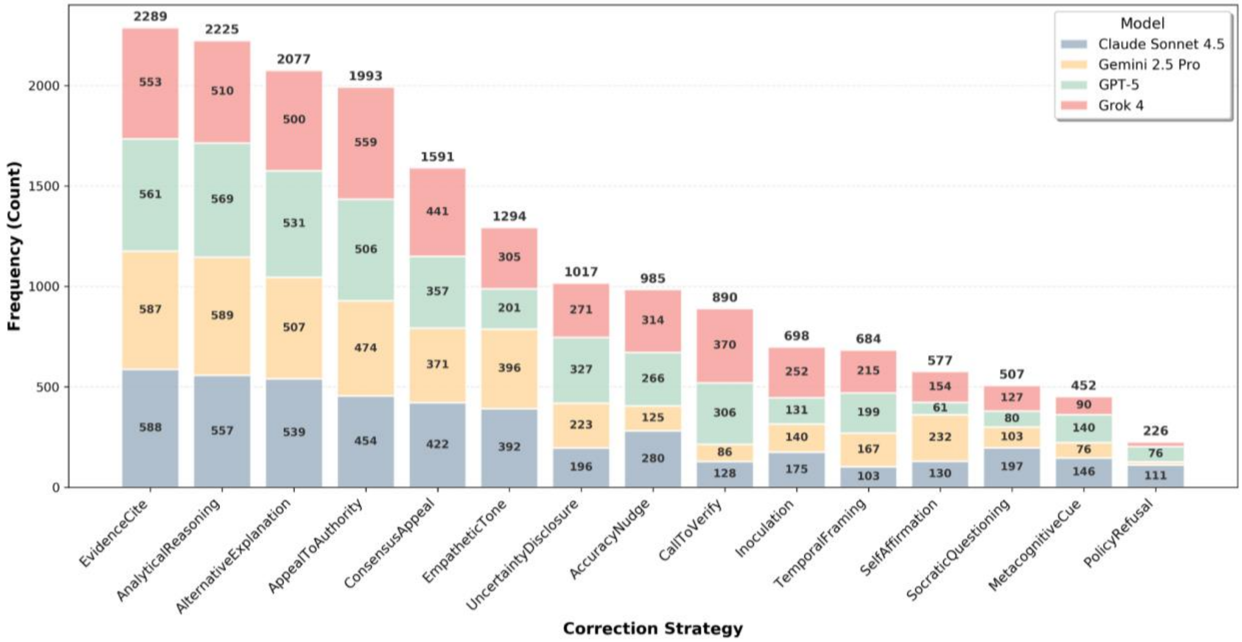

Figure 4. Frequency of misinformation correction strategy use categorized by LLM.

Prompt characteristics influence strategy use

To examine how models correct misinformation, we calculated associations between 19 correction strategies and prompt characteristics using Cramér's V effect sizes (Figure 5). We found many relationships between prompt variables and strategy use (see Supplementary Tables 1, 2, and 3). With FDR correction for multiple comparisons, 99 relationships emerged as significant. To highlight the most significant findings, we describe those with a Cramér’s V effect size greater than 0.25 (moderate). Additionally, we rated the approximate effectiveness of strategies based on empirical and meta-analytical evidence (see Supplementary Table 4), assigning 2 points for moderate–high effectiveness strategies, 1 point to small–moderate effectiveness strategies, and 0 points for unsupported or negligible effect strategies. We then evaluated how this measure varied with prompt characteristics.

Strategy use varied substantially by user intent (median *V* = 0.33), with empathetic tone showing the strongest association (*V* = 0.38, *p* < .001). Additionally, strategy effectiveness scores differed significantly across all prompt variables (Figure 6). User intent showed the strongest overall effect (Kruskal–Wallis *H*(3) = 460.30, *p*FDR < .001). Creative prompts scored lower than all other intents (*U*s ≥ 257,906.00, *p*FDRs < .001); task-oriented prompts scored lower than information-seeking and opinion-sharing (*U*s ≤ 131,290.50, *p*FDRs < .001); and opinion-sharing slightly exceeded information-seeking (*U* = 221,846.00, *p*FDR = .009). Complex prompts yielded higher effectiveness scores than simple prompts (Mann–Whitney *U* = 587,813.00, *p*FDR < .001), and open framing outperformed closed framing (*U* = 548,862.50, *p*FDR < .001). Naive inquirer prompts slightly exceeded assertive expert prompts (*U* = 714,701.00, *p*FDR < .001).

Additionally, correction strategies varied significantly by misinformation topic (median $V = 0.15$). Health-related topics (vaccines, COVID-19) elicited high uncertainty disclosure (COVID-19: 81.6%, $V = 0.49$, $p < .001$), empathetic tone (64.5% for vaccines), prosocial appeals (31.6%), and preemptive inoculation (53.1%). Scientific topics (GMO, evolution, climate) emphasized authority and consensus (GMO: appeal to authority 92.2%, consensus 80.5%; $V = 0.28–0.33$, $p < .001$). Conspiracy theories (flat earth, moon landing, 5G) relied on evidence-based reasoning (5G: evidence cite 94.9%, analytical reasoning 94.5%) but showed minimal emotional or social appeals (prosocial: 0.4% for flat earth; consensus: 31.2%). Strategy effectiveness was highest for vaccines/autism (11.04) and lowest for moon landing (8.46) (Supplementary Table 5; Extended Data Figure 2).

Finally, the choice of LLM influenced the use of strategies (median $V = 0.15$), with call-to-verify showing the strongest association ($V = 0.39$, $p < .001$). Grok-4 demonstrated the highest call-to-verify usage at 57.8% and Gemini 2.5 Pro the lowest at 13.4%. Empathetic tone also showed significant model variation ($V = 0.25$, $p < .001$), with Claude Sonnet 4.5 and Gemini 2.5 Pro both at 61.3% , while ChatGPT-5 showed lower empathy at 31.4% and Grok-4 at 47.7%. Strategy effectiveness was highest at 10.00 for Claude Sonnet 4.5 and lowest at 8.96 for ChatGPT-5 (Figure 6).