prosocial appeals, self affirmation, uncertainty disclosure, temporal framing, humor or sarcasm, and metacognitive clues.

To assess how strategy use varied across experimental conditions, we computed Cramér's V effect sizes for associations between each strategy and six categorical variables: prompt complexity, user role, user intent, open-mindedness, topic, and model. Cramér's V was chosen for its suitability in measuring associations between categorical variables with differing numbers of levels. Data are shown for misinformation trials only, and only significant relationships (pFDR < .05) are shown.

Strategy Effectiveness Scores

Responses were scored for effectiveness using a literature-based scheme (see Supplementary Table 4 for descriptions). This method necessarily simplifies and should be interpreted as an approximate measure of strategy effectiveness. Strategies associated with moderate-to-high effect sizes received 2 points (Evidence Cite, Alternative Explanation, Consensus Appeal, Inoculation, Analytical Reasoning, Empathetic Tone), those with small–moderate effect sizes received 1 point (Accuracy Nudge, Appeal To Authority, Social Norm Appeal, Prosocial Appeal, Self Affirmation, Metacognitive Cue), and strategies with negligible or unsupported effects received 0 points (Uncertainty Disclosure, Call To Verify, Redirect, Socratic Questioning, Temporal Framing, Policy Refusal, Humor or Sarcasm). The total effectiveness score for each response was the sum of points for all strategies present. We treat the resulting effectiveness score as a heuristic index of how many empirically supported strategies a response deploys, not as a psychometric scale. Accordingly, we focus on relative comparisons across conditions rather than the absolute magnitude of scores.

Mean effectiveness scores were compared across experimental factors using non-parametric tests. Two-level variables (complexity, open-mindedness, user role) were analyzed with Mann–Whitney U tests; Four-level variables (user intent, model) were analyzed with a Kruskal–Wallis test followed by pairwise Mann–Whitney U tests. All p-values were adjusted using the Benjamini–Hochberg false discovery rate (pFDR), with thresholds at pFDR < 0.05, 0.01, and 0.001. Because empirical estimates for some strategies come from different contexts, we interpret absolute effectiveness scores cautiously and focus on comparative differences between conditions.

Validation by human coders

To evaluate the reliability of ChatGPT's stance score coding, we had two human coders evaluate a subset of responses (Figure 1D). Coders each evaluated 128 responses, 4 outputs per the 32 prompt types (from the original 2×4×2×2 factorial design). Before doing so, they were trained on a pilot set of 10 questions. Instructions (see Supplementary Materials) were clarified and expanded upon based on coder feedback and performance. These improved instructions were used for the true coding phase. Each response was generated using a pseudorandomly assigned misinformation scheme, ensuring that all 10 topics of misinformation were each included at least 6 times.

Coders used an abbreviated version of the 11-point ordinal scale used by ChatGPT, with only 7 points for ease of use (see Supplementary Materials). To compare ratings, we harmonized GPT's

11-point scores to the human coder’s 7-point scale using nearest-category mapping based on midpoints between allowed values (0, 2, 3, 4, 6, 7, 10). For example, GPT scores of 8 or 9 were mapped to 10, while scores of 5 were mapped to 4. This ensured both raters were evaluated on the same ordinal scale.

Agreement between coders and between coders and ChatGPT was calculated using Weighted Cohen’s Kappa with quadratic weighting, which accounts for the degree of disagreement between ordinal categories. Coders were not informed which LLM produced each response, reducing the risk that perceptions of specific systems influenced stance ratings.

Acknowledgments

We thank the human coders, Celestine Rosales and Jestine Cabiles, who evaluated a subset of model responses for stance. Their careful work was essential for validating our coding scheme and ensuring the reliability of our measures. Furthermore, we are grateful to Dan Pilat and Marielle Montenegro for their feedback on early drafts.

References

1 Chatterji, A. *et al.* How People Use ChatGPT. (National Bureau of Economic Research, 2025).

2 Mayer, H., Yee, L., Chui, M. & Roberts, R. Superagency in the workplace: Empowering people to unlock AI's full potential. (2025).

3 Heitzman, A. How People Search Today: A Study on Evolving Search Behaviors in 2025. (HigherVisibility., 2025).

4 Miller, M. New User Trends on Wikipedia. (Wikimedia, 2025).

5 Press Gazette. *Top 50 news websites in the US in September: BBC and NBC News among big winners*, https://pressgazette.co.uk/media-audience-and-business-data/media_metrics/most-popular-websites-news-us-monthly-3/ (2025).

6 Rousmaniere, T., Li, X., Zhang, Y. & Shah, S. Large Language Models as Mental Health Resources: Patterns of Use in the United States. (2025).

7 Keyes, K. M. *et al.* Stigma and Treatment for Alcohol Disorders in the United States. *American Journal of Epidemiology* **172**, 1364-1372 (2010). https://doi.org/10.1093/aje/kwq304

8 Schomerus, G. & Angermeyer, M. C. Stigma and its impact on help-seeking for mental disorders: what do we know? *Epidemiologia e Psichiatria Sociale* **17**, 31-37 (2008). https://doi.org/10.1017/s1121189x00002669

9 Thornicroft, G. Stigma and discrimination limit access to mental health care. *Epidemiologia e Psichiatria Sociale* **17**, 14-19 (2008). https://doi.org/10.1017/s1121189x00002621

10 Lewandowsky, S. *et al. The Debunking Handbook*. (2020).

11 Southwell, B. G. & Thorson, E. A. The Prevalence, Consequence, and Remedy of Misinformation in Mass Media Systems. *Journal of Communication* **65**, 589-595 (2015). https://doi.org/10.1111/jcom.12168

12 Chan, M.-P. S. & AlbarracíN, D. A meta-analysis of correction effects in science-relevant misinformation. *Nature Human Behaviour* **7**, 1514-1525 (2023). https://doi.org/10.1038/s41562-023-01623-8

13 Ecker, U. K. H. *et al.* The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology* **1**, 13-29 (2022). https://doi.org/10.1038/s44159-021-00006-y

14 Van Der Linden, S., Leiserowitz, A., Rosenthal, S. & Maibach, E. Inoculating the Public against Misinformation about Climate Change. *Global Challenges* **1**, 1600008 (2017). https://doi.org/10.1002/gch2.201600008

15 Costello, T. H., Pennycook, G. & Rand, D. G. Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385** (2024). https://doi.org/10.1126/science.adq1814

16 Zhou, X., Sharma, A., Zhang, A. X. & Althoff, T. Correcting misinformation on social media with a large language model. *arXiv* (2024). https://doi.org/10.48550/arxiv.2403.11169

17 Costello, T. H., Rabb, N., Stagnaro, M. N., Pennycook, G. & Rand, D. G. Reducing belief in conspiracy theories as they unfold. *PsyArXiv* (2025).