18 Mosier, K. L., Skitka, L. J., Burdick, M. D. & Heers, S. T. Automation Bias, Accountability, and Verification Behaviors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **40**, 204-208 (1996). https://doi.org/10.1177/154193129604000413

19 Logg, J. M., Minson, J. A. & Moore, D. A. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* **151**, 90-103 (2019). https://doi.org/10.1016/j.obhdp.2018.12.005

20 Rathje, S. *et al.* Sycophantic AI increases attitude extremity and overconfidence. *PsyArXiv* (2025). https://doi.org/https://doi.org/10.31234/osf.io/vmyek_v1

21 Orgad, H. *et al.* LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. *arXiv* (2024). https://doi.org/10.48550/arxiv.2410.02707

22 Marvin, G., Hellen, N., Jjingo, D. & Nakatumba-Nabende, J. Prompt Engineering in Large Language Models in *Algorithms for Intelligent Systems*. 387-402 (Springer Nature Singapore, 2024).

23 Mishra, S., Khashabi, D., Baral, C., Choi, Y. & Hajishirzi, H. Reframing Instructional Prompts to GPTk's Language. *arXiv* (2022). https://doi.org/10.48550/arxiv.2109.07830

24 Sharma, M. *et al.* Towards understanding sychophancy in large language models. *arXiv* (2024). https://doi.org/10.48550/arxiv.2310.13548

25 Brucks, M. & Toubia, O. Prompt architecture induces methodological artifacts in large language models. *PLOS One* **20**, e0319159 (2025). https://doi.org/10.1371/journal.pone.0319159

26 Anh-Hoang, D., Tran, V. & Nguyen, L.-M. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence* **8** (2025). https://doi.org/10.3389/frai.2025.1622292

27 Nadeem, A., Dras, M. & Naseem, U. Steering Towards Fairness: Mitigating Political Bias in LLMs. *arXiv* (2025). https://doi.org/10.48550/arxiv.2508.08846

28 Vinay, R., Spitale, G., Biller-Andorno, N. & Germani, F. Emotional prompting amplifies disinformation generation in AI large language models. *Frontiers in Artificial Intelligence* **8** (2025). https://doi.org/10.3389/frai.2025.1543603

29 Wang, Z. *et al.* A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More. *arXiv* (2024). https://doi.org/10.48550/arxiv.2407.16216

30 Proma, A. M. *et al.* How LLMs Fail to Support Fact-Checking. *arXiv* (2025).

31 Fanous, A. *et al.* SycEval: Evaluating LLM Sycophancy. *arXiv* (2025). https://doi.org/10.48550/arxiv.2502.08177

32 Lyons, B., King, A. J. & Kaphingst, K. A. Overconfidence in ability to discern cancer misinformation: a conceptual replication and extension. *Human Communication Research* (2025). https://doi.org/10.1093/hcr/hqaf017

33 Mang, V., Fennis, B. M. & Epstude, K. Source credibility effects in misinformation research: A review and primer. *advances.in/psychology* **2**, e443610 (2024). https://doi.org/10.56296/aip00028

34 Rapp, D. N. & Withall, M. M. Confidence as a metacognitive contributor to and consequence of misinformation experiences. *Current Opinion in Psychology* **55**, 101735 (2024). https://doi.org/10.1016/j.copsyc.2023.101735

35 Broda, E. & Strömbäck, J. Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association* **48**, 139-166 (2024). https://doi.org/10.1080/23808985.2024.2323736

36 Suzgun, M. *et al.* Language models cannot reliably distinguish belief from knowledge and fact. *Nature Machine Intelligence* (2025). https://doi.org/10.1038/s42256-025-01113-8

37 Clark, N., Shen, H., Howe, B. & Mitrav, T. Epistemic Alignment: A Mediating Framework for User-LLM Knowledge Delivery. *arXiv* (2025). https://doi.org/10.48550/arxiv.2504.01205

38 Sarkar, U. E. Evaluating alignment in large language models: a review of methodologies. *AI and Ethics* **5**, 3233-3240 (2025). https://doi.org/10.1007/s43681-024-00637-w

39 Lee, D., Hwang, Y., Kim, Y., Park, J. & Jung, K. Are LLM-Judges Robust to Expressions of Uncertainty? Investigating the effect of Epistemic Markers on LLM-based Evaluation. *arXiv* (2024). https://doi.org/10.48550/arxiv.2410.20774

40 Lior, G., Nacchace, L. & Stanovsky, G. WildFrame: Comparing Framing in Humans and LLMs on Naturally Occurring Texts. *arXiv* (2025). https://doi.org/10.48550/arxiv.2502.17091

41 Traberg, C. S., Harjani, T., Roozenbeek, J. & Van Der Linden, S. The persuasive effects of social cues and source effects on misinformation susceptibility. *Scientific Reports* **14** (2024). https://doi.org/10.1038/s41598-024-54030-y

42 Chi, M. T. H., Glaser, R. & Farr, M. J. *The nature of expertise*. (Lawrence Erlbaum Associates, Inc., 1988).

43 Kahneman, D. & Klein, G. Conditions for intuitive expertise: A failure to disagree. *American Psychologist* **64**, 515-526 (2009). https://doi.org/https://doi.org/10.1037/a0016755

44 Van Der Linden, S. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine* **28**, 460-467 (2022). https://doi.org/10.1038/s41591-022-01713-6

45 Kashima, Y., Perfors, A., Ferdinand, V. & Pattenden, E. Ideology, communication and polarization. *Philosophical Transactions of the Royal Society B: Biological Sciences* **376**, 20200133 (2021). https://doi.org/10.1098/rstb.2020.0133

46 Stapleton, C. E. & Wolak, J. Political Self-Confidence and Affective Polarization. *Public Opinion Quarterly* **88**, 79-96 (2024). https://doi.org/10.1093/poq/nfad064

47 Geng, J. *et al.* (Association for Computational Linguistics).

48 Schneider, S. Chatbot Epistemology. *Social Epistemology* **39**, 570-589 (2025). https://doi.org/10.1080/02691728.2025.2500030

49 Peter, C. & Koch, T. Countering misinformation: Strategies, challenges, and uncertainties. *Studies in Communication and Media* **8**, 431-445 (2019). https://doi.org/10.5771/2192-4007-2019-4-431

50 Amazeen, M. A. & Krishna, A. Refuting misinformation: Examining theoretical underpinnings of refutational interventions. *Current Opinion in Psychology* **56**, 101774 (2024). https://doi.org/10.1016/j.copsyc.2023.101774

51 Zheng, E. L. & Lee, S. S.-J. The Epistemological Danger of Large Language Models. *The American Journal of Bioethics* **23**, 102-104 (2023). https://doi.org/10.1080/15265161.2023.2250294

52 Lin, S., Hilton, J. & Evans, O. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv* (2022). https://doi.org/10.48550/arxiv.2109.07958

53 Roozenbeek, J., Van Der Linden, S., Goldberg, B., Rathje, S. & Lewandowsky, S. Psychological inoculation improves resilience against misinformation on social media. *Science Advances* **8** (2022). https://doi.org/10.1126/sciadv.abo6254

54 Kim, Y. & Lim, H. Debunking misinformation in times of crisis: Exploring misinformation correction strategies for effective internal crisis communication. *Journal of Contingencies and Crisis Management* **31**, 406-420 (2023). https://doi.org/10.1111/1468-5973.12447

55 Guan, T., Liu, T. & Yuan, R. Facing disinformation: Five methods to counter conspiracy theories amid the Covid-19 pandemic. *Comunicar* **29**, 71-83 (2021). https://doi.org/10.3916/c69-2021-06

56 Roozenbeek, J., Culloty, E. & Suiter, J. Countering Misinformation. *European Psychologist* **28**, 189-205 (2023). https://doi.org/10.1027/1016-9040/a000492

57 Wittenberg, C. & Berinsky, A. J. in *Social media and democracy: The state of the field, prospects for reform* 163-198 (2020).

58 Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N. & Cook, J. Misinformation and Its Correction. *Psychological Science in the Public Interest* **13**, 106-131 (2012). https://doi.org/10.1177/1529100612451018