## Updated References

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
2. Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
3. Chen, A., Phang, J., Parrish, A., Padmakumar, V., Zhao, C., Bowman, S. R., & Cho, K. (2023). Two Failures of Self-Consistency in the Multi-Step Reasoning of Large Language Models. *arXiv preprint arXiv:2305.14279*.
4. Curvo, P. M. P. (2025). The Traitors: Deception and Trust in Multi-Agent Language Model Simulations. *arXiv preprint arXiv:2505.12923*.
5. Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-Verification Reduces Hallucination in Large Language Models. *Findings of the Association for Computational Linguistics: ACL 2024*, 3563–3578.
6. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
7. Jia, J., Yuan, Z., Pan, J., McNamara, P., & Chen, D. (2024). Decision-making behavior evaluation framework for LLMs under uncertain context. *arXiv preprint arXiv:2406.05972*.
8. Liu, Y., Guo, Z., Liang, T., Shareghi, E., Vulić, I., & Collier, N. (2024). Aligning with Logic: Measuring, Evaluating and Improving Logical Preference Consistency in Large Language Models. *arXiv preprint arXiv:2410.02205*.
9. Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 3214–3229.
10. OpenAI. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
11. Park, J. S., O’Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17.
12. Runje, D., & Shankaranarayana, S. M. (2023). Constrained Monotonic Neural Networks. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 29338–29353). PMLR.
13. Sartor, D., Sinigaglia, A., & Susto, G. A. (2025). Advancing Constrained Monotonic Neural Networks: Achieving Universal Approximation Beyond Bounded Activations. *arXiv preprint arXiv:2505.02537*.
14. Sivaraman, A., Khandelwal, K., & Ravikumar, P. (2020). Counterexample-Guided Learning of Monotonic Neural Networks. In *Proceedings of the 8th Workshop on Explainable AI at International Conference on Learning Representations (ICLR)*.
15. Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Advances in Neural Information Processing Systems*, 36, 74952–74962.
16. Yao, Z., Liu, Y., Chen, Y., Chen, J., Fang, J., Hou, L., Li, J., & Chua, T.-S. (2025). Are Reasoning Models More Prone to Hallucination? *arXiv preprint arXiv:2505.23646*.