Reference

1. Tonmoy, S. M., Zaman, S. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.
2. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
3. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024). Hallucination of Multimodal Large Language Models: A Survey. *arXiv preprint arXiv:2404.18930*.
4. 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.
5. Chen, K., Chen, Q., Zhou, J., He, Y., & He, L. (2024). DiaHalu: A Dialogue-level Hallucination Evaluation Benchmark for Large Language Models. *arXiv preprint arXiv:2403.00896*.
6. Li, J., Cheng, X., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2023, December). Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6449-6464).
7. Guan, J., Dodge, J., Wadden, D., Huang, M., & Peng, H. (2023). Language Models Hallucinate, but May Excel at Fact Verification. *arXiv preprint arXiv:2310.14564*.
8. Manakul, P., Liusie, A., & Gales, M. J. (2023). Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.
9. Lin, S., Hilton, J., & Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.
10. Muhlgay, D., Ram, O., Magar, I., Levine, Y., Ratner, N., Belinkov, Y., Abend, O., Leyton-Brown, K., Shashua, A., & Shoham, Y. (2023). Generating Benchmarks for Factuality Evaluation of Language Models. *Conference of the European Chapter of the Association for Computational Linguistics*.
11. Liu, T., Zhang, Y., Brockett, C., Mao, Y., Sui, Z., Chen, W., & Dolan, B. (2021). A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704*.
12. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
13. Zhao, Y., Yan, L., Sun, W., Xing, G., Meng, C., Wang, S., ... & Yin, D. (2023). Knowing What LLMs DO NOT Know: A Simple Yet Effective Self-Detection Method. *arXiv preprint arXiv:2310.17918*.
14. Du, L., Wang, Y., Xing, X., Ya, Y., Li, X., Jiang, X., & Fang, X. (2023). Quantifying and attributing the hallucination of large language models via association analysis. *arXiv preprint arXiv:2309.05217*.
15. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J. R. (2023). Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
16. Rawte, V., Sheth, A., & Das, A. (2023). A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.
17. Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., ... & Levy, O. (2024). Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, *36*.
18. Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., ... & Wang, G. (2023). Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.
19. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744.
20. Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, *35*, 24611-24624.
21. Gunjal, A., Yin, J., & Bas, E. (2024, March). Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 16, pp. 18135-18143).
22. Wu, J., Hobbs, J., & Hovakimyan, N. (2023). Hallucination improves the performance of unsupervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16132-16143).
23. Himakunthala, V., Ouyang, A., Rose, D., He, R., Mei, A., Lu, Y., ... & Wang, W. (2023, December). Let’s think frame by frame with VIP: A video infilling and prediction dataset for evaluating video chain-of-thought. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 204-219).
24. Kulal, S., Brooks, T., Aiken, A., Wu, J., Yang, J., Lu, J., ... & Singh, K. K. (2023). Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 17089-17099).
25. Doh, S., Choi, K., Lee, J., & Nam, J. (2023). Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*.
26. Chen, X., Song, D., Gui, H., Wang, C., Zhang, N., Yong, J., Huang, F., Lv, C., Zhang, D., & Chen, H. (2023). FactCHD: Benchmarking Fact-Conflicting Hallucination Detection. *ArXiv, abs/2310.12086*.
27. Chaudhury, S., Swaminathan, S., Gunasekara, C., Crouse, M., Ravishankar, S., Kimura, D., ... & Gray, A. (2022, December). X-factor: A cross-metric evaluation of factual correctness in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 7100-7110).
28. Keller, N., & Rosemarin, A. (2021, June). Mind the middle layer: the HADES design strategy revisited. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques* (pp. 35-63). Cham: Springer International Publishing.
29. Lin, S., Hilton, J., & Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.
30. Yu, T., Zhang, S., & Feng, Y. (2024). Truth-Aware Context Selection: Mitigating the Hallucinations of Large Language Models Being Misled by Untruthful Contexts. *arXiv preprint arXiv:2403.07556*.
31. Chen, K., Chen, Q., Zhou, J., He, Y., & He, L. (2024). DiaHalu: A Dialogue-level Hallucination Evaluation Benchmark for Large Language Models. *arXiv preprint arXiv:2403.00896*.
32. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., & Bing, L. (2023). Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*.
33. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., ... & Chen, E. (2023). Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.
34. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Shi, S. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.
35. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., ... & Yu, N. (2023). Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*.
36. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., & Bing, L. (2023). Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*.
37. Zhu, Z., Xue, Y., Chen, X., Zhou, D., Tang, J., Schuurmans, D., & Dai, H. (2023). Large Language Models can Learn Rules. *ArXiv, abs/2310.07064*.
38. Zhang, Y., Cui, L., Bi, W., & Shi, S. (2023). Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*.
39. Tonmoy, S. M., Zaman, S. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.
40. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2024). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, *36*.
41. Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022). Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.
42. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199-22213.
43. Zhao, W. X., Liu, J., Ren, R., & Wen, J. R. (2024). Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, *42*(4), 1-60.
44. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.
45. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
46. Work, W. M. I. C. L. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?.
47. Shi, W., Han, X., Lewis, M., Tsvetkov, Y., Zettlemoyer, L., & Yih, S. W. T. (2023). Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.
48. Chuang, Y. S., Xie, Y., Luo, H., Kim, Y., Glass, J., & He, P. (2023). Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.
49. Li, K., Patel, O., Viégas, F., Pfister, H., & Wattenberg, M. (2024). Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, *36*.
50. Yoon, S., Yoon, E., Yoon, H. S., Kim, J., & Yoo, C. D. (2022). Information-theoretic text hallucination reduction for video-grounded dialogue. arXiv preprint arXiv:2212.05765.
51. Elaraby, M., Lu, M., Dunn, J., Zhang, X., Wang, Y., & Liu, S. (2023). Halo: Estimation and reduction of hallucinations in open-source weak large language models. arXiv preprint arXiv:2308.11764.