Check for updates

Preventing and Detecting Misinformation Generated by Large Language Models

Aiwei Liu
Tsinghua University
Beijing, China
liuaw20@mails.tsinghua.edu.cn

Qiang Sheng
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
shengqiang18z@ict.ac.cn

Xuming Hu*
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
xuminghu@hkust-gz.edu.cn

ABSTRACT

As large language models (LLMs) become increasingly capable and widely deployed, the risk of them generating misinformation poses a critical challenge. Misinformation from LLMs can take various forms, from factual errors due to hallucination to intentionally deceptive content, and can have severe consequences in high-stakes domains.This tutorial covers comprehensive strategies to prevent and detect misinformation generated by LLMs. We first introduce the types of misinformation LLMs can produce and their root causes. We then explore two broad categories: **Preventing misinformation generation**: a) AI alignment training techniques to reduce LLMs' propensity for misinformation and refuse malicious instructions during model training. b) Training-free mitigation methods like prompt guardrails, retrieval-augmented generation (RAG), and decoding strategies to curb misinformation at inference time. **Detecting misinformation after generation**, including a) using LLMs themselves to detect misinformation through embedded knowledge or retrieval-enhanced judgments, and b) distinguishing LLM-generated text from human-written text through black-box approaches (e.g., classifiers, probability analysis) and white-box approaches (e.g., watermarking). We also discuss the challenges and limitations of detecting LLM-generated misinformation.

CCS CONCEPTS

• **Computing methodologies** → *Natural language processing*; • **Security and privacy** → **Social aspects of security and privacy**.

KEYWORDS

Large Language Models, Misinformation, Hallucination

ACM Reference Format:
Aiwei Liu, Qiang Sheng, and Xuming Hu. 2024. Preventing and Detecting Misinformation Generated by Large Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3626772.3661377

*Corresponding author

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3661377

1 INTENDED AUDIENCE

This tutorial is intended for researchers, practitioners, and policymakers interested in understanding and addressing the challenges of misinformation generated by large language models (LLMs). Attendees should have a basic understanding of natural language processing and large language models. The tutorial will be accessible to a broad audience, including researchers and practitioners from academia and industry, as well as policymakers and journalists. Attendees will gain a comprehensive understanding of the types of misinformation LLMs can produce, the root causes of misinformation, and the state-of-the-art strategies to prevent and detect misinformation generated by LLMs.

2 PRESENTERS

Aiwei Liu[1] is a Ph.D. student at the School of Software, Tsinghua University, supervised by Prof. Lijie Wen. He received his Bachelor's degree from the Software Institute, Nanjing University in 2020. His research interests focus on large language models, particularly in the areas of security and trustworthiness, including red-teaming, secure alignment, and watermarking techniques for large models. He has published numerous papers in top-tier conferences and journals, such as ICLR, SIGKDD, ACL, EMNLP, SIGIR, and TKDE. Additionally, he serves as a reviewer for several prestigious conferences, including ACL, EMNLP, EACL, and WWW.

Qiang Sheng[2] is an Assistant Professor at the Media Synthesis and Forensics Lab, Institute of Computing Technology, Chinese Academy of Sciences. He received his Ph.D. from the University of Chinese Academy of Sciences under the supervision of Prof. Juan Cao. His research interests include fake news detection, fact-checking, natural language understanding, social media mining, and AI safety. He has published in top-tier conferences and journals such as ACL, AAAI, SIGIR, WWW, and TKDE. He has served as an area chair, reviewer, or PC member for conferences including MM, AAAI, ACL, EMNLP, IJCAI, KDD, and WWW, and as a reviewer for journals such as ACM TOIS, IEEE TASLP, and IP&M.

Xuming Hu[3] is an Assistant Professor at the Hong Kong University of Science and Technology (Guangzhou). He obtained his Ph.D. from the School of Software at Tsinghua University in 2024. His research interests encompass trustworthy large language models, multimodal large language models, and AI for science. His work has been published in numerous top-tier international conferences, such as SIGIR, SIGKDD, ACL, EMNLP, ICLR, NAACL, and TKDE.

[1] https://scholar.google.com/citations?user=UCOOmcEAAAAJ
[2] https://scholar.google.com/citations?user=8BEUQ9UAAAAJ
[3] https://scholar.google.com/citations?user=dbBKbXoAAAAJ

Furthermore, he serves as an Area Chair for prestigious conferences, including ACL, NAACL, and EACL.

3 TOPIC AND RELEVANCE

3.1 Motivation

The rapid development of large language model technologies has led to the emergence of increasingly powerful models, including open-source ones like LLaMA [51], Falcon [2], and Mixtral [29], as well as commercial products such as GPT-4 [1], Claude [3], and Gemini [45]. These models have not only made tremendous progress in traditional NLP tasks like question answering [48], translation [42], and information extraction [55] but have also demonstrated new capabilities in areas such as code generation [43]. Some researchers even refer to GPT-4 as an early version of Artificial General Intelligence (AGI) [8]. As more and more text on the internet is generated by LLMs, concerns about the credibility and authenticity of LLM-generated content have grown [16, 53, 62]. On one hand, LLMs themselves suffer from hallucination [27, 44], which can lead to the unintentional generation of false information. Some studies suggest that hallucination in large models may even be inevitable [59]. On the other hand, due to the strong instruction-following capabilities of LLMs [17], malicious users can manipulate them to generate false information through carefully crafted prompts [41]. Research has shown that fake information generated by LLMs is more difficult for both humans and detectors to identify [10], suggesting that LLM-generated misinformation may have a more deceptive style and potentially cause greater harm.

Therefore, mitigating the harm caused by LLM-generated misinformation is a crucial issue. In this tutorial, we will discuss two main aspects: how to prevent LLMs from generating misinformation and how to detect misinformation generated by LLMs.

- Prevention strategies: Current research on prevention strategies focus more on how to mitigate the hallucination phenomenon in LLMs [50, 61], enabling them to generate more accurate and reliable information. This can be achieved by modifying the input prompts of LLMs, including retrieving information from external knowledge bases to guide LLMs in generating more accurate and reliable content, known as the Retrieval-Augmented Generation (RAG) method [15, 52, 57]. Additionally, designing appropriate prompting strategies can also guide LLMs to generate more accurate and reliable content [14, 28]. Furthermore, research on LLMs' decoding strategies can help generate more accurate content [13]. Moreover, some studies focus on aligning LLMs during the training phase to ensure that the content generated by LLMs better aligns with human values and preferences [5]. This not only mitigates the hallucinations generated by LLMs [60] but also enables LLMs to refuse some malicious instructions to generate misinformation [6]. However, it is important to note that for intentionally generated misinformation, prevention strategies cannot guarantee complete prevention of LLMs from generating misinformation [10, 59]. Therefore, we also need detection strategies to identify misinformation.
- Detection strategies: We will discuss the detection strategies for LLM-generated misinformation from two perspectives. The first perspective focuses on whether we can detect text generated by LLMs [49]. This can be achieved through black-box approaches, such as training a classifier to distinguish between LLM-generated and human-written text [26] or exploring the probabilities assigned by LLMs to the generated text [40]. Alternatively, white-box approaches can be employed, such as incorporating watermarks into the text generation process of LLMs to facilitate detection [30, 36]. The second perspective is to attempt to use LLMs themselves to detect misinformation [19, 20, 25]. This can be done by utilizing the knowledge embedded within the LLMs to make judgments [10, 56] or by employing retrieval methods to leverage external knowledge bases for retrieval-enhanced judgments [11, 12]. Lastly, some works openly discuss whether misinformation generated by LLMs can be detected at all [10]. This is an important consideration, as the increasing sophistication of LLMs may make it challenging to distinguish between genuine and misleading content.

In this tutorial, we aim to provide a comprehensive overview of the current progress in both prevention and detection strategies for misinformation generated by LLMs.

Necessity and timely of this tutorial. Mitigating and detecting misinformation generated by large language models (LLMs) is a relatively new topic that has become particularly important recently as the quality of LLM-generated text continues to improve. In recent years, a significant number of relevant papers have been published at top international conferences (such as ACL, ICML, NeurIPS, SIGIR, EMNLP, etc.), with one paper receiving the ICML Outstanding Paper Award [30]. We believe that more valuable work will emerge in this field in the future. This tutorial is highly necessary to provide a comprehensive understanding to researchers, practitioners, and decision-makers in this field, helping them better carry out relevant work.

3.2 Objectives

The objective of this tutorial is to provide a comprehensive overview of the current strategies for preventing and detecting misinformation generated by large language models. It aims to equip attendees with the knowledge and skills needed to develop and implement effective techniques to mitigate the harm caused by LLM-generated misinformation. Furthermore, this tutorial seeks to foster discussion, raise awareness, inspire new research directions, and contribute to the responsible development and deployment of LLM technologies in various domains.

3.3 Relevance

This tutorial is highly relevant to the SIGIR community, as information retrieval and large language models have become closely intertwined fields that can significantly benefit from each other. Previous tutorials at SIGIR 2022 [9] and ACL 2023 [4] have discussed retrieval-augmented generation and the generative capabilities of LLMs. Additionally, tutorials at EMNLP 2023 [32, 58] have discussed the potential harms associated with LLMs.

What sets our tutorial apart is its specific focus on mitigating the harm caused by misinformation generated by LLMs, which presents a more targeted and challenging problem. Retrieval techniques play a crucial role in this context, and with the increasing application of LLMs in the retrieval domain, this tutorial offers a valuable opportunity for SIGIR attendees to gain a deeper understanding of how LLMs are being utilized in their field.

3.4 Format and Schedule

This tutorial will be held offline, and all speakers will present it on site. The tentative schedule of this half-day tutorial (3 hours plus breaks) is as follows:

- `9:00-9:30`: Introduction
  - Large Language Models (LLMs) and their capabilities [1–3, 29, 45, 51]
  - Misinformation generated by LLMs
    * Unintentional generated misinformation. (Hallucination) [27, 44, 59]
    * Intentional generated misinformation. [10, 41]
  - Overview of prevention and detection strategies
- `9:30-10:30`: Prevention strategies
  - Retrieval augmented generation (RAG) [15, 52, 57]
  - Prompting techniques [14, 28]
  - Decoding based methods [13]
  - LLM alignment training [5, 6, 60]
- `10:30-11:00`: Coffee break
- `11:00-12:10`: Detection strategies
  - LLM generated text detection
    * Black-box detection [26, 40]
    * White-box detection [30, 36]
  - Misinformation detection with LLMs [10–12, 19, 25, 31, 56]
  - Could LLM-Generated Misinformation be Detected? [10]
- `12:10-12:25`: Open problems, future directions and conclusions
  - Misinformation in Multimedia Content
  - Transferability of Solutions
  - Policy and Governance
  - Conclusions
- `12:25-12:30`: Q&A

3.5 Qualification of presenters

We have been working on preventing and detecting Misinformation generated by large language models for a long time and have published a series of related works in top-tier conferences [7, 18, 20–24, 34, 35, 37–39, 46, 47, 63], including ICLR 2024, AAAI 2024, SIGIR 2023, EMNLP 2022, TKDE, MM 2023, ACL 2022, SIGIR 2022 and ACL 2020. These studies range from research on hallucinations in large models [22], to retrieval-augmented approaches[24], to alignment techniques [33], and to using large models for misinformation detection [7, 20, 21, 46, 47, 63] and retrieval of text generated by large models [35, 36]. We are very familiar with both lines of research on this topic and have contributed surveys on knowledge retrieval for LLM [54] as well as watermarking techniques for LLMs [37].

4 TURORIAL MATERIALS

- **Duration:** The tutorial is planned as a 3-hour tutorial.
- **Interaction style:** This is a lecture-style tutorial.
- **Turorial materials:** The slides will be released at https://sigir24-llm-misinformation.github.io/.
- **Organization details:** The tutorial can be conducted through both in-person and online formats, with all presenters intending to be physically present and lead the tutorial. If needed, we can offer a pre-recorded lecture as well. Additionally, with permission, we can also live-stream the tutorial using well-known video streaming platforms.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
[2] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867* (2023).
[3] Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Online. Available: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.
[4] Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Tutorial Proposal: Retrieval-based Language Models and Applications. In *The 61st Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. 41.
[5] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).
[6] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).
[7] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Combating online misinformation videos: Characterization, detection, and future directions. In *Proceedings of the 31st ACM International Conference on Multimedia*. 8770–8780.
[8] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
[9] Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 3417–3419.
[10] Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788* (2023).
[11] I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv:2307.13528* (2023).
[12] Tsun-Hin Cheung and Kin-Man Lam. 2023. FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 846–853.
[13] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883* (2023).
[14] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495* (2023).
[15] Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024. Retrieve Only When It Needs: Adaptive Retrieval Augmentation for Hallucination Mitigation in Large Language Models. *arXiv preprint arXiv:2402.10612* (2024).
[16] Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246* (2023).
[17] Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Lida Chen, Xintao Wang, Yuncheng Huang, et al. 2023. Can Large Language Models Understand Real-World Complex Instructions? *arXiv preprint arXiv:2309.09150* (2023).
[18] Zhiwei He, Binglin Zhou, Hongkun Hao, Aiwei Liu, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, and Rui Wang. 2024. Can Watermarks Survive Translation? On the Cross-lingual Consistency of Text Watermark for Large Language Models. *arXiv preprint arXiv:2402.14007* (2024).
[19] Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Leveraging ChatGPT for efficient fact-checking. *PsyArXiv. April* 3 (2023).

[20] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2023. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *arXiv preprint arXiv:2309.12247* (2023).

[21] Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. Learn over Past, Evolve for Future: Forecasting Temporal Trends for Fake News Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. 116–125.

[22] Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2024. Do Large Language Models Know about Facts?. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=9OevMUdods

[23] Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and S Yu Philip. 2022. CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3362–3376.

[24] Xuming Hu, Zhaochen Hong, Chenwei Zhang, Aiwei Liu, Shiao Meng, Lijie Wen, Irwin King, and S Yu Philip. 2023. Reading broadly to open your mind improving open relation extraction with search documents under self-supervisions. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[25] Yue Huang and Lichao Sun. 2023. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. *arXiv preprint arXiv:2310.05046* (2023).

[26] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650* (2019).

[27] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[28] Ziwei Ji, YU Tiezheng, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[29] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).

[30] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*. PMLR, 17061–17084.

[31] Sai Koneru, Jian Wu, and Sarah Rajtmajer. 2023. Can Large Language Models Discern Evidence for Scientific Hypotheses? Case Studies in the Social Sciences. *arXiv preprint arXiv:2309.06578* (2023).

[32] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Mitigating Societal Harms in Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*. 26–33.

[33] Aiwei Liu, Haoping Bai, Zhiyun Lu, Xiang Kong, Simon Wang, Jiulong Shan, Meng Cao, and Lijie Wen. 2024. Direct Large Language Model Alignment Through Self-Rewarding Contrastive Prompt Distillation. *arXiv preprint arXiv:2402.11907* (2024).

[34] Aiwei Liu, Xuming Hu, Li Lin, and Lijie Wen. 2022. Semantic enhanced text-to-sql parsing via iteratively learning schema linking graph. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1021–1030.

[35] Aiwei Liu, Leyi Pan, Xuming Hu, Shuang Li, Lijie Wen, Irwin King, and Philip S. Yu. 2024. An Unforgeable Publicly Verifiable Watermark for Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=gMLQwKDY3N

[36] Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. 2023. A semantic invariant robust watermark for large language models. *arXiv preprint arXiv:2310.06356* (2023).

[37] Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Lijie Wen, Irwin King, and Philip S Yu. 2023. A survey of text watermarking in the era of large language models. *arXiv preprint arXiv:2312.07913* (2023).

[38] Aiwei Liu, Honghai Yu, Xuming Hu, Li Lin, Fukun Ma, Yawen Yang, Lijie Wen, et al. 2022. Character-level White-Box Adversarial Attacks against Transformers via Attachable Subwords Substitution. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 7664–7676.

[39] Yijian Lu, Aiwei Liu, Dianzhi Yu, Jingjing Li, and Irwin King. 2024. An Entropy-based Text Watermarking Detection Method. *arXiv preprint arXiv:2403.13485* (2024).

[40] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*. PMLR, 24950–24962.

[41] Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661* (2023).

[42] Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780* (2023).

[43] Russell A Poldrack, Thomas Lu, and Gašper Beguš. 2023. AI-assisted coding: Experiments with GPT-4. *arXiv preprint arXiv:2304.13187* (2023).

[44] Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922* (2023).

[45] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

[46] Qiang Sheng, Juan Cao, H Russell Bernard, Kai Shu, Jintao Li, and Huan Liu. 2022. Characterizing multi-domain false news and underlying user effects on Chinese Weibo. *Information Processing & Management* 59, 4 (2022), 102959.

[47] Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating pattern- and fact-based fake news detection via model preference learning. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 1640–1650.

[48] Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can ChatGPT replace traditional KBQA models? An in-depth analysis of the question answering performance of the GPT LLM family. In *International Semantic Web Conference*. Springer, 348–367.

[49] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205* (2023).

[50] SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313* (2024).

[51] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[52] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214* (2023).

[53] Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2023. Disinformation capabilities of large language models. *arXiv preprint arXiv:2311.08838* (2023).

[54] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521* (2023).

[55] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205* (2023).

[56] Guangyang Wu, Weijie Wu, Xiaohong Liu, Kele Xu, Tianjiao Wan, and Wenyi Wang. 2023. Cheap-fake Detection with LLM using Prompt Engineering. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 105–109.

[57] Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Cheng Niu, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396* (2023).

[58] Qiongkai Xu and Xuanli He. 2023. Security challenges in natural language processing models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*. 7–12.

[59] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817* (2024).

[60] Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation. *arXiv preprint arXiv:2402.09267* (2024).

[61] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* (2023).

[62] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[63] Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2120–2125.