Citation: A Key to Building Responsible and Accountable Large Language Models

Jie Huang **Kevin Chen-Chuan Chang**
Department of Computer Science, University of Illinois at Urbana-Champaign
{jeffhj, kcchang}@illinois.edu

Abstract

Large Language Models (LLMs) bring transformative benefits alongside unique challenges, including intellectual property (IP) and ethical concerns. This position paper explores a novel angle to mitigate these risks, drawing parallels between LLMs and established web systems. We identify "citation" as a crucial yet missing component in LLMs, which could enhance content transparency and verifiability while addressing IP and ethical dilemmas. We further propose that a comprehensive citation mechanism for LLMs should account for both non-parametric and parametric content. Despite the complexity of implementing such a mechanism, along with the inherent potential pitfalls, we advocate for its development. Building on this foundation, we outline several research problems in this area, aiming to guide future explorations towards building more responsible and accountable LLMs.

1 Introduction

The landscape of artificial intelligence is undergoing rapid transformation, spurred by the emergence of large language models (LLMs) such as ChatGPT/GPT-4 (OpenAI, 2022, 2023) and PaLM (Chowdhery et al., 2022; Anil et al., 2023). These models, recognized for their striking ability to generate human-like text, have shown enormous potential in various applications, from information provision to personalized assistance. Nonetheless, their capabilities bring along substantial risks, including intellectual property (IP) and ethical concerns (Carlini et al., 2021, 2023; Huang et al., 2022; Shao et al., 2023; Li et al., 2023; Lee et al., 2023; Frye, 2022; Chesterman, 2023; Bender et al., 2021; Brown et al., 2022; El-Mhamdi et al., 2022).

Research by Carlini et al. (2021); Huang et al. (2022), for instance, reveals that LLMs are prone to memorizing extensive segments of their training data, including sensitive information. This can result in violations of IP and ethical guidelines. Furthermore, studies by El-Mhamdi et al. (2022); Brown et al. (2022) suggest that current protective measures fail to provide a comprehensive and meaningful notion of safety for LLMs, making it seemingly impossible to develop safety-preserving, high-accuracy large language models even when trained on public corpora.

While the notion of building an entirely safe LLM might appear daunting, it is crucial to acknowledge that many well-established systems, such as the Web, grapple with similar challenges and have not yet reached absolute safety. Recent legislation like the Online News Act[1], which requires online search engines to compensate Canadian online news outlets for their content, underscores the ongoing issues around content use and compensation on the Web. Furthermore, the Web continues to be a breeding ground for both sensitive information and misinformation. Hence, expecting a completely risk-free LLM may be an over-ask. Instead, our focus should be on accurately quantifying these risks and developing effective mitigation strategies. It is not about achieving absolute security, but about responsibly managing and minimizing risks in an ethically sound manner.

[1] https://www.canada.ca/en/canadian-heritage/services/online-news.html

Guided by these insights, we propose to examine the problem through a different lens: *Can we draw parallels between the risks inherent to LLMs and those experienced by established systems such as search engines and the Web? Can we devise strategies to decrease these risks by aligning with the practices of these mature systems?*

In examining systems like the Web and search engines, we observe a common and robust practice employed to manage IP and ethical concerns: the use of "citations". Broadly defined, a "citation" refers to the act of mentioning or referencing a source or piece of evidence. For example, search engine results also serve as a form of citation, with each entry typically consisting of a title, URL, and brief description. These components collectively cite the webpage's content, offering the user an overview and inviting them to explore the source in greater depth. Citations thus act as anchors for accountability and credit in these systems, providing traceability, preventing plagiarism, and ensuring credit is correctly attributed. They also contribute to transparency, allowing users to verify the information's source.

Upon reflection, it becomes clear that LLMs lack this critical functionality. When LLMs generate content without citations, their output is perceived as independent and self-derived. This creates two significant issues: firstly, when the model produces valuable information, it fails to credit the source it relies on; secondly, when it generates harmful content, it becomes challenging to assign accountability. Incorporating the ability to cite could not only address these ethical and legal conundrums but also bolster the transparency, credibility, and overall integrity of the content generated by LLMs.

However, implementing a "citation" mechanism in LLMs is not as straightforward as it might seem. Unlike the Web, which explicitly links and references sources, LLMs internalize the information and transform it into an abstract representation, making accurate citation a significant technical challenge. Although some strides have been made in this direction, as seen in systems like New Bing, Bard, and Perplexity AI, they fall short on several fronts. First, the citations provided in the response of existing systems are often inaccurate (Liu et al., 2023; Gao et al., 2023). Moreover, these systems typically only cite *non-parametric* content, i.e., content directly retrieved from external sources such as the Web. However, they neglect *parametric* content, the knowledge embedded in the model parameters, which also needs appropriate credit attribution and consideration for potential harm.

This position paper embarks on an exploratory journey into the potential of integrating a citation mechanism within large language models, examining its prospective benefits, the inherent technical obstacles, and foreseeable pitfalls. We delve into approaches to cite both non-parametric and parametric content, considering the unique characteristics of each type. We also identify and discuss potential setbacks, such as reduced creativity, dissemination of sensitive information, and citation bias. Building on this foundation, we lay bare the hurdles in our path, presenting them as enticing problems for future research. Through this endeavor, we aim to stimulate further discussion and research towards building responsible and accountable large language models.

2 "Citation" in LLMs

Figure 1 illustrates model generations with and without citations. In the absence of a citation, there is a potential risk of misunderstanding, leading one to believe that the claim is an opinion or statement formulated by the LLM itself. This not only fails to appropriately credit the original authors, but could also result in ethical dilemmas if the claim is inaccurate or misrepresented.

LLMs memorize a lot of training data.	LLMs memorize a lot of training data [1].
Women are better suited for caregiving roles than men.	According to [1], women are better suited for caregiving roles than men.

Figure 1: Examples without (left) and with (right) citations. In the first case, citations serve as a way to appropriately credit authors. In the second case, citing the original source of a biased statement ensures that the bias is not misconstrued as the model's opinion.

On the other hand, the inclusion of citations can act as a multifaceted solution to these concerns. Primarily, it helps to mitigate intellectual property and ethical disputes by signaling that the information is not a product of the LLM's "opinion", but a reflection of the cited source. Additionally, citations enhance the transparency and verifiability of the LLM's output. By indicating the source from which the information is derived, they provide a clear pathway for users to independently verify the validity and context of the information.

3 RoadMap

In this section, we embark on exploring the potential of incorporating a "citation" mechanism within LLMs. We start our exploration by defining when it would be ideal for an LLM to provide a citation, drawing insights from established practices in academia and existing systems like search engines and the Web. We then delve into discussing the possible strategies for effectively implementing citations in LLMs, confronting the methodological and technical intricacies this endeavor involves.

3.1 When to Cite?

In academic or professional writing, a citation is typically required when using someone else's ideas, concepts, data, or specific language. For LLMs, determining when to provide a citation is a considerably more challenging task. Given the vast and varied range of queries posed to LLMs, it is crucial to establish when a citation would be appropriate or necessary.

A fundamental rule could be that any fact, idea, or concept that is not general knowledge should be cited. This mirrors the existing conventions on the Web, where sources for specific information are typically provided. For instance, widely known facts like "The Earth revolves around the Sun" would not necessitate a citation, while a less well-known fact like "The fastest spinning stars can rotate more than 600 times per second" would warrant one.

Moreover, the need for a citation could also depend on the nature of the task LLMs are performing. Certain tasks may not necessitate citations, particularly if the output is a reformulation or reinterpretation of the input. For example, in summarization tasks, LLMs condense the input data without introducing new information. The resultant summary is hence an interpretation of the input, and typically, a citation may not be needed for such tasks. Similarly, translation tasks involve converting content from one language to another, without the introduction of novel information.

In essence, while the need for citations in LLMs is task-dependent and context-specific, guiding principles should be the commitment to knowledge integrity, respect for intellectual property, and adherence to ethical norms. These are similar principles that guide the management of intellectual property and ethical concerns on the Web and in search engines.

3.2 How to Cite?

Incorporating citations in LLMs ideally involves connecting outputs to the original source of information. However, this presents a notable technical challenge. During LLMs' training, information is internalized and transformed into an abstract representation, unlike search engines which possess indices to track and retrieve information. In the case of LLMs, this index is absent, which makes referencing the original source a daunting task. In this section, we delve into the consideration of citations for both *non-parametric* and *parametric* content.

3.2.1 Citation for *non-parametric* content

As a potential solution to prevailing challenges, one could design a hybrid system that merges large language models with information retrieval (IR) systems. In this approach, the model is trained to discern when a citation might be required. Subsequently, the IR system is utilized to retrieve relevant sources, namely, *non-parametric* content. The LLM can then incorporate these sources into its responses as citations. We identify two strategies for citing non-parametric content:

- ***Pre-hoc citation***: This approach involves first identifying the need for a citation in the upcoming dialogue or content generation. Once this requirement is recognized, the LLM triggers the IR system to retrieve the necessary information. The LLM then generates its response, seamlessly incorporating the retrieved non-parametric content as citations. This technique can be associated with the broader body of research that augments language models with retrieval (Guu et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021; Borgeaud et al., 2022; Izacard et al., 2022; Shi et al., 2023; Wang et al., 2023; Menick et al., 2022; Huang et al., 2023).
- ***Post-hoc citation***: Conversely, in this strategy, the LLM initially produces a response. An evaluation process then scrutinizes the generated content to ascertain whether a citation is necessary. If a citation is deemed necessary, the IR system is used to locate the appropriate non-parametric content, which is subsequently inserted into the existing text as a citation. Related research includes

measuring or requiring attribution in LLMs (Rashkin et al., 2023; Gao et al., 2022; Honovich et al., 2022; Yue et al., 2023; Liu et al., 2023; Gao et al., 2023).

In practical applications, a combination of both *pre-hoc* and *post-hoc* citation methods could be adopted for an optimized method. This mixed approach would employ the initial identification and retrieval of potential citations in line with the *pre-hoc* method, followed by a *post-hoc* evaluation to refine the integration of citations based on the generated content. This blend of proactive retrieval and reactive refinement could facilitate the creation of robust, accurate, and well-supported content, while also mitigating intellectual property and ethical concerns surrounding LLMs.

3.2.2 Citation for *parametric* content

In addition to the *non-parametric* content, i.e., content directly retrieved from external sources such as the Web, *parametric* content, which refers to information internalized from the training data, also needs appropriate credit attribution and consideration for potential harm. However, crafting a citation strategy for *parametric* content presents its own set of unique challenges.

The fundamental challenge is the underlying nature of how LLMs process and internalize information. During training, LLMs assimilate vast amounts of data and transform them into an intricate, high-dimensional space that represents learned patterns and structures. The transformation process, rooted in complex mathematical operations, does not inherently retain any clear mapping back to individual data points in the training set. Consequently, generated content cannot easily be traced back to specific training data (Koh and Liang, 2017; Bae et al., 2022; Park et al., 2023; Grosse et al., 2023).

This situation is further complicated by the fact that an output generated by LLMs is typically influenced by a multitude of training data points, rather than a single source. This is due to the multi-faceted and context-sensitive nature of language understanding and generation, where a single output can be influenced by a diverse range of linguistic patterns and structures. Thus, the task of accurately attributing a generated output to specific training data pieces is a complex and multifaceted problem that involves unpacking the high-dimensional representations in the model.

Despite these challenges, potential solutions exist. A conceivable approach involves ***training the model with source identifiers***, essentially tags that link specific pieces of information back to their original sources. During training, the model could then be encouraged to retain these identifiers. This would provide a more transparent lineage of information, thereby enhancing accountability. A relevant attempt in this direction was made by Taylor et al. (2022), which used special reference tokens to wrap citations and trained models to predict these citations. However, it exhibited certain limitations, such as citation inaccuracy and confinement to academic citations. The successful execution of this method would likely call for advancements in model architecture and training techniques, thereby highlighting intriguing directions for future research.

4 Pitfalls of Citation in LLMs

While citations in LLMs can potentially mitigate risks such as IP and ethical issues, as well as improve transparency and verifiability, it is crucial to consider potential pitfalls.

Over-Citation and Sensitive Information Dissemination. The implementation of a citation system in LLMs poses the risk of over-citation, where the excessive use of references might expose more information than necessary. This overexposure could lead to information overload, diluting the significance of critical citations. Moreover, over-citation might inadvertently elevate the risk of disseminating sensitive information Huang et al. (2022); Shao et al. (2023); Li et al. (2023). An ill-intentioned user could exploit these extensive citations to gather sensitive or private information.

Inaccurate Citations. Another potential pitfall of implementing citations in LLMs is the risk of inaccurate citations (Liu et al., 2023; Gao et al., 2023). Given that LLMs may not possess a deep understanding of the content they are trained on or the sources they are citing, there is a chance that they could incorrectly attribute information to a source that does not actually contain that information. Inaccurate citations could mislead users, causing them to believe that a piece of information is verified and supported by a credible source, when in fact, it is not.

Outdated Citations. With the continuous expansion and evolution of knowledge, there is a risk that the sources an LLM cites may become outdated or irrelevant over time. This is particularly likely in

fast-evolving fields where new discoveries or advancements quickly supersede existing knowledge. As LLMs are trained on a fixed dataset, their generated content and the sources they cite may not reflect the most current or accurate information. Therefore, there is a potential for LLMs to propagate outdated knowledge, misleading users who rely on the generated content and the cited sources for information.

Propagation of Misinformation. The risk of propagation of misinformation presents a significant concern in the application of LLMs (Pan et al., 2023). As LLMs generate output based on the data they have been trained on, there is a chance they could inadvertently cite or echo unreliable or misleading sources, thereby spreading misinformation. This problem could potentially be amplified by the addition of a citation mechanism. A misinterpreted or incorrect citation could be perceived as an authoritative endorsement, inadvertently lending credibility to inaccurate or misleading content.

Citation Bias. Implementing citations in LLMs can also lead to citation bias (Jannot et al., 2013; Greenberg, 2009). Models may prefer to cite certain types of sources over others, either due to the characteristics of the training data or inherent biases in the retrieval mechanism of the IR system. This could lead to an over-reliance on certain types of information and unintentional promotion of certain viewpoints.

Potential for Diminished Creativity. The integration of citations could inadvertently cause a decrease in the creative outputs of the model. When prompted to generate innovative text or propose creative solutions, LLMs might become over-reliant on existing, citable information, thus stifling their novel content generation.

Legal Implications. The utilization of citations could also bring forth legal implications. The introduction of citations could imply a level of responsibility and accountability that the LLM, as an artificial entity, is not equipped to handle. Legal systems worldwide are yet to reach a consensus on how to treat legal issues related to artificial intelligence and its outputs, and the inclusion of citations could further complicate these discussions.

5 Barriers and Research Problems

Building on the potential solutions and pitfalls discussed above, we delve into the primary barriers and corresponding research problems.

Determining When to Cite. Deciding when an LLM should cite its sources is a complex issue. While it may be intuitive to suggest that LLMs should always cite sources when they generate information that is not common knowledge (§3.1), defining what constitutes "common knowledge" is itself a difficult task. Furthermore, as discussed in §4, it is essential to consider the potential risks associated with over-citation, particularly the increased risk of sensitive information dissemination Huang et al. (2022); Shao et al. (2023); Li et al. (2023). LLMs may inadvertently expose sensitive information or contribute to information overload if they include unnecessary or excessive citations. Balancing the need for transparency and accountability with the need to protect privacy and prevent information overload is a critical challenge that needs to be addressed.

Addressing Hallucination in Citation. Hallucination in large language models refers to the phenomenon where the models generate information not grounded in their training data, and that cannot be verified or is simply incorrect (Ji et al., 2023). The incorporation of a citation feature can both alleviate and exacerbate this issue. On the one hand, requiring LLMs to link generated information to a tangible source can serve as a form of external verification, potentially restraining the model from generating completely baseless or hallucinated content. The requirement for a source may encourage the model to better align its output with the available data, thereby reducing the likelihood of hallucination. On the other hand, the citation mechanism itself can potentially hallucinate. If not meticulously designed and implemented, it may end up citing incorrect or non-existent sources (Liu et al., 2023; Gao et al., 2023). This presents a twofold challenge: Not only is the generated content incorrect, but the citation misleads users into believing that the content is verified and substantiated by the cited source. This issue necessitates the development of techniques to enhance the model's ability to accurately represent the information present in the source, and equip the model to cross-check the consistency of the generated content with the content of the cited source.

Maintaining Temporal Relevance of Citations. In the pursuit of an effective citation mechanism within LLMs, it is essential to address the need for the model's ability to stay updated with the most

recent and relevant knowledge. One potential approach towards this challenge is inspired by the operational principles of search engines. In their bid to stay relevant, search engines continuously update their indexes and ranking algorithms to reflect the latest web. A similar approach could be adopted for LLMs, where they could be designed for ongoing training on updated datasets, and their citation mechanisms could favor more recent sources. However, executing this in practice presents a significant research problem, considering the scale and complexity of continuously training LLMs and updating their citation mechanisms. Exploring efficient techniques for model training and designing citation mechanisms capable of consistently prioritizing the most recent and relevant sources will require substantial research and development.

Evaluating Source Reliability. Another important challenge is evaluating the reliability of sources used for training data and citations. As mentioned in §4, LLMs could potentially propagate misinformation if they cite unreliable or misleading sources. While search engines face similar challenges, they are equipped with advanced algorithms to evaluate the reliability and relevance of web pages (Page et al., 1999). Implementing analogous systems within the framework of LLMs presents an interesting and crucial direction for further exploration.

Mitigating Citation Bias. Citation bias in LLMs, as discussed in §4, can result in the uneven representation of information, leading to the propagation of certain viewpoints while others are neglected. Formulating strategies to curtail such tendencies is paramount. To begin with, sourcing a more balanced selection of training data can mitigate bias at the inception stage. Ensuring diversity in terms of viewpoints and topics in the training data can reduce bias to some extent. During citation retrieval, LLMs should utilize an impartial mechanism that does not favor specific types of sources. The underlying algorithms should be optimized to retrieve citations based on their relevance and credibility rather than the prominence of the source or its frequency in the training data. Finally, the development and application of effective evaluation techniques can help identify and measure any residual bias in LLM outputs. Quantifying the extent of bias enables more targeted corrective measures and provides an objective measure of their efficacy.

Balancing Existing Content with Novel Content Generation. Another intriguing area of research centers on striking a balance between the frequency of citing existing content and generating novel content. LLMs are admired for their capacity to generate creative and unique content (Franceschelli and Musolesi, 2023), as well as their reasoning ability (Wei et al., 2022; Huang and Chang, 2023). An over-reliance on citations could potentially inhibit these attributes, reducing the model to a mere aggregator of existing knowledge rather than a generator of new ideas. Research into this would involve the development of techniques that allow for appropriate citation without hampering the model's creativity. One potential approach could be to create models that are capable of determining the novelty of their generated content and adjusting their citation behavior accordingly. For instance, if a model is generating content based heavily on its training data or the retrieved content, it should provide appropriate citations. Conversely, if the model is generating content that is significantly different from its training data and the retrieved content, it might deem citation unnecessary. Developing such capabilities would require significant advancements in understanding how LLMs generate novel content and how to quantify the 'novelty' of such content.

Navigating Copyright and Fair Use Laws. The application of citation mechanisms in LLMs opens up a new array of legal challenges. Understanding and complying with copyright and fair use laws when citing sources is a complex issue. For instance, how much quoted material from a source would be considered fair use and under what conditions can it be used? In many jurisdictions, the law is not completely clear, especially as it applies to the use of AI technology. Thus, research in the legal aspects of using LLMs for generating text with citations is crucial to ensure legal compliance.

6 Discussion and Conclusion

While introducing a citation mechanism in LLMs presents an exciting opportunity for enhancing responsibility and accountability, implementing such a system is not without its technical challenges. Our paper introduces this concept with a hopeful perspective, but readers should be cognizant of the numerous technical hurdles that must be overcome, as highlighted in Section 4 and Section 5. Nevertheless, these challenges also represent valuable areas for future research and innovation. By addressing these issues head-on, we believe there is potential to unlock the true benefits, leading to more responsible and accountable LLMs.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *ArXiv preprint*, abs/2305.10403.

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B Grosse. 2022. If influence functions are the answer, then what is the question? In *Advances in Neural Information Processing Systems*, volume 35, pages 17953–17967. Curran Associates, Inc.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6.

Simon Chesterman. 2023. Ai-generated content is taking over the world. but who owns it? *But Who Owns it*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311.

El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyên Hoang, Rafael Pinot, and John Stephan. 2022. On the impossible safety of large ai models. *ArXiv preprint*, abs/2209.15259.

Giorgio Franceschelli and Mirco Musolesi. 2023. On the creativity of large language models. *ArXiv preprint*, abs/2304.00008.

Brian L Frye. 2022. Should using an ai text generator to produce academic writing be plagiarism? *Fordham Intellectual Property, Media & Entertainment Law Journal, Forthcoming*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2022. Rarr: Researching and revising what language models say, using language models. *ArXiv preprint*, abs/2210.08726.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *ArXiv preprint*, abs/2305.14627.

Steven A Greenberg. 2009. How citation distortions create unfounded authority: analysis of a citation network. *Bmj*, 339.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Jie Huang, Wei Ping, Peng Xu, Mohammad Shoeybi, Kevin Chen-Chuan Chang, and Bryan Catanzaro. 2023. Raven: In-context learning with retrieval augmented encoder-decoder language models. *arXiv preprint arXiv:2308.07922*.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv*, 2208.

Anne-Sophie Jannot, Thomas Agoritsas, Angèle Gayet-Ageron, and Thomas V Perneger. 2013. Citation bias favoring statistically significant studies was present in medical research. *Journal of clinical epidemiology*, 66(3):296–301.

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. *ACM Computing Surveys*, 55(12):1–38.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2023. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, pages 3637–3647.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. 2023. Multi-step jailbreaking privacy attacks on chatgpt. *ArXiv preprint*, abs/2304.05197.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *ArXiv preprint*, abs/2304.09848.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *ArXiv preprint*, abs/2203.11147.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

OpenAI. 2023. Gpt-4 technical report.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

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*, abs/2305.13661.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, pages 1–66.

Hanyin Shao, Jie Huang, Shen Zheng, and Kevin Chen-Chuan Chang. 2023. Quantifying association capabilities of large language models and its implications on privacy leakage. *ArXiv preprint*, abs/2305.12707.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *ArXiv preprint*, abs/2301.12652.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *ArXiv preprint*, abs/2211.09085.

Boxin Wang, Wei Ping, Peng Xu, Lawrence McAfee, Zihan Liu, Mohammad Shoeybi, Yi Dong, Oleksii Kuchaiev, Bo Li, Chaowei Xiao, et al. 2023. Shall we pretrain autoregressive language models with retrieval? a comprehensive study. *ArXiv preprint*, abs/2304.06762.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *ArXiv preprint*, abs/2305.06311.