Literature Review

Filip Dimitrieviski, Akram Moustafa

April 2025

I Overview

A Purpose

This is a literature review paper, and it presents as a review of a current research problem that needs to be addressed and requires further attention. This paper focuses on an existing research topic that is becoming increasingly prominent and investigates the connection between two different areas of research. The primary objective is to examine the current research state to highlight a research gap for investigation. We focus our evaluation and goals on identifying the research problem, without providing specific methodologies for implementation.

B Limitations

This review has several limitations that should be considered when interpreting the findings of future research directions. The scope of the paper is restricted to sources published in English language, which might fail to provide full insights from different language sources. Most of the sources used in the review are from recent studies, limiting the engagement with earlier studies in the certain topics. Because our preview is focused on finding the connection between different aspects, while providing detailed description of the component in the analysis, it limits our understanding of the wide representations of different perspectives. Lastly, there are only few studies focused on exploring the connection between these aspects, which limits our understanding of their relationship.

C Components

First, we defined Large Language Models as artificial intelligence models that are used to communicate with humans. We show that those models are still evolving with chatGPT-4 being released, which shocked the public with its capabilities. We show that those models are becoming manifested in different important areas and wide use, such as healthcare. Despite that those models are becoming advanced there are still challenges associated with them. Due to these problems the number of research papers associated with LLMs are still increasing. We then defined those accuracy issues as hallucinations, in which they provide results that are not backed up by ground truth. We also showed that hallucinations are complex based on several existing different types. We extend the complexity issue by showing that existing strategies to detect and mitigate hallucinations don't work since most of the studies of the studies were not able to show the sources of those hallucinations. Because of the reliability issues associated with these techniques, we tried to find an actual source of hallucinations, and we found that biases is one of the hidden sources based on direct relationship. We gave background on biases as definition on how they exist as hidden biases, and we showed mitigation and detection strategies. Then we found our research gap that while studies showed and detected hallucinations induced biases they did not provide mitigation strategies.

D Funnel Model

We used the funnel model in identifying our research idea. We found our interest in studying and investi-

gating the field of LLMs. Then we found that there are many problems that do exist in those models. We identified the problems that do exist as inaccuracy and hallucinations. Then we found that another problem exists, where the existing strategies don't focus on the sources of hallucinations while trying to detect and mitigate them. Then we narrowed down our understanding that there are some other types of hallucinations that are caused by biases. We then found that those existing studies don't focus on mitigating those types of hallucinations. Rather, they only detect and find the relationship. We identified this as our research gap.

II Introduction to Large Language Models

A Large Language Models are Becoming Increasingly Advanced as a Result of Decades of Research and Development

Large Language Models (LLMs) are artificial intelligence models that are designed to understand and generate human-like language [1], [2]. These models are advanced as they are trained on hundreds of billions of parameters, which include books, websites, articles, and other sources [1]. Due to the large amount of pretraining data, LLMs can perform complex tasks, such as answering questions, translating languages, or writing text [1], [2]. In particular, the neural networks in LLMs mimic how the human brain works, in which they help the models process entire sequences of text instead of focusing on individual words [2]. LLMs advancements are the accumulation of ongoing text for decades in areas such as Artificial intelligence, language understanding, and neural networks, starting from the 1950s [1].

B Evolving capabilities of Large Language Models

Today we have more advanced models, like ChatGPT, Bert, Llama and more models. The capabilities of LLMs are constantly growing with GPT-4 being released in 2024, which showed that models can have human like performance [3]. GPT-4 can also get different forms of inputs that include texts and images, and achieved high scores on different benchmarks related to professional and academic fields [3]. LLMs have been growing and becoming more reliable.

C Applications for Large Language Models

LLMs are being used in many fields, such as healthcare, education, and finance. For example, a study was made on experts' opinions on LLMs use in healthcare. The experts were gathered from health informatics, nursing informatics, and medical natural language processing. They suggested that LLMs can be used in areas such as handling data extraction, automation of processes, improving quality of health care services, personalized care, diagnosis and treatment processes, and improving interactions between patients and health care professionals [4].

III Hallucinations in LLMs

A Challenges with Accuracy in LLMs

Because of the advancements in LLMs and the wide use in many areas, the number of research papers has surged to reach 28400 research papers with the keyword LLMs [1]. While these research papers show the ongoing evolution of LLMs, they also reflect some challenges associated with LLMs, such as accuracy issues that need to be addressed as they were found in the most recent models [1], [2], [3]. An experiment was done on GPT-4 with the benchmark TruthfulQA used for evaluating LLMs, and although GPT-4 showed better results in most areas than previous models, it still generated inaccurate results in some tasks, showing that LLMs still have the tendency to provide inaccuracy in their output [2]. Inaccuracy generated answers (hallucinations) happen when the model provides output that is not related to ground truth according to some studies [5], [6], [7], [8], [9]. Some other studies extended the definition by stating

that hallucinations happen when the models give factually incorrect results with high confidence [6], [7]. However, generalizing the definition of hallucinations can be difficult, as there are many types.

B Hallucinations Types

Today exist many types of hallucinations including, Entity-error hallucination that involves mistakes in facts like names, dates, places, or objects. Relation-error hallucination refers to incorrect relationships between entities. Incompleteness hallucination happens when answers are missing or partial. Outdatedness hallucination involves generating facts that were once true but are now outdated. Overclaim hallucination includes exaggerated or absolute claims, and unverifiability hallucination describes content that cannot be verified by real sources [10]. For some other studies also addressed two main types addressed hallucinations as intrinsic and extrinsic [11]. Intrinsic hallucinations happen when the model's answer is opposite from ground truth, while extrinsic hallucinations happen when the model's answer cannot be verified by an external source [11]. Lastly, the other two types of hallucinations are studied as factual and non-factual hallucinations [10]. Factual hallucinations occur as models give wrong facts, and it can be one of the least noticeable source of hallucinations [8]. Now we are able to identify different types of hallucinations, are we able to get rid of them or mitigate them at least?

C Techniques for Mitigating and Detecting Hallucinations

Hallucinations have been shown to persist in generating hallucinations in their output, and some studies claim that they are inevitable. The persistent relation between hallucinations and LLMs has been shown using some math equations in the proof [8]. However, studies on LLMs have shown that it's still possible to increase accuracy in models and decreasing hallucinations through finding techniques for reducing certain types of hallucinations [2], [5], [6], [7], [12], [10], [11]. A comprehensive empirical study created a benchmark called HaluEval 2.0, which is used to address factual hallucinations. The mitigation process was done by sending other model's responses, such as Alpaca, Vicuna, ChatGPT, Llama 2-Chat, and Claude to ChatGPT4 for evaluation and mitigation. ChatGPT4 was used mainly to extract the facts from the response based on factual ground truth. Mitigating hallucinations include three strategies, human feedback, refining the answer using outer models, and advanced decoding [10]. While some studies still show that decreasing hallucinations is possible, the detection of hallucinations is still difficult. SelfCheckGPT is a technique created solely to detect hallucinations. SelfCheckGPT is based on allowing LLMs to detect hallucinations before generating the final output. The importance of addressing hallucinations is found as there are many other techniques associated with SelfCheckGPT to limit hallucinations. These techniques include, Selfcheck with Question Answering, Prompt, Natural Language Response, and Ngram model [11]. Factual hallucinations can be detected by finding and testing the key concepts. The key concepts are usually the keywords in the sentence that show whether the facts are based on ground truth or are incorrect [13]. Some of the existing studies were trying to address the original presence of hallucinations in LLMs. However, those studies failed to capture the underlying sources of hallucinations.

D Importance for Addressing Hidden Sources of Hallucinations

Some studies claimed that the sources of hallucinations do exist due to model's uncertainty in the generated answers [12]. In other words, studies associated models' hallucinations with high certainty. However, models sometimes hallucinate with high confidence and existing data, which made it difficult for these studies to find the hidden patterns and reasons why hallucinations still persist in those models. One study focused on proving that hallucinations can happen when the model has existing knowledge and facts in its pretraining data, and it named this type of hallucination as Certain Hallucinations Overriding Known Evidence (CHOKE). Detecting those hallucinations based on the model's certainty can be im-

possible because the models show high certainty percentage in their answers. The framework consists of injecting the dataset with a dataset for finding when the model gives the right answer consistently. The dataset the model is tested on is called TriviaQA for detecting hallucinations, and it has open ended questions. Then the input becomes trickier so that the model starts hallucinating using approaches such as Child Setting (mimicking the tone of a child) and Alice-Bob Setting (mimicking that of a student). The technique focused only on the output generated when the model's technique for choosing the words with the highest certainty. The study found that hallucinations can happen even when the model has a certainty that passes a certain threshold [12]. Since finding the source of hallucinations can be puzzling, some studies moved to different techniques for finding the source of hallucinations. One of the reasons that the sources of hallucinations remain unfound is that most of the previous studies have tried to figure out hallucinations' sources using black box models [5], [6], [7]. These model do not show the underlying components or the steps they took to reach a certain answer, and they focus on the output of the model for hallucinations evaluation. Because of the failure in identifying the original sources of hallucinations, their studies became unreliable [14]. Therefore, some studies moved to investigate the sources of hallucinations using reasoning process, use white-box models, and focus on inference tasks [9], [15], [14]. One study used white box models to focus on how the inner components of models change when hallucinations and when producing factual information [9]. It found that one source of hallucinations is when the models shows animalities in the state transition, which means when the models move through different states when hallucinating [9]. Other studies also provide clear steps on a direct relationship between hallucinations and some other aspects in models' evaluation.

IV Social Bias as a Possible Source of Hallucinations

A Exploring the Direct Relationship between Hallucinations and Biases

A study was made that tries to understand social bias affecting unfairness hallucinations through a causal relationship. In other words, the study focused on identifying how certain types of bias might directly induce hallucinations in LLMs. The idea is removing all the aspects that generate bias alone or generate hallucinations alone, which is called cofounders. Then the study tried to find how changing from different biases might influence hallucinations. Different bias states include Anti-stereotype (Reverse stereotype), Pro-stereotype (common stereotype), and Non-stereotype (no stereotype). The experiments involved using the BBQ dataset, mentioned earlier, that had person 1 and person 2, both of them had attributes for reverse stereotypes (or different stereotypes). The study also created another dataset called BID, as a subset of BBQ that can find precise measurements of biases. It encompasses 9 social biases with ten thousand entries. Based on evaluating seven LLMs, the study found that social biases directly affect different types of hallucinations, such as common and unfairness hallucinations [16].

B Biases Affect the Reasoning Process of Models

A study was conducted on DeepSeek-R1, which focused on addressing or finding one of the hidden sources of hallucination that could be affected by biases. The study focused on implicit bias effects on hallucinations. It tried to solve the question; how implicit bias affects the reasoning process of factual hallucinations without fine tuning the models. Deep Seek-R was given dataset for finding biases called BBQ dataset, which contains MCQ questions, with some of the questions being ambiguous. The model was tested based on the steps the model gave to generate the answer. The authors prompted ChatGPT with those steps and received a score between 0 and

4, with 0 being neutral and 4 being extreme bias. The evaluation was also done on different models, including open-source instruction-tuned models, proprietary API-based models, and state of-the-art. It was shown that when the model did not show biases in the reasoning steps it took to generate an answer, the model's accuracy was high. On the other hand, the accuracy decreased when the model showed biases in reasoning process [15]. Now we can understand how hallucinations might be affected by biases, there are some other studies that show that hallucinations are directly affected by biases.

C Understanding Biases

To understand the issue of biases further in LLMs, there exist many studies that show how biases exist and affect LLMs' performance [17], [18], [19], [20], [21], [22], [23]. Studies have shown that recent LLMs are gender biased [19]. Biases happen when LLMs show a direct relationship between a specific demographic group and a biased term associated with that group [24]. For example, LLMs might associate men with "powerful" and "strong," while associating women with "weak" and "submissive". These biases usually occur based on human datasets [17]. There have been many documents and existing datasets where LLMs have been trained on, and this results in transferring social biases from humans to LLMs. These biases can cause harm to some social groups and in sensitive fields [19], [23]. For example, LLMs have been tested and growing in writing recommendation letters for individuals. In that case, models might associate that person with specific social bias, resulting in problems in writing that letter [23]. This example shows that the newer models do not show biases in their output, and these biases remain hidden as traces in models [23]. Models such as SEAT and WEAT exist for bias detection in the dataset. WEAT handles bias detection in the dataset for individual words, while SEAT finds the biases by evaluating the whole sentence [18]. WEAT and SEAT can also handle complicated bias detection in datasets. This includes examples, such as "Angry Black Women," in which the stereotype is that biases is related to only to black woman and is not generalized for women or black people [18]. Some datasets also detected models showing biases Bias Benchmark for Question Answering that detects explicit bias in models, and there are other extensions for it to capture the hidden biases, such as HBBQ [23]. Mitigation techniques for biases also include self-reflection and debiasing the models' answer [17], [18].

V Identifying Research Gap

A Mitigating Hallucinations Induced Biases

As we studied, there are many studies that focus on identifying and detecting hallucinations and biases separately. Other studies focused on mitigating hallucinations without focusing on the source of hallucinations. Hallucinations are complicated in LLMs, and our understanding of that phenomena is limited. However, finding the source is important as most of the previous studies did not focus on the source of hallucinations, while trying to mitigate them, which made those studies unreliable. Therefore, we used other studies that show that hallucinations and biases have direct relationships and showed that biases are an extreme source of hallucinations. To the best of our knowledge, the studies that focused on biases effects on hallucinations only focused on detection and direct correlation but discarded using any mitigation strategies. We identified our research gap as extending the work of previous studies through mitigating hallucinations induced biases. We focus our approach on reducing hallucinations that result from social biases in the pretraining data. We also focus on unfairness hallucinations to make sure that we are logically narrowing our research idea.

In conclusion, while Large Language Models showed capabilities across diverse and sensitive fields such as healthcare, education, and finance, they still show some flaws. Hallucinations, with many types, including factual, intrinsic, extrinsic, or bias-induced, are still a major challenge that affects the reliability of LLMs. Sources of hallucinations are another issue which can affect how hallucinations are addressed, detected, and mitigated. Discarding sources in ad-

dressing hallucinations can impact the faithfulness of studies. The funnel model gave us insights into finding biases as a hidden source in hallucinations, while guided us in finding the research gap on mitigating biased induced hallucinations. Our identified research gap only focuses on extending the work of other studies to arrive at a solution in improving Large Language Models.

References

[1] S. Minaee, T. Mikolov, N. Nikzad, et al., "Large language models: A survey," arXiv preprint arXiv:2402.06196, Feb. 2024, Available: https://arxiv.org/abs/2402.06196.

[2] L. Wang, J. Li, and Y. Zhang, "A comprehensive overview of large language models," arXiv preprint arXiv:2307.06435, 2023, Available: https://arxiv.org/pdf/2307.06435. [Accessed: 02-Apr-2025].

[3] T. B. et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023, Available: https://arxiv.org/pdf/2303.08774. [Accessed: 02-Apr-2025].

[4] K. Denecke, R. May, LLMHealthGroup, and O. R. Romero, "Potential of large language models in health care: Delphi study," Journal of Medical Internet Research, vol. 26, e52399, 2024. DOI: 10.2196/52399.

[5] Q. Liu, X. Chen, Y. Ding, S. Xu, S. Wu, and L. Wang, "Attention-guided self-reflection for zero-shot hallucination detection in large language models," arXiv preprint arXiv:2501.09997, Jan. 2025.

[6] P. Manakul, A. Liusie, and M. J. F. Gales, "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models," arXiv preprint arXiv:2303.08896, 2023, Available: http://arxiv.org/pdf/2303.08896.

[7] N. Varshney, W. Yao, H. Zhang, J. Chen, and D. Yu, "Detecting and mitigating hallucinations via self-critiquing," arXiv preprint arXiv:2307.03987, 2023, Available: https://arxiv.org/abs/2307.03987.

[8] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," arXiv preprint arXiv:2401.11817, Jan. 2024, Available: https://arxiv.org/pdf/2401.11817.

[9] Y. Fu, X. Wang, J. Chen, and M. Zhang, "On large language models' hallucination with regard to known facts," arXiv preprint arXiv:2403.20009, Mar. 2024, Available: https://arxiv.org/pdf/2403.20009v1.

[10] J. Li, Q. Dong, J. Liu, Z. Ma, and W. Zhang, "The dawn after the dark: An empirical study on factuality hallucination in large language models," in Proc. 62nd Annual Meeting of the Association for Computation Bangkok, Thailand, 2024, pp. 10 879–10 899.

[11] N. Varshney, W. Yao, H. Zhang, J. Chen, and D. Yu, "A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation," arXiv preprint arXiv:2307.03987, 2023, Available: https://arxiv.org/abs/2307.03987.

[12] A. Simhi, I. Itzhak, F. Barez, G. Stanovsky, and Y. Belinkov, "Trust me, i'm wrong: High-certainty hallucinations in llms," arXiv preprint arXiv:2502.12964, 2025, Available: https://arxiv.org/pdf/2502.12964.

[13] N. Varshney, W. Yao, H. Zhang, J. Chen, and D. Yu, "Detecting and mitigating hallucinations via self-critiquing," arXiv preprint arXiv:2307.03987, 2023, Available: https://arxiv.org/abs/2307.03987.

[14] D. Zhu, D. Chen, Q. Li, et al., "Pollmgraph: Unraveling hallucinations in large language models via state transition dynamics," arXiv preprint arXiv:2404.04722, Apr. 2024.

[15] X. Wu, J. Nian, Z. Tao, and Y. Fang, "Evaluating social biases in llm reasoning," arXiv, 2025, Available: https://arxiv.org/pdf/2502.15361. [Accessed: Mar. 30, 2025].

[16] J. Tang, Y. Tang, and J. Yao, "Exploring causal effect of social bias on faithfulness hallucinations in large language models," in Proceedings of the 2024 International Conference on Learning Representations (ICLR), Available: https://openreview.net/pdf?id=aOiUOohUBw, 2024.

[17] I. O. Gallegos, R. A. Rossi, J. Barrow, et al., "Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes," arXiv preprint arXiv:2402.01981, 2024, Available: https://arxiv.org/pdf/2402.01981.

[18] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, "On measuring social biases in sentence encoders," arXiv preprint arXiv:1903.10561, 2019.

[19] A. Scherrer and R. West, "Kelly is a warm person, joseph is a role model: Gender biases in llm-generated reference letters," arXiv preprint arXiv:2310.09219, 2023, Available: https://arxiv.org/pdf/2310.09219.

[20] W. Thong and C. G. M. Snoek, "Bias-aware classifier for generalized zero-shot learning," arXiv preprint arXiv:2008.11185, 2020, Available: https://arxiv.org/abs/2008.11185.

[21] D. Anderson and M. D. Smith, "Inherent bias in large language models: A random sampling approach," J. Digital Health, vol. 10, pp. 1–8, 2024.

[22] P. Anantaprayoon, M. Kaneko, and N. Okazaki, "Intent-aware self-correction for mitigating social biases in large language models," arXiv preprint arXiv:2503.06011, Mar. 2025, Available: https://arxiv.org/pdf/2503.06011.

[23] J. Pan, C. Raj, Z. Yao, and Z. Zhu, "Beneath the surface: How large language models reflect hidden bias," arXiv preprint arXiv:2502.19749, 2025, Available: https://arxiv.org/abs/2502.19749.