Can I trust You?
LLMs as conversational agents

Marc Döbler, Raghavendran Pownraju Mahendravarman, Anna Moskvina
Ella Lab – Gesellschaft für künstliche Intelligenz mbH
Cologne
Germany
{marc.doebler,raghavendran.mahendravarman,anna.moskvina}@ella-lab.io

Abstract

With the rising popularity of LLMs in the public sphere, they have become more and more attractive as a tool for doing one's own research without having to rely on search engines or specialized knowledge of a scientific field. But using LLMs as a source for factual information can lead one to fall prey to misinformation or hallucinations dreamed up by the model. In this paper we examine the gpt-4 LLM by simulating a large number of potential research queries and evaluate how many of the generated references are factually correct as well as existent.

1 Introduction

One of the main functions of the system of mass media, according to Niklas Luhmann, consists in constituting a sort of short-term memory for society by providing and processing information about the world (Luhmann, 1995). To fulfill this function and to perpetuate its own existence, the mass media constantly generates and communicates information. However, the truth of this information is not the most important factor, even when it comes to news. Truth is only relevant insofar as it emphasizes the sensational value of any given message and averts the risk of being accused of deception (Luhmann, 1995).

But this is in contrast to how we use news in our everyday lives where we rely on them for their accuracy.

With the advent of LLMs in the larger public sphere (Yang et al., 2023; Thirunavukarasu et al., 2023; Baidoo-Anu and Ansah, 2023; Dwivedi et al., 2023) , this tension between the individual's desire for factually correct information and the mass media's preference for mere communication grows more pronounced. When ordinary users utilize LLMs as more advanced internet search engines to answer questions that are not easily answered by sites like Wikipedia, they expect a truthful response. But the model's primary function lies in communication. It might be dissatisfying for users when an LLM hallucinates instead of providing correct answers to a query, but this is not an error on a technical level. A technical error would mean a complete failure to communicate (i.e. a blank output or an incomprehensible non-sequitur). To counteract this discrepancy, filters and fact-checking processes have been implemented, but these are additional mechanisms added on top of the base model to steer the result in the desired direction. The issue of communication being valued more highly than truth still remains. While this might not be a problem on a larger systemic level, it is an issue when individuals expect information to be factual. This is why we need to examine the quantitative and qualitative nature of false information produced by LLMs.

Hypothesis

Our hypothesis is that *LLMs are more focused on facilitating communication than factual accuracy*. This means that we expect the model to generate relevant-sounding answers in significantly more cases than it generates answers that are actually supported by the facts.

2 Related Work

Despite recent developments in artificial intelligence and the emergence of different LLMs such as PaLM (Chowdhery et al., 2023), OpenAI's ChatGPT (OpenAI, 2022), and GPT-4 (OpenAI, 2023), Google's Bard (Manyika, 2023), Meta's LLaMa (Touvron et al., 2023) all of the established LLMs are infamous for hallucinations (Dziri et al., 2022; Ji et al., 2023) . This fact prompted research to examine the ways of how we can facilitate the "made up" nature of some of the outputs of the models. As well as overall analysis of whether the LLMs can be trusted. Workshops such as *Workshop on*

Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024), pages 71–75

Large Language Models' Interpretability and Trustworthiness (Saha et al., 2023) or *TrustNLP: Workshop on Trustworthy Natural Language Processing* (Ovalle et al., 2023) nudge the scientific community to investigate explainability and trustworthiness of different predictive models and LLMs.

Most of the research on the trustworthiness of the LLMs is either based on surveys and analysis of different requirements (such as fairness, explainability, accountability, reliability to name a few) that can be used to asses the output of a model in general (Kaur et al., 2022; Liu et al., 2023) or in specific fields, for instance in Healthcare (Ahmad et al., 2023). Other studies cover general guidelines for establishing a trustworthy model (Litschko et al., 2023) or provide an overview of methods available to detect the fairness of the output of the model based on their toxicity, bias and value-alignment (Huang et al., 2023).

Though there are several studies that are focused on investigating how factually correct the output of the models is (Zhao et al., 2023; Min et al., 2023), the usual methods of testing factual correctness is to check the facts present in the output. Only very few researchers studied the references that were provided by a model for the generated output, i.e. checked the validity of the source that the output was based on (Shi et al., 2023).

This positional paper is designed to make it more apparent if we should trust the output of the LLMs just because they provide us with official looking sources.

3 Methodology

To test our hypothesis that Large Language Models are in their nature trained to be a conversational partner first and a "fact provider" second, we have conducted a series of experiments that were designed to examine how much LLMs can be trusted as a source of information. For the first part of the experiment we automatically generated scientific questions from different branches. Overall, 985 questions were generated for 231 scientific branches. The most questions were generated for the fields of Medicine (40), Computer Science (38), Environmental Science (37), Earth Science (35), Physics (35), Mathematics (35) and Chemistry (34).

The generated questions were then used as a query for the gpt-4 LLM with a task of further generating a corresponding concise explanation which was to be based on real scientific references (in form of a list of links). For example, the query "How do children acquire syntactic knowledge in their native language?" received the explanation "Children acquire syntactic knowledge in their native language through a combination of innate abilities and environmental input <. . . >" and a list of 4 links to different scientific journals. The query "How can we model complex systems with mathematics?" received the explanation "Complex systems can be modeled mathematically using various approaches, depending on the nature of the system and the phenomena being studied. Here are some of the common mathematical frameworks and methods used to model complex systems <...>" based on 10 different links.

Previous research has shown that mapping back to the original documents or providing sources to the generated texts can potentially invoke a feeling of trust from a user towards a model (Bohnet et al., 2022). Therefore the generated references were investigated for their relevance. In order to do that, we counted how many of the generated references actually exist at the moment of the experiment, and how many of the links provided the information that was relevant to the question.

The workflow of the experiment is given in Figure 1.

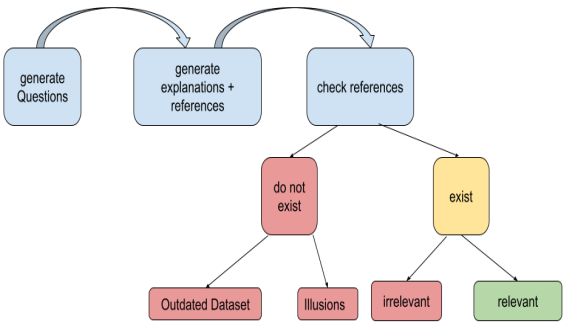

Figure 1: Automated Workflow for testing the hypothesis

LangChain (Harrison, 2022) was used for generating both questions and explanations.

Though our experiments resemble the study per-

formed by Shi et al., 2023, our approach focuses more on generating specific questions within the scientific domain to simulate the workflow and the research process of a human.

4 Evaluation

To further investigate whether our hypothesis holds we have conducted a quantitative (statistical) analysis of the results.

Overall, for 985 different questions we generated 985 explanations with 4434 references (with an average of 5 references per explanation).

Out of the 4434 references 2967 links led to pages that did not exist anymore at the time the experiments were conducted (middle of December 2023). One of the reasons could be that the submitted paper that existed at the time the LLM was trained, was retracted, deleted or moved to another website (so-called illusions (Shi et al., 2023)). Another reason could be that the link was hallucinated by the model and did not exist in the first place.

Out of the 1467 references that led to real existing scientific sources 1376 were relevant to the question. Table 1 gives a short overview of the findings.

We also examined the scientific branches that suffered most from the sources that were not fully available. Out of 40 questions and explanations for the field of Medicine, 37 had missing references. 37 out of 37 explanations for questions from the field of Environmental Science had incomplete references. 2 out of 2 questions and explanations from the area of Zymology had referenced non-existing sources. In general, only approximately 25% of explanations were covered by the references that were marked as existing at the moment of the experiment. After the relevance check 88% of explanations had not only verifiable but also related sources.

5 Discussion

Our findings show that a large number of the generated sources are relevant to the question that was asked, but do not exist. This is either because the model hallucinated these citations or because its data was outdated, thus providing a dead link.

It is not entirely clear if this supports our hypothesis. If the citations are indeed real, but have been deleted or retracted since the model was trained, this finding may only be a reflection of the model's outdated training data. But if the model did indeed hallucinate these sources, this finding would support our hypothesis that the model puts greater value on communication and providing answers that are superficially satisfactory than on factually correct information.

However, many of the generated sources were irrelevant to the query. These were generated by the model to fulfill the demand for sources in the query, but the sources provided were selected with little regard for relevancy. This supports our hypothesis that the model puts greater emphasis on communicating successfully than on responding to the query correctly.

6 Limitations

A limitation of this paper is that we only tested the gpt-4 LLM. Our findings might only be relevant to models similar to this one, but not for models that are very different from it or those that have a greater focus on factual accuracy.

We were also not able to clearly differentiate between sources that did not exist anymore because they were removed or retracted and those that were hallucinated entirely.

7 Ethical Considerations

As far as ethical considerations go, our findings illustrate that only in 88% of cases the cited sources were relevant as well as existent. This suggests that LLMs are not primarily concerned with providing accurate and up-to-date information. Individual users that seek to use LLMs as a tool for an in-depth net search that search engines can't provide should be very cautious to double check the information they receive.

8 Future work

We strongly believe that further investigation into the phenomenon of illusions, hallucinations and unrelated sources is needed. Understanding why the model outputs references that are not indeed relevant to the generated answers as well as a way of identifying types of illusions and hallucinations is crucial for building reliable, personalized and trustworthy LLMs.

9 Acknowledgment

This study was conducted within the framework of developing a commercial paraphrasing tool for German by employees of Ella Lab – Gesellschaft für künstliche Intelligenz mbH. Furthermore, we

Sources	Does not Exist	Exists irrelevant	Exists relevant
4434	2967	91	1376

Table 1: Overall number of sources given as references to the generated explanation and their distribution based on relevance

would like to acknowledge the guidance and provided by Nasrin Saef, Head of Machine Learning Team.

References

Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. 2023. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv preprint arXiv:2311.01463*.

David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Chase Harrison. 2022. *LangChain*.

Yue Huang, Qihui Zhang, Lichao Sun, et al. 2023. Trustgpt: A benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507*.

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.

Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. 2022. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, 55(2):1–38.

Robert Litschko, Max Müller-Eberstein, Rob Van Der Goot, Leon Weber, and Barbara Plank. 2023. Establishing trustworthiness: Rethinking tasks and model evaluation. *arXiv preprint arXiv:2310.05442*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Niklas Luhmann. 1995. *Die realität der massenmedien*. Springer.

James Manyika. 2023. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*, 2.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta. 2023. Proceedings of the 3rd workshop on trustworthy natural language processing (trustnlp 2023). In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*.

Tulika Saha, Debasis Ganguly, Sriparna Saha, and Prasenjit Mitra. 2023. Workshop on large language models' interpretability and trustworthiness (llmit). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5290–5293.

Xiang Shi, Jiawei Liu, Yinpeng Liu, Qikai Cheng, and Wei Lu. 2023. Know where to go: Make llm a relevant, responsible, and trustworthy searcher. *arXiv preprint arXiv:2310.12443*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language

models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Ruochen Zhao, Xingxuan Li, Yew Ken Chia, Bosheng Ding, and Lidong Bing. 2023. Can chatgpt-like generative models guarantee factual accuracy? on the mistakes of new generation search engines. *arXiv preprint arXiv:2304.11076*.