LLMSDI: OSS LLMs and decentralized systems for Search, Discovery and Indexing on the Internet

Aurora González-Vidal
Department of Information and Communication Engineering
University of Murcia
Murcia, 30100
aurora.gonzalez2@um.es

Mirko Presser
Digital Business Development
Aarhus University
Aarhus, Denmark
mirko.presser@btech.au.dk

***Abstract*—Large Language Models (LLMs) are at the forefront of a transformative era in internet information retrieval. Trained on extensive textual data, these Artificial Intelligence (AI) models exhibit remarkable capabilities in comprehending intricate queries, interpreting context, and generating human-like text across diverse scenarios. Their effectiveness in processing vast information has positioned them as indispensable tools for search engines, virtual assistants, and other retrieval systems, revolutionizing the search and discovery experience.**

This paper explores the challenges and solutions associated with LLMs and Knowledge Graphs (KGs). Computational demands, hallucination, common sense limitations, multilingual challenges, and the imperative need for auditability and explainability form the primary challenges addressed. Subsequently, the paper delves into a range of subtopics, including the democratization of legal information, combating fake news through LLMs and Natural Language Processing (NLP) mechanisms, sustainability in LLMs, hallucination detection and mitigation, decentralized and federated LLMs, application-tailored LLMs, innovative Knowledge Graph designs, probabilistic and logical thinking incorporation, transfer learning for LLMs, and cost-efficient training methodologies.

Through an in-depth examination of these challenges and subtopics, this paper aims to contribute to the evolving landscape of LLMs and KGs, providing insights into the current state of research and potential future directions.

***Index Terms*—Large Language Models, Knowledge Graphs, Artificial Intelligence, Search and Discovery, Indexing, Internet**

I. Introduction

The advent of Large Language Models (LLMs) has marked a revolutionary phase in the realm of internet information retrieval. These sophisticated AI models, having undergone extensive training on massive text datasets, have demonstrated unparalleled proficiency in understanding complex queries, contextual interpretation, and generation of human-like text. Their significance lies in their ability to process vast volumes of information, thereby reshaping the landscape of search engines, virtual assistants, and other information retrieval systems.

As we delve into the challenges and solutions surrounding LLMs and Knowledge Graphs (KGs), it becomes apparent that their potential is not without hurdles. Computational demands, susceptibility to hallucination, common sense limitations, multilingual intricacies, and the critical need for auditability and explainability pose substantial challenges to the seamless integration of LLMs into information retrieval systems.

To dissect these challenges further, this paper explores a spectrum of subtopics, ranging from the democratization of legal information to the incorporation of probabilistic and logical thinking into LLMs. Each subtopic sheds light on specific facets, providing a comprehensive understanding of the current landscape and paving the way for future developments.

Through this exploration, we aim to contribute valuable insights to the ongoing discourse on LLMs and KGs, fostering a deeper understanding of the challenges at hand and envisioning a more robust future for information retrieval systems.

II. State Of The Art

In recent times, several AI-based search and content-generating tools are gaining popularity among people and researchers. Large language models are ushering in a revolutionary era in the way we search for and discover information on the internet. These advanced AI-powered models have been trained on massive amounts of text data, allowing them to understand and generate human-like text in a wide range of contexts. It is the case of tools such as chatGPT, that interact with the user in a conversational way and is able to generate new content, having amazing potential as well as raising concerns about progress, critical analysis [3] and bias. GitHub copilot is another example of an AI-based content generator. It is a language model trained over open-source GitHub code that assists humans in coding, and given this characteristic, its commercial use may be questionable[1] as well as its skills on advance programming tasks [4] and the repetition of bugs [20]. AI-based writing tools such as Rytr, Copy AI, NeuroFlash, etc., AI-based art generators: Jasper Art, Stable Diffusion are other examples.

LLMs tend to generate non-factual and fab- ricated content [24], [27] that is not grounded in reality or supported by factual evidence, which is usually referred to “hallucination” and makes the generated text not trustworthy. Recently, a new generation paradigm, Verifiable Generation [6], [7]. li2024

Previous work on verifiable information revolves around two key aspects:

Identify applicable funding agency here. If none, delete this.

[1] https://www.nytimes.com/2022/11/23/technology/copilot-microsoft-ai-lawsuit.html

- Modeling: Nakano et al. [19], Menick et al. [17], Liu et al. [16], and Qin et al. [7] explore training large language models (LLMs) to browse web pages and answer long-form questions with supporting evidence. Gao et al. [6] propose research-and-revision to retrieve evidence for LLM's output, fixing unsupported content. Gao et al. [8] utilize demonstrations to enable LLMs to generate text with citations of supporting documents. Li et al. [13] extend the evidence source to knowledge graphs, proposing the retrieval of entity sub-graphs as supporting evidence.
- Verifiability Evaluation: Rashkin et al. [23] introduce the Attributable to Identified Sources (AIS) framework for human evaluation, assessing if the answer can be supported by evidence. Gao et al. [5] define auto-AIS, using a strong natural language inference model [9] to approximate human AIS judgments. Liu et al. [15] propose citation-recall and citation-precision for evaluating generative search engines. Gao et al. [8] propose a framework automating recall and precision evaluation for citations. Min et al. [18] introduce FactScore for automatic evaluation of verifiability in generated text.

Works such as LLatrieval [14] can retrieve supporting documents for verifiable generation, which lets the LLM iteratively refine the retrieval result until the LLM verifies the retrieved documents can support answering the given question

Another problem that LLM have is that they sometimes lack common sense. To this regard, for enhancing LLMs' performance, [2] proposed a novel approach: prompting the model with a few input–output exemplars that demonstrate the task. This method, proven effective for various simple question-answering tasks, was further extended by [26], who introduced concrete chain-of-thought prompting. This approach involves providing LLMs with input–output examples along with explicit reasoning steps, demonstrating notable improvements in arithmetic reasoning, common-sense reasoning, and symbolic reasoning tasks.

In the pursuit of bolstering LLMs' zero-shot performance, [11] introduced zero-shot-CoT prompting, which involves enhancing the input prompt with the sentence ”Let's think step by step.” Despite significant achievements on various tasks, [25] found that LLMs still struggle with simple decision-making tasks, such as the blocksworld, which requires arranging a set of given blocks in a particular order.

III. CORE

A. Verif.ai: Towards an Open-Source Scientific Generative Question-Answering System with Referenced and Verifiable Answers [12]

The Verif.ai project aims to develop an open-source scientific generative question-answering system with referenced and verified answers. The system consists of three main components: an information retrieval system utilizing semantic and lexical search techniques on scientific papers (PubMed), a fine-tuned generative model (Mistral 7B) that generates answers with references to the source papers, and a verification engine that checks the generated claims against the abstracts or papers from which they are derived. The goal is to mitigate the risk of hallucinations, or the inadvertent generation of false or misleading information, in scientific generative language models. The project emphasizes user trust and productivity in scientific environments, where misinformation is intolerable.

The methodology involves a toolbox with an information retrieval engine based on OpenSearch, fine-tuning the Mistral 7B model using PubMedQA dataset, and a verification engine using XLM-RoBERTa and DeBERTa models to check for hallucinations. The user interface allows users to ask questions, review answers, and provide feedback for continuous improvement. Preliminary evaluations include qualitative assessments of information retrieval methods and the fine-tuned Mistral 7B model's performance, as well as the evaluation of the verification models on the SciFact dataset. The project acknowledges challenges, including the evolving landscape of large language models, and aims to incorporate user feedback and improve hallucination detection methods.

B. Science Checker Reloaded: A Bidirectional Paradigm for Transparency and Logical Reasoning [22]

The authors propose a two-block approach to enhance language understanding in sparse retrieval and provide comprehensive answers to complex questions in long documents. The first block focuses on query expansion using an ontology-oriented approach to improve relevance in sparse retrieval. The second block involves a hybrid search system, combining sparse and dense retrieval methods, to deepen answers by extracting relevant passages. The authors emphasize transparency, logical reasoning, and comprehensive understanding in scientific information retrieval through bidirectional engagement with users.

The authors discuss the limitations of existing methods, such as semantic divergence, vocabulary gaps, and issues with large language models (LLMs). They propose a system that prunes inefficient processes, aiming for an optimal trade-off between performance and cost. The paper presents an evaluation of the document retrieval block, demonstrating the system's performance against various benchmarks. While the system shows promising results, the authors acknowledge the need for additional evaluations, including user tests, to assess the overall system's usability, transparency, and performance in scientific and technical domains.

C. Building Authentic AI: The OriginTrail Decentralized Knowledge Graph and Knowledge Mining [21]

The OriginTrail Decentralized Knowledge Graph (DKG) and the ChatDKG framework represent a groundbreaking development in the realm of trusted AI frameworks. OriginTrail's DKG focuses on ensuring discoverability, verifiability, and value of digital and physical assets. The introduction of ChatDKG goes beyond typical large language model (LLM)-based chatbots by incorporating both extractive and generative methods, drawing from the rich and verified Knowledge

Assets in the DKG. This approach significantly improves the sourcing, evaluation, and application of trusted knowledge in AI applications. By guaranteeing data origin and authenticity, DKG and ChatDKG contribute to the creation of more reliable and ethical AI systems, emphasizing the critical role of data integrity in the era of advanced technology.

To establish a foundation for trustworthy AI systems, OriginTrail's DKG merges blockchains and knowledge graphs, creating a neutral, open-source infrastructure for discovering and verifying information. This integration forms a reliable knowledge base, aiding AI applications in discovering information across the DKG, verifying data provenance, and assigning ownership to Knowledge Assets. The ChatDKG framework, with its retrieval-augmented generation, enhances the synergy between DKG and AI. Through crowdsourcing efforts and global collaboration, ChatDKG aims to cultivate a trusted knowledge base. OriginTrail further introduces knowledge mining, a crowdsourcing method, to build the world's largest public knowledge base. The V8 Foundation represents a future phase, emphasizing autonomous knowledge mining with AI, positioning OriginTrail as a dynamic platform for ethical and reliable AI development.

D. Generative-AI and Elections: Are Chatbots a Reliable Source of Information for Voters? [1]

AI Forensics is a European non-profit organization that is investigating influential algorithms, conducted a study focusing on Bing Chat, now called Copilot. This study aimed to assess the accuracy of Bing Chat's election-related answers, evaluate political imbalance, and measure the impact of the queries' language during the Swiss federal election in October 2023 and the Hesse and Bavaria federal states' elections in Germany in the same month. The study, conducted in collaboration with AlgorithmWatch, employed a comprehensive methodology involving prompt generation, data collection through web scraping, data preparation, labeling by experts, and qualitative analysis. The findings revealed a concerning 30% of answers containing factual errors, indicating structural issues in the chatbot's reliability despite being based on a large language model and deployed on every Windows machine.

The study's methodology included prompt generation through workshops with experts, data collection using web scraping and multiple web browser instances, and labeling by annotators from AI Forensics and AlgorithmWatch. The labeled data highlighted a substantial percentage of factual errors in Bing Chat's responses. The analysis also indicated language-specific variations, with English providing the most accurate results, French producing the most evasions, and German showing more factual errors. The presentation concluded with plans for future work, including improving data collection infrastructure, extending the study to other AI models like ChatGPT and Gemini, and addressing polarized content in Copilot's suggestions. Overall, the study raises concerns about the reliability of generative AI, particularly chatbots, in delivering accurate information during critical events like elections.

IV. Challenges

With regards to logical reasoning, there are several challenges on hallucination and verification [10].

- Challenges in Long-form Text Generation As the length of generated content increases, so does the likelihood of hallucinations and the existing benchmark tend to analyse factoid questions, neglecting long-form text generation. The evaluation metrics are limited and struggle with nuanced, open-ended, and debatable content, hindering real-world applications.
- Challenges in Large Vision-Language Models (LVLMs) LVLMs can generate inconsistent responses related to visual elements, including non-existent objects and incorrect relationships. They are prone to severe performance drops due to over-reliance on strong language priors and vulnerability to inappropriate inputs
- Challenges in Retrieval Augmented Generation (RAG). RAG introduces the risk of propagating irrelevant evidence into the generation phase and occasionally suffers from citation inaccuracies, impacting the traceability of information. Balancing diversity and factuality in RAG poses a challenge, influencing the reliability of responses.

And also, several open questions:

- The effectiveness of LLMs' self-correction mechanisms in mitigating reasoning hallucinations remains uncertain
- Capturing and understanding LLMs' knowledge boundaries, especially in recognizing their own limits, requires ongoing research.
- Striking a balance between creativity and factuality in LLMs, particularly in multi-modal and vision generation tasks, is an unresolved challenge.

References

[1] Generative ai and elections: Are chatbots a reliable source of information for voters? In *INTERNET 2024, The Sixteenth International Conference on Evolving Internet*. Athens, March 10-14 2024.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[3] W. Castillo-Gonzalez. Chatgpt y el futuro de la comunicación científica. *Metaverse Basic and Applied Research*, 1:8–8, 2022.

[4] A. M. Dakhel, V. Majdinasab, A. Nikanjam, F. Khomh, M. C. Desmarais, and Z. Ming. Github copilot ai pair programmer: Asset or liability? *arXiv preprint arXiv:2206.15331*, 2022.

[5] L. Gao, X. Dai, P. Pasupat, and et al. Rarr: researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, page 16477–16508. Association for Computational Linguistics, 2023.

[6] L. Gao, Z. Dai, P. Pasupat, A. Chen, A. T. Chaganty, Y. Fan, and et al. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, July 2023.

[7] T. Gao, H. Yen, J. Yu, and D. Chen. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*, 2023.

[8] T. Gao, H. Yen, J. Yu, and D. Chen. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*, 2023.

[9] O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, and et al. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*, 2022.

[10] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, and T. Liu. A survey on hallucination in large language models: Principles. taxonomy, challenges, and open questions. 2023.

[11] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems 35*, page 22199–22213, 2022.

[12] M. Košprdić, A. Ljajić, B. Bašaragin, D. Medvecki, and N. Milošević. Verif.ai: Towards an open-source scientific generative question-answering system with referenced and verifiable answers. In *INTERNET 2024, The Sixteenth International Conference on Evolving Internet*. Athens, March 10-14 2024.

[13] X. Li, Y. Cao, L. Pan, Y. Ma, and A. Sun. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. *arXiv preprint arXiv:2310.05634*, 2023.

[14] X. Li, C. Zhu, L. Li, Z. Yin, T. Sun, and X. Qiu. Llatrieval: Llm-verified retrieval for verifiable generation. 2023. arXiv:2311.07838.

[15] N. F. Liu, T. Zhang, and P. Liang. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*, 2023.

[16] X. Liu, H. Lai, H. Yu, and et al. Webglm: Towards an efficient web-enhanced question answering system with human preferences. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, page 4549–4560. ACM, 2023.

[17] J. Menick, M. Trebacz, V. Mikulik, and et al. Teaching language models to support answers with verified quotes. *CoRR*, 2022. arXiv:2203.11147.

[18] S. Min, K. Krishna, X. Lyu, M. Lewis, W. T. Yih, P. W. Koh, and et al. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. 2023. arXiv:2305.14251.

[19] R. Nakano, J. Hilton, S. Balaji, and et al. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, 2021. arXiv:2112.09332.

[20] H. Pearce, B. Ahmad, B. Tan, B. Dolan-Gavitt, and R. Karri. An empirical cybersecurity evaluation of github copilot's code contributions. *arXiv e-prints*, 2021. arXiv:2108.

[21] B. Rakić, T. Levak, and Ž. Drev. The origintrail decentralized knowledge graph and knowledge mining. In *INTERNET 2024, The Sixteenth International Conference on Evolving Internet*. Athens, March 10-14 2024.

[22] L. Rakotoson, S. Massip, and F.A.A. Laleye. Science checker reloaded: A bidirectional paradigm for transparency and logical reasoning. In *INTERNET 2024, The Sixteenth International Conference on Evolving Internet*. Athens, March 10-14 2024.

[23] H. Rashkin, V. Nikolaev, M. Lamm, and et al. Measuring attribution in natural language generation models. *CoRR*, 2021. arXiv:2112.12870.

[24] V. Rawte, A. Sheth, and A. Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

[25] K. Valmeekam, M. Marquez, A. Olmo, S. Sreedharan, and S. Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. 2023.

[26] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, and et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems 35*, page 24824–24837, 2022.

[27] Y. Zhang and et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

Acknowledgment

The NGI Search project is funded by the European Union under the Horizon Europe Programme, grant agreement 101069364. This was also supported by the ThinkInAzul program by MCIN with funding from European Union NextGenerationEU (PRTR-C17.I1) by Comunidad Autónoma de la Región de Murcia - Fundación Séneca as well as by the HORIZON-MSCA-2021-SE-01-01 project Cloudstars (g.a. 101086248).