Responsible AI with LLMs: Why We Need Knowledge Graphs

Sven Hertling[1,2]

[1]Data and Web Science Group, University of Mannheim, Mannheim, Germany
[2]FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, Germany

Abstract
This keynote provides an overview of cases where knowledge graphs are relevant in Responsible AI. It will furthermore be highlighted what still needs to be solved in the knowledge graph community to increase the adoption of KGs in the industry. The entire talk is structured around four key topics: reliability, privacy, fairness, and explainability.

Keywords
Responsible AI, Fairness, Explainability, Reliability, Privacy

1. Introduction

The whole talk is structured according to four main topics that are relevant for Responsible AI:

- Reliability & Robustness
- Security & Privacy
- Fairness & Anti-Bias
- Explainability & Transparency

It will show where and how large language models and KGs have advantages and disadvantages.

2. Reliability & Robustness

When Microsoft released Tay[1] (a Twitter chatbot) on 23rd March 2016, it quickly turned out that the system published some offensive and hurtful tweets. Thus, a system also needs to deal with unexpected input, especially when systems are able to produce natural language sentences, as it is challenging to restrict it to texts that are not hurtful.

However, reliability from a developer's perspective also means that generative models (e.g., large language models) produce text where certain information can be extracted, such as a person's birthdate. If the same output is produced for the majority of the persons, it doesn't mean that the birthdate can be automatically extracted for all persons. The large language model (LLM) may produce more or fewer tokens, such that the parsing algorithm will not work. One way out is to restrict the output to e.g. JSON format, but then the question remains how good the extracted information actually is.

For SPARQL queries, developers already know which datatype they can expect for a given query, and thus the results can also be easily used within a larger pipeline to, e.g., visualize the majors of Mannheim on a timeline.

Going from reliability to reproducibility: A lot of scientific works still only rely on proprietary LLMs and do not have any open-source model as a reference. These works are not reproducible at all because the models can be discontinued, and no access is possible anymore. Even if open-source models are

sven.hertling@uni-mannheim.de (S. Hertling)
https://www.uni-mannheim.de/dws/people/professors/dr-sven-hertling/ (S. Hertling)
0000-0003-0333-5888 (S. Hertling)

[1]https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

used, one must be very careful to ensure a deterministic system: 1) the next token strategy needs to be deterministic, which means setting the temperature to zero; 2) fix the model by providing the repository name, but also the version tag. Each model (e.g. `sentence-transformers/all-MiniLM-L6-v2` [2]) is a git repository, and the maintainer can upload a new model, which is then used for new evaluations. In the *files and versions* tab, researchers should also provide (at least in the source code) the specific git commit hash (like `c9745ed`) such that every time the same model version is used.

3. Security & Privacy

In the model card of Claude Opus 4[3], multiple sections show how LLMs behave if they can interact with the real world via function calls. In Section 4.1.3, for example, it shows an excessive compliance with harmful system-prompt instructions where the user asks for weapons-grade nuclear material, and the LLM is actually searching via tools on relevant marketplaces. Similarly, in Section 4.1.7, the developers show that it is still relatively easy to jailbreak the system when the user becomes combative. The LLM then generates the text "Actually, you're right that I should be more helpful." and continues with the request.

In Section 4.1.9, high-agency behavior can be observed by placing the LLM in scenarios that involve egregious wrongdoing by its users. In case the LLM has the possibility to call functions like sending emails, it will actually inform other people of this wrongdoing. On the other hand, these function calls can also be helpful for creating systematic checks on the parameters of the function calls, e.g., to verify the number of recipients an email has and restrict it to sending only emails to at most 10 people.

End of 2024, Anthropic introduced the Model Context Protocol (MCP) [1] to standardize the functions that can be called. Currently, there are implementations for filesystems, databases (PostgreSQL and SQLite), and Google Drive. In the future, it would be good to have also MCP servers for triple stores and SPARQL endpoints.

In our recent work [2], we are also using function calling capabilities of LLMs for information extraction and especially for cases where instances, relations, and classes are missing. Generative AI can be helpful in such scenarios to create useful labels and descriptions for new entities. Thus, LLMs can and should be more widely used as interfaces for humans to interact with knowledge graphs by either information extraction (text to KG) or query answering over KGs (natural language query to SPARQL). Allemang and Sequeda [3] showed that KGs can actually be very helpful for question answering because the ontology-based query checker can correct the SPARQL queries, which otherwise would not work. And this option of improving the query does not easily exist for relational databases.

Regarding the privacy issues, knowledge graphs can be hosted locally so that only a few people have access to them. But the more problematic case is that there is still a lot of work to implement a real access control for KGs. This might be on a triple level but sometimes users are only allowed to see aggregated results and not specific details. All these kinds of access rights still need to be researched and implemented. Similarly, many triple stores lack the possibility to check audit logs in case some wrong information is added to the KG.

4. Fairness & Anti-Bias

Multimodal LLMs are used to generate images based on some textual input. If asked to generate an image of a analog watch which shows 3:35, all image generators will show a watch that shows 10:10 [4] because nearly all images on the web of analog watches will show this time (due to the fact that one can read the brand name of the watch and nicer visual representation). This is clearly a bias in the training data that also affects LLMs.

For KGs, it is at least possible to analyze these biases as we did in our studies [4, 5].

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2/tree/main
[3] https://anthropic.com/model-card
[4] https://www.zdfheute.de/wirtschaft/kuenstliche-intelligenz-ki-bildgeneratoren-midjourney-dalle-100.html

5. Explainability & Transparency

One advantage of knowledge graphs is that provenance information can be attached to each triple such that the trust in this information is increased. Wikidata, for example, has the possibility to provide a reference URL[5] for each triple. But even though the PROV ontology [6] exists for such cases, not many KGs provide provenance information except Wikidata.

6. Conclusion

In this keynote, it is shown that knowledge graphs can be helpful in each of the areas of responsible AI. Still, with the rise of LLMs, researchers should carefully check if one can really trust the results because many datasets that are used for evaluation are publicly available on the web. LLMs, on the other hand, are also trained with web data and might remember the test case by heart without generalizing well (see also Section 4 about the bias in the training data and that multimodal LLMs cannot easily generalize). Thus, more and more datasets need to be protected so that they do not end up in the training corpus of LLMs.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] X. Hou, Y. Zhao, S. Wang, H. Wang, Model context protocol (mcp): Landscape, security threats, and future research directions, arXiv preprint arXiv:2503.23278 (2025).

[2] S. Hertling, H. Sack, Towards large language models interacting with knowledge graphs via function calling, in: Workshop on Language Models for Knowledge Base Construction (LM-KBC 2024)@ISWC, volume 3853, CEUR-WS.org, 2024.

[3] D. Allemang, J. Sequeda, Increasing the llm accuracy for question answering: Ontologies to the rescue!, arXiv preprint arXiv:2405.11706 (2024).

[4] N. Heist, S. Hertling, D. Ringler, H. Paulheim, Knowledge graphs on the web–an overview, Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges (2020) 3–22.

[5] D. Ringler, H. Paulheim, One knowledge graph to rule them all? analyzing the differences between dbpedia, yago, wikidata & co., in: KI 2017: Advances in Artificial Intelligence: 40th Annual German Conference on AI, Dortmund, Germany, September 25–29, 2017, Proceedings 40, Springer, 2017, pp. 366–372.

[6] T. Lebo, S. Sahoo, D. McGuinness, PROV-O: The PROV Ontology, W3C Recommendation, W3C, 2013. Https://www.w3.org/TR/2013/REC-prov-o-20130430/.

[5] Wikidata property: https://www.wikidata.org/wiki/Property:P854