LLMs Meet Knowledge Graphs For A More Transparent Conversational AI: Wien Energie Chatbot

Umutcan Serles*
University of Innsbruck
Technikerstrasse 21a 6020 Innsbruck Austria
umutcan.serles@sti2.at

Ioan Toma
Onlim GmbH
Weintraubengasse 22 1020 Vienna Austria
ioan.toma@onlim.com

Abstract

This paper describes an industrial use case and a work in progress to enhance an LLM chatbot used by Wien Energie for customer service in terms of transparency and trustworthiness. We foresee that this will be achieved by a solution that can use both LLMs and knowledge graphs to produce answers.

1 Introduction

Large Language Model (LLM)-based conversational agents significantly simplified question answering in various topics. With state-of-the-art approaches like Retrieval Augmented Generation (RAG) [6], even in domain-specific scenarios, question answering over text can be provided. Knowledge graphs are large semantic networks that can integrate heterogeneous data sources and can provide explicit knowledge [4]. Both LLMs and knowledge graphs have their advantages and disadvantages [10]. For example, knowledge graphs provide explicit and well-integrated knowledge; however, they can be incomplete. LLMs are susceptible to well-established issues like hallucinations but also suffer from the drawbacks of sub-symbolic AI applications like lack of transparency and consequently trustworthiness. These issues have a negative impact on industrial applications as companies who provide their services over conversational AI need to ensure a certain level of quality and trust for the customer. Incomplete knowledge or hallucinated and inexplainable answers may harm the relationship with the customer and can even have financial and legal consequences.

Wien Energie[2] is an Austrian utility supplier and offers various services such as electricity, gas, heating and internet. Its customers can consume online content about Wien Energie services and take certain actions through a chatbot powered by Onlim technology.

In this paper, we present our work in progress for improving the transparency and trustworthiness of Wien Energie Chatbot. Our core hypothesis is that combining LLMs and knowledge graphs can achieve a better customer experience in terms of transparency and trustworthiness. Unlike traditional approaches that try to fuse knowledge graphs into LLMs at different stages (e.g., training or verification), we want to combine the strength of both approaches by deciding which questions can be answered from which source (e.g., LLM/documents, knowledge graphs) and providing suitable explanations.

The remainder of the paper is structured as follows: Section 2 describes the current implementation of the Wien Energie Chatbot, Section 3 presents the future implementation. Section 4 makes a brief literature review of approaches that augment LLM results with knowledge graphs. Finally, Section 5 provides a summary and the next steps.

*Corresponding author. Also affiliated with Onlim GmbH
[2]https://wienenergie.at

The First Austrian Symposium on AI, Robotics, and Vision (AIROV24).

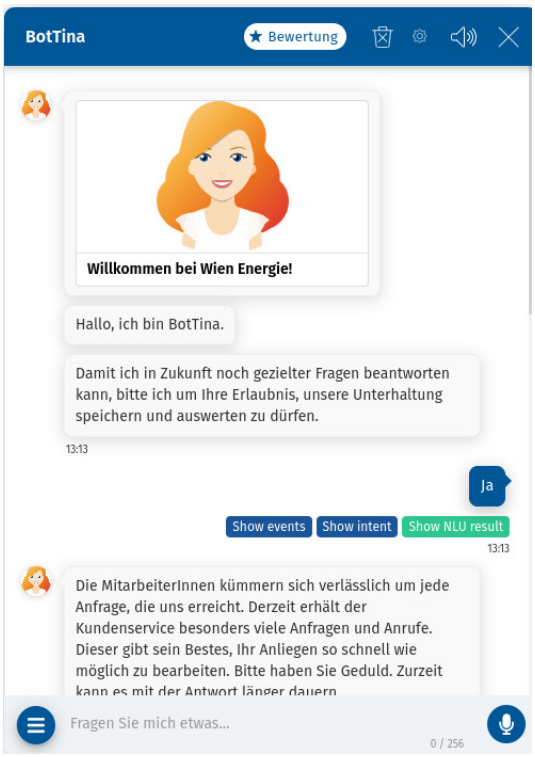

Figure 1: BotTina - Wien Energie chatbot

2 Wien Energie Chatbot

The Wien Energie chatbot, called BotTina, is a chatbot that provides detailed information about the services and products offered by Wien Energie as well as actions or processes associated with them. In addition, the chatbot can also answer generic FAQs about the company. The chatbot has been in operation since 2017 and has answered more than 2000000 customer inquiries. Typical questions handled by the chatbot are related to utility registration and cancellation, energy tariffs, invoices and payments, etc. Figure 1 shows the chat widget, the Wien Energie customers can use to interact with the chatbot BotTina.

The current version of BotTina is powered by a mix of technological solutions. Questions related to the type, contact point, opening hours, URLs, and basic description of the Wien Energie products and services are answered with knowledge from the Wien Energie Knowledge Graph. Other information that is currently not modeled as part of the Knowledge Graph related to these services and products, such as pricing information, technical details, contract conditions, etc. are provided as unstructured data i.e., documents and being answered following a classical RAG (Retrieval Augmented Generation) approach using LLMs. Last but not least there are generic FAQs, related to the company, small-talk, etc. that are implemented as static intents. In case of questions that cannot be answered by the chatbot, there is a live-chat functionality available in the platform that the human administrators of the chatbot can activate to take over conversations.

3 Wien Energie Chatbot in the Future

Currently, BotTina can answer very well questions that fall strictly in one of the three topic areas i.e. (i) product and services (using Knowledge Graph), (ii) pricing and technical details (using RAG with Document Stores and LLMs) and (iii) generic FAQs (using static intents), there are still more complex questions that can not be answered properly at the moment. These are questions for which combining knowledge from the three systems (Knowledge Graph, Document Store/LLMs, and static intents/answers) is required. Our approach is to use LLMs to decide and build a query plan i.e. decide which of the 3 systems mentioned above to be used and when, for which part of the user question,

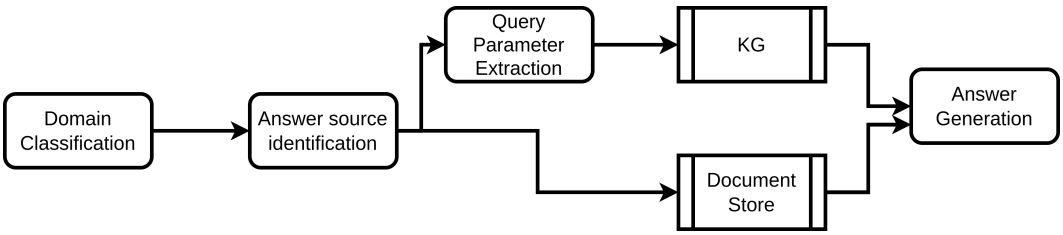

Figure 2: A high-level overview of the core of RAG/KG approach

how to query the three systems, and last but not least how to combine the partial results and build the final answer. We call this approach RAG/KG. The approach is being implemented as a chain of prompts using LLM agents.

Figure 2 shows a high-level overview of the RAG/KG approach. The first step is domain classification. Here the question is classified to (sub)domains supported by the chatbot. The available subdomains correspond to major types supported by the knowledge graph and collected from a set of SHACL shapes. This classification will be done via a zero- or few-shot LLM prompt. Afterward, the answer source of the question is identified. This step uses the schema information corresponding to the identified domain (e.g., target classes and properties supported by the SHACL shape for that domain). If the LLM agent thinks the question can be answered via the knowledge graph based on the given schema, then the question is redirected to the knowledge graph after query parameters recognized. If the LLM agent thinks the schema does not support the question, then it redirects it to the Document Store, where the documents are stored as vectors for RAG. These documents are typically PDF files, FAQs or other unstructured content from the Wien Energie website that is not semantically annotated. Note that in some cases, the LLM agent for answer source identification may decide that the question should be answered from both knowledge graph and document store. In this case, the question is split in multiple parts and each part is sent to the suitable answer source. In the final step, an answer is generated by another LLM agent using the context provided by the knowledge graph, document store or both.

With this approach not only the decision flow can be easily tracked, explained, and understood, but also how the three systems are being queried and how they provide results (knowledge graphs are known for being white box approaches). In this way, we provide explanations for the inner workings of our chatbots and increase their overall transparency and trustworthiness. Take the following question as an example: "Can you tell me about the fiber gas offers and how much they cost?". In this case, our system would first identify that the question asks about two things: (a) name and description of gas offers and (b) prices of those offers. Our query planner would identify that the first part (a) of the question is something that can be answered by the knowledge graph and the second part (b) from the document store. After an answer is provided, an explanation similar to Listing 1 will be generated.

```
The general descriptions of the gas offers are generated from the
    following entities in the knowledge graph: <dereferable uris
    of the entites>. The price information is generated by an LLM
    based on the following documents <uris of the documents>. Note
     that the information generated by an LLM may not be always
    accurate.
```

Listing 1: An indicative explanation example for the future Wien Energie Bot

4 Related Work

Classical RAG approaches help with providing domain-specific knowledge significantly, i.e. context-specific information, in comparison to using pre-trained generative large language models like GPT-3.5 out of the box [5]. There are still major drawbacks, for example, the lack of integrated knowledge, which makes it challenging to answer questions if the information is spread across different documents or pieces of documents. Therefore, the research on combining LLM answers with knowledge graphs is growing rapidly as evident from many recent surveys [1, 10]. The main

goal of many of these approaches is to improve the correctness of the answers, providing the model with encoded knowledge graphs in some form [8] e.g. as part of the prompt [2] or verifying the answers via a knowledge graph [9].

A common feature of these existing approaches is that they assume an existing knowledge graph and use it for enhancing LLM answers in different ways. However, they do not really discover the possibility of using knowledge graphs directly to answer questions and adopt LLMs only for natural language understanding and generation. In many cases, answering a question directly with a knowledge graph may be more efficient than trying to improve the answer the LLM provides before, during, or after answer generation. They also do not explore answering complex questions that may be suitable for both LLMs and knowledge graphs.

Transparency of sub-symbolic AI is an important and interesting research topic [7]. However, in the context of LLMs, the research is still being established. Current research mostly focuses on making LLMs explain their answers (cf.[3]) with the help of a knowledge base. In terms of transparency, an interesting approach is MindMap [11], which uses LLMs to show the thought process of answering a question utilizing a knowledge graph. The approach is quite promising in terms of contributing to verbalizing the provenance of answers in a user-friendly way. We will study this approach in more detail to see if we can adopt it to some extent in our solution.

5 Conclusion and Future Work

In this paper, we presented our vision and solution in progress for creating chatbots that benefit from both LLMs and knowledge graphs. The idea is not something new, in fact, it is one of the most researched areas. However, the current implementations try to enhance or verify LLM answers with knowledge graphs. We see these approaches as a bit problematic. Training an LLM with knowledge graphs (either during pre-training or as fine-tuning) may improve the answers of an LLM in domain-specific settings, however does not help with the transparency and also increases the risk of hallucination even for the questions that could have been answered by a knowledge graph. Using knowledge graphs as a way of verifying LLMs is just unnecessary work: if you already have the correct knowledge in the knowledge graph, why not directly answer it from there? We foresee that the key to combining LLMs with knowledge graphs is to identify the right source for the right question (or the parts of a question) and provide suitable explanations for the answer. In this paper, we presented our vision for this hypothesis, and in future work, we will test it with our RAG/KG implementation.

Acknowledgments and Disclosure of Funding

This work is partially supported by the following projects: FAIR-AI (FFG Projektnummer: FO999904624), Fortsetzung WisNat (FFG Projektnummer: 897468) co-funded by the Forschungs- und Förderungsgesellschaft (FFG) and PERKS (Project: 101120323) co-funded by the European Union.

References

[1] Agrawal, G., Kumarage, T., Alghami, Z., and Liu, H. (2023). Can knowledge graphs reduce hallucinations in llms?: A survey. *arXiv e-prints*, pages arXiv–2311.

[2] Baek, J., Aji, A. F., and Saffari, A. (2023). Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *ACL 2023 Workshop on Matching Entities*.

[3] Chen, Z., Chen, J., Gaidhani, M., Singh, A., and Sra, M. (2023). Xplainllm: A qa explanation dataset for understanding llm decision-making. *arXiv*, 2311.08614.

[4] Fensel, D., Simsek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., and Wahler, A. (2020). *Knowledge graphs*. Springer.

[5] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

[6] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

[7] Liao, Q. V. and Vaughan, J. W. (2023). Ai transparency in the age of llms: A human-centered research roadmap. *arXiv*, 2306.01941.

[8] Moiseev, F., Dong, Z., Alfonseca, E., and Jaggi, M. (2022). SKILL: Structured knowledge infusion for large language models. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1581–1588, Seattle, United States. Association for Computational Linguistics.

[9] Mountantonakis, M. and Tzitzikas, Y. (2023). Real-time validation of chatgpt facts using rdf knowledge graphs. *ISWC Demo Paper*. https://hozo.jp/ISWC2023_PD-Industry-proc/ISWC2023_paper_408.pdf.

[10] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., and Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20.

[11] Wen, Y., Wang, Z., and Sun, J. (2023). Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models.