Towards Explainability in Retrieval-Augmented LLMs

Joel Rorseth
University of Waterloo
jerorset@uwaterloo.ca

Parke Godfrey
York University
godfrey@yorku.ca

Lukasz Golab
University of Waterloo
lgolab@uwaterloo.ca

Divesh Srivastava
AT&T Chief Data Office
divesh@research.att.com

Jaroslaw Szlichta
York University
szlichta@yorku.ca

2024 IEEE 40th International Conference on Data Engineering (ICDE) | 979-8-3503-1715-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDE60146.2024.00466

***Abstract*—In an era where artificial intelligence (AI) is reshaping countless aspects of society, we present a forward-looking perspective for enhancing the explainability of large language models (LLMs), with a particular focus on the retrieval-augmented generation (RAG) prompting technique. We motivate the urgency for developing techniques to explain LLM decision-making behaviour, especially as these models are deployed in critical sectors. Central to this effort is RAGE, our novel explainability tool that can trace the provenance of an LLM's answer back to external knowledge sources provided via RAG. RAGE builds upon established explainability techniques to recover citations for LLM answers, identify context biases, and mine answer rules. Through our novel explainability formulations and practical use cases, we chart a course toward more transparent and trustworthy AI technologies.**

I. Question

In the rapidly evolving field of artificial intelligence (AI), large language models (LLMs) such as OpenAI's ChatGPT are growing increasingly capable, as evidenced by their groundbreaking proficiency across various tasks. However, LLMs continue to exhibit significant issues, such as the hallucination of plausible yet incorrect information. Despite these potential concerns, LLM adoption is increasingly widespread, extending into critical sectors where issues may carry significant consequences. As LLMs are adopted for high-stakes applications (e.g., healthcare, safety, employment), explanations of their outputs are paramount to ensure trustworthiness. Although LLMs can be prompted to explain their answers (e.g., using *chain-of-thought* prompting [1]), these self-explanations may be hallucinated and unfaithful [2], suggesting a need for external tools that can audit and explain LLM systems.

Retrieval-augmented generation (RAG), an ability that enables LLMs to learn from external knowledge sources, has been pivotal in reducing their tendency to hallucinate. Using RAG, LLMs can augment their *internal* (trained) knowledge using *external* knowledge sources (e.g., web search results), drawing upon *both* to formulate their output. External sources are typically provided to an LLM through its input prompt (*context*), which must be engineered to contextualize the sources with respect to some instructions (e.g., to answer a given question using these sources). This prompt engineering method, which exploits the LLM's in-context learning capability, encourages the LLM to ground its output in this knowledge, leading to significant improvements in performance and trustworthiness. Nevertheless, integrating external knowledge introduces new complexities into the provenance of the LLMs' outputs, as there is no inherent method to determine which knowledge informed the output.

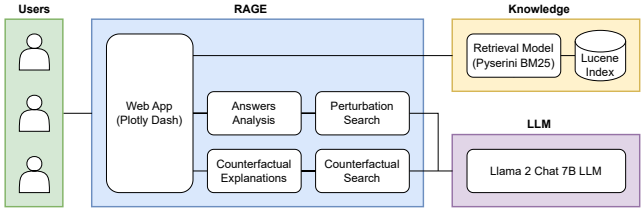

Fig. 1. The architecture of RAGE.

When considered under the lens of RAG, the issue of LLM provenance is twofold. On the one hand, by blending the model's internal knowledge with external knowledge sources, there is ambiguity in determining which of the two was used to formulate the output. On the other hand, when providing multiple external knowledge sources, there is ambiguity in determining which of the external sources were used. Furthermore, recent research has uncovered a "lost in the middle" bias in many popular LLMs [3], where the consideration given to sources can be influenced by their number and relative order in the prompt. Although RAG appears to exacerbate the challenge of explaining LLMs, its deliberate focus on external knowledge sources enable unique explainability insights through our ability to ablate, reorder, or modify the sources.

II. Methods

We propose an interactive tool, RAGE [4], which deduces provenance and salience for external knowledge sources using counterfactual explanations. Users pose a question to RAGE, which is then posed to a BM25 retrieval model to fetch relevant knowledge sources (documents from a separate index). These external knowledge sources are concatenated alongside the question into a single prompt, for which the LLM produces an answer. Under this formulation, RAGE replicates a typical LLM setup for the open-book question answering (QA) task, a prominent beneficiary of the RAG architecture. RAGE analyzes LLM answers over a strategic sample of perturbed prompts, deriving insights that implicate different sources (and source orders) in the LLM's determination of different answers.

In RAGE, the original ranked order of sources serves as the basis for all insights. These insights are mined by analyzing the

2375-026X/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDE60146.2024.00466

LLM's answers for a sample of source perturbations, namely *combinations* (subsets) or *permutations* (reorderings) of the retrieved sources. For combinations, RAGE draws a fixed-size random sample. For permutations, the Fisher-Yates shuffle can be used to sample, or variants of Murty's algorithm can find permutations that place relevant sources in high-attention positions (to counteract lost in the middle bias). As illustrated in Figure 2, RAGE summarizes the answers produced across sampled perturbations, grouping perturbations by answer and checking for valid rules that hold within-group.

To compute counterfactual explanations, RAGE strategically navigates the space of all perturbations to identify minimal perturbations that change the LLM's answer. By analyzing answers for minimal source combinations, RAGE produces citations for all answers observed. By analyzing source permutations, RAGE identifies minimal source reorderings that unexpectedly lead to different answers. RAGE uses source relevance scores (from the retriever) and attention (from the LLM) to prioritize the evaluation of perturbations more likely to admit a given answer, thereby reducing the search space.

III. Results

Across various domains, RAGE reveals insights that illustrate how LLMs resolve ambiguity, rationalize inconsistencies, and assess the saliency of data over time. Formalizing this taxonomy, we present qualitative experiments to recover meaningful insights from RAGE.

1) **Use Case 1: Ambiguous Answers.** When posed with an ambiguous or unrefined question, such as "Who is the best tennis player in The Big Three" (a famous trio of tennis players), RAGE will highlight minimal source combinations that support each player.
2) **Use Case 2: Inconsistent Sources.** When sources are contradictory, such as when recent and out-of-date tennis statistics sources are mixed, RAGE will discover whether similar source permutations (shuffling old and new sources) can confuse the LLM (i.e., exhibit lost in the middle bias).
3) **Use Case 3: Timelines.** When a question is posed over sources that form a timeline, such as yearly tennis statistics, RAGE will report how answers change as alternative timelines are considered.

IV. Conclusions

In this work, we outline our vision for explainability in LLMs, and for RAG in particular. Our novel tool, RAGE, takes the first steps towards explaining RAG and in-context learning in LLMs, exploiting the architectural shift towards the use of external knowledge in LLM inference. Under this retrieval paradigm, we can continue to build upon the current work in explainable AI and explainable information retrieval (IR), such as our recent work formulating counterfactual explanations for document ranking models [5]. Specific extensions to RAGE could explore the duality in IR between query and documents (i.e., question and knowledge sources), and devise complementary perturbations for the query. Beyond these IR inputs, novel insights could be discovered by perturbing the prompt in other ways, such as by modifying the tone or prompting strategy used in the prompt instructions.

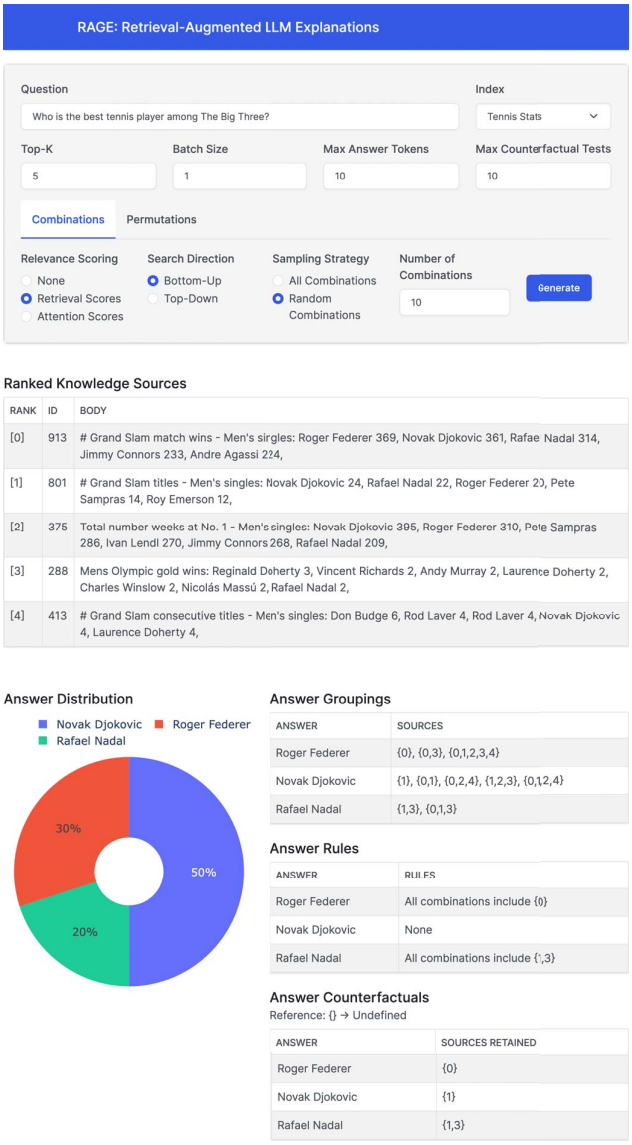

Fig. 2. Combination insights for a query about The Big Three, a famous trio of male tennis players.

References

[1] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *NeurIPS*, vol. 35, pp. 24 824–24 837, 2022.

[2] M. Turpin, J. Michael, E. Perez, and S. Bowman, "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting," *NeurIPS*, vol. 36, pp. 74 952–74 965, 2023.

[3] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *TACL*, vol. 12, pp. 157–173, 2024.

[4] J. Rorseth, P. Godfrey, L. Golab, D. Srivastava, and J. Szlichta, "RAGE against the machine: Retrieval-augmented LLM explanations," in *ICDE*, 2024.

[5] J. Rorseth, P. Godfrey, L. Golab, M. Kargar, D. Srivastava, and J. Szlichta, "CREDENCE: Counterfactual explanations for document ranking," in *ICDE*, 2023, pp. 3631–3634.