Explainable LLM-powered RAG To Tackle Tasks In The Unstructured-structured Data Spectrum

Darío Garigliotti[1]

[1]*University of Bergen, Norway*

Abstract

In the context of multiple spaces of research and application in text and information processing dominated by Large Language Models (LLMs), Retrieval-augmented Generation (RAG) provides a general framework with which to integrate external, explicit knowledge into the vast parametric knowledge of LLMs. In this paper, we present a crosspoint of tasks of diverse nature, maturity and level of cognitive challenge for an intelligent system, that nevertheless share in their analogies the suitability for being addressed by a similar RAG approach. Based on observations from several of our recent works, we reflect on the RAG framework, in particular about methods where the LLM is prompted with strategies to explain its generation output, across these tasks with components ranging from unstructured to structured data.

1. Introduction

Awaiting its due time to be eventually judged as a landmark, the paradigm of Generative AI is nowadays meaning an ever-consolidating leap forward in the possible solutions that machine learning techniques enable, and so too new creative challenges to be further solved. Playing with a rather sweet ambiguity –no more than what is needed–, a new generation (or era) of technology seems to be emerging thanks to a new *generation* level (or skill set), the one recently brought to mainstream consideration by powerful Large Language Models (LLMs) [1, 2]. The come-to-dominance of representation –or deep– learning, across multiple fields within areas such as image and text processing, meant at its time the combination of fundamental theoretical models in artificial neural networks, higher computational power by advancements in the dedicated hardware, and the availability of large amounts of data to effectively train those models in those infrastructures. A similar spot finds together in synergy a few key factors driving the successful capabilities of LLMs: transformers-based neural models learning auto-regressively over massive volumes of text scrapped from web pages, plus multi-task supervision over several datasets deemed relevant to approach linguistic and cognitive tasks –such as machine translation, summarization, and conversational search, to name a few–, and complemented with techniques like fine-tuning to continuously improving underperforming cases, and Reinforcement Learning with Human Feedback for alignment [3]. This paper describes the work in assessing the capabilities of LLMs in a series of research problems in overlapping areas such as Natural Language Processing (NLP), Information Extraction (IE) and Retrieval (IR), and Knowledge Representation (KR). The problems themselves meet in a common ground where their analogies enable a similar kind of approaches to address them and compare the corresponding outcomes.

The billions of learnable parameters that these neural architectures are equipped with implicitly encode the knowledge that their models capture. Their higher-abstraction and wider-context levels of information belong to the core of the seemingly emergent abilities of LLMs increasingly associated with various notions of intelligence [1, 2]. However, for many tasks that vary sufficiently in their underlying data distributions –as it is the case of shifting the information domains or capturing new behaviours in the expected machine-learned outcomes–, this implicit parametric data is not enough. In these scenarios, additional knowledge provided explicitly in the input, i.e. by *augmenting* the prompt to the LLM, allows its generation process to have access to this external context [4]. A prevalent umbrella of

approaches within this general strategy is Retrieval-augmented Generation (RAG) [5]. This framework is, at a high-level overview, a three-stage pipeline. Its first stage, that of retrieval, obtains ranked knowledge items to be the relevant contexts, typically textual excerpts that are identified with some retrieval method. The second stage, augmentation, integrates the contexts in a well-engineered prompt as the input for the LLM to trigger, in the last stage, the generation containing the kind of expected output at which these models excel. In this work, we use RAG as the general approach to address an ensemble of research problems. Characteristic aspects of each of these problems allow for bringing them here together by considering their analogies. We have approached each of these problems separately, instantiating them via a similar strategy based on RAG framework. The objective of this paper is then to present the similarities in the respective problem definitions that allows and remark our observations on comparing, too, their experimental results.

2. Experimental Space: Tasks, Datasets and Settings

Although it is possible to put these tasks in correspondence within a common approach given some key analogies, the research problems that we have studied present their unique aspects that make their ensemble diverse. Each task emerges in a particular context or domain, with earlier or later stages of consolidation in the space of research, and put in value by distinct users whose expectations for the kind of intelligent abilities from these assessed models vary too. Their common ground of lending themselves into being expressed in natural language allows for approaching all of them via language models, in particular, Large Language Models. Their dynamic character –by which new items of interest continuously emerge– calls for a mechanism to challenge the still-limited knowledge skills of LLMs. This, alongside a requirement to often overcome privacy constraints in the manipulation of data, suits well with the common umbrella of RAG here applied. We determined a handful of parameters in the experimental setting of the general RAG umbrella, so that each fixed combination of values for the parameters, i.e. a configuration, is a RAG-based method. Some of the recurrent parameters we experiment with are:

- The retrieval method, and the ranking length, used during the first stage of RAG;
- The order of the retrieved items when integrated to the prompt at augmentation phase –as some artefacts in the LLMs tend to memorize, in this case, the order–;
- The set of examples provided as few-shot learning also during prompting to illustrate, in particular, in what format the LLM should produce its expected output;
- The actual LLM finally invoked during generation.

2.1. Self-supported Question Answering

Question Answering (QA) is the first of the research problems in our work. Having been for long at the heart of the developments in NLP, QA is the basic instantiation of the intent of information seeking in natural language. Given how ubiquitous the need for search is, it lives as a fundamental in related areas such as Information Retrieval and Databases, yet QA is the primitive in a challenge like the Turing Test that has come to define, for many, what intelligent behaviour is. Here we consider our work on assessing several LLMs, from open to close commercial models, on performing QA in the context of proprietary data [6]. This type of data abounds in many organizations that intend to apply LLM-powered business intelligence systems to their vast collections. Yet the developers are typically legally restricted from sharing this data given how valuable it becomes, so inputting entire documents to an external state-of-the-art LLM is out of possibility. We experiment with a test collection built from corporate news articles that were published after the cut-off date for all the LLMs of interest, emulating the scenario where these documents participate when prompting for generation as external knowledge not present yet e.g. during LLM training or fine-tuning. Passages from these documents are retrieved for the question, and incorporated during prompt augmentation as contexts to generate the final answer.

Our setting actually addresses Self-Supported Question Answering (SQA), as the LLM is also requested to cite the passages that support the correctness of the generated answer, and answer and citation are evaluated. This evaluation can become a challenge in itself, as relevant benchmarking in literature may present discrepancies in the way that the typical retrieval evaluation possibly propagates through the entire RAG assessment while the LLM has access only to a subset of the universe of evidence [7].

In order to study how to extend the abilities of this framework to support explainable generations, we equip it at prompting stage with simple mechanisms that elicit interpretations from the model itself [8]. Specifically, we extend the prompt with an additional question in the augmentation phase. In our experiments, a couple of additional questions are used: (i) one that directly requests to the model for explaining the reasons behind the its generation output, and (ii) one that counterfactually proposes to the LLMM an alternative scenario where changes –irrelevant to the correct answer– are made in a given experimental parameter. We find some approach configurations exhibiting improvements while also others where the model generates differently from when it is not prompted with a request for explainability.

2.2. SDG Evidence Identification and Target Detection

The second task we consider in this paper is the one with least volume of research development, as it deals with the recency of (i) digitalization in the area of environmental impact assessment (EIA) and of (ii) the Sustainable Development Goals (SDG) framework which large proportion of the efforts within EIA are and will be guided by. Our task is actually a tandem of dual problems: *Evidence Identification (EI)* of textual excerpts from EIA reports where a given SDG target is addressed, and, symmetrically, *Target Detection (TD)* that consists in identifying relevant SDG targets addressed in a given textual excerpt [9]. The tandem captures a recurrent interplay between EI and TD in the EIA practice. This task pushes the limit of the apparent cognitive abilities of LLMs given that it concerns with more than information extraction and requires a comprehensive understanding of the complex space of factors affecting an environment and their relation with a SDG target. Our instantiation of RAG involves here a target as a query to retrieve its excerpts for EI task, and the other way around for TD task where SDG targets are indexed as a collection for retrieval. This domain of information is understood to be much less available digitally, and hence less likely to have been incorporated to the training routines for the LLMs of interest. Privacy requirements often apply to EIA reports too.

In this space of tasks, we also experiment with the RAG framework modified with similar strategies of interpretation elicitation as used for the SQA task [10].

2.3. Query Target Type Identification

The last of our three tasks is *Query Target Type Identification (TTI)* [11]. This problem has its roots in web search, where users were increasingly interested in obtaining focused, direct-access units of information, *entities*, beyond the "ten blue links" pointing to relevant documents; the Entity Retrieval task [12] and the whole Entity-oriented Search field [13] were born. The semantic class, or type, of an entity is known to improve the effectiveness of an entity retriever [14]. TTI is the problem of predicting, for a user query, the type information of its expected relevant entities. This research problem relates unstructured, keyword queries with types as structured entries in the ontology associated to the knowledge base storing the uniquely identified entities. Here, the types themselves retrieved for the query are assessed by the LLM after being added to the dedicated prompt [15]. Specific parameters that we assess in this problem include the usage of the textual type descriptor and/or that one of actual relevant entities in the augmented prompt.

Our interest for explainable generation gears in this task towards example-based explainability. We have adapted the RAG methods to augment the prompt further, by incorporating known relevant entities for the input query. Then, it is possible to compare its final RAG performance with the respective one without entity examples, as a proxy to justify why the method generates an answer when it is or it is not in the presence of these examples [16].

2.4. Conclusion

After having approached these tasks with similar RAG-based methods, our results shed light on some common observations over the assessed parameters as well as more distinctive aspects in the characterization of each research problem. In particular, we have described strategies towards equipping the common approach with explainability across this space of tasks in the unstructured-structured data spectrum.

Acknowledgments

References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.

[2] H. T. et al., Llama 2: Open Foundation and Fine-Tuned Chat Models, ArXiv abs/2307.09288 (2023).

[3] Y. Elazar, A. Bhagia, I. Magnusson, A. Ravichander, D. Schwenk, A. Suhr, P. Walsh, D. Groeneveld, L. Soldaini, S. Singh, H. Hajishirzi, N. A. Smith, J. Dodge, What's in my big data?, 2024. arXiv:2310.20707.

[4] A. Asai, M. Gardner, H. Hajishirzi, Evidentiality-guided generation for knowledge-intensive NLP tasks, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 2226–2243. URL: https://aclanthology.org/2022.naacl-main.162. doi:10.18653/v1/2022.naacl-main.162.

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 9459–9474.

[6] D. Garigliotti, B. Johansen, J. V. Kallestad, S.-E. Cho, C. Ferri, EquinorQA: Large Language Models for Question Answering over proprietary data, in: ECAI 2024 - 27th European Conference on Artificial Intelligence - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024), IOS Press, 2024.

[7] D. Garigliotti, On the Relevant Set of Contexts for Evaluating Retrieval-Augmented Generation Systems, in: Retrieval-Augmented Generation Enabled by Knowledge Graphs, co-located with ISWC 2024, CEUR-WS.org, 2024.

[8] D. Garigliotti, Explaining LLM-based Question Answering via the self-interpretations of a model, in: Advances in Interpretable Machine Learning and Artificial Intelligence, co-located with ECML-PKDD 2024, Springer Nature Switzerland, 2024.

[9] D. Garigliotti, SDG target detection in environmental reports using retrieval-augmented generation with LLMs, in: D. Stammbach, J. Ni, T. Schimanski, K. Dutia, A. Singh, J. Bingler, C. Christiaen, N. Kushwaha, V. Muccione, S. A. Vaghefi, M. Leippold (Eds.), Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 241–250. URL: https://aclanthology.org/2024.climatenlp-1.19.

[10] D. Garigliotti, Self-Explanatory Retrieval-Augmented Generation for SDG Evidence Identification, in: Advances in Conceptual Modeling, Springer Nature Switzerland, 2024.

[11] D. Garigliotti, F. Hasibi, K. Balog, Target type identification for entity-bearing queries, in:

Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17, 2017, pp. 845–848.

[12] K. Balog, M. Bron, M. de Rijke, Query modeling for entity search based on terms, categories, and examples, ACM Trans. Inf. Syst. 29 (2011) 1–31.

[13] K. Balog, Entity-Oriented Search, volume 39 of *The Information Retrieval Series*, Springer, 2018.

[14] D. Garigliotti, F. Hasibi, K. Balog, Identifying and exploiting target entity type information for ad hoc entity retrieval, Information Retrieval Journal 22 (2019) 285–323.

[15] D. Garigliotti, Retrieval-Augmented Generation for Query Target Type Identification, in: Retrieval-Augmented Generation Enabled by Knowledge Graphs, co-located with ISWC 2024, CEUR-WS.org, 2024.

[16] D. Garigliotti, Entity Examples for Explainable Query Target Type Identification with LLMs, in: Intelligent Data Engineering and Automated Learning – IDEAL 2024, Springer Nature Switzerland, 2024.