On the Relevant Set of Contexts for Evaluating Retrieval-Augmented Generation Systems

Darío Garigliotti[1]

[1]*University of Bergen, Norway*

Abstract

The recent interest in approaching more language and knowledge processing tasks via the Retrieval-Augmented Generation (RAG) framework allows for the consideration of evaluation criteria that can lead to a discrepancy in the way that the set of relevant results is determined for assessing retrieval-based performances. In this work, we describe and reflect on the consequences of such a discrepancy, and present basic results from experimentation over a RAG-based benchmark for Question Answering.

1. Introduction

Retrieval-Augmented Generation (RAG) [1] has been increasingly applied to a larger number of language processing tasks and domains, given its ability to combine explicit knowledge external to a Large Language Model (LLM) with the implicit knowledge that this LLM contains spreading across its billions of parameters [2, 3]. Throughout the three stages of its pipeline, a RAG-based system typically aims to address a problem of Question Answering (QA) nature, where some initial query or question is to be answered. For this question, RAG allows (i) to first retrieve a set of textual passages such that they become highly relevant contexts where the answer should be obtained from, and then (ii) to augment a well-engineered prompt by incorporating the contexts, so that (iii) to input this augmented prompt to a LLM generator. By the way that these stages work, and given the main involved items in a RAG pass −question, (retrieved and relevant) context(s), and (generated and correct) answer(s)−, several evaluation criteria are considered in related works [2]. The most prominent criteria are the question-context relevance (the paradigmatic Information Retrieval scenario), the context-answer relevance (where the prompt engineering aspects enter into play, and the so-claimed abilities of the LLM as well as its drawbacks, prominently hallucinations [4, 5]), and the question-answer correctness [6]. In the interplay of these criteria, an incompatibility arises when different relevant sets are used, specifically, by aiming to evaluate the end-to-end performance of a RAG-based method against items which, while correct, are never presented to the generator so they are irretrievable by it. The rest of this work discusses this incompatibility in detail and illustrates it with experimental results from a Question Answering benchmark.

RAGE-KG'24: Retrieval-Augmented Generation enabled by Knowledge Graphs workshop at ISWC 2024
dario.garigliotti@uib.no (D. Garigliotti)
0000-0002-0331-000X (D. Garigliotti)

CEUR Workshop Proceedings (CEUR-WS.org)

2. The problem of determining the relevant set of contexts

Let $\mathcal{D}$ denote the universe of all documents under consideration in a retrieval scenario. In our case, this is an indexed collection of uniquely identified textual passages with respect to which the first RAG stage, retrieval, obtains a ranking of contexts to use during its second stage, augmentation, to build the prompt that elicits the last stage, generation. Given a question q, a retrieval method m is used to score documents from $\mathcal{D}$ and return a ranking $Out(q)$ of the top k documents for q. The question q can be, among the most typical cases, a (primitive) query in the traditional sense within Information Retrieval (IR), a *wrapper* of a primitive query that accompanies it in an augmented prompt asking to the generator for ranking with respect to the query, or a question within a Question Answering task for which the obtained answers are to be evaluated with some retrieval criteria (e.g. answer recall, or exact matching, in QA [7]). The ranking $Out(q)$ is evaluated with respect to a set of relevant documents for q, $Rel(q) \subseteq \mathcal{D}$. For short, let us refer to this sets by Out (the retrieved results) and Rel (the relevant results), respectively. Two set-based metrics, precision (P) and recall (R), are typically used as fundamental criteria to evaluate a retrieval system. They are defined, in terms of Out and Rel, as $P = |Out \cap Rel|/|Out|$ and $P = |Out \cap Rel|/|Rel|$.

Given the way that the relevant set is usually obtained when building a test collection, a selection of M methods $\mathcal{M} = \{m_i\}_{i=1}^{M}$ are used to retrieve a ranking for q, specifically top k_i results with each method m_i. As typically the cut-off $k_i = K$ for all i, a maximum of $M * K$ documents are *pooled* for q and judged according to some relevance criteria. In the case of binary relevance, this leads to the set Rel defined above, $Rel(q) = \{d \in \mathcal{D} : d \text{ is relevant to } q\} = \bigcup_i Rel_{m_i,k_i}(q)$, where $Rel_{m_i,k_i}(q)$ is the subset of documents in $Out_{m_i,k_i}(q)$ such that they are relevant to q, and $Out_{m_i,k_i}(q)$ the top k_i documents for q retrieved by m_i that were put in the pool to build Rel.

For simplicity, let us assume that (1) the method used during the retrieval phase of a RAG system S is $m' \in \mathcal{M}$, and let us also assume that (2) the top K' results ranked by m' is such that $K' < K$. This is a common situation since for test collections it is often the case that $K = 100$ or $K = 1,000$ while K' is way smaller, in the order of tens (e.g. a recently devised benchmark [8] uses $K' =$ 3, 5, 10 and 20). From (1) and (2), it follows that (3) $Rel' = Rel_{m',K'}(q) \subsetneq Rel$. The K' contexts in Rel' are then used to augment the prompt in the second RAG stage.

When the prompt is finally input to the LLM in the third RAG stage, generation, the LLM, unlike any of the methods in $\mathcal{M}$ such as m', does not consider $\mathcal{D}$ as the universe to retrieve from, but Rel', the set of contexts that it *has seen* or *is aware of* for q. The rest of the documents in Rel that are not *seen* during generation, this is, $Rel - Rel'$, cannot be retrieved by LLM.[1] Hence, for any context $d \in Rel - Rel'$, $d \notin Out_{LLM}(q)$, with $Out_{LLM}(q)$ –Out_{LLM}, for short– the set of results returned by the LLM in its generated answer. And because of (4) $Out_{LLM} \subseteq Rel'$, then from it and (3) it follows that (5) $Out_{LLM} \subseteq Rel$, and from this, (6) $Out_{LLM} \cap Rel' = Out_{LLM} \cap Rel$. This means that the measurement of the performance of S (the LLM performance as a re-ranker of the augmented contexts Rel') in terms of precision does not change if measured with respect to the original relevant set Rel, $P(S) = |Out_{LLM} \cap$

[1] Unless due to some hallucination that generates in the answer a (n identifier of a) context that has not seen in the prompt.

$Rel'|/|Out_{LLM}| = |Out_{LLM} \cap Rel|/|Out_{LLM}|$. However, **recall measurement does change** since the denominator in its expression changes from Rel to its proper subset Rel'.

This discrepancy in the criterion over the relevant set of contexts has a direct effect in the entire evaluation, since the use of Rel instead of Rel' for recall-oriented measures involves to divide by a larger size of relevant set and hence leads to observing an underperforming system. Recent literature aimed to establish a benchmark [8] for vanilla instances of the RAG framework, for example, in a paradigmatic task like QA, presents an evaluation within the Rel-based criterion that we criticize in our work. In the next section, we describe its methodology and demonstrate the discrepancy by the means of experimental results with this benchmark.

3. Experimental results over ALCE benchmark

The ALCE benchmark [8] studies RAG performance over three QA datasets. One of them, QAMPARI [9], requires to answer each question with one or more entities −in most cases, with multiple entities− from a knowledge graph (KG) of reference. When asked via the augmented prompt to answer the instance question, the surface form −with which each of the entities appear in one or more of the retrieved contexts provided during augmentation− is to be found in the generated answer as an exactly delimited substring. This evaluation metric is referred to as exact matching (EM) [7], and essentially corresponds to a recall-oriented measurement where all the known relevant entities are contrasted with those actually present, i.e. *retrieved* by the LLM as a re-ranker, in the generated answer.

Within the family of QA problems, ALCE assesses the instantiation of *attributed QA* or *self-supported QA* [10, 5], a task that seeks to answer a question and complement it with evidence cited from the contexts that augment the prompt, in order to support the correctness of said answer. This citation of the passages supporting the answer is evaluated with the basic set-oriented metrics of P and R. It often happens that when augmenting the prompt with only the top $K' = 5$ or 10 of the retrieved contexts, the RAG-based system is not providing the generator with a set of available contexts among which all the relevant entities can be supported, but instead only a subset. As we analyze it in the previous section, evaluating the recall-based measurements with respect to Rel (the set of relevant contexts for the first RAG stage) can be misguiding in reporting the true performance when taking into account only those contexts available at generation time. There are then two aspects of the performance that are evaluated by a recall-based metric: answer recall and citation recall, i.e., two criteria possibly affected by the discrepancy previously analyzed. For the latter metric, the phenomenon happens with Rel^C, the set of relevant contexts for the query at retrieval stage, and Rel'^C, its subset when restricting it to only the contexts used during augmentation. For the former, answer recall, an analogous situation is led to by the set of known correct answers, Rel^A, and the subset Rel'^A induced by said set of relevant contexts available during augmentation where correct answers appear.[2]

We experiment with a setup for a naive RAG framework based on the setting used by ALCE Gao et al. [8]. Our experiments assess parameter configurations in all the three stages,

[2] This is a simplification of the less likely, general case, where a context is deemed relevant by a first-pass retrieval method, i.e. retrieved in ranking $Out(q)$, yet it does not support any relevant answer for q.

Table 1
Experimental results for RAG configurations with GPT-3.5 and GPT-4 as the generators, over our set of 60 selected QAMPARI instances. In all these experiments, retrieval is dense, and the prompt includes two examples. In each block by LLM, the best performance on a metric is shown in **bold**.

QAMPARI instances: 60 questions. Relevance set: Rel.					
LLM	Retrieval cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall
GPT-3.5	5	By ranking	**0.1046**	0.7056	0.1563
		Random	0.0993	0.6403	0.1227
	10	By ranking	0.1043	**0.7303**	**0.1655**
		Random	0.0883	0.6315	0.1145
GPT-4	5	By ranking	0.1563	0.6842	0.2031
		Random	0.1364	0.6833	0.1677
	10	By ranking	**0.1817**	0.6404	**0.2156**
		Random	0.1786	**0.7039**	0.2111

QAMPARI instances: 60 questions. Relevance set: Rel'.					
LLM	Retrieval cutoff	Passage order	Answer Recall	Citation Precision	Citation Recall
GPT-3.5	5	By ranking	**0.4596**	0.7056	**0.5425**
		Random	0.4018	0.6403	0.4956
	10	By ranking	0.4085	**0.7303**	0.4706
		Random	0.3393	0.6315	0.3795
GPT-4	5	By ranking	**0.624**	0.6842	0.6714
		Random	0.5937	0.6833	**0.6753**
	10	By ranking	0.5931	0.6404	0.6115
		Random	0.5806	**0.7039**	0.6264

including, for example, the retrieval method and cut-off, the order of passages from retrieval into augmentation and the number of shot examples, and the generator model. The parameters, and the values experimented with for each of them, are all very similar to the ones used in ALCE –in some cases identical, as it is the case for retrieval and generation parameters–. A major distinction is in our prompt, that is longer than its ALCE counterpart and accounts for more instructions, to better and more politely specify the expected answer. The main request in the prompt for both the benchmark and our experiments is to answer the question and refer to the passages that support such a response. In this way, our work is an approximation to reproduce part of the experiments developed in the benchmark. We work with a selection of 60 questions from QAMPARI dataset, 20 from each of the three question groups (simple, intersection, and complex, all based on hopping criteria over the KG of reference), selected to make all the possible parameter configurations meaningful. In particular, we ensure that there is a relevant answer occurring in the top 3 (and so in the top 5 and 10) retrieved passages for each method for each question.

Table 1 reports our experimental results for configurations where the first RAG stage is performed by a dense retriever and the prompt does two-shot learning with provided examples. The LLMs used in these configurations are GPT-3.5 (gpt-3.5-turbo-0125) [11] and GPT-4 [12]. As we described in the previous section, the performance in terms of precision is not affected by the involved factors, except for a possible larger degree of non-determinism in an LLM that here is not required. Instead, for recall-based measurements of answer and citation, we can observe that our obtained results when considering, as in ALCE, Rel as the denominator of R are in the magnitude of the values reported in ALCE. In the alternative evaluation w.r.t. Rel' –bottom half of Table 1–, where the performance of the LLM generation is assessed only with respect to the contexts that were available at prompting, the performances are substantially higher and, as we argue, report a more principled assessment of the respective RAG configuration.

4. Conclusion

This paper has reflected on the aspect of determining appropriately the set of relevant contexts with which to evaluate the performance of a RAG system. Our observations call for considering its determination in terms of the subset of retrieved contexts that is available during augmentation such that the evaluation does not arrive to claim an underperforming method, while also uses a more principled criterion.

Acknowledgments

This work was funded by the Norwegian Research Council grant 329745 Machine Teaching for Explainable AI.

References

[1] 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.

[2] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-Augmented Generation for Large Language Models: A Survey, 2024. arXiv:2312.10997.

[3] 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.

[4] N. Liu, T. Zhang, P. Liang, Evaluating verifiability in generative search engines, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 7001–7025.

[5] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, N. McAleese, Teaching language models to support answers with verified quotes, ArXiv abs/2203.11147 (2022). URL: https://api.semanticscholar.org/CorpusID:247594830.

[6] S. Es, J. James, L. Espinosa Anke, S. Schockaert, RAGAs: Automated evaluation of retrieval augmented generation, in: N. Aletras, O. De Clercq (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 150–158. URL: https://aclanthology.org/2024.eacl-demo.16.

[7] I. Stelmakh, Y. Luan, B. Dhingra, M.-W. Chang, ASQA: Factoid questions meet long-form answers, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 8273–8288. URL: https://aclanthology.org/2022.emnlp-main.566. doi:10.18653/v1/2022.emnlp-main.566.

[8] T. Gao, H. Yen, J. Yu, D. Chen, Enabling large language models to generate text with citations, in: H. Bouamor, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 6465–6488.

[9] S. Amouyal, T. Wolfson, O. Rubin, O. Yoran, J. Herzig, J. Berant, QAMPARI: A benchmark for open-domain questions with many answers, in: S. Gehrmann, A. Wang, J. Sedoc, E. Clark, K. Dhole, K. R. Chandu, E. Santus, H. Sedghamiz (Eds.), Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), Association for Computational Linguistics, Singapore, 2023, pp. 97–110. URL: https://aclanthology.org/2023.gem-1.9.

[10] B. Bohnet, V. Tran, P. Verga, R. Aharoni, D. Andor, L. B. Soares, M. Ciaramita, J. Eisenstein, K. Ganchev, J. Herzig, K. Hui, T. Kwiatkowski, J. Ma, J. Ni, T. Schuster, L. S. Saralegui, W. W. Cohen, M. Collins, D. Das, D. Metzler, S. Petrov, K. Webster, Attributed question answering: Evaluation and modeling for attributed large language models, 2022. URL: https://arxiv.org/abs/2212.08037.

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

[12] OpenAI, GPT-4 Technical Report, 2024. arXiv:2303.08774.