The Affect of Embedding Methods on Faithfulness of Retrieval-Augmented Generation (RAG)

Amandeep Kaur Singh
amandeepkaur@umass.edu
University of Massachusetts
Amherst, USA

Chang Jin
changjin@umass.edu
University of Massachusetts
Amherst, USA

Reagan Keeney
rkeeney@umass.edu
University of Massachusetts
Amherst, USA

ACM Reference Format:
Amandeep Kaur Singh, Chang Jin, and Reagan Keeney. 2024. The Affect of Embedding Methods on Faithfulness of Retrieval-Augmented Generation (RAG). In *COMPSCI 646: Advanced Information Retrieval, Fall 2024, Amherst MA*. ACM, New York, NY, USA, 2 pages.

1 Problem Statement

Large Language Models (LLMs) are powerful tools capable of question answering over broad contexts, offering impressive natural language generative capabilities. LLMs are trained on vast volumes of data to enable them to use billions of parameters to generate original output for tasks such as answering questions, translating between languages, and completing sentences. However, due to the black-boxed nature and non-deterministic features of LLMs, their results are not consistent or interpretable and are vulnerable to 'hallucinations', where the model incorrectly answers questions with false information instead of reporting a lack of adequate topic knowledge.

Retrieval-augmented generation (RAG) is a technique that optimizes a large language model's output, utilizing information retrieval to consult documents from a reliable knowledge base outside of its training data sources before generating a response using an LLM. Ensuring that these LLM-generated responses are accurate, pertinent, and helpful is important. Current research faces challenges in assuring the faithfulness of the generated responses—that is, that the output stays correct and in line with the acquired data—a significant problem with RAG[5].

The selection of embedding techniques, which translate textual data into numerical vectors used to compute similarities and retrieve pertinent documents, is a key determinant of a Retrieval-Augmented Generation model's faithfulness and performance. Different embedding strategies can lead to drastically different retrieval accuracy, quality, and reliability of the generated responses. For instance, dense embeddings like BERT can capture semantic relationships, improving relevance and potentially maintaining faithfulness [1]; sparse embeddings, on the other hand, yield faster results but may sacrifice contexts, which can lead to less faithful responses [6].

Given these variations in embedding models, it is important to optimize RAG for applications that need high factual consistency and alignment. Our study thus attempts to answer the following question: Which embedding technique best maintains faithfulness in responses produced by RAG models on a domain-specific dataset? We hope that our findings will support applying RAG for high-stake applications where maintaining faithfulness and building user trust (through interpretability) is crucial.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

2 Motivation

Retrieval-Augmented Generation (RAG) has shown significant promise in enhancing the performance of large language models with the integration of retrieved information. The performance of RAG models can vary depending on the choice of embedding models used for information retrieval.

As such, while Retrieval Augmented Generation (RAG) offers a solution to some of the major shortcomings of Large Language Models (LLMs) for question-answering tasks, however, optimization of this pairing of systems remains inadequately explored. By systematically evaluating and comparing the faithfulness of RAG with different embedding models for a domain-specific dataset, we hope to gain valuable insights into how the choice of embeddings impacts the quality of the generated responses.

3 Related Work

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks' by Lewis et. al[8] introduced Retrieval-Augmented Generation as a novel approach combining information retrieval (IR) with text generation for knowledge-intensive Natural Language Processing (NLP) tasks in 2020. The proposed approach includes two types of memories:

(1) Non-parametric Memory: Document index which can be 'hot-swapped' in order to update knowledge as needed.
(2) Parametric Memory: Parameters of a generative model such as BART.

A key finding of this paper was that the proposed RAG models perform better and tend to be more 'grounded' i.e., they generate "more specific, diverse and factual language" than the then state-of-the-art parametric-only seq2seq baseline chosen for the study (BART). Lewis et. al concluded their findings with a discussion of the broader impact of their work, suggesting applications for medical information retrieval. This is particularly relevant for our proposed project, where we plan to evaluate the impact of the choice of embedding models on the performance of RAG by applying it to the RAG-Mini-BioASQ dataset.

More recently, Şakar and Emekci[12] performed 23,625 grid-search iterations to obtain results comparing several RAG architectures and techniques. The embedding models compared included

OpenAI's text-embedding-v3-large model, Beijing Academy of Artificial Intelligence's (BAAI) open-source bge-en-small model, and Cohere's cohere-en-v3 model. As discussed in the results section of [12], the choice of embedding models significantly affects the performance of the RAG systems in terms of hardware utilization (runtime, CPU usage, and memory usage), and performance (evaluated with median similarity scores). While Şakar and Emekci's research included a direct comparison of embedding models, the focus was on efficiency with regard to runtime and hardware requirements, not on the quality of generated outputs. Which embedding model is most optimal seems to vary by use case [4, p.9], leaving ambiguity on the correct conditions for a given model.

As such, evaluating the retrieval quality and the generated output of RAG models remains a challenging task. Evaluating whether a model's outputs contain the correct information requires specific evaluation metrics and strategies. [4] explores some of the challenges in evaluating RAG systems and proposes 'A Unified Evaluation Process of RAG (AUEPORA) as a structured approach to RAG evaluation.

For the purposes of this research, we focus on 'faithfulness' as a measure of generated output quality. Broadly, faithfulness is the level of agreement between the generated text and the contexts over all claims. [3]

4 Baselines

To determine the set of embedding models to test, we refer to Muennighoff et al.'s Massive Text Embedding Benchmark [9] and the corresponding leaderboard.

As our baseline, we intend to use the Beijing Academy of Artificial Intelligence's open source BGE-SMALL-EN-V1.5[11] model, which is the current version of the top performing model in Şakar and Emekci's resource efficiency comparison [12].

5 Dataset(s)

For the purpose of our study, we plan to use a subset of the BioASQ 11b dataset [10] in the domain of biomedical question answering. Specifically, the pre-generated RAG-Mini-BioASQ dataset[2], which is designed for Retrieval-Augmented Generation use cases.

5.1 Key Features of the RAG-Mini-BioASQ dataset:

(1) Subset of BioASQ dataset: Given the limited resources and timeline of the project, using a smaller dataset is advantageous for ensuring the scope is achievable while the information ne still retain the complexity from the original dataset.
(2) Structured for RAG-based approaches/experiments: The RAG-Mini-BioASQ dataset already has a correct response to the question, which could be useful for evaluation purposes. The dataset consists of the following:
    1. Text-Corpus: Consists of biomedical text passages and their IDs.
    2. Question-Answer-Passages: Consists of questions (representing the users' information need), a correct answer to those questions, and a list of relevant passage IDs from the Text-Corpus data.
(3) Biomedical focus: The chosen dataset is domain-specific; enabling us to perform a more accurate assessment of various embedding models for this specific use case.
(4) Complexity: The questions are designed to be particularly challenging[7], which may make disparities between the faithfulness of different embeddings more apparent.

In particular, in critical domains such as biomedical question-answering, hallucinations from generative language models can raise significant safety concerns. Thus it is uniquely important to understand, evaluate, and ensure faithfulness to verified correct reference information in such fields.

6 Evaluation

Given that our research centers on the quality of the generated text, so do our evaluation metrics. We aim to use the RAG-specific evaluation of faithfulness, as defined in the RAGAs framework [3]. This faithfulness metric is evaluated using all claims in the LLMs generated answer cross-referenced with the retrieved context documents. We intend to use faithfulness as a metric to monitor the degree to which the LLM integrates niche knowledge from the corpus to produce correct answers. A high faithfulness score would represent both integration and a lack of hallucination, both of which are important to ensure for trustworthy models.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Jill Burstein, Christy Doran, and Thamar Solorio, (Eds.) Association for Computational Linguistics, Minneapolis, Minnesota, (June 2019), 4171–4186. DOI: 10.18653/v1/N19-1423.

[2] enelpol. 2024. Rag-mini-bioasq. (2024). https://huggingface.co/datasets/enelpol/rag-mini-bioasq.

[3] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Nikolaos Aletras and Orphee De Clercq, (Eds.) Association for Computational Linguistics, St. Julians, Malta, (Mar. 2024), 150–158. https://aclanthology.org/2024.eacl-demo.16.

[4] Yunfan Gao et al. 2023. Retrieval-augmented generation for large language models: a survey. *arXiv preprint arXiv:2312.10997*.

[5] Or Honovich et al. 2022. True: re-evaluating factual consistency evaluation. (2022). https://arxiv.org/abs/2204.04991 arXiv: 2204.04991 [cs.CL].

[6] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, (Eds.) Association for Computational Linguistics, Online, (Nov. 2020), 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550.

[7] Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: a manually curated corpus for biomedical question answering. *Scientific Data*, 10, 1, 170.

[8] Patrick Lewis et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (NIPS '20) Article 793. Curran Associates Inc., Vancouver, BC, Canada, 16 pages. ISBN: 9781713829546.

[9] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: massive text embedding benchmark. (2023). https://arxiv.org/abs/2210.07316 arXiv: 2210.07316 [cs.CL].

[10] n.d. 2023. Bioasq training 11b. (2023). http://participants-area.bioasq.org/datasets/.

[11] Beijing Academy of Artificial Intelligence. [n. d.] Bge-small-en-v1.5. (). https://huggingface.co/BAAI/bge-small-en-v1.5.

[12] Tolga Şakar and Hakan Emekci. 2024. Maximizing rag efficiency: a comparative analysis of rag methods. *Natural Language Processing*, 1–25.