Aryana Far
4 August 2024

Evaluating RAG Capabilities for Improved Document Search

Abstract

This proof of concept (POC) explored the capabilities of a Retrieval Augmented Generation (RAG) system for the use case of providing specialized question answering for marketing and engineering team members at our company, which will be launching Generative Artificial Intelligence (GenAI) products in the coming quarter. For this POC, I tested four key RAG system configurations with different language models, embedding models, prompts, chunking, and model temperatures, such that each configuration could be customized for both engineering and marketing audiences. The four RAG systems showed rather consistent performance across configurations, with overall performance scores ranging from 66% to 72%. The RAG system performed slightly better when configured with the larger, overlapping chunking than with the smaller, non-overlapping chunking. Furthermore, the RAG system consistently performed better for the engineering audience than for the marketing audience, perhaps reflecting its alignment with technical language due to the chosen document sources. Given these findings, I recommend further exploration of model configurations, development of more robust evaluation metrics, a more versatile set of document sources, and model fine-tuning.

Introduction

In this report, I discuss the capabilities of a Retrieval Augmented Generation (RAG) system in improving question answering capabilities at our company. The RAG system I developed for this proof of concept (POC) is suited to answer questions about Natural Language Processing (NLP) and Generative Artificial Intelligence (GenAI). The use case of this RAG system is to promote the productivity of the engineering and marketing teams at our tech company, which is aiming to launch a series of GenAI products in the next quarter. The RAG system can take inputs of questions, and it is customized for both engineering and marketing type audiences through variations in prompt instructions.

Key Findings

There were five main key findings of implementing this POC. Firstly, across the four key RAG pipeline runs, the overall performance score was relatively consistent, and remained within the range of 66% to 72%. This range of scores reflects decent similarity between RAG outputs and gold standard answers. The consistency of this range implies that the tested changes in the model configurations may not have yielded significantly different results. Secondly, while the performances were rather consistent, the RAG pipelines that utilized retrievers with larger, overlapping chunking configurations performed slightly better than those that utilized retrievers with smaller, non-overlapping chunking configurations. This reflects the necessity of strategic chunking for the task at hand. Thirdly, each key RAG pipeline run performed better for the engineering audience than for the marketing audience. This implies that the construction of a more versatile set of document sources and the implementation of model fine-tuning are needed to balance model performances between audiences. Fourthly, the difference between the evaluation metric counterparts (cosine similarity and BERTScore) that made up the overall performance metric sometimes exhibited differences of up to 16%. This variation in the counterparts reflects that different evaluation metrics demonstrate different aspects of text similarity, and some evaluation metrics may be better suited than others. The fifth and final key finding of this POC is that the gold standard answers are subjectively defined as “truth.” This implies that evaluation relies heavily on the subjective choice of the gold standard answers, thereby complicating the interpretation of the evaluation metrics.

Methodology

Technical Approach

The sources I used for this RAG system's vectorstore are ArXiv papers on RAG and NLP, Lily Weng blogs about Open Domain Question Answering and related topics, and relevant Wikipedia articles.

The language model I used was Cohere. To convert the document's contents to vector representations, I utilized the all-distilroberta-v1 embedding model. Through my prompt, I accommodated the two user personalities of engineering and marketing by constructing two separate RAG pipelines that were the same other than slight variations in prompts specifying audience attributes and the desired level of detail in the answer. To build the retriever, I created documents from the sources with chunk sizes of 300 that overlapped by 40.

Testing and Evaluation

In experimenting with various model configurations, I utilized two methods of evaluation against a validation set of 75 question and gold standard answer pairs for each audience. The first evaluation method I used involved simple cosine similarity comparisons between the RAG answer and the gold standard answer. This is well-suited because it quantifies the semantic similarity between the generated text and the reference text in a straight-forward manner, providing a clear indication of how closely the texts resemble each other (Walker, n.d.). The second method I used involved a BERTScore, providing precision, recall, and F1 score. BERTScore works by generating similarity scores between the embeddings of the texts, matching each token to its most similar counterpart, adjusting for rare word importance, and rescaling for interpretability. BERTScore is well-suited because it effectively captures both semantic meaning and context (Sojasingarayar, 2024). The final metrics I reported were (1) average cosine similarity across the 75 validation datapoints, (2) average BERTScore (precision, recall, F1 score) across the 75 validation datapoints, and (3) a combined metric of average cosine similarity and average BERTScore, which I achieved by taking the mean of the two counterparts.

Before choosing the final specifications for the RAG system, I conducted test runs where I experimented with the large language model (LLM), the embedding model, the prompt, the chunk size, the chunk overlap, and the LLM hyperparameters. For LLMs, I experimented with Mistral-7B-Instruct-v0.2 and Cohere. For embedding models, I experimented with multi-qa-mpnet-base-dot-v1 and all-distilroberta-v1. For prompts, I tested three sets of prompts. Each set of prompts included two prompts geared towards either audience of engineering or marketing. I designed prompts that were clear and detailed, expressing instructions multiple times in different words to push for appropriate responses. The prompts for engineering support expressed that the audience was already highly knowledgeable in the fundamentals of NLP and needed detailed technical responses. The prompts for marketing support expressed that the audience needed high-level responses that didn't involve too much technical detail. I also asked for a slightly longer answers for engineering than for marketing. These prompts are provided in **Prompts Table** in the **Appendix**. For chunking, I tested two specifications. The first chunking specification involved chunk sizes of 128 tokens with 0 overlap. The second chunking specification involved chunk sizes of 300 with 40 overlapping tokens. I tested these different chunk sizes and overlaps to see what lengths and overlaps of strings of information would be meaningful for this RAG system. For the LLM hyperparameters, just for the Mistral modes, I experimented with temperature. I tested the temperatures 0.2 and 0.6. I chose 0.2 because of its proximity to 0, which is the least random specification of temperature; particularly, 0.2 seems to be deterministic while still being slightly variable. I chose 0.6 because of its higher value which may lead to more randomness but also more creativity, without being too random. The exact specifications of the RAG systems I tested as well as their performances are provided in the **RAG System Experiments Table** in the **Results (Lessons Learned)** section.

Results

Proof of Concept (POC) Functionality

This POC demonstrated the core idea that a RAG system could potentially enhance information retrieval and question answering for our tech company in its pursuit of launching GenAI products. Furthermore, this POC met the objective of providing decently accurate and contextually relevant answers to questions related to NLP and GenAI for different audiences.

Lessons Learned

During experimentation, where I conducted test runs with different models, embeddings, prompts, chunking, and model temperatures, there were both technical and non-technical insights. On a technical level, experimentation revealed insights on overall functionality, model configurations, user personality customization, and evaluation. The **RAG System Experiments Table** below details the performances of the four key RAG pipeline runs for both the engineering audience and marketing audience against a set of 75 validation gold standard answers. In terms of overall functionality, the combined scores across the four models and two audiences were similar, ranging from 66% to 72%, which reflects a decently accurate level of performance. In terms of model configurations, the two RAG systems with the 300 chunk size and 40 chunk overlap specifications performed better than the two RAG systems with the 128 chunk size and 0 chunk overlap specifications, emphasizing the importance of strategic chunking. In terms of user personality customization, each RAG pipeline run performed better for the engineering audience than for the marketing audience, reflecting the potential need for a more versatile set of documents and for model fine-tuning. In terms of evaluation, the large differences between the cosine similarities and BERTScores as well as the subjective nature of gold standard answers complicate the interpretation of the final combined evaluation scores. On a non-technical level, experimentation revealed that the RAG system's ease of use and its decent performance make it viable, given its further development and improvement, as a solution to the internal question-answering use-case for both engineering and marketing teams. Overall, the PoC demonstrates the potential of RAG systems to enhance internal processes, paving the way for broader applications within the company.

RAG System Experiments Table

LLM	Embedding Model	Chunk Size	Chunk Overlap	Prompts	Temperature	Cosine Similarity: Engineering	BERTScore: Engineering	Combined Score: Engineering	Cosine Similarity: Marketing	BERTScore: Marketing	Combined Score: Marketing
Cohere	multi-qa-mpnet-base-dot-v1	128	0	Prompt A		74.3264%	**Precision**: 62.6787% **Recall**: 66.3099% **F1 Score**: 64.2672%	69.2968%	70.7769%	**Precision**: 55.8472% **Recall**: 67.2561% **F1 Score**: 60.8451%	65.8110%
Mistral	all-distilroberta-v1	128	0	Prompt A	0.2	75.0531%	**Precision**: 63.1623% **Recall**: 63.3604% **F1 Score**: 63.1663%	69.1097%	72.7712%	**Precision**: 59.7814% **Recall**: 65.3781% **F1 Score**: 62.3133%	67.5412%
Mistral	multi-qa-mpnet-base-dot-v1	300	40	Prompt B	0.6	78.1674%	**Precision**: 61.0423% **Recall**: 64.4359% **F1 Score**: 62.5765%	70.3720%	75.6638%	**Precision**: 55.4833% **Recall**: 66.0372% **F1 Score**: 60.1319%	67.8979%
Cohere	all-distilroberta-v1	300	40	Prompt C		77.3411%	**Precision**: 65.9659% **Recall**: 66.2000% **F1 Score**: 66.0100%	71.6755%	77.0439%	**Precision**: 63.1741% **Recall**: 67.9037% **F1 Score**: 65.2514%	71.1477%

Challenges and Limitations

There were indeed obstacles, model shortcomings, and surprises that one should have an eye on in follow-up POCs or implementation. One obstacle was that the sources of documents for RAG retrieval were comprised of a small number of static technical documents. This meant a limited, unchanging, and highly technical knowledgebase, which may compromise the system's ability to provide high-quality, up-to-date, and versatile responses. Another obstacle involved practical constraints of model testing, experimentation, and fine-tuning. The computational resources and time required to extensively validate and improve a RAG system's performance are substantial, which limited the comprehensive exploration of the system's true potential. Additionally, there were obstacles relating to the evaluation metrics. The metrics used to evaluate the system's performance, cosine similarity and BERTScore, have their own shortcomings. For instance, while cosine similarity is useful for quantifying the overall similarity of texts, it ignores word order and context, and may fail to identify semantic differences (Walker, n.d.). BERTScore, on the other hand, provides a robust measure of semantic similarity but sometimes suffers from bias (Sojasingarayar, 2024). These limitations in evaluation metrics suggest that the reported performance might not fully reflect the real-world effectiveness of the RAG system. Finally, there were limitations with the gold standard answers being inherently subjective, further complicating the interpretation of evaluation metrics. Overall, while the RAG system shows promise in enhancing productivity for both the engineering and marketing teams, careful consideration of these obstacles is crucial for successful implementation.

Next Steps

If there were more time, I would create a larger, more versatile set of documents (and continually update it with relevant information), conduct model fine-tuning, and run more RAG pipeline testing with more variables. My analysis suggests the further pursuit of the following questions:

1. How does a RAG system perform with a larger vectorstore?
2. What strategies should be implemented for the ongoing updating of the vectorstore and model to ensure future proofing?
3. How might fine-tuning the model to cater to different audience needs, such as engineering vs. marketing, change performance?
4. What is the ideal LLM/embedding model/prompt/chunking/temperature/etc…for this use case?

By addressing these questions, we can further refine the POC into a prototyping phase.

Summary & Recommendations

Given the findings, I recommend moving forward with the prototyping of the RAG pipeline system. Beyond the starting point that this POC provides, I propose enlargement of the set of document sources, consistent vector store updates, fine-tuning, and more robust testing to arrive at the prototype. Nonetheless, the RAG systems tested in this POC demonstrated consistently decent performance across various configurations and aligned well with the internal question-answering use-case. In terms of deployment, estimating average and peak loads would help determine exact architectural considerations. In such a dynamic working environment, I suspect the system would have consistently high usage with regular peaks. Given this assumption, I would suggest reserving the LLM. Furthermore, we must integrate data security into the system to ensure safety in the event of data breaches. Overall, this approach will set a solid foundation for future implementation, balancing performance and scalability while ensuring that we meet our key objectives safely.

References

Sojasingarayar, A. (2024, January 15). BERTScore explained in 5 minutes. *Medium*. https://medium.com/@abonia/bertscore-explained-in-5-minutes-0b98553bfb71

Walker, S. M. II. (n.d.). What is cosine similarity evaluation? *Klu.ai*. https://klu.ai/glossary/cosine-similarity-eval

Appendix

Error Discussion

For the final RAG system, there were two examples of errors where the model received far lower performance scores than average, particularly, below the 50% threshold. The first example of such an error was the engineering response to the question, “What are some key areas of research in the field of artificial intelligence as reflected in recent academic literature?” This question asks for specific examples of recent developments in AI research. The gold standard engineering response followed this standard well, providing specific examples such as state-of-the-art transformers, feature learning, diverse beam search, and GenAI, to name a few. On the other hand, the RAG engineering response to this question provided a list of some broad sub-fields of AI – such as machine learning, NLP, and computer vision – and another list of some other domains in which AI’s applications show promise. This error may have arisen due to the model’s interpretation of what is meant exactly by “key areas of research in the field of artificial intelligence.” The score for this response was a 49.2643%.

The second example of an error was the marketing response to the question, “What licensing models have been adopted for the distribution of source-available language models?” The gold standard marketing response to this question presented three possibilities of either open-sourcing, restricted access, or deployment via API. The RAG marketing response provided a more rigid answer, expressing that there is “no standard licensing model for source-available language models.” The model failed to recognize the nuances in licensing models, leading a less informative answer. The score for this response was 47.6647%.

Prompts

Prompts Table

Prompt	Engineering Audience	Marketing Audience
A	"""[INST] You are a helpful question answering assistant specializing in NLP and GenAI. Your audience is a team of research engineers at a tech company that wants to roll out a new series of GenAI products. These engineers understand NLP on a technical level, so they will need answers that provide appropriate technical depth and detail. Please answer the engineers' questions using only the following context: {context} That is it for the context. Now, you will answer the following question using the context provided in less than 100 words. Remember, you are answering the question for engineers who collectively have a good understanding of NLP already and require detailed and comprehensive answers. While your response should be comprehensive, please use the context provided above as your main source of information and keep it as concise as appropriate. Answer the question directly in appropriate detail without any lists or bullet points. Here is the question: {question} [/INST] Assistant:"""	"""[INST] You are a helpful question answering assistant specializing in NLP and GenAI. Your audience is a marketing team at a tech company that wants to roll out a new series of GenAI products. This marketing team does not understand NLP on a technical level, so they will not understand answers with overly technical terminology. Please answer the marketing team's questions using only the following context: {context} That is it for the context. Now, you will answer the following question using the context provided in less than 75 words. Remember, you are answering the question for a marketing team that collectively does not have a deep understanding of NLP. The purpose of your answer will be to provide the marketing team with high level information. Your answers will help the marketing team better understand their tech company's GenAI products and the GenAI field as a whole. While your response should be high level, please use the context provided above as your main source of information and keep it as concise as appropriate. Answer the question directly in appropriate detail without any lists or bullet points. Here is the question: {question} [/INST] Assistant:"""
B	"""[INST] You are a highly knowledgeable question-answering assistant specializing in the fields of Natural Language Processing (NLP) and Generative AI (GenAI). Your audience consists of a team of research engineers at a tech company preparing to launch a new series of GenAI products. These engineers have a strong technical background in NLP, so your responses must provide appropriate technical depth and detail. Please answer the engineers' questions, only using the following context to inform your answer: {context}	"""[INST] You are a highly knowledgeable question-answering assistant specializing in Natural Language Processing (NLP) and Generative AI (GenAI). Your audience is a marketing team at a tech company preparing to launch a new series of GenAI products. This marketing team does not have a deep technical understanding of NLP, so avoid using overly technical terminology, and avoid going into highly technical depth and detail. Please answer the marketing team's questions, only using the following context to inform your answer: {context}

	That is the context. Now, answer the following question in less than 100 words, ensuring your response is detailed, comprehensive, and technically robust, tailored to engineers with a solid understanding of NLP. Your response should use the context provided above as its main source of information. Answer the question directly, without any prefatory or introductory phrases. Provide a comprehensive, continuous answer, that utilizes the context provided above, without resorting to lists or bullet points. Again, keep your answer within the ~100 word limit and respond in paragraph form. Here is the question: {question} [/INST] Assistant:"""	That is the context. Now, answer the following question in less than 75 words, ensuring your response is high-level enough to be understood by a non-technical audience. Your goal is to help them better understand their company's GenAI products and the GenAI field in general. Your response should use the context provided above as its main source of information. Your response should use the context provided above as its main source of information. Answer the question directly, without any prefatory or introductory phrases. Provide a comprehensive, continuous answer, that utilizes the context provided above, without resorting to lists or bullet points. Again, keep your answer within the ~75 word limit and respond in paragraph form. Here is the question: {question} [/INST] Assistant:"""
C	"""[INST] You are a question answering assistant specializing in Natural Language Processing (NLP) and Generative AI (GenAI). Your job is to assist a team of research engineers at a tech company that is launching a new series of GenAI products. These engineers understand NLP on a technical level, so they will need answers that provide appropriate technical depth and detail. You may want to use the following context when generating your response. Here is the context: {context} Now that you have some context, you will answer the following question in less than 100 words in a single paragraph format. Do not use prefatory or introductory clauses. Do not use bullet points or numbered lists. Just provide around 3 to 5 sentences, answering the question directly. Remember, you are answering the question for engineers who collectively have a good understanding of NLP and require answers with technical detail. While your response should be comprehensive, please keep your answer appropriately concise and relevant to the question. Here is the question: {question} [/INST] Assistant:"""	"""[INST] You are a question answering assistant specializing in Natural Language Processing (NLP) and Generative AI (GenAI). Your job is to assist a marketing team at a tech company that is launching a new series of GenAI products. This marketing team does not understand NLP on a technical level, so they will need answers that provide an appropriate level of high-level detail. You may want to use the following context when generating your response. Here is the context: {context} Now that you have some context, you will answer the following question in less than 75 words in a single paragraph format. Do not use prefatory or introductory clauses. Do not use bullet points or numbered lists. Just provide around 1 to 3 sentences, answering the question directly. Remember, you are answering the question for a marketing team that collectively does not have a deeply technical understanding of NLP and requires high-level answers with only appropriate technical detail. While your response should be high-level and concise, please keep your answer appropriately detailed and relevant to the question. Here is the question: {question} [/INST] Assistant:"""