Using Retrieval Augmented Generation to Extract Information and Answer Questions

Rita Tu

Abstract

This study explores the potential of a Retrieval-Augmented Generation (RAG) system implemented with LangChain, focusing on comparing different large language models and configuration parameters. Specifically, we investigate the impact of utilizing Cohere and Mistral language models on the model's ability to generate relevant and accurate responses to questions. This study also examines how varying chunk sizes and chunk overlaps influence the document retrieval process and overall response quality. The results demonstrated there are trade-offs between computational costs, context coverage and coherence of the responses. Based on the findings, we recommend utilizing Cohere with a moderate chunk size and overlap to balance the trade-offs. Our research serves as a proof-of-concept for optimizing RAG systems for question answering tasks.

1 Introduction

The availability of papers and resources in the Generative AI space is growing exponentially over time as the field continues to develop at a fast pace. This rapid growth presents both opportunities and challenges for teams seeking to utilize the latest GenAI technologies to improve their internal processes. While information is readily available, sifting through the papers and websites to find relevant information to answer specific questions is laborious. The McKinsey Global Institute estimated that workers spent approximately one day each work week searching for and gathering information.[1] To address this challenge, this paper presents a proof of concept to evaluate the potential of RAG systems in automatically retrieving relevant documents to generate responses to questions.

The inputs of a RAG system include a prompt to guide the model, a document database which is stored into a vector store, a chunk size and overlap to determine how the documents are split, and a large language models (LLM) to assist with generating responses. The document database in this study includes a few papers from ArXiv on RAG and NLP, a few Lily Weng blogs that discuss Open Domain Questions and related topics, and a few relevant Wikipedia articles. Therefore, this model is equipped to answer questions about GenAI.

This study leverages the Cohere and Mistral LLMs with LangChain's framework to demonstrate how a RAG system can be tailored for different audiences. The primary objective of this study is to provide insights into the applications of RAG for document search and question answering. By focusing on the unique needs of engineering and marketing teams, this study seeks to demonstrate the potential of RAG systems to enhance productivity and support informed decision-making.

2 Key Findings

- Carefully crafted prompts with important details enhanced the quality of generated responses. Providing the audience allowed the model to craft two responses catered for a technical and non-technical audience.

- Very small chunk sizes improved retrieval accuracy by providing more precise document segments and yielding better performance, but they sometimes also led to less coherence. Very large chunk sizes improved the coherence and yielded good performance, but the answers were sometimes wordier and contained more irrelevant information.
- Moderate chunk overlap (20%) preserved context across chunks and improved the model's performance without significantly increasing processing time.
- With the chosen chunk size and overlap, Cohere more effectively captured the context of gold answers than Mistral.
- Smaller chunks increased retrieval precision but have a higher computational cost due to the higher numbers of chunks processed. Larger chunks with overlaps helped mediate the trade-offs.

3 Methodology

3.1 Technical Approach

Prompt Engineering: A carefully crafted prompt can guide LLMs to generate relevant and accurate responses. In this study, the prompt was iteratively tested against the gold answers and refined to improve the model's performance using the selected metrics discussed below.

The final prompt specifically defined the objective and audience of the RAG system and provided an example of responses tailored for the different audiences. The same prompt was used for both the engineering and marketing teams, except when defining the audience to generate team-tailored responses. As a result of this design, the engineering responses are more technical whereas the marketing responses provide a general overview while remaining engaging.

LLM: Once the RAG prompt was honed, the two LLMs, Cohere and Mistral, were evaluated against each other. The former is a large, commercial model, which is more versatile for different topics. The latter LLM is smaller and thus more efficient. However, Mistral is highly focused on specific research and niche applications,[5] therefore it may not be broad enough for this study. In comparing a large LLM with a small LLM, we can determine the scale of LLM required for this study.

Parameters: Chunking and vectorization was used to store and utilize the document collection. This study examined how the model's performance was affected with different chunk sizes and overlaps. The trade-offs were examined, and the top performing parameters were selected for the final model.

3.2 Testing and Evaluation

Metrics: This study utilizes the metrics of cosine similarity and ROUGE-L scores to provide a comprehensive understanding of similarity between the generated responses and gold answers by examining both the content and the structure. Cosine similarity offers a general content match using vector representation. A score of 0 indicates there is no similarity in the two sentences represented by vectors and a score of 1 represents substantial similarities between the two sentence.[2] The ROUGE-L score, more specifically, ROUGE-L's F1-score focuses on the longest common subsequence (LCS) and captures both the content and the order of words, which can provide additional insight on contextual information.[3] By examining the two metrics together, we can understand how well the generated responses capture the content and sequence of words of the gold answers.

Test Runs: We utilized a random 10% sample of the entire validation set for test runs, as this subset yielded metrics within 10% of the entire validation set. Subsequently, a prompt was designed through iterative testing with the validation set. As the prompt was written to be more specific, the model's performance improved. The model was then run to compare Cohere and Mistral. We also examined small and large chunk sizes and overlaps. The metrics from these test runs allowed us to understand how different parameters affected the model's performance and select parameters to optimize the results.

4 Results

This study explores how a combination of prompt engineering, LLMs, and parameters affected the performance of the RAG system using LangChain in generating responses to questions. In using the validation set with questions and gold answers, we were able to explore the combination that would yield the highest scores and most closely related answers.

When comparing the Cohere and Mistral LLMs, results illustrate that using Cohere yielded better generated responses that matched the gold answers more so than those generated by Mistral. This could be because Mistral is not general enough to yield better performance on the validation set. Additionally, this could also be due to the moderate chunk size and overlap used. A smaller chunk size would have produced more contextually dense information and possibly yielded better performance compared to the chunk size selected.

This study found that very small chunk sizes, which focus more on relevant information due to their conciseness[4], are effective in retrieving information from the stored documents. Very large chunk sizes yield results that are on par with very small chunk sizes, possibly because they can capture more context. Intermediate chunk sizes are not as effective in balancing the relevant details with noise.

No chunk overlap yielded similar metrics with an overlap of 20%. This could possibly be explained by the fact that no chunk overlap was able to capture the relevant context whereas an overlap of 20% of the chunk size balanced context and continuity across chunks. Other overlaps settings yielded lower performance, possibly because they contained too much repetitive information across chunks.

Table 1 below displays the results yielded after running the model with the entire validation set after experimenting with the parameters using a subset of the validation set. The high cosine similarity score indicates that the model was able to generate answers that capture most of the meaning and context present in the gold answers for both engineering and marketing teams. However, the low ROUGE-L score indicates the exact wording or order of words of the generated responses differed than the gold answers.

			ROUGE-L Score		Cosine Similarity Score	
Model	**Chunk Size**	**Overlap**	**Research**	**Marketing**	**Research**	**Marketing**
Cohere	2000	400	0.255	0.243	0.803	0.798

Table 1 *Results from the RAG system using the optimal parameters*

4.1 Challenges and Limitations

Challenges this model has include:

- inability to replicate the sequence of words in the gold answers, resulting in a low ROUGE-L F1-Score.
- the risk of inaccurate retrieval: if the most relevant document is not retrieved, the generated response may be less correct or useful. Additionally, the dataset is limited – if the information does not exist in the library of papers, then the response may be missing key components.

Limitations this model has include:

- The model only knows the information in the database provided and does not continuously update. Therefore, the model may not be up-to-date with the most recent research and findings in the NLP field.
- A grid search was not performed for the chunk size and chunk overlap as this study is a proof of concept to understand how these parameters might affect the model's performance.

The mentioned challenges and limitations can be observed in the engineering response to question 23, which had the worst ROUGE-L (0.118) and cosine similarity score (0.583) of all the questions in the validation set. Upon investigation, the generated response did not capture much of the content or sequence of the gold answer. On a deeper dive, the model did not retrieve information from the documents which the gold answer drew from, leaving room for model improvements.

4.2 Next Steps

For the next steps, one thing we can do is fine-tune the LLM. Fine-tuning of the LLM could improve the model's ability to retrieve better matched relevant documents to generate responses more closely aligned posed questions. We would use the latest Generative AI research papers and interviews with Generative AI experts as data for fine-tuning. With the additional data, we can perform supervised fine-tuning with the validation set. However, the trade-off between resources and the margin of improvement in the results would have to be considered as fine-tuning with more data could be computationally costly. Additionally, in this study, only the chunk size and chunk overlap were adjusted in RAG and only two LLMs were considered. In the future, other embedding models can be explored, such as OpenAI's GPT or BERT. These steps would be advised if more time and resources were to be invested in this proof-of-concept.

5 Conclusion

This model has the potential to effectively answer questions related to Generative AI for the engineering and marketing teams and should be implemented after making further improvements. While the model was able to generate answers that captured the essence of the gold answers, it was not able to effectively do so with the same verbiage. Addressing the latter would be crucial in creating an effective model.

For deployment, it would be worthwhile to consider a hybrid by reserving the LLM for the average load and using a pay-per-use deployment for the quarterly product launches, which is expected to be the peak load.

6 References

1. "The Economic Potential of Generative AI: The Next Productivity Frontier." *McKinsey & Company*, 2024, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#business-value. Accessed 2 Aug. 2024.
2. Prakash, Arjun. "Understanding Cosine Similarity: A Key Concept in Data Science." *Medium*, 27 July 2023. https://medium.com/@arjunprakash027/understanding-cosine-similarity-a-key-concept-in-data-science-72a0fcc57599. Accessed 2 Aug 2024.
3. OpenAI. ChatGPT. 2024. OpenAI, https://www.openai.com/
4. Theja, Ravi. "Evaluating the Ideal Chunk Size for a RAG System using LlamaIndex" LlamaIndex, 5 October 2023. https://www.llamaindex.ai/blog/evaluating-the-ideal-chunk-size-for-a-rag-system-using-llamaindex-6207e5d3fec5. Accessed 2 Aug 2024.
5. Azhar labs. "Exploring Mistral Large with Function Calling — Code." *Medium,* 29 February 2024. https://medium.com/ai-insights-cobet/exploring-mistral-large-with-function-calling-code-1aa0894db4d4

7 Appendix

7.1 Determining Experiment Sample Size

Model	**Chunk Size**	**Overlap**	**ROUGE-L**		**Cosine Similarity**	
			Research	**Marketing**	**Research**	**Marketing**
Cohere (5 samples)	1500	150	0.304	0.256	0.83	0.782
Cohere (all samples)	1500	150	0.257	0.241	0.807	0.801
Difference (%)			-18.2879	-6.22407	-2.85006	2.372035
Model	**Chunk Size**	**Overlap**	**ROUGE-L**		**Cosine Similarity**	
			Research	**Marketing**	**Research**	**Marketing**
Cohere(10 samples)	1500	150	0.28	0.262	0.846	0.802
Cohere (all samples)	1500	150	0.257	0.241	0.807	0.801
Difference (%)			-8.94942	-8.71369	-4.83271	-0.12484

A sample of size of 10 was used to experiment with other configurations for computational cost and time purposed. This sample size's performance was within 10% of the entire validation set.

7.2 Determining the LLM: Cohere vs Mistral

Model	**Chunk Size**	**Chunk Overlap**	**ROUGE-L F1-Score**		**Cosine Similarity Score**	
			Research	**Marketing**	**Research**	**Marketing**
Cohere	2000	400	0.272	0.29	0.842	0.815
Mistral	2000	400	0.179	0.143	0.754	0.749

Cohere was selected as the LLM in this model.

7.3 Determining Chunk Size

Model	Chunk Size	Overlap	ROUGE-L Score		Cosine Similarity Score	
			Research	Marketing	Research	Marketing
Cohere	50	0	0.2746	0.2581	0.8201	0.81
Cohere	128	0	0.264	0.218	0.802	0.773
Cohere	512	0	0.239	0.224	0.794	0.819
Cohere	2000	0	0.27	0.285	0.836	0.807
Cohere	5000	0	0.283	0.2459	0.8189	0.7851

A chunk size of 2000 was selected as a balance between maintaining context and coherence, while not inflicting too much computation overhead.

7.4 Determining Chunk Overlap

Model	Chunk Size	Overlap	ROUGE-L Score		Cosine Similarity Score	
			Research	Marketing	Research	Marketing
Cohere	2000	0	0.27	0.285	0.836	0.807
Cohere	2000	200	0.26	0.257	0.818	0.771
Cohere	2000	150	0.262	0.241	0.801	0.805
Cohere	2000	400	0.272	0.29	0.842	0.815
Cohere	2000	600	0.261	0.257	0.801	0.794

With a chunk size of 2000, a chunk overlap of 400 (20%) was selected as a balance between maintaining context and coherence, while not inflicting too much computation overhead.

7.5 Examining the Worst Generated Response

Entry for key '23':

Question: How do subsequent versions of Claude compare in terms of their likelihood to produce false statements?

Generated Response (Research): Claude, a large language model, has undergone iterations to mitigate false statement generation. Subsequent versions employ enhanced verification mechanisms, reducing false statement likelihood. Each iteration refines fact-checking, improving accuracy and reliability.

Gold Answer (Research): Claude Instant is a faster and lighter version of Claude, with an input context length of 100,000 tokens. In contrast, Claude 3 has faced criticism for its stringent ethical alignment, leading to a debate over the 'alignment tax' in AI development. Users have been refused assistance with benign requests, which has sparked discussions on balancing ethical considerations and practical functionality. This suggests that Claude Instant may have a lower likelihood of producing false statements compared to Claude 3 due to its focus on usability and performance.

Generated Response (Marketing): Claude, our large language model, has undergone several iterations, and we're proud to say that with each version, we've significantly reduced the likelihood of false statements being generated. The latest version of Claude is our most advanced yet, and it boasts a sophisticated fact-checking mechanism that verifies the accuracy of its responses. This means you can trust Claude to provide reliable and accurate information to your customers, ensuring your marketing campaigns are built on a foundation of truth and integrity.

Gold Answer (Marketing): Claude Instant is a faster, less expensive, and lighter version of Claude with a shorter input context length. Claude 3 has faced criticism for ethical alignment issues that may affect usability and performance.

Entry for key '59':

Question: Considering the structure and content of the provided text, what guidelines should be used to evaluate the effectiveness of a summary or chatbot response in this context?

Generated Response (Research): To evaluate the effectiveness of a summary or chatbot response for the engineering team, the following guidelines can be used:

- Technical Accuracy: Ensure that the response is technically accurate and aligns with the underlying mechanisms, processes, and technologies discussed in the provided text.

- Consistency: Verify that the response maintains consistency with the structure and content of the source material. This includes checking for any discrepancies or contradictions in the information presented.

- Clarity and Conciseness: Assess whether the response is clear, concise, and free-flowing. As the audience is the engineering team, a certain level of technical jargon is expected, but the response should still be communicated effectively without unnecessary complexity.

- Coverage: Check that the response covers the key points and addresses the essential aspects of the provided text. This includes evaluating if the summary or response captures the main themes, arguments, or findings presented in the source material.

- Contextual Relevance: Ensure that the response remains relevant to the context and scope of the provided text. The evaluation should consider whether the response appropriately reflects the intended focus and purpose of the source material.

These guidelines will help assess the quality and effectiveness of a summary or chatbot response, ensuring that it accurately represents the content and structure of the provided text while catering to the expectations and needs of the engineering team audience.

Gold Answer (Research): To evaluate the effectiveness of a summary or chatbot response in this context, guidelines should include assessing the faithfulness of the answer to the retrieved context, the relevance of the answer to the question, and the focus of the retrieved context. Additionally, consider using quality metrics such as answer relevancy to rank responses based on how directly they address the question and avoid redundant or incomplete information. Lastly, take into account the performance of different tasks such as summarization, citation prediction, and passage ranking to determine the overall effectiveness of the response.

Generated Response (Marketing): To evaluate the effectiveness of a summary or chatbot response for the marketing team, consider the following guidelines:

- Clarity and Conciseness: Ensure the response is easy to understand and free of jargon. It should convey complex ideas in a simple and concise manner.

- Relevance: The response should directly address the question or query, providing relevant and useful information without unnecessary details.

- Engagement and Readability: The content should be structured and written in a way that captures the interest of the target audience and encourages them to read or listen.

- Tone and Language: Use a friendly and approachable tone that aligns with the brand's voice and the audience's expectations. The language should be accessible and free of complex technical terms.

- Call to Action: Consider including a call to action (CTA) in the response, encouraging the audience to take the next step, whether it's to learn more, sign up for a newsletter, or try a product.

- Consistency: Ensure that the response aligns with the brand's messaging and positioning. It should reinforce the brand's values and promises.

- Feedback and Testing: Evaluate the response by gathering feedback from a small focus group or through A/B testing. This will help gauge the effectiveness of the message and identify areas for improvement.

- Metrics and Analysis: Define key performance indicators (KPIs) to measure the success of the response. This could include engagement metrics such as click-through rates, time spent reading, or user satisfaction scores.

By following these guidelines, you can create effective summaries or chatbot responses that resonate with the marketing team's audience and achieve the desired marketing goals.

Gold Answer (Marketing): Answer: Evaluate based on faithfulness, answer relevance, and context relevance.