## Enhanced Document Retrieval and Audience-Specific Response Generation

## For the Engineering and Marketing Team

## Author: Jack Lucas Chang

Date: April 3rd 2025

**Executive Summary**This report details the development and evaluation of a proof-of-concept (POC) system designed to enhance document search and question answering capabilities within a tech company, catering to both engineering and marketing teams. The system, leveraging Mistral’s large language model (LLM), constructs a retrieval-augmented generation (RAG) pipeline to deliver contextually relevant and audience-tailored answers from research documents. Mistral's performance only slightly surpassed Cohere in the metrics: relevance, faithfulness, groundedness, and role relevance. The result is a reasonable RAG system enabling accurate and nuanced responses for diverse audiences. I recommend implementing this system while conducting scalability testing and refining retrieval mechanisms, such as re-ranking and semantic chunking, to improve internal information access.

**Introduction**Our tech company, aiming to accelerate engineering activity and product/marketing efforts while exploring GenAI-based products, requires an efficient system to organize search and question-answering capabilities. This POC addresses this need by implementing a RAG system designed to provide accurate and audience-specific answers to questions related to Generative AI and LLM concepts. The system ingests research documents, splits them into manageable chunks, and creates vector embeddings to facilitate semantic search. Users, comprising 300 engineers and 40 marketing staff, can pose questions, and the system retrieves relevant document excerpts, which are then used to generate answers tailored to either an engineering or marketing perspective. The system is intended to support product releases by providing quick, accurate information to the relevant teams.

**Key Findings**

1. Model Performance Varies by Use Case – Mistral slightly outperforms Cohere in Relevance, Faithfulness, and Groundedness, making it the preferred choice. However, the performance gap is minimal, so both models are viable options.
2. Prompt Engineering is Important – Complex prompts with excessive instructions led to inconsistent or confused responses. Simplified, audience-specific prompts improved clarity and ensured responses aligned with engineering vs. marketing needs.
3. Chunking Strategy Affects Context Retention – Using a fixed chunk size of 256 balances processing speed and memory efficiency, but semantic chunking could further improve coherence by preserving related context.
4. Re-ranking Can Improve Response Quality – The system currently retrieves chunks without prioritization. Introducing cross-encoder re-ranking could ensure the most relevant and faithful context is used for generating responses.
5. Scalability & API Constraints Impact Feasibility – Mistral was chosen over Cohere due to API limits and free-tier constraints, but with unlimited resources, model selection should be revisited for potential improvements in accuracy and efficiency.

**Methodology: Model Selection, Test Runs and Evaluation Approach**  
The pipeline uses a chunk size of 256, which strikes a balance between processing speed and memory utilization (see Appendix B, Figure 1). This chunk size enables the model to handle data efficiently without overwhelming the system or causing timeouts, particularly when working with a limited API. The multi-qa-mpnet-base-dot-v1 embedding model was selected based on its better performance in Relevance and Marketing, where it scored significantly higher than other models. In particular, it achieved a high Relevance score of 2.468 compared to lower scores from other models, making it ideal for generating contextually accurate responses for marketing tasks. Additionally, it demonstrated strong performance in Groundedness and Faithfulness, ensuring the generated content aligns with both marketing and engineering needs (see Appendix B, Figure 2).

For the Mistral Model Parameters (see Appendix B, Figure 3 for detailed results), max\_new\_tokens=1250: allows the model to generate responses with a sufficient length. This is important for detailed answers, especially when processing open-ended questions. temperature=0.6: This parameter controls the randomness of the model’s responses. A value of 0.6 ensures a balanced level of creativity while maintaining coherence and relevance. top\_p=0.98: Top-p sampling is used to limit the token sampling to the top 98% of likely next tokens. This helps avoid irrelevant or low-probability words, improving the overall relevance and quality of the response. repetition\_penalty=1.2: This parameter penalizes the model for repeating phrases, helping to reduce redundancy in generated answers.

Initially, I experimented with more complex prompts, like the one designed for AI Engineers seeking in-depth GenAI explanations, which included specific instructions for brevity, algorithmic detail, performance metrics, and source referencing. However, these intricate prompts often led to inconsistent or confused responses from the model (see Appendix B, Figure 4). The added layers of instruction seemed to overwhelm the model's ability to generate coherent and accurate answers, highlighting the need for simpler, more direct prompts (see Appendix A).

While Mistral slightly outperforms Cohere across several key metrics, the difference between the two models is relatively small, indicating that both models are strong contenders. Mistral achieves a higher overall score of 2.9638 compared to Cohere's 2.7207, and performs better in metrics such as Relevance (2.6987 vs. 2.4968) and Faithfulness (2.8397 vs. 2.5164), but the margins are slim. Additionally, Mistral's score in Groundedness (2.9835 vs. 2.5623) and Role (3.3333 vs. 3.3072) are slightly higher, suggesting that while Mistral has a small edge, the performance gap is not drastic. Overall, both models perform quite similarly, but Mistral's marginally better results in these areas make it a slightly more favorable choice (see Appendix B, Figure 5). I have decided to use Mistral for this project primarily due to the limitations on API key usage and resource allocation in the free trial of Cohere.

**Metrics**

While traditional models like BERT excel in classification and comprehension, they lack the generative capacity to assess answer quality in the way an LLM can. LLMs, designed to generate coherent, context-aware text, are ideal for evaluating the richness, relevance, and consistency of responses. By leveraging LLMs, the evaluation aligns more closely with human-like judgment. This is essential for this task because it requires high-quality, context-sensitive responses suitable for humans.

These metrics are well-suited for evaluating an LLM as an evaluator because they collectively measure key aspects of response quality and alignment with user needs. Relevance ensures that the retrieved information directly addresses user questions, which is critical for useful and contextually appropriate responses. Faithfulness helps detect hallucinations by comparing generated answers to the gold answers, ensuring the model remains accurate. Groundedness assesses whether responses are well-supported by retrieved context, reinforcing the model's reliability. Role evaluation examines whether the model adapts appropriately to different contexts, which is especially important for assessing domain-specific responses in marketing versus engineering. Together, these metrics provide a well-rounded evaluation of an LLM's effectiveness in retrieving, generating, and aligning responses with user expectations.

**Results and Findings: Proof of Concept Functionality**The proof of concept (POC) successfully implemented a structured pipeline that generates contextually relevant responses tailored to different audiences, meeting its objectives by providing engineers with detailed technical explanations and marketing teams with high-level summaries. It balanced accuracy and creativity by fine-tuning model settings (e.g., temperature=0.6, top\_p=0.98) for relevant and engaging outputs. The POC demonstrated efficient data processing with a chunk size of 256, optimizing memory usage and processing speed under API constraints, and used the multi-qa-mpnet-base-dot-v1 embedding model for strong performance in both engineering and marketing contexts. However, issues with context retrieval resulted in some irrelevant or noisy data, impacting response quality, highlighting the need for further improvements in this area (see Appendix C). Overall, the POC showcased the potential of a scalable pipeline while identifying key areas for refinement in context retrieval.

**Lessons Learned and Challenges**Prompt complexity significantly impacts model performance, as overly intricate instructions lead to confused or inconsistent responses, reinforcing the need for simpler, more direct prompts that align with the model’s capabilities. While Mistral slightly outperformed Cohere in Relevance, Faithfulness, and Groundedness, the difference was marginal, suggesting that both models are viable depending on the use case. Ultimately, resource constraints played a key role in model selection, with Mistral being chosen over Cohere due to API limits and free-tier restrictions, highlighting the importance of scalability and cost in practical deployment decisions. Another resource constraint was the use of Colab to run experiments, which required over 250 credits along with a Pro account due to frequent session crashes on the T4 GPU. This instability made it difficult to maintain performance during long tasks, so I switched to the A100 for more power. However, this introduced a new challenge: my computer would overheat when handling the more intensive workloads. To address these resource constraints and ensure stable performance, a dedicated budget would be needed to secure sufficient credits, a Pro account, and potentially additional hardware or cloud resources for handling intensive workloads.

Non-technical insights highlighted the need to tailor content to different audience needs, where engineers require technical explanations to support implementation, while marketing teams benefit from high-level, accessible summaries. The final prompt structure successfully accommodated these different user personalities without explicitly dictating writing style. Balancing accuracy and creativity was another key challenge; the chosen settings (e.g., temperature=0.6, top\_p=0.98) ensured responses were creative and relevant, but maintaining faithfulness to source data while producing engaging, natural responses was difficult.

**Next Steps**Given more time, I would explore additional testing with other models (e.g., Claude, GPT-4, Llama) to determine if better alternatives exist for specific use cases, despite Mistral's slight edge over Cohere. Currently, retrieved chunks are used as-is, which may introduce noise or irrelevant context; implementing a re-ranking step could prioritize the most relevant information before passing it to the model. Additionally, the current pipeline uses a fixed chunk size of 256, which may split relevant information across multiple chunks, leading to incomplete or disjointed responses. Instead of using naive fixed-length chunking, I would implement semantic chunking techniques like sentence boundary detection to prevent sentences from being cut off when retrieved.

**Summary**This proof-of-concept (POC) successfully demonstrates the feasibility and effectiveness of a Retrieval-Augmented Generation (RAG) system, powered by the Mistral language model, to enhance document search and question-answering capabilities for a tech company's engineering and marketing teams. By tailoring responses to specific audience needs, the system significantly improves information accessibility and relevance, crucial for accelerating both technical development and marketing strategies, particularly in the context of Generative AI and LLM concepts.

My recommendation is to proceed with a pilot deployment for a limited user base to validate real-world performance, while conducting additional testing to refine prompt effectiveness and ensure the model consistently produces accurate, relevant responses. An architectural consideration is assessing the system’s ability to handle increased traffic, especially under API constraints as well as ensuring the deployment infrastructure is scalable and resilient to potential bottlenecks.

**Appendix A, Chosen Prompt**

"""

[INST]Please answer the question below only based on the context information provided.

\n\nHere is a context:\n{context} \n\nHere is a question: \n{question}.

\n\nMake the answer for detailed response appropriate for an engineer.[/INST]"""

"""

[INST]Please answer the question below only based on the context information provided.

\n\nHere is a context:\n{context} \n\nHere is a question: \n{question}.

\n\nMake the answer for a high level response appropriate for a marketing team and supporting staff.[/INST]

"""

**Appendix B Experiment Results**

Figure 1

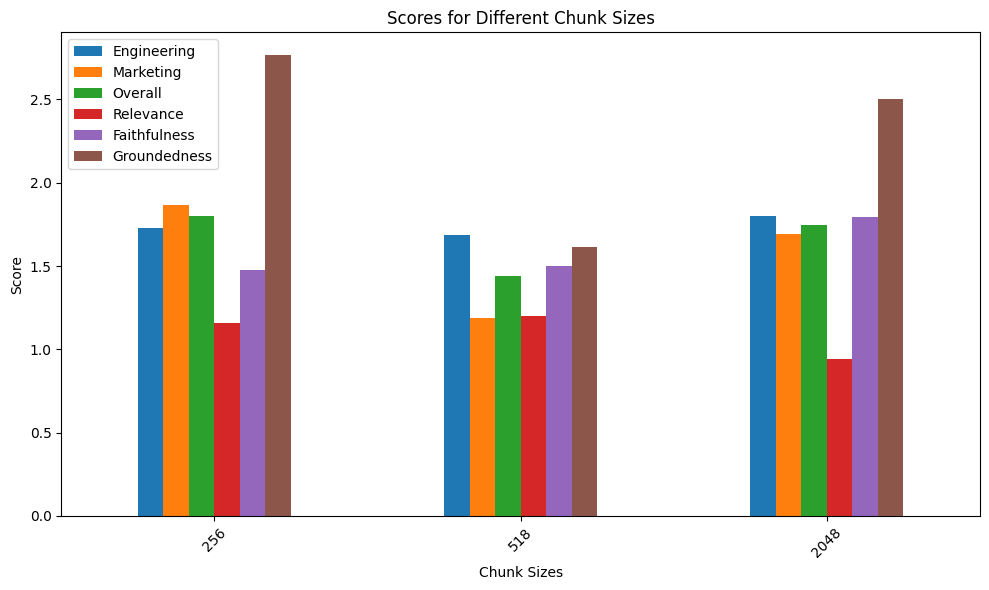


Figure 2

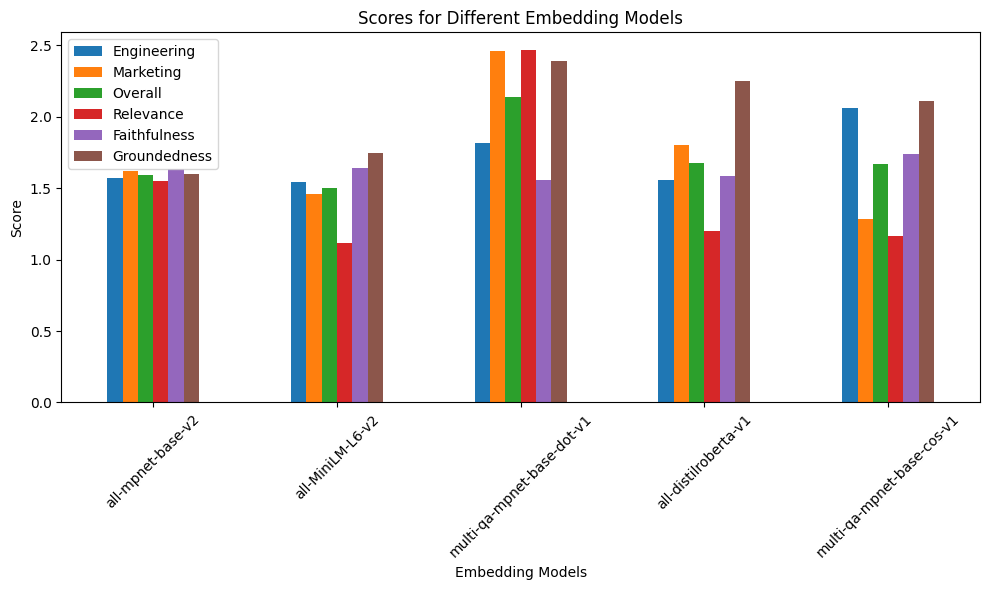


Figure 3

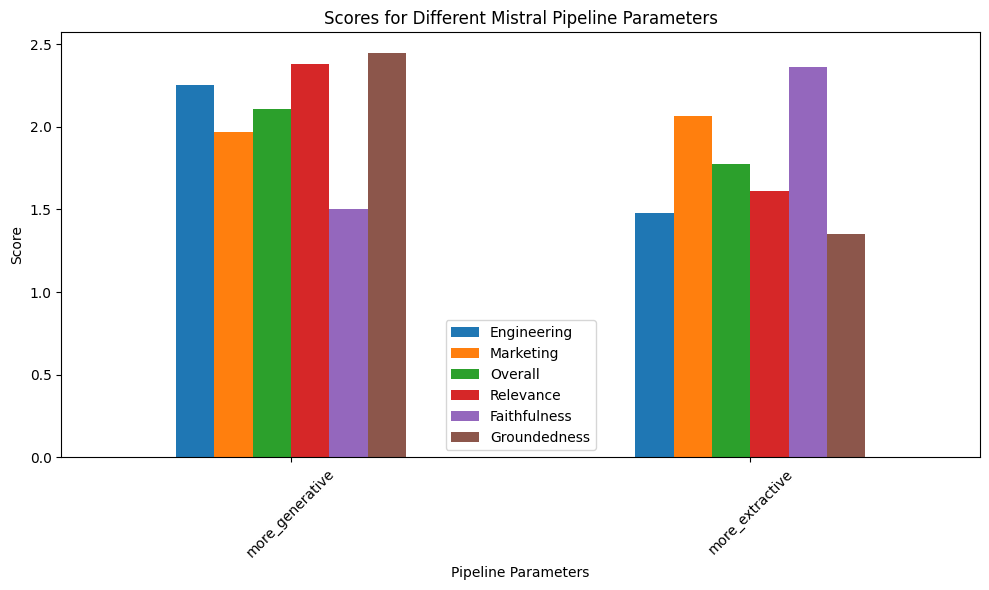


Figure 4

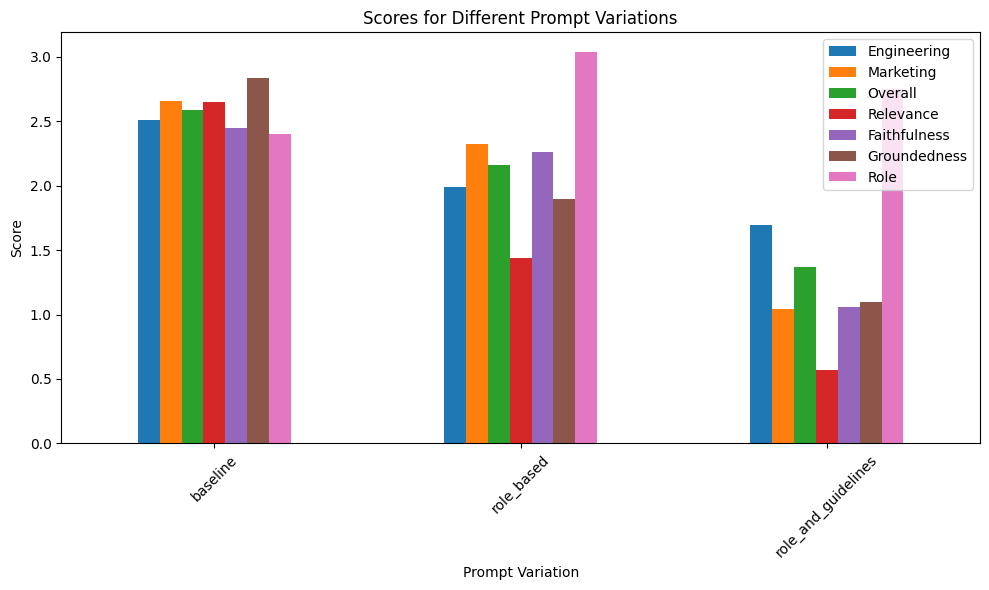
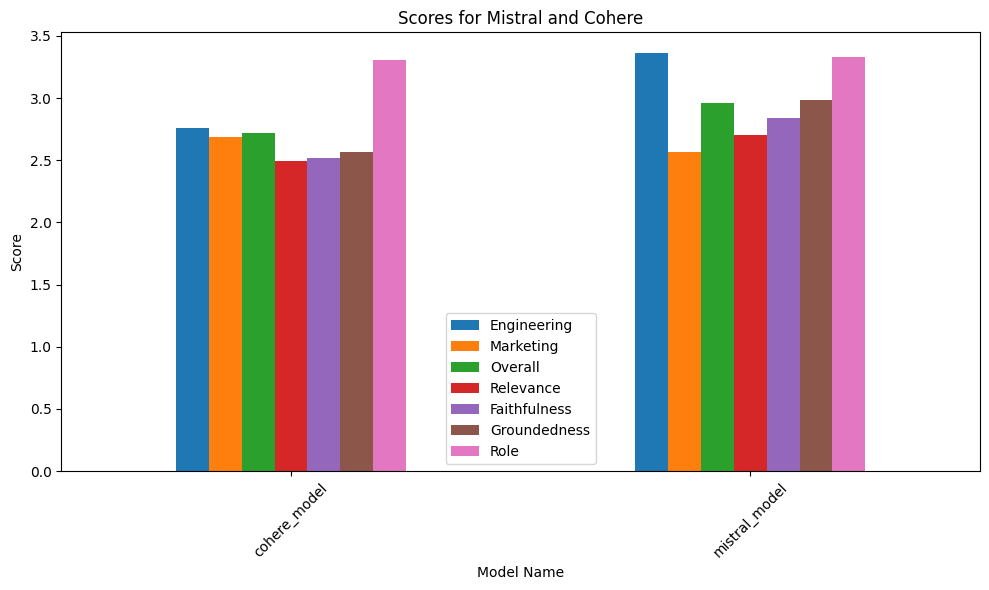


Figure 5



**Appendix C, Example question**

Question:

What are the differences between soft and hard attention in the context of image caption generation?

Context:

"Soft vs Hard Attention

This paper first proposed the distinction between “soft” vs “hard” attention, based on whether the attention has access to the

Attention is, to some extent, motivated by how we pay visual attention to different regions of an image or correlate words in

Here are a summary of broader categories of attention mechanisms:"

Engineering Answer:

"In the context of image caption generation, ""soft"" and ""hard"" attention refer to two different ways that model's attention mechanism interacts with input data.

Hard attention, also known as fixed or spatial attention, focuses on specific regions or pixels of an image by assigning equal weight to each feature within that region during caption generation. This approach does not change its focus throughout the process. It can be seen as a form of localization where the model pays attention to certain parts of the image consistently.

On the other hand, Soft attention, also called dynamic or temporal attention, allows the model to adaptively adjust its focus across the entire image sequence while generating captions. The weights assigned to various features (pixels or words) change dynamically depending on which part of the image contributes most to understanding the given context at any particular moment. This method enables more nuanced understanding and better accuracy in image-caption pairs.

As engineers working on this problem, our goal would be to understand these attention types deeply and analyze their impact on performance metrics such as BLEU scores, METEOR scores, or ROUGE scores. We may experiment with implementing both approaches, comparing computational resources required, processing time efficiency, and overall quality of generated captions. Ultimately, determining the best balance between hard and soft attention could lead to improved models for automatic image captioning systems."

Marketing Answer:

"In the context of image caption generation, ""soft"" and ""hard"" attention refer to two distinct types of attention mechanisms used in neural networks.

Soft attention allows the model to selectively focus on specific parts of an input (such as an image) while generating the output (in this case, a caption). It does so by assigning weights to each part of the input based on its relevance to the output. This results in more nuanced and accurate captions that take into account the spatial relationships within images.

On the other hand, hard attention fixes the attention mechanism on one particular location throughout the entire process of generating the output. While it may simplify the computational requirements compared to soft attention, it can result in less precise captions as they don't adapt to changes in the input.

For a marketing team and their supporting staff, understanding these differences can lead to better-informed decisions regarding which type of attention mechanism to use when developing AI solutions for tasks like image captioning. Soft attention might be preferred for applications where precision and accuracy matter most, such as e-commerce product description generation. Meanwhile, hard attention could prove useful in situations where speed and simplicity are prioritized over flexibility, such as social media post processing."

### Why the model made these choices:

### Engineering Answer

The engineering prompt explicitly asks for a detailed response appropriate for an engineer, leading the model to prioritize technical accuracy, implementation details, and evaluation metrics. The model:

* Explained attention mechanisms in depth, introducing concepts such as how attention weights are assigned and adjusted dynamically.
* Emphasized performance trade-offs, discussing metrics like BLEU, METEOR, and ROUGE that are relevant for assessing model effectiveness.
* Mentioned computational considerations, noting that engineers might experiment with both attention types to balance processing efficiency, accuracy, and resource constraints.
* Adopted a structured, analytical tone, reflecting how engineers approach problem-solving—through optimization, benchmarking, and trade-off analysis.

### Marketing Answer

The marketing prompt requests a high-level response appropriate for a marketing team and supporting staff, so the model shifts its focus to broad concepts, accessibility, and real-world applications. The model:

* Avoided excessive technical detail, ensuring that the explanation remains digestible for a non-technical audience.
* Used simpler language, explaining attention mechanisms in a way that highlights their functional impact rather than the underlying algorithms.
* Framed the explanation around practical applications, such as e-commerce product descriptions and social media post generation, making the topic relevant to marketing professionals.
* Contrasted soft and hard attention in terms of business benefits, emphasizing that soft attention improves nuance and accuracy, while hard attention prioritizes speed and simplicity.

### Conclusion

The model’s choices stem directly from the explicit instructions in the prompt. For the engineering response, it optimized for depth, precision, and technical rigor, whereas for the marketing response, it emphasized clarity, business applications, and ease of understanding. By structuring the prompt this way, the responses remain audience-appropriate without requiring additional intervention or manual rewriting.