Utilization of a RAG System to Support Engineering Research and Marketing Productivity

Kevin Chow - April 14th, 2025

Executive Summary:

This successful proof of concept (POC) demonstrates that a retrieval-augmented generation (RAG) system has the potential to increase productivity and save company money. Further refinement needs to be done to uncover this potential before deployment. Additionally, a re-examination of the need for marketing specific responses should be considered.

Introduction:

New technology utilizing large language models (LLM), specifically generative artificial intelligence (GenAI), is the future of the tech industry and human productivity. In alignment with our company's vision, this POC features a RAG system to better optimize our internal question answering and search capabilities. With this tool, I aim to elevate employee productivity by providing intelligent, insightful answers to LLM and GenAI related questions.

This initial POC is constructed utilizing the LangChain software framework [1]. Its knowledge base is a pre-selected set of LLM and GenAI related documents stored in a vector store. The system retrieves contexts related to the inputted query. Based on these, Mistral is then utilized to generate answers that deepen our employees' understanding and empower them to act. With our target audiences of engineering research and marketing in mind, specific prompt templates have been tested for ability to provide tailored information to each team.

Key Findings:

- The tuned model consistently outperforms the baseline model across audiences, demonstrating its potential and ability to provide cross-functional value.
- Prompt engineering plays an important role in response quality. It can also be used to tailor outputs to audiences and help reduce hallucinations.
- Document quality and content is crucial. Ongoing curation and updating of source documents is essential for maintaining system reliability and factual accuracy [2].
- Model tuning is ready to move beyond linear experimentation. Future testing should circle back to different parameters and fine-tune with greater detail.
- Evaluation metrics are not definitive. Re-examination of evaluation metrics may provide a clearer picture of model performance and better guide future improvements.

Methodology:

The approach to this POC is guided by three core objectives: showcase the potential of a RAG system, demonstrate the opportunity to tailor its performance to specific engineering and marketing needs, and execution of the first two objectives through a structured results-driven approach [2]. As a proof of concept, the final iteration of this system serves as evidence that those objectives can be accomplished. It is not a perfectly-tuned final product, but rather the foundation upon which a production ready RAG system can be built.

A subset of questions chosen from the provided dataset. This was done to allow for more rapid experimentation and iteration. Questions were selected through manual inspection for usefulness in evaluating different aspects. For example, question 60 was selected for its two part structure. Question 76 was selected to examine compare and contrast ability. Question 63 was selected for its differences in answers between research and marketing, namely length and technical depth. The final subset of questions is as follows: [0, 18, 19, 44, 60, 63, 76, 78].

Calculating similarity between texts is a key aspect of RAG systems. For this system, the final choice was cosine similarity. Cosine similarity is a measurement of the angle between two vectors, quantifying how similar those texts are without penalizing for differences in magnitude (i.e. text length) [2]. It is also fast to compute, scales well, and straight-forward in interpreting. The embedding model used, *multi-qa-mpnet-base-cos-v1*, was designed for semantic search though training with cosine similarity as its similarity function [3]. The vector store also utilizes cosine as its distance strategy. The evaluation metric comparing semantic similarity between retrieved contexts and answers, as well as the metric comparing semantic similarity between answers, both utilize cosine similarity of embeddings to score the results.

To showcase the potential of tailoring the RAG system for different audiences, certain technical choices were made. Two pipelines were utilized and results were analyzed separately. This allowed for a targeted assessment of answer quality. The resulting design decisions balanced the benefits of diverging versus the pragmatic implications of doing so. That is, the cost to benefit of separation was assessed. One example in the system's design is unique prompt templates for each. In other cases, the cost outweighed the benefits and engineering answer quality was prioritized. This is due to their larger size, 300 vs. 40, the effect that high-quality technical answers can have on increasing engineering productivity, and the availability of free online methods of accessing less technical answers.

Evaluation via LLMs to measure context metrics, in a manner similar to RAGAS [4], and semantic similarity was implemented, but determined to be less effective due to their non-deterministic nature; as such, these LLM-based metrics were discarded from use. The final score is a weighted average (0 to 100) of four metrics.

The first is a context quality metric, Precision at K (P@K), that utilizes the same embedding model as the vector store to quantify semantic similarity. If the resulting score is above the threshold, the context is considered relevant. An initial threshold of 0.5 for both audiences was used, determined through manual inspection. This initial threshold had a confirmed positive correlation with the semantic similarity of gold vs. generated answers. This threshold was re-evaluated after tuning to find better, audience-specific thresholds of {research: 0.55, marketing: 0.7} on a more stable pipeline and context signal (see App. A). This new threshold is a stronger measurement of context relevance, making it more useful for future evaluation and model improvements. A weight of 0.2 was given to this metric to give it a solid, but not overwhelming, effect on the final evaluation score.

The second metric is a measure of readability via the Flesch Reading Ease test. This was included, at a weight of 0.05, as a consideration for the research and marketing audiences, slightly penalizing or rewarding for answer readability. Consideration was given to comparing the Flesch Reading Ease scores between the gold and generated answers. However, this was decided against to provide a more easily interpretable metric, scoring the answer in a manner emulating an employee querying the RAG system.

The third metric is a surface level, n-gram based approach, METEOR score. It was chosen for its higher complexity, when compared to similar n-gram metrics like BLEU or ROUGE, and balancing of precision and recall. However, in acknowledgment of this metric's shortcomings in evaluating RAG answers, it is only given a weight of 0.05.

The fourth and final metric is a measure of semantic similarity between the gold answer and generated answer, calculated using the same embedding model and distance measurement as both the vector store and P@K metric for consistency and performance. Semantic similarity via cosine, weighted at 0.70, was included at a high weight for its ability to deterministically capture semantic meaning without penalizing for length.

Experiments were done through use of the question subset outlined above. The full series of tests examine embedding models, chunk size, overlap, text splitters, number of contexts retrieved, the retriever search type, re-ranker inclusion (and varying the number of contexts retrieved and returned), prompt parts, job specific prompt parts, and the LLM. At each step, the best performing configuration was selected to move forward through a mix of manual examination and analysis of evaluation metrics. While this methodology provides a structured approach, it doesn't fully acknowledge the interdependent nature of RAG parameters and components. Before tuning, the whole gold validation dataset was run for use as a baseline.

The parameters and LangChain parts tested were chosen for their strong influences on RAG response quality. Different embedding models were examined because a bad embedding model leads to poorly vectorized text pieces. Chunk size and overlap are important parameters to test to ensure useful contexts can be retrieved. The text splitter is important for the same reason. The NLTK text splitter was implemented, but rejected due to its propensity to add chunks of hugely inconsistent sizes. The number of contexts retrieved reflects the quantity of info fed to the LLM. Retriever search type also directly affects what contexts are passed. The inclusion of a re-ranker was decided upon to improve retrieved context quality. Prompt templates used can also have unique effects on responses generated and are useful in uncovering potential to tailor responses. Finally, the LLM itself plays a key role as they can vary in aspects like training datasets or task specialization, leading to each having a unique sensitivity to certain prompt wordings or structures.

Results and Findings:

Comparing the baseline and final tuned models showed improvement across the board. For the researcher audience, the tuned model demonstrated a 4.25% increase in mean score, a 1.94% increase in standard deviation, and a 7.2% increase in median score. For the marketing audience, the tuned model showed a 7.2% increase in mean score, a 7.3% increase in standard deviation, and a 10.43% increase in median score (See App. B). These successful results demonstrate RAG's potential to improve document search and question answering.

The system's successful tuning showcases the potential for future improvement and serves as a foundation for future iterations. Key insights include: evaluating specific embedding models, similarity measurements, specific values and ranges for hyperparameters, the testing of various LangChain components, prompt engineering opportunities, and considerations of different LLMs. There is a large opportunity for improvement via prompt tuning. Prompts can be reworded, restructured, and adjusted to include different aspects. This testing covered some aspects, but left many promising avenues unexplored. Furthermore, the results also shed light on the opportunity to tailor responses to specific audience needs. Each of the experiments highlight areas to investigate the benefits of further tailoring towards each audience.

This POC also revealed the importance of document quality. With the pace at which the GenAI field is evolving, documents need to be regularly evaluated and updated from a wide variety of sources to ensure response accuracy. For specific facts or definitions, quality of response is directly related to the presence of that specific information.

In cases of inadequate context, the LLM was forced to draw on its own knowledge base to answer. This can lead to responses that are out of date or incorrect. In some cases, this may lead to hallucinations rather than grounded answers. One method that helped mitigate hallucination was using the phrase “only based on the context”. It encouraged responses that acknowledged an inability to provide a grounded answer. This should be kept in mind for future iterations, as failing to provide an answer is better than providing a factually incorrect answer.

It is also important to note that the experiments performed utilized a small subset of the validation data. Future work should utilize different subsets, larger subsets, or the whole validation set. Instead of tuning individual parameters successively, future testing should directly examine the interactions between different parameters. Furthermore, each test’s chosen values focused on capturing a wide view, rather than focusing on fully optimized values, to provide a solid foundation for future fine-tuning. For example, the chunk size values chosen were [128, 256, 384, 512]. The boundary parameter values were chosen through manual inspection and prioritized chunks that were neither too small nor too large. The final chunk size, 256, is both a starting point and range for future work. Rather than settling for a local optimum set of specifications, a more detailed approach can be taken to identify the global optimum [2].

The evaluation metrics utilized are not definitive. Potential future work could re-examine the metrics and find ways to better represent answer quality. For example, the P@K relevancy threshold was re-evaluated to make it stronger. The inclusion of answer comparison in the readability metric should be explored. METEOR, though helpful as a surface level comparison, could be removed. New metrics could be added, such as the official implementation of RAGAS, if budgeted for.

Summary & Recommendations:

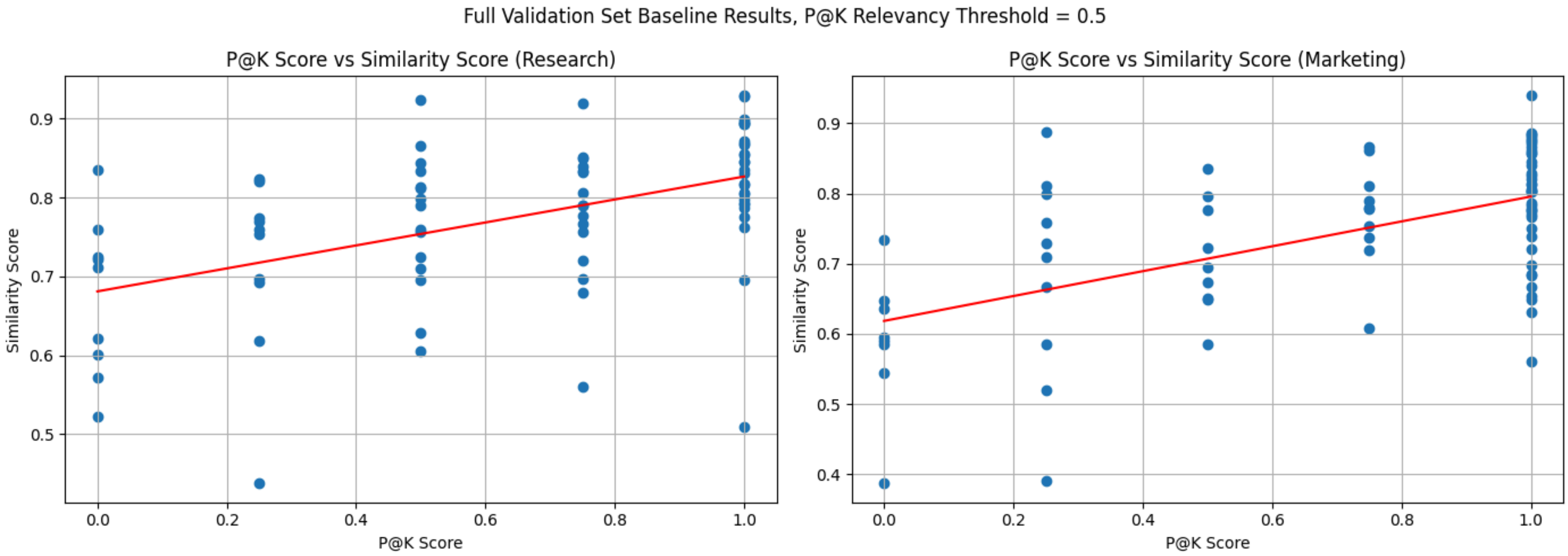
Based on these results and learnings, the recommended action is to follow this work with another POC with different testing and data, only implementing it for employee use once a more optimal configuration has been found. There is strong potential for improved productivity, cost savings, and an increase in employee understanding of our products. After a few months, the released RAG system’s user metrics should be evaluated to examine these impacts. It is also recommended to investigate the true need and benefits of designing a marketing targeted version.

While engineers require in-depth technical information to be effective, marketing may only require surface level information that is likely accessible for free through existing LLMs (such as ChatGPT) or googling, rendering any specialization for the smaller team of 40 to be moot.

References

1. “LangChain Python API Reference — LangChain documentation,” Langchain.com, 2023. https://python.langchain.com/api_reference/. Accessed: Apr. 14, 2025 [Online]
2. OpenAI. (2025). *ChatGPT (April 2025 version)* [Large language model]. Retrieved from https://chat.openai.com. Assistance from ChatGPT was used to improve word choice, sentence structure, and phrasing. The core content, analysis, and research remain the original work of the author.
3. “sentence-transformers/multi-qa-mpnet-base-cos-v1 · Hugging Face,” Huggingface.co, 2024. https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-cos-v1 Accessed: Apr. 14, 2025. [Online]
4. S. Es, J. James, L. Espinosa-Anke, S. Schockaert, and E. Gradients, “RAGAS: Automated Evaluation of Retrieval Augmented Generation,” 2024. Accessed: Apr. 14, 2025. [Online]. Available: https://aclanthology.org/2024.eacl-demo.16.pdf

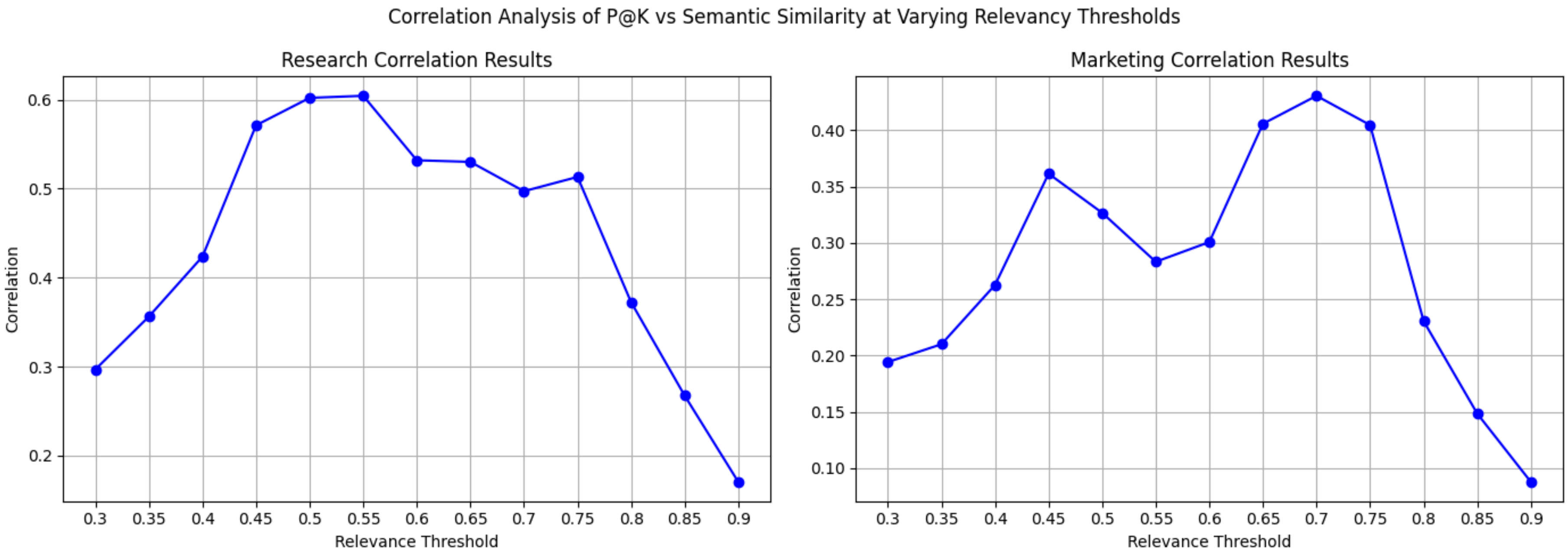
Appendix A, Initial and Improved P@K Relevancy Thresholds:



Full Validation Set Baseline Results, P@K Relevancy Threshold = 0.5

P@K Score vs Similarity Score (Research)

P@K Score vs Similarity Score (Marketing)

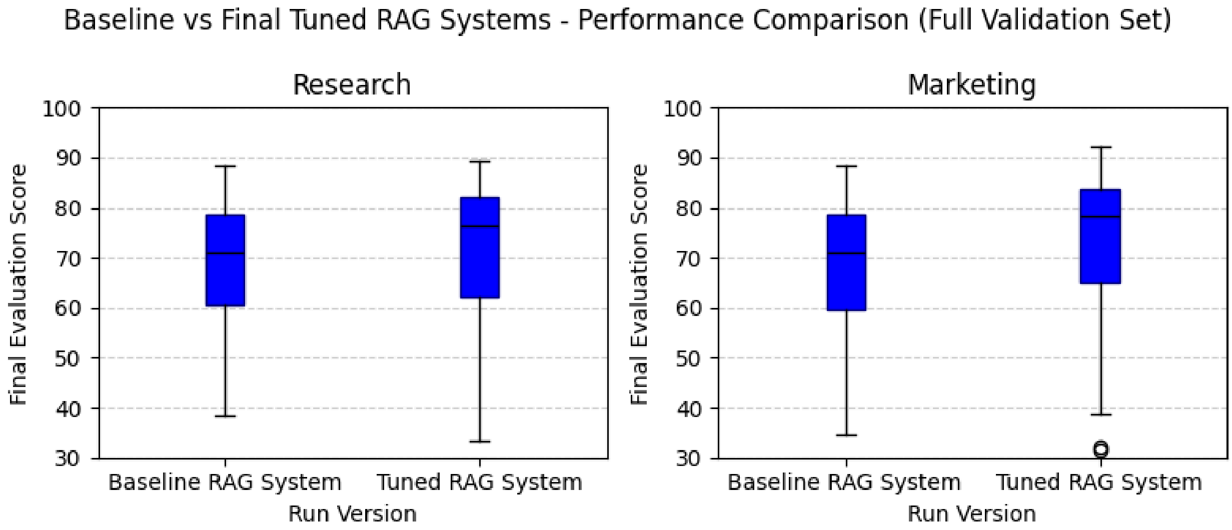


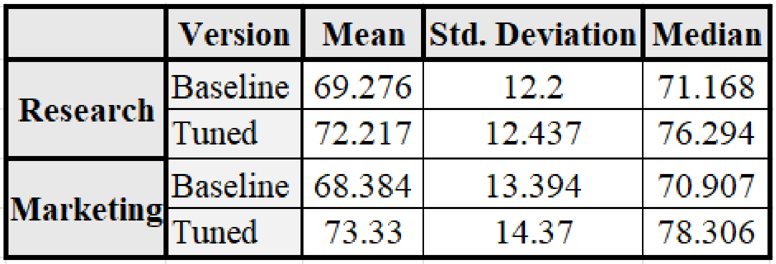
Correlation Analysis of P@K vs Semantic Similarity at Varying Relevancy Thresholds

Research Correlation Results

Marketing Correlation Results

Appendix B, Baseline versus Tuned RAG System Performance





	Version	Mean	Std. Deviation	Median
Research	Baseline	69.276	12.2	71.168
	Tuned	72.217	12.437	76.294
Marketing	Baseline	68.384	13.394	70.907
	Tuned	73.33	14.37	78.306