HIGH SCHOOL EDITION
Journal of Student Research

Volume 12 Issue 4 (2023)

A Retrieval-Augmented Generation Based Large Language Model Benchmarked on a Novel Dataset

Kieran Pichai

Menlo School

ABSTRACT

The evolution of natural language processing has seen marked advancements, particularly with the advent of models like BERT, Transformers, and GPT variants, with recent additions like GPT and Bard. This paper investigates the Retrieval-Augmented Generation (RAG) framework, providing insights into its modular design and the impact of its constituent modules on performance. Leveraging a unique dataset from Amazon Rainforest natives and biologists, our research demonstrates the significance of preserving indigenous cultures and biodiversity. The experiment employs a customizable RAG methodology, allowing for the interchangeability of various components, such as the base language model and similarity score tools. Findings indicate that while GPT performs slightly better when given context, Palm exhibits superior performance without context. The results also suggest that models tend to perform optimally when paired with similarity scores from their native platforms. Conclusively, our approach showcases the potential of a modular RAG design in optimizing language models, presenting it as a more advantageous strategy compared to traditional fine-tuning of large language models.

Introduction

The evolution of natural language processing models has seen significant strides from rule-based approaches in the early stages of language understanding, eventually leading to the advent of neural networks. However, the full potential of these neural networks awaited the computational infrastructure to catch up. The pivotal moment arrived with the emergence of neural machine translation (NMT), exemplified by Google Translate, which marked a turning point in machine language comprehension (Bahdanau, 2016). Subsequently, a plethora of advanced models, including BERT, Transformers, GPT-2, and GPT-3, have emerged, driving the field forward. Recent notable additions to this landscape are models like GPT and Bard (Devlin, 2018) (Vaswani, 2017) (Radford, 2018). While fine-tuning such models has proven to be a challenging endeavor, it has become evident that Retrieval-Augmented Generation (RAG) offers a promising alternative (Lewis, 2020) (Siriwardhana, 2023) (Yu, 2022).

Curiously, little attention has been devoted to dissecting the individual components of RAG and their respective impacts on overall performance. In response to this gap, our paper undertakes a comprehensive investigation of the RAG framework and embarks on the design of RAG models from the ground up, with a focus on the modularity and replaceability of its constituent modules. This research seeks to contribute to a deeper understanding of the mechanisms underlying RAG and its potential for enhancing natural language understanding and generation. These Large Language Models (LLMs) exhibit a remarkable proficiency in replicating human language styles, achieving a level of linguistic verisimilitude that borders on the impeccable. In light of these capabilities, it is prudent to delve into the profound significance of the Amazon rainforest, which equates to the importance of any ethnically or racially diverse nation across the globe. Within the vast expanse of the Amazon, an intricate tapestry of life unfolds, where millions of distinct species intermingle. Each of these species, as rare as the other, holds a unique and intrinsic value to the indigenous populations who have made this ecosystem their home. The Amazon rainforest is not only a cradle of biological diversity but also a sanctuary for an array of religions and cultures, many of which teeter on the brink of oblivion. Preserving the Amazon is not merely an environmental imperative; it is an act of justice to the indigenous communities whose ancestral lands are enshrined within its boundaries. It is a call to safeguard the memories of the land, the traditions that have evolved within its embrace, and the very essence of their cultures. However, certain regions of the Amazon remain shrouded in obscurity, their flora and fauna so rare that reliable and readily available information is conspicuously lacking in the vast repository of knowledge available on the internet. In this context, advanced LLMs play an instrumental role in addressing this deficit by facilitating the dissemination of indigenous narratives and thereby amplifying awareness and appreciation of the rich tapestry of beliefs, practices, and traditional knowledge that these communities hold dear. They serve as a bridge connecting the indigenous Amazonian cultures with the global community, emphasizing the paramount importance of preserving the cultural diversity interwoven within this vast rainforest. In sum, the overarching mission of this endeavor is twofold: to document and educate the Western world about hitherto unknown cultures while concurrently ensuring the enduring preservation of these invaluable facets of human heritage and biodiversity.

Proposed Experiment

Background and Importance

The intrinsic value of indigenous knowledge, especially from regions as biodiverse and culturally rich as the Amazon Rainforest, cannot be overstated. This knowledge, passed down through generations, encompasses not only cultural and religious beliefs but also practical insights into the local flora and fauna. As the modern world encroaches on these lands, this wisdom is in peril of being lost forever. Recognizing this, our proposed experiment aims to employ a state-of-the-art Retrieval-Augmented Generation (RAG) framework to capture and leverage this vast, yet vulnerable, knowledge base.

Our dataset, derived from interviews with Amazon Rainforest natives and biologists, is unparalleled in its depth and breadth. It includes detailed discussions on religious practices, cultural nuances, and the integral role of the surrounding ecosystem in the daily lives of these communities. This data is not just a scientific or anthropological resource; it is a repository of living history and an urgent call to action for preservation efforts.

By integrating this unique dataset into the RAG framework, we anticipate not only the preservation of knowledge but also the generation of responses that reflect the rich tapestry of Amazonian life. The experiment is designed to evaluate how different components within the RAG setup—such as base language models and similarity scoring algorithms—can be optimized to reflect the nuances captured within our dataset. In doing so, we aim to bridge the gap between advanced language models and the profound human insights found within the Amazon.

The central objective of our experiment is twofold: to analyze the performance implications of modular design within the RAG framework and to demonstrate the profound capability of such a system to preserve and communicate the wealth of indigenous knowledge. We hypothesize that a customizable RAG model will not only facilitate a deeper understanding of the data but also allow us to fine-tune the system for optimal performance across different configurations. To achieve this, we will systematically explore the interchangeability of various RAG components. We will assess different base language models such as GPT and Palm and compare the efficacy of similarity scoring tools from diverse platforms. The experiment will rigorously test these combinations, identifying which synergies most effectively capture the essence of the dataset.

The ultimate goal is to showcase the potential of a modular RAG system in processing culturally significant information, paving the way for future applications that can benefit from such tailored language models. We anticipate that our findings will contribute significantly to the fields of computational linguistics and cultural preservation, demonstrating a novel approach to the application of large language models.

Source and Composition

Our proprietary dataset stands as the cornerstone of this experiment. It is a rich compendium of verbal histories, interviews, and ecological insights gathered from the indigenous peoples of the Amazon Rainforest, as well as from biologists and ecologists dedicated to studying this unique biome. The dataset is characterized by its diversity, comprising narratives that elucidate the intricate relationship between the natives and their environment, including the religious and cultural significance of plant and animal life.

The data collection was an extensive process, where linguists and researchers engaged in deep conversations with the natives, recording their dialects, translating their stories, and documenting their knowledge of the ecological system. Similarly, biologists contributed their decades of research on the flora and fauna, providing a scientific perspective to the indigenous narratives. The data thus forms a confluence of traditional wisdom and modern scientific understanding, offering a 360-degree view of the Amazon Rainforest's ecosystem.

Cultural and Environmental Significance

The urgency of preserving indigenous knowledge is akin to conserving an endangered species. It is a race against time, as globalization and environmental degradation threaten to erase unique cultures and the wisdom they hold. Our dataset serves as a digital ark, a means to preserve and perpetuate the knowledge that has sustained the Amazon's communities for millennia.

The environmental significance of the Amazon Rainforest cannot be overstated—it is a keystone of global biodiversity. By documenting the intricate knowledge, the natives have of their environment, we are also chronicling the ecological interdependencies that are vital for the rainforest's survival. This dataset, therefore, is not just an academic or technological asset; it is a critical record for environmental conservationists and policymakers.

Through our experiment, we aim to amplify the voices of the Amazon's indigenous peoples, whose understanding of their habitat is unmatched. By integrating their knowledge into the RAG framework, we hope to create a model that not only responds with information but also with wisdom that respects the interconnectedness of life and culture.

Retrieval-Augmented Generation Framework

The heart of our experiment lies in the Retrieval-Augmented Generation (RAG) framework, a sophisticated algorithm that enables the deconstruction of the language model into discrete, interchangeable components. This framework integrates a retriever model that sources relevant context and a generator model that synthesizes the retrieved information into coherent responses.

In mathematical terms, given an input query q, the retriever model searches a knowledge base $\mathcal{K}$ and retrieves a set of relevant documents $D = \{d_1, d_2, \ldots, d_k\}$. Each document d is represented as a vector $\mathbf{v}_d$ in a high-dimensional space, obtained from an embedding layer. This process transforms the raw text data into a structured form amenable to computational manipulation.

To examine the effects of component interchangeability, we adopt various base language models and similarity scoring mechanisms. For instance, if Edenotes the embedding function, and s and t represent the input and target text sequences, respectively, their vector representations would be $\mathbf{v}_s = E(s)$ and $\mathbf{v}_t = E(t)$. We employ cosine similarity as the basis for our similarity score, defined by the formula:

$$\text{similarity}(\mathbf{v}_s, \mathbf{v}_t) = \frac{\mathbf{v}_s \cdot \mathbf{v}_t}{||\mathrm{v}_s|| \, ||\mathrm{v}_t||}$$

Here, $\cdot$ denotes the dot product between the two vectors, and $||\cdot||$ denotes the Euclidean norm. This score quantifies the closeness of the semantic meaning represented by the vectors, with a value of 1 indicating identical directionality and thus, maximal similarity.

The experiment tests different configurations by substituting E with embedding functions from various models (e.g., GPT, Palm), allowing us to discern the impact of the embedding layer on the final similarity score. By comparing the performance of different E choices, we can identify which embeddings yield the most semantically rich representations for our unique dataset.

Experiment Setup

The experiment commences with the training of the language models using our unique dataset. For the training phase, we define the following:
$\mathcal{L}$: The base language model, which can be either GPT or Palm.
$\mathcal{D}$: The training dataset, consisting of pairs (q_i, a_i) where q_i is a query from the dataset and a_i is the corresponding answer.

The language model $\mathcal{L}$ is fine-tuned on $\mathcal{D}$, optimizing the weights to minimize the loss function, typically a cross-entropy loss between the predicted and actual answers.
Following training, the question-answering process involves feeding a new query q' to the trained model and retrieving the answer a'. This answer is then compared to a predefined list of correct answers using the similarity score, which is fundamental to evaluating the model's performance.

Benchmarking and Evaluation

The evaluation metric for our experiment is based on the similarity scores between the generated responses and a set of reference answers. Let $A = \{a'_1, a'_2, \dots, a'_n\}$ be the set of answers generated by the model, and $A_{\text{ref}} = \{a^*_1, a^*_2, \dots, a^*_n\}$ be the set of reference answers. We define the average similarity score as follows:

$$\text{Score}_{\text{avg}} = \frac{1}{\text{n}} \sum_{i=1}^{n} \text{similarity}(\mathbf{v}_{a'_i}, \mathbf{v}_{a^*_i})$$

This average score acts as the primary benchmark for comparing different model configurations. We systematically record the scores across various combinations of language models and similarity scoring mechanisms to assess which configurations yield the highest average similarity, indicating the most effective model setup for our dataset.

Additionally, we account for the presence or absence of context in the model's training and response generation. This is critical, as the presence of context has been shown to significantly influence model performance, particularly in the domain of indigenous knowledge and biodiversity, where context provides essential background information that can drastically affect the meaning and relevance of a response.

Through this meticulous experimental setup, we aim to illuminate the intricate dynamics between different components of the RAG framework and their collective impact on the model's ability to accurately replicate and convey the richness of the Amazon Rainforest's cultural and ecological knowledge.

Pre-Experiment Performance Expectations and Discussion

In the landscape of varying configurations, we hypothesize that certain setups will yield higher average similarity scores than others, indicative of more nuanced and accurate language generation. Particularly, we expect that:
The similarity scores for models trained with contextual data $\mathbf{v}_{\text{context}}$ will surpass those trained without, due to the enriched understanding and background the model has of the subject matter.

When aligning models with their native embeddings (e.g., GPT with OpenAI Embed, Palm with Palm Embed), the semantic vector representations $(\mathbf{v}_s, \mathbf{v}_{t)}$ should align more closely, thus producing higher similarity scores.
The modular nature of the RAG setup will reveal that certain combinations of base language models and similarity scoring mechanisms are more effective than others, depending on whether context is included or not.

We denote the expected performance increase due to context as Δ_{context}, and the alignment of native embeddings as Δ_{native}. Mathematically, we can represent our hypothesis as:

$$\text{Score}_{\text{avg,context}} > \text{Score}_{\text{avg,no context}} + \Delta_{\text{context}}$$
$$\text{Score}_{\text{avg,native}} > \text{Score}_{\text{avg,non-native}} + \Delta_{\text{native}}$$

These hypotheses will be tested through a series of experiments, allowing us to determine the optimal model configuration for processing and generating responses reflective of the Amazon Rainforest dataset.

Potential Implications for LLMs

The results of this experiment are expected to have significant implications for the development and fine-tuning of Large Language Models (LLMs). By identifying the most effective configurations, we can offer insights into the adaptability of these models to specialized datasets, which is crucial for applications that require a high degree of cultural and contextual sensitivity.

Moreover, the experiment is poised to challenge the prevailing approach to LLM training and fine-tuning, which often relies on static, one-size-fits-all models. Our findings could suggest a shift towards a more dynamic, component-based approach, allowing for greater flexibility and precision in model performance across diverse domains.

The potential success of the RAG framework in this context may also pave the way for more granular improvements in LLMs, beyond the standard metrics of accuracy and fluency. It may, for instance, enhance the models' ability to engage with and preserve less-represented languages and dialects, fostering greater inclusivity and diversity in the realm of natural language processing.

Implications for Indigenous Knowledge Preservation

The significance of our experiment extends beyond the technical accomplishments within the field of natural language processing. It serves as a testament to the power of advanced computational techniques in preserving the rich tapestry of human culture, particularly the imperiled knowledge of the Amazon Rainforest's indigenous peoples. By successfully training a language model to accurately reflect and communicate this knowledge, we not only preserve it for future generations but also validate the importance of linguistic and cultural diversity in our global ecosystem.

This experiment, should it succeed, will demonstrate a practical application of LLMs in the service of cultural preservation. It emphasizes the role that technology can play in safeguarding intangible heritage, a mission that aligns with the broader objectives of UNESCO's Intangible Cultural Heritage initiatives. It serves as a model for how communities around the world can leverage technology to protect and share their unique cultural identities and knowledge systems.

Advancements in RAG Framework

From a methodological standpoint, our experiment is poised to contribute to the advancement of the RAG framework within the realm of AI language models. By dissecting the RAG components and examining their interplay, we will gain insights into the mechanics of modular design in language models, offering a blueprint for future research and development.

The outcomes of this experiment could lead to the evolution of RAG into a more nuanced and adaptable framework, one that can be customized for specialized datasets and applications. This adaptability is critical as the demand for LLMs expands into increasingly varied and complex domains, from legal and medical to historical and anthropological.

Furthermore, the experiment's focus on modularity could inspire a new wave of research into component-based architectures for LLMs. Such architectures may provide a more sustainable and efficient pathway to model improvement, as opposed to the computationally intensive process of training large models from scratch.

In conclusion, the proposed experiment holds the potential to make significant contributions to both the field of AI and the preservation of human cultural heritage. The insights gained could lead to a more inclusive and representative future for LLMs, where the voices of all communities can be heard and understood.

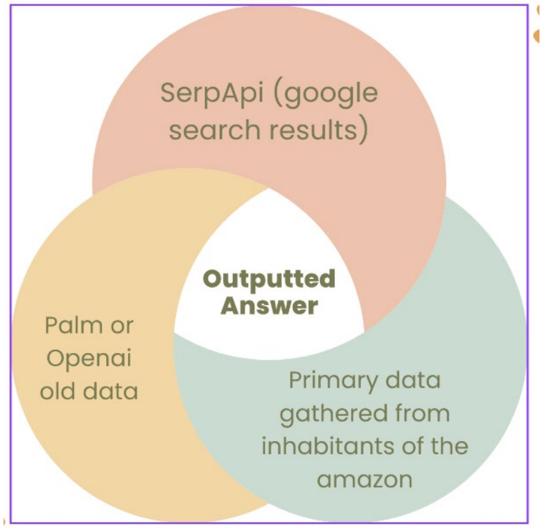
SerpApi (google search results)
Outputted Answer
Palm or Openai old data
Primary data gathered from inhabitants of the amazon

Figure 1. Venn Diagram of Data Sources for RAG. This figure represents a venn diagram of 3 sources of information (google search results, OpenAI/Palm, proprietary data collected by the author) combined in order to create the "outputted answer."

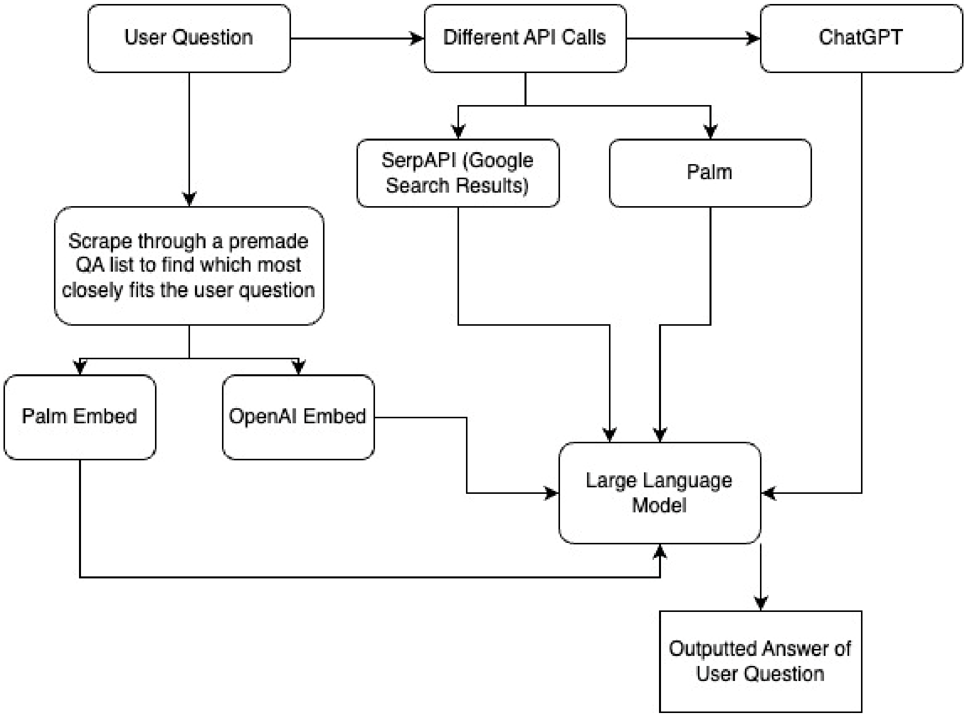
User Question
Different API Calls
ChatGPT
SerpAPI (Google Search Results)
Palm
Scrape through a premade QA list to find which most closely fits the user question
Palm Embed
OpenAI Embed
Large Language Model
Outputted Answer of User Question

Figure 2. Executive Diagram of Proposed RAG. This diagram outlines the various steps and procedures of the RAG algorithm from the input of the "user question" to the "outputted answer of the user question."

Experimental Results and Discussion

The purpose of this section is to lay down the different steps and customizations used within our experiment in order to demonstrate the conclusive results of this experiment to the reader; our experiment using a RAG methodology accurately shows how each component of a LLM positively or negatively affects the accuracy of the outcome itself.
In the initial world of LLM, in order to incrementally increase its performance engineers of these models would have to fine tune them then retrain which took immense amounts of power and large amounts of null results. However, now as they become more and more complex to tune models like OpenAI's GPT and Google's Bard have been plateauing performance wise.

Table 1. Experimental Results. This table represents the various different combinations of LLM components with respect to the average similarity score they each produced.

	Context		LLM		Embed for Similarity Score		
	Yes	No	GPT	Palm	OpenAI Embedding	Palm Embedding	Score
1		x	x		x		0.75
2	x		x			x	0.92
3	x			x		x	0.93
4		x		x		x	0.88
5	x		x		x		0.997
6	x			x	x		0.996
7		x		x	x		0.91
8		x	x			x	0.897

This paper produces a new solution to the slowing improvement of LLM in the form of RAG, a way to componentize the models and break them down into smaller sections. This allows the user to add certain parts / combinations to test the performance of those then to substitute different modules in to see which leads to the largest performance increase over the other. These customizable steps allow you to see minute differences in performance that slowly tuning a model couldn't have shown you previously. This is a novel way to approach the tuning of LLM and will only serve to increase their accuracy as time moves forward.

Another major component of our RAG methodology is the ability to switch out which embedding layer you use. The standard embedding (OpenAI) or Palm's embed. When choosing between both of those some tradeoffs are made.

When using no context, Palm's embedding layer seems to perform much better across the board, allowing for a much higher average similarity score, however this drastically shifts when given context as now OpenAI's embedding layer performs much more soundly. The evidence for these claims is discussed later in this paper.

Additionally, another beneficial feature of the RAG optimization and breakdown style is the ability to customize which similarity score the LLM uses to decide which answer to base its response off from the Q-A list. To go into further detail, the code when prompted with a user question compares the user question to the Q-A list and reorders the list based off highest similarity score to lowest, this allows the LLM to select the top 2-3 answers to the highest ranked questions and continue generating its own response from there.

The first choice of similarity score was STS OpenAI Score while the second was STS Palm Score. In terms of the data when GPT (for the purposes of precision all of the following results include context) and STS OpenAI were combined, you got an average similarity score of 0.997. If you instead pair this with Palm STS Score instead, the average score drops to 0.92, a 0.077 decrease in performance. A similar effect when using Palm with Palm STS and Palm with OpenAI STS (0.996 versus 0.93, respectively). This data demonstrates that both Palm and OpenAI are able to reach very high accuracy levels when paired with a similarity score calculated from the same program (this means GPT worked better with OpenAI STS Score and that Palm worked better with Palm STS Score). What is also interesting to note is that although Palm produced a 0.001 lower performance than GPT it seemed to be more flexible, working better with its competitor (OpenAI STS Score + Palm produced 0.93) than the GPT with its competitor (Palm STS Score + GPT produced 0.92).

To recap on the experimental setup, this code uses an interchangeable piece of a LLM so you can swap or replace things like the embedding layer used, the similarity score used, and the base language model used, also whether it was given context from the Q-A database or not. This is a novel and important way to be able to break down LLM and the data collected speaks a lot to the importance of each aspect of an LLM.

In terms of expected results, two things were noticed, first that Palm had a slightly lower performance than GPT (0.996 versus 0.997 respectively) on their top runs, however, there was also some contradictory data as it seemed that Palm worked significantly better when given no context to work with compared to GPT, Palm produced an average score of 0.88 and 0.91 when given no context while GPT produced an average score of 0.75 and 0.897. Although GPT may perform much better when given context, Palm seems to beat it out just given its own proprietary dataset (no context).

Some unexpected results occurred with pairing GPT and Palm with their opposite embeds, for example, pairing Palm with OpenAI embed. While on paper it makes sense that Palm would work better with Palm embed, it actually performed better when paired with OpenAI's embed. 0.91 (with OpenAI embed) versus 0.88 (Palm embed)—note that this is without context given. A similar effect was noticed going the other way around as well, without context, GPT performed much better with Palm embedding layer than with its own OpenAI embedding layer (0.897 versus 0.75 respectively). This data shows how Palm embedding layer tends to perform much better given no context when compared to OpenAI's embed. Similarly, to explored above, it is the opposite when given context, however. OpenAI's embedding layer performs a bit better when given context across the board than Palm embed.

Most notably from this experiment was two realizations. First, that Palm Embedding layer tends to work much better when just given its own proprietary dataset (and no context), when compared to OpenAI's embed. Additionally, when given context, the playing field switches: Palm tends to perform much worse when given context when compared to OpenAI. Lastly, it is important to note that a combination of Palm/GPT with OpenAI's embedding layer and context yielded extremely accurate results when its similarity scores were averaged.

Conclusion

Building upon the foundation laid by our initial findings, it is paramount to recognize the exceptional performance of the Palm model when utilizing its proprietary dataset in conjunction with the Palm embed. This specificity in data and technology synchronization has shown that Palm outshines OpenAI in terms of model accuracy in a context-free environment. However, the landscape shifts when contextual data is integrated. In such scenarios, the combination of GPT with its native OpenAI embedding layer excels, leveraging the additional context to produce responses of remarkable accuracy that resonate with the cultural and ecological nuances of the Amazon.

This pivot in performance based on context underscores the significance of tailored datasets and embedding mechanisms in the optimization of Large Language Models (LLMs). The adaptability of the Retrieval-Augmented Generation (RAG) framework emerges as a cornerstone for future enhancements in LLMs. By enabling the seamless interchange of model components, RAG presents an evolutionary leap in the fine-tuning of language models, catering to the intricate demands of culturally rich and contextually complex datasets.

In light of these advancements, our research signifies a pivotal moment for LLMs. The evidence suggests that when models are finely tuned with an awareness of the dataset's inherent context and the corresponding embedding layers, they reach new heights of linguistic precision. Therefore, the path forward for LLMs lies in embracing the modular and contextually aware RAG framework, which promises to refine the capabilities of language models to an unprecedented degree, ensuring the preservation and celebration of the world's diverse linguistic and cultural heritage.

Limitations

The results of the "outputted answer" of this algorithm largely reflect the quality of the data. If the LLM is trained off low-quality data, then the answer will reflect this bias. The results of the experiment will fluctuate with different results should a different dataset be used to train the LLM.

In the future there is a lot of potential to expand on this research by breaking down the RAG algorithm into even more separate components to further see the differences in average similarity score that adding or removing each component makes.

References

Bahdanau, D. C. (2016). End-to-end attention-based large vocabulary speech recognition. *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4945-4949.

Devlin, J. C. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805.

Lewis, P. P. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 9459-9474.

Radford, A. N. (2018). Improving language understanding by generative pre-training. *OpenAI.*

Siriwardhana, S. W. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 1-17.

Vaswani, A. S. (2017). Attention is all you need. *Advances in neural information processing systems*.

Yu, W. (2022). Retrieval-augmented generation across heterogeneous knowledge. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 52-58.