IMPROVING QUESTION-ANSWERING CAPABILITIES IN LARGE LANGUAGE MODELS USING RETRIEVAL AUGMENTED GENERATION (RAG): A CASE STUDY ON YORUBA CULTURE AND LANGUAGE

Adejumobi Joshua*
Department of Computer Science
Federal University of Agriculture Abeokuta
aesther661@gmail.com

ABSTRACT

This study addresses the phenomenon of hallucination in large language models (LLMs), particularly in GPT-3.5 turbo, when tasked with processing queries in Yoruba—a low resource language. Hallucination refers to the generation of incorrect information, often occurring due to the model's unfamiliarity with specific content or languages not extensively covered during its pretraining phase. We propose a novel methodology that incorporates Retrieval-Augmented Generation (RAG) techniques to mitigate this issue. Our method utilizes an exclusive dataset derived from a Yoruba-centric blog, covering an array of subjects from the language's learning resources to its folklore. By embedding this data into an open-source chroma database, we improve GPT-3.5 turbo's ability to deliver responses that are not only linguistically and factually correct but also resonate with the cultural nuances of the Yoruba heritage. This enhancement marks a significant step towards the creation of a chatbot aimed at promoting and disseminating knowledge about the Yoruba culture and language.

1 INTRODUCTION

Addressing the issue of hallucination is crucial in the field of Generative Question Answering (GQA), where the primary objective is to furnish users with accurate and factual responses. The presence of hallucinated content in answers not only misleads users but also significantly undermines the reliability of GQA systems.(Ji et al., 2023). To mitigate this, research has introduced innovations such as the Knowledge-Enriched Answer Generator (KEAG) by (Bi et al., 2019), which synthesizes answers by weaving together facts from multiple sources. Similarly, (Li et al., 2021) developed the Rationale-Enriched Answer Generator (REAG), enhancing answer accuracy by incorporating rationale extraction at the encoding phase, thus ensuring the decoder bases its response on both the extracted rationale and the initial input. Despite significant advancements, including contributions from (Fan et al., 2019), (Krishna et al., 2021), (Nakano et al., 2021), and (Su et al., 2022), the evaluation of GQA systems frequently involves human judgment to ascertain the veracity of answers, essentially correlating factual accuracy with a lower propensity for hallucination.

This research specifically investigates the challenge of hallucinations in GPT-3.5 turbo when responding to queries in Yoruba, a low resource language, utilizing the RAG framework introduced by (Lewis et al., 2020). We compile a dataset imbued with the cultural essence of Yoruba from a dedicated blog and integrate this into a chroma database, thereby elevating GPT-3.5 turbo's performance to produce responses that are factually correct. Similar to the broader field of GQA, this study employs human evaluation to measure the model's output, as the best automatic metric to evaluate this work has not been decided upon, for example RAGAs efficiency on yoruba data has not been fully determined.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies. Funding acknowledgements go at the end of the paper.

2 DATA COLLECTION AND PROCESSING

The dataset for this study was collected through a web scraping process from `https://www.theyorubablog.com/`, utilizing the Beautiful Soup library in Python for parsing HTML content. The website is dedicated to the preservation and promotion of the Yoruba culture, we specifically extracted text related to cultural traditions, learning the basics of the language, Yoruba Culture, food, and Names. Following the collection, the raw data underwent a series of preprocessing steps to ensure its suitability for analysis.

3 RETRIEVAL AUGMENTED GENERATION

3.1 CHUNKING AND EMBEDDING

After saving the scraped Yoruba blog data to a text file, we utilized Langchain's recursive text splitter to divide the content of the website into chunks. This helps to improve the processing efficiency and prepare the data for the model to be used as gpt 3.5 turbo has a maximum token length 4,096. The recursive text splitter's ability to keep related text together also helped maintain the coherence and cultural relevance of the information. To effectively handle the Yoruba language, we employed the Language-Agnostic BERT Sentence Embedding(LaBSE) model by Google. Given LaBSE's training on a diverse set of languages, including low-resource ones, it helped in creating embeddings that capture the semantic nuances of the Yoruba sentences. The embeddings generated were then stored in a Chroma database providing a structured and efficient storage solution.

3.2 RETRIEVAL AND GENERATION

The retrieval mechanism within a Chroma database is predicated on the similarity of embeddings. When a user poses a query, Chroma transforms this query into an embedding using the same model or framework specified for the stored data, in our case LaBSE. The cosine similarity between the query embedding and the embeddings stored within the specified collection is then computed. The system retrieves the entries with the highest similarity scores, effectively using embeddings as a bridge between raw textual or unstructured data and the user's informational needs. The LangChain Retrieval QA module was used for generation, this module is an advanced framework for building question-answering systems that leverage large language models (LLMs). Upon retrieving relevant documents from the chroma database, the retrieved content is carefully merged with the initial user query, crafting an enriched prompt. This enriched prompt is subsequently fed into an LLM, such as OpenAI's GPT-3.5 turbo, which adeptly synthesizes the integrated insights to formulate a response. This process benefits from the detailed context provided by the retrieved documents, as well as the model's extensive pre-trained knowledge, to generate a nuanced and informed reply.

4 EXPERIMENTAL RESULTS, CONCLUSION AND FUTURE WORK

Upon integrating the Retrieval-Augmented Generation (RAG) framework, the performance of GPT in responding to certain queries showed noticeable improvement, as assessed by human evaluators comparing GPT's outputs with and without the implementation of RAG. This evidence supports the hypothesis that RAG can indeed mitigate the issue of producing erroneous or irrelevant content, particularly in responses to queries on specific contents in low resource languages. For future endeavors, expanding the database to include audio data from the site is proposed to enrich the dataset further. Additionally, combining RAG with fine-tuning techniques and employing a well-defined automatic metric for assessment are suggested to enhance the evaluation process and overall system performance.

REFERENCES

Bin Bi, Chen Wu, Ming Yan, Wei Wang, Jiangnan Xia, and Chenliang Li. Incorporating external knowledge into machine reading for generative question answering. *arXiv preprint arXiv:1909.02745*, 2019.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*, 2019.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. Hurdles to progress in long-form question answering. *arXiv preprint arXiv:2103.06332*, 2021.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, and Songfang Huang. Addressing semantic drift in generative question answering with auxiliary extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 942–947, 2021.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Read before generate! faithful long form question answering with machine reading. *arXiv preprint arXiv:2203.00343*, 2022.

A Appendix

The images below shows some of the queries posed to the model, with and without RAG.

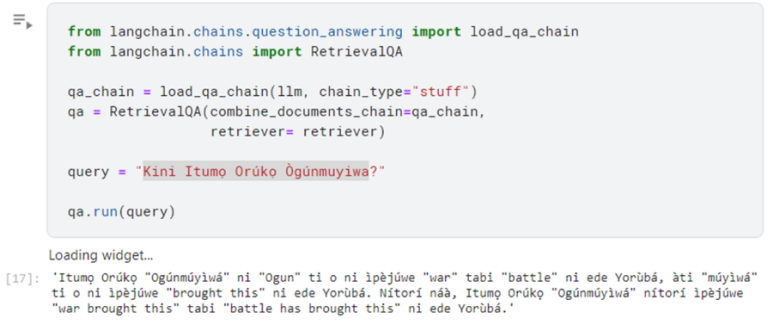

Figure 1: With RAG

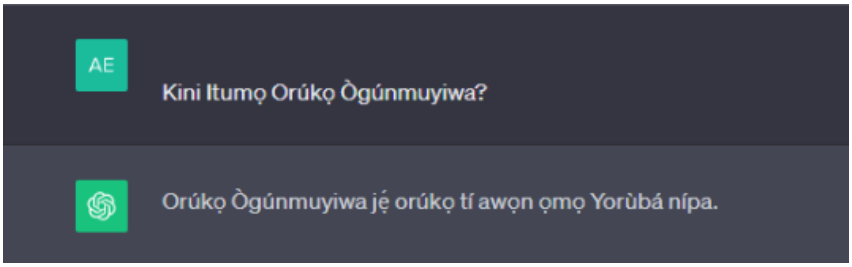

Figure 2: Without RAG

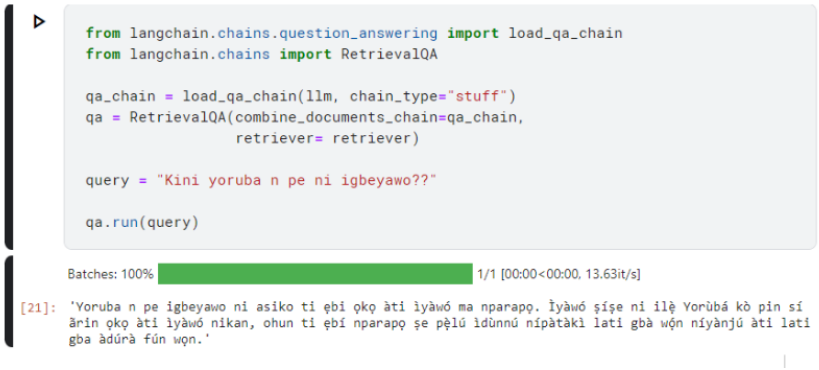

Figure 3: With RAG

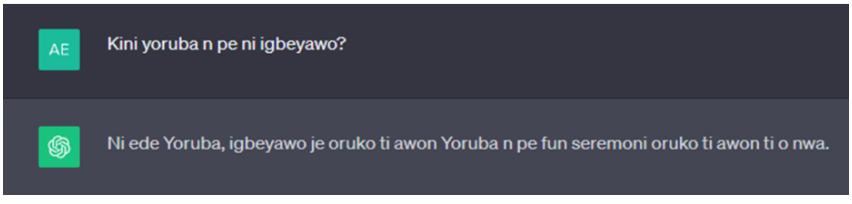

Figure 4: Without RAG