Evaluating Retrieval-Augmented Generation for Medication Question Answering on Nigerian Drug Labels in Yorùbá

Aramide Adebesin[1], **Zainab Tairu**[1]

[1]Obafemi Awolowo University

{akadebesin, zainabtairu}@student.oauife.edu.ng

Abstract

Large Language Models (LLMs) have the potential to improve healthcare information access in Nigeria, but they risk generating unsafe or inaccurate responses when used in low-resource languages such as Yorùbá. Retrieval-Augmented Generation (RAG) has since emerged as a promising approach to mitigate hallucinations by grounding LLM outputs in verified knowledge sources. To assess its effectiveness in low-resource contexts, we construct a controlled Yorùbá QA dataset derived from Nigerian drug labels, comprising 460 question–answer pairs across 92 drugs which was used to evaluate the impact of different retrieval strategies: hybrid lexical–semantic retrieval, Hypothetical Document Embeddings (HyDE), and Cross-Encoder re-ranking. Our results show that hybrid retrieval strategies, combining lexical and semantic signals, generally yield more reliable and clinically accurate responses, while other advanced re-ranking approaches show inconsistent improvements. These findings hereby underscore the importance of effective retrieval design for safe and trustworthy multilingual healthcare QA systems.

1 Introduction

Large language models (LLMs) have transformed healthcare by enabling natural language interaction and on-demand information access, helping to reduce the spread of incorrect medical information in Nigeria. LLMs can improve health information accessibility, particularly for individuals with limited health literacy. However, LLMs are prone to hallucinations, potentially generating misleading or unsafe medical content when responses are not grounded in verified sources, posing significant risks in healthcare settings.

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to mitigate such hallucinations by grounding generated responses in external knowledge bases, thereby improving factual consistency and reliability. As a result, RAGs have received increasing attention for healthcare information access.

Despite this innovation, most existing work do not sufficiently account for the linguistic realities of low-resource settings. Although English is Nigeria's official language, many individuals primarily communicate in indigenous languages, such as Yorùbá, which limits the usability of LLMs and RAG systems in rural communities. These language barriers disproportionately affect vulnerable populations, including the elderly, individuals with low literacy level, and persons with visual impairments, for whom accessible and localized information is especially critical.

While RAG has been shown to improve reliability in healthcare, its performance in low-resource, multilingual contexts remains underexplored. Retrieval failures in such settings can directly compromise generation quality, undermining the trustworthiness of medical information provided to users.

This raises the research question: **How do variations in embedding models and retrieval strategies affect Yorùbá QA accuracy in RAG systems?**

Addressing this question is essential for assessing the practical viability of RAG healthcare information systems in Nigeria and similar contexts.

To bridge this gap, we construct a controlled Yorùbá QA dataset used to evaluate several retrieval strategies to give insights on the system that offers the best results.

2 Literature Review

Given the increasing adoption of LLMs across multiple industries, especially healthcare, where they support tasks such as clinical documentation, question answering, and the summarization of complex medical research.

Despite these advances, a fundamental limitation of LLMs is their tendency to hallucinate, producing plausible but incorrect information (Brown et al., 2020). To address this challenge, Retrieval-Augmented Generation (RAG) was introduced as a non-parametric memory framework that grounds model outputs in external knowledge sources to increase the factual accuracy of these models. (Lewis et al., 2021).

Following its introduction, RAG has been increasingly applied in healthcare settings, where accuracy is paramount due to the direct impact of generated information on human lives and to minimize medical errors. However, while RAG substantially mitigates hallucinations, the same reliance on external knowledge that underpins its effectiveness also represents its primary vulnerability. This dependence exposes RAG systems to risks such as poisoning attacks, which can compromise the reliability of downstream generations (Zou et al., 2024). This has led to a lot of questioning in these technologies and their efficiency in information retrieval

The healthcare domain provides a particularly suitable testbed for evaluating these RAG systems, as it offers well-defined ground truth sources and stringent correctness requirements. Recent studies have therefore focused on Evaluating RAG in medical settings, assessing their retrieval quality, factual accuracy, and overall clinical usefulness (Xiong et al., 2024). Alongside these evaluations, several methods have been proposed to improve retrieval effectiveness and generation reliability. For instance, Hypothetical Document Embeddings (HyDE) (Gao et al., 2022) address challenges in zero-shot learning and relevance encoding by generating a hypothetical document from a query using an instruction-following language model, which is then embedded for retrieval. Cross-Encoders have also been used to re-rank candidate passages based on joint query-document representations (Déjean et al., 2024), and hybrid retrieval strategies combining cosine similarity with keyword-based search have been proposed to balance semantic relevance and exact-match precision (Lee et al., 2023).

However, the majority of these evaluations and architectural improvements remain centered on high-resource languages and regions. Recent studies have shown that LLMs exhibit degraded performance on Yorùbá QA tasks, highlighting the challenges of applying these models to low-resource languages (Okewunmi et al., 2025). In Nigerian healthcare contexts, access to information is often constrained by limited curated datasets, linguistic diversity, and infrastructural challenges. Within this setting, the implementation of RAG systems has demonstrated the potential to improve access to drug information and reduce medication errors (AI et al., 2025), and has also shown promising results in enhancing accessibility across indigenous languages, while revealing persistent challenges related to tonal complexity and language-specific retrieval fidelity (Ismail et al., 2025).

To aid further exploration of Retrieval-Augmented Generation (RAG) systems for healthcare access in Nigeria, systematic evaluation is essential in identifying configurations that offer appropriate trade-offs between retrieval accuracy and generation reliability when working with indigenous languages. This need is further heightened by the architectural complexity of RAG systems and their sensitivity to design choices. To address this, we evaluate several RAG strategies on a Yorùbá QA drug dataset.

3 Dataset

The collation process for our Yorùbá QA dataset began by capturing images of drugs in a pharmacy, reflecting a realistic and accessible way users may seek information about medications, whether through images or text. From these images, both brand and generic drug names were extracted and used to construct a trusted dataset tailored to the Nigerian medical context, which served as the ground truth.

To closely mirror real-world user interactions, questions were formulated to reflect how users typically engage with large language models when seeking medical information.

The dataset comprises drug brand names, generic names, typical user questions, verified answers in English, and translations of both the questions and answers into Yorùbá. The Yorùbá translations were manually curated by native speakers and supplemented using the Spitch (Spitch, 2025) translation model , ensuring accuracy and consistency throughout the dataset.

The dataset covers 92 distinct drugs, with five question–answer pairs per drug, resulting in a total of 460 QA pairs. On average, this corresponds to five questions per drug, each designed to reflect common information needs in real-world medical

Table 1: A sample of the Yoruba QA dataset

Category	English (Question & Answer)	Yoruba (Ìbéèrè àti Ìdáhùn)
Usage	**What is Unicid used for?** It treats hyperacidity, heartburn and stomach ulcers.	**Kínì ni a máa n lò Unicid fún?** Ó ń tọju hyperacidity, iná ọkan ati awọn ọgbẹ inu.
Dosage	**What is the adult dose for Unicid?** Chew one or two tablets one hour after meal.	**Kí ni iye tí àgbàlagbà gbọ́dọ̀ lo fún Unicid?** Jẹ tabulẹti kan tabi meji ni wakati kan lẹhin ounjẹ.
Side Effects	**What are common side effects?** Mild diarrhea, constipation, fatigue and very rarely weight loss.	**Kí ni àwọn àbájáde àìlera tí ó wọ́pọ̀?** Àwọn ìrẹ̀pò pẹ̀lu igbẹ gbuuru, àìrígbẹyà, rirẹ ati iwuwo pipadanu.
Contra-indications	**Who should NOT take Unicid?** Patients with decreased gastrointestinal motility and kidney disease.	**Ta ni ko yẹ ki o máa lo Unicid?** Àwọn aláìsàn to ni ìdàámú ètò ofífo ati àwọn tó ní àrùn kíndìnrín.
Administration	**How should Unicid be taken?** It should be taken by mouth.	**Báwo ni kí a ṣe lo Unicid?** Ó yẹ kí ó mú nípasẹ̀ ẹnu.

inquiries, as shown in Table 1,

Each question–answer pair is provided in both English and Yorùbá, yielding a balanced multilingual dataset. English serves as the source language for verified medical content, while Yorùbá translations capture realistic low-resource language usage in the Nigerian context. For every question, corresponding answers are available in both languages, ensuring a 1:1 language split and enabling controlled evaluation of cross-lingual and multilingual retrieval strategies.

The resulting question–answer pairs were subsequently validated by certified pharmacists for medical accuracy and by Yorùbá language experts for linguistic correctness. This validated data constitutes the Yorùbá question–answer dataset used to evaluate the retrieval strategies evaluated in this study.

4 Methodology

4.1 Knowledge Base

For the knowledge base, we employ the EMDEX pharmaceutical reference due to a study by (Ogbonna and Ijeoma, 2021) that found that one of Nigeria's most common sources of drug information is the Nigerian Essential Medicines Index (EMDEX). EMDEX provides structured, clinically relevant information on drugs, including indications, dosage guidelines, contraindications, and adverse effects. Its local relevance and medical reliability make it particularly suitable for grounding responses in a Nigerian medical context, especially when developing systems for low-resource language applications.

4.2 Data Processing

The EMDEX knowledge base is originally distributed in PDF format, which is not directly suitable for efficient retrieval. To enable preprocessing, chunking, and embedding, all documents are first converted into plain text (`.txt`) format.

The knowledge base consists of EMDEX drug monographs, where each document corresponds to a single drug. After conversion, minimal preprocessing is applied: leading and trailing whitespace is removed, and newline characters are replaced with spaces to preserve textual continuity.

Given the structured nature of EMDEX drug monographs, a two-stage chunking strategy is adopted to maintain semantic coherence while supporting fine-grained retrieval. In the first stage, documents are segmented by generic drug name, producing a collection of drug-specific text files. This ensures that each document contains information about only one drug and prevents cross-drug information leakage during retrieval.

In the second stage, each drug-specific text file is further divided into fixed-length, character-based chunks of 1200 characters using deterministic slicing. To better capture local context, chunks are generated with overlap. Each resulting chunk is assigned a unique global contiguous integer identifier(chunk_id), which is used consistently throughout indexing, retrieval, and evaluation.

This hierarchical chunking approach allows the retrieval system to match queries to specific sections like dosage, contraindications, or adverse effects, rather than entire drug documents, thereby improving retrieval precision.

4.3 Embeddings

To support multilingual retrieval across English and Yorùbá queries, we select embedding models based on the linguistic characteristics of the dataset. For Yorùbá-language text, we employ two embedding models: a BERT-based model fine-tuned specifically for Yorùbá (Davlan, 2021) and the multilingual-E5 model (Wang et al., 2024). These models are chosen as they ahve been proven to capture language-specific semantics while also supporting cross-lingual representations.

For English-language text, we use the MiniLM sentence transformer, which provides an effective balance between computational efficiency and semantic representation quality.

We generate embeddings for all chunks using the SentenceTransformers framework with a batch size of 32. For the multilingual-E5 model, passages are prefixed with `"passage:"` prior to encoding, following the model's recommended input format. This prefix is necessary because multilingual-E5 was trained on labeled text, where `"query:"` indicates questions and `"passage:"` indicates documents, helping the model encode chunks in a way that aligns with retrieval queries. Other models, such as MiniLM and Davlan BERT-Yorùbá, do not require such prefixes because they are trained as standard sentence encoders. All embeddings are L2-normalized before indexing.

Once the embeddings were generated and normalized, we performed dense retrieval using FAISS with an `IndexFlatIP` index, where inner product similarity corresponds to cosine similarity. We relied on exact nearest-neighbor search without approximate indexing to ensure precise matches. For each embedding model, a separate FAISS index was built, and chunk metadata was stored in the same order as vectors were added, maintaining a stable mapping between FAISS vector IDs and chunk identifiers. This setup enabled efficient similarity-based retrieval of chunks, allowing queries to accurately match relevant sections across the knowledge base.

4.4 Retrieval Strategies

For response generation, we first conducted a preliminary evaluation of several candidate LLMs using the LLM-as-a-Judge framework across five metrics: fluency, accuracy, completeness, factuality, and clinical safety. As shown in Table 2, **Mistral-7B** achieved the highest overall scores, demonstrating superior performance in medication label interpretation. Based on these results, Mistral-7B was selected as the base LLM for our RAG system.

Table 2: LLM-as-a-Judge comparison of candidate models.

Model	Flu	Acc	Comp	Fact	Safety
Mistral-7B	4.8	4.5	4.5	4.9	4.9
Llama-3	4.5	2.0	2.5	4.9	4.9
Qwen2.5-7B	4.7	4.0	4.0	4.9	4.9
Mixtral-8x7B	4.8	3.5	3.7	4.9	4.9

To provide Mistral-7B with contextually grounded and medically accurate information, we evaluated three retrieval strategies: **hybrid retrieval**, **HyDE retrieval**, and **Cross-Encoder re-ranking**. These strategies differ in their balance between semantic coverage, exact matching, and computational cost, and allow us to understand trade-offs in low-resource multilingual settings.

4.4.1 Hybrid Retrieval

We adopt hybrid retrieval as the baseline, which integrates a combination of dense vector search and lexical keyword search, which are merged using weighted Reciprocal Rank Fusion (RRF). Dense retrieval is performed over L2-normalized embeddings using FAISS, capturing semantic similarity between queries and document chunks. Lexical search uses BM25 with Unicode-aware tokenization to ensure that Yorùbá diacritics are preserved. To prioritize exact matches for drug names, cleaned generic drug names are repeated in the BM25 query.

For each query, the top 30 candidates are retrieved independently from both the dense and lexical searches. These candidates are then combined using RRF (k=60, BM25 weight=2.0, dense weight=1.0), which balances exact term matches with semantic relevance. Finally, the top 5 chunks with the highest aggregated RRF scores are returned for use in downstream response generation.

4.4.2 HyDE Retrieval

HyDE (Hypothetical Document Embeddings) retrieval is designed to improve semantic matching, especially when queries may not share exact words with relevant documents. Instead of using the original query directly, HyDE first generates a short, hypothetical answer to the query using the base LLM (Mistral-7B). This answer is intended to capture the underlying intent and key information of the

query in natural language, effectively transforming the question into a document-like form that is easier to match with relevant chunks.

The generated hypothetical passage is produced in English with deterministic decoding (temperature=0) to ensure consistency across retrievals. The generated passage is converted into an embedding using the query encoder. For multilingual-E5, the `"query:"` prefix is added to indicate that it is a query. The resulting embedding represents the semantic content of the query in a form that can be directly compared with the embeddings of the knowledge base chunks. Finally, the top K=5 chunks with the highest similarity scores are retrieved from the FAISS index and passed to the response generation model.

By creating a query surrogate in this way, HyDE allows the system to retrieve semantically relevant information even when the original query does not share exact vocabulary with the target documents, improving coverage and retrieval accuracy in challenging scenarios.

4.4.3 Cross-Encoder Re-ranking

Cross-encoder re-ranking is a two-step process designed to improve retrieval precision. First, a set of up to 50 candidate chunks is retrieved from the FAISS index using the same bi-encoder employed for initial dense retrieval. This step ensures a broad set of potentially relevant documents is considered.

In the second step, the retrieved candidates are scored and re-ranked using the multilingual cross-encoder `cross-encoder/mmarco-mMiniLMv2-L12-H384-v1`. This model jointly considers the query and each candidate chunk, producing a more accurate relevance score than the bi-encoder alone. It was selected because it supports non-English text and was trained for passage re-ranking, making it suitable for bilingual retrieval. The top K=5 re-ranked chunks are then returned for use in response generation.

After implementing the retrieval strategies, we integrate them into a Retrieval-Augmented Generation (RAG) system, where the top K retrieved chunks for each query are provided as context to Mistral-7B. The model then generates responses grounded in these documents, ensuring accurate and contextually relevant answers. For Yorùbá output, the generated English response is translated using a fine-tuned English-to-Yorùbá model. The hybrid retrieval strategy serves as the baseline, combining keyword-based and vector-based search to capture both exact matches for critical medical terms and semantically relevant chunks when lexical overlap is limited. This approach ensures that retrieved documents are effectively leveraged, allowing Mistral-7B to generate accurate and medically grounded responses in both English and Yorùbá.

We evaluate the three retrieval strategies—hybrid retrieval, HyDE, and Cross-Encoder re-ranking—to assess their effectiveness in a low-resource, multilingual setting and to understand the trade-offs between retrieval accuracy, semantic coverage, and computational cost.

4.5 Metrics

We employ different evaluation metrics to assess the quality and semantic faithfulness of generated responses across different retrieval strategies.

BERTScore: To evaluate semantic similarity beyond exact lexical matches, we compute BERTScore between the model-generated responses and the ground-truth answers.We chose BERTScore as a metric because it leverages contextual embeddings from pre-trained transformer models to measure similarity unlike BLEU (Papineni et al., 2002), which relies on n-gram overlap, and METEOR (Lavie and Agarwal, 2007), which depends primarily on lexical matching and predefined linguistic resources and are less effective at accounting for paraphrasing and semantic equivalence.

LLM-as-a-Judge Evaluation: To address the limitations of automatic metrics, particularly for Yorùbá, we adopt an LLM-as-a-Judge evaluation framework to perform human-like assessments of response quality. We used OpenAI GPT-4o (OpenAI, 2024) as the judge model to evaluate whether the generated responses demonstrate fluency, maintain medical correctness, is factual and ensures clinical safety. This approach enables a more nuanced evaluation of response quality, especially in cases where surface-level similarity metrics fail to capture semantic degradation or hallucinations.

Hit Rate@k (HR@k): We choose HR@k because it directly reflects the retriever's ability to include pertinent documents among its top-ranked results, making it a suitable and intuitive metric for assessing retrieval quality in RAG systems with-

out relying on model-specific or generation-level assumptions. The metric measures how often relevant documents are successfully retrieved within the top k results returned by the retriever. (He et al., 2017)

Table 3: Evaluation of RAG system across languages, retrieval strategies, and embedding models.

Language	Retrieval Strategy	Embedding Model	BERTScore (F1)	LLM-as-a-Judge				
				Flu.	Acc.	Fact.	Comp.	Safety
English	Hybrid Retrieval	all-MiniLM-L6-v2	0.73	4.67	2.11	2.84	2.02	3.75
English	HyDE	all-MiniLM-L6-v2	0.72	4.64	1.82	2.68	1.81	3.62
English	Cross Encoder	all-MiniLM-L6-v2	0.72	4.62	1.88	2.54	1.88	3.49
Yorùbá	Hybrid Retrieval	Davlan/bert-yoruba	0.70	4.27	1.80	2.28	1.75	3.15
Yorùbá	Hybrid Retrieval	multilingual-e5	**0.71**	**4.37**	**2.02**	**2.18**	**1.81**	**3.15**
Yorùbá	HyDE	Davlan/bert-yoruba	0.71	4.22	1.63	1.87	1.58	2.72
Yorùbá	HyDE	multilingual-e5	0.70	4.29	1.73	1.94	1.65	2.96
Yorùbá	Cross Encoder	Davlan/bert-yoruba	0.70	4.19	1.54	1.71	1.46	2.45
Yorùbá	Cross Encoder	multilingual-e5	0.71	4.29	1.84	2.08	1.74	2.95

5 Results

Across all evaluated retrieval and generation strategies, we observe a clear and systematic performance gap between the high-resource English setting and the low-resource Yorùbá setting (Table 3). To isolate the source of this gap, we further analyze retrieval behavior using a retrieval-only evaluation with manually annotated gold evidence (Table 4). This analysis allows us to disentangle retrieval quality from downstream generation and translation effects.

5.1 English Retrieval Performance

English retrieval exhibits strong and stable performance across strategies. Hybrid Retrieval with the all-MiniLM-L6-v2 embedding consistently yields the most reliable and medically accurate responses. The retrieval-only evaluation confirms that relevant evidence is frequently retrieved at very high ranks, achieving an HR@1 of 0.67 and an HR@5 of 0.91. This indicates that for the majority of English queries, the correct supporting evidence is retrieved immediately or within the top few results.

Although HyDE and Cross-Encoder approaches achieve comparable semantic similarity scores, they do not meaningfully improve retrieval coverage or downstream factual correctness. The high HR@1 and narrow gap between HR@1 and HR@5 suggest that English retrieval errors are primarily limited to rare misses rather than ranking deficiencies. Overall, these results demonstrate that in high-resource languages, strong sentence embeddings combined with a well-designed hybrid retrieval pipeline are sufficient, and increasingly complex retrieval mechanisms yield diminishing returns.

Table 4: Retrieval-only HR@K with Gold Evidence

Language	HR@1	HR@3	HR@5
English	0.67	0.85	0.91
Yorùbá	0.32	0.73	0.87

5.2 Yorùbá Retrieval Performance

In contrast, Yorùbá retrieval appears more sensitive to ranking quality and coverage. While relevant evidence is often present in the retrieved set, it is frequently ranked lower, as reflected by the disparity between HR@1 (0.32) and HR@5 (0.87). This suggests that improvements in ranking could substantially enhance retrieval effectiveness in this language.

Hybrid retrieval paired with the multilingual-e5 embedding provides the most consistent balance, achieving broad coverage while maintaining medical relevance. However, using a language-specific or multilingual embedding alone does not guarantee better performance. In our current setup, Yorùbá queries are translated into English before retrieval, which may result in information loss. As a result, retrieval errors in Yorùbá can propagate to downstream generation, reducing answer completeness and factual accuracy.

These findings highlight a key difference between high- and low-resource settings, English retrieval is largely saturated, whereas Yorùbá retrieval remains constrained by coverage and ranking. In low-resource medical QA, expanding retrieval coverage and improving ranking are likely to yield the most significant gains in downstream answer quality and reliability.

5.3 HyDE and Cross-Encoder Behavior in Yorùbá

Advanced strategies such as HyDE and Cross-Encoder show inconsistent performance in Yorùbá. HyDE can improve retrieval with certain embeddings but may reduce downstream medical correctness with others. Cross-Encoder is the most unstable, sometimes producing incomplete or unsafe responses despite similar semantic similarity. These results suggest that computationally intensive reranking does not reliably improve performance in low-resource languages and may even harm medical reliability if retrieval coverage is insufficient.

5.4 Retrieval Accuracy and Clinical Safety Trade-offs

Across both languages, higher retrieval coverage correlates strongly with improved LLM-as-a-Judge scores for accuracy, factuality, and clinical safety (Table 3). BERTScore alone is insufficient, as similar semantic similarity scores can mask large differences in medical correctness. Overall, these results demonstrate that in low-resource healthcare QA, robust and reliable retrieval strategies are more critical than sophisticated re-ranking mechanisms. Hybrid lexical–semantic retrieval offers the best balance between retrieval effectiveness and downstream medical safety, emphasizing the importance of careful retrieval design when deploying RAG systems for indigenous languages.

6 Conclusion

Overall, our study demonstrates the potential of RAG systems to improve access to medication information in low-resource languages like Yorùbá. By evaluating different retrieval strategies, we highlight the importance of effective retrieval design for ensuring reliable and medically grounded responses, particularly in multilingual healthcare contexts.

7 Limitations

Our RAG system exhibits challenges in retrieving accurate information for certain drugs due to the inherent structure of the knowledge base. Despite employing careful chunking strategies to segment the documents semantically, some drug information remains fragmented or inconsistently represented. This affected the downstream accuracy and completeness of the generated responses

Another technical hurdle was the system's difficulty in correctly retrieving information for drugs with multi-component generic names. These limitations highlight key areas that could significantly enhance retrieval reliability and response quality.

7.1 Future Work

Several directions could improve the coverage, reliability, and multilingual performance of our system. The current evaluation uses only five questions per drug, which is quite limited. Medication labels involve many complex scenarios, including pediatric dosing, pregnancy and lactation considerations, drug interactions, and renal or hepatic dose adjustments. Expanding the dataset to cover these corner cases and additional clinical categories would provide a more realistic assessment for deployment.

Establishing a detailed error taxonomy, selecting the wrong drug, providing incorrect dosages, missing contraindications, or hallucinating adverse effects could help identify specific weaknesses in the system. Linking these errors to retrieval failures would clarify the impact of retrieval quality on overall performance.

Incorporating native Yorùbá generation and separately evaluating machine translation quality would strengthen the system's multilingual support. Using metrics such as BLEU or ChrF, along with targeted human spot-checks, can help distinguish errors arising from machine translation versus retrieval or generation.

References

Axum AI, :, J. Owoyemi, S. Abubakar, A. Owoyemi, T. O. Togunwa, F. C. Madubuko, S. Oyatoye, Z. Oyetolu, K. Akyea, A. O. Mohammed, and A. Adebakin. 2025. Open-source retrieval augmented generation framework for retrieving accurate medication insights from formularies for african healthcare workers. *Preprint*, arXiv:2502.15722.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Davlan. 2021. bert-base-multilingual-cased-finetuned-yoruba.

Hervé Déjean, Stéphane Clinchant, and Thibault Formal. 2024. A thorough comparison of cross-encoders and llms for reranking splade. *Preprint*, arXiv:2403.10407.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *Preprint*, arXiv:2212.10496.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. *Preprint*, arXiv:1708.05031.

Ahmad Ismail, Bashirudeen Ibrahim, Olubayo Adekanmbi, and Ife Adebara. 2025. Retrieval-augmented generation meets local languages for improved drug information access and comprehension. pages 108–114.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

Dohyeon Lee, Seung-won Hwang, Kyungjae Lee, Seungtaek Choi, and Sunghyun Park. 2023. On complementarity objectives for hybrid retrieval. pages 13357–13368.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Brian Ogbonna and Okoye Ijeoma. 2021. Utilization of drug information services in selected tertiary hospitals in enugu state, nigeria. *Journal of Biomedical Research*, 1:17–26.

Paul Okewunmi, Favour James, and Oluwadunsin Fajemila. 2025. Evaluating robustness of LLMs to typographical noise in Yorùbá QA. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 195–202, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Spitch. 2025. Spitch documentation. `https://docs.spitch.app/`. Accessed: 2025-12-18.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *Preprint*, arXiv:2402.13178.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. *Preprint*, arXiv:2402.07867.