Benchmarking Open-Source Large Language Models on Code-Switched Tagalog-English Retrieval Augmented Generation

Aunhel John M. Adoptante[1,*], Jasper Adrian Dwight V. Castro[1], Micholo Lanz B. Medrana[1], Alyssa Patricia B. Ocampo[1], Elmer C. Peramo[1], and Melissa Ruth M. Miranda[2]

[1]Department of Science and Technology, Computer Software Division, Advanced Science and Technology Institute, Diliman, Quezon City, Philippines
[2]Pamantasan ng Lungsod ng Maynila, Intramuros, Manila, Philippines
Email: aunheljohn.adoptante@asti.dost.gov.ph (A.J.M.A.); jasperadriandwight.castro@asti.dost.gov.ph (J.A.D.V.C.); michololanz.medrana@asti.dost.gov.ph (M.L.B.M.); alyssapatricia.ocampo@asti.dost.gov.ph (A.P.B.O.); elmer@asti.dost.gov.ph (E.C.P.); mrmmiranda15@gmail.com (M.R.M.M.)
*Corresponding author

***Abstract*—Code-switching, the alternation between languages within a sentence or discourse, poses significant c hallenges in Natural Language Processing (NLP). Effective NLP systems must employ advanced modeling techniques to accurately process and generate code-switched text, capturing the linguistic nuances and contextual dependencies. This study evaluates five o pen-source L arge L anguage M odels (LLMs)—Mistral, SeaLLM, Falcon, Phi-3-mini, and Gemma—on their performance with code-switched Tagalog-English (Taglish) queries in a retrieval-augmented generation (RAG) task. Mistral and SeaLLM outperformed the others in generating contextually grounded and relevant answers, attributed to their advanced architectures and extensive multilingual training data. In contrast, Falcon, Phi-3-mini, and Gemma struggled with effective code-switching handling. The models performed best on English queries, followed by Taglish, with the lowest performance on Tagalog queries, highlighting the need for more balanced training data for low-resource languages. Additionally, models retrieved context better from `.pdf` documents compared to `.txt` files. F uture r esearch s hould f ocus o n a nalyzing language composition in datasets, investigating real-world code-switching behavior, and studying the effects of code-switching on model understanding to enhance NLP accessibility for underrepresented languages.**

***Keywords*—Large Language Models (LLM), code switching, retrieval augmented generation**

I. Introduction

The emergence of Large Language Models (LLMs) has ushered the unrelenting improvement of Natural Language Processing (NLP), enabling significant advancements across a variety of applications [1], from machine translation [2], [3] to conversational agents [4], [5]. These models, trained on vast amounts of text data, have demonstrated an unprecedented ability to understand and generate human-like text, making them invaluable tools for numerous NLP tasks. Among these applications, Retrieval-Augmented Generation (RAG) stands out as a promising approach that combines the strengths of retrieval-based and generation-based models to produce more accurate and contextually relevant responses [6]. By capitalizing on external knowledge bases during the generation process, RAG models can retrieve pertinent information and seamlessly integrate it into their responses, enhancing the overall quality and coherence of the generated text.

However, the challenge of code-switching—where speakers alternate between two or more languages within a single conversation or sentence—poses unique difficulties for LLMs [7]. Code-switching introduces linguistic complexity that goes beyond simple bilingualism, requiring models to not only understand multiple languages but also to navigate the nuanced and dynamic transitions between them [8]. This phenomenon is particularly prevalent in multilingual societies, such as the Philippines, where Tagalog-English code-switching (commonly known as "Taglish") is a common mode of communication. Taglish reflects the fluid and adaptive nature of language use in such communities, where speakers effortlessly blend elements of both languages in response to social, cultural, and contextual cues [9]. Additionally, these linguistic complexities also increase the likelihood of hallucination, where the model generates plausible but factually incorrect or irrelevant responses [10]. Hallucination is especially concerning in code-switched environments, such as Tagalog-English, where the fluid transition between languages can confuse the model, leading to the generation of content that does not accurately reflect the retrieved information.

The ability to effectively process and generate code-switched text is crucial for creating NLP systems that can serve multilingual populations in a manner that respects and reflects their linguistic practices [11]. In the context of the Philippines, where Taglish is widely used in everyday communication, media, and even formal settings, developing LLMs capable of handling code-switching can significantly enhance user experience and accessibility. Moreover, the insights gained from studying Taglish can be extended to other multilingual contexts, offering broader implications for the global development of NLP technologies.

In one study, LLM-based QA systems were tested on their abilities to generate answers in the language of the question, regardless of the language of the retrieved context [12].

Manuscript received August 28, 2024; revised September 30, 2024; accepted November 11, 2024; published February 17, 2025.

These QA systems used mBERT, mT5-base, mLUKE, and XLM-RoBERTa, which are multilanguage-capable models employed in the context of cross-lingual open-retrieval question answering. The best-constrained system for tagalog was mLUKE with FiD which achieved 20.8 F1 on Tagalog. This is significantly better than other systems with nearly zero F1 scores. The improvement in Tagalaog performance was attributed to using entity-aware retrieval representations and the inclusion of the Tagalog Wikipedia passages for retrieval while the poor performance of the other systems is likely due to them not incorporating sufficient Tagalog-specific resources for retrieval and generation.

This paper aims to benchmark the performance of open-source large language models on code-switched Tagalog-English retrieval-augmented generation tasks. Despite the widespread use of code-switching in multilingual communities, research on LLMs' ability to handle such mixed-language inputs remains limited. By focusing on Tagalog-English code-switching, this study seeks to address a significant gap in the current understanding of LLMs' capabilities and limitations.

This work will evaluate several state-of-the-art LLMs, examining their effectiveness in the generation phase. Through comprehensive benchmarking, this research aims to provide valuable insights into the optimization of open-source LLMs for code-switched environments, thereby enhancing their applicability in real-world, multilingual settings.

As the demand for multilingual and culturally adaptive AI grows, understanding and improving LLMs' performance in mixed-language contexts becomes increasingly critical. This research not only addresses a pertinent technical challenge but also supports the broader goal of fostering inclusive and accessible AI technologies for diverse linguistic communities.

II. Related Work

A. A Brief Overview of Retrieval Augmented Generation (RAG)

Pre-trained language models such as general purpose seq2seq models are powerful tools in NLP which are capable of capturing extensive world knowledge within their parameters. While these models exhibit strong performance across a range of tasks, they occasionally produce plausible but incorrect information. This issue arises from inherent difficulties in accessing and updating their underlying knowledge base. On the other hand, retrieval-based methods, which use externally retrieved texts, provide a precise and easily updatable knowledge access mechanism which is useful for many NLP tasks. In this area, dense retrieval techniques are increasingly outperforming traditional information retrieval methods that rely on heuristics [13]. However, training dense retrievers require labels for supervision and incorporating them to downstream models usually require task specific architectures. RAG combines the strengths of seq2seq models with explicit knowledge retrieval to potentially address these limitations. This is achieved by jointly learning end-to-end latent retrieval and generation [14].

The process of RAG is defined as the interaction between two components, namely the *retriever* and the *generator*, in which the retriever *augments* the output of the generator by providing contextually relevant information in at least one step of the generation pipeline [15]. A basic RAG pipeline proceeds with the ingestion phase where raw documents are parsed and broken down into more manageable chunks. These chunks are converted into meaningful embeddings using pre-trained transformer-based models [16]. These embeddings are indexed in a vector store to facilitate efficient retrieval when a query is made. A user or system-generated query is converted into embeddings using the same pre-trained transformer-based model. These are then queried in the vector store using similarity search to retrieve the top k most relevant chunks. The original query and the top k retrieved chunks are fed into an LLM where the combined information is synthesized and a coherent and contextually relevant response is generated [17].

A common use case of RAG is connecting LLMs with an organization's proprietary data [18]. This integration allows users to obtain specific information more quickly and accurately by only parsing through the relevant sections of documents from the organization's database. In these use cases where RAG is productionized, improved retrieval techniques and evaluation metrics are paramount.

B. Large Language Models for RAG

Retrieval-Augmented Generation (RAG) bring new advancements to enhance LLMs by incorporating external sources of knowledge to improve the factual accuracy of the generated content [6]. This incorporation has become crucial for advancing the capabilities of LLMs in real-world applications, particularly in complex and knowledge-intensive tasks. The widespread accessibility of LLMs, mainly due to their open-source nature, has made experimental research more feasible. As a result, many studies have explored the effects of RAG on these models. This literature review assesses the current state and application of RAG in various LLMs, discussing the benefits and challenges that come with this technology.

Integrating RAG with many LLMs has led to notable improvements in accuracy. Gemini models, particularly Gemini Pro's (GPro) factual accuracy and context-specific generation are improved by RAG utilizing relevant external information and temporal ordering [19]. Without RAG, GPro performs well but lacks the same level of factual grounding and contextual accuracy. In multimodal tasks, such as image understanding and speech recognition, Gemini models show notable improvements with RAG.

The accuracy improvement is also seen with BLOOM models. Integrating RAG with BLOOM models enhances their performance, particularly in handling long-context scenarios [20]. The effectiveness of RAG in providing additional contextual information and reducing interference from irrelevant data are noticeable due to the accuracy improvements are consistent across different BLOOM model sizes. Similarly, the Mistral model also benefits significantly from RAG integration, showing improvements in precision, recall, F1-score, BLEU, and ROUGE scores. The use of external knowledge sources like Wikipedia helps reduce hallucinations and generate more factually accurate and coherent responses [21].

The use of RAG also improves LLaMA-2's performance noticeably [22]. The fine-tuned LLaMA-2 without additional context answered over half of the questions correctly,

while LLaMA-2 with RAG showed a slight improvement. However, the best performance was achieved with the Tree-RAG (T-RAG) implementation, which combines RAG with an entities tree for hierarchical context, significantly enhancing accuracy. T-RAG also demonstrated superior resilience in the "Needle in a Haystack" test compared to RAG alone, highlighting the benefits of integrating structured contextual information. With that, it is worth noting that there is a model called SeaLLM-13B-v1, trained with Southeast Asian languages [23]. This model was based from the LLaMA-2-13B architecture.

Interestingly, not all models improve with the incorporation of RAG. GPT-4's integration with RAG shows mixed results when it comes to accuracy [24]. The model's accuracy decreases slightly when RAG is used, indicating that external context can sometimes introduce errors. The model exhibits a high context bias, sometimes relying on external context, even when it might be incorrect. However, GPT-4 has a low prior bias, rarely disregarding correct external context for its own incorrect prior knowledge.

In contrast, Claude Opus outperforms other models with the highest accuracy and the lowest context bias [25]. It adheres to incorrect contextual information less than GPT-4. Claude Opus maintains a high accuracy with and without RAG, showing strong performance regardless of external context. The model also exhibits minimal prior bias, indicating effective use of correct external context. Meanwhile, PaLM achieves moderate accuracy with RAG, indicating that a substantial portion of the responses generated were correct when using retrieved chunks [26]. Interestingly, PaLM's accuracy increases without RAG, suggesting that directly providing evidence improves accuracy more effectively than relying on retrieval components.

The integration of RAG with LLMs significantly enhances their performance by providing additional context and reducing the incidence of hallucinations. While RAG still comes with challenges, such as context bias and potential errors from external sources, the overall benefits in accuracy and factual correctness contribute greatly to LLM technology. Continued research and development in RAG techniques and technology are essential for enhanced capabilities of LLMs in complex and real-world applications.

III. Methodology

A. Selection of Large Language Models

A diverse set of large language models were selected for performance evaluation on code-switched Tagalog-English retrieval-augmented generation question answering task. These models were chosen based on relevance and effectiveness in question answering. The selected models are Mistral AI's Mistral-7B-Instruct-v0.3 [27], SeaLLMs' SeaLLMs-v3-1.5B-Chat [23], TII's Falcon-7b-instruct [28], Microsoft's Phi-3-mini-4k-instruct [29], and Google's Gemma-7b-it [30]. These open-source models were implemented in the RAG evaluation pipeline as discussed in the succeeding sub-sections.

Each model has distinct architectural innovations to enhance performance in language processing tasks. Mistral integrates transformers and RNNs to manage long-term dependencies and improve response accuracy [31]. SeaLLM employs advanced transformer variants with enhanced attention mechanisms and is fine-tuned for Southeast Asian languages, excelling in multilingual and code-switched scenarios [23]. Falcon features optimizations like FlashAttention and multi-query attention in its transformer-based architecture, enhancing memory efficiency and scalability [32]. Phi-3-mini uses a dense decoder-only transformer with 3.8 billion parameters, focusing on parameter efficiency to generate high-quality text [29]. Gemma builds on the original transformer architecture with positional encodings and multi-head attention, boosting contextual understanding and generation efficiency [30].

B. RAG Pipeline

Langchain's *RecursiveCharacterTextSplitter* module [33] was used to split the documents into chunks of size 2000 and overlap equal to 100 characters. The document splits and user queries were embedded using OpenAI embeddings. The split embeddings were stored using Meta's FAISS [34] vector database. Fig. 1 illustrates the prompt template used in the pipeline.

```
You are a smart assistant capable of answering
questions given a context derived from multiple
documents.

You will be given a question and relevant
excepts from different documents related to the
question.

Please provide short and clear answers based on
the provided context. Be polite and helpful.


Context: {context}

Question: {question}

Answer:
```

Fig. 1. Prompt template used for the Retrieval Augmented Generation (RAG) Pipeline.

C. Evaluation Dataset

The evaluation dataset used for this study was derived from curated `.txt` and `.pdf` files spanning various document types, including news articles, short stories, and academic literature, each written in English, Tagalog, or Taglish. From each file, 15 questions were manually extracted along with the corresponding answers and the actual chunk of text from where the questions were derived. Each of the 15 question and answer pairs are composed of three groups of five questions, with each query group written in English, Tagalog, or Taglish, irrespective of the language of the source document. Adopting this approach allowed for the dataset to encompass a variety of linguistic contexts and complexities typical of code-switched interactions.

D. Evaluation Metrics

The similarity between LLM-generated responses and human-generated responses was assessed using the ROUGE and BLEU metrics from the evaluate module. These metrics quantify the overlap between the model's output and the reference (human) output. For ROUGE, we obtained the F1-scores for unigram, bigram, and longest common subsequence overlaps for each model. In the case of BLEU,

TABLE I. BLEU Scores

LLM	BLEU	1-gram	2-gram	3-gram	4-gram	BP	LR
Gemma-7b-it	0.106744	0.207503	0.10953	0.082933	0.068879	1	1.733348
Mistral-7B-Instruct-v0.3	0.07296	0.15571	0.079069	0.054448	0.042271	1	2.366448
SeaLLMs-v3-1.5B-Chat	0.071617	0.149566	0.077573	0.054071	0.041933	1	3.379995
Phi-3-mini-instruct	0.015101	0.044451	0.014785	0.010182	0.007772	1	4.038158
falcon-7b-instruct	0.013154	0.048097	0.014359	0.007968	0.00544	1	4.187401

both the overall BLEU score and the precision scores for 1-gram to 4-gram overlaps were extracted. Moreover, the performance of each LLM's integration into the pipeline were also evaluated using three metrics, namely the *Context Relevance* (CR), *Answer Relevance* (AR), and *Groundedness* (G) metrics based on the implementation in BeyondLLM [35]. Each of these metrics form a benchmark that measures the effectiveness of both the generated outputs based on their relevance with their corresponding query and generated context as evaluated by an LLM.

An integer score on the scale of 0 to 10 is assigned based on an assessment of relevance made by the LLM on each metric, where 0 corresponds to a generated result that is least relevant to the context and contains no connections to it, and 10 is a generated result that is most relevant and fully supported by the context. The Context Relevance metric is a rating of the relevance of the generated context to the initial query, while the Answer Relevance metric assesses the relevance of the generated answer to the initial query. The Groundedness metric, on the other hand, is a measure of the relevance of the generated answer to its corresponding generated context. The evaluation pipeline was built on LangChain's *ChatPromtTemplate* class [33]. The LLM that was used for the evaluation of the pipeline is OpenAI's gpt-4o-mini [36] with temperature equal to zero to avoid randomness. Each question and their corresponding generated outputs for each of the selected LLMs has its own scores for Context Relevance, Answer Relevance, and Groundedness, which are then used to aid the discussion in the succeeding section.

IV. Results and Discussion

A. BLEU

The summary of results is shown in Table I. The BLEU metric reveals a nuanced hierarchy among evaluated models. Gemma-7b-it consistently achieves the highest scores, surpassing both Mistral-7b-instruct-v0.3 and seallms-v3-1.5b-chat, which are closely matched. An important observation is that Gemma-7b-it generates the shortest responses in comparison to the other models. This result is closely tied to BLEU's nature as a precision-oriented metric, which tends to penalize excessively verbose outputs. As the length of the generated text increases relative to the reference, the likelihood of producing non-matching n-grams also rises, thereby diluting the proportion of correct n-grams and reducing the overall precision score. Thus, brevity can lead to improved BLEU scores by limiting potential mismatches with the reference text.

B. ROUGE

The summary of results is shown in Table II. The Seallms-v3-1.5b-chat model emerged as the top performer across all three scores, demonstrating a commendable ability to extract and present key information. This was evidenced by a moderate level of content similarity with reference answers. It was closely followed by the Mistral-7b-instruct-v0.3 and Gemma-7b-it models. Conversely, Falcon-7b-instruct and Phi-3-mini-4k-instruct lagged behind, producing more unintelligible responses than coherent ones. Both Falcon and Phi struggled with fact recall and contextual understanding, resulting in outputs that did not align well with the reference materials. Notably, Mistral-7b-instruct-v0.3 performed best when the questions and contexts were presented in plain English, followed by Gemma-7b-it and Seallms-v3-1.5b-chat. However, when both questions and contexts were in Tagalog, Seallms-v3-1.5b-chat took the lead, with Gemma-7b-it and Mistral-7b-instruct-v0.3 trailing by a significant margin. Additionally, Seallms-v3-1.5b-chat excelled in code-switched Tagalog-English scenarios, outperforming the competition and followed by Mistral-7b-instruct-v0.3 and Gemma-7b-it.

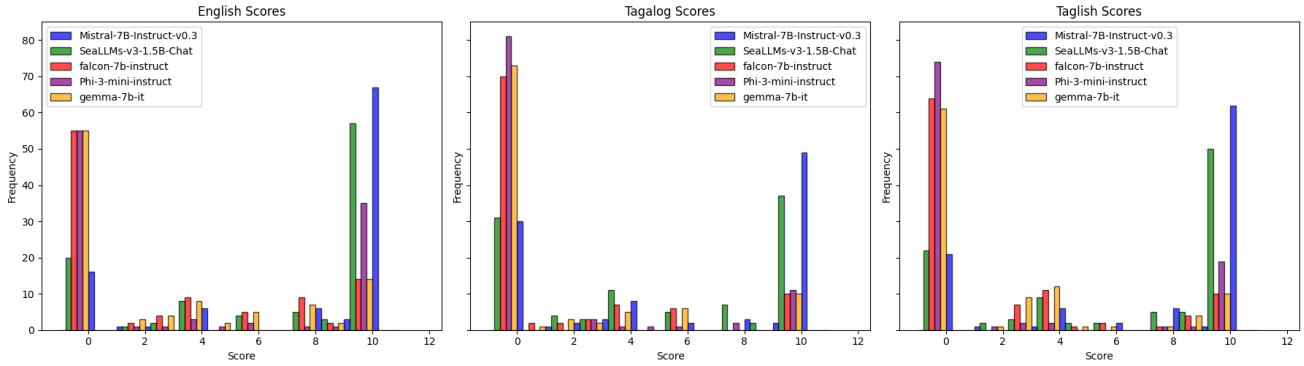
TABLE II. ROUGE Scores

LLM	Rouge-1	Rouge-2	Rouge-L
SeaLLMs-v3-1.5B-Chat	0.234581	0.137146	0.208251
Mistral-7b-instruct-v0.3	0.210862	0.121331	0.197133
Gemma-7b-it	0.205109	0.121382	0.187456
Falcon-7b-instruct	0.075681	0.22684	0.069577
Phi-3-mini-instruct	0.054189	0.022935	0.048623

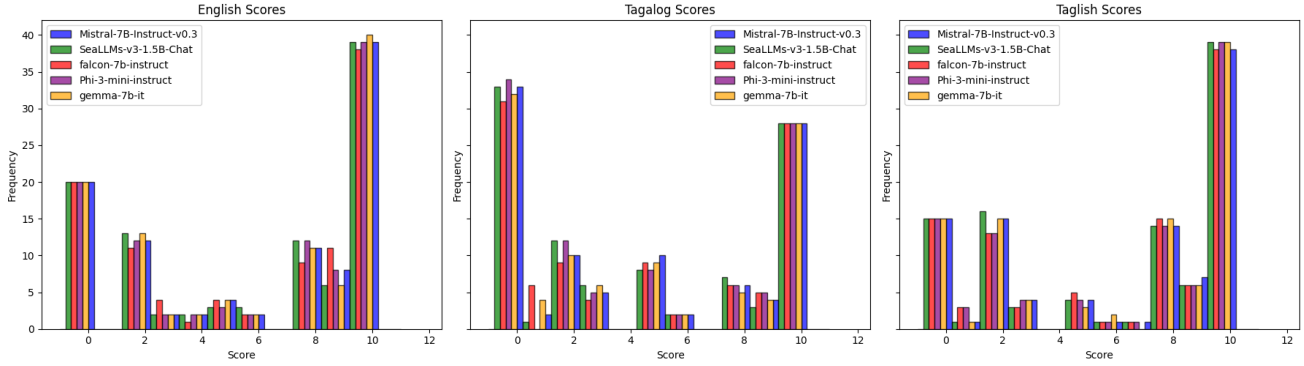
C. RAG Triad

Since BLEU and ROUGE have their limitations, the RAG triad was also employed for evaluation to account for semantic meaning and sentence structure. The distribution of scores across three languages are visualized in Fig. 2, with Fig. 2a, Fig. 2b, and Fig. 2c showing Groundedness, Context Relevance, and Answer Relevance, respectively. Most of the generated responses for both Mistral and SeaLLM show good alignment with the given context, suggesting that both models are generally successful in generating grounded responses. On the other hand, the models Falcon, Phi-3-mini, and Gemma typically exhibit subpar performance in this aspect, with the majority of their scores centered around zero. The remainder of the scores for these models exhibit a sparse distribution across various values, implying a deficiency in producing answers that are pertinent and contextually grounded.

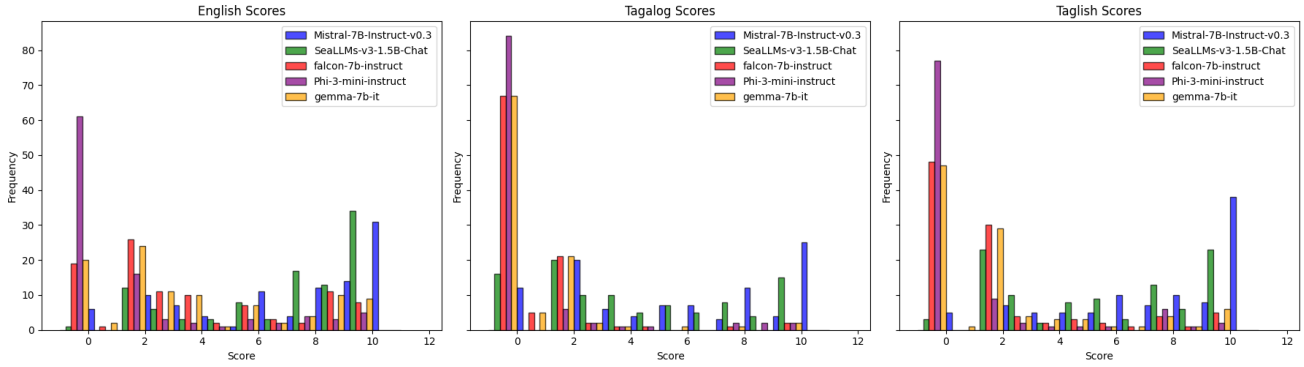
All five LLMs could retrieve relevant chunks for most queries particularly when fed with English and Taglish queries, as evidenced by the high Context Relevance for these languages in Fig. 2b. This high context retrieval performance may be helped by the emergent multilinguality of LLMs, due to the fact that they are pre-trained on inherently multilingual data [37], although skewed toward



(a) Groundedness (G)



(b) Context Relevance (CR)



(c) Answer Relevance (AR)

Fig. 2. Histograms of metric scores over the five open-source LLMs evaluated using the benchmark metrics of Groundedness (G), Context Relevance (CR) and Answer Relevance (AR). Each column corresponds to the query language of English, Tagalog, and Taglish respectively. The scores of each query fall on an integer scale of 0 to 10, where a score of 10 corresponds to a generated output being completely relevant to the context or query, while a score of zero corresponds to the absence of any relevance between the generated output and the context or query.

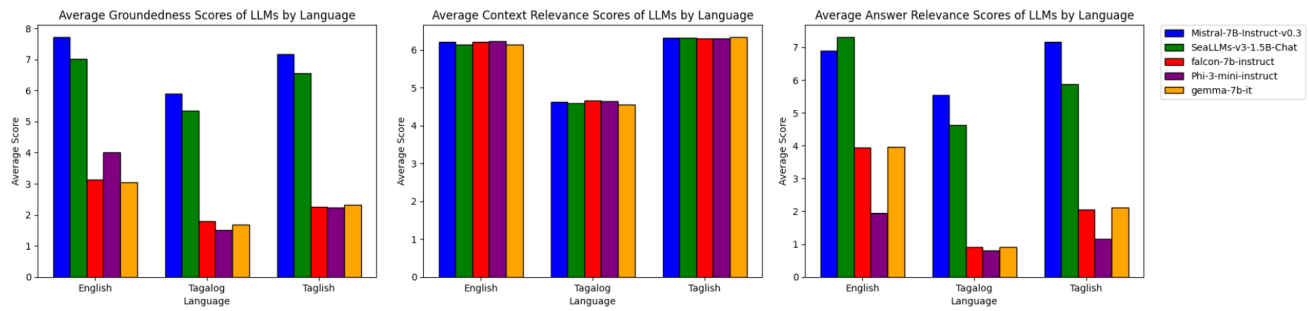


Fig. 3. Average Groundedness, Context Relevance, and Answer Relevance of each LLM.

high-resource languages that are well represented in the training data, such as English [38], [39]. The retrieval performance for purely Tagalog queries highlights this gap in monolingual performance, with a greater frequency value of zero-rated Context Relevance than other scores.

The scores for Answer Relevance in Fig. 2c vary widely between two groups of LLMs, with Mistral and SeaLLM exhibiting a higher frequency of high scores than those of Falcon, Phi-3-mini, and Gemma, whose Answer Relevance score frequency skews toward zero. A low Answer Relevance score reflects the model's inability to generate an appropriate answer to the given query [35], which indicates lacking performance in what's regarded as an important component of a question answering task. The low scores of the latter group of three LLMs are manifested similarly to their low Groundedness score, further highlighting issues in the generated answers for these LLMs.

D. Performance Across Languages

The benchmark results of each LLM on their integration into the pipeline are presented in Fig. 3. It reveals a clear trend in the performance of the models across different languages, particularly in terms of Groundedness and Answer Relevance. The models generally perform best on English queries, followed by Taglish, with Tagalog queries yielding the lowest performance. Notably, Mistral-7B-Instruct-v0.3 demonstrates a strong ability to provide query-relevant answers when the queries are in Taglish. Mistral and SeaLLM stand out as the top performers in both Groundedness and Answer Relevance across languages. In contrast, when considering Context Relevance, the models are largely comparable, with only a slight edge observed in Taglish queries over English ones. Regarding consistency across languages, Gemma-7b-it exhibits the most stable performance in Groundedness, while Mistral-7b-instruct-v0.3 is the most consistent in Context Relevance. Phi-3-mini-instruct is most consistent in Answer Relevance, as shown in Table III. However, Falcon-7b-instruct emerges as the overall most consistent model across all three metrics, demonstrating balanced performance regardless of the query language.

TABLE III. Standard Deviation of Groundedness, Context Relevance, and Answer Relevance for Various LLMs

LLM	Standard Deviation		
	G	CR	AR
Mistral-7b-instruct-v0.3	4.2243	4.1599	3.3926
SeaLLMs-v3-1.5B-Chat	4.2268	4.1665	3.3816
Falcon-7b-instruct	3.5969	4.1673	2.9903
Phi-3-mini-instruct	4.1781	4.1817	2.7140
Gemma-7b-it	3.5648	4.1736	3.0498

E. Comparative Performance Across Document Types

To investigate whether the file format of the original document may affect the performance of each LLM in the pipeline, Fig. 4 presents the average metric scores of each LLM based on document type, whether `.txt` or `.pdf`. The differences in metric scores between document types across all LLMs are normally distributed based on Shapiro-Wilk Tests. However, based on paired t-tests, the models are generally better at retrieving query-relevant contexts from `.pdf` documents. In terms of Groundedness and Answer Relevance, the differences in scores for these metrics based on file format are not significant.

F. Impact of Code-Switching on Performance

For this section, the impact of intra-sentence code-switching and the combination of intra-sentence and intra-word code-switching on answer relevance is examined first. The summary is shown in Table IV. Among the models tested, mistral-7b-instruct-v0.3 demonstrated the best overall performance across different languages, maintaining consistency in both English and code-switched Taglish queries. SeaLLMs-v3-1.5b-chat followed closely, with consistent performance on both Tagalog and Taglish inputs. Meanwhile, Gemma-7b-it and falcon-7b-instruct showed comparable performance, while phi-3-mini-instruct performed the poorest among the models. In general, a decline in performance was observed when the models were presented with code-switched questions compared to questions posed in pure English. However, the models performed better on Taglish (code-switched) questions than on questions presented entirely in Tagalog. Additionally, performance slightly improved when the questions involved both intra-sentence and intra-word code-switching, as opposed to intra-sentence code-switching alone. Notably, mistral-7b-instruct-v0.3 showed enhanced responses to questions containing both types of code-switching. It is also worth noting that 4 out of the 9 documents used for this evaluation contained code-switching between English and Tagalog.

This time, the effect of code-switching in general to RAG triad results is examined. As seen in Fig. 3, there is a high average performance for English language queries as opposed to Tagalog or queries across all LLMs. This is an expected result for high-resource languages such as English when compared to medium- or low-resource languages such as Tagalog, where a large gap exists between different resource levels in their relative performance of basic NLP tasks [40]. Table V compares the performance of English, Tagalog, and code-switched Taglish queries across three metrics: Groundedness, Context Relevance, and Answer Relevance. The results indicate that models generally perform best on purely English queries, especially regarding Groundedness and Answer Relevance. This superior performance is likely due to the extensive availability of English training data, which enables models to generate more accurate and contextually relevant responses.

Interestingly, models also perform better on Taglish queries compared to pure Tagalog queries. This enhanced performance can be attributed to the integration of English terms and syntactic structures within Taglish, which aligns more closely with the models' training data. In contrast, the performance on pure Tagalog queries is comparatively lower, reflecting the limited availability of training data in Tagalog, which hinders the models' ability to produce precise and relevant answers. Although these values are fairly significant, ANOVA suggests that only context relevance is significantly affected by code-switching with a p-value of 5.04×10^{-17}. Tukey's HSD test further suggests that the models' performance on Taglish queries is significantly better than on both English and Tagalog queries.

TABLE IV. Impact of Code-switching Types on Performance

LLM	Intra-sentence			Intra-sentence and Intra-word		
	English	Tagalog	Taglish	English	Tagalog	Taglish
Mistral-7B-Instruct-v0.3	7.416667	0.625	1.125	3.8	0.8	1.5
SeaLLMs-v3-1.5B-Chat	7.5	5.041667	5.833333	7.1	5.3	5.6
falcon-7b-instruct	2.958333	0.625	1.125	3.8	0.8	1.5
gemma-7b-it	2.916667	0.583333	1.125	3.9	0.85	1.9
Phi-3-mini-instruct	0.541667	0.875	0.75	2.8	0.7	1.1

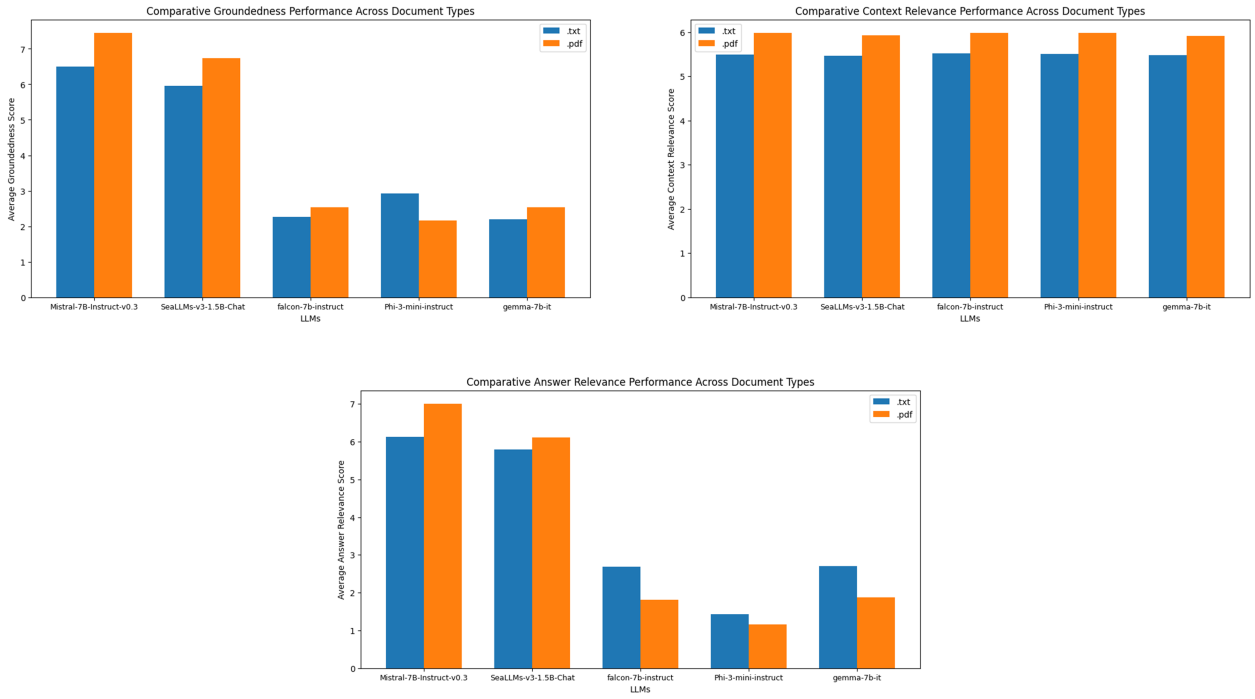


Fig. 4. Performance of the five open-source LLMs on different file types, namely `.txt` and `.pdf` files.

G. Model Size and Performance

The selection of LLMs in this work represents various model sizes and complexities among the current and state-of-the-art. It is prudent to assess the possibility of a correlation between model size and RAG performance, and whether 'lighter' models that can run on minimal or constrained resources can perform as well as "heavier" models, which may require more resources, incur higher operating costs or necessitate a reliance on paid third-party or private entities. For instance, the two LLMs with high average Groundedness and Answer Relevance across each language according to Fig. 3 are Mistral and SeaLLM. Both have different model sizes at 7 billion and 1.5 billion parameters respectively, while maintaining comparably similar performance over all evaluation metrics.

Overall, the performance gap between the two groups of LLMs (Mistral and SeaLLM vs. Falcon, Phi-3-mini, and Gemma) across all three metrics—groundedness, context relevance, and answer relevance—highlights a few critical insights:

Multilingual Capabilities: Models pre-trained on multilingual data, particularly Mistral and SeaLLM, show better performance in code-switched scenarios like Tagalog-English. Their higher Groundedness and relevance scores suggest that they are better equipped to handle linguistic transitions in mixed-language environments, making them more effective for multilingual NLP tasks.

Handling Lower-Resource Languages: The low scores of all models on purely Tagalog queries emphasize the persistent challenges in handling lower-resource languages. This issue is especially apparent in context relevance, where models often fail to retrieve or generate relevant information for Tagalog queries, likely due to the underrepresentation of the language in training data.

Hallucination Risk: Lower Groundedness and answer relevance scores for models like Falcon and Phi-3-mini point to a higher risk of hallucination, where the model generates responses that are either factually incorrect or contextually inappropriate. Improving Groundedness by fine-tuning these models on domain-specific or code-switched datasets may help mitigate this issue.

H. Text Generation Errors

The five models demonstrate varying error types that highlight different areas for improvement in text generation and comprehension. Common error types include contextual inaccuracies, semantic misinterpretation of context, and issues with syntactic coherence and relevance. Contextual

TABLE V. Comparative Performance of Monolingual English (EN) and Tagalog (TL) Queries versus Taglish Queries

Model	Groundedness		Context Relevance		Answer Relevance	
	EN vs Taglish	TL vs Taglish	EN vs Taglish	TL vs Taglish	EN vs Taglish	TL vs Taglish
Mistral-7B-Instruct-v0.3	−7.12%	21.53%	1.77%	36.29%	3.92%	29.06%
SeaLLMs-v3-1.5B-Chat	−6.42%	22.62%	2.94%	37.77%	−19.70%	27.06%
falcon-7b-instruct	−27.56%	26.26%	1.45%	35.19%	−47.97%	127.78%
Phi-3-mini-instruct	−44.39%	48.67%	1.12%	35.78%	−40.51%	43.21%
gemma-7b-it	−23.68%	37.28%	3.26%	39.34%	−46.46%	132.97%

inaccuracies can be observed across all models where they provide specific but inaccurate answers. Generally, semantic and contextual misinterpretations are common where models sometimes struggle with matching the details in the context with the specific queries. The higher tier models such as Mistral, SeaLLM, and Gemma usually exhibit this by stating that the answer was not found in the given context. Phi3mini and Falcon on the other hand often fail to extract or correctly infer information from the provided context. These models tend to struggle particularly when the questions require detailed extraction or precise interpretation. Additionally, syntactic coherence and generation issues vary in severity but are especially pronounced in Phi3mini, Falcon, and Gemma, which often produces fragmented or repetitive responses that do not answer the questions directly or does not make any sense at all. These indicate breakdowns in the generation process. These errors are more pronounced when the queries or contexts are in Tagalog or Taglish where the generated answers are gibberish. Syntactic errors can be mitigated through fine-tuning with language-specific grammar-focused data which emphasizes proper grammar and sentence structure such as grammar correction corpora or educational materials designed for language learners. Contrastive learning can also be employed to help the model distinguish between syntactically correct and incorrect sentences. Semantic errors on the other hand can be mitigated through better embedding process. Improperly parsed documents may lead to semantically inaccurate context chunks which can greatly influence the LLM's response.

V. Conclusion

In this work, five open-source LLMs are evaluated for their RAG performance on code-switched English-Tagalog queries. Our results have led to a greater understanding of the dynamics of LLM-based RAG pipelines on both varied query language, code-switching, and model complexity, prompting the selection of appropriate models based on their performance in this particular task. The study revealed Mistral and SeaLLM consistently outperformed the others in generating contextually grounded and relevant answers, attributable to their advanced architectures and extensive multilingual training data. Performance analysis indicated that the models excelled with English queries, followed by Taglish, and performed the worst with Tagalog queries, underscoring the need for more balanced training data that includes low-resource languages. Additionally, models generally retrieved query-relevant contexts more effectively from `.pdf` documents compared to `.txt` files. The evaluation considered three criteria: Groundedness, Context Relevance and Answer Relevance, with Mistral and SeaLLM excelling in all three. ANOVA analysis demonstrated that Context Relevance is significantly affected by code-switching, with Taglish queries outperforming both English and Tagalog queries, indicating a statistically significant difference.

The researchers suggest focusing on key areas to improve language models in multilingual and code-switched contexts. First, a detailed analysis of the language composition within training datasets is essential to identify and mitigate potential biases, particularly by quantifying the proportions of English, Tagalog, and Taglish present. Expanding the Tagalog dataset by including diverse sources like literature, news, and social media can enhance the model's accuracy and relevance for Tagalog queries. Additionally, studying the effects of code-switching in Taglish on models' understanding could reveal how mixed-language elements impact comprehension and retrieval, guiding the development of more effective multilingual models. Investigating user code-switching behavior will provide insights into real-world language use, enabling the design of models better suited to actual usage patterns. Among further potential research directions, a survey of RAG performance for queries and documents of other code-switched language pairs, dialects, creole languages, or low-resource languages remains relatively unexplored. Research efforts in these areas would greatly contribute to the accessibility of RAG among different NLP applications, benefiting the broader community of underrepresented language users.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

AJMA proposed and guided the overall research direction, performed initial benchmarking experiments, and supervised all revisions; JADVC conducted literature review and dataset curation, performed experiments, performed data analysis, and conducted benchmarking experiments; MLBM conducted preliminary research; APBO conducted data collection and initial experiments; ECP conducted literature reviews, as well as reviewed the revisions to the paper; MRMM conducted literature reviews and dataset curation. All authors co-authored the paper. All authors had approved the final version.

Funding

This research has been funded by the Department of Science and Technology - Advanced Science and Technology Institute through the iTANONG project budget allocated from the agency's fund from the General Appropriations Act (GAA) FY 2024.

ACKNOWLEDGMENT

The authors would like to thank and acknowledge the Computer Software Division of DOST-ASTI, primarily the iTANONG project staff, for their unwavering support and contributions to this study. Additionally, we would like to acknowledge our interns Giovanna Jaden Buitizon and Rica Mae Tugad from the Philippine Science High School system for their contributions to the curation of the dataset used in the study. Finally, we extend our gratitude to DOST-ASTI management for allocating the budget for the iTANONG project from the agency's GAA fund.

REFERENCES

[1] W. X. Zhao, K. Zhou, J. Li *et al.*, "A survey of large language models," *ArXiv*, vol. abs/2303.18223, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257900969

[2] Y. Moslem, R. Haque, J. D. Kelleher *et al.*, "Adaptive machine translation with large language models," in *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, M. Nurminen, J. Brenner, M. Koponen *et al.*, Eds. Tampere, Finland: European Association for Machine Translation, Jun. 2023, pp. 227–237. [Online]. Available: https://aclanthology.org/2023.eamt-1.22

[3] W. Zhu, H. Liu, Q. Dong *et al.*, "Multilingual machine translation with large language models: Empirical results and analysis," in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 2765–2781. [Online]. Available: https://aclanthology.org/2024.findings-naacl.176

[4] G. Lee, V. Hartmann, J. Park *et al.*, "Prompted LLMs as chatbot modules for long open-domain conversation," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4536–4554. [Online]. Available: https://aclanthology.org/2023.findings-acl.277

[5] M. Döbler, R. Mahendravarman, A. Moskvina *et al.*, "Can I trust you? LLMs as conversational agents," in *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, A. Deshpande, E. Hwang, V. Murahari *et al.*, Eds. St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 71–75. [Online]. Available: https://aclanthology.org/2024.personalize-1.5

[6] Y. Gao, Y. Xiong, X. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *ArXiv*, vol. abs/2312.10997, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266359151

[7] S. Cahyawijaya, H. Lovenia, and P. Fung, "Llms are few-shot in-context low-resource language learners," *ArXiv*, vol. abs/2403.16512, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268680591

[8] L. Nguyen, Z. Yuan, and G. Seed, "Building Educational Technologies for Code-Switching: Current practices, difficulties and future directions," *Languages*, vol. 7, no. 3, p. 220, 8 2022. [Online]. Available: https://doi.org/10.3390/languages7030220

[9] M. L. Visperas, C. J. Borjal, A. J. M. Adoptante *et al.*, "iTANONG-DS : A collection of benchmark datasets for downstream natural language processing tasks on select Philippine languages," in *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, M. Abbas and A. A. Freihat, Eds. Online: Association for Computational Linguistics, Dec. 2023, pp. 316–323. [Online]. Available: https://aclanthology.org/2023.icnlsp-1.34

[10] Z. Bai, P. Wang, T. Xiao *et al.*, "Hallucination of multimodal large language models: A survey," *arXiv preprint arXiv:2404.18930*, 2024.

[11] S. Sitaram, K. R. Chandu, S. K. Rallabandi *et al.*, "A survey of code-switched speech and language processing," *CoRR*, vol. abs/1904.00784, 2019. [Online]. Available: http://arxiv.org/abs/1904.00784

[12] A. Asai, S. Longpre, J. Kasai *et al.*, "MIA 2022 shared task: Evaluating cross-lingual open-retrieval question answering for 16 diverse languages," in *Proceedings of the Workshop on Multilingual Information Access (MIA)*, A. Asai, E. Choi, J. H. Clark *et al.*, Eds. Seattle, USA: Association for Computational Linguistics, Jul. 2022, pp. 108–120. [Online]. Available: https://aclanthology.org/2022.mia-1.11

[13] V. Karpukhin, B. Oğuz, S. Min *et al.*, "Dense passage retrieval for Open-Domain question answering," *arXiv (Cornell University)*, 1 2020. [Online]. Available: https://arxiv.org/abs/2004.04906

[14] P. Lewis, E. Perez, A. Piktus *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks," *arXiv (Cornell University)*, 1 2020. [Online]. Available: https://arxiv.org/abs/2005.11401

[15] R. Zhao, H. Chen, W. Wang *et al.*, "Retrieving multimodal information for augmented generation: A survey," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4736–4756. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.314

[16] P. Finardi, L. Avila, R. Castaldoni *et al.*, "The Chronicles of RAG: The Retriever, the Chunk and the Generator," *arXiv e-prints*, p. arXiv:2401.07883, jan 2024.

[17] D. Rau, S. Wang, H. Déjean *et al.*, "Context embeddings for efficient answer generation in RAG," *arXiv (Cornell University)*, 7 2024. [Online]. Available: https://arxiv.org/abs/2407.09252

[18] S. R. Ahmad, "Enhancing multilingual information retrieval in mixed human resources environments: a RAG model implementation for multicultural enterprise," *arXiv (Cornell University)*, 1 2024. [Online]. Available: https://arxiv.org/abs/2401.01511

[19] I. Ahmed and R. Islam, "Gemini-the most powerful llm: Myth or truth," mar 2024. [Online]. Available: http://dx.doi.org/10.36227/techrxiv.171177477.70151414/v1

[20] K. Shi, X. Sun, Q. Li *et al.*, "Compressing long context for enhancing rag with amr-based concept distillation," *ArXiv*, vol. abs/2405.03085, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269605470

[21] J. Kirchenbauer and C. Barns, "Hallucination reduction in large language models with retrieval-augmented generation using wikipedia knowledge," may 2024. [Online]. Available: http://dx.doi.org/10.31219/osf.io/pv7r5

[22] M. Fatehkia, J. Lucas, and S. Chawla, "T-rag: Lessons from the llm trenches," *ArXiv*, vol. abs/2402.07483, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267627486

[23] X.-P. Nguyen, W. Zhang, X. Li *et al.*, "SeaLLMs - large language models for Southeast Asia," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Y. Cao, Y. Feng, and D. Xiong, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 294–304. [Online]. Available: https://aclanthology.org/2024.acl-demos.28

[24] P. Finardi, L. Avila, R. Castaldoni *et al.*, "The chronicles of rag: The retriever, the chunk and the generator," 2024. [Online]. Available: https://arxiv.org/abs/2401.07883

[25] K. Wu, E. Wu, and J. Zou, "ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence," *arXiv e-prints*, p. arXiv:2404.10198, Apr. 2024.

[26] Y. Tang and Y. Yang, "MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries," *arXiv e-prints*, p. arXiv:2401.15391, Jan. 2024.

[27] A. Q. Jiang, A. Sablayrolles, A. Mensch *et al.*, "Mistral 7B," *arXiv e-prints*, p. arXiv:2310.06825, Oct. 2023.

[28] E. Almazrouei, H. Alobeidli, A. Alshamsi *et al.*, "The Falcon Series of Open Language Models," *arXiv e-prints*, p. arXiv:2311.16867, Nov. 2023.

[29] M. Abdin, J. Aneja, H. Awadalla *et al.*, "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone," *arXiv e-prints*, p. arXiv:2404.14219, Apr. 2024.

[30] G. Team, T. Mesnard, C. Hardin *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[31] A. Do and S. Tran, "Improving context awareness of transformer networks using retrieval-augmented generation," Ph.D. dissertation, KTH, School of Engineering Sciences (SCI), 2024. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-348512

[32] K. Carolan, L. Fennelly, and A. F. Smeaton, "A review of multi-modal large language and vision models," *arXiv preprint arXiv:2404.01322*, 2024.

[33] "Langchain," 10 2022. [Online]. Available: https://github.com/langchain-ai/langchain

[34] M. Douze, A. Guzhva, C. Deng *et al.*, "The Faiss library," *arXiv e-prints*, p. arXiv:2401.08281, Jan. 2024.

[35] T. Jain, A. Hegde, and M. Taha, "BeyondLLM by AI Planet." [Online]. Available: https://github.com/aiplanethub/beyondllm

[36] OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. [Online]. Available: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[37] S. Minaee, T. Mikolov, N. Nikzad *et al.* Large Language Models: A Survey. [Online]. Available: http://arxiv.org/abs/2402.06196

[38] P. Joshi, S. Santy, A. Budhiraja *et al.*, "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 6282–6293. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.560

[39] N. Chirkova, D. Rau, H. Déjean *et al.* Retrieval-augmented generation in multilingual settings. [Online]. Available: http://arxiv.org/abs/2407.01463

[40] M. A. Hedderich, L. Lange, H. Adel *et al.*, "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer *et al.*, Eds. Association for Computational Linguistics, pp. 2545–2568. [Online]. Available: https://aclanthology.org/2021.naacl-main.201