Adapting LLMs to New Domains: A Comparative Study of Fine-Tuning and RAG strategies for Portuguese QA Tasks

Leandro Yamachita da Costa[1], João Baptista de Oliveira e Souza Filho[1]

[1]Programa de Engenharia Elétrica
Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro, RJ – Brazil
leandro.yamachita@coppe.ufrj.br, jbfilho@poli.ufrj.br

Abstract. *The rise of Large Language Models (LLMs) represented a significant advance in text generation applications. However, LLMs face challenges in domains outside the scope of their original training. This study investigates the following two approaches to adapt LLMs to new domains in the context of generative question-answering (QA) with data in Portuguese: fine-tuning and Retrieval-Augmented Generation (RAG). The experiments carried out in this study demonstrate the effectiveness of incorporating external data sources, even in models that had not been adjusted for the specific domain. Furthermore, the combination of supervised fine-tuning with RAG proved to be the most effective approach.*

1. Introduction

The rise of Large Language Models (LLMs) marked a significant advance in Natural Language Processing (NLP), especially in text generation tasks [Brown et al., 2020; Achiam et al., 2023]. These models, which are trained with large volumes of data, can retain vast amounts of knowledge implicitly in their parameters. However, LLMs face challenges in domains outside the scope of their original training data, such as areas of specialized knowledge or current affairs [Kandpal et al., 2023, Kasai et al., 2024]. This issue accentuates the need to adapt LLMs to specific contexts, especially for smaller models with limited memory capacity.

This study explores the adaptation of LLMs to new domains in the generative question-answer (QA) task, a scenario where the model generates answers based on questions provided to it. Fine-tuning is a common approach to adjust LLMs to new domains by modifying the model parameters with training on application-specific data. In the QA task, fine-tuning can be done with question-answers pairs in a "closed-book" scenario [Zhang et al., 2024], where the model does not have access to external information. Nonetheless, this approach may require considerable computational power and extensive data annotation work [Guo et al., 2023].

A widely adopted alternative is the integration of external knowledge sources, such as documents and books, in a setting known as "open-book" [Zhang et al., 2024]. A typical strategy of this approach is the Retrieval-Augmented Generation

(RAG) [Lewis et al., 2020], which combines an information retriever (IR) model, aimed at searching relevant data on an external data source, with a language model to generate answers based on the information retrieved. This strategy allows LLMs to adapt to new domains without the need for fine-tuning.

This work analyzed these two approaches to adapt LLMs to specific domains in the QA task with data in Portuguese. We analyzed both fine-tuning and RAG configurations, in addition to the integration of the two. The experiments demonstrated the effectiveness of incorporating external data sources for improved results. Moreover, different fine-tuning strategies have shown to be particularly effective when combined with the inclusion of external data, even with a reduced volume of training data. For all analysis, we considered a scenario with limited computational resources, where we only used a general-purpose GPU (Nvidia GeForce RTX 4090). The fine-tuning of the models was performed using the QLoRA technique [Dettmers et al., 2023] with quantized models.

2. Background and Related Work

In this session we briefly discuss RAG and LLMs fine-tuning.

2.1. Retrieval-augmented Language Models

Conditioning LLM responses to information from external sources has proven effective in adapting these models to specific domains in several NLP tasks. The RAG approach has been successfully applied in areas such as agriculture [Balaguer et al., 2024], scientific literature [Lála et al., 2023], and medical data [Zakka et al., 2024]; attesting its utility in improving the accuracy and relevance of answers. Additionally, it can reduce the occurrence of hallucinations [Borgeaud et al., 2022; Shuster et al., 2021], improve the model's ability to manage the gradual decline in its knowledge over time [Vu et al., 2023], and enhance the interpretability of the answers [Lewis et al., 2020; Izacard et al., 2023].

The effectiveness of RAG depends on the quality of the retrieval mechanism used, which impacts the relevance of the contextual information obtained. Among traditional retrieval mechanisms, those based on term frequency stand out, which employ sparse representations of text passages, such as TF-IDF [Sparck Jones, 1972] and BM25 [Robertson and Zaragoza, 2009]. Alternatively, more recent approaches employ dense representations of texts, such as Dense Passage Retriever (DPR) [Karpukhin et al., 2020] and ColBERT [Khattab and Zaharia, 2020].

2.2. Fine-tuning

To adapt QA models to a new domain, fine-tuning seeks to adjust the model to respond according to the pattern observed in the training data. Furthermore, it is expected that with fine-tuning the model will acquire domain-specific knowledge, enhancing its capacity to provide more accurate answers.

Fine-tuning large language models (LLMs) typically requires significant computational resources. Parameter Efficient Fine Tuning (PEFT) [Xu et al., 2023] addresses this by freezing the model's parameters and adjusting only the newly added ones. Among PEFT methods, Low-Rank Adaptation (LoRA) [Hu et al., 2021] reduces the number of trainable parameters using low-rank matrices. The Quantized Low-Rank Adaptation (QLoRA) [Dettmers et al., 2023] takes this approach a step further by applying LoRA to quantized models. Studies indicate that PEFT-tuned models often perform comparably to those fully fine-tuned [Li et al., 2023].

2.3. RAG and Fine-tuning

The effectiveness of RAG and fine-tuning strategies has been extensively studied. [Balaguer et al., 2024] compare these methods in a QA model for agriculture, while [Ovadia et al., 2023] extend the comparison to various topics with a multiple-choice QA model. Many RAG models undergo fine-tuning, such as [Lewis et al., 2020], where both the model and retriever are adjusted together. [Zhang et al., 2024] introduce RAFT (Retrieval-Augmented Fine-Tuning), which helps the model ignore irrelevant documents. Overall, pre-trained models require additional fine-tuning to learn specific reading comprehension tasks, which is essential for the effectiveness of RAG. This instruction fine-tuning does not always need to be done with domain-specific data.

3. Methodology

3.1. Language Models

For the experiments in this study, we used a Portuguese-adapted version of the model T5 [Raffel et al., 2020], called PTT5 [Carmo et al., 2020], and two versions of the Llama-3 8B model [AI at Meta, 2024].

PTT5 was pre-trained on the BrWac [Wagner et al., 2018], a dataset composed of millions of Internet pages in Brazilian Portuguese. This study used the base version of the model, which contains 220M parameters. This choice was due to its relatively small size, Portuguese pre-training, and encoder-decoder architecture, distinguishing it from Llama 3. Llama 3, which was developed by Meta, uses a decoder-only architecture and is available in both pre-trained and instruction-tuned versions. Despite being trained mainly on English data, Llama 3 is multilingual and was evaluated exclusively with Portuguese data in this study. The 8-billion-parameter version of Llama was chosen for its suitability to limited computational resources and its popularity as a widely used open-source model.

3.2. Fine-tuning Approach

For fine-tuning the models, we used two different techniques: full parameter fine-tuning and the QLoRa technique [Dettmers et al., 2023]. Full fine-tuning was

applied only to the PTT5 model, due to its reduced size and the limited availability of computational resources. For the Llama 3-8B models, we chose the QLoRa technique due to the large number of parameters in these models.

3.3. RAG Setup

We adopted a basic RAG setup, in which we used the retriever and language models with no changes to their original architectures. Three retriever models were evaluated: BM25 [Robertson and Zaragoza, 2009], Dense Passage Retriever (DPR) [Karpukhin et al., 2020], and ColBERT [Khattab and Zaharia, 2020]. With DPR, the embeddings were generated by a BERT-based [Devlin et al., 2019] model known as Sentence-BERT [Reimers and Iryna, 2019]. In the case of ColBERT, we specifically used its second version - ColBERTv2 [Santhanam et al., 2021]. All IR models were used without any type of training on the data under study. The texts retrieved for each query were concatenated and inserted into the input prompt of the LLMs as support texts.

3.4. Evaluation Setup

We considered four metrics to evaluate the results of the experiments: (i) Rouge-1 and (ii) Rouge-L [Lin, 2004], which are based on word overlap; (iii) BERTScore [Zhang et al., 2019], which employs embeddings generated by a BERT model; and (iv) a specific metric developed with the use of GPT (GPT 4o mini).

GPT was used to verify the accuracy of the answers generated by the models. To achieve this, we developed the GPTScore metric, which evaluates whether the models' answers align with the content of the reference answers, even when they differ in wording, length, or style. This evaluation used the prompt shown in Figure 1. The metric was computed by tallying the number of "yes" or "no" answers provided by GPT.

```
Você está avaliando a saída de um modelo de linguagem de perguntas e respostas. O modelo recebe uma pergunta e um contexto
com informações que ajudam a responder à pergunta. O modelo responde com base no contexto oferecido. Abaixo estão
apresentados o contexto, a pergunta, a resposta dada pelo modelo e a resposta correta de referência. Você deve avaliar se a
resposta dada pelo modelo contém informações equivalentes às da resposta de referência. As respostas não precisam ser idênticas
e podem apresentar a mesma informação de formas diferentes; você deve avaliar somente se as informações apresentadas na
resposta são equivalentes e não a forma de apresentação das informações. Responda apenas com "sim" ou "não".

Contexto: {contexto}
Pergunta: {pergunta}
Resposta do modelo: {resposta_modelo}
Resposta referência: {resposta_referência}
```

Figure 1. Prompt used to obtain the GPTScore.

4. Experiments and Results

This section outlines the experimental setup and presents the results obtained using the RAG and fine-tuning strategies.

4.1. Experimental Setup

We utilized two datasets in Portuguese. The first one, Pirá 2.0 [Paschoal et al., 2021; Pirozelli et al., 2024], focuses on topics related to the Brazilian coast, oceans and climate change. This dataset contains questions, answers, and support texts derived from the abstracts of scientific papers and specialized reports on the aforementioned topics, all of which have a version in Portuguese. Comprising 2258 samples, this dataset was split as follows: 80% for training, 10% for validation, and 10% for testing. We selected this dataset because it includes texts in Portuguese and focuses on a specific domain. The second dataset is a Portuguese translation of the Databricks-Dolly dataset [Conover et al., 2023], which consists of pairs of instructions and answers across various task categories generated by Databricks employees. For this study, we only kept the records classified as "closed QA" task, which involves questions and answers based on excerpts from Wikipedia. This dataset contains 1766 records, distributed as follows: 70% for the training, 15% for validation, and 15% for test. It was included in this study because it contains questions from a broader domain that still require supporting texts for answer formulation.

The experiments conducted in this study aimed to explore different strategies for adapting LLMs to new domains, particularly focusing on fine-tuning and RAG-based approaches. We investigated various settings regarding the use of supporting texts, both in the fine-tuning process and during the validation stage. For the fine-tuning experiments, we evaluated two approaches: one that includes supporting texts and question-answer pairs in the input prompt, termed "RAG FT"; and another that utilizes only question-answer pairs, referred to as "QA FT". For validation, the scenarios in which support texts were included in the input prompt were called "RAG". For settings in which only the question was used as input, the following prompt was utilizes: “Responda à pergunta de forma sucinta.\n\nPergunta: {question}”. In cases where support texts were also included, the prompt was modified to: “Responda à pergunta de forma sucinta e com base no contexto dado. Contexto: {context}\n\n Pergunta: {question}”.

In our experiments, we employed a "greedy" decoding strategy, selecting the token with the highest probability during generation. We established a maximum output limit of 100 tokens, while the input limit was set at 1024 tokens to accommodate most of the supporting texts without truncation. Fine-tuning of the Llama models employed the QLoRa method with 4-bit quantization over 10 epochs, utilizing a batch size and gradient accumulation of 4 to optimize hardware capacity. The model generated answers at 16-bit precision. Meanwhile, the PTT5 model underwent full fine-tuning for 60 epochs, using a batch size of 8 and gradient accumulation of 4. All training was conducted using Hugging Face libraries on an NVIDIA GeForce RTX 4090 GPU.

To select the best retriever for the RAG experiments, we used GPT 4o mini to answer the questions in each dataset based on the context provided by each retriever. For each question, the texts retrieved by each model were concatenated and added to the prompt used by GPT to generate the answer. For the Pirá dataset, we used the four most relevant passages identified by each retriever, while for the Dolly dataset, we used the three most relevant passages. ColBERT outperformed all other models across the evaluated metrics and was, therefore, chosen for this study. Table 1 summarizes these results and includes a hypothetically ideal retriever, simulated by using context texts that are always correct for each question.

Table 1. Evaluation of the retriever methods (see text).

GPT4o-mini		Rouge-1 (F1)	Rouge-L (F1)	BertScore (F1)	GPTScore (Acc)
BM25	Pirá	0.2660	0.2344	0.7404	0.5947
	Dolly	0.3273	0.2807	0.7539	0.5819
DPR	Pirá	0.2142	0.1823	0.7146	0.4273
	Dolly	0.3373	0.2924	0.7594	0.6328
ColBERT	Pirá	**0.2723**	**0.2457**	**0.7476**	**0.6872**
	Dolly	**0.3410**	**0.2959**	**0.7612**	**0.6525**
Ideal	Pirá	0.3185	0.2831	0.7670	0.9295
	Dolly	0.4145	0.3658	0.7947	0.9068

4.2. Models without fine-tuning

The experiments with models without fine-tuning, considering only the question in the input prompt, may reveal their level of prior knowledge about the datasets' domain. The results for this setting, referred to as "No FT, No RAG" in Table 2, indicate that these models have low prior knowledge of the Pirá dataset's domain and moderate knowledge of the Dolly dataset. In this analysis, results for PTT5 models are not reported, as we were unable to obtain satisfactory answers from this model without fine-tuning.

In the case of RAG experiments with models without fine-tuning, referred to as "No FT, RAG", we observed an increase in GPTScore and a more modest rise in Rouge metrics. This behavior is expected, as Rouge metrics, which assess term matching, are more influenced by the style of the answers – particularly their length and vocabulary. Since the models were not fine-tuned to the datasets of interest, the generated answers may not align with the answer patterns from the dataset. It was observed that the pre-trained model often answered the questions and then continued generating question-answer pairs indefinitely until it reached the maximum number of output tokens. The Llama Instruct model performed significantly better on both the GPTScore and BERTScore due to its prior fine-tuning, which enhanced its reading comprehension abilities. This suggests that

models with advanced comprehension skills, even if trained in domains different from those being tested, can substantially benefit from the use of supporting contexts to leverage their performance. We can also observe that models without prior domain knowledge and that did not explore RAG were the worst performers.

4.3. Models fine-tuned solely with question-answer pairs

This setting aims to evaluate whether the models can internalize knowledge about the domains through the fine-tuning process, specifically based on questions and answers from the datasets, referred to as "QA FT, No RAG" in Table 2. The results show that for both datasets, fine-tuning does not provide a significant improvement when the model is tested without RAG. However, when the model includes RAG, referred to as "QA FT, RAG" in Table 2, we observed a meaningful gain in some evaluation scenarios. This suggests that the fine-tuning process helps the model learn the style of the answers - the length of the answers becomes more similar to that observed in the dataset - but does not necessarily enable it to retain domain knowledge. It is worth noting that the limited amount of training data may hinder the model's ability to learn effectively through the fine-tuning process.

4.4. Models fine-tuned with question, answer and context

In this scenario, the models were fine-tuned with the addition of contexts in the training prompts. In the validation setting without RAG, referred to as "RAG FT, No RAG", all models performed poorly. This result is expected, as the primary purpose of fine-tuning with added contexts is to train the model to generate answers based on the context itself. Since this setting does not include the support texts in the input prompts, fine-tuning did not appear to achieve the desired outcome.

In the setting that includes RAG, referred to as "RAG FT, RAG", the models achieved the best results across all metrics for the two datasets analyzed. It is worth noting that for the Llama Instruct model, which was already fine-tuned for the reading comprehension task, all settings that utilized RAG performed well according to GTPScore. However, for Rouge metrics, the models with fine-tuning on domain-specific data showed superior performance. This experiment suggests that even if the model is capable of extracting answers from the context, fine-tuning on problem-specific data may be beneficial for generating answers in a format more closely aligned with that found in the dataset. We also observed that fine-tuning the pre-trained Llama model allowed it to achieve results comparable to those of the Llama Instruct model, despite the latter being previously fine-tuned with a significantly larger amount of data. This result indicates that fine-tuning with context texts, even when performed with a reduced dataset, can enhance the model's ability to extract relevant information from context. In this setting, we also observed a significant improvement in the PTT5 results, which were clearly surpassed by those obtained with the Llama models, likely to their much larger number of parameters.

Table 2. Experimental results (see text).

Llama 3 PT (Pirá)	Rouge-1 (F1)	Rouge-L (F1)	BertScore (F1)	GPTScore (Acc)	Llama 3 PT (Dolly)	Rouge-1 (F1)	Rouge-L (F1)	BertScore (F1)	GPTScore (Acc)
No FT, No RAG	0.1091	0.0967	0.6305	0.0264	No FT, No RAG	0.1638	0.1420	0.6233	0.1412
No FT, RAG	0.1772	0.1525	0.6795	0.3624	No FT, RAG	0.2375	0.2029	0.6799	0.3505
QA FT, No RAG	0.1784	0.1628	0.7245	0.1101	QA FT, No RAG	0.2807	0.2464	0.7477	0.1554
QA FT, RAG	0.3322	0.3105	0.7802	0.4978	QA FT, RAG	0.3579	0.3238	0.7713	0.4718
RAG FT, No RAG	0.1420	0.1301	0.7101	0.0352	RAG FT, No RAG	0.2489	0.2204	0.7386	0.1554
RAG FT, RAG	**0.3711**	**0.3521**	**0.7905**	**0.5903**	RAG FT, RAG	**0.3815**	**0.3489**	**0.7805**	**0.5339**
Llama 3 Instruct (Pirá)					**Llama 3 Instruct (Dolly)**				
No FT, No RAG	0.1603	0.1358	0.6940	0.0529	No FT, No RAG	0.2498	0.2159	0.7305	0.2599
No FT, RAG	0.2824	0.2541	0.7458	**0.6035**	No FT, RAG	0.3166	0.2850	0.7498	**0.5678**
QA FT, No RAG	0.1830	0.1661	0.7236	0.0969	QA FT, No RAG	0.2798	0.2433	0.7502	0.1695
QA FT, RAG	0.3538	0.3307	0.7831	0.5859	QA FT, RAG	0.3698	0.3385	0.7756	0.5424
RAG FT, No RAG	0.1587	0.1438	0.7136	0.0793	RAG FT, No RAG	0.2600	0.2268	0.7438	0.1667
RAG FT, RAG	**0.3699**	**0.3520**	**0.7939**	**0.6035**	RAG FT, RAG	**0.3967**	**0.3593**	**0.7867**	0.5597
PTT5 (Pirá)					**PTT5 (Dolly)**				
QA FT, No RAG	0.1859	0.1708	0.7230	0.0925	QA FT, No RAG	0.2228	0.1971	0.7212	0.0169
QA FT, RAG	0.1379	0.1188	0.6666	0.0396	QA FT, RAG	0.1892	0.1524	0.6647	0.1158
RAG FT, No RAG	0.1282	0.1164	0.6929	0.0132	RAG FT, No RAG	0.2028	0.1837	0.7110	0.0198
RAG FT, RAG	**0.3028**	**0.2859**	**0.7642**	**0.3744**	RAG FT, RAG	**0.3134**	**0.2713**	**0.7478**	**0.2429**

5. Conclusion

This work analyzed various methods for adapting LLMs to specific domains in QA tasks, including fine-tuning the model and integrating external data through RAG. The experiments demonstrated that incorporating external data generally improves the models' performance, regardless of whether fine-tuning is applied. The results also showed that fine-tuning, even when conducted with a reduced dataset, can enhance the models' performance. Additionally, we observed that while the best results were achieved by models specifically tuned to domain data, a model with previously fine-tuned instructions produced similar outcomes, with the clear advantage of not requiring any additional fine-tuning.

The experiments presented here were conducted using a basic RAG architecture, without any additional training of the retrievers on the datasets of interest. Future work could explore the same settings with adjustments to the retrievers as well.

6. Acknowledgments

To CNPq, FAPERJ, and CAPES. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001.

7. References

Brown, Tom B. "Language models are few-shot learners." *In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)* (2020).

Achiam, Josh, OpenAI et al. "GPT-4 technical report." *arXiv preprint arXiv:2303.08774* (2023).

Kandpal, Nikhil, et al. "Large language models struggle to learn long-tail knowledge." *International Conference on Machine Learning*. PMLR, (2023).

Kasai, Jungo, et al. "REALTIME QA: What's the Answer Right Now?" *Advances in Neural Information Processing Systems* 36 (2024).

Zhang, Tianjun, et al. "RAFT: Adapting Language Model to Domain Specific RAG." *arXiv preprint arXiv:2403.10131,* (2024).

Guo, Kunpeng, et al. "Fine-tuning Strategies for Domain Specific Question Answering under Low Annotation Budget Constraints." *IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, (2023).

Lewis, Patrick, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems* 33: 9459-9474, (2020).

Dettmers, Tim et al. "QLoRA: Efficient Finetuning of Quantized LLMs." *ArXiv* abs/2305.14314, (2023).

Balaguer, Angels, et al. "RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture." *arXiv e-prints* (2024): arXiv-2401.

Lála, Jakub, et al. "PaperQA: Retrieval-augmented generative agent for scientific research." *arXiv preprint arXiv:2312.07559,* (2023).

Zakka, Cyril, et al. "Almanac—retrieval-augmented language models for clinical medicine." *NEJM AI* 1.2 (2024): AIoa2300068.

Borgeaud, Sebastian, et al. "Improving language models by retrieving from trillions of tokens." *International conference on machine learning*. PMLR, (2022).

Shuster, Kurt, et al. "Retrieval augmentation reduces hallucination in conversation." *arXiv preprint arXiv:2104.07567,* (2021).

Vu, Tu, et al. "FreshLLMs: Refreshing large language models with search engine augmentation." *arXiv preprint arXiv:2310.03214,* (2023).

Izacard, Gautier, et al. "Atlas: Few-shot learning with retrieval augmented language models." *Journal of Machine Learning Research* 24.251 (2023): 1-43.

Sparck Jones, Karen. "A statistical interpretation of term specificity and its application in retrieval." *Journal of documentation* 28.1 (1972): 11-21.

Robertson, Stephen E. and Hugo Zaragoza. “The Probabilistic Relevance Framework: BM25 and Beyond.” *Found. Trends Inf. Retr.* 3 (2009): 333-389.

Karpukhin, Vladimir, et al. "Dense passage retrieval for open-domain question answering." *arXiv preprint arXiv:2004.04906* (2020).

Khattab, Omar, and Matei Zaharia. "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval* (2020).

Xu, Lingling, et al. "Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment." *arXiv preprint arXiv:2312.12148* (2023).

Hu, Edward J., et al. "LoRA: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).

Li, Yixiao, et al. "LoftQ: Lora-fine-tuning-aware quantization for large language models." *arXiv preprint arXiv:2310.08659* (2023).

Ovadia, Oded, et al. "Fine-tuning or retrieval? comparing knowledge injection in LLMs." *arXiv preprint arXiv:2312.05934* (2023).

Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of machine learning research* 21.140 (2020): 1-67.

Carmo, Diedre, et al. "PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data." *arXiv preprint arXiv:2008.09144* (2020).

“Introducing Meta Llama 3: The most capable openly available LLM to date.” AI at Meta. (2024). https://ai.meta.com/blog/meta-llama-3/

Wagner Filho, Jorge A., et al. "The brWaC corpus: A new open resource for Brazilian Portuguese." *Proceedings of the eleventh international conference on language resources and evaluation* LREC (2018).

Devlin, Jacob et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *North American Chapter of the Association for Computational Linguistics* (2019).

Reimers, Nils and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” *Conference on Empirical Methods in Natural Language Processing* (2019).

Santhanam, Keshav et al. “ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction.” *North American Chapter of the Association for Computational Linguistics* (2021).

Lin, Chin-Yew. “ROUGE: A Package for Automatic Evaluation of Summaries.” *Annual Meeting of the Association for Computational Linguistics* (2004).

Zhang, Tianyi, et al. "BERTScore: Evaluating text generation with BERT." *arXiv preprint arXiv:1904.09675* (2019).

Paschoal, André FA, et al. "Pirá: A bilingual portuguese-english dataset for question-answering about the ocean." *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (2021).

Pirozelli, Paulo, et al. "Benchmarks for Pirá 2.0, a Reading Comprehension Dataset about the Ocean, the Brazilian Coast, and Climate Change." *Data Intelligence* 6.1 (2024): 29-63.

Conover, Mike, et al. "Free Dolly: Introducing the world’s first truly open instruction-tuned LLM." *Company Blog of Databricks* (2023). https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm