






Pt-HotpotQA: Evaluating Multi-Hop Question Answering on Original and Portuguese-translated Datasets Using LLMs

Sérgio S. Mucciaccia   [Federal University of Espírito Santo | sergio.mucciaccia@ufes.br]


Thiago M. Paixão  [Federal Institute of Espírito Santo | thiago.paixao@ifes.edu.br]

Filipe Mutz  [Federal University of Espírito Santo | filipe.mutz@ufes.br]

Alberto F. De Souza  [Federal University of Espírito Santo | alberto@lcad.inf.ufes.br]

Claudine S. Badue  [Federal University of Espírito Santo | claudine@lcad.inf.ufes.br]

Thiago Oliveira-Santos  [Federal University of Espírito Santo | todsantos@inf.ufes.br]

 Postgraduate Program in Computer Science, Federal University of Espírito Santo, Av. Fernando Ferrari, 514, Goiabeiras, Vitória, ES, 29075-910, Brazil

Received: 31 March 2025 • Accepted: 17 July 2025 • Published: 06 October 2025

Abstract Multi-hop Question Answering (MHQA) advances Natural Language Processing by pushing models to combine information from multiple sources in a series of reasoning steps. Despite substantial advancements in MHQA for English, resources for evaluating Large Language Models (LLMs) in Portuguese remain scarce. To address this gap, we introduce a publicly available Portuguese translation of the HotpotQA dataset, a well-established English MHQA benchmark. We systematically evaluate several variants of the Llama multilingual LLM across both the original and translated datasets, analyzing performance variations by language. Our findings demonstrate that multilingual models consistently perform better in English than in Portuguese, though this gap narrows with increased model size. Additionally, we show the impact of fine-tuning on improving MHQA performance in Portuguese. This study provides valuable insights into optimizing LLMs for multilingual contexts and contributes a relevant benchmark for Portuguese-language MHQA research.

Keywords: Multi-hop QA, Portuguese NLP, Large Language Models, Dataset Translation, Cross-lingual Evaluation

1 Introduction

Multi-hop Question Answering (MHQA) is the task of answering natural language queries by extracting and combining several pieces of information through multiple reasoning steps. This capability can substantially enhance the usefulness of NLP systems and can be applied in many areas. In healthcare, predicting interactions between drugs is a critical challenge in molecular biology. This problem has been effectively addressed by MHQA systems, which are capable of integrating information from multiple sources. These systems analyze various articles that detail the properties of different drugs and can infer potential interactions, even when no single article explicitly investigates them [Gao *et al.*, 2024]. In education, MHQA systems enhance the learning process by linking information across multiple sources, enabling deeper understanding. These systems navigate educational knowledge graphs to provide accurate, context-aware answers. This approach fosters critical thinking and supports personalized learning, making MHQA valuable for complex subjects [Fang *et al.*, 2023]. In the legal domain, integrating and reasoning over multiple pieces of information is crucial for effective question answering when a decision depends on multiple, possibly scattered, regulations [Martinez-Gil, 2023].

The performance of Large Language Models (LLMs) can vary significantly for different languages, largely due to the amount of language-specific data available during training [Zhao *et al.*, 2024]. In areas such as law, literature, and language learning, question answering systems should rely on

documents in the native language since translation may introduce biases. These practical cases underscore the importance of developing and evaluating MHQA systems across multiple languages. While English-centric models tend to perform well due to the abundance of training data, many real-world applications demand language-specific capabilities to effectively serve diverse populations [Mucciaccia *et al.*, 2025]. Consequently, creating high-quality resources in various languages is crucial for expanding access to effective MHQA systems and promoting fairness across different linguistic contexts.

In this study, we developed a semi-automated pipeline for translating the HotpotQA dataset into Portuguese, making the translation publicly available as a resource for training and evaluating Portuguese language models for MHQA applications. Furthermore, we test several Llama-3 models in both the original and translated dataset to evaluate the effect of language in their performance. We also fine-tune a multilingual variant of Llama-3 exclusively on the translated Portuguese dataset to assess whether language-specific adaptation improves performance. Finally, we analyze questions for which model performance differs between languages and elaborate hypothesis to why this can happen and how to avoid this. To summarize, the main contributions of this study are:

- A methodology for translating the HotpotQA dataset from English to Portuguese.
- A comparison of state-of-the-art LLMs in the original and translated datasets.

- A Portuguese fine-tuned LLM variant and an analysis of its performance on both Portuguese and English MHQA tasks.
- A critical analysis of why different performances are found in these two languages.

The remainder of the work is organized as follows: Section 2 presents a review of related works, discussing previous research and methodologies relevant to our study. In Section 3, we describe the dataset translation process, including preprocessing steps and adaptation strategies. Section 4 details the proposed approach, outlining the techniques and models employed in our study. The results of our experiments and evaluations are reported in Section 5 along with an analysis of the findings, highlighting their implications and potential limitations. In Section 6, we discuss ethical considerations, including licensing, potential biases, environmental impact, and misuse risks that arise from both the dataset and our experiments. Finally, Section 7 summarizes our contributions and suggests directions for future research.

2 Related Works

QA benchmarks are an excellent tool to evaluate LLMs and as a goal of improvement to LLM developers. The SQuAD 1.0, for example, was widely used and consists of over 100,000 question-answer pairs created by crowdworkers based on passages from Wikipedia [Rajpurkar *et al.*, 2016]. After language models started to get excellent performance in this benchmark the SQuAD 2.0 was launched introducing adversarial question generation, where questions that made available LLMs fail were generated and also added unanswerable questions where the models have to detect that the question does not have an answer [Rajpurkar *et al.*, 2016].

With the continuous improvements of LLMs it is increasingly difficult to benchmark them using single-hop questions. Datasets like HotpotQA introduced the MHQA problems, where LLMs have significantly improved performance, but still pose challenges to current models. In its distractor setting hotpotQA consists of a question and 10 text snippets of which some are necessary to answer the question and some are irrelevant. To answer the question it is necessary to find which snippets are relevant and integrate the relevant information. The dataset provides 3 difficulty levels: easy, medium and hard. And also two types of multi-hop questions: bridge and comparison [Yang *et al.*, 2018].

Contrasting the rich benchmark resources for LLMs in English, there are much fewer resources in Portuguese. Among them, the Napolab is a collection of Portuguese datasets designed for evaluating LLMs. It encompasses various tasks, including question answering, and adheres to guidelines ensuring naturalness, reliability, public availability, human annotation, and general applicability [Rodrigues, 2023].

Pirá is a bilingual dataset containing questions and answers about the ocean and the Brazilian coast, available in both Portuguese and English. It supports multiple benchmarks, such as machine reading comprehension, information retrieval, open question answering, and answer triggering. The dataset also includes a multiple-choice QA extension with five candidate answers per question [Paschoal *et al.*, 2021].

SQuAD-BR is the Portuguese translation of the Stanford Question Answering Dataset (SQuAD) version 1.1. This dataset has been utilized to fine-tune BERT-based Portuguese language models, achieving state-of-the-art performance in extractive QA tasks.

MilkQA is a dataset comprising consumer questions related to the dairy domain, written in Portuguese. It contains 2,657 question-answer pairs and is dedicated to the study of consumer questions, particularly focusing on the task of answer selection [Criscuolo *et al.*, 2017].

PORTULAN ExtraGLUE is a benchmark suite developed for Portuguese, comprising 14 datasets for various language processing tasks, including question answering. These datasets were machine-translated from mainstream English benchmarks to support the development of neural language models for Portuguese [Osório *et al.*, 2024].

Mintaka is a complex, natural, and multilingual dataset designed for end-to-end question answering. It comprises 20,000 question-answer pairs collected in English, which have been professionally translated into eight languages, including Portuguese, resulting in a total of 180,000 samples. The dataset includes eight types of complex questions, such as superlative, intersection, and multi-hop questions, all naturally elicited from crowd workers [Sen *et al.*, 2022].

MultiWOZ-PT is a manually translated and culturally adapted version of the MultiWOZ task-oriented dialogue corpus for European Portuguese. It introduces 1,000 dialogues about services in the city of Coimbra and has already been used to benchmark intent recognition and dialog-state-tracking models [Ferreira *et al.*, 2024].

We did not find any translations of the HotpotQA available in the literature nor any benchmarks in Portuguese dedicated exclusively to multi-hop questions. Some datasets like Mintaka and Pirá contain a small number of multi-hop questions, but this is not their main focus. Baseline runs on multilingual LLMs over Mintaka show that the best model achieves 38% hits@1 in English and 31% hits@1 multilingually. This shows that current multilingual models perform significantly better in English than other high resource languages. One study fine-tuned Llama-65B on diverse Portuguese datasets, obtaining Sabiá-65B that showed better performance than the original Llama over several Portuguese benchmarks including the number of questions in the ENEM exam. It was also noted that Sabiá-65B performed worse than Llama-65B in English benchmarks, showing that there is a trade-off between language performances. A second version of the model was launched improving even more the results on Portuguese benchmarks. These effects are also observed in bilingual students, who normally have better performance on exams in their mother tongue compared to second language. Many questions are still open, for example:

- Does the difference in performance between languages increases or decreases with more difficult questions?
- Does model size correlate with better multilingual multi-hop reasoning, or are architectural modifications needed to improve performance across languages?
- Are certain types of reasoning tasks, such as bridging or comparisons, more affected by language differences than others?

- To what extent do different training strategies, such as multilingual training versus language-specific finetuning, affect model performance on MHQA tasks?
- What role does cultural knowledge play in MHQA across languages, and how can models be trained to integrate such knowledge more effectively?

The empirical LLM-size experiment (Section 5, Table 2) is designed to answer our first three research questions, probing how language gaps evolve with difficulty, how model scale affects multilingual MHQA, and whether bridge versus comparison reasoning is differentially impacted. The fine-tuning experiment (Section 5, Table 3) tackles the fourth question by contrasting a multilingual model with a Portuguese fine-tuned model. Finally, a manual qualitative analysis addresses the fifth question, highlighting the role of cultural knowledge and other subtle factors that quantitative metrics may overlook.

3 Pt-HotpotQA Dataset

To investigate how language affects multi-hop reasoning in LLMs, we created a Portuguese version of the widely used HotpotQA benchmark. This section details the process of selecting HotpotQA [Yang *et al.*, 2018] as the source dataset, the methodology adopted for its translation, and the steps taken to ensure the quality and consistency of the resulting resource.

3.1 Dataset selection

Since the objective of this study is to advance the understanding of cross-lingual performance in LLMs, the ideal scenario would involve using the same dataset in both English and Portuguese. Therefore, the first step was to identify a suitable English benchmark dataset for LLM evaluation. To achieve this, a literature review was conducted using Google Scholar and the CAPES Portal (Caf  ) to find scientific papers that proposed benchmark datasets for language models.

For each identified paper, both the references and the articles that cited it were analyzed to uncover additional relevant datasets. Titles were initially screened, and if they appeared promising, the abstracts were read. The inclusion criteria were: (i) published after 2010, (ii) open-source availability, (iii) containing at least 1,000 questions, and (iv) being benchmarks used for LLM evaluation. Once a set of candidate datasets was collected, a search was performed to find existing Portuguese translations using Hugging Face, GitHub, Google Scholar, and the CAPES Portal. The findings are summarized in Table 1.

After a thorough analysis of the candidate datasets, HotpotQA was selected as the most suitable benchmark for this study. Several factors contributed to this decision. First, there is a notable scarcity of Portuguese datasets that focus specifically on MHQA, which is a more complex and informative task for evaluating reasoning capabilities in LLMs. While some existing Portuguese datasets include a small proportion of multi-hop questions, none are exclusively dedicated to this modality. In contrast, HotpotQA is designed explicitly for multi-hop reasoning, requiring models to combine information from multiple sources to answer a question accurately.

Second, HotpotQA is one of the largest available QA datasets, containing over 113,000 questions, which ensures statistical robustness in the evaluation. Its scale also makes it more suitable for fine-tuning or adaptation processes, especially when working with LLMs that benefit from large and diverse training sets.

Third, the dataset is richly annotated, providing not only answers but also supporting facts and Wikipedia context paragraphs, enabling more fine-grained evaluation of a model’s reasoning path and interpretability. This is particularly useful for analyzing whether LLMs truly understand the underlying facts or are simply guessing based on surface-level patterns.

Fourth, the dataset includes difficulty-level annotations (easy, medium, and hard), which allow for a more nuanced evaluation of LLM performance across varying levels of complexity. This enables a more detailed understanding of model strengths and limitations and helps in identifying which reasoning levels pose greater challenges.

Lastly, HotpotQA has been widely used in the NLP community and cited in numerous benchmarking studies, increasing its credibility and making it easier to compare results with prior work. Its open-source availability and well-documented format also facilitate adaptation and translation efforts.

Together, these features make HotpotQA an ideal choice for evaluating cross-lingual performance of LLMs, particularly in underrepresented languages like Portuguese.

3.2 Dataset translation

After selecting the dataset, the next step was determining the most suitable translation method. Our goal was to ensure high translation quality while maintaining consistency and contextual integrity across all fields of the dataset.

The 2023 Conference on Machine Translation provided a systematic comparison between LLMs and specialized translation systems. ChatGPT-4 notably achieved first place in three language pairs and consistently ranked among the top three for all pairs evaluated. Remarkably, it even surpassed human reference translations in five of the eight language pairs assessed through Direct Assessment (DA). DA involves ratings from native or fluent speakers on a 0–100 scale, evaluating the accuracy and fluency of translations against the original sentences.

Although the English-Portuguese language pair was not explicitly evaluated in the referenced conference, existing literature categorizes Portuguese as a high-resource language, indicating ample data availability. Since ChatGPT’s performance strongly correlates with the volume of training data, comparable performance for the English-Portuguese language pair can reasonably be expected. Moreover, ChatGPT has consistently demonstrated excellent performance in English-Portuguese translation benchmarks.

For our specific translation needs, we selected the ChatGPT-4o model, a newer iteration expected to surpass the performance of the earlier ChatGPT-4 model due to continuous advancements in LLM capabilities. Additionally, we chose GPT-based translation over expert human translators primarily due to cost-effectiveness. While expert human translators can deliver high-quality results, the significant expense associated with expert services was a limiting factor. GPT

Table 1. Datasets used to benchmark LLMs in English

Dataset	Context	Answer type	Size	Translated	Citation
SQuAD 1.1	Wikipedia	Extractive span	100k	✓	Rajpurkar <i>et al.</i> [2016]
SQuAD 2.0	Wikipedia	Extractive span	150k	✓	Rajpurkar <i>et al.</i> [2018]
Natural Questions	Wikipedia	Extractive span	323k		Kwiatkowski <i>et al.</i> [2019]
NewsQA	CNN articles	Extractive span	120k		Trischler <i>et al.</i> [2017]
TrivialQA	Web + Wikipedia	Free-form text	95k		Joshi <i>et al.</i> [2017]
BioASQ	PubMed (biomedical)	Extractive span	5k		Tsatsaronis <i>et al.</i> [2015]
WebQuestions	Freebase	Free-form text	5.8k		Berant <i>et al.</i> [2013]
CuratedTREC	Freebase	Free-form text	2.2k		Baudiš and Šedivý [2015]
Open-NQ	Wikipedia	Extractive span	79k		Lee <i>et al.</i> [2019]
QuAC	Wikipedia	Extractive span	98k		Choi <i>et al.</i> [2018]
CoQA	Multi-domain	Extractive span	127k		Reddy <i>et al.</i> [2019]
RACE	School exams	Multiple-choice	97k		Lai <i>et al.</i> [2017]
MCTest	Short stories	Multiple-choice	660		Richardson <i>et al.</i> [2013]
CommonsenseQA	Commonsense	Multiple-choice	12k	✓	Talmor <i>et al.</i> [2019]
OpenBookQA	Sci. + commonsense	Multiple-choice	6k		Mihaylov <i>et al.</i> [2018]
ARC	Science (grades)	Multiple-choice	7.8k		Clark <i>et al.</i> [2018]
PIQA	Physical commonsense	Multiple-choice	16k	✓	Bisk <i>et al.</i> [2020]
BoolQ	Wikipedia	Boolean	16k	✓	Clark <i>et al.</i> [2019]
FEVER	Wikipedia claims	Boolean	185k	✓	Thorne <i>et al.</i> [2018]
GSM8K	Math (grade school)	Numeric	8.5k	✓	Cobbe <i>et al.</i> [2021]
MATH	Adv. math (HS/comp)	Numeric	12.5k		Hendrycks <i>et al.</i> [2021]
SVAMP	Arithmetic reasoning	Numeric	1k		Patel <i>et al.</i> [2021]
HotpotQA	Wikipedia (multi-hop)	Extractive span	113k		Yang <i>et al.</i> [2018]
MuSiQue	Wikipedia (multi-hop)	Extractive span	25k		Trivedi <i>et al.</i> [2022]
QASC	Science facts	Multiple-choice	10k		Khot <i>et al.</i> [2020]
WikiHop	Wikipedia	Multiple-choice	51k		Welbl <i>et al.</i> [2018]
NarrativeQA	Books + scripts	Free-form text	46k		Kočický <i>et al.</i> [2018]
ELI5	Reddit (long-form)	Free-form text	270k		Fan <i>et al.</i> [2019]
DROP	Wikipedia + reasoning	Numeric	96k		Dua <i>et al.</i> [2019]
HellaSwag	Sentence completion	Multiple-choice	70k		Zellers <i>et al.</i> [2019]

models offer a more economically viable alternative without substantially compromising translation quality. Nonetheless, we recognize the inherent limitations of LLM-based translations, such as potential inaccuracies or biases. We deemed these manageable within the scope and objectives of our research.

Each entry in the HotpotQA dataset consists of exactly seven top-level fields: `id`, `question`, `answer`, `type`, `level`, `supporting_facts`, and `context`. Among these, only `question`, `answer`, `supporting_facts`, and `context` required translation. The remaining fields (`id`, `type`, and `level`) were retained as is, both to reduce translation overhead and to maintain consistency with the original structure. Additionally, we preserved the original English field names in the translated dataset to facilitate code reuse.

Rather than translating each field individually, we opted to translate the relevant fields together as a single structured JSON object. This approach provided two key benefits:

- **Contextual consistency:** Terms that appeared across multiple fields (e.g., proper names, entities, or repeated phrases) were more likely to be translated consistently.
- **Improved translation quality:** By providing the entire context, the model better understood the structure and purpose of the dataset, leading to more accurate and coherent translations.

A dedicated prompt was crafted to guide ChatGPT-4o, combining translation instructions with the JSON data to be translated. After translation, each output underwent a multi-step validation process:

1. **JSON format validation:** We used a Python JSON parser to ensure the output was syntactically valid.
2. **Structural validation:** We verified that all fields were present and that list and sublist lengths matched the original.
3. **Translation completeness check:** For each string field, we compared the number of unique words in English and Portuguese. If the counts were identical, suggesting the string might not have been translated, it was flagged for manual review.

Entries that passed all validation checks had the untranslated fields (`id`, `type`, and `level`) reinserted and were added to the Portuguese dataset. If validation failed, a new translation attempt was initiated using a refined prompt that included a one-shot example. While this increased the translation cost, it significantly improved success rates for edge cases. In rare instances where the second automatic translation also failed, the corresponding entries were temporarily marked as untranslatable. These cases were later revisited manually after the automated process was complete, with custom prompt adjustments applied individually to obtain valid translations.

This semi-automated, quality-controlled pipeline enabled the efficient translation of the entire HotpotQA dataset in its distractor configuration, preserving the integrity of the original content while ensuring linguistic consistency and scalability in the Portuguese adaptation.

It is important to note that our translation effort focused exclusively on the distractor configuration of the HotpotQA dataset. In the fullwiki configuration, models are expected to retrieve relevant information from the entire English Wikipedia, rather than relying on a predefined set of context paragraphs. As a result, translating this configuration would require not only translating the dataset’s predefined fields but also the entire Wikipedia corpus used for retrieval, an impractical undertaking given its scale. Considering our goal of producing a Portuguese version of HotpotQA suitable for controlled evaluation and training, the distractor configuration offered a more practical and consistent foundation for translation.

Nevertheless, many fields, such as `question`, `answer`, and `supporting_facts`, are shared between the distractor and fullwiki configurations. As a result, the translated versions of these fields can be directly reused in a potential future translation of the fullwiki setting, reducing effort and ensuring consistency across configurations.

3.3 Automated translation quality test

To ensure that differences in LLM performance between English and Portuguese were due to genuine language-based reasoning challenges, and not artifacts introduced by translation, we conducted quality assessment of the translated dataset using a back-translation approach.

We created a diagnostic subset named the Back-translation Sample Set (BTSS), consisting of 400 randomly selected questions from the translated Portuguese HotpotQA dataset. These questions were retranslated back into English using the same methodology employed during the original translation, leveraging the ChatGPT-4o model with structured prompts to ensure consistency. Each question in the BTSS is paired with a corresponding original English question, forming what we refer to as the Original English Sample Set. Together, these two aligned subsets enable a controlled comparison between original and retranslated questions.

The purpose of this sets is to assess whether the translation process itself introduces distortions that could affect model performance. By comparing responses to the original and retranslated questions, we aim to isolate the effects of translation and validate the reliability of the Portuguese dataset for cross-lingual evaluation.

To complement this automated check, a manual inspection was conducted on the sample sets to assess the quality of the translations. No instances were found where the translation quality prevented a question from being answered correctly. In a small number of cases (18 out of 400), alternative word choices could have improved the wording of the question, but these were subjective preferences rather than objective issues.

4 Methodology

This section outlines the experimental setup used to evaluate how language influences the performance of LLMs on MHQA tasks. We describe the prompting strategy used in the evaluations, the evaluation metrics adopted, the models selected for testing, the setup of the experiments and the fine-tuning procedure. Furthermore, we include a qualitative analysis to better understand the sources of performance variation across languages.

4.1 Prompting Strategy

The prompt used in the experiments adopts the one-shot learning strategy. Each prompt begins with a complete illustrative example consisting of an instruction, context, question, and the corresponding answer. This example serves as a guide to demonstrate the expected structure and output format, helping the model understand how to respond to the target question. After this example, the model is presented with the actual input, comprising the instruction, context, and question, while being instructed to return only the direct answer without any additional explanation or justification.

The overall structure of the prompt is as follows:

```
{one-shot example}
<instruction>
Given the context below, answer the question
concisely, providing only the direct
answer without additional explanations or
justifications.
</instruction>
<context>
Christ the Redeemer is located in the city
of Rio de Janeiro, which is one of the most
famous cities in Brazil. The statue was
inaugurated in 1931.
</context>
<question>
In which city is Christ the Redeemer located?
</question>
<answer>
Rio de Janeiro
</answer>
```

In the example above, the text highlighted in blue represents the input to the language model, while the text in red corresponds to the output expected from the LLM. Although the example is intentionally simple to maximize clarity, the actual evaluation instances are more complex: the context consists of 10 Wikipedia paragraphs, and the questions require multi-hop reasoning to answer correctly.

For the Portuguese evaluation, the entire prompt (including the one-shot example, instruction, context, and question) was translated into Portuguese. This ensured that the model received inputs entirely in the target language. The same structure and format were maintained across both English and Portuguese evaluations to allow for a fair and standardized comparison between models and languages.

4.2 Evaluation Metrics

To evaluate the model performance, we adopted the **Exact Match (EM)** score as our primary metric. EM is a widely used measure in question-answering benchmarks that calculates the percentage of predictions that exactly match the ground truth answers. To ensure robustness against minor formatting variations, a normalization step is applied prior to comparison. This step includes:

- Converting all characters to lowercase,
- Removing all punctuation marks, and
- Replacing sequences of whitespace characters with a single space.

If the model’s normalized prediction matches the normalized reference answer exactly, the prediction is considered correct; otherwise, it is considered incorrect, even if the difference is as small as a single character.

A secondary metric is also introduced: **Back-Translation Agreement Rate (BTAR)**. The goal of this metric is to serve as a diagnostic tool for assessing if variations in LLM performance across languages result from genuine language understanding challenges rather than being caused by translation artifacts or changes in prompt wording.

To compute the metric, each question in the Original Sample Set is paired with its corresponding question in the BTSS using the unique id field. Then, for a given model, the responses for each pair are compared to categorize the outcomes as follows:

- **Correct-Correct (CC)**: Correct in both sets,
- **Incorrect-Incorrect (II)**: Incorrect in both sets,
- **Correct-Incorrect (CI)**: Correct in the Original Sample Set but not in the BTSS,
- **Incorrect-Correct (IC)**: Incorrect in the Original Sample Set but correct in the BTSS.

From these counts, BTAR was calculated using the standard accuracy formula:

$$\text{BTAR} = \frac{CC + II}{CC + II + CI + IC} \quad (1)$$

It is important to note, that the BTAR is calculated by comparing model performance on the Original English Sample Set to its counterpart that was first translated into Portuguese and then back into English. As such, it reflects the impact of two sequential translation steps. Therefore, the agreement rate of an LLM between the English and Portuguese datasets is expected to be correlated with, but not equivalent to, the BTAR.

Together, EM and BTAR provide a comprehensive assessment framework, enabling both absolute performance evaluation and cross-lingual comparison.

4.3 LLM size experiment

The objective of this experiment is to investigate whether the performance gap between multilingual LLMs in English and Portuguese multi-hop questions increases, decreases, or remains stable as the number of model parameters varies. This

analysis will provide insights into the relationship between model size and multilingual reasoning capabilities, particularly in multi-hop tasks, and help determine if larger models inherently perform better across languages or if additional fine-tuning and language-specific training are necessary.

We selected the Llama model family due to its popularity, proven effectiveness, availability in various sizes, and widespread use as a benchmark in multilingual NLP tasks. Specifically, we evaluated the following variants:

- Llama-3.2-1B-instruct
- Llama-3.2-3B-instruct
- Llama-3.2-11B-vision-instruct
- Llama-3.3-70B-instruct

These versions were chosen for being the most recent available representatives of their respective size categories. The vision-instruct model was preferred over a 3.1 instruct variant for being more recent, despite the possibility that its vision-specific training could influence performance.

Each model was evaluated on the complete HotpotQA validation set, which comprises 7,405 questions in English, as well as on the corresponding Pt-HotpotQA validation set containing an equivalent 7,405 questions in Portuguese. For each question in both datasets, the respective llama model was prompted using the one-shot format described in Section 4.1, and the model’s answer was recorded for evaluation. The answers were then compared against the ground-truth references using the EM metric detailed in Section 4.2, yielding the EM score for each model on both the English and Portuguese validation sets.

Additionally, we analyzed performance across different difficulty levels and question types (bridge and comparison questions), providing deeper insights into how model size affects performance variability within and across languages.

To further interpret the results, we conducted a detailed qualitative analysis on specific questions where significant performance discrepancies occurred between languages. This allowed us to explore the potential underlying factors contributing to the performance gap, such as linguistic nuances, syntactic complexity, or cultural context differences.

Ultimately, this experimental setup aims to identify the trade-offs involved in using different sizes of LLM for multilingual multi-hop question-answering tasks.

4.4 Portuguese fine-tuning experiment

This experiment aims to investigate whether fine-tuning a multilingual LLM specifically on Portuguese-language data can enhance its performance on MHQA tasks in both Portuguese and English. The Llama family is also selected for this experiment:

- The original Llama 3.2 1B model, pre-trained on a multilingual corpus predominantly consisting of English data.
- A variant of the Llama 3.2 1B model fine-tuned using the training set of Pt-HotpotQA.

The fine-tuned variant was obtained by training the Llama 3.2 1B model on the Pt-HotpotQA distractor training set, which contains 90,447 questions. Fine-tuning was performed

Table 2. Exact Match performance of Llama models of different sizes on the English and Portuguese HotpotQA datasets. The * symbol denotes a vision-instruct model trained for both image and text inputs, which may influence performance.

Model	English	Portuguese	BTAR
Llama 3.2 1B	42%	34%	91%
Llama 3.2 3B	51%	44%	93%
Llama 3.2 11B*	50%	49%	93%
Llama 3.3 70B	62%	60%	96%

for a single epoch using a batch size of 2. Optimization employed the AdamW algorithm with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. The learning rate followed a linear decay schedule from its initial value of 2×10^{-5} down to zero over the course of training. Training was performed entirely in full FP32 precision, without the use of mixed-precision techniques. All experiments were conducted on a single NVIDIA A100-80GB GPU. While this setup suffices to demonstrate the feasibility of language-specific fine-tuning, more extensive training using greater computational resources could likely yield better performance gains.

Both models were evaluated in the validation sets of the translated Portuguese version of the HotpotQA dataset and the original English dataset to determine cross-lingual impacts. Performance metrics were computed separately for each language, with special attention given to differences in accuracy between languages, as well as performance gaps between the original and fine-tuned models.

This comparative evaluation allowed us to assess whether fine-tuning exclusively in Portuguese could yield generalizable performance benefits or whether improvements remained confined to the Portuguese language. Furthermore, qualitative analyses were conducted to identify which types of multi-hop questions, such as bridging and comparison, were affected the most by fine-tuning.

Ultimately, this experimental approach provided insights into the potential trade-offs between language-specific fine-tuning and broader multilingual applicability in LLM-based multi-hop QA tasks.

5 Results

This section presents the results obtained from our experiments conducted with different LLM variants on the original English and the translated Portuguese HotpotQA datasets.

5.1 LLM Size Experiment

Table 2 presents the results of our investigation into how model size impacts performance across two languages. We evaluated four variants of Llama (1B, 3B, 11B, and 70B parameters) on both the English and Portuguese versions of HotpotQA.

A clear upward trend in performance is observed as model size increases. The smallest model (1B parameters) achieved 42% accuracy on English and 34% on Portuguese, while the largest model (70B parameters) reached 62% and 60%, respectively. This trend held for both languages, though the absolute performance gap between English and Portuguese

Table 3. Comparison of original and Portuguese fine-tuned Llama model performance on HotpotQA datasets.

Model	English	Portuguese	BTAR
Original Llama 1B	42%	34%	91%
Fine-tuned Llama 1B	36%	40%	92%

narrowed as model size increased, suggesting that larger models are more robust in multilingual settings.

An interesting observation arises from the performance of the LLaMA 3.2 11B vision-instruct model. Although it has more parameters than the 3B version, it performed slightly worse on English questions (50% vs. 51%) but notably better on Portuguese (49% vs. 44%). While the model supports multilingual text input, English is the only language it fully supports in vision-language tasks. This English-centric focus for image-text pairings might have inadvertently biased the model toward English representations in multimodal contexts, possibly reducing its performance on the purely textual English MHQA task. In contrast, since Portuguese was not part of the vision training, its textual reasoning abilities in this language likely remained unaffected.

Moreover, the BTAR improved consistently as model size increased, from 91% in the 1B model to 96% in the 70B model. This indicates that larger models are better at producing consistent responses to semantically equivalent questions, even when phrased in different languages or retranslated. Another noteworthy insight is the diminishing returns observed in scaling. While the jump from 1B to 3B led to a 9% gain in English MHQA, the leap from 3B to 70B yielded an 11% gain, suggesting that, although beneficial, scaling exhibits a tapering effect in performance improvement.

These findings highlight the dual benefit of scaling: enhanced overall performance and improved cross-lingual consistency. However, the results also emphasize that even the largest models are far from perfect on challenging MHQA tasks, with performance topping out at 62%.

5.2 Portuguese Fine-tuned LLM Experiment

Table 3 presents the results of the Portuguese fine-tuning experiment, comparing the original multilingual LLaMA 3.2 1B model with a fine-tuned variant trained exclusively on the Portuguese version of the HotpotQA dataset. Both models were evaluated on the English and Portuguese validation sets to assess the effects of language-specific adaptation. The results clearly illustrate the impact of language-specific fine-tuning. Initially, the original multilingual model, predominantly trained on English data, performed better on English questions (42%) than on Portuguese questions (34%). After fine-tuning on Portuguese data, the model demonstrated a substantial improvement in Portuguese performance, rising to 40%, indicating the effectiveness of targeted language adaptation.

However, this enhancement in Portuguese proficiency coincided with a notable decrease in English performance, which fell to 36%. This decline suggests a phenomenon known as catastrophic forgetting, where fine-tuning the model exclusively on Portuguese data adversely affected its previously acquired English linguistic capabilities. Given the limited capacity of smaller models like Llama 3.2 1B, this trade-off

is expected, as the model likely reallocates resources towards Portuguese representations at the expense of maintaining English proficiency.

These findings support observations from earlier studies, such as experiments with the Sabiá-65B model, indicating a consistent trade-off between specialized language training and multilingual generalization. Thus, while language-specific fine-tuning offers significant performance benefits for targeted languages, it may reduce a model’s effectiveness in other languages. Consequently, these results emphasize the importance of developing language-specific models to effectively address linguistic nuances and maximize performance for targeted applications.

5.3 Qualitative Analysis of Performance Differences

When qualitatively analyzing the outputs of the LLMs based on the input prompts, several interesting observations were made. While the Llama 1B model occasionally produced obviously incorrect answers to questions, it also displayed impressive insight at times, demonstrating unexpected competence in tackling complex multi-hop reasoning tasks. When asked to explain its reasoning behind certain answers, the model often provided coherent and accurate justifications, indicating a partial grasp of complex concepts. However, it was also noted that minor changes in the wording of the prompt, without altering its overall meaning, could sometimes cause the model to shift from providing correct answers to instead outputting repeated XML tags.

One illustrative example is the question: *“Which song did Eminem and Rihanna collaborate on after their other collaboration song in the studio album ‘Unapologetic?’”* This question was posed to the LLM using the prompt strategy described in Section 4.1. In the provided context, which consisted of 55 sentences, only two were actually needed to answer correctly:

- “Numb” is a song by Barbadian singer Rihanna from her seventh studio album “Unapologetic” (2012).
- “The Monster” marks the fourth collaboration between Eminem and Rihanna, following “Love the Way You Lie”, its sequel “Love the Way You Lie (Part II)” (2010), and “Numb” (2012).

The correct answer is “The Monster.” However, the Llama 1B model incorrectly responded with “Love the Way You Lie” in the English version of the question, but answered correctly when the question was posed in Portuguese. Rephrasing the question to *“Point out the track Eminem and Rihanna collaborated on after their work on a song in the ‘Unapologetic’ album”*, without changing anything else in the prompt, resulted in Llama 1B returning “Love the Way You Lie (2010)” and “The Monster (2013),” which, while not entirely correct, did include the right answer. Interestingly, when the context snippet mentioning “Love the Way You Lie” was removed, Llama 1B then answered correctly, indicating that the mistake stemmed from contextual distraction rather than misuse of prior knowledge.

In contrast, the larger Llama 70B consistently answered the question correctly and was robust to rewordings that preserved the question’s meaning. This highlights a key difference in prompt sensitivity between smaller and larger models. Larger models made fewer errors, offered deeper reasoning, and were significantly less sensitive to changes in prompt wording, consistently giving the same answer when the wording was altered but the meaning remained the same. It was also observed that for both small and large models, when the model gave an incorrect answer, informing it of the mistake and asking for another response often led to the correct answer, with this occurring more frequently in larger models, though small models also showed similar behaviour.

The fine-tuning of the model on Portuguese data appeared to negatively impact its performance in English. After fine-tuning, the model was noticeably more likely to produce odd or incoherent responses to simple English prompts, suggesting a degradation in its reasoning abilities in that language. Interestingly, this degradation was not observed in Portuguese, where the model continued to perform normally. The issue seems isolated to English, implying that the fine-tuning process may have unintentionally disrupted some of the model’s previously strong capabilities in that language. This highlights a potential limitation of multilingual training in smaller models and suggests that models dedicated to a single language might yield better performance.

6 Ethical Considerations

6.1 Data Licensing and Intellectual Property

The original HotpotQA dataset is released under a permissive license that allows redistribution and derivative works. Our translated version (*Pt-HotpotQA*) preserves all original licensing terms and is distributed under the same Creative Commons Attribution 4.0 International License. We explicitly release only the translated JSON files and no proprietary language-model outputs or additional copyrighted material, so that downstream users can verify provenance and remain compliant with open-source requirements.

6.2 Translation Fidelity and Biases

LLMs can perpetuate or amplify biases present in their training data. Although we adopted quality-control steps (Section 3), subtle semantic shifts or culturally specific connotations may still have slipped through automated translation. We therefore (i) publish the full source code for the translation pipeline to enable community audits, (ii) provide sentence-level alignment so that annotators can flag problematic cases, and (iii) welcome pull requests that supply corrections or culturally sensitive rewrites. Future work should incorporate human-in-the-loop evaluation, ideally by domain experts, to minimise bias and ensure inclusive language.

6.3 Fairness and Performance Disparities

Our experiments revealed systematic performance gaps between English and Portuguese (Section 5). Deploying models that underperform for certain language communities risks

unequal access to reliable information. We caution practitioners against using our baseline scores as justification for production-level deployment; additional fairness audits (e.g., stratified by topic, nationality, or demographic references) are advised before high-stakes use.

6.4 Privacy and Sensitive Content

The original HotpotQA context paragraphs are public Wikipedia articles, which mitigates privacy concerns. Nonetheless, all intermediate artefacts in our pipeline were processed and stored locally, and no personal identifiers were publicized.

6.5 Environmental Impact

The computational aspects of this study involved the use of LLMs through cloud APIs and local GPU processing. Specifically:

- **OpenAI GPT-4o** A total of 106,282 prompts were processed using OpenAI’s GPT-4o API, with an average of 1,509 input tokens and 1,652 output tokens, consumed an estimated 141 KWh.
- **Amazon Bedrock** Additional 61,840 prompts were processed via Amazon Bedrock API, with an average of 1,951 input tokens and 32 output tokens, consuming an estimated 18 kWh.
- **Local GPUs** Local computations included approximately 210 hours on an NVIDIA RTX 4060 Ti GPU and 30 hours on an NVIDIA A100 GPU equating approximately 46kWh

The total estimated energy consumption was approximately 205kWh, corresponding to a carbon footprint of about 82kg, using a global average emission factor of 0.4 kg CO₂ per kWh.

6.6 Misuse Potential

Pt-HotpotQA can be employed to improve downstream reasoning abilities in Portuguese LLMs. Positive applications include educational tools and accessible information retrieval; negative uses could involve more sophisticated generation of misinformation or persuasive content in Portuguese. We therefore (i) release the dataset under a license that requires attribution, (ii) include a clear statement discouraging malicious use, and (iii) notify the community of known misuse vectors in our project repository’s SECURITY.md.

7 Conclusion

We presented Pt-HotpotQA, the first large-scale Portuguese adaptation of the HotpotQA benchmark, and used it to probe how language affects the multi-hop reasoning abilities of modern LLMs. Our study yields five key takeaways, each directly answering one of the research questions posed at the end of Section 2:

- **Difficulty may reduce the language gap.** The performance gap between English and Portuguese tended to be smaller for harder questions, although the difficulty levels are somewhat subjective.
- **Bigger LLMs have lower language gap.** Scaling Llama from 1 B to 70 B parameters boosts exact-match (EM) from 42 % to 62 % in English and from 34 % to 60 % in Portuguese, reducing the cross-lingual gap from 8 to 2 percentage points.
- **Bridge questions showed a larger gap than comparison questions.** The performance drop from English to Portuguese was greater for bridge type questions, suggesting that this reasoning type is more sensitive to cross-lingual limitations.
- **Language-specific fine-tuning is a double-edged sword.** One-epoch training on Pt-HotpotQA yields a 4 pp gain in Portuguese EM but a 8 pp loss in English, confirming catastrophic forgetting in low-capacity models.
- **Cultural and linguistic nuances play a minor role.** Manual error analysis revealed that while some translation artifacts or culturally specific expressions affected a small subset of questions, they were not the primary driver of cross-lingual performance differences.

This study presents practical implications for the NLP community. The public release of the translated Portuguese HotpotQA dataset provides a valuable benchmark to foster further research and development of Portuguese-specific MHQA systems. It also serves as a platform to test multilingual capabilities of emerging LLMs, driving improvements in model architecture and multilingual training strategies.

Nevertheless, our study has several limitations. First, the reliance on GPT-based translation, despite high quality, might introduce subtle inaccuracies or biases influencing model evaluation. Second, our analysis was restricted to Llama models; thus, findings might vary with alternative architectures or pre-training approaches. Lastly, the scope of qualitative analysis was limited; a broader examination could yield additional insights into linguistic and cultural impacts.

Our findings demonstrate significant benefits from language-specific fine-tuning while also highlighting inherent trade-offs and complexities involved in multilingual QA tasks. Further research exploring sophisticated multilingual training paradigms and culturally-sensitive datasets is warranted to enhance model performance across languages and ultimately create more equitable NLP technologies.

Declarations

Acknowledgements

The authors gratefully acknowledge the volunteers who contributed to reviewing the translations.

Authors' Contributions

Mucciaccia is the main contributor and writer of this manuscript. Oliveira-Santos played a key role in funding acquisition, project administration, and supervision. All authors contributed equally to conceptualization, data curation, formal analysis, investigation, and methodology.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil) - Finance Code 001; by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil); and Fundação de Amparo à Pesquisa do Espírito Santo (FAPES, Brazil) – grants 2021-07KJ2, 2022-NGKM5 and 2021-GL60J.

Availability of data and materials

The code, documentation, and the Portuguese translation of the HotpotQA dataset are available on GitHub at the following repository link: <https://github.com/i2ca/pt-hotpot-qa>. This dataset is a derivative of the original HotpotQA dataset [Yang et al., 2018], which is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0). Our translated version is released under the same license available at this link: <https://creativecommons.org/licenses/by-sa/4.0/>

References

- Baudiš, P. and Šedivý, J. (2015). Modeling of the question answering task in the yodaqa system. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 222–228. Springer International Publishing. DOI: 10.1007/978-3-319-24027-5_20.
- Berant, J., Chou, A., Frostig, R., and Liang, P. (2013). Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544. DOI: 10.18653/v1/d13-1160.
- Bisk, Y., Zellers, R., Bras, R. L., Gao, J., and Choi, Y. (2020). Piqa: Reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7432–7439. DOI: 10.1609/aaai.v34i05.6239.
- Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184. DOI: 10.18653/v1/d18-1241.
- Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. (2019). BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics. DOI: 10.48550/arXiv.1905.10044.
- Clark, P., Cowhey, S., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafford, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. In *arXiv preprint arXiv:1803.05457*. DOI: <https://doi.org/10.48550/arXiv.1803.05457>.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, J., Jun, H., Kaiser, L., Miller, J., Plappert, M., Tworek, J., Hilton, J., Schulman, J., Salakhutdinov, R., and Amodei, D. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*. DOI: 10.48550/arxiv.2110.14168.
- Criscuolo, M., Fonseca, E. R., Aluisio, S. M., and Speranca-Criscuolo, A. C. (2017). MilkQA: A Dataset of Consumer Questions for the Task of Answer Selection. In *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 354–359, Los Alamitos, CA, USA. IEEE Computer Society. DOI: 10.1109/BRACIS.2017.12.
- Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. (2019). Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2368–2378. DOI: 10.48550/arxiv.1903.00161.
- Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M. (2019). ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics. DOI: 10.48550/arXiv.1907.09190.
- Fang, Y., Deng, J., Zhang, F., and Wang, H. (2023). An intelligent question-answering model over educational knowledge graph for sustainable urban living. *Sustainability*, 15(2). DOI: 10.3390/su15021139.
- Ferreira, P., Pais, F., Silva, C., Alves, A., and Oliveira, H. G. (2024). Multiwoz-pt um conjunto de diálogos orientados a tarefas em português. *Linguamática*, 16(2):75–90. DOI: 10.21814/lm.16.2.431.
- Gao, P., Gao, F., Ni, J., Wang, Y., Wang, F., and Zhang, Q. (2024). Medical knowledge graph question answering for drug-drug interaction prediction based on multi-hop machine reading comprehension. *CAAI Transactions on Intelligence Technology*, 9(3):1217–1228. DOI: 10.1049/cit2.12332.
- Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset. *NeurIPS*. DOI: 10.48550/arxiv.2103.03874.
- Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L.

- (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611. DOI: 10.48550/arXiv.1705.03551.
- Khot, T., Sabharwal, A., and Clark, P. (2020). Qasc: A dataset for question answering via sentence composition. *AAAI*. DOI: 10.1609/aaai.v34i05.6319.
- Kočíský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. (2018). The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*. DOI: 10.1162/tacl_a00023.
- Kwiatkowski, T., Palomaki, J., Redfield, O., et al. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466. DOI: 10.1162/tacl_a00276.
- Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. (2017). RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics. DOI: 10.18653/v1/D17-1082.
- Lee, K., Chang, M.-W., and Toutanova, K. (2019). Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096. DOI: 10.18653/v1/p19-1612.
- Martinez-Gil, J. (2023). A survey on legal question–answering systems. *Computer Science Review*, 48:100552. DOI: 10.1016/j.cosrev.2023.100552.
- Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. (2018). Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391. Association for Computational Linguistics. DOI: 10.48550/arXiv.1809.02789.
- Mucciaccia, S. S., Meireles Paixão, T., Wall Mutz, F., Santos Badue, C., Ferreira de Souza, A., and Oliveira-Santos, T. (2025). Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2246–2260, Abu Dhabi, UAE. Association for Computational Linguistics. Available at: <https://aclanthology.org/2025.coling-main.154/>.
- Osório, T. F., Leite, B., Lopes Cardoso, H., Gomes, L., Rodrigues, J., Santos, R., and Branco, A. (2024). PORTULAN ExtraGLUE datasets and models: Kick-starting a benchmark for the neural processing of Portuguese. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, pages 24–34, Torino, Italia. ELRA and ICCL. DOI: 10.7202/1042710ar.
- Paschoal, A. F. A., Pirozelli, P., Freire, V., Delgado, K. V., Peres, S. M., José, M. M., Nakasato, F., Oliveira, A. S., Brandão, A. A. F., Costa, A. H. R., and Cozman, F. G. (2021). Pirá: A bilingual portuguese-english dataset for question-answering about the ocean. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, page 4544–4553, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3459637.3482012.
- Patel, A., Bhattamishra, S., and Goyal, N. (2021). Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2021.naacl-main.168.
- Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don’t know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 784–789. Association for Computational Linguistics. DOI: 10.48550/arXiv.1806.03822.
- Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics. DOI: 10.18653/v1/D16-1264.
- Reddy, S., Chen, D., and Manning, C. D. (2019). Coqa: A conversational question answering challenge. In *Transactions of the Association for Computational Linguistics*, volume 7, pages 249–266. DOI: 10.1162/tacl_a00266.
- Richardson, M., Burges, C., and Renshaw, E. (2013). Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203. DOI: 10.18653/v1/d13-1020.
- Rodrigues, R. C. (2023). Lessons learned from the evaluation of portuguese language models. Master’s thesis, University of Malta, Msida, Malta. Available at: <https://www.um.edu.mt/library/oar/handle/123456789/120557>.
- Sen, P., Aji, A. F., and Saffari, A. (2022). Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. DOI: 10.48550/arXiv.2210.01613.
- Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2019). Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4149–4158. DOI: 10.48550/arxiv.1811.00937.
- Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819. DOI: 10.48550/arXiv.1803.05355.
- Trischler, A., Wang, T., Yuan, X., et al. (2017). Newsqa: A machine comprehension dataset. In *Proceedings of the*

- 2nd Workshop on Representation Learning for NLP, pages 191–200. DOI: 10.48550/arXiv.1611.09830.
- Trivedi, H., Balasubramanian, N., Talmor, A., Gardner, M., and Tafford, O. (2022). Musique: Multihop questions via single-hop question composition. In *Proceedings of the 60th Annual Meeting of the*, doi = 10.1162/tacl_a00475.
- Tsatsaronis, G., Balikas, G., Malakasiotis, P., et al. (2015). An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138. DOI: 10.1186/s12859-015-0564-6.
- Welbl, J., Stenetorp, P., and Riedel, S. (2018). Constructing datasets for multi-hop reading comprehension across documents. In *Transactions of the Association for Computational Linguistics*. DOI: 10.1162/tacl_a00021.
- Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics. DOI: 10.18653/v1/D18-1259.
- Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800. DOI: 10.48550/arXiv.1905.07830.
- Zhao, Y., Zhang, W., Chen, G., Kawaguchi, K., and Bing, L. (2024). How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. DOI: 10.48550/arxiv.2402.18815.

A Prompts

This appendix details the exact prompts used in our workflow, together with a brief report on their effectiveness.

A.1 Translation Prompt

Each example from the original HotpotQA dataset was converted into a structured JSON object with only the fields that required translation (see Section 3). We then sent the object to GPT-4o with the following system message:

```
Translate only the text values (e.g., inside
quotes) from the following JSON object from
English to Portuguese.
Do not change the structure, keys,
punctuation, or formatting of the JSON.
Do not replace double quotes " with single
quotes '.
The result must be a valid JSON.
Return only the translated JSON object and
nothing else.
{json_object}
```

Out of all items in the distractor configuration, the first pass translation succeeded for 97.1%. A second pass with a refined, one-shot prompt corrected an additional 2.8%. The

remaining 0.1% (118 items) were translated by manually adjusting the prompt and verifying the output to ensure accurate translations.

After the translation the Llama LLMs were used to answer the questions with the prompt below (that uses a one-shot example and in this case there were no retries, if something was wrong with the output the LLM answer is considered wrong):

A.2 Question–Answering Prompt

During model evaluation we employed a one-shot format. The complete prompt template is reproduced below, with placeholders shown in braces.

```
<prompt>
Given the context below, answer the question
concisely, providing only the direct
answer without additional explanations or
justifications.
</prompt>
<context>
{example_context}
</context>
<question>
{example_question}
</question>
<answer>
{example_answer}
</answer>
<prompt>
Given the context below, answer the question
concisely, providing only the direct
answer without additional explanations or
justifications.
</prompt>
<context>
{context}
</context>
<question>
{question}
</question>
<answer>
```

The first part serves as an illustrative example to prime the model (one-shot learning). The part block contains the actual instance to be answered. No retries were issued for this prompt, answers failing to match the ground-truth after normalization (Section 4.2) were marked incorrect.

B Stratified results

The HotpotQA validation set contains only hard questions, so we built a stratified subset from the training set to enable broader analysis. Questions were grouped by difficulty level (easy, medium, hard) and reasoning type (bridge, comparison), forming six combinations. We randomly sampled 1,000 questions per combination, yielding a balanced set of 6,000.

Table 4 presents the exact match (EM) scores of the LLaMA 3.2 1B model on this set, in both English and Portuguese.

Table 4. EM scores of LLaMA 3.2 1B, stratified by reasoning type and difficulty level.

Type	Level	English	Portuguese
Bridge	Easy	47%	39%
Bridge	Medium	45%	36%
Bridge	Hard	30%	25%
Comparison	Easy	38%	32%
Comparison	Medium	49%	47%
Comparison	Hard	37%	35%