Small vs. Large Language Models: A Comparative Study on Multiple-Choice Question Answering in Portuguese

Marilia Rosa Silveira
mrsilveira@inf.ufpel.edu.br
Federal University of Pelotas (UFPel)
Graduate Program in Computer Science (PPGC)
Artificial Intelligence Innovation Hub (H2IA)
Pelotas, Rio Grande do Sul, Brazil

Guilherme Dallmann Lima
gdlima@inf.ufpel.edu.br
Federal University of Pelotas (UFPel)
Graduate Program in Computer Science (PPGC)
Artificial Intelligence Innovation Hub (H2IA)
Pelotas, Rio Grande do Sul, Brazil

Carlos Augusto Calage Silveira Júnior
Federal University of Pelotas (UFPel)
Graduate Program in Computer Science (PPGC)
Artificial Intelligence Innovation Hub (H2IA)
Pelotas, Rio Grande do Sul, Brazil

Larissa Astrogildo Freitas
larissa@inf.ufpel.edu.br
Federal University of Pelotas (UFPel)
Graduate Program in Computer Science (PPGC)
Artificial Intelligence Innovation Hub (H2IA)
Pelotas, Rio Grande do Sul, Brazil

Ulisses Brisolara Corrêa
ulisses@inf.ufpel.edu.br
Federal University of Pelotas (UFPel)
Graduate Program in Computer Science (PPGC)
Artificial Intelligence Innovation Hub (H2IA)
Pelotas, Rio Grande do Sul, Brazil

ABSTRACT

Generative models are widely used for Multiple-Choice Question Answering (MCQA). While performance often improves with model size, prior work reports inconsistencies depending on task, prompting strategy, and language. We evaluate eleven open models ranging from millions to billions of parameters, both monolingual and multilingual, on Portuguese MCQA built from college entrance exams, under six prompting strategies: zero-shot, one-shot, few-shot, shuffled-order (to probe positional effects), and two per-option label-only settings. We also quantify positional bias using a normalized positional-bias coefficient (BPC). Overall, performance increases with parameter count; however, the magnitude varies across strategies. LLaMA-3.1-Storm-8B achieves the best average accuracy, and Sabiá-7B, a model trained with a strong focus on Portuguese, performs competitively among models of similar size. Smaller models (e.g., Tucano-2B, Qwen2-0.5B) attain solid results in specific settings, particularly with per-option scoring. These findings suggest that, although larger models are generally more robust, carefully chosen prompting can make smaller models viable under resource constraints. In summary, performance scales with size but depends on prompting—per-option configurations reduce the SLM–LLM gap, and positional bias is measurable via BPC; future work includes multi-shuffle BPC estimation, calibration, and log-likelihood baselines for per-option scoring and extensions to additional domains and languages.

KEYWORDS

Small Language Models, Low-Resource Languages, Evaluation Metrics

1 INTRODUCTION

Generative models have become central tools in natural language processing (NLP), particularly in tasks such as Multiple-Choice Question Answering (MCQA), where models must analyze the context, infer intent, and select an alternative from a list of options [10, 19]. The prevailing assumption has been that larger language models with more parameters and greater pretraining exposure offer superior performance across most NLP benchmarks, a view formalized in the widely cited *scaling law* [8]. The law proposes a near-linear correlation between model size and performance, especially on tasks requiring deep comprehension and reasoning.

However, this assumption has shown inconsistency across languages and task types. Recent evidence suggests that model efficiency, architecture, and language alignment may outweigh raw size in determining performance, especially in low-resource or morphologically rich languages such as Portuguese [9, 20]. While large multilingual models like LLaMA-8B have shown strong zero-shot performance on Portuguese MCQA datasets despite not being trained specifically for the language, monolingual models like Sabiá-7B still struggle with complex inferential tasks [15].

Simultaneously, the community has increasingly focused on smaller and more efficient models for practical deployments, particularly those requiring low latency or edge processing capabilities [29].

This paper investigates Portuguese Multiple-Choice Question Answering (MCQA) by comparing larger models (e.g., LLaMA-8B, Sabiá-7B) and lightweight alternatives (e.g., TTL-160M, TTL-460M, Samba-1.1B) across two datasets—ENEM-Challenge and BLUEX. We evaluate models under zero-shot and one-shot prompting to

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

approximate realistic usage, covering reading comprehension, encyclopedic knowledge, and domain-specific reasoning.

We consider monolingual and multilingual models spanning 160M to 8B parameters. Our goals are threefold: (i) assess how far scaling trends hold for MCQA in Portuguese; (ii) examine the size–efficiency trade-offs relevant to deployment; and (iii) understand how language-specific characteristics interact with model architecture and training data. The overarching aim is to provide practical guidance for designing affordable, context-aware language models suitable for linguistically diverse and resource-constrained settings.

This paper is structured as follows. We begin by discussing key background concepts in Section 2 and an overview of related work in Section 3. Section 4 outlines our methodology, and Section 5 presents and analyzes the results. We conclude with final remarks in Section 6.

2 BACKGROUND

In this section, we present fundamental concepts, addressing the distinction and relationship between language models and the multiple-choice question-solving task.

2.1 Language Models

Large language models (LLMs) are deep learning systems trained on vast textual corpora, capable of capturing linguistic patterns for a wide range of natural language understanding and generation tasks [18]. Their performance is closely linked to the number of parameters: larger models, in theory, can extract more complex patterns and better handle morphology and context. However, this increased capability comes with higher computational demands, making deployment challenging on resource-constrained devices [24]. For example, a large language model such as LLaMA-3-70B [13] may require up to 168 GB of memory in full precision, while a smaller language model like TTL-460M [2] fits within approximately 1 GB, highlighting the feasibility gap for real-world applications.

Language models can currently be divided into two categories: LLMs and SMLs. SMLs are typically defined as models with up to 2 billion parameters, while LLMs have more than 2 billion parameters [2, 3, 6, 7, 11, 21, 27].

2.2 Multiple Choice Question Answering

Within the field of NLP, a rapidly growing area of interest involves multiple-choice questions (MCQs), commonly referred to as MCQA. This field has gained prominence in recent years, driven by advances in language models. The MCQA task can be divided into two main approaches: MCQ generation and MCQ solving. In MCQ generation, a context is provided, and the model generates both the questions and the possible answer choices [17]. In contrast, MCQ solving involves presenting a question prompt followed by a set of alternatives, requiring the system to select the correct answer based on its understanding of the content [22]. Unlike simpler approaches that rely solely on keyword matching, MCQA demands that models truly comprehend the meaning of the text, make inferences, and even connect information from different sources [23].

This task is directly related to complex linguistic challenges such as semantic ambiguity, contextualization, and logical reasoning, making it one of the most demanding and promising frontiers in the field. As applications continue to advance in supporting education and assessment, MCQA stands out as an efficient solution for automating educational processes, personalizing learning, and providing more effective support to both students and educators.

The MCQA task remains a challenge for language models, as it requires deep semantic understanding. It often involves lengthy or ambiguous contexts, demands domain-specific knowledge, or even draws upon multiple areas of expertise. For these reasons, achieving high performance on this task remains difficult for many models. Precisely because of its high level of complexity, MCQA has recently been adopted as a benchmark for evaluating language models [25], highlighting its relevance in comparing models of different scales, such as small and large ones.

3 RELATED WORKS

Research in LLM models has been driven by an intense exploration of scaling laws. On one hand, massive models continue to break performance records. On the other hand, there is a growing interest in smaller versions, aiming to find a balance between performance and computational efficiency [1]. This pursuit has led to a paradigm shift in training philosophy, moving from a brute-force approach of "more data is always better" to a focus on the quality and nature of the training data [4].

A key insight towards this shift is that Small Language Models (SLMs) can achieve surprising reasoning and language capabilities if trained on a highly curated dataset. The Phi series of models from Microsoft is a representative example of this shift. In *Text Are All You Need* [5], the authors demonstrate that Phi-1, a 1.3B parameter model, achieved state-of-the-art performance among SLMs by training on a small corpus of meticulously selected textbook data. This model follows the approach explored by Eldan and Li [4], which showed that even a minuscule model could learn coherent language structures when trained on synthetic, simple stories, achieving a high-quality compact dataset.

Recent research has explored the development of SLMs. These models aim to strike a balance between efficiency and performance, providing alternatives to large-scale LLMs. For example, TinyLLaMA and MiniCPM demonstrate how compact architectures can approach or even rival the performance of larger models in selected tasks, enabling new solutions under resource constraints [2]. Regarding smaller models, the work by [23] proposes an efficient approach to improve the performance of smaller models in multiple-choice (MCQ) tasks with few-shot learning, using LLMs for synthetic data generation and knowledge distillation. Given the high computational cost of LLMs and the scarcity of annotated datasets, the authors seek ways to transfer knowledge from larger models to lighter models that are feasible in resource-limited environments. The study demonstrates that lighter models can achieve competitive performance when combined with approaches such as few-shot learning, distillation, and transfer learning, making their application viable in environments with limited computational resources.

Regardless of training philosophy, a credible comparison of model capabilities depends on robust evaluation. Multiple-Choice Question Answering (MCQA) is attractive for its straightforward

implementation and automated scoring, but recent work has surfaced reliability concerns. A growing literature documents positional bias, whereby the order of answer choices alone can strongly affect predictions. Zheng et al. [28] examined the robustness of LLMs and found preferences for specific identifiers (e.g., A/B), with substantial accuracy shifts when the correct option's position was changed—for instance, on MMLU, *gpt-3.5-turbo* varied from 60.9% to 74.2% depending on position. Wang et al. [26] likewise reported systematic positional bias and employed the Bias by Position of Choice (BPC) metric to quantify whether relocating the correct answer alters model outputs. Consistent with these findings, Li et al. [12] showed that randomizing option order leads to significant changes in decisions, underscoring the need for caution when relying solely on MCQA-based benchmarks.

Our work is situated at the intersection of these challenges. We aim to compare models that embody these different philosophies within the context of the Portuguese language. Critically, to ensure a valid comparison, we embed an evaluation of positional bias into our methodology. In doing so, our work offers a dual contribution: (1) a comprehensive performance overview of various models on Portuguese MCQA tasks, and (2) a critical analysis of how robustness to evaluation bias correlates with model scale, providing a more equitable and insightful assessment of their true capabilities.

4 METHODOLOGY

This section describes the methodology adopted in this study. Initially, the models for evaluation in the MCQA task were selected and organized into four categories. The first category consists of small models specialized in Portuguese (a): **Samba-1.1B**[1], **TTL-160M**[2], **TTL-460M**[3], and **Tucano-2B4**[4]. The second category includes the selected small multilingual models (b): **Qwen2-0.5B**[5], **TinyLlama-v1.1**[6], and **LLaMA-3.2-1B**[7]. The third category covers the large models specialized in Portuguese (c): **Sabiá-7B**[8] and **Bode-7B**[9]. Finally, the fourth category consists of the large multilingual models (d): **LLaMA-3.1-Storm-8B**[10] and **LLaMA-3.2-3B**[11].

a) Small Language Model Monolingual

Samba-1.1B is a model trained on Portuguese language data, based on TinyLlama-1.1B, a version of LLaMA-2 with 1.1 billion parameters. TTL160M and TTL-460M, in turn, are models based on the transformer architecture, pre-trained for Portuguese. They consist of two compact models for text generation, with 160 million and 460 million parameters, respectively. The Tucano series, on the other hand, includes decoder-transformer models pre-trained natively in Portuguese, all trained on GigaVerbo, a corpus of deduplicated Portuguese texts, totaling 200 billion tokens. This model has 2 billion parameters.

[1]Available at: https://huggingface.co/lrds-code/samba-1.1B
[2]Available at: https://huggingface.co/nicholasKluge/TeenyTinyLlama-160m
[3]Available at: https://huggingface.co/nicholasKluge/TeenyTinyLlama-460m
[4]Available at: https://huggingface.co/TucanoBR/Tucano-2b4
[5]Available at: https://huggingface.co/Qwen/Qwen2-0.5B
[6]Available at: https://huggingface.co/TinyLlama/TinyLlama_v1.1
[7]Available at: https://huggingface.co/meta-llama/Llama-3.2-1B
[8]Available at: https://huggingface.co/maritaca-ai/sabia-7b
[9]Available at: https://huggingface.co/recogna-nlp/bode-7b-alpaca-pt-br-no-peft
[10]Available at: https://huggingface.co/akjindal53244/Llama-3.1-Storm-8B
[11]Available at: https://huggingface.co/meta-llama/Llama-3.2-3B

b) Large Language Model Monolingual

Sabiá-7B is a language model trained for text generation, with 7 billion parameters, using a dataset extracted from the Portuguese portion of ClueWeb2022 [14]. The model was also designed to understand other Latin languages. Bode is an LLM developed for Portuguese based on LLaMA 2, through fine-tuning on the Alpaca dataset. The model has 7 billion parameters.

c) Small Language Model Multilingual

Qwen2-0.5B is the new series of models from Qwen. This model includes various base models and instruction-tuned models, with 500 million parameters. The TinyLlama-1.1B model is based on a Llama model with 1.1 billion parameters and 3 trillion tokens. LLaMA-3.2-1B, here referred to as LLaMA-1B, is a multilingual generative language model, pre-trained and instruction-tuned, with a size of 1 B. The instruction-tuned, text-only models of LLaMA-1B are optimized for use cases in multilingual dialogues, including information retrieval tasks and automatic summarization.

d) Large Language Model Multilingual

The model LLaMA-3.1-Storm-8B, which we refer to in this study as LLaMA-8B, and the LLaMA-3.2-3B, here referred to as LLaMA-3B, are models developed with the goal of enhancing the capabilities of medium-sized language models. LLaMA-8B represents an evolution in Meta AI's model line, having been built through an automated data curation and model fusion process.

The LLaMA-3.1-8B-Instruct model was developed with a focus on optimizing the quality of training data and fine-tuning efficiency, especially in contexts with limited computational resources. The LLaMA-3B is a generative language model, pre-trained and instruction-tuned, with 1B and 3B parameter variants, designed for text input and output tasks. It has been optimized for multilingual dialogue scenarios, including specific tasks such as information retrieval and text summarization. The training of the LLaMA 3.2 version involved up to 9 trillion tokens from public sources, and the model incorporated the logits of the LLaMA 3.1 models in the 8B and 70B parameter versions. With 3 billion parameters, LLaMA 3.2 is capable of handling various languages, including Portuguese.

Dataset

Initially, for model evaluation, we selected the ENEM dataset [16]. This dataset contains a total of 180 admission exam questions, known as the Exame Nacional do Ensino Médio (ENEM). The ENEM is used for admitting high school students to higher education. The questions in the dataset refer to the year 2023. As a preprocessing step, we excluded questions that involved images in their context, resulting in a total of 132 questions. Subsequently, for model evaluation, we included the BLUEX datasets, which contain a total of 1,261 multiple-choice questions. This set was constructed from questions from the entrance exams of the UNICAMP and USP universities, covering the period from 2014 to 2024. As with the ENEM dataset, we removed questions that required image comprehension or multilingual understanding, keeping only those written in Portuguese. The final set contains 245 questions from UNICAMP and 255 from USP, totaling 500 questions.

Experiments

For the experiments, we used six different approaches to analyze the models. The approaches consist of: using a prompt with the statement and all the possible alternatives in the modes (1) zero-shot, (2) one-shot, and (3) few-shot; modifying the order of the alternatives in the mode (4) zero-shot; and presenting the statement along with each alternative individually, selecting the one with the highest probability assigned by the model, in modes (5) zero-shot and (6) one-shot.

For approaches 1, 2, 3, and 4, we provided each question with all possible alternatives and instructed the model to respond with only one of the provided alternatives, without writing any explanation for the choice or any additional text. Specifically, in the 4th approach, we performed the reordering of alternatives to evaluate a possible positional bias that the models might exhibit. Below is an example of the prompt we used for the respective approaches.

Prompt

Question:
{Question}

Options:

A) *{option1}*
B) *{option2}*
C) *{option3}*
D) *{option4}*
E) *{option5}*

Instructions:

- Answer only with the letter corresponding to the correct option for the question (A, B, C, D, or E).
- Do not write any explanation or additional text.

Answer: ___

Figure 1: Prompt template used for experiments 1, 2, 3, and 4.

In approaches 5 and 6, each answer option was iteratively presented along with the question prompt. The model is shown the question and one candidate option and is asked to output a probability in [0,1] for that option; we repeat this for all options and choose the one with the highest probability. This approach was employed to ensure that the model would not exhibit positional bias and would retain the central context of the question, given that smaller models have more limited context windows compared to larger models. Below is an example of the prompt used for these approaches.

Evaluation Metrics

To evaluate model performance on the MCQA task using the ENEM and BLUEX datasets, we report accuracy, since in multiple-choice questions, answers are either correct or incorrect, with no scope for partial credit. We additionally compute the Bias by Position of Choice (BPC) to assess model behavior when the order of answer options is permuted in approach (4); for the remaining approaches, we report accuracy only. This metric, grounded in Total Variation Distance, measures how far a model's response distribution departs from a perfectly uniform distribution. Let $p = (p_1, \ldots, p_k)$ be the

Prompt

Question:
{Question}

Option:

- *{option}*

Probability: _____

Figure 2: Prompt template used for experiments 5 and 6

empirical distribution of chosen positions across items when the option order is shuffled. We compute

$$\mathrm{BPC}(p) = \frac{1}{1-\frac{1}{k}} \cdot \frac{1}{2} \sum_{i=1}^{k} \left| p_i - \tfrac{1}{k} \right|$$

where 0 indicates no positional bias and 1 indicates always choosing the same position. In our setup, we apply one shuffle per item (Method 4) and aggregate choices across items to estimate p. In this formulation, a BPC value close to 0 indicates that the model distributes its responses uniformly across the alternatives, suggesting little or no positional bias, whereas a high BPC reflects a strong preference for specific positions, indicating greater susceptibility to this type of bias.

5 RESULTS AND DISCUSSION

The following results show accuracy metrics increasing as model size grows, although other patterns can also be observed. In Tables 1, we present the macro-average percent accuracies of each model across the two datasets used in this study (ENEM and BLUEX). The evaluation over six prompting strategies: (1) zero-shot, (2) one-shot, (3) few-shot, (4) zero-shot with shuffled alternatives used for computing the Bias by Position of Choice (BPC), (5) zero-shot with individual alternative scoring, and (6) one-shot with individual alternative scoring. The final column shows the overall macro-average accuracy of each model across all prompting strategies. These results allow us to compare not only the general performance of the models, but also their sensitivity to variations in prompting configuration. In what follows, we discuss the key findings, considering the effects of model size, language support (monolingual vs. multilingual), positional biases, and the impact of different prompting strategies.

5.1 Scaling Laws and Model Size Effects

Analyzing Table 1, it is possible to infer that the scaling law [8] applies to MCQA tasks for the evaluated models. Although smaller models exhibit limitations in most tested scenarios, there are cases in which they achieve competitive performance. This occurs in part due to the strategies employed; for example, in setting (6), an example is provided and each option is presented individually alongside the question stem, enabling the model to better capture the context and produce a more accurate response. Under this configuration, the Tucano-2B model attains 37.12% accuracy, approaching the 39.39% achieved by the larger models under the prompt strategy (6).

Models	(1)	(2)	(3)	(4)	(5)	(6)	Average	BPC
TTL-160m	17.30%	25.69%	23.08%	19.69%	21.78%	28.03%	22.59%	0.7040
TTL-460m	18.84%	23.91%	23.72%	20.92%	24.10%	28.03%	23.25%	0.8844
Qwen2-0.5B	21.50%	17.54%	12.36%	20.68%	26.96%	30.30%	21.56%	0.5100
Samba-1.1B	19.66%	24.17%	22.46%	20.92%	26.27%	31.81%	24.21%	0.9678
Llama-1B	22.03%	21.93%	21.07%	21.63%	28.51%	31.06%	24.37%	0.7171
TinyLlama_v1.1B	21.00%	22.04%	21.59%	22.06%	26.62%	34.84%	24.69%	0.9975
Tucano-2b	23.81%	22.43%	20.58%	19.17%	33.34%	37.12%	24.82%	0.5065
LLaMA-3B	30.72%	40.53%	45.88%	31.73%	34.70%	32.63%	36.03%	0.5934
Bode-7B	28.83%	22.63%	25.09%	27.54%	29.81%	32.38%	27.71%	0.4366
Sabiá-7B	48.94%	50.02%	52.47%	50.83%	35.62%	**39.39**%	46.21%	**0.2452**
LLaMA-8B	**69.94** %	**68.77**%	**68.67**%	**69.52**%	**40.36**%	**39.39**%	**59.44**%	0.3135

Table 1: Average accuracies (in %) across the two datasets (ENEM and BLUEX) for six prompting strategies: (1) zero-shot, (2) one-shot, (3) few-shot, (4) zero-shot with shuffled alternatives , (5) zero-shot with individual alternative scoring, and (6) one-shot with individual alternative scoring. The BPC (Bias by Position of Choice) column presents a metric based on the Total Variation Distance, which quantifies positional bias. Lower BPC values indicate greater robustness of the model to changes in the order of answer options. The last column presents the macro-average accuracy of each model across the six prompting settings.

Analyzing the Figure 3, it becomes evident that the scaling law holds for MCQA tasks in Portuguese, as there is a clear trend of improved average accuracy with an increasing number of parameters. This trend is particularly pronounced, with significant performance gains observed around the 3B and 7-8B parameter counts. However, the figure also highlights crucial nuances beyond raw scale. We observe notable dispersion among models of similar size, such as the 7B models. In this category, Sabiá-7B (monolingual) demonstrates superior performance over Bode-7B (multilingual), suggesting that language-specific pre-training and data alignment can provide a significant advantage, compensating for potential architectural differences. This finding supports the argument that the quality and nature of pre-training data are as critical as scale. Furthermore, the plot illustrates how prompting strategies, particularly per-option scoring in setting (6), can narrow the performance gap. For instance, as noted in our results, smaller models like Tucano-2B can achieve performance levels close to their larger counterparts under these specific configurations. Nonetheless, when considering the macro-average performance across all evaluated strategies, the effect of model scale remains the predominant factor.

5.2 Positional Bias

Evaluating positional bias by altering the order of answer options revealed clear differences across models. In particular, four models showed marked contrasts across prompting configurations: LLaMA-3.1-Storm-8B (large multilingual), Sabiá-7B (large monolingual), Qwen2-0.5B (small multilingual), and Tucano-2B (small monolingual).

LLaMA-3.1-Storm-8B, the largest model, achieved 69.94% (zero-shot), 68.77% (one-shot), 68.67% (few-shot), and 69.52% (shuffled alternatives), with a variation of only 1.3 pp, indicating high robustness regardless of strategy. Similarly, Sabiá-7B remained consistent, with accuracies between 48.94% and 52.47% across the same four strategies (variation of 3.6 pp).

By contrast, smaller models varied more. Tucano-2B ranged from 19.17% (shuffled) to 23.81% (zero-shot), a 4.64 pp difference. Qwen2-0.5B ranged from 12.36% (few-shot) to 21.50% (zero-shot), a 9.1 pp difference. This instability indicates greater sensitivity to prompt design, especially in smaller models; we did not vary the number of answer options, so we refrain from drawing conclusions beyond this scope.

Figure 4 shows the distribution of selected alternatives for these four models under zero-shot, one-shot, and few-shot. In zero-shot, Qwen2-0.5B concentrated its responses on A, suggesting positional bias, whereas Tucano-2B spread choices across A, B, and C. In one-shot, Qwen2-0.5B increased diversity (A, C, and E), while Tucano-2B reduced variety (B and C). In few-shot, Qwen2-0.5B became dispersed again, and Tucano-2B maintained the one-shot pattern. Across approaches, larger models displayed a more uniform distribution over alternatives, suggesting greater generalization and lower position sensitivity, while Qwen2-0.5B was less stable, and Tucano-2B remained intermediate. These findings reinforce that positional bias is critical—especially in smaller models—and can undermine MCQA reliability when no mitigation (e.g., randomization or per-option scoring) is applied.

As shown in Table 1, BPC values (lower is better) vary widely, reflecting different levels of susceptibility to answer-position bias. Models with high BPC, such as TinyLlama_v1.1B (0.9975) and TTL-460m (0.8844), are strongly influenced by option order, making measured accuracy sensitive to presentation. In contrast, higher-performing models like Sabiá-7B (0.2452) and LLaMA-3.1-Storm-8B (0.3135) show substantially lower BPC. Overall, we observe a negative association between average accuracy and BPC: more robust models tend to be less position-sensitive and rely more on semantic content than on ordering heuristics. The trends observed in Figure 3 are further supported by our positional bias analysis. The higher-performing models, such as LLaMA-8B and Sabiá-7B, which occupy the top positions in the plot, also exhibit the lowest BPC values (0.3135 and 0.2452, respectively). This negative association between average accuracy and BPC suggests that models with greater scale and language alignment are not only more accurate but also more robust to simple variations in prompt design, relying more on semantic content than on ordering heuristics.

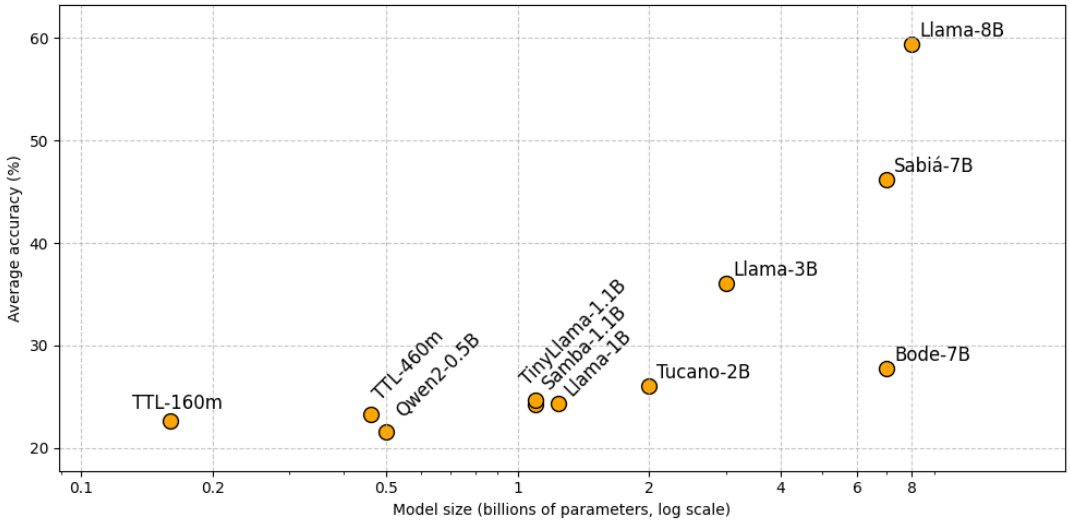

Figure 3: Average accuracy (%) as a function of model size (billions of parameters, log scale). Each point represents a model; accuracies are averaged over the two datasets (ENEM and BLUEX) and the six prompting strategies.

Therefore, reporting BPC alongside accuracy is essential to assess not only absolute performance but also robustness to simple presentation changes.

Based on these results, we can infer that in the larger models, the balanced distribution among the correct alternatives is closely related to their ability to identify the correct answer with greater accuracy. This indicates that these models are more effective in scenarios with a wide range of possible answers, due to their greater reasoning and generalization capabilities. In contrast, for the smaller models, a broader distribution among the alternatives tends to reflect lower accuracy in the responses, suggesting that these models are more efficient in contexts with a limited number of options, where the inference complexity is lower.

5.3 Impact of Prompting Strategy

The analysis of mean accuracy across strategies reveals that while larger models tend to be more stable, smaller models exhibit greater fluctuations depending on the prompt configuration. For larger models such as LLaMA-8B and Sabiá-7B, accuracy remained relatively stable across the zero-shot, one-shot, and few-shot settings. This stability indicates lower sensitivity to configuration variations, reinforcing the robustness of these models.

In contrast, smaller models exhibited more pronounced fluctuations. Qwen2-0.5B, for example, showed inconsistent accuracy in settings such as few-shot, where its performance dropped significantly. Moreover, in some few-shot cases we observed a tendency to replicate the alternatives provided in the example, which reduced the diversity of responses. These findings indicate that, whereas larger models demonstrate robustness largely independent of prompting type, smaller models are more sensitive to the choice of prompting strategy, underscoring the central role of prompt design for their practical use.

The dependence on prompting strategies becomes even more evident when examining approaches (5) and (6), which evaluate each alternative individually by probability. Under these configurations, some smaller models—such as Tucano-2B and Qwen2-0.5B—substantially narrowed the gap to larger models. Tucano-2B, for instance, reached 37.12% in strategy (6), even surpassing Sabiá-7B under strategy (5). These findings suggest that, when guided by well-crafted, task-specific prompts, smaller models can deliver competitive performance, reinforcing the decisive role of prompt design in unlocking their potential in resource-constrained settings.

5.4 Monolingual vs. Multilingual

When comparing monolingual and multilingual models, it becomes clear that linguistic alignment is crucial for MCQA performance in Portuguese. Although the Sabiá-7B relies on an older architecture, its pretraining on large-scale Portuguese corpora allows it to achieve competitive results, in some cases surpassing newer multilingual models such as LLaMA-3B. A notable example occurs under strategy (6), where Sabiá-7B reaches an accuracy of 39.39%, matching the score obtained by LLaMA-8B, the most powerful model in our evaluation. This suggests that language-specific pretraining can compensate for architectural limitations, particularly in morphologically rich languages where tokenization and semantic disambiguation are challenging. In contrast, smaller multilingual models like Qwen2-0.5B and LLaMA-1B consistently underperform, indicating that small-scale multilingual pretraining may dilute linguistic coverage and hinder reasoning capabilities. These findings highlight the importance of balancing model size and linguistic specialization, suggesting that, in certain contexts, intermediate-sized monolingual models can be a more robust and cost-effective solution than larger multilingual alternatives, while also representing a promising direction for future research.

Based on the results obtained, we selected representative models from each category, focusing on the approaches with the greatest performance discrepancies (1, 2, 3, and 4). The best representatives from each group were: Tucano-2B (Small Monolingual Language

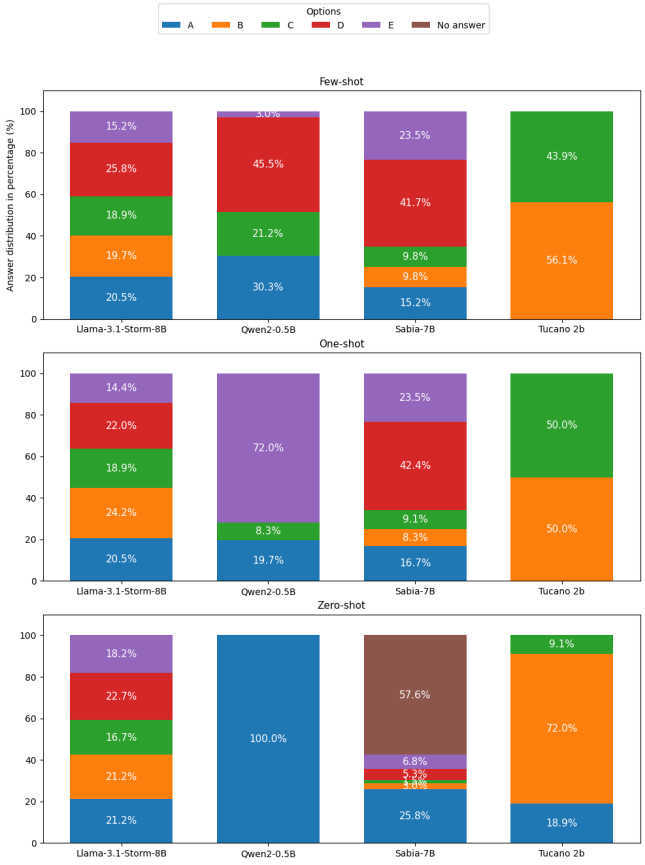

Figure 4: Percentage distribution of selected responses by different models. Each chart presents the evaluated models under a specific configuration (shot type): zero-shot, one-shot, or few-shot. The stacked bars represent the relative frequency of each alternative (A–E), including no-response cases.

Model), Qwen2-0.5B (Small Multilingual Language Model), Sabiá-7B (Large Monolingual Language Model), and LLaMA-8B (Large Multilingual Language Model). When evaluating the responses generated by these models across all alternatives, we observed that the larger models—which also achieved the best performances—exhibited a broader distribution among the inferred alternatives, whereas the smaller models concentrated their answers on a more limited set of options.

6 CONCLUSION

The results obtained in this study highlight the strong correlation between the number of parameters in language models and their performance on comprehension and reasoning tasks, particularly in a morphologically rich language such as Portuguese. The LLaMA-8B model stood out significantly, outperforming the others across nearly all evaluated approaches, underscoring the importance of modern and scalable architectures.

On the other hand, the Sabiá-7B model demonstrated that pretraining specifically focused on the Portuguese language can mitigate architectural limitations and lead to competitive performance. This observation highlights the value of linguistic and cultural alignment in the training of LLMs tailored for specific languages.

Smaller models, such as Tucano-2B and Qwen2-0.5B, showed satisfactory results in specific contexts, particularly in approaches 5 and 6. This suggests that, although they have lower generalization capacity, these models can still provide effective responses when properly aligned with the application context. Furthermore, their behavior in settings such as zero-shot, one-shot, and few-shot revealed important differences in how they infer answers, especially regarding diversity and sensitivity to positional bias.

The results emphasize that relying solely on the MCQA task to validate a model's knowledge acquisition is not robust, as the factors influencing answer accuracy metrics are broad and multifaceted. Nevertheless, the analyzed data demonstrate that larger models tend to be more robust, versatile, and generalist, whereas smaller models can serve as viable and efficient alternatives in more constrained contexts or under limited computational resources.

As future work, we propose enhancing smaller models through fine-tuning techniques specifically tailored to the Portuguese language, aiming to improve their performance in zero-shot and few-shot settings; conducting qualitative analyses of the generated responses to better understand the reasoning strategies employed and the most frequent types of errors; and expanding the experiments to other morphologically rich languages to assess the generalizability of the obtained results.

REFERENCES

[1] Zhengyu Chen, Siqi Wang, Teng Xiao, Yudong Wang, Shiqi Chen, Xunliang Cai, Junxian He, and Jingang Wang. 2025. Revisiting Scaling Laws for Language Models: The Role of Data Quality and Training Strategies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 23881–23899. doi:10.18653/v1/2025.acl-long.1163

[2] Nicholas Kluge Corrêa, Sophia Falk, Shiza Fatimah, Aniket Sen, and Nythamar De Oliveira. 2024. Teenytinyllama: open-source tiny language models trained in brazilian portuguese. *Machine Learning with Applications* 16 (2024), 100558.

[3] Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759* (2023).

[4] Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How Small Can Language Models Be and Still Speak Coherent English? arXiv:2305.07759 [cs.CL] https://arxiv.org/abs/2305.07759

[5] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks Are All You Need. arXiv:2306.11644 [cs.CL] https://arxiv.org/abs/2306.11644

[6] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. [n. d.]. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024. *URL https://arxiv. org/abs/2404.06395* ([n. d.]).

[7] Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog* 1, 3 (2023), 3.

[8] Jared Kaplan, Sam McCandlish, Tom Henighan, et al. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[9] Aisha Khatun and Daniel G Brown. 2024. A Study on Large Language Models' Limitations in Multiple-Choice Question Answering. *arXiv preprint arXiv:2401.07955* (2024).

[10] Yassine Labrak, Adrien Bazoge, Romain Dufour, et al. 2023. FrenchMedMCQA: A French multiple-choice question answering dataset for medical domain. *arXiv*

preprint arXiv:2304.04280 (2023).

[11] Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Zhijing Wu, and Yiqun Liu. 2025. Blade: Enhancing black-box large language models with small domain-specific models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 24422–24430.

[12] Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of LLMs? *arXiv preprint arXiv:2403.17752* (2024).

[13] Meta AI. 2024. Meta LLaMA 3 70B: An open-weight large language model. https://huggingface.co/meta-llama/Meta-Llama-3-70B. Accessed: 2025-05-07.

[14] Arnold Overwijk, Chenyan Xiong, Xiao Liu, Cameron VandenBerg, and Jamie Callan. 2022. ClueWeb22: 10 Billion Web Documents with Visual and Semantic Information. arXiv:2211.15848 [cs.IR] https://arxiv.org/abs/2211.15848

[15] Rafael Pires, Heitor Abonizio, Thiago Santos Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. *Brazilian Conference on Intelligent Systems* (2023), 226–240.

[16] Ramon Pires, Thales Sales Almeida, Hugo Abonizio, and Rodrigo Nogueira. 2023. Evaluating GPT-4's Vision Capabilities on Brazilian University Admission Exams. arXiv:2311.14169 [cs.CL]

[17] Mamillapalli Chilaka Rao, P Sreedhar, M Bhanurangarao, and G Sujatha. 2022. Automatic multiple-choice question and answer (MCQA) generation using deep learning model. In *International Conference on Information and Management Engineering*. Springer, 1–8.

[18] Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353* (2022).

[19] Roberto Rodriguez-Torrealba, Enrique Garcia-Lopez, and Antonio Garcia-Cabot. 2022. End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications* 208 (2022), 118258.

[20] Giorgio Sarti and Malvina Nissim. 2024. It5: Text-to-text pretraining for italian language understanding and generation. *arXiv preprint arXiv:2402.13513* (2024).

[21] Timo Schick and Hinrich Schütze. 2020. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118* (2020).

[22] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine* (2025), 1–8.

[23] Patrick Sutanto, Joan Santoso, Esther Irawati Setiawan, and Aji Prasetya Wibawa. 2024. Llm distillation for efficient few-shot multiple choice question answering. *arXiv preprint arXiv:2412.09807* (2024).

[24] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.

[25] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024. LLMs May Perform MCQA by Selecting the Least Incorrect Option. *arXiv preprint arXiv:2402.01349* (2024).

[26] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926* (2023).

[27] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385* (2024).

[28] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882* (2023).

[29] Lianmin Zheng, Wei-Lin Chiang, Yizhong Sheng, et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685* (2023).