Textbook-Enriched Training for Language Models: Boosting Answer Quality in Specialized Contexts

Lucas B. Bulcão Mota[1], Larrissa Dantas[1], Daniela Barreiro Claro[1], Aline Paes[2], Claudia Freitas[1,3], Marlo Souza[1], Helena Caseli[4], Livy Real[5,6]

[1]FORMAS Research Center on Data and Natural Language
Institute of Computing, Federal University of Bahia (UFBA)
Salvador, BA, Brazil
{lucasbulcao,larrissasilva,dclaro,msouza1}@ufba.br
[2]Institute of Computing, Fluminense Federal University (UFF)
Niterói, RJ, Brazil
alinepaes@ic.uff.br
[3]Institute of Letters, Federal University of Bahia (UFBA)
Salvador, BA, Brazil
maclaudia.freitas@gmail.com
[4]Federal University of São Carlos (UFSCar), São Carlos, SP, Brazil
helenacaseli@ufscar.br
[5]Kunumi Institute, Belo Horizonte, MG, Brasil
[6]Federal University of Amazonas, Manaus, AM, Brasil.
livy@kunumi.com

Abstract. *The use of textbooks as primary sources of information has increasingly given way to tools based on Large Language Models (LLMs), raising concerns about the reliability of generated answers. This study investigates how different adaptation strategies shape the behavior of small language models in educational Question Answering (QA) tasks in Portuguese. To support this analysis, we built a question-answer dataset derived from an NLP textbook and compared base models and the Retrieval-Augmented Generation (RAG) pipeline with models adapted through supervised fine-tuning and Continued Pretraining. The evaluation relies on questions from the LARI dataset, which has been validated by human specialists, and combines automatic and qualitative assessment procedures. The findings indicate that small models tuned with structured instructional knowledge achieve stronger semantic alignment and produce more pertinent answers in Portuguese educational QA scenarios.*

1. Introduction

Large Language Models (LLMs) have increasingly been incorporated into educational settings as tools for information retrieval and knowledge generation [Yan et al. 2024, Xing et al. 2025]. Although these models can produce fluent and context-aware responses, recent studies report limitations in factuality, *grounding*, and the reliability of generated content, particularly in technical domains [Ji et al. 2023, Huang et al. 2025]. In *Question Answering* (QA) tasks, such limitations become especially problematic, since conceptual inconsistencies may directly affect the learning process.

Strategies such as *Supervised Fine-Tuning* (SFT), aimed at model specialization, and *Retrieval-Augmented Generation* (RAG), which incorporates external knowledge during generation, have been employed to enhance the performance of LLMs in specialized

domains [Lewis et al. 2020, Gao et al. 2023]. Even so, the behavior of small language models under these adaptation settings in Portuguese educational scenarios remains insufficiently studied.

At the same time, technical textbooks have been used as sources of specialized knowledge for language models [Gunasekar et al. 2023]. In the context of Natural Language Processing (NLP) in Portuguese, the book *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português* [Caseli and Nunes 2026], developed by the Brasileiras em PLN (BPLN) group, stands out as one of the main reference materials in the field. Alongside this initiative, the LARI *dataset* [Junqueira et al. 2026] was built from the BPLN NLP textbook and contains context-question-answer triples evaluated by human specialists.

In this work, we rely exclusively on the third edition of the textbook [Caseli and Nunes 2024] as the knowledge source for both the construction of an instruction-based question-answer pairs dataset, named here as *Carapicu-textbook*, and the adaptation of small language models to act as experts into the book. We investigate two adaptation strategies: supervised fine-tuning (SFT) using *Carapicu-textbook* and continued pretraining (CPT) on the raw content of the book followed by SFT (CPT-SFT). The experiments include base models of different scales, adapted variants of these models, and a RAG pipeline built from the textbook content.

Although recent studies have examined LLM adaptation and contextual retrieval for QA in Portuguese [Costa 2024], analyses involving small models adapted with structured instructional knowledge and evaluated on educational *benchmarks* validated by human specialists are still scarce. The results indicate that adapted small models can outperform larger non-specialized models in semantic alignment and answer relevance metrics.

2. Related Work

LLM adaptation strategies have been widely studied as a way to align pretrained models with specialized domains. Among these approaches, *Supervised Fine-Tuning* (SFT) incorporates domain knowledge directly into model parameters [Ouyang et al. 2022, Gururangan et al. 2020] through instruction-based datasets. Research on parametric knowledge indicates that LLMs can store factual information within their parameters [Petroni et al. 2019, Roberts et al. 2020]. Even so, gains in automatic evaluation metrics do not necessarily correspond to higher factual fidelity or semantic consistency [Ji et al. 2023]. Models adapted exclusively through parametric learning also remain prone to generating incorrect information, a limitation that becomes particularly problematic in *Question Answering* (QA) tasks.

As an alternative, *Retrieval-Augmented Generation* (RAG) approaches integrate external retrieval mechanisms into the text generation process [Lewis et al. 2020]. Recent studies report improvements in factuality and *grounding* when retrieval-based strategies are adopted [Gao et al. 2023, Mallen et al. 2023]. At the same time, these methods still face issues related to irrelevant document retrieval and the handling of long contextual inputs [Liu et al. 2024, Asai et al. 2024].

Within the Brazilian context, recent studies have started to examine the behavior of LLMs in Portuguese for tasks such as summarization and QA [Assis et al. 2025].

At the same time, the growing availability of small *open-weight* models has intensified discussions around computational cost and model adaptation in resource-constrained environments.

QA *benchmarks* play a central role in the evaluation of language models. Resources such as SQuAD [Rajpurkar et al. 2016] helped establish this evaluation paradigm, although they were originally designed for English. In this setting, the LARI *dataset* [Junqueira et al. 2026] introduces a Portuguese educational QA resource derived directly from the BPLN NLP textbook [Caseli and Nunes 2024]. The dataset consists of context–question–answer triples reviewed by human specialists and was designed specifically for educational scenarios in Brazil.

Despite recent progress in LLM adaptation, contextual retrieval, and QA in Portuguese [Costa 2024], studies examining small models adapted with structured instructional knowledge and evaluated on educational *benchmarks* validated by human specialists remain scarce. The influence of controlled didactic materials on the textual and conceptual behavior of adapted models for educational QA in Portuguese also remains largely unexplored.

3. Resources and Datasets

This section presents the main resources used in the experiments, including the textbook adopted as the knowledge source (Section 3.1), the evaluation *dataset* (Section 3.2), and the models considered in the study (Section 3.3).

3.1. Textbook: BPLN book

As the primary knowledge source for model adaptation and evaluation, we used the third edition of the book *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*[1] [Caseli and Nunes 2024], hereafter referred to as the BPLN NLP textbook. The book covers theoretical foundations, linguistic resources, and contemporary NLP applications focused on Brazilian Portuguese, and has become one of the main educational references in the field.

The choice of this textbook was motivated by the presence of structured technical content reviewed by specialists and closely aligned with NLP education in Portuguese. Unlike heterogeneous corpora automatically collected from the Web, the textbook presents greater terminological and conceptual consistency across chapters and topics. Table 1 reports general statistics of the book computed using the Qwen3 0.6B tokenizer.

In the context of this study, the content of the BPLN NLP textbook was used as the sole knowledge source for both parametric model adaptation and the document collection employed in the RAG-based *pipeline*. The textual content of the book was organized at the chapter and section levels while preserving its original thematic structure. Details regarding text processing and the use of this resource in each experimental strategy are presented in Section 4.

[1] https://brasileiraspln.com/livro-pln/3a-edicao/

Table 1. General statistics of the textbook computed using the Qwen3 0.6B tokenizer

Metric	Value
Total number of chapters	33
Total number of sections	125
Total number of textbook tokens	126103
Average number of tokens per chapter	3821.3
Average number of tokens per section	1008.82

3.2. LARI Dataset

For model evaluation, we used the LARI *dataset*[2] [Junqueira et al. 2026], developed specifically for educational *Question Answering* tasks in Brazilian Portuguese. The dataset was derived directly from the BPLN NLP textbook and consists of context-question-answer triples related to the NLP domain.

Beyond the construction of the triples, LARI also includes a human evaluation stage conducted by the textbook authors, in which each instance received scores ranging from 1 to 5 according to quality and pedagogical adequacy criteria.

Unlike automatically translated *datasets*, LARI was originally created in Portuguese and validated by human specialists in the field. We used exclusively the questions that received the highest score (5), resulting in 238 instances out of a total of 464. This decision was made to reduce potential ambiguities in the evaluation set and to ensure that the benchmark consisted only of instances considered to have the highest pedagogical quality by the dataset creators. Consequently, the evaluation focuses on high-quality reference questions, making the comparison between models more reliable and reducing the influence of potential noise in the benchmark. The *dataset* was employed solely for evaluation purposes and was not used during any stage of model training or adaptation. This separation between adaptation and evaluation data was intentionally designed to prevent data leakage and to assess the models ability to generalize beyond the instructional material seen during training.

3.3. Evaluated Models

We select Qwen3 0.6B, Qwen3 8B, and Sabiá 7B models to investigate how model scale, linguistic specialization, and adaptation strategies influence educational QA tasks in Portuguese. The Qwen3 models were selected because they are open models belonging to the same architectural family [Yang et al. 2025], allowing a more controlled comparison across different parameter scales. While Qwen3 0.6B represents a compact and computationally accessible model, Qwen3 8B offers greater capacity for textual representation and reasoning. The model family also includes multilingual support and capabilities targeted at generative and text understanding tasks.

We also evaluated Sabiá 7B, a commercial model developed for Brazilian Portuguese [Pires et al. 2023b]. Its inclusion makes it possible to analyze the effects of linguistic specialization in tasks involving technical NLP concepts in Portuguese [Assis et al. 2025].

[2]https://huggingface.co/datasets/jjuliar/plndataset

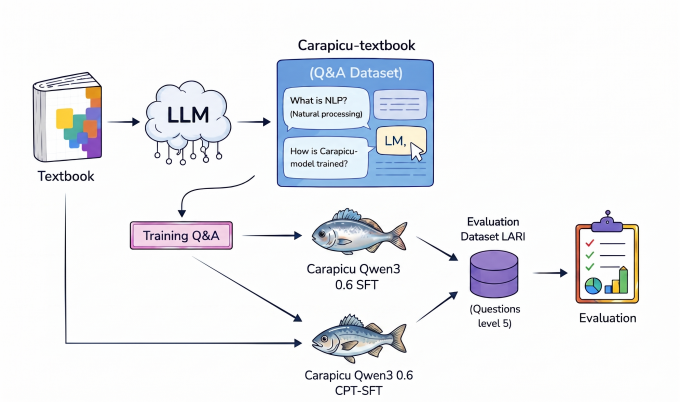

Figure 1. Methodological pipeline adopted in this study.

Beyond the base models, we evaluated the performance of Qwen3 0.6B within a RAG pipeline built on top of the BPLN NLP textbook content, using Qwen3 Embedding 0.6B for question vectorization and Qwen3 Reranker 0.6B to retrieve the most semantically related textbook sections.

4. Methodology: Carapicu Pipeline

Figure 1 presents an overview of the proposed *pipeline*. The BPLN NLP textbook [Caseli and Nunes 2024] was used as the sole knowledge source for the construction of the *carapicu-textbook* dataset, which was later employed for model adaptation through supervised fine-tuning and, in one variant, continued pretraining.

After the adaptation stage, the models were evaluated exclusively using LARI *dataset* questions that received the maximum score (5), resulting in a total of 238 triples. The generated answers were analyzed using BERTScore and an *LLM-as-a-Judge* setup, in which Sabiá-4 evaluated factual correctness, relevance, and completeness on a Likert scale ranging from 1 to 5. The following subsections describe the procedures adopted in each stage of the *pipeline*.

4.1. *Dataset Carapicu-textbook*

The *Carapicu-textbook* dataset was constructed exclusively from the third edition of the book *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português* [Caseli and Nunes 2024], which served as the sole knowledge source. The use of a single educational resource allows us to investigate how language models behave when exposed to structured and domain-specific knowledge in Portuguese. Unlike heterogeneous corpora automatically collected from the Web, the textbook presents stronger conceptual consistency, terminological standardization, and pedagogical organization.

The textual content was initially extracted and organized in JSON format while preserving the original chapter and section structure. Each section was then submitted to the GPT-4o-mini API together with a *prompt*[3] designed to generate 20 question-answer pairs divided into three thematic groups: (i) foundations and concepts; (ii) techniques and models; and (iii) applications and challenges.

[3]https://github.com/FORMAS/carapicu-model

The resulting resource, named *Carapicu-textbook*, consists of a collection of question-answer pairs derived from the textbook content and was used during the model adaptation stage. Table 2 presents general statistics of the *dataset*.

In addition to the automatic generation of question-answer pairs, we conducted a qualitative analysis of the *Carapicu-textbook* dataset through manual inspection of a representative sample. This analysis aimed to verify conceptual consistency, semantic alignment, and adherence to the content of the BPLN NLP textbook. Overall, the generated pairs exhibited thematic coherence and alignment with the concepts presented in their respective sections. We also observed a predominance of short and objective answers, a characteristic that was later reflected in the behavior of the adapted models. Unlike the LARI dataset, whose reference answers are generally more detailed, the Carapicu-textbook emphasizes concise responses. This difference should be taken into account when interpreting the completeness scores reported in the evaluation. Although a full manual validation of the entire dataset was not performed, the sampling-based review allowed us to identify occasional inconsistencies and assess the overall quality of the resource used in the experiments.

Table 2. General statistics of the Carapicu-Textbook dataset

Metric	Value
Total number of QA pairs	1,875
Total number of *dataset* tokens	220,726
Average number of user tokens (question)	18.47
Average number of model tokens (answer)	40.25

4.2. Model Adaptation

The model adaptation stage was designed to investigate how different specialization strategies affect small language models in educational QA tasks in Portuguese. For this purpose, we selected Qwen3 0.6B, an SLM with lower computational cost and inference time, using the *Carapicu-textbook* dataset as the main adaptation resource.

In this work, we considered two strategies: supervised fine-tuning (SFT) and *continued pretraining* (CPT) followed by SFT (CPT-SFT), both applied to Qwen3 0.6B.

In the first strategy, named Carapicu-Qwen3-0.6B-SFT, the model was fine-tuned using the question–answer pairs from *Carapicu-textbook*. Hyperparameter configuration was obtained through Bayesian Optimization and HyperBand (BOHB) [Falkner et al. 2018] using the Weights & Biases (WandB)[4] platform.

In the second strategy, named Carapicu-Qwen3-0.6B-CPT-SFT[5], we conducted a CPT stage using the raw textbook content in order to expose the model to NLP terminology and concepts before supervised fine-tuning. The resulting model was then submitted to the same SFT procedure using the question–answer pairs from *Carapicu-textbook*.

The two strategies allow a comparison between a model directly adapted for QA and another previously exposed to the textual domain of the technical textbook before

[4]`https://wandb.ai`

[5]https://huggingface.co/FORMAS/Carapicu-Qwen3-0.6B-CPT-SFT

supervised fine-tuning. Table 3 presents the final hyperparameters obtained during the optimization process, allowing for the reproduction of the configuration used in both adaptation strategies.

Table 3. Training hyperparameters of the Carapicu models

model	lr	epochs	batch size	eval/loss
Carapicu-Qwen3-0.6B-SFT	3e-5	4	2	1.50
Carapicu-Qwen3-0.6B-CPT-SFT7	3e-5	3	2	1.58

4.3. Retrieval-Augmented Generation

We also developed a *Retrieval-Augmented Generation* (RAG) pipeline to verify whether Qwen3 0.6B would be able to answer adequately when provided with the textbook sections that were semantically closest to the query. We adopted an *Advanced RAG* approach incorporating a *reranker* model to refine the retrieved passages before inference.

The vector database was built using ChromaDB, all sections of the BPLN NLP textbook were indexed and embedded using Qwen3 Embedding 0.6B. We then employed Qwen3 Reranker 0.6B to reorder the retrieved passages: initially, 10 sections were retrieved, and the 3 highest-ranked sections were used as context for answer generation.

During inference, we used a *prompt*[6] instructing the model to answer exclusively based on the provided context. We also set the temperature to 0 and limited the generated responses to a maximum of 512 *tokens*.

4.4. Evaluation

The answers generated by the models were evaluated using BERTScore [Zhang et al. 2019] and a qualitative assessment based on the *LLM-as-a-Judge* paradigm [Lin et al. 2021]. This strategy was adopted to analyze not only textual similarity between generated and reference answers, but also aspects related to conceptual adequacy, relevance, and completeness of the responses produced by the models. Since reference answers are available, traditional automatic metrics can be reliably used for model ranking, as prior work has shown strong alignment between these metrics and human judgments in ranking settings [Thakur et al. 2024].

4.4.1. BERTScore

As the main automatic evaluation metric, we adopted BERTScore [Zhang et al. 2019], an approach based on contextual *embeddings* that measures semantic similarity between texts using representations produced by Transformer models. Unlike metrics based solely on lexical overlap, such as BLEU [Papineni et al. 2002] and ROUGE [Lin 2004], BERTScore can capture semantic similarity even in scenarios involving paraphrases and textual reformulations.

In this work, BERTScore was computed using BERTimbau [Souza et al. 2020] to compare the answers generated by the models against the reference answers from the LARI *dataset*, allowing us to examine the semantic alignment of the produced responses.

[6] https://github.com/FORMAS/carapicu-model

4.4.2. Sabiá-4 as Judge

In addition to automatic metrics, we adopted an *LLM-as-a-Judge* strategy [Zheng et al. 2023], in which a language model acts as an automatic evaluator of the generated responses. Sabiá-4 was the judge model because it was specifically developed for Brazilian Portuguese, offering stronger linguistic and terminological alignment for the evaluation of educational answers in Portuguese [Pires et al. 2023a]. The use of an external model also helps avoid self-evaluation bias [Xu et al. 2024].

Sabiá-4 assigned scores on a Likert scale [Likert 1932] ranging from 1 to 5 for three criteria: (i) factual correctness; (ii) answer relevance with respect to the question; and (iii) completeness. During the qualitative analysis, we observed that the adapted models predominantly generated short and objective answers, reflecting the characteristics of the *Carapicu-textbook* used during training. In contrast, the LARI evaluation dataset generally contains more detailed reference answers. Therefore, the lower completeness scores observed for the adapted models likely reflect differences between the response style learned during adaptation and the style expected by the evaluation benchmark, rather than an actual lack of conceptual knowledge.

5. Results

Table 4 presents the results obtained with BERTScore. In general, the models adapted using the *Carapicu-textbook* dataset achieved the highest semantic similarity scores when compared to the non-specialized base models. The best performance was obtained by Carapicu-Qwen3-0.6B-CPT-SFT, which reached an F1 score of 0.5354, followed by Carapicu-Qwen3-0.6B-SFT with an F1 score of 0.5313. These results indicate that both supervised fine-tuning and the additional continued pretraining stage contributed to bringing the generated answers closer to the reference content contained in the LARI *dataset*.

The results also indicate that smaller models adapted to the educational domain represented by the BPLN NLP textbook can outperform larger non-specialized models. Although Qwen3-8B has substantially higher parametric capacity, its BERTScore performance remained below that of the adapted models, reaching an F1 score of 0.3978. This behavior supports the idea that supervised adaptation using structured instructional knowledge may exert a stronger influence on answer alignment than simply increasing the parameter scale of the model. It is also worth mentioning that the maximum generation length during inference was fixed at 512 *tokens*. Under this restriction, Qwen3-8B frequently exhausted the available budget while producing its reasoning chain (through *think* blocks), leading to truncated outputs and preventing the model from generating a final answer.

We also observed that the RAG pipeline achieved competitive performance in terms of semantic *recall*, obtaining the highest *recall* value among all evaluated models (0.5823). This behavior suggests that retrieving passages from the BPLN NLP textbook favors broader semantic coverage of the generated answers, although this does not always translate into the highest overall alignment measured by F1.

The results obtained by Sabiá-7B indicate that linguistic specialization alone is insufficient to guarantee stronger performance in specialized educational *Question Answering* tasks. Although the model was developed specifically for Brazilian Portuguese, its

BERTScore performance remained below that of the models adapted with the *Carapicu-textbook* dataset. During qualitative analysis, we observed signs of textual degeneration, including repetitive loops, semantically incoherent sentences, and the generation of random characters. It is worth noting that the evaluated version corresponds to the first public release of the Sabiá family available on Hugging Face.

Table 4. Average BERTScore values obtained by the evaluated models

Model	Precision	Recall	F1
Carapicu-Qwen3-0.6B-CPT-SFT	0.5267	0.5526	0.5354
Carapicu-Qwen3-0.6B-SFT	0.5200	0.5513	0.5313
Answer_RAG	0.4529	0.5823	0.5028
Qwen3-0.6B	0.3995	0.5443	0.4575
sabia-7b	0.4284	0.4938	0.4534
Qwen3-8B	0.3453	0.4825	0.3978

Table 5 presents the results of the qualitative evaluation based on the *LLM-as-a-Judge* approach. Unlike BERTScore, which measures semantic alignment with respect to the reference answers, this evaluation examines aspects related to *factual correctness*, *relevance*, and *completeness* of the generated responses.

The results show that the adapted models achieved stronger performance mainly in the relevance dimension. Carapicu-Qwen3-0.6B-CPT-SFT obtained the highest average score for this criterion (3.92), followed by Carapicu-Qwen3-0.6B-SFT (3.85).

In the factual correctness dimension, the retrieval-based *pipeline* achieved the best overall result (3.30), indicating a greater ability to produce factually consistent answers. The adapted models also obtained competitive performance in this criterion, outperforming the base models from the Qwen3 family.

With respect to completeness, we observed more modest results for the adapted models. During qualitative analysis, we found that these models frequently produced shorter and more objective answers, a behavior possibly associated with the characteristics of the *Carapicu-textbook* dataset used during training. For this reason, short answers were not automatically considered inadequate during qualitative evaluation as long as they were conceptually sufficient to answer the proposed question correctly.

We also observed that larger models did not necessarily achieve stronger qualitative performance. Despite its higher parametric capacity, Qwen3-8B obtained the lowest average scores among the evaluated Qwen models across all three analyzed criteria.

Table 5. *LLM-as-a-Judge* results using Sabiá-4. Answers between a Likert scale ranging from 1-strongly disagree to 5-strongly agree.

model	factual correctness	SD	relevance	SD	completeness	SD
Answer_RAG	3.30	±1.56	3.76	±1.62	3.16	±1.46
Qwen3-0.6B	2.78	±1.50	3.34	±1.73	2.67	±1.40
Qwen3-8B	2.13	±1.78	2.08	±1.75	1.85	±1.42
Carapicu-Qwen3-0.6B-SFT	2.97	±1.31	3.85	±1.34	2.70	±1.10
Carapicu-Qwen3-0.6B-CPT-SFT	3.04	±1.28	**3.92**	±1.29	2.79	±1.11
Sabia-7B	1.31	±0.92	1.40	±1.11	1.17	±0.61

6. Discussions

Our results indicate that *fine-tuning* with structured instructional knowledge positively affected the semantic alignment of the generated answers. Both Carapicu models achieved the highest BERTScore F1 values, outperforming the base models and Sabiá-7B, suggesting that direct exposure to the textbook content contributed to bringing the generated answers closer to the LARI reference responses at the semantic level. The results also indicate that small models can benefit considerably from specialization strategies when the domain is well delimited and the training material presents consistent conceptual organization. The comparison between Qwen3-0.6B and the Carapicu models reinforces this interpretation, since all models share the same architectural foundation.

Another aspect observed in the experiments is that increasing parameter scale did not necessarily lead to better performance. Qwen3-8B obtained lower results than the adapted models and even lower scores than its smaller counterpart, Qwen3-0.6B. This behavior suggests that larger models may require a greater number of *tokens* to structure their reasoning chains. In practice, we observed that Qwen3-8B frequently exhausted the entire generation budget during internal processing (*think* blocks), resulting in truncated and incomplete answers. This finding carries critical implications for resource-constrained environments. In local deployment scenarios, where GPU memory, power budget, and latency bounds are strictly limited, larger models not only incur higher computational overhead, but their tendency to produce verbose internal reasoning steps can actively degrade usability.

The results obtained with Sabiá-7B support the interpretation that specialized fine-tuning plays a central role in this scenario. Although the model was natively trained for Brazilian Portuguese, its performance remained below that of the Carapicu variants, indicating that linguistic specialization alone does not guarantee the conceptual alignment required by a technical domain. In addition, qualitative analysis of the Sabiá-7B outputs revealed severe textual degeneration issues. In many interactions, the model produced anomalous outputs characterized by repetitive loops, semantically incoherent sentences, and random character generation.

The *LLM-as-a-Judge* evaluation complements the BERTScore results by revealing qualitative aspects that are not fully captured by traditional metrics. The Carapicu models achieved stronger performance mainly in the relevance dimension, indicating closer alignment between the generated answers and the proposed questions. This behavior is consistent with the adopted training procedure, since *Carapicu-textbook* was constructed directly from question-answer pairs derived from the textbook content.

With respect to completeness, the adapted models achieved lower results than those observed for relevance and factual correctness. One possible explanation lies in the characteristics of the *Carapicu-textbook dataset* used during training. Since this *dataset* was constructed from objective question–answer pairs, most of its answers are relatively short and focused on the core information. Consequently, the adapted models learned to reproduce this pattern during inference, favoring concise responses even when the *LARI dataset* contained more detailed reference answers. Therefore, the lower completeness scores appear to reflect differences between the writing style learned during adaptation and the response format expected by the evaluation benchmark, rather than indicating lower conceptual quality of the generated answers.

It is also important to note that the performance achieved by the RAG strategy in this evaluation may indicate a complementary relationship between retrieval-based methods and parametric adaptation in the studied task, suggesting a viable path for developing higher-quality systems with lower computational costs.

In general, our results indicate that *fine-tuning* with structured instructional knowledge is a promising strategy for adapting small language models to educational QA within the specific NLP domain investigated in this work. Since the experiments were conducted using a single textbook as the knowledge source, further studies are necessary to verify whether these findings generalize to other educational domains and instructional materials.

7. Conclusion

In this work, we investigated how different adaptation strategies influence small language models in educational QA tasks in Portuguese. Our findings indicate that the adapted models achieved stronger semantic alignment and higher answer relevance than the evaluated non-specialized base models. Within the investigated setting, Carapicu-Qwen3-0.6B-SFT and Carapicu-Qwen3-0.6B-CPT-SFT outperformed larger models across different evaluation metrics, suggesting that supervised adaptation with structured instructional knowledge can be an effective strategy for domain-specific educational QA.

The experiments also indicate that the characteristics of the training dataset directly influence the textual behavior of the adapted models, particularly regarding answer objectivity. These findings reinforce the importance of carefully designing instructional datasets for model specialization. More broadly, this work contributes to ongoing discussions on LLM adaptation in Portuguese, educational QA, and the use of structured instructional knowledge for specializing small language models.

Future work should investigate whether the observed behavior remains consistent when other textbooks, knowledge domains, educational resources, and languages are considered. We also plan a human evaluation of the questions in the Carapicu-textbook dataset, aiming to identify the types of questions that language models answer well or struggle with. Based on a taxonomy of question types, we intend to investigate how different categories affect model performance and use these findings to support the practical adoption of this resource as educational material within the BPLN textbook. We also plan to extend the dataset and experiments to the recently released fourth edition of the BPLN textbook.

Limitations

Some limitations of this work should be described. First, the experiments relied exclusively on the BPLN NLP textbook as the knowledge source, restricting the analysis to a specific educational domain within Natural Language Processing. Therefore, the findings should be interpreted within the investigated educational setting and do not necessarily generalize to other textbooks, educational resources, or knowledge domains.

Another limitation is related to the size and characteristics of the *Carapicu-textbook* dataset. Since the resource is composed predominantly of short and objective answers, the adapted models tended to reproduce this pattern during inference, particularly affecting the completeness results.

In addition, the evaluation based on the *LLM-as-a-Judge* paradigm carries limitations inherent to the use of language models as automatic evaluators, potentially reflecting preferences associated with writing style or answer length [Chen et al. 2024]. Furthermore, the judge model employed in this study belongs to the *Sabiá* family, while one of the evaluated models also belongs to the same family. This setup may introduce *self-recognition bias*, a phenomenon in which models tend to favor outputs generated by themselves or by closely related model families [Panickssery et al. 2024]. Although the *Sabiá*-based judge did not explicitly demonstrate systematic preference toward answers generated by the *Sabiá-7B* model, it is possible that some degree of this bias may still have influenced the evaluation results. In future work, we will complement this approach using human evaluation.

Finally, the experiments were conducted only with models from the Qwen3 family and with Sabiá-7B. Future work may extend this analysis by considering other architectures, adaptation strategies, and additional educational datasets in Portuguese.

Ethical Considerations Statement

This work relies exclusively on public academic textual resources for the construction of the *datasets* and the execution of the experiments. The content used for model adaptation was derived from the BPLN NLP textbook, licensed under the Creative Commons BY-NC-ND 3.0 BR license and used in this work with permission from the editors, while the LARI *dataset* was employed solely for evaluation purposes, maintaining an explicit separation between training and testing data. The experiments did not involve the collection of personal or sensitive data. We acknowledge, however, that language models may produce incorrect or biased answers. For this reason, the presented results should be interpreted as experimental analyses within a specific educational setting rather than as a replacement for instructional materials or specialized human evaluation.

Statement on the Use of Generative AI

Generative Artificial Intelligence tools were used as support during the development of this work, including assistance with grammatical revision and translation from Portuguese into English. All methodological decisions, experimental analyses, result interpretations, and final content validation, however, were carried out by the authors. Language models were also employed as part of the proposed experimental methodology, specifically during the evaluation stage based on the *LLM-as-a-Judge* paradigm.

Acknowledgments

The authors acknowledge the support provided by the undergraduate research scholarship from UFBA (2025/2026 PIBIC program), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 (process no. 88887.970806/2024-00), CNPq (grant 307088/2023-5), FAPERJ (processes SEI-260003/002930/2024, SEI-260003/000614/2023) and the Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (grant 408490/2024-1).

References

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint arXiv:2310.11511*.

Assis, G., Freitas, C., and Paes, A. (2025). Exploring brazil's llm fauna: Investigating the generative performance of large language models in portuguese. *Journal of the Brazilian Computer Society*, 31(1):939–971.

Caseli, H. M. and Nunes, M. G. V., editors (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 3 edition.

Caseli, H. M. and Nunes, M. G. V., editors (2026). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 4 edition.

Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. (2024). Humans or LLMs as the judge? A study on judgement bias. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Costa, Leandro e Souza-Filho, J. B. d. O. e. (2024). Adaptação de llms a novos domínios: um estudo comparativo de estratégias de ajuste fino e rag para tarefas de qa em português. In Claro, Daniela Barreiro e Pagano, A., editor, *Anais do 15º Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, Belém do Pará, Brasil. Associação para Linguística Computacional.

Falkner, S., Klein, A., and Hutter, F. (2018). Bohb: Robust and efficient hyperparameter optimization at scale. In *International conference on machine learning*, pages 1437–1446. PMLR.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., et al. (2023). Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Junqueira, J. d. R., Freitas, L. A. d., and Corrêa, U. B. (2026). LARI dataset: A native Portuguese question answering dataset from brasileiras em PLN. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1, pages 1055–1061, Salvador, Brazil. Association for Computational Linguistics.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, S., Hilton, J., and Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. (2023). When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. *arXiv preprint arXiv:2212.10511*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Panickssery, A., Bowman, S. R., and Feng, S. (2024). Llm evaluators recognize and favor their own generations. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802. Curran Associates, Inc.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Pires, R., Abonizio, H., Almeida, T., and Nogueira, R. (2023a). Sabiá: Portuguese large language models. In *Anais da XII Brazilian Conference on Intelligent Systems*, pages 226–240, Porto Alegre, RS, Brasil. SBC.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023b). Sabiá: Portuguese large language models. *arXiv preprint arXiv:2304.07880*.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Roberts, A., Raffel, C., and Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426. Association for Computational Linguistics.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., and Hupkes, D. (2024). Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

Xing, W., Nixon, N., Crossley, S., Denny, P., Lan, A., Stamper, J., and Yu, Z. (2025). The use of large language models in education. *International Journal of Artificial Intelligence in Education*, 35(2):439–443.

Xu, W., Zhu, G., Zhao, X., Pan, L., Li, L., and Wang, W. (2024). Pride and prejudice: Llm amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., and Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

Yang, A., Li, A., Yang, B., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.