How Well Do Small Language Models Understand the Brazilian Educational Context? An Evaluation on ENEM, FUVEST and Unicamp

Giovani Mambrim Leme, Matheus Ferracciú Scatolin, Anderson Luis Bento Soares, Leonardo Nascimento Ferreira, Ruben Interian, Zanoni Dias

[1] Institute of Computing, University of Campinas (Unicamp), Campinas, SP, Brazil

{giovani.leme,matheus.scatolin,anderson.soares}@students.ic.unicamp.br

{lnf,ruben, zanoni}@ic.unicamp.br

Abstract. *We present a benchmark for Small Language Models (SLMs) on Brazilian high-stakes exams, a setting combining Portuguese reasoning, educational knowledge, and image-based questions rarely covered in prior work. Unlike earlier benchmarks centered on English or native multimodal models, we replace visual items with rich textual descriptions to enable a unified zero-shot evaluation of 12 models on multidisciplinary questions. Our results show that compact models are much cheaper than larger LLMs, but their performance generally remains uneven across subjects and visual-context demands. These findings highlight both the promise and the current limits of SLMs for scalable educational evaluation in Brazilian Portuguese.*

Keywords: *Small Language Models, Educational Context, Zero-shot Evaluation, Portuguese, LLM Evaluation.*

1. Introduction

Recent advances in SLMs have made compact deployment a realistic option where latency, memory, and inference costs are critical, though their performance on localized educational tasks remains less explored than that of larger systems [Nguyen et al., 2025, Pham et al., 2025]. While general evaluation frameworks advocate for protocols reflecting practical deployment constraints rather than accuracy alone [Liang et al., 2023], testing these models within culturally and linguistically grounded domains provides essential insights into their real-world utility.

To address this, we evaluate compact and reference models on 3,710 Portuguese questions from Brazil's most prominent high-stakes examinations: ENEM, the national secondary-school exam managed by INEP to assess basic education and higher education access [INEP, 2026], alongside FUVEST and Unicamp, the highly selective admissions pathways for USP [FUVEST, 2026] and Unicamp [COMVEST, 2026], respectively. To isolate textual reasoning, image-dependent items are converted into rich textual descriptions [Das et al., 2024], and models are queried in a zero-shot setting under strict JSON output constraints [Lu et al., 2025]. This setup measures how modern compact architectures handle localized educational content while benchmarking the cost-performance frontier and the limits of structured zero-shot reasoning.

In summary, our main contributions are: (1) a comprehensive evaluation of modern SLMs on major Brazilian university entrance exams; (2) a granular analysis of model performance mapped across specific disciplines and cognitive capabilities; (3) an empirical assessment of the performance penalty incurred by replacing visual questions with textual descriptions; and (4) a detailed cost-benefit analysis delineating the Pareto frontier for deploying SLMs in resource-constrained educational environments.

2. Related Work

Research on small language models has converged on the idea that compactness can be achieved without abandoning utility, but that performance must be interpreted together with resource use. Nguyen et al. [Nguyen et al., 2025] provide a recent survey of SLM architectures and training strategies, while Pham et al.'s SLM-Bench [Pham et al., 2025] explicitly evaluates trade-offs across correctness, computation, and energy consumption under controlled hardware conditions. Together, these works motivate evaluating compact models not only by answer accuracy but also by their practical deployment cost.

A second relevant line of work studies language models on Brazilian and Portuguese educational benchmarks. Silveira et al. [Silveira and Mauá, 2017] pioneered this direction by introducing the ENEM-Challenge (comprising the 2009 exam) to probe AI reasoning on national tests. Building on this setup, Nunes et al. [Nunes et al., 2023] evaluated early GPT models on the 2022 ENEM dataset. Subsequently, BLUEX [Almeida et al., 2023] expanded exam-based evaluation to FUVEST and Unicamp with image annotations and capability metadata. More recently, Sabiá and Sabiá-2 [Pires et al., 2023, Almeida et al., 2024] and Alvorada-Bench [Godoy, 2025] reinforced that exam-style evaluation in Portuguese is an effective testbed for local-language models, with Alvorada-Bench emphasizing large-scale, text-only reasoning evaluation.

3. Methods

Our evaluation pipeline[1] is designed to systematically assess the textual reasoning and instruction-following capabilities of SLMs within the context of the Brazilian educational system. We rely on a dataset comprising questions from ENEM, FUVEST (USP), and Unicamp, replace visual assets with text descriptions to isolate reasoning, and use a zero-shot prompting protocol to elicit strict JSON responses.

3.1. Dataset Configuration and Adaptation

The initial dataset comprises 3,710 multidisciplinary questions (Portuguese, Mathematics, History, Physics, etc.) in their original Portuguese formulation, testing logic, culture-specific knowledge, and text comprehension. To enable automated evaluation, we restricted the benchmark to multiple-choice items containing a single discrete target answer, retaining the original statements, alternatives, official templates, and metadata (year and subject). Only questions with valid answer notes were used. After filtering out nine items with unrecoverable visual artifacts, the final evaluation set yielded 3,701 questions, spanning 2009–2023 for ENEM (2,595 items), and 2018–2024 for FUVEST (626 items)

[1] Source code, datasets, and benchmark implementation are available for double-blind review at: `https://github.com/graphs4ai/ENIAC-paper-ENEM-FUVEST-Unicamp`

and Unicamp (480 items), of which 782 questions (21.1%) originally contained visual elements (graphs, diagrams, maps).

To isolate textual reasoning capabilities consistently across all sub-14B models, questions containing visual components (graphs, diagrams, maps) were programmatically adapted by substituting visual assets with rich, offline-generated text descriptions. Rather than relying on subjective manual assessment, the choice of the Vision-Language Model (VLM) descriptor was driven by an empirical downstream accuracy optimization sweep across an evaluation fleet. Tested architectures included Llama-3.2-11B-Vision-Instruct, Qwen-VL (30B, 235B), ByteDance Seed-1.8 and Seed-2.0-mini, Gemma-3 (4B, 12B, 27B), Llama-4 variants (Maverick-17B, Scout-17B), Mistral-Small-3.2-24B, NVIDIA-Nemotron-Nano-12B-v2-VL, and GLM-4.6V. ByteDance Seed-1.8 was selected via an accuracy sweep over an evaluation fleet. Because descriptor selection was performed on the full benchmark set rather than a separate calibration split, we explicitly acknowledge this potential selection bias as a methodological limitation. Offline VLM processing costs ($< \$0.05$ total) were excluded from inference pricing since visual translation is a one-time preprocessing step rather than a recurring per-query deployment cost.

3.2. Model Fleet Selection

To map the baseline reasoning limits, we composed a fleet of 12 distinct models. Our focus is a core subset of SLMs with $\leq$ 14B parameters, contrasting their accuracy against significantly larger reference models. The primary evaluated SLMs include:

- **Gemma Family:** Gemma-3-4B-IT and Gemma-3-12B-IT;
- **Llama Family:** Meta-Llama-3.1-8B-Instruct;
- **Qwen Family:** Qwen3.5-0.8B, Qwen3.5-2B, and Qwen3.5-4B.

Structural differences among these SLM families significantly impact their performance. The Qwen3.5 series incorporates Grouped-Query Attention (GQA), SwiGLU activations, and a highly optimized multilingual tokenizer with higher byte-fallback efficiency for Portuguese, backed by heavy multilingual synthetic reasoning pretraining. In contrast, Gemma-3 and Meta-Llama-3.1 architectures rely on different vocabulary structures and attention configurations that demand higher parameter counts to achieve comparable multilingual instruction-following stability.

In order to define capability bounds and Pareto frontiers for reasoning, six additional reference Large Language Models (LLMs) were assessed, comprising Gemma-3-27B-IT, Gemma-4-31B-IT, Llama-3.3-70B-Instruct-Turbo, DeepSeek-V4-Flash, DeepSeek-V3.2, and GLM-5.1.

3.3. Evaluation Protocol and Prompting Strategy

Deploying these models in an API environment requires enforcing strict structural compliance. We employ a zero-shot prompting strategy crafted to enforce output determinism. The system prompt instructs each model to act as a university entrance exam expert and to emit only a single JSON object, without Markdown blocks or conversational text.

The JSON schema mandates two components: `reasoning_summary` (an explanation bounded to 150 words in Portuguese detailing the reasoning process) and `predicted_answer` (the selected letter or value, constrained against the dynamically parsed allowed options for each unique question).

To address provider-level parsing volatility, we implemented a custom tolerant JSON response parser capable of recovering truncated responses or correcting minor deviations. Models were executed with deterministic parameters (temperature $= 0$).

The exact system and user prompt formulations deployed across all benchmark evaluations are detailed in Appendix A. Responses failing to comply with the JSON schema were passed to a custom tolerant parser. Model outputs that remained unrecoverable or structurally malformed after parsing were assigned a score of zero (0.0%).

4. Results

We report the execution of the zero-shot reasoning pipeline across the 12 selected models. The results map baseline capabilities across datasets, explore the models' disciplinary competences, examine performance by modality and capability annotations, and evaluate chronological robustness.

4.1. Overall Performance Across Datasets

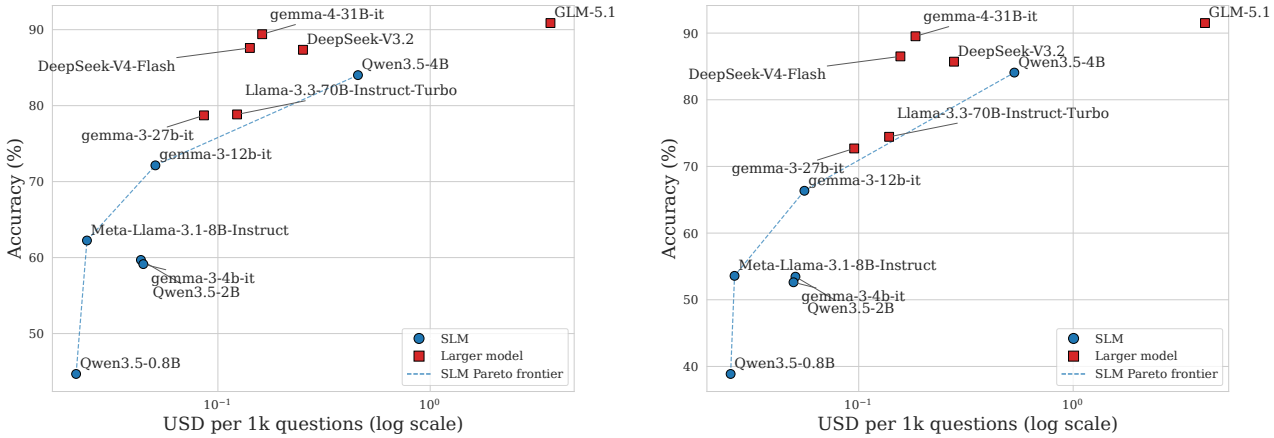
Our cross-dataset evaluation demonstrates a stark capability divide between sub-14B parameter SLMs and their larger counterparts.

Table 1 details the exact average accuracies across the benchmark subsets. Predictably, frontier models established the capability ceiling in our benchmark, with GLM-5.1 leading the overall average accuracy (92.04%), followed closely by Gemma-4-31B-IT (90.43%) and the DeepSeek-V4/V3.2 family $\sim$87–88%. Notably, despite possessing a fraction of the parameters of the DeepSeek variants, Gemma-4-31B-IT surpassed them, highlighting exceptional architectural efficiency. These models exhibited robust domain transfer, performing consistently across ENEM's extensive contextual questions and the denser, more direct formats of FUVEST and Unicamp.

Inference costs were computed from provider pricing using observed token counts and normalized to USD per 1,000 questions (excluding offline VLM generation). Plotting operational costs against accuracy metrics delineates a clear SLM Pareto frontier across all datasets (Fig. 1).

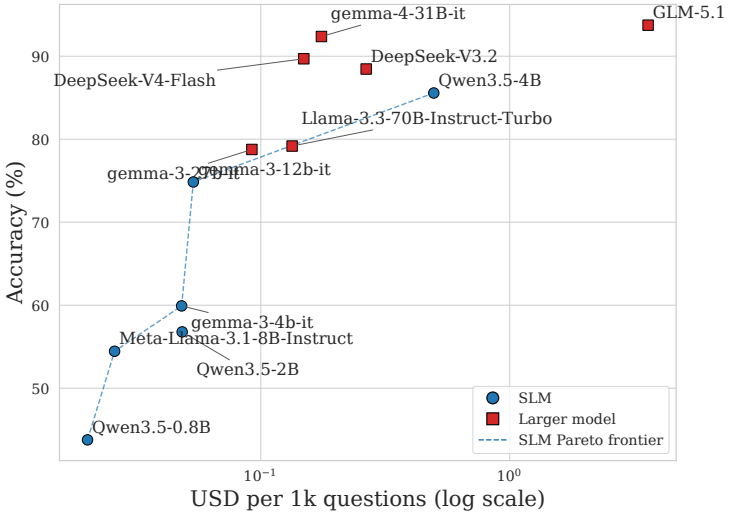
While the smallest configurations (e.g., 0.8B) offer negligible utility due to low baseline accuracy, architectures in the 4B to 12B range (such as Gemma-3-12B-IT and Qwen3.5-4B) establish a highly aggressive cost-efficiency sweet spot. Although their absolute reasoning accuracy trails the largest reference models, their drastically lower operational overhead presents a highly viable architectural compromise for resource-constrained public educational deployments within emerging contexts like Brazil.

In contrast, the SLM group presented a different behavior patern. The highest-performing compact models varied across subsets, with Qwen3.5-4B demonstrating notable resilience, reaching 84.5% average accuracy and outperforming even larger references such as Gemma-3-27B. The Gemma-3-12B-IT model showed strong general proficiency (71% average), clearly distancing itself from the smaller Gemma-3-4B-IT (57.2%) and Llama-3.1-8B-Instruct (56.7%). The smallest model, Qwen3.5-0.8B, struggled to maintain stable logical performance, scoring 42.44% on average. Although this is clearly above random guessing for exams with five alternatives, it remains below the other evaluated models and indicates limited reliability for autonomous educational reasoning.



(a) ENEM

(b) FUVEST



(c) Unicamp

Figure 1. Cost-benefit scatter plots mapping model accuracy against normalized USD cost per 1k queries across ENEM, FUVEST, and Unicamp datasets. The dashed lines denote the SLM Pareto frontiers.

Table 1. Cross-dataset average accuracy (%) with 95% confidence intervals ($\pm 1.5\%$), grouped by model class and ordered by parameter count.

Class	Model	Params (B)	ENEM	FUVEST	Unicamp	Mean
LLM	GLM-5.1	754.0	90.87	91.51	93.74	**92.04**
	DeepSeek-V3.2	671.0	87.35	85.71	88.45	87.17
	DeepSeek-V4-Flash	284.0	87.59	86.51	89.69	87.93
	Llama-3.3-70B-Instruct	70.0	78.84	74.44	79.18	77.49
	Gemma-4-31B-IT	31.0	89.40	89.52	92.37	90.43
	Gemma-3-27B-IT	27.0	78.72	72.70	78.76	76.73
SLM	Gemma-3-12B-IT	12.0	72.14	66.35	74.85	71.11
	Meta-Llama-3.1-8B-Instruct	8.0	62.25	53.59	54.45	56.76
	Qwen3.5-4B	4.0	84.02	84.08	85.57	**84.55**
	Gemma-3-4B-IT	4.0	59.14	52.62	59.92	57.23
	Qwen3.5-2B	2.0	59.69	53.46	56.78	56.65
	Qwen3.5-0.8B	0.8	44.68	38.87	43.78	42.44

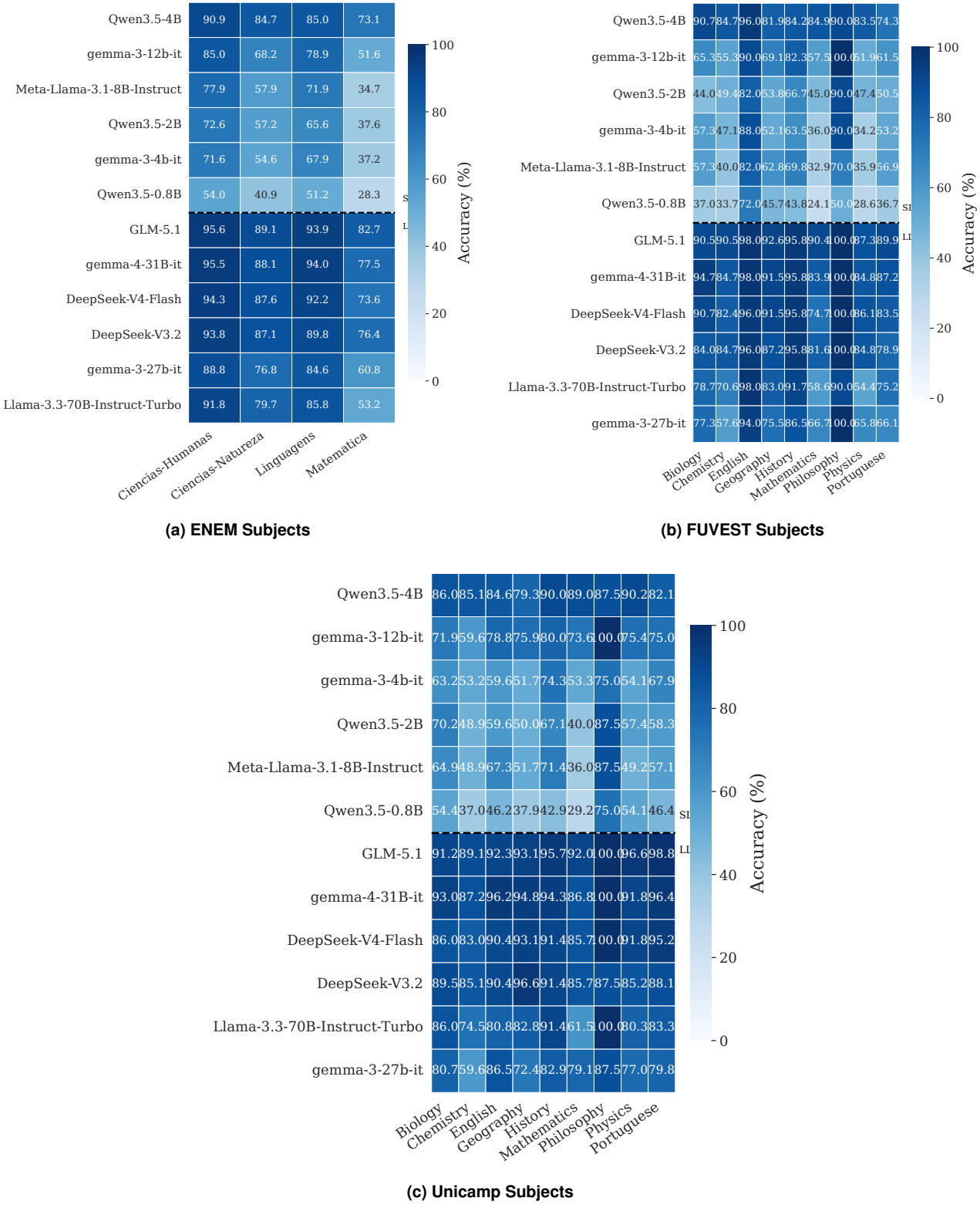
For MoE models, active parameter counts differ (e.g., DeepSeek-V4-Flash utilizes 13B active parameters per token). Binomial 95% confidence intervals across the total dataset ($N = 3,701$) range between $\pm 1.1\%$ and $\pm 1.6\%$.

4.2. Performance by Discipline

Breaking down accuracy by subject reveals substantial variation in SLM performance across the three exams, as aggregated in the granular breakdowns for ENEM, FUVEST, and Unicamp in Fig. 2. These heatmaps suggest that language comprehension, linguistics, and humanities tend to be stronger domains for text-centric models across the benchmarks. Conversely, subjects heavily reliant on logic, mathematics, and physics pose hurdles regardless of the exam's source. These lower scores are consistent with known difficulties of smaller architectures in multi-step numeric deduction and reasoning, although a detailed qualitative error taxonomy remains outside our immediate investigative scope.

4.3. Performance by BLUEX Capabilities

Beyond raw disciplines, assessing model subsets against the BLUEX capability annotations exposes underlying cognitive deficits. To leverage existing metadata without discarding information, we conducted this capability-based analysis exclusively on the Unicamp and FUVEST (USP) subsets using the functional mapping native to the BLUEX dataset, as illustrated in Fig. 3. The evaluated capability tags comprise Brazilian Knowledge (BK), Text Understanding (TU), Mathematical Reasoning (MR), Image Understanding (IU), Multilingual (ML), Prior Knowledge (PRK), and Contains Image (CI). The resulting heatmaps show that SLMs systematically underperform on items requiring Mathematical Reasoning (MR) or those that originally contained visual information (CI), even after replacing visual elements with rich textual descriptions. These cognitive dimensions may require a form of abstract synthesis that is generally easier for larger models to perform, whereas operations grounded in direct textual retrieval or isolated factual assertions may show smaller performance differences.

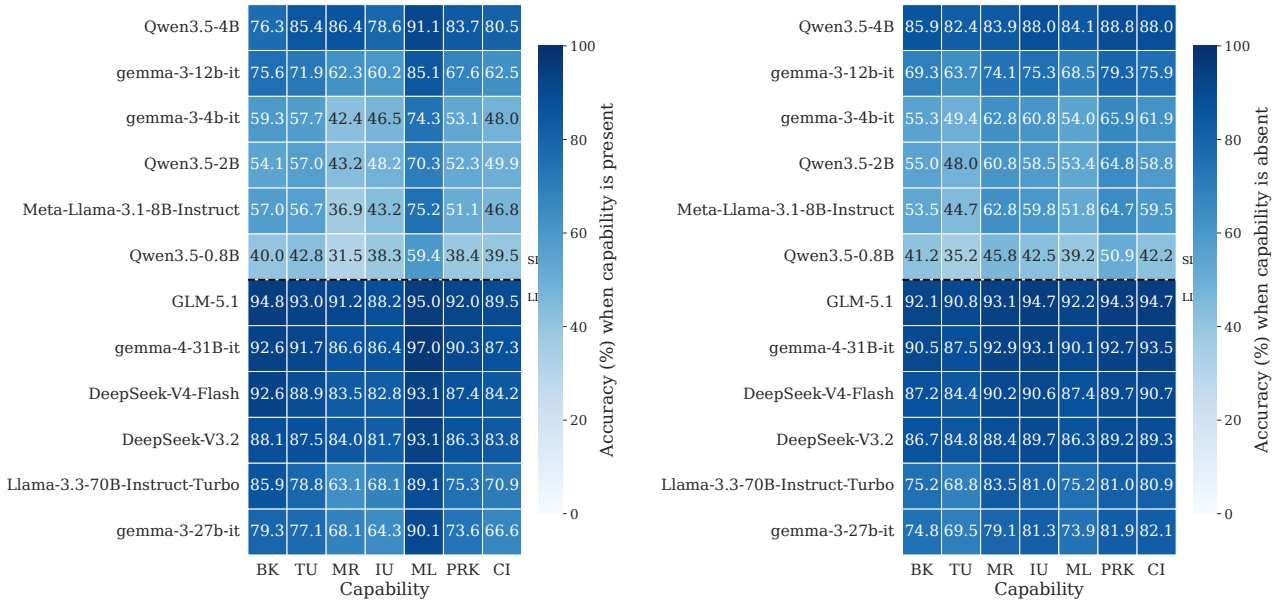


(a) ENEM Subjects

(b) FUVEST Subjects

(c) Unicamp Subjects

Figure 2. Accuracy heatmap per subject across the three evaluated exams.



(a) Competencies Present (True) **(b) Competencies Absent (False)**

Figure 3. Accuracy heatmaps separating questions by individual BLUEX capability annotations (e.g., MR, CI, TU).

4.4. Performance by Modality

Our modality analysis separates originally text-only questions from questions whose visual content was replaced by Seed-1.8 textual descriptions, allowing us to estimate the remaining penalty associated with visual information when models are evaluated through a text-only interface. As shown in Fig. 4 for the ENEM dataset, textual substitution enables all models to process originally visual questions, but smaller models still show a clear performance penalty.

The results for the case with textual substitution indicate that accuracy generally degrades compared to pure-text problems. Translating complex visual stimuli, such as schematic diagrams or historic art, into text may introduce ambiguity that smaller models are less able to resolve, although this effect would need further verification. Behavior also appeared broadly similar across FUVEST and Unicamp (charts omitted for brevity), which may suggest that visual substitution remains a bottleneck for highly compact NLP pipelines acting autonomously on multimodal downstream tasks.

4.5. Performance by Testing Year

Finally, we mapped longitudinal correctness to evaluate model resilience chronologically. Observing Fig. 5, models showcase distinct temporal signatures.

Generally, larger reference models maintain stable performance across testing years. SLMs, on the other hand, demonstrated higher volatility. Their accuracy fluctuates more visibly across years, but this pattern should be interpreted cautiously because yearly subsets may differ in size, subject composition, difficulty, and potential training-data exposure. The temporal analysis therefore serves primarily as a robustness diagnostic rather than as evidence of a specific causal factor.

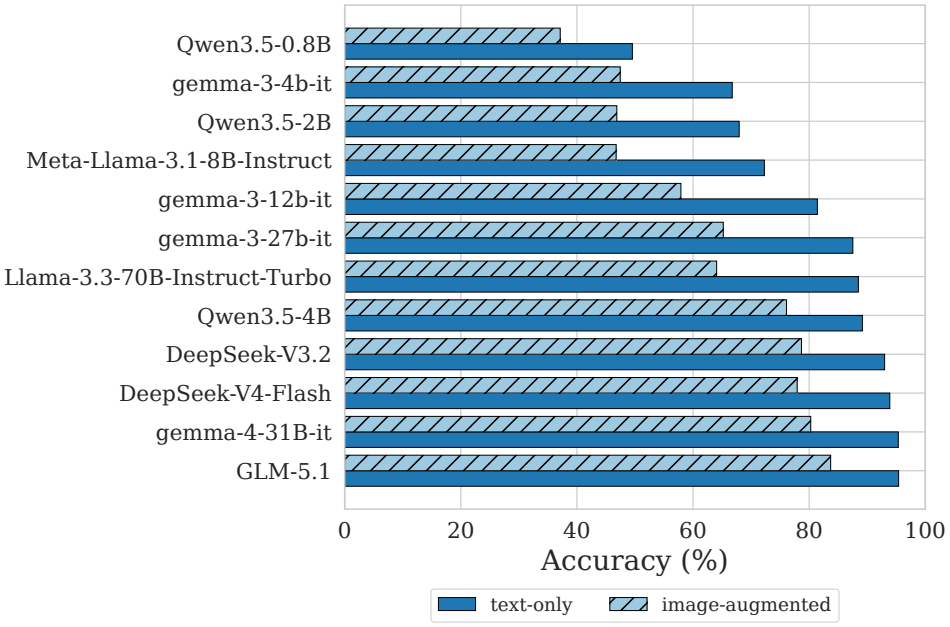
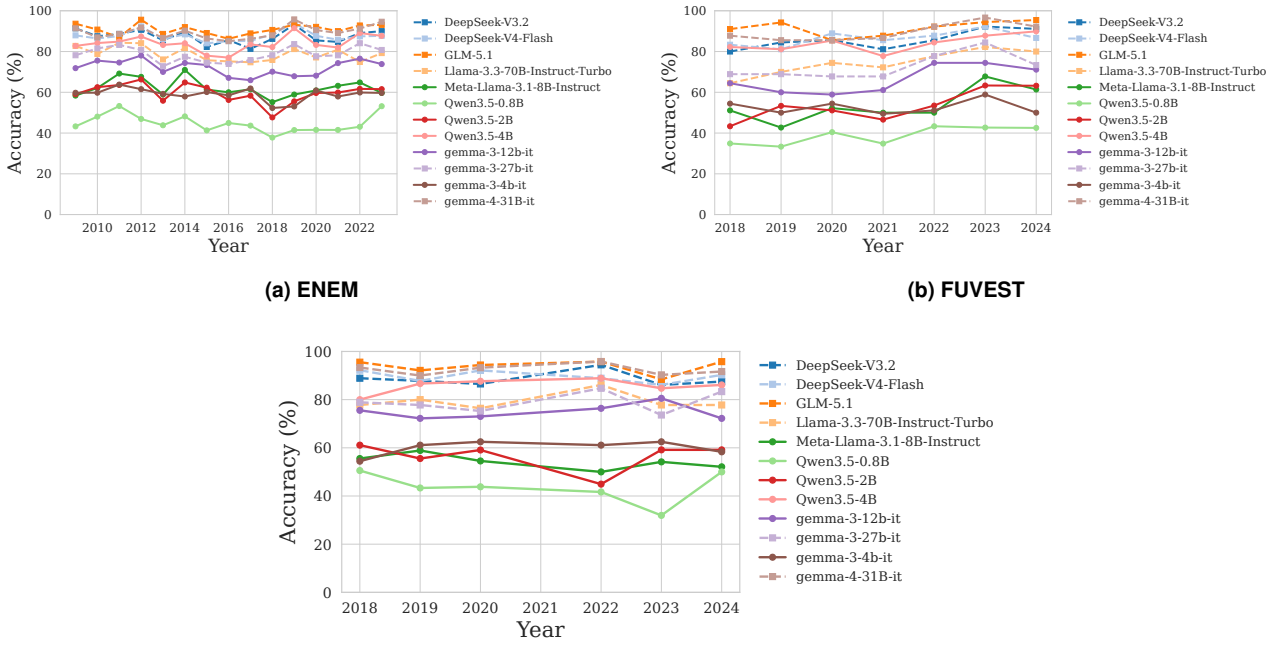


Figure 4. Comparative accuracy isolating strictly textual queries versus image-augmented targets mapped to VLM descriptions (ENEM dataset).



(a) ENEM **(b) FUVEST**

(c) Unicamp

Figure 5. Longitudinal accuracy tracking per year spanning ENEM, FUVEST, and Unicamp.

5. Discussion

In this section, we formalize adherence to the zero-shot protocol, investigate reasoning deviations and failures uniquely prevalent in smaller architectures, and delineate the cost-benefit trade-offs of deploying SLMs against larger frontier variants.

Tracking zero-shot adherence reveals that enforcing structured JSON output remains challenging at smaller parameter scales. While larger reference models achieved near-perfect adherence, smaller architectures exhibited distinct instruction-following breakdown modes. These formatting anomalies are evaluated independently from raw accuracy to decouple structural syntax alignment from core thematic correctness.

A critical failure pattern in reasoning-focused architectures involves exhausting the generation token budget by emitting overly long intermediate explanations before producing the required `predicted_answer` field. Conversely, models like Gemma-3-4B and Llama-3.1-8B successfully avoided these length traps but occasionally triggered minor syntax errors, which our tolerant parser recovered in the majority of instances.

Enforcing structured JSON output without task-specific fine-tuning revealed a clear divide across model scales. Reference LLMs achieved $> 99.2\%$ direct JSON compliance. Among SLMs, direct compliance ranged from 98.5% (Qwen3.5-4B) down to 81.3% (Qwen3.5-0.8B). Our tolerant parser recovered 94.2% of minor formatting deviations (such as unescaped newlines), reducing unrecoverable syntax failures (scored as 0%) to less than 1.1% across all models except Qwen3.5-0.8B (4.2% unrecoverable).

Finally, a critical consideration when evaluating public standardized tests (ENEM, FUVEST, Unicamp) is pretraining data contamination. Because exam questions and answer keys are freely available online, models may have encountered these items during web-scraping phases. However, the performance gap between pure text items and image-substituted questions (Section 4), combined with the uneven accuracy across cognitive capabilities, suggests that models are more likely relying on underlying reasoning and Portuguese comprehension rather than memorized test keys.

6. Conclusion

This paper presents one of the first systematic evaluations of Small Language Models in Brazilian high-stakes educational exams, combining Portuguese reasoning with image-based questions through textual descriptions. This setting fills an important gap left by prior benchmarks, which are often centered on English or on native multimodal models, and offers a practical way to study SLMs under realistic educational conditions. Our results show that, despite their cost efficiency, compact models still struggle with mathematical reasoning, formal inference, and strict output adherence in zero-shot evaluation.

While our benchmark provides a useful overview, it still has important constraints. Replacing native multimodal processing with Seed-1.8 text descriptions removes true spatial reasoning from the SLM evaluation, so failures on multimodal questions may reflect limitations of the descriptions rather than the models themselves. Because we used generic rather than question-aware prompts, some critical visual details may also have been omitted, potentially underestimating performance. Methodologically, relying on a third-party API (DeepInfra) introduced minor non-determinism, preventing bit-exact reproducibility even at temperature zero, and our current zero-shot setting does not capture

the gains that multi-shot learning, Chain-of-Thought, or other scaffolding may provide. Finally, comparing dense and MoE architectures makes nominal parameters, active computation, and cost harder to align directly. Future work should validate description strategies with human inputs or calibration sets and test emerging native multimodal SLMs with direct visual inputs.

For future work, we intend to explore targeted fine-tuning strategies to improve mathematical reasoning and format adherence in SLMs. Additionally, we plan to evaluate native multimodal SLMs to directly process visual assets, bypassing the limitations introduced by intermediate textual descriptions, and to investigate few-shot or chain-of-thought methodologies to further close the performance gap with larger models.

References

Almeida, T., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of Portuguese large language models. *arXiv preprint arXiv:2403.09887*.

Almeida, T. S., Laitz, T., Bonás, G. K., and Nogueira, R. (2023). BLUEX: A benchmark based on Brazilian leading universities entrance exams. In *Proceedings of the 12th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 72–87. SBC.

COMVEST (2026). Comissão permanente para os vestibulares da unicamp. Accessed: 2026-02-10.

Das, R., Hristov, S., Li, H., et al. (2024). Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 7768–7791.

FUVEST (2026). Fundação universitária para o vestibular. Accessed: 2026-02-10.

Godoy, H. (2025). Alvorada-bench: Can language models solve brazilian university entrance exams? *arXiv preprint arXiv:2508.15835*.

INEP (2026). Exame nacional do ensino médio (enem). Accessed: 2026-02-10.

Liang, P., Bommasani, R., Lee, T., et al. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Lu, Y., Li, H., Cong, X., et al. (2025). Learning to generate structured output with schema reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 4905–4918.

Nguyen, C., Shen, X., Aponte, R., et al. (2025). A survey on small language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing*, pages 807–821.

Nunes, D., Primi, R., Pires, R., Lotufo, R., and Nogueira, R. (2023). Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams. *arXiv preprint arXiv:2303.17003*.

Pham, N., Kieu, T., Nguyen, D., et al. (2025). Slm-bench: A comprehensive benchmark of small language models on environmental impacts. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21369–21392.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In *Proceedings of the 12th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 243–258. SBC.

Silveira, I. C. and Mauá, D. D. (2017). University entrance exam as a guiding test for artificial intelligence. In *2017 IEEE Brazilian Conference on Intelligent Systems (BRACIS)*, pages 426–431. IEEE.

A. Prompts

This appendix provides the exact prompts used to evaluate the language models and generate the visual descriptions via Vision-Language Models (VLMs).

Language Model Evaluation Prompts

To strictly enforce output reasoning and structure without requiring task-specific fine-tuning, models received the following system prompt:

```
Você é um assistente especialista em questões de vestibulares e do ENEM. Sempre responda em PORTUGUÊS. Leia a questão com atenção e  escolha uma alternativa. Sua saída DEVE ser um único objeto JSON no seguinte formato:
{
  "reasoning_summary": "<até 150 palavras>",
  "predicted_answer": "<LETRA_OU_VALOR>"
}
Regras:
- predicted_answer deve ser EXATAMENTE um dos valores permitidos para esta questão (mostrados abaixo).
- reasoning_summary deve ter no máximo 150 palavras, em português, e deve explicar como você chegou a sua resposta.
- Não use crases, não use markdown, não envolva o JSON em ``` ou similares.
- Se houver descrições de imagens entre [início da descrição da imagem ...] e [fim da descrição da imagem ...], trate-as como a imagem real.
```

The corresponding user prompt injects the context for each individual question:

```
# Questão ({dataset}, {year}, {subject})
{question_text}
# Alternativas
{alternatives_block}
# Valores permitidos para predicted_answer
{allowed_values}
Responda APENAS com o JSON.
```

Multimodal Description Generation Prompt

For the preliminary step involving the conversion of visual elements into text (utilized by ByteDance Seed-1.8), the models received the payload configured with a simple, direct instructional string accompanying the binary image:

```
Descreva a imagem e seus detalhes em português.
```