BLUEX Revisited: Enhancing Benchmark Coverage with Automatic Captioning

João Guilherme Alves Santos[1] 0000-0001-5307-5338 , **Giovana Kerche Bonás**[1,2] 0009-0001-9460-8353 , **Thales Sales Almeida**[1,2] 0009-0006-9568-9331

[1]Instituto de Computação (IC) – Universidade Estadual de Campinas (UNICAMP)
Campinas – SP – Brazil

[2]Maritaca AI
maritaca.ai

j199624@dac.unicamp.br, g216832@dac.unicamp.br, t224732@dac.unicamp.br

Abstract. *With the growing capabilities of Large Language Models (LLMs), there is an increasing need for robust evaluation methods, especially in multilingual and non-English contexts. We present an updated version of the BLUEX dataset, now including 2024-2025 exams and automatically generated image captions using state-of-the-art models, enhancing its relevance for data contamination studies in LLM pretraining. Captioning strategies increase accessibility to text-only models by more than 40%, producing 1,422 usable questions, more than doubling the number in the original BLUEX. We evaluated commercial and open-source LLMs and their ability to leverage visual context through captions.*

1. Introduction

Large Language Models (LLMs) have made remarkable progress in recent years, demonstrating impressive capabilities across a wide range of natural language processing tasks, such as code generation and assistance [Nam et al. 2024, Liu et al. 2023, Chen et al. 2021, Zhang et al. 2023], question answering [Singhal et al. 2025, Petroni et al. 2019, Lazaridou et al. 2022, Almeida et al. 2025], open-ended conversation [OpenAI et al. 2024, Grattafiori et al. 2024, Abonizio et al. 2024, Ouyang et al. 2022], and summarization [Zhang et al. 2024a, Zhang et al. 2024b, Chang et al. 2023]. As these models become increasingly capable of performing complex reasoning and generating coherent, context-aware responses, robust benchmarks play a central role in assessing their true capabilities. Beyond measuring surface-level accuracy, well-designed benchmarks can help uncover how models handle ambiguity, reason through multiple steps, and generalize across diverse topics and linguistic styles.

Considering real-world utility, one domain that naturally demands advanced understanding and complex reasoning is standardized education. High-stakes exams often require students to interpret complex textual information –sometimes alongside diagrams, graphs, illustrations, or images– and respond to questions that require factual knowledge and inferential thinking. These settings provide rich and authentic challenges where textual understanding and reasoning across domains directly impact task success. Because such assessments are carefully designed to evaluate specific cognitive skills and knowledge, they offer a grounded, purpose-driven context to evaluate the capabilities of large language models.

In this work, we introduce an expanded and updated version of BLUEX [Almeida et al. 2023], a benchmark comprising over 1,000 multiple-choice questions from the entrance exams of Brazil's top universities and the top 500 worldwide [Times Higher Education 2024, ShanghaiRanking Consultancy 2024], Unicamp and USP, administered between 2018 and 2023. Each question includes text, answer options, and associated images. Our main objective is to evaluate how large language models perform in multimodal educational tasks, and to investigate the impact of different image captioning strategies on their performance. To do this, we expand the original dataset both quantitatively and temporally, including two additional years (2024 and 2025) and generate image captions for all visual elements using GPT4o.

We generate image descriptions using GPT-4o under two distinct conditions: *Blind captions*, where captions are generated solely based on visual content without access to accompanying text; and *Context captions*, where captions are generated with access to the associated question and alternatives, allowing context-aware interpretations. This setup not only broadens the dataset with image-associated questions that were previously unanswerable by non-multimodal LLMs, but also provides a controlled experimental setting to analyze how context-aware image captioning affects LLM performance, under the hypothesis that contextualized captions –though often shorter– may lead to equal or even superior model accuracy by focusing on task-relevant visual elements.

Our findings show that many current models now reach scores high enough to surpass the admission thresholds of over 90% of undergraduate programs at Unicamp and USP, highlighting the rapid progress of LLMs in handling complex reasoning tasks. By releasing this updated dataset[Datasets 2025], along with the evaluation code[Almeida 2025], we aim to provide a more realistic and multimodal benchmark to evaluate LLMs in educational contexts, while offering empirical insights into how different captioning strategies influence their ability to interpret and reason about image-based content. Additionally, by converting image-based questions into caption-based ones, our benchmark extends its applicability to non-multimodal models, enabling broader participation in multimodal tasks and allowing researchers to isolate the effects of visual grounding through controlled captioning strategies.

2. Related work

2.1. LLM standard test evaluation

A growing body of work has evaluated large language models (LLMs) using standardized exams as proxies for human-level reasoning and real-world task performance. Such as example, MMLU (Massive Multitask Language Understanding) [Hendrycks et al. 2021] contains 57 tasks –that include STEM questions, humanities, social sciences and other fields– and represents a standard reference for measuring academic proficiency in LLMs, evaluating factual knowledge, reasoning and problem-solving abilities enabling evaluate models through both zero-shot and few-shot settings.

Moreover, AGIEval[Zhong et al. 2023] also presents a similar landscape, evaluating LLMs in real-world cognitive challenges through standardized exams, such as college entrance exams, math competitions, and lawyer qualification tests. Furthermore, GPQA[Rein et al. 2024] raises the bar for evaluating LLMs by introducing a highly challenging dataset of 448 multiple-choice questions authored by specialists from a range

of academic domains. These questions are designed to be extremely difficult –even for experts– and resist simple retrieval strategies, making GPQA a valuable benchmark for assessing deep reasoning and scalable oversight in advanced AI systems.

Delving into Portuguese-language evaluations, a few large-scale benchmarks exist to assess LLMs in Portuguese. But among them, BLUEX [Almeida et al. 2023] stands out as a comprehensive and challenging dataset focused on Brazilian university entrance exams. Comprising more than 1,000 multiple-choice questions from Brazil's Unicamp and USP entrance exams (2018–2023), it is designed to reflect the complexity of real-world educational assessments, including a significant portion (40%) of questions with visual components. This benchmark serves as the foundation for our present work, which extends BLUEX by adding two more years of exam data and incorporating a captioning pipeline, enabling a richer evaluation, particularly for models without native vision capabilities.

Other important efforts include benchmarks based on structured assessments traditionally used in Brazil, such as the Brazilian Bar Examination (OAB) [Delfino et al. 2017] and the National High School Exam (ENEM)[Silveira and Mauá 2017], both of which have been used to explore reasoning and comprehension of language models.

2.2. Multimodal Benchmarks

Recent advances in multimodal large language models (MLLMs) have highlighted the need for specialized benchmarks that can rigorously assess their visual and reasoning abilities. To meet this demand, recent studies have introduced diagnostic evaluations that go beyond raw performance, aiming to uncover specific strengths and weaknesses of models in complex multimodal settings. Benchmarks such as SEED-Bench [Li et al. 2023] and its successor studies[Li et al. 2024] have focused on systematically evaluating the multimodal reasoning capabilities of large language models. The benchmark consists of 24,000 multiple-choice questions covering 27 evaluation dimensions, including topics such as charts, visual mathematics, and free-form interleaved image-text reasoning. By restricting outputs to A/B/C/D choices and combining human annotation with automated filtering, it provides a scalable and objective evaluation framework, revealing that even top models achieve only around 60% accuracy and highlighting major challenges in complex visual and reasoning tasks.

Similarly, the Perception Test [Patraucean et al. 2023] extends multimodal evaluation into the video domain, proposing a diagnostic benchmark that measures fine-grained perceptual abilities and temporal reasoning across thousands of annotated videos. Instead of static images, models are challenged with dynamic visual scenes requiring continuous understanding, spatial awareness, and causal inference. In parallel, Importantly, MM-SafetyBench [Liu et al. 2024b] reveals an even deeper layer of vulnerability, showing that MLLMs can be easily compromised through query-relevant images, even when their underlying LLMs are safety-aligned. This demonstrates that multimodal safety demands specific, dedicated attention, beyond what is currently done for text-only models.

Building upon these insights, this expanded dataset provides a valuable resource for future studies aiming to assess linguistic models in Portuguese and in authentic educational contexts. In this way, our contribution supports the development of benchmarks that move beyond English-centric paradigms, enriching the tools available for cross-linguistic

and culturally grounded multimodal evaluation.

2.3. Caption generation by Multimodal Models

As a strategy to navigate towards vision and language modalities, recent works have explored replacing or augmenting images with textual descriptions. A notable example, *"Evaluating GPT-4's Vision Capabilities on Brazilian University Admission Exams"*[Pires et al. 2023], focuses on Brazil's national high school exam (ENEM), evaluating GPT-4 in recent ENEM editions, incorporating both textual and visual elements. A key finding is that text captions transcribing visual content often outperform the direct use of images. This highlights the potential of captioning as a powerful bridge between visual data and language models, suggesting that carefully crafted textual descriptions can unlock complex reasoning in models without requiring full visual processing. Building on this insight, our work integrates a captioning pipeline into the BLUEX, enabling a more targeted and scalable evaluation of LLMs on visually grounded academic tasks.

Some of these efforts can also be represented by projects such as Image Textualization (IT) [Pi et al. 2024], a framework that collaborates with multiple expert vision models with MLLMs to automatically generate high-quality and detailed image descriptions. In a similar vein, Bianco et al. [Bianco et al. 2023] propose a method to enhance the quality of image captions by integrating the output of various state-of-the-art models using LLMs, providing richer captions. In the same direction, TIFA [Hu et al. 2023] also explores that field by making an evaluation that demonstrates that high-quality captions can significantly impact model performance in visual tasks.

3. Methodology

This section describes the pipeline to create the benchmark, which is illustrated in Figure 1.

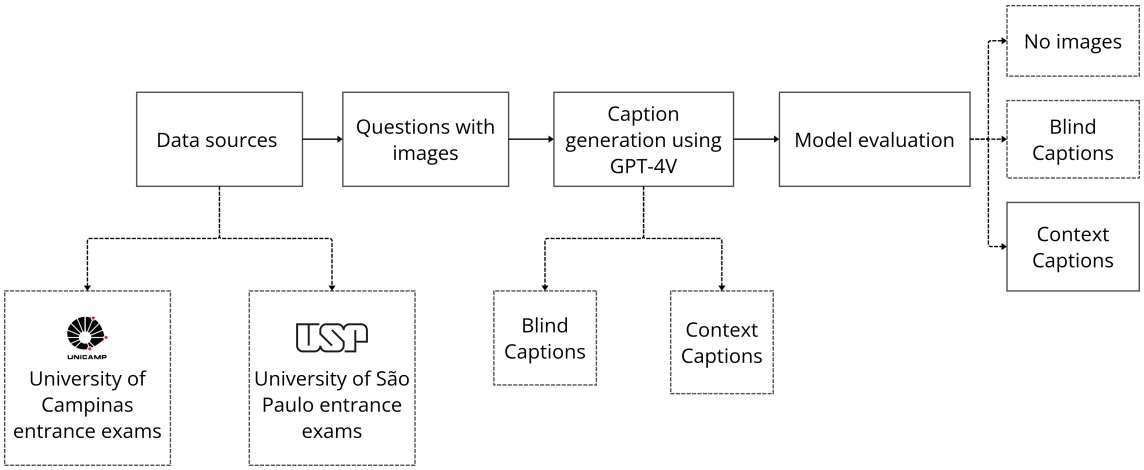


Figure 1. Overview of the benchmark construction pipeline.

The updated dataset is publicly available [Datasets 2025] –as well as the evaluation code [Almeida 2025]– and includes the original visual content, two types of generated captions, full question text, answer choices, correct answer, and metadata relevant to the nature of the assessment for each question. All data were manually curated and classified. This structure enables both rigorous evaluation and reproducible experimentation with multimodal and text-only models.

3.1. Dataset Selection

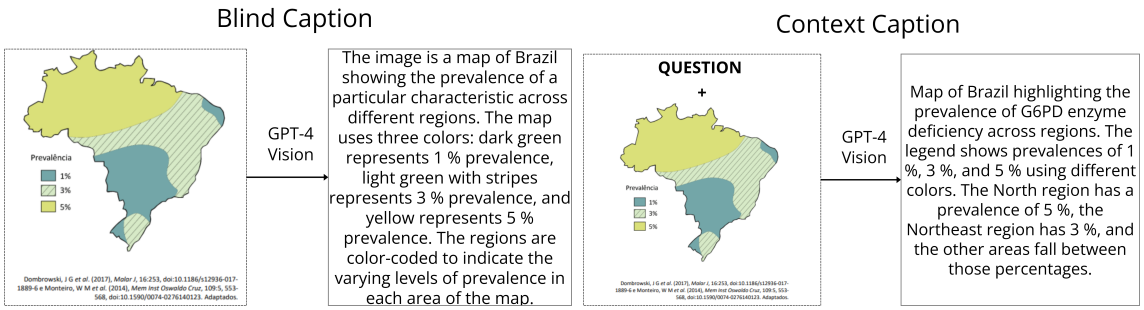
We use the entire Brazilian Leading Universities Entrance eXams (BLUEX) collection introduced by [Almeida et al. 2023], designed to address the scarcity of high-quality datasets in Portuguese by compiling entrance exam questions from Brazil's two most competitive universities. Additionally, we expanded the dataset to incorporate the most recent exams from 2024 and 2025 and provided new generated captions for visual content in Portuguese.

3.2. Caption Generation

Roughly 43% of the original BLUEX questions included images and were therefore inaccessible to text-only LLMs. To make these items evaluable, we generate Portuguese textual descriptions for every image with GPT-4o under two settings [OpenAI 2024]:

1. Blind Captions: GPT-4o generates captions based solely on the image, without any contextual information from the associated question.
2. Context Captions: GPT-4o is provided with both the image and its associated question, enabling it to generate context-aware descriptions.

Figure 2 illustrates a comparative example of the two captioning strategies, exposing that context captions such as shown in Figure 2b tended to be significantly shorter than blind captions in Figure 2a. This phenomenon likely arises because GPT-4o, when given access to the question, focuses on the visual elements most relevant to answering the task.



(a) Blind caption **(b) Context caption**

Figure 2. Comparative example of blind and context captioning. *Captions were generated originally in Portuguese; presented in English for convenience.*

3.3. Model Evaluation

To assess how effectively language models leverage the captions when answering exam questions, we conducted evaluations across three distinct experimental conditions. These conditions vary the amount and type of information provided to the models. Specifically, we evaluated models using the following configurations:

1. No images: The model receives only the textual content of the question, without any accompanying visual information (neither images nor captions). This baseline measures performance in purely textual scenarios.
2. Blind Captions: In place of actual images, the model is provided with the corresponding blind captions –descriptive captions generated without question context.

3. Context Captions: Here, the models receive context-aware captions –captions generated by GPT-4o using both the question and image as inputs.

Through these experiments, we aim to quantify how visual context and caption specificity affect model performance, thereby offering clearer insight into each model's effectiveness.

3.4. Caption and Image Statistics

Figure 3 presents the distribution of the caption lengths produced by GPT-4V under the two prompting strategies. As expected, context captions are markedly shorter than blind captions, because the question context guides the model to mention only the information most relevant for solving the problem.

Table 1 organizes all BLUEX questions into the four ENEM macro-areas: Natural Sciences (Biology, Chemistry, Physics), Human Sciences (History, Geography, Philosophy, Sociology), Languages (Portuguese and English), and Mathematics - and reports, for each year from 2018 to 2025, the number of items that do and do not include associated images. The proportion of image-based questions is consistently high across the entire period, underscoring the value of the captioning step. Once textual descriptions are available, every image question becomes accessible to text-only models. Per-area subtotals do not equal the grand total because a single question can be annotated with more than one subject, reflecting its inherently interdisciplinary character.

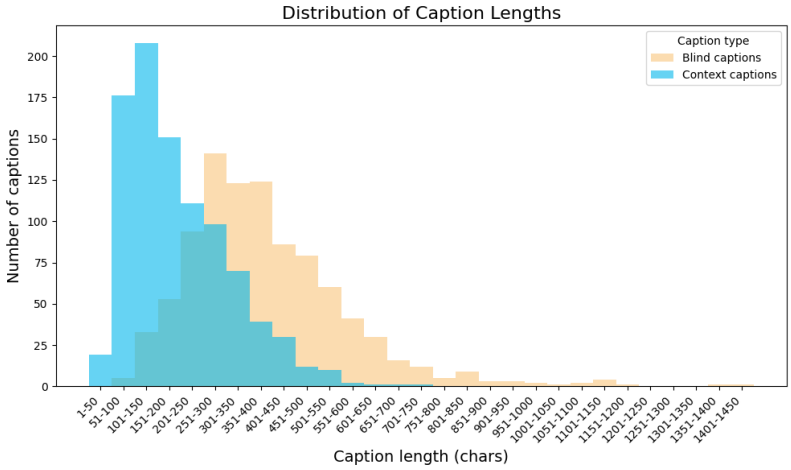


Figure 3. Distribution of caption lengths for blind and context captions.

3.5. Implementation Details

For each question, the input prompt consists of the full textual content (question statement and multiple-choice alternatives) –formatted as it originally appeared in the exam. When image captions are included in the evaluation, they are inserted into the prompt in the same position as the original image in the source exam layout. We use lm-evaluation-harness [Gao et al. 2024] to evaluate open-source models, using FP16 precision. Small models (up to 14B parameters) were executed on a machine with 2× A6000 GPUs. Llama 3.3 (70B) and Qwen 2.5 (72B) were executed on a machine with 2× A100 80GB GPUs. The Deepseek-v3 and Llama 4 models were executed via the Together AI API [1] due to their high parameter count.

[1] https://www.together.ai/

Table 1. Number of BLUEX questions by ENEM macro area and image distribution

	Subject Distribution				Image distribution		
Year	**Natural Sciences**	**Human Sciences**	**Languages**	**Mathematics**	**With Images**	**Without Images**	**Total**
2018	72	49	46	25	67	113	180
2019	68	53	50	31	90	90	180
2020	69	51	50	33	69	111	180
2021	82	64	64	40	85	149	234
2022	51	48	43	26	74	88	162
2023	59	50	41	24	75	87	162
2024	52	57	42	26	76	86	162
2025	53	62	54	25	74	88	162
Total	506	434	390	230	610	812	1422

4. Results

We evaluated several language models, separating them into two categories, small (every model with 14B parameters or less) and big models (70B and more parameters), as well as commercial models in sabiazinho-3, sabia-3 [Abonizio et al. 2024] from Maritaca AI, and GPT-4o and GPT-4o-mini [OpenAI 2024] from openAI. Our complete results are presented in Table 2, which reports the results for three partitions of the dataset: questions without images, questions with images, and considering all questions in the benchmark. For partitions that involve images, we test using both context-aware and blind captions, as well as providing no caption at all.

Table 2. Performance of all the tested models in the BLUEX benchmark

Model	Questions without images	Questions with images			All Questions		
	No Caption	**Context Captions**	**Blind Captions**	**No Caption**	**Context Captions**	**Blind Captions**	**No Caption**
Comercial models							
Sabia-3 [Abonizio et al. 2024]	**0.852**	0.701	0.695	0.616	**0.787**	**0.784**	0.750
GPT-4o [OpenAI 2024]	0.807	**0.718**	**0.729**	**0.683**	0.769	0.774	**0.754**
GPT-4o-mini [OpenAI 2024]	0.785	0.642	0.627	0.589	0.724	0.717	0.701
Sabiazinho-3 [Abonizio et al. 2024]	0.756	0.642	0.652	0.627	0.707	0.712	0.701
Large Open source models							
DeepSeek-V3 [Liu et al. 2024a]	**0.841**	**0.739**	**0.741**	**0.668**	**0.797**	**0.798**	**0.767**
Llama-4-Scout	0.758	0.644	0.665	0.589	0.709	0.718	0.685
Llama-4-Maverick	0.820	0.731	0.731	0.658	0.781	0.781	0.750
Llama-3.3-70B-Instruct [Dubey et al. 2024]	0.769	0.647	0.657	0.603	0.717	0.721	0.697
Qwen2.5-72B-Instruct [Yang et al. 2024]	0.796	0.695	0.693	0.650	0.752	0.752	0.733
Small open source models							
Qwen 2.5-14B [Yang et al. 2024]	**0.745**	**0.637**	**0.640**	**0.614**	**0.699**	**0.700**	**0.689**
Qwen 2.5-7B [Yang et al. 2024]	0.676	0.568	0.558	0.547	0.630	0.626	0.621
Qwen 2.5-3B [Yang et al. 2024]	0.587	0.530	0.507	0.483	0.563	0.553	0.542
Qwen 2.5-1.5B [Yang et al. 2024]	0.524	0.427	0.409	0.384	0.482	0.475	0.464
Falcon-10B [Team 2024]	0.669	0.575	0.560	0.527	0.628	0.622	0.608
Falcon-7B [Team 2024]	0.612	0.534	0.524	0.509	0.578	0.574	0.568
Falcon-3B [Team 2024]	0.455	0.422	0.392	0.389	0.441	0.428	0.427
Falcon-1B [Team 2024]	0.266	0.266	0.253	0.240	0.266	0.260	0.255
Llama-3.1-8B [Dubey et al. 2024]	0.595	0.473	0.466	0.448	0.542	0.539	0.532

Sabia-3 achieves the best performance among commercial models on questions without images, surpassing GPT-4o, the second-best commercial model, by 5 points.

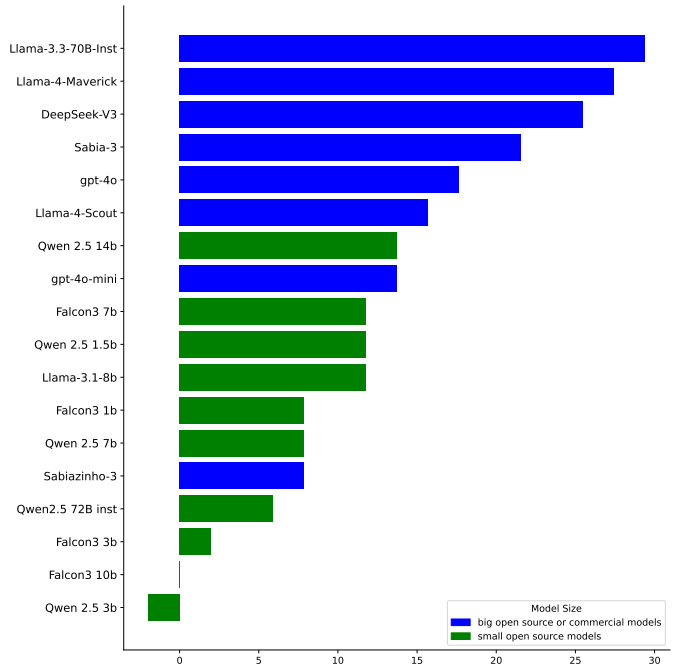


Figure 4. Accuracy gain in BLUEX questions that have images as alternatives. The graph shows the accuracy gain for each model by providing context captions for the images, compared to the blind performance.

However, this performance lead does not hold when looking at questions with images, where GPT-4o shows the best performance. Sabia-3 and GPT-4o perform similarly across all questions, with Sabia-3 leading by a small margin..

Within large open-source models, DeepSeek-V3 is the most robust model across all categories, presenting the highest accuracy in questions without images and performing consistently well across caption scenarios in questions with images. LLama-4-maverick ranks as the second-best performance. Note that both are mixture of experts(MOE) models with more than 400B total parameters. Qwen-2.5-72B [Yang et al. 2024] ranks as the third best model among large open source models, a very competitive performance for a model given it has less than 25% total parameters compared to the first two models.

Among small open-source models, Qwen 2.5-14B [Yang et al. 2024] stands out significantly, showcasing strong capabilities in all scenarios. This model notably surpasses other smaller models, such as Falcon and Llama-3. Generally, the Qwen 2.5 family of models performs better than their counterparts of the same size.

In general, we observe a slight variation between the performance of models with context or blind captions, even with the context captions being on average half the size, this indicates that the additional context provided when creating context captions did not necessarily make the caption more informative, but rather allowed it to be more concise and focused only on the relevant aspects of the image. Meanwhile, the blind description of the image used in the blind captions does seem to provide enough information to answer some questions, at the expense of longer captions.

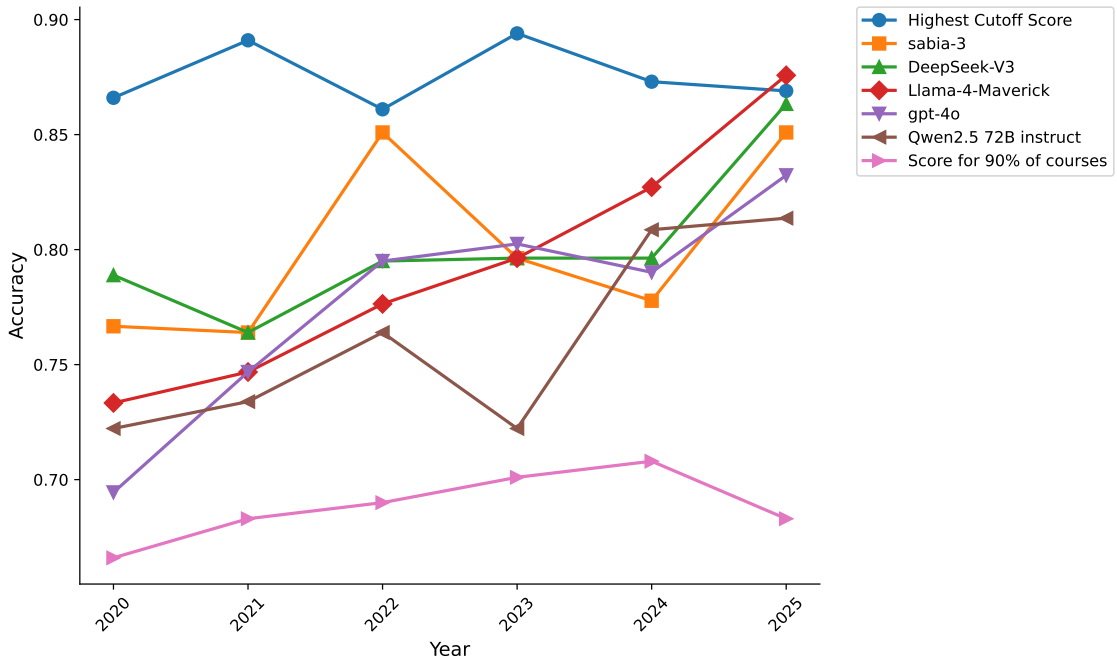


Figure 5. Performance of the top 5 tested models in each year, alongside the Highest Cutoff score and Score for passing in 90% of the offered courses.

4.1. How Effectively Models Use Captions

To better understand the impact of providing captions to models, we specifically analyzed questions whose alternatives consisted solely of images –questions that would be impossible to answer correctly without additional visual context. For these cases, we evaluated how much each model's performance improved when provided with context-aware captions.

Figure 4 illustrates these improvements. Upon receiving captions, most models exhibited a performance gain of at least 10 accuracy points. Additionally, we observed that larger models tended to benefit more significantly from captions, likely because interpreting and reasoning from captions require higher cognitive capabilities that scale with model size.

4.2. Comparing Models with Human Performance

All larger models evaluated already achieve scores sufficient to gain admission to approximately 90% of undergraduate courses at both USP and UNICAMP. Figure 5 illustrates the accuracy over the past six years for the top five models tested when provided with context-aware captions. The pink line represents the threshold score necessary to gain admission to 90% of courses. The models consistently meet or exceed this bar throughout the observed period, with a slight performance increase noted in recent years. Importantly, considering the exams' recency (particularly the 2025 tests), it is unlikely that most models had encountered these questions during training. This indicates their strong performance stems from genuine reasoning abilities rather than merely memorizing publicly available exam solutions.

However, despite their generally high performance, most models do not achieve

scores sufficient to enter the most competitive undergraduate program[2], represented by the 'Highest Cutoff Score' in the graph. An exception is LLaMA 4 Maverick, which attains a qualifying score in the 2025 exam, though not in previous years.

5. Conclusion

In this study, we expanded the BLUEX benchmark dataset by introducing image captions, enabling broader question accessibility for language models without multimodal capabilities. By exploring two captioning strategies–blind and context-aware captions –we found comparable model performance, despite observed differences in caption length. Context captions provided shorter yet more targeted descriptions compared to blind captions, which, although more detailed, contained less selectively relevant information.

Our evaluation across various commercial and open-source language models revealed several insights. Among commercial models, Sabia-3 excelled in text-only scenarios, whereas GPT-4o demonstrated superior performance when handling image-based questions. Within large open-source models, DeepSeek-V3 consistently showed strong results across all evaluated conditions, highlighting the effectiveness of models incorporating large parameter counts or mixture-of-expert architectures. Among smaller models, Qwen 2.5-14B displayed remarkable competitiveness, surpassing its peers significantly.

Importantly, providing captions, especially context-aware ones, notably improved model accuracy on questions that inherently required image interpretation. This performance enhancement was more pronounced in larger models, indicating a correlation between model size and the effective utilization of the textual descriptions.

Our work further develops the Portuguese benchmark landscape by effectively doubling the number of questions usable for non-multimodal LLMs compared to the original benchmark and demonstrates the current capabilities of state-of-the-art models.

Additionally, further research could expand the evaluation scope to include open-ended, dissertative questions, which are also part of the examined exams. This would broaden the assessment of LLM capabilities beyond the current multiple-choice format, providing a richer evaluation framework. Future work could also incorporate exams from other university entrance tests, further diversifying the benchmark.

References

Abonizio, H. et al. (2024). Sabi\'a-3 technical report. *arXiv preprint arXiv:2410.12049*.

Almeida, T. S. (2025). Revisited bluex benchmark - code repository. `https://github.com/ZanezZephyrs/bluex_eval`. Accessed: 2025-08-07.

Almeida, T. S. et al. (2025). Tiebe: Tracking language model recall of notable worldwide events through time. *arXiv preprint arXiv:2501.07482*.

Almeida, T. S. et al. (2023). Bluex: A benchmark based on brazilian leading universities entrance exams. In *Brazilian Conference on Intelligent Systems*, pages 337–347. Springer.

Bianco, S. et al. (2023). Improving image captioning descriptiveness by ranking and llm-based fusion. *arXiv preprint arXiv:2306.11593*.

[2]Typically, the most competitive undergraduate program is Medicine.

Chang, Y. et al. (2023). Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Chen, M. et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Datasets, P. B. (2025). Bluex: Brazilian undergraduate entrance exams benchmark. `https://huggingface.co/datasets/portuguese-benchmark-datasets/BLUEX`. Accessed: 2025-08-07.

Delfino, P. et al. (2017). Passing the brazilian oab exam: data preparation and some experiments. In *Legal knowledge and information systems*, pages 89–94. IOS Press.

Dubey, A. et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gao, L. et al. (2024). The language model evaluation harness.

Grattafiori, A. et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hendrycks, D. et al. (2021). Measuring massive multitask language understanding.

Hu, Y. et al. (2023). Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Lazaridou, A. et al. (2022). Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*.

Li, B. et al. (2024). Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*.

Li, B. et al. (2023). Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Liu, A. et al. (2024a). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, J. et al. (2023). Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36:21558–21572.

Liu, X. et al. (2024b). Mm-safetybench: A benchmark for safety evaluation of multimodal large language models.

Nam, D. et al. (2024). Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

OpenAI (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI et al. (2024). Gpt-4 technical report.

Ouyang, L. et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Patraucean, V. et al. (2023). Perception test: A diagnostic benchmark for multimodal video models. In Oh, A. et al., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 42748–42761. Curran Associates, Inc.

Petroni, F. et al. (2019). Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Pi, R. et al. (2024). Image textualization: An automatic framework for generating rich and detailed image descriptions. *Advances in Neural Information Processing Systems*, 37:108116–108139.

Pires, R. et al. (2023). Evaluating gpt-4's vision capabilities on brazilian university admission exams. *arXiv preprint arXiv:2311.14169*.

Rein, D. et al. (2024). Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

ShanghaiRanking Consultancy (2024). Academic ranking of world universities 2024. Accessed: 2025-04-25.

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

Singhal, K. et al. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Team, T. (2024). Falcon 3 family of open foundation models.

Times Higher Education (2024). World university rankings 2024. Accessed: 2025-04-25.

Yang, A. et al. (2024). Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhang, S. et al. (2023). Planning with large language models for code generation. *arXiv preprint arXiv:2303.05510*.

Zhang, T. et al. (2024a). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Zhang, Y. et al. (2024b). A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Zhong, W. et al. (2023). Agieval: A human-centric benchmark for evaluating foundation models.