EduBench: A Portuguese Benchmark for Open-Ended Discursive Question Answering

Pedro Henrique Paiola[1], **Luís Gabriel Damiati Mendes**[1], **Bruno de Oliveira Monchelato**[1], **André da Fonseca Schuck**[1], **Gabriel Lino Garcia**[1], **Douglas Rodrigues**[1], Helena de Medeiros Caseli[2], João Paulo Papa[1]
[1]São Paulo State University (UNESP), [2]Federal University of São Carlos (UFSCar)
[1]Av. Eng. Luís Edmundo Carrijo Coube, 14-01 - Bauru - SP - Brazil
[2]Rodovia Washington Luis, km 235 - São Carlos - SP - Brazil
{pedro.paiola, lg.mendes, bruno.monchelato, andre.schuck,
gabriel.lino, d.rodrigues, joao.papa}@unesp.br
helenacaseli@ufscar.br

Abstract

Evaluating open-ended text generation in large language models remains challenging, particularly for non-English languages. We introduce EduBench, a comprehensive Portuguese-language benchmark comprising 3,149 discursive questions from Brazilian university entrance examinations spanning 2015–2025. Unlike multiple-choice or extractive QA benchmarks, EduBench requires extended, argumentative responses across diverse domains, including Humanities, Exact and Natural Sciences, and Languages. Each question includes expert-curated reference answers from official sources, rich metadata, and automated image descriptions to support text-only evaluation. We establish baseline results using nine contemporary models, ranging from 4B-parameter SLMs to state-of-the-art reasoning-capable LLMs, and evaluate them using complementary metrics (BLEU, BERTScore, G-Eval). Our results reveal substantial metric disagreement and highlight the complexity of assessing discursive generation, with models achieving 54–71% alignment with expert answers. We release EduBench publicly to support research on Portuguese NLP and open-ended generation evaluation.

Keywords: discursive question answering, open-ended generation, Portuguese, NLP, LLM evaluation, educational assessment

1. Introduction

Large language models (LLMs) have rapidly redefined the state of the art across a broad spectrum of natural language processing (NLP) tasks, surpassing earlier milestones established by recurrent neural network (RNN) architectures. Their remarkable versatility has driven substantial progress in text generation, reasoning, and cross-domain knowledge transfer, positioning them as powerful general-purpose language technologies.

Despite these achievements, systematically evaluating the open-ended text generation capabilities of LLMs remains a fundamental challenge. Traditional benchmarks have largely emphasized objective and well-defined tasks, including text classification, multiple choice answers, and span-based extraction. Although such tasks enable stable and reproducible evaluation, they capture only a limited subset of the competencies expected from contemporary LLMs. Critically, they overlook essential abilities such as structured multistep reasoning, discursive explanation, and the capacity to follow complex, multifaceted instructions.

Unlike tasks with clearly defined ground truth, such as classification, multiple choice answering, or factual recall, open-ended generation requires assessing dimensions such as coherence, factual grounding, reasoning depth, and stylistic appropriateness. These aspects are inherently difficult to measure automatically and consistently, resulting in a persistent gap between leaderboard performance and real-world utility.

As LLM capabilities continue to advance, there is a growing need for benchmarks that directly assess open-ended, discursive generation rather than relying on indirect proxies such as option selection or span extraction (Wang et al., 2024; Srivastava et al., 2023). Multiple-choice benchmarks such as the Massive Multitask Language Understanding (MMLU) and its more discriminative successor, MMLU-Pro, span a wide range of domains and provide repeatable evaluation protocols. However, they primarily test recognition and short-form recall, rather than the ability to construct extended arguments or elaborate explanations (Wang et al., 2024). In contrast, extractive QA datasets such as SQuAD and large-scale reading comprehension benchmarks like MS MARCO target answerability but constrain outputs to short spans, limiting their ability to evaluate rich and discursive reasoning (Rajpurkar et al., 2016; Bajaj et al., 2018).

Long-form QA datasets such as ELI5 move closer to assessing extended, open-ended genera-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4587–4596
11–16 May 2026.

tion, though they often inherit noise and inconsistency from crowd-sourced annotations (Fan et al., 2019). Instruction-following benchmarks like IFEval provide a complementary perspective by introducing verifiable constraints (e.g., length requirements, keyword inclusion), which allow for automatic evaluation of controllability and format adherence (Zhou et al., 2023). More recently, model-based evaluators have emerged as practical surrogates for human judgment in open-ended tasks, yet they raise critical concerns regarding bias, reliability, and robustness (Zheng et al., 2023).

In this work, we pursue two primary objectives. First, we introduce a new Portuguese-language benchmark for open-ended, discursive question answering, derived from the second phase of major Brazilian university entrance examinations. The benchmark includes ground-truth reference answers curated from official solutions and expert commentaries, ensuring both linguistic richness and evaluative fidelity. Second, we present preliminary experiments with contemporary LLMs on this benchmark, exploring evaluation protocols tailored explicitly to discursive generation, and establishing baseline performance results.

By releasing the dataset and accompanying code[1], we aim to bridge a critical gap in evaluation resources for open-ended generation in Portuguese and to provide a solid foundation for future research on discursive reasoning in LLMs.

2. Related Work

2.1. Benchmarks

Multiple-choice evaluations remain widely adopted due to their scalability, ease of automation, and inherent inter-rater reliability. The MMLU benchmark popularized large-scale knowledge testing across diverse domains. Its successor, MMLU-Pro, introduces more challenging items, featuring ten-option questions and greater discriminative power, thereby mitigating the saturation effects observed in earlier benchmarks (Wang et al., 2024).

For Portuguese, the BLUEX benchmark compiles multiple-choice questions from major Brazilian university entrance examinations, including USP and UNICAMP. The dataset is enriched with metadata and image annotations and focuses on high-school curricular knowledge in Portuguese. BLUEX has emerged as a widely adopted baseline for evaluating language models in Portuguese (Almeida et al., 2023).

More recently, the *Exame Nacional do Ensino Médio* (ENEM) has also been explored as a valuable benchmark. As one of Brazil's primary university admission exams, ENEM covers a wide curricular range through four objective tests and a writing essay (Locatelli et al., 2025). In addition to its breadth, the release of anonymized student responses and socioeconomic data enables meaningful comparisons between LLM and human performance across demographic groups.

While such resources ensure breadth, standardization, and comparability, they ultimately evaluate option selection rather than the ability to generate coherent, multi-paragraph arguments. As a result, they fail to capture the discursive reasoning skills required in open-ended, essay-style tasks.

Datasets such as SQuAD (Rajpurkar et al., 2016) and FaQuAD (Sayama et al., 2019) for Portuguese formulate question answering as a span selection problem over Wikipedia, enabling precise automatic evaluation using metrics such as Exact Match and token-level F1. MS MARCO extends this paradigm by pairing real-world search queries with relevant passages and human-written answers, thereby promoting research on reading comprehension and retrieval-augmented generation (Bajaj et al., 2018). Despite their significance, these corpora typically yield short answers or brief justifications, which fail to capture the complexity of discursive reasoning required in essay-style settings.

Long-form QA benchmarks explicitly aim to elicit multi-sentence or paragraph-level responses. ELI5, for instance, uses Reddit questions that require layperson-friendly explanations, encouraging multi-sentence, self-contained answers and supporting document retrieval as part of the generation process (Fan et al., 2019). NarrativeQA, on the other hand, evaluates deeper narrative comprehension by posing questions about books and movie scripts, which must be answered using summaries or complete texts (Kočiský et al., 2018).

These corpora represent an essential step beyond short-span answers toward extended, discursive generation. However, they vary considerably in domain and source quality. For example, the community-sourced responses in ELI5 often exhibit inconsistencies in factual rigor, whereas the literature-oriented NarrativeQA operates in a more specialized domain with narrower applicability.

2.2. LLMs

Benchmark suites designed to probe LLMs' broader capabilities offer complementary perspectives on generative evaluation. BIG-bench, for instance, aggregates over 200 tasks spanning diverse linguistic phenomena, including generative tasks with rubric-based scoring to assess reasoning and emergent

[1] Dataset available at https://huggingface.co/datasets/recogna-nlp/EduBench. Source code and evaluation scripts available at https://github.com/pedropaiola/EduBench

behaviors at scale (Srivastava et al., 2023). Orthogonally, IFEval introduces automatically verifiable constraints, such as output length and keyword inclusion, to evaluate instruction adherence alongside content accuracy (Zhou et al., 2023). Together, these resources underscore that generation quality is a multidimensional construct encompassing both content-related factors (factuality, reasoning, relevance) and form-related aspects (style, structure, constraint satisfaction). This motivates the development of evaluation protocols that integrate both dimensions.

Human evaluation remains the gold standard for assessing open-ended text generation but is inherently costly and time-consuming. Recent studies have explored using strong LLMs as evaluators, exemplified by benchmarks such as MT-Bench and Chatbot Arena, which report high agreement with human judgments while also exposing biases, such as position and verbosity effects, and potential mitigation strategies (Zheng et al., 2023). In exam-style discursive response settings, reference-based comparisons with expert solutions can be combined with rubric-driven evaluation criteria. In some instances, carefully controlled LLM-as-judge protocols may also be employed, provided their limitations are explicitly acknowledged and mitigated.

3. EduBench

3.1. Data Collection

We compiled a large-scale benchmark of discursive question–answer pairs from the second-stage entrance examinations of three of the most prestigious Brazilian universities: UNESP, USP, and UNICAMP. The dataset spans a decade of exams, from 2015 to 2025, and focuses exclusively on open-ended, essay-style items. Whenever available, official commented solutions were used as reference answers. For editions lacking official commentaries, we relied on publicly available annotated solutions from *Colégio Objetivo*, a well-known high school network in Brazil. All source materials were obtained in PDF format.

Each question record retains essential metadata, including the institution, year, and subject when available. This design ensures contextual traceability while emphasizing linguistically grounded and reasoning-oriented evaluation.

3.2. Extraction and Formatting

Our data processing pipeline comprises three stages: (i) PDF text and image extraction, (ii) figure captioning, and (iii) LLM-assisted structuring.

(i) Text and image extraction: We extracted textual content and visual assets from the original PDFs, reconstructing question statements together with their associated materials, for example, excerpts, diagrams, and charts. Mathematical expressions present in the original documents were preserved whenever they were explicitly encoded.

(ii) Image description: To keep the dataset self-contained for text-only evaluation and to facilitate automatic scoring, each extracted figure was paired with a concise natural-language caption. Captions were generated using a lightweight multimodal LLM, GPT-4o-mini, configured to produce neutral, factual, and self-contained descriptions suitable for replacing the original figures when necessary.

(iii) LLM-assisted structuring: We used a higher-capability model, GPT-4o, to organize the extracted content into a structured schema. This stage involved:

- segmenting exams into individual questions and sub-questions, treating each one as an independent item while preserving shared context,
- normalizing question prompts and supporting texts, including generated figure captions, by removing PDF artifacts while maintaining semantic fidelity and source attribution,
- attaching ground-truth reference answers from the commented solutions, and deriving a short-form answer, for example, categorical or numerical, when appropriate to support specific evaluation tasks,
- assigning a broad subject label such as Languages, Mathematics, Natural Sciences, or Human Sciences.

This structured representation ensures consistency across institutions and exam editions, enabling downstream evaluation without manual per-exam adjustments.

3.3. Post-Processing and Quality Assurance

Both PDF extraction and LLM-based structuring are susceptible to errors like broken formatting, missing assets, or unintended transformations. To mitigate these issues, we implemented a semi-automatic quality assurance pipeline that combines lightweight NLP checks with targeted LLM assistance, followed by human verification on flagged cases.

(i) Missing figures and cross-references: We identified potential missing visual elements by cross-referencing textual mentions such as "see figure" or "as shown in the graph", against the set of extracted captions. Items with missing references were flagged for manual review.

(ii) Normalization of commentary-style answers: Commented solutions often describe what candidates should state rather than providing canonical answers. In these cases, the LLM rewrote the solutions into concise, self-contained reference answers while preserving factual accuracy and detail. Both the original commentary and the normalized version were retained for quality assurance to prevent semantic drift.

(iii) Consistency and formatting: We standardized typography, list structures, and mathematical delimiters where applicable, ensuring internal consistency among prompts, supporting texts, and reference answers. A human curator manually reviewed all flagged items and a stratified sample of unflagged ones to evaluate residual noise and confirm fidelity to the source materials. This manual verification step included checking segmentation boundaries, validating the semantic faithfulness of LLM-rewritten reference answers against the original commented solutions, and ensuring consistency between generated figure captions and their corresponding visual content.

Although the pipeline leverages LLMs for structuring and normalization, final inclusion of items in EduBench required explicit human oversight for all automatically flagged cases and targeted validation of randomly sampled instances, ensuring that the released corpus preserves the integrity of the original examination materials.

(iv) Human verification and community curation: In addition to automated checks, we performed a manual pass to correct segmentation errors, misplaced sub-items, malformed mathematical notation, and missing cross-references. Although extensive quality control was conducted, residual errors may still exist. To address this, the dataset will be released publicly under a versioned scheme, and future corrections and improvements will be integrated into documented updates. We invite the research community to contribute to this ongoing curation effort.

(v) Readability assessment: To characterize the linguistic complexity of questions and answers, we computed the Flesch Reading Ease Index adapted for Portuguese (Martins et al., 1996). Initially developed by Flesch (Flesch, 1948) for English, the formula assesses text readability based on average sentence length and syllable count per word. The Portuguese adaptation by Martins et al. (1996) adjusts the coefficients to account for Portuguese's morphological and syntactic properties, yielding scores from 0 (very difficult) to 100 (very easy). The adapted formula is:

$$\text{Flesch}_{\text{PT}} = 248.835 - 1.015 \times \left(\frac{W}{S}\right) - 84.6 \times \left(\frac{Sy}{W}\right), \quad (1)$$

where S denotes the number of sentences, W the number of words, and Sy the number of syllables. This metric provides an objective, language-specific measure of textual complexity that complements human-annotated difficulty ratings and enables correlation analysis with model performance.

(vi) Reproducibility note: The dataset construction pipeline prioritizes transparency and reproducibility. The corpus is derived entirely from publicly available PDFs, GPT-4o-mini generates figure captions, and GPT-4o performs question structuring and answer normalization. This modular design allows future reprocessing with updated models without altering the underlying methodology or data provenance. To allow replicability, the pipeline code will be made available with the final version of this paper.

3.4. Corpus Statistics

The resulting dataset, called **EduBench**, comprises 3,149 discursive questions collected from second-stage entrance examinations administered by USP, UNICAMP, and UNESP between 2015 and 2025. It represents, to our knowledge, the most significant benchmark for Portuguese-language essay-style question answering, capturing a decade of authentic, high-stakes assessment material that demands argumentative, multi-paragraph responses.

To illustrate the structure and content of EduBench, we present a representative example from the 2020 USP examination (Biology, difficulty level 7), translated into English:

> **Question:** Individuals who are lactose intolerant cannot digest this sugar present in milk. The main cause of lactose intolerance is the reduction in lactase enzyme production, especially in adulthood. [...] b) The coding region of genes is preceded and controlled by a regulatory region. A mutation (C to T) in the area highlighted in the table occurred about 10,000 years ago in people from Northern Europe. It was conserved, resulting in continued gene expression in adulthood and, consequently, the persistence of the ability to digest lactose. In which region of the gene did this mutation

occur? Based on this mutation, what is the inheritance pattern of the trait “Lactose tolerance in adulthood”?

Supporting text: [Table showing genetic sequences for 19,000 nucleotides upstream of the first exon, with alleles from four individuals exhibiting tolerant/intolerant phenotypes]

Reference answer: According to the information in the table, the mutation occurs in a region located 19,000 nucleotides upstream of the first exon. Thus, the mutation occurred in a non-coding region of the gene, which is related to gene regulation. The inheritance pattern of lactose tolerance is autosomal dominant.

Short answer: Non-coding region; Autosomal dominant inheritance

This example illustrates several key characteristics of EduBench: (i) questions require domain-specific knowledge and multi-step reasoning to be answered, (ii) supporting materials provide essential context, (iii) reference answers demonstrate the expected depth and structure of responses, and (iv) short-form answers offer an alternative evaluation target for specific metrics.

We now summarize the main corpus-level statistics and design choices underlying EduBench:

- **Overall composition (Fig. 1):** EduBench is balanced across the three major Brazilian institutions: 1,465 questions from USP (46.5%), 963 from UNICAMP (30.6%), and 721 from UNESP (22.9%). Across disciplines, Humanities dominate the dataset (32.1%), followed by Exact Sciences (16.7%), Portuguese Language and Literature (14.4%), Biology (13.9%), Physics (11.9%), Chemistry (10.2%), and English (0.8%). The dataset spans eleven examination cycles (2015–2025) and captures evolving curricular emphases and question design patterns. Difficulty levels, reported on a 0–10 scale, exhibit a realistic distribution centered around intermediate values (mean 6.12, median 6.00), with most items concentrated between levels 5 and 7. These difficulty scores were derived from two complementary sources: in some cases, the official exam resolutions explicitly reported difficulty levels, often based on the proportion of students who correctly solved each question; in others, difficulty was inferred by an LLM through textual and contextual cues in the question and reference answer. As such, this field may be refined in future work through systematic human validation and cross-model calibration.

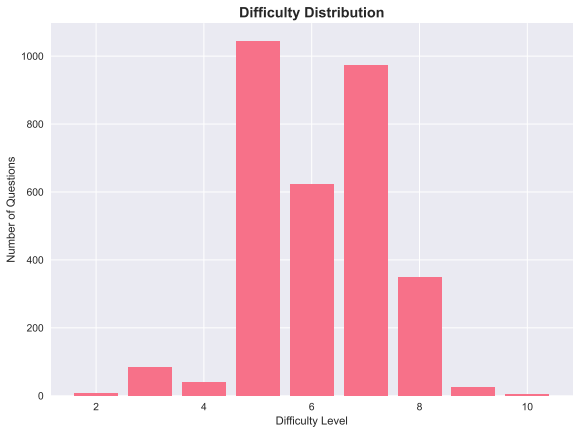
- **Temporal variation:** The dataset spans eleven years, reflecting fluctuations in the number and composition of questions per exam. In Brazilian university entrance examinations, candidates typically undergo two stages: an initial multiple-choice phase followed by a second stage composed of open-ended, discursive questions. For instance, the second stage of UNESP was replaced by multiple-choice tests in some years (e.g., 2021–2023), while USP and UNICAMP maintained consistent two-stage formats. These variations introduce temporal diversity in question design, subject emphasis, and difficulty distribution.

- **Statement and answer complexity (Fig. 2):** EduBench exhibits significant variability in question and answer length, reflecting the complexity of the reasoning tasks. Prompts contain on average 59.0 words (median 40), while answers average 69.2 words (median 59). Length distributions are heavy-tailed, with a substantial fraction exceeding 150 words. Differences across disciplines are pronounced: Physics questions are the longest (101.8 words on average), followed by Chemistry (83.4) and Exact Sciences (76.3), whereas Humanities and Portuguese tend to be more concise (34.2 and 36.8 words, respectively). Such variation provides a challenging benchmark for models tasked with generating contextually rich, discursive responses.

- **Visual and multimodal content:** Visual elements are pervasive in EduBench: 1,633 questions (51.9%) include at least one figure, and 296 (9.4%) require direct image interpretation. The prevalence of images varies significantly across exams and subjects: USP features the highest proportion (55.7%). At the same time, Exact Sciences (65.3%) and Physics (60.4%) show the most frequent use of visual context compared to Humanities (44.0%). Most questions contain one or two images, but a small number include four or more, illustrating a wide range of multimodal complexity.

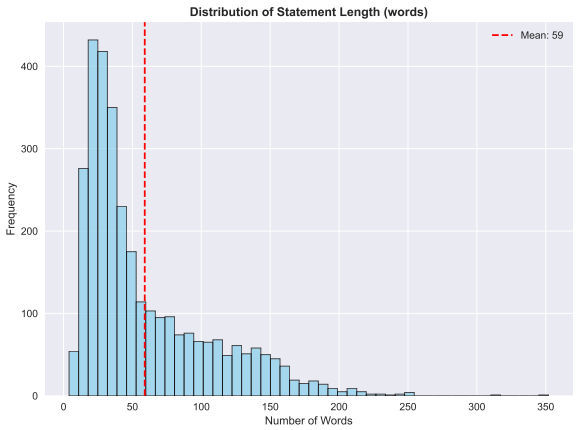
- **Difficulty landscape:** Difficulty analysis reveals a realistic distribution across levels (2–10 scale), with most questions requiring intermediate reasoning skills (mean 6.12, median 6.00). Chemistry (6.36), Exact Sciences (6.30), and Physics (6.28) are the most challenging subjects, while English (4.84) and Portuguese (5.80) are easier on average, as shown in Figure 3. Difficulty values were obtained from two complementary sources: in some cases, official exam resolutions reported question difficulty based on student performance data; in others, such a difficulty was inferred by an LLM from textual and contextual information. This estimation approach introduces some variability, suggesting that future work could refine

(a) Temporal Distribution

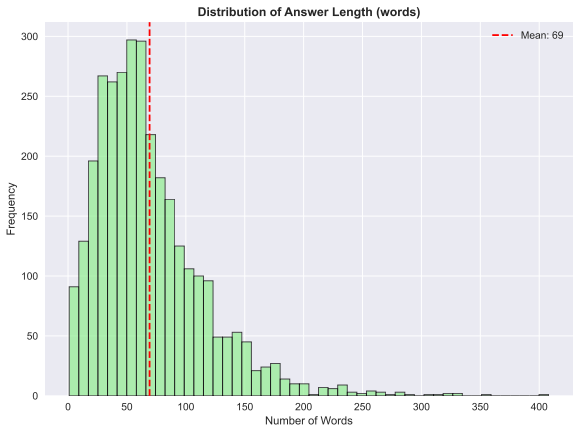


(b) Difficulty Distribution

Figure 1: Overview of questions across exams, subjects, years, and difficulty levels.



(a) Distribution of Statement Length (words)



(b) Distribution of Answer Length (words)

Figure 2: Length analysis of statements and answers: histograms showing the distribution of word counts.

this field through systematic human validation. Difficulty also shows temporal variation, with slight increases observed in recent years. Such diversity enables fine-grained evaluation of reasoning performance across subjects and over time.

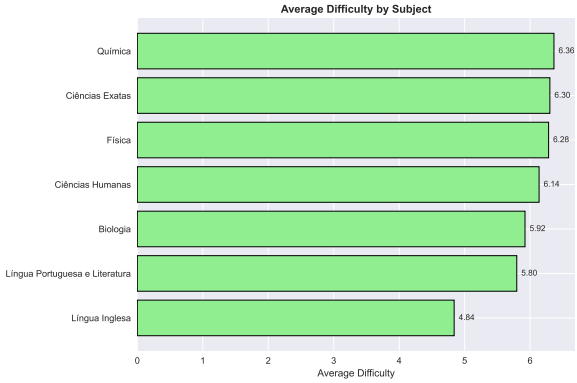


Figure 3: Average difficulty (2–10) by subject in EduBench (2015–2025).

- **Readability and textual complexity:** Flesch$_{PT}$ scores for question statements average 46.2 (median 48.1), indicating moderately difficult text suitable for high school graduates, which aligns with the target audience of university entrance examinations. Reference answers exhibit slightly higher complexity, with a mean Flesch$_{PT}$ score of 42.8 (median 44.3), reflecting the technical precision and argumentation density expected in expert responses. Notably, Flesch$_{PT}$ scores correlate negatively with difficulty ratings (Spearman $\rho = -0.34$, $p < 0.001$), confirming that linguistically complex questions tend to be perceived as more challenging. This correlation, while significant, is weak, suggesting that difficulty depends on factors beyond textual complexity alone, including domain knowledge requirements and reasoning depth.

EduBench captures a rich diversity of discursive tasks across subjects, difficulty levels, and modal-

ities. Its scale, linguistic variety, multimodal complexity, and temporal depth make it a uniquely challenging benchmark for evaluating open-ended generation and reasoning in Portuguese. By reflecting the structure and demands of real-world educational assessments, it provides a rigorous testbed for analyzing model performance across multiple dimensions of language understanding and production.

4. EduBech validation

To evaluate the preliminary performance of LLMs on the EduBench dataset, nine language models were selected and categorized into two main groups: state-of-the-art (SOTA) proprietary LLMs and open-source small language models (SLMs), including models with and without fine-tuning for Brazilian Portuguese. Model specifications and generation parameters are presented in Table 1.

4.1. Setup and Protocol

We unify configurations to keep the focus on the dataset. Proprietary models run with sampling and `temperature=1`. For models with native reasoning, GPT 5, DeepSeek v3.2-Exp, Claude Sonnet 4, and Gemini 2.5 Pro, we standardize a comparable budget using `reasoning_effort=low` and, where applicable, `budget_tokens=3000`. Sabiá 3.1 does not expose native reasoning. Open-source SLMs run via HuggingFace Transformers on Colab L4 with greedy decoding, `torch.manual_seed(4)`, and `max_new_tokens=32000`.

Prompts are zero-shot and primarily in Brazilian Portuguese. Each prompt has three parts: (i) a brief contextual framing, (ii) the question statement delimited by XML tags, and (iii) the task description. When available, supporting texts and image captions, also XML-delimited, appear between the statement and the task. A Jinja2 template ensures consistent rendering across instances while preserving question-specific content. Metadata such as university, year, and subject are intentionally excluded.

All nine models answer the 3,149 EduBench questions, yielding 28,341 responses. We then compute the metrics described below. Extended implementation notes and ablations appear in the appendix to keep this section centered on the dataset.

4.2. Evaluation Metrics

We combine three complementary metrics, each targeting a distinct facet of answer quality.

- **BLEU** (Papineni et al., 2002) measures n-gram precision against references. It is widely used in n-gram based assessment but correlates weakly with human judgment in open-ended generation, since correct answers may use different wording. We report it as a standard lexical baseline.
- **BERTScore** (Zhang et al., 2020) uses contextual embeddings to score semantic similarity. It captures paraphrases beyond surface overlap and is better suited to discursive generation. We use the multilingual BERT model for Portuguese text.
- **G-Eval** (Liu et al., 2023) employs a strong LLM as judge with explicit criteria. In our experiments, G-Eval was implemented using GPT-4o with temperature set to 0 to ensure deterministic scoring. We score coherence, fluency, and relevance from 0 to 100 and average them. While LLM-as-judge protocols have been shown to correlate with human assessments in prior work, they may inherit model-specific biases and should be interpreted accordingly. This approximates human judgment more closely than purely automatic metrics, although it may inherit LLM biases (Chen et al., 2024).

4.3. Results

Table 2 reports scores for all models. Overall, open-ended discursive answers produce low lexical overlap and substantial variation in semantic and judge-based metrics, with proprietary models generally ahead and small open models showing competitive pockets.

- **Metric disagreement.** The metrics rank models differently. BLEU is uniformly low, 2.43–6.04%, reflecting limited n-gram overlap in open-ended answers. GPT 5 and DeepSeek v3.2-Exp lead in lexical overlap; BERTScore shows a similar preference with greater spread, 11.45–22.26%. G-Eval assigns higher absolute scores, 54.40–70.90%, and places Claude Sonnet 4 and Gemini 2.5 Pro at the top, highlighting complementary dimensions.
- **Reasoning capabilities.** Native reasoning in Claude Sonnet 4, Gemini 2.5 Pro, DeepSeek v3.2-Exp, and GPT 5 aligns with stronger G-Eval scores, suggesting that controlled deliberation improves coherence and relevance, with smaller gains on purely automatic metrics. Sabiá 3.1, without native reasoning, it is competitive but does not lead.
- **Small vs. large models.** SLMs, Gemma-3-4B, Qwen3-4B, Qwen3-8B, and Gaia-PT-BR-4B lag behind proprietary models, which is consistent with scale and training data differences.

Model Type	Model Name	Model Version	PT-BR Fine-Tuning	Consumption Method	Generation Configuration				
					Decoding Strategy	Max New Tokens	Temperature	Candidate Count	Reasoning
SOTA Proprietary LLMs	Claude Sonnet 4	2025-05-14	×	Anthropic SDK (Batch API)	Sampling	32,000	1	1	✓ †
	DeepSeek v3.2-Exp	2025-09-29	×	OpenAI SDK (API)	Sampling	32,000	1	1	✓ ‡
	Gemini 2.5 Pro	2025-06-17	×	GenAI SDK (Batch API)	Sampling	32,000	1	1	✓ †
	GPT 5	2025-08-07	×	OpenAI SDK (Batch API)	Sampling	32,000	1	1	✓ ‡
	Sabiá 3.1	2025-05-08	✓	OpenAI SDK (Batch API)	Sampling	32,000	1	1	×
Open-Source SLMs	Gemma-3-4B-it	2025-03-12	×	HF Transformers (Colab L4)	Greedy	32,000	—	—	×
	Gaia-PT-BR-4B-it	2025-06-10	✓	HF Transformers (Colab L4)	Greedy	32,000	—	—	×
	Qwen3-4B-Instruct	2025-08-05	×	HF Transformers (Colab L4)	Greedy	32,000	—	—	×
	Qwen3-8B	2025-04-28	×	HF Transformers (Colab L4)	Greedy	32,000	—	—	×

† Budget tokens set to 3,000.
‡ Reasoning effort configured as low.

Table 1: Models selected for preliminary evaluation and their generation parameters.

Model	BLEU (%)	BERTScore (%)	G-Eval (%)
Claude Sonnet 4	3.95	15.87	**70.90**
DeepSeek v3.2-Exp	**6.04**	<u>21.15</u>	70.24
Gemini 2.5 Pro	4.58	18.50	<u>70.73</u>
GPT 5	<u>5.21</u>	**22.26**	66.53
Sabiá 3.1	4.36	16.83	67.14
Gemma-3-4B-it	3.52	14.95	55.24
Gaia-PT-BR-4B-it	2.43	11.45	54.40
Qwen3-4B-Instruct	3.54	14.17	65.79
Qwen3-8B	4.05	13.84	64.59

Table 2: Model performance on EduBench across three evaluation metrics. Bold indicates best performance; underline indicates second best.

Still, Qwen3-4B-Instruct reaches 65.79% on G-Eval, close to Sabiá 3.1 (67.14%), indicating viability in resource-constrained scenarios.

- **Difficulty and readability.** Across difficulty ranges 1–3, 4–6, and 7–10, mixed patterns emerge, including occasional gains on harder items, possibly due to richer reference answers. Correlation with $\text{Flesch}_{\text{PT}}$ is positive but weak for G-Eval (mean $\rho = 0.12$, $p < 0.01$), higher in proprietary models ($\rho = 0.15$–0.19) than in SLMs ($\rho = 0.06$–0.10). BERTScore shows negligible correlation ($\rho < 0.05$), indicating orthogonality to surface-level readability.

These results establish baselines on EduBench and reinforce the challenge of evaluating open-ended discursive generation. The disagreements between metrics and the model-specific patterns motivate deeper validation of LLM-as-judge protocols against human ratings, with EduBench offering a controlled, Portuguese-language testbed centered on authentic exam questions.

5. Conclusions

We introduced EduBench, a large-scale Portuguese-language benchmark for open-ended, discursive question answering comprising 3,149 questions from major Brazilian university entrance examinations spanning 2015–2025. Unlike existing Portuguese benchmarks that focus on multiple-choice or short answers, EduBench requires extended, argumentative responses across diverse academic subjects, providing a more comprehensive evaluation of language models' generative and reasoning capabilities.

Our contributions are threefold. First, we present a transparent methodology for building discursive QA datasets from authentic educational materials, including automated extraction, multimodal handling via image captioning, and LLM-assisted structuring. Second, we provide corpus statistics and analyses, including readability measured with the Portuguese-adapted Flesch index, which reveals systematic relationships between textual complex-

ity and perceived difficulty. Third, we establish baseline results using *nine* contemporary models, from 4B-parameter SLMs to state-of-the-art reasoning-capable LLMs, evaluated with complementary metrics (BLEU, BERTScore, G-Eval).

5.1. Limitations and Future Work

EduBench has limitations. The corpus is drawn exclusively from Brazilian Portuguese entrance examinations, which may limit generalizability to other Portuguese varieties or educational contexts. Despite careful quality assurance, residual issues from automated extraction or structuring may persist. The current experimental evaluation relies exclusively on automatic metrics, including embedding-based similarity and an LLM-as-judge protocol, without a dedicated large-scale human evaluation of model outputs. Consequently, while the reported results provide useful baseline signals, they should not be interpreted as definitive indicators of discursive answer quality.

A systematic human study assessing coherence, factual accuracy, argumentative completeness, and pedagogical adequacy would strengthen conclusions regarding model performance on EduBench. Such validation is an important direction for future work and would also enable calibration of automatic metrics in the context of Portuguese discursive question answering.

Future work will address these points. Human studies comparing automatic metrics with expert assessments can validate protocols and guide metric refinement. Expanding the corpus to other Portuguese-speaking regions and educational levels can improve representativeness. Developing evaluation criteria tailored to discursive generation in educational settings may increase assessment fidelity. An in-depth assessment of coherence, factual basis, depth of reasoning, and stylistic appropriateness can also highlight the full potential of EduBench when evaluating open generation. Finally, fine-tuning smaller models on EduBench can inform strategies for resource-efficient educational AI.

By releasing EduBench and the accompanying code, we aim to catalyze research in Portuguese NLP, evaluation of open-ended generation, and educational AI. We invite the community to extend, refine, and build upon this resource to advance robust evaluation and improvement of language models for discursive tasks.

6. Bibliographical References

Thales Sales Almeida, Thiago Laitz, Giovana K. Bonás, and Rodrigo Nogueira. 2023. Bluex: A benchmark based on brazilian leading universities entrance exams.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering.

Rudolf Flesch. 1948. *A New Readability Yardstick*, volume 32. American Psychological Association.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Marcelo Sartori Locatelli et al. 2025. *Examining the Behavior of LLM Architectures within the Framework of Standardized National Exams in Brazil*, page 879–890. AAAI Press.

Teresa Bertholdo Freire Martins, Claudete Moreno Ghiraldelo, Maria das Graças Volpe Nunes, and Osvaldo Novais de Oliveira Junior. 1996. *Readability Formulas Applied to Textbooks in Brazilian Portuguese*. Ph.D. thesis, Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo, São Carlos, Brazil. Notas do ICMSC-USP, Série Computação, n. 28.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Hélio Fonseca Sayama, Anderson Viçoso Araujo, and Eraldo Rezende Fernandes. 2019. Faquad: Reading comprehension dataset in the domain of brazilian higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448.

Aarohi Srivastava et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models.