Beyond Aggregate Scores: A Diagnostic Analysis of Portuguese LLM Evaluation Suites

Gustavo Hochgraf[1], **Gabriel Assis**[2], **Aline Paes**[2], **Fabio G. Cozman**[1]

[1]Escola Politécnica, Universidade de São Paulo, São Paulo, SP, Brasil

[2]Instituto de Computação, Universidade Federal Fluminense, Niterói, RJ, Brasil

gusthoch@gmail.com, assisgabriel@id.uff.br, alinepaes@ic.uff.br, fgcozman@usp.br

Abstract. *Benchmarks play a central role in evaluating large language models (LLMs), providing standardized comparisons across models, tasks, and adaptation strategies. However, aggregate benchmark scores often compress performance into a single number, obscuring important variation across tasks. This issue is especially relevant for less-resourced languages like Portuguese, where benchmarks may combine native tasks with translated datasets spanning heterogeneous categories and subareas. In this work, we examine this issue using PoETa v2, a broad Portuguese evaluation benchmark, as a diagnostic setting to evaluate three Qwen3 1.7B variants under different Portuguese adaptation settings. Our results show that, although overall scores differ by less than one percentage point across models, disaggregated analyses reveal substantially different patterns across task origin, category, and subarea. In particular, gains on native Portuguese tasks may coexist with losses on translated tasks, while different adaptation corpora redistribute performance in distinct ways. Taken together, our findings highlight the importance of complementing aggregate scores with multi-level analyses when evaluating Portuguese LLMs, particularly in the context of language-specific adaptation.*

1. Introduction

In recent years, large language models (LLMs) have become a central component of Natural Language Processing (NLP), reshaping how language technologies are developed, adapted, and evaluated [Paes et al. 2026, Real et al. 2026]. As these models are increasingly used across different domains and linguistic contexts, their evaluation has become essential for understanding not only their overall capabilities, but also their limitations and patterns of behavior [Real et al. 2026]. Benchmarks play an important role in this process, since they provide standardized points of comparison across models and adaptation strategies. However, model performance is not uniform across languages, domains, or cultural contexts, and this variation is particularly relevant for languages that remain less represented in the development and evaluation pipelines of widely used LLMs [Huang et al. 2023].

A common practice in LLM evaluation is to compare models through a single aggregate score [Rodriguez et al. 2021]. Although this form of reporting is convenient, it can hide substantial variation across tasks and lead to incomplete or even misleading conclusions about model behavior. Two models with similar overall scores may display very different strengths and weaknesses depending on the task type, the origin of the dataset, or the linguistic and cultural phenomena being evaluated. This is especially important in multilingual and language-specific evaluation, where benchmarks may combine native

tasks, translated datasets, and heterogeneous task categories under the same final score. In such cases, aggregate scores may suggest that two models behave similarly, even when their performance is redistributed in substantially different ways across the benchmark.

Building on this motivation, we examine how aggregate scores may obscure qualitatively different performance patterns in Portuguese LLM evaluation. We instantiate this discussion with PoETa v2 [Almeida et al. 2025b], a benchmark composed of 44 tasks that covers different linguistic and cultural dimensions of Portuguese. This setting allows us to move beyond a single overall number and inspect how results vary according to task origin, category, and subarea. As a case study, we analyze three Qwen3 1.7B [Yang et al. 2025] variants under different Portuguese adaptation settings and investigate whether disaggregating PoETa results changes the interpretation suggested by the aggregate score. This analysis therefore frames aggregate scores as useful but insufficient summaries, especially when evaluating the effects of language-specific adaptation.

2. Related Work

This section provides a brief overview of prior work on Portuguese LLM benchmarks and language-specific adaptation.

2.1. Benchmarks in Portuguese

PoETa is an early benchmark for evaluating LLMs in Portuguese [Pires et al. 2023]. Its second version, PoETa v2 [Almeida et al. 2025b], extends this effort to 44 tasks and more than 20 models, highlighting how computational scale and language-specific adaptation affect performance in Portuguese. Its taxonomy includes 12 tags that may overlap across tasks, and the benchmark also supports analyses by task origin, distinguishing native and translated tasks, as well as by broader capability groupings.

Other Portuguese benchmarks target specific domains or evaluation settings, reinforcing the importance of analyses that go beyond a single aggregate score [Chaves Rodrigues 2023, Severino et al. 2025, Assis et al. 2025]. How aggregate scores are computed also varies across leaderboards: some report a single mean, while others split results by capability or origin, as PoETa v2 does. Some studies have argued that single-number metrics can hide important failures and proposed reporting results in more disaggregated ways [Real et al. 2026, Rodriguez et al. 2021].

2.2. Portuguese-Specific LLM Adaptation

Continued pre-training can improve target-language performance but may also induce forgetting [Gururangan et al. 2020]. In Portuguese, this strategy has been explored through continued pre-training and instruction tuning [Larcher et al. 2023, Almeida et al. 2025a, Garcia et al. 2024, Cruz-Castañeda e Amadeus 2025]. Since these efforts are often assessed with Portuguese-centered benchmarks, how benchmark scores are aggregated becomes a central interpretive issue.

3. Methodology

This section describes the experimental setup used to analyze how different Portuguese adaptation settings affect model performance on PoETa v2. We first present the model variants and corpora considered in the study, and then describe the benchmark subset, evaluation protocol, and reporting scheme used throughout the experiments.

Table 1. Variants evaluated and continued pre-training characteristics.

Variant	Continued pre-training	Corpus	Tokens
Qwen3 1.7B Base	None	–	–
+GigaVerbo adapted	1 epoch	GigaVerbo	∼100 B
+Carolina adapted	1 epoch	Carolina	∼200 M

3.1. Models and Corpora

We evaluate three variants of Qwen3 1.7B, summarized in Table 1. We selected this base model because, within its model-size tier, it combines the strongest previously reported PoETa [Almeida et al. 2025b] results with compatibility with our available computational resources. The first variant is the original Qwen3 1.7B checkpoint. The second is a variant tuned to the GigaVerbo corpus [Corrêa et al. 2025], a large-scale Portuguese collection assembled from multiple publicly available sources and dominated by native Portuguese text. The third is a variant tuned to the Carolina corpus [Crespo et al. 2023], a substantially smaller resource composed of texts in contemporary Brazilian Portuguese with varied typology. In both adaptation settings, training starts from the same Qwen3 1.7B model [Yang et al. 2025] and applies one additional epoch over the respective corpus. We deliberately fix the base model and vary only the adaptation corpus so that observed differences can be attributed to corpus choice rather than to architectural or pre-training variation across distinct baselines; the 1.7B size keeps continued pre-training tractable under our computational budget. This design allows us to compare how the benchmark is interpreted under different adaptation settings.

3.2. Evaluation Setup and Reporting

Following [Almeida et al. 2025b], we first report the benchmark's aggregate metrics, **All**, **Native**, and **Translated**. These correspond, respectively, to the average across all evaluated tasks, the subset of native Portuguese tasks, and the subset of translated tasks. In our experiments, all three aggregates are computed over a fixed public subset of 40 tasks from PoETa v2[1]. We do not impute or reweight missing tasks; therefore, the aggregate values reported in this paper should be interpreted with respect to this 40-task subset rather than to the original 44-task composition of PoETa v2.

The benchmark includes tasks evaluated with accuracy (`acc`), textual similarity (`pearson`), exact match for free-form answers (`best_em`), and F1. Our experiments follow the few-shot setting with dynamic example selection, in which each test instance is evaluated with a small set of automatically selected in-context examples from the same task. To keep evaluation computationally manageable while maintaining coverage across tasks, we limit each task to 300 samples.

Beyond the standard aggregate view, we introduce an additional diagnostic layer by grouping the selected tasks into six mutually exclusive categories: (i) Brazil/exams/culture, (ii) text understanding/QA/classification, (iii) social/safety, (iv) reasoning, (v) math, and (vi) code/other. This grouping is not intended to replace the original PoETa v2

[1]The excluded tasks are `poscomp_greedy`, `arc_challenge_greedy_pt`, `arc_easy_greedy_pt`, and `ethics_commonsense_test_hard_greedy`. These tasks appear in the original full benchmark configuration but are not included in the public subset available.

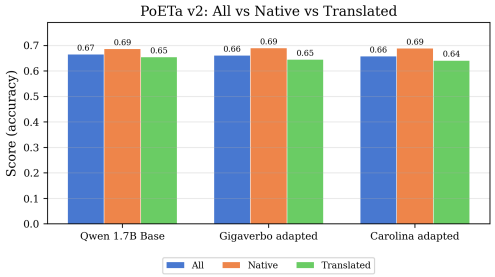

Figure 1. Normalized aggregate scores per variant.

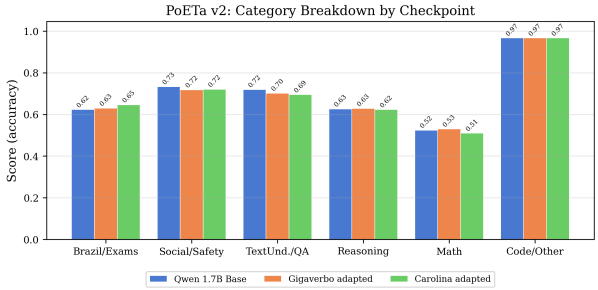

Figure 2. Average score by category for each variant.

taxonomy. Rather, it collapses the benchmark's overlapping tags and, when needed, exam-specific metadata into one primary reporting group per task, allowing us to summarize results at an intermediate level between the overall score and individual tasks. Each task also retains its original native/translated label.

For the Brazil/exams/culture group, we also report results by subarea. Model variants are compared using deltas in percentage points (p.p.) relative to the base model. Together, these category-, subarea-, and delta-based summaries constitute the main analytical extension adopted in this work.

To quantify sampling uncertainty, we accompany reported differences with 95% confidence intervals from a paired nonparametric bootstrap. Because every variant is evaluated on the same 300-example subset of each task, we resample examples with replacement (10,000 replicates) using identical indices for the three variants and recompute task scores and aggregates on each replicate, so deltas are paired at the example level. For the three tasks without recoverable per-example scores (BLUEX, FaQuAD, and MKQA) we use a normal approximation based on the harness-reported standard errors, with unpaired deltas; a difference is significant when its 95% CI excludes zero.[2]

4. Results and Discussion

4.1. Aggregate scores

Figure 1 presents the aggregated scores (All, Native, Translated) for the three variants. We observe that the All score varies by only 0.75 p.p. across models. In contrast, the Native and Translated aggregates move in opposite directions, and the bootstrap separates signal from noise in this movement: the losses on translated tasks are statistically robust (-0.93 p.p., 95% CI $[-1.69, -0.16]$, for +GigaVerbo adapted and -1.32 p.p., $[-2.27, -0.36]$, for +Carolina adapted), whereas the corresponding native gains ($+0.4$ p.p., $[-0.72, +1.47]$, and $+0.29$ p.p., $[-1.00, +1.56]$) are small relative to sampling uncertainty and not individually significant, even though specific native tasks and categories do shift significantly (Sections 4.2 and 4.3). The aggregate score therefore conceals statistically significant degradation on translated tasks behind an apparently stable overall number.

[2]Per-task scores, the task-to-category mapping, and the bootstrap analysis code are available at https://github.com/GustavoHochgraf/brasa-eval.

4.2. Analysis by category

Figure 2 shows the average scores by category. The variant adapted with the Carolina corpus stands out in Brazil/exams/culture (+2.33 p.p., 95% CI $[+0.64, +4.03]$, vs. the base), which is consistent with the stronger presence of Brazilian-oriented content in that corpus. In turn, the GigaVerbo-adapted model shows modest advantages in reasoning and math, a pattern compatible with the scale and heterogeneity of GigaVerbo. In contrast, the base model performs better in text understanding/QA/classification, which aggregates reading comprehension, NLI, sentiment, and fact-verification tasks and in social/safety, which combines hate-speech detection with bias and ethical-reasoning tasks. This indicates that the relative gains from adaptation are not uniform across categories. We note that both of these groupings are intentionally broad to preserve enough tasks per group for stable averages; a finer subdivision would likely surface additional within-category patterns and is left as future work. Under the bootstrap, the Brazil/exams/culture gain of +Carolina adapted and the losses of both adapted variants in text understanding/QA/classification (-1.76 p.p., $[-2.99, -0.51]$, and -2.45 p.p., $[-3.97, -0.98]$) are significant, as is the +GigaVerbo adapted loss in social/safety (-1.48 p.p., $[-2.90, -0.07]$); the movements in reasoning and math are not.

4.3. Largest task-level shifts

Table 2 lists the five largest improvements and drops of each variant relative to the base model. These task-level shifts are particularly revealing because the two adapted variants were exposed to different additional corpora during continued pretraining: GigaVerbo is much larger and more heterogeneous, whereas Carolina is smaller and more centered on contemporary Brazilian Portuguese. For the GigaVerbo-adapted model, the largest gains appear in reasoning tasks (BB Empirical Judgments, BB Mathematical Induction) and in one Brazilian-culture task (BRoverbs History→Proverb), while the main losses are observed in translated reading-comprehension tasks such as BoolQ and IMDb. For the Carolina-adapted model, the largest improvements occur in BB Empirical Judgments and ENEM 2022, which align with the stronger Brazilian orientation of those corpora, whereas the largest losses again occur in translated tasks such as IMDb and StoryCloze. Most of the largest losses fall on translated tasks, with IMDb being the only one to appear in both lists. A possible explanation is that these tasks rely on linguistic cues (e.g., sentiment in IMDb, pronoun resolution in WSC-285) that translation can distort. The results also indicate that corpus alignment may be more influential than corpus size for certain task families, as the smaller but more targeted Carolina corpus yields comparable or stronger gains in Brazilian-specific tasks. Taken together, these task-level results reinforce the paper's main point: different adaptation settings may lead to different performance distributions even when aggregate scores remain close.

4.4. Drill-down on sub-areas ENEM/BLUEX

To better understand the *Brazil/exams/culture* category, we analyzed the ENEM 2022 and BLUEX exams by subarea. Figure 3 shows the average scores for each subarea. In ENEM 2022, the largest gains appear in *Natural Sciences* and *Mathematics*, especially for the Carolina adapted variant (+11.5 p.p. and +9.1 p.p.). In contrast, the *Languages* subarea (which includes Portuguese plus an English or Spanish reading test) declines after Portuguese adaptation (–6.1 p.p. for GigaVerbo adapted), highlighting trade-offs in

Table 2. Top 5 improvements and losses by task relative to Qwen3 1.7B Base (p.p.). † 95% paired-bootstrap CI excludes zero.

GigaVerbo adapted		Carolina adapted	
Task	**Δ**	**Task**	**Δ**
Biggest improvements			
BB Empirical Judgments	+6.1	BB Empirical Judgments	+8.1†
BB Mathematical Induction	+5.7	ENEM 2022	+4.2
BRoverbs History→Proverb	+3.7†	BRoverbs History→Proverb	+3.7
Balanced COPA	+3.3†	BRoverbs Proverb→History	+3.7
BB Analogical Similarity	+2.3	Math MC	+3.3
Biggest losses			
BoolQ	–14.7†	IMDb	–7.3†
AGIEval SAT Math	–3.6	BB Mathematical Induction	–5.7
IMDb	–3.3†	StoryCloze	–5.7†
WSC-285	–3.2	FaQuad	–5.0
BB BBQ	–3.0†	BB BBQ	–4.7†

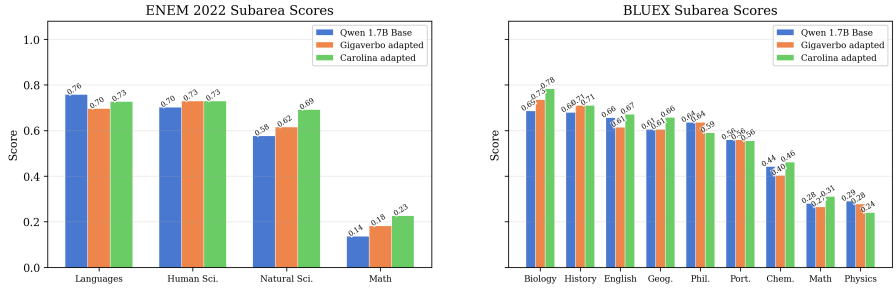

Figure 3. ENEM 2022 and BLUEX scores by subarea, averaged across tasks.

bilingual capability. In BLUEX, we observe consistent gains in *Biology*, *History*, and *Geography*, whereas quantitative subareas such as *Physics* and *Chemistry* show mixed results. Given the small number of questions per subarea, we treat this drill-down as exploratory and do not attach significance claims to subarea-level differences.

Taken together, these results show that aggregate reporting alone is insufficient for interpreting benchmark behavior in Portuguese LLMs. In our case study, models that remain close under **All** display substantially different performance redistributions once the benchmark is broken down by origin, category, subarea, and task. This is precisely why diagnostic reporting is more informative than relying only on a single summary score.

5. Conclusion

This paper presented a diagnostic evaluation of three Qwen3 1.7B variants with PoETa v2 at multiple levels of granularity. We showed that disaggregating results by origin, category, and subarea reveals trade-offs that are not visible under the aggregate score alone. In particular, the adapted variants obtain gains on native tasks while losing performance on translated tasks, with distinct redistribution patterns depending on the adaptation corpus. These findings reinforce the idea that aggregate scores provide a useful starting point for Portuguese LLM evaluation, but should not serve as the main basis for interpretation. Looking ahead, they support treating PoETa v2 less as a single benchmark and more as a set of complementary diagnostics, characterizing each model's capability profile (e.g., robustness to translation or domain sensitivity) beyond its average score.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback, and the BR-LLM group at the Center for Artificial Intelligence (C4AI-USP) for the computational infrastructure and discussions. The C4AI is supported by the São Paulo Research Foundation (FAPESP, grant #2019/07665-4) and by the IBM Corporation. F. G. C. is partially supported by CNPq Pq grant #305753/2022-3. G. A. and A. P. also acknowledge support from the CNPq, grant #307088/2023-5; FAPERJ, processes SEI-260003/002930/2024 and SEI-260003/000614/2023; and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Finance Code 001.

Limitations

This study has some limitations. It is restricted to Qwen3 1.7B variants obtained after one epoch of continued pretraining on two Portuguese corpora, and we do not attempt to isolate the causal effect of each training variable. Sampling uncertainty is quantified with paired bootstrap confidence intervals over 300-example subsets; at this sample size several aggregate differences, including the Native gains, are not individually significant, and the subarea analysis remains exploratory. The analysis also depends on the coverage and design choices of PoETa v2 itself.

Ethics Statement

This research does not involve human subjects. All models were evaluated on public tasks from the PoETa v2 benchmark, and the corpora used in continued pretraining (GigaVerbo and Carolina) are publicly available. The main ethical issue addressed in this work concerns evaluation practice: aggregate scores can obscure meaningful losses in specific task groups and may therefore mislead decisions about model development, comparison, and release. In this sense, more diagnostic reporting can contribute to more transparent assessments of Portuguese LLMs. We also note that both benchmarks and training corpora may reproduce social, cultural, and linguistic biases present in their source data.

Use of Generative Artificial Intelligence

Generative AI tools (Claude, Anthropic) were used to assist with grammatical revision, code generation for data analysis, and LaTeX manuscript formatting. All scientific content, analysis, and interpretation of results are the responsibility of the authors.

References

Almeida, T. S., Nogueira, R., e Pedrini, H. (2025a). Curió-edu 7b: Examining data selection impacts in llm continued pretraining. *arXiv preprint arXiv:2512.12770*.

Almeida, T. S., Pires, R., Abonizio, H., Nogueira, R., e Pedrini, H. (2025b). PoETa v2: Toward more robust evaluation of large language models in Portuguese. *arXiv preprint arXiv:2511.17808*.

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

Chaves Rodrigues, R. (2023). Lessons learned from the evaluation of Portuguese language models. Master's thesis, University of Malta.

Corrêa, N. K., Sen, A., Falk, S., e Fatimah, S. (2025). Tucano: Advancing neural text generation for Portuguese. *Patterns*, 6(1):101325.

Crespo, M. C. R. M., Rocha, M. L. d. S. J., Sturzeneker, M. L., Serras, F. R., Mello, G. L. d., Costa, A. S., Palma, M. F., Mesquita, R. M., Finger, M., Sousa, M. C. P. d., Namiuti, C., e Monte, V. M. d. (2023). Carolina: A general corpus of contemporary Brazilian Portuguese with provenance, typology and versioning information.

Cruz-Castañeda, W. A. e Amadeus, M. (2025). Large languages models in brazilian portuguese: A chronological survey. *Journal of the Brazilian Computer Society*, 31(1):1167–1186.

Garcia, G. L., de Medeiros, I. S., e de Oliveira, V. G. (2024). Introducing bode: A fine-tuned large language model for portuguese prompt-based tasks. *arXiv preprint arXiv:2401.02909*.

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

Huang, H., Tang, T., Zhang, D., Zhao, X., Song, T., Xia, Y., e Wei, F. (2023). Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Bouamor, H., Pino, J., e Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Larcher, C., Piau, M., Finardi, P., Gengo, P., Esposito, P., e Caridá, V. (2023). Cabrita: Closing the gap for foreign languages. *arXiv preprint arXiv:2308.11878*.

Paes, A., Vianna, D., e Rodrigues, J. (2026). Modelos de linguagem. In Caseli, H. M. e Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 2, book chapter 1. BPLN, 4 edition.

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

Real, L., Carvalho, A., e Silva, A. S. d. (2026). Avaliação de grandes modelos de linguagem: Fundamentos, métodos tradicionais e desafios atuais. In Caseli, H. M. e Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 2, book chapter 4. BPLN, 4 edition.

Rodriguez, P., Barrow, J., Hoyle, A., Lalor, J. P., Jia, R., e Boyd-Graber, J. (2021). Evaluation examples are not equally informative: How should that change NLP leaderboards? In Zong, C., Xia, F., Li, W., e Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Severino, J. V. B. et al. (2025). Benchmarking open-source large language models on Portuguese revalida multiple-choice questions. *BMJ Health Care Informatics*, 32(1):e101195.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., e Qiu, Z. (2025). Qwen3 technical report.