On the Trail of a Green and Golden Generative AI: Assessing LLM Text Generation in Brazilian Portuguese

Gabriel Assis & Aline Paes *(advisor)*

Instituto de Computação, Universidade Federal Fluminense (UFF)
Niterói, Rio de Janeiro, Brazil
assisgabriel@id.uff.br, alinepaes@ic.uff.br

Abstract. *Large Language Models (LLMs) are increasingly used in practical applications, yet their evaluation remains heavily centered on narrow discriminative benchmarks and predominantly English-language settings. The dissertation proposes a comprehensive framework for evaluating LLMs in Brazilian Portuguese across four Natural Language Generation tasks: summarization, text simplification, generative question answering, and material fact controllable text generation, a financial-domain task proposed in this work. To support evaluation in this specialized domain, the dissertation also introduces LLM4FinComm and its creation methodology, resulting in a dataset that aligns official corporate disclosures with expert financial commentaries in Portuguese. The framework is used to evaluate thirteen LLM families, including Portuguese-oriented models, open-weight multilingual models, and commercial systems, through automatic metrics, LLM-as-a-judge evaluation, human assessment, and analyses of computational efficiency and equivalent CO_2 emissions. Results show that GPT-4o models achieve the strongest overall generative performance, while Sabiá-3 closely approaches state-of-the-art results. Open-weight models present competitive efficiency trade-offs but vary substantially across tasks. Overall, the dissertation advances the evaluation of generative LLMs in Portuguese and contributes resources and methods for language- and domain-specific assessment, with its broader research agenda reflected in 21 papers and collaborations developed throughout the master's program.*

Application Information:

Institution where the degree was obtained: Universidade Federal Fluminense (UFF)

Advisor: Prof. Aline Marins Paes Carvalho, D.Sc. – Instituto de Computação (IC/UFF), Universidade Federal Fluminense (UFF), Niterói, Rio de Janeiro, Brasil

Track: MSc **Defense date:** 17 March 2026

Examination Committee:

1. Prof. Aline Marins Paes Carvalho, D.Sc., Universidade Federal Fluminense (UFF), Niterói, RJ, Brazil — *President of the Examination Committee*
2. Prof. Alexandre Plastino de Carvalho, D.Sc., Universidade Federal Fluminense (UFF), Niterói, RJ, Brazil – *Internal Member of the Examination Committee*
3. Prof. Altigran Soares da Silva, D.Sc., Universidade Federal do Amazonas (UFAM), Manaus, AM, Brazil – *External Member of the Examination Committee*
4. Prof. Arnaldo Cândido Júnior, D.Sc., Universidade Estadual Paulista (UNESP), Brazil – *External Member of the Examination Committee*

Link to the complete Master's dissertation:https://bit.ly/icuff-dissertacao-GabrielAssis

Pointers to publications containing derived work[1]**:** The research underpinning the dissertation resulted in 21 publications, including eight first-authored papers and thirteen co-authored works, published in journals, conferences, and workshops in Artificial Intelligence (AI), Natural Language Processing (NLP), Information Retrieval (IR), Data Mining, and Financial AI (FinAI). Its methodological core is reflected in JBCS [Assis et al. 2025d] and ACM ICAIF [Assis et al. 2025c], which consolidate PT-BR generative LLM evaluation and introduce the financial-domain generation task. This agenda was further developed toward generative QA, finance, and sustainability-aware LLM development through LREC [Assis et al. 2026b], ICLR FinAI WS [Assis et al. 2026a], and ACM ICAIF [Assis et al. 2025e]. Further works applied this evaluation perspective to Portuguese-centered generative tasks in STIL [Assis et al. 2024a, Stellet et al. 2025, Cavalcanti et al. 2025] and PROPOR [Masid et al. 2026, Moura et al. 2026]. Other contributions broadened this agenda across Portuguese and Lusophone language processing with language models, including ACM SAC [Amorim et al. 2025b], SBBD [Paiva et al. 2025], FUSION [Amorim et al. 2025a], and KDMiLe [Amorim et al. 2024]. Finally, publications in BRACIS [Nepomuceno et al. 2025], NTCIR [Dutra et al. 2025], KDMiLe [Martinho et al. 2025, Romero et al. 2025], ENIAC [Assis et al. 2025a], and SemEval [Assis et al. 2025b] address adjacent topics in financial AI, multilingual evaluation, and LLM capability analysis.

Links to Lattes CV, Google Scholar, and DBLP: http://lattes.cnpq.br/1303631444067208 | https://scholar.google.com/citations?user=XJYgl8YAAAAJ&hl | https://dblp.org/pid/371/5059.html

1. Introduction

Large Language Models (LLMs) have reshaped a broad spectrum of tasks and applications in Natural Language Processing (NLP), with impact across domains such as law [Malaquias-Junior et al. 2024, Pereira et al. 2025], finance [Xie et al. 2024], healthcare [Schneider et al. 2021, Paiola et al. 2024b], and creative fields [Assis et al. 2025a, Yuan et al. 2024]. Once largely confined to academic and industrial research settings, LLMs are now embedded in widely adopted technologies, including Google's search infrastructure and chat-based systems such as ChatGPT [Assis et al. 2025d].

This impact also resonates in Brazil, where the development and adaptation of local LLMs have accelerated in recent years. In the dissertation, this emerging ecosystem is referred to as the *LLM Fauna*, inspired by the convention of naming models after animals, such as LLaMA [Touvron et al. 2023], Vicuna [Chiang et al. 2023], and Falcon [Almazrouei et al. 2023]. Brazilian LLMs often follow this pattern with names drawn from native fauna, including Bode [Paiola et al. 2025], Boto [Santa Brígida 2024a], Tucano [Corrêa et al. 2024b], and Sabiá [Pires et al. 2023].

However, the increasing deployment of LLMs has not been matched by an equivalent expansion in their evaluation. Current practices often remain centered on con-

[1]A detailed discussion of the contributions of each work and their connections to the dissertation is provided in §4.2 and Appendix A.

trolled, task-specific, and discriminative benchmarks originally designed for academic settings [Zellers et al. 2019, Wang et al. 2019, Chen et al. 2021, Hendrycks et al. 2021, Rein et al. 2024]. Although such benchmarks provide valuable evidence of factual accuracy, reasoning, and task-solving ability, they offer limited insight into generative properties such as coherence, adequacy, writing quality, pragmatic appropriateness, and sensitivity to context [Assis et al. 2025d]. This limitation is particularly relevant as text increasingly serves as both the analytical substrate and the main interface through which LLMs receive instructions, process information, and produce outputs.

The issue becomes even more significant in native languages and local contexts. In these settings, models must not only operate linguistically, but also align with communicative norms, pragmatic conventions, and culturally grounded expectations [Pawar et al. 2025]. Prior work in multilingual NLP shows that models trained or adapted to specific linguistic and cultural contexts tend to produce outputs that are more appropriate, intelligible, and aligned with user expectations [Hovy and Spruit 2016, Joshi et al. 2020, Pawar et al. 2025]. Yet, evaluation practices remain largely English-centric, limiting our understanding of how LLM-powered applications behave across linguistic and cultural settings. The dissertation addresses this gap by focusing on Brazil and its official language, Brazilian Portuguese (PT-BR).

This perspective is especially relevant in domains where language, culture, and local practices intersect. In finance, for example, LLMs may help democratize access to information by generating commentaries aimed at informing investors, a type of service often paid for or hidden behind paywalls [Assis et al. 2025c]. However, effectiveness in this context requires more than fluent generation; it also depends on alignment with country-specific financial realities, including market structures, regulatory expectations, investment portfolios, and prevailing investment practices.

At the same time, LLM training and deployment require substantial computational and infrastructural resources [Corrêa et al. 2024b], which developing countries such as Brazil often struggle to establish and sustain [Lehdonvirta et al. 2025]. This disparity is illustrated by planned investments: Meta announced approximately $65 billion USD in AI infrastructure spending for 2025[2], whereas Brazil's national AI plan [Government of Brazil 2025] allocates less than $5 billion USD for 2024–2028. Beyond financial constraints, large-scale LLMs also raise environmental concerns related to electricity demand, freshwater consumption, and CO_2 emissions [Li et al. 2023], which remain underexamined even in English-language contexts [Luccioni et al. 2025].

In this context, the dissertation addressed the following research question:

RQ *What is the text generation performance of LLM-powered solutions for PT-BR?*

The work addresses this question as part of a broader path toward *Green & Golden Generative AI*. The notion of *Green* is grounded in Green AI, extended here beyond environmental responsibility to include ethical development, transparency, and accountability [Henderson et al. 2020]. The notion of *Golden* reflects the need for strong performance, since effectiveness remains essential for practical adoption. Accordingly, the analysis of generative quality is complemented by measurements of computational time,

[2] https://www.nytimes.com/2025/01/24/technology/meta-data-center.html

energy consumption, and equivalent CO_2 emissions [Budennyy et al. 2023].

From a generative evaluation perspective, the dissertation focuses on four PT-BR tasks: summarization [Souza et al. 2024], text simplification [Scalercio et al. 2025], generative question answering [Paschoal et al. 2021], and Material Fact Controllable Text Generation (MF-CTG) [Assis et al. 2024b], a financial-domain task proposed as part of the work. These tasks serve as complementary proxies for assessing content selection, meaning preservation, knowledge use, factual grounding, controlled generation, and domain-specific adequacy.

Concretely, the dissertation evaluates 13 LLM families across those four PT-BR text generation tasks, including Portuguese-oriented systems, open-weight multilingual models, and proprietary international models. The evaluation is carried out through the comprehensive framework designed in the work, combining automatic metrics, LLM-as-a-judge protocols, human assessment, and measurements of execution time, energy consumption, and equivalent CO_2 emissions. Taken together, these elements support a more situated, responsible, and comprehensive understanding of generative LLMs for Brazilian Portuguese.

2. Related Work: Summary and Research Gaps Addressed

Existing work on LLM evaluation can be organized into three main strands relevant to the dissertation. The first concerns general-purpose benchmark suites and evaluation protocols. Benchmarks such as HellaSwag, SuperGLUE, MMLU, BIG-bench, TruthfulQA, HumanEval, GPQA, SafetyBench, and ChatBot Arena are widely used to assess factual knowledge, reasoning, truthfulness, safety, instruction following, code generation, and conversational proficiency [Zellers et al. 2019, Wang et al. 2019, Hendrycks et al. 2021, Srivastava et al. 2023, Lin et al. 2022, Chen et al. 2021, Rein et al. 2024, Zhang et al. 2024, Chiang et al. 2024, Zheng et al. 2023]. Other benchmarks target specialized domains, including finance, law, and healthcare [Xie et al. 2024, Guha et al. 2023, Singhal et al. 2023]. These resources support systematic model comparison, but they are still mostly based on controlled, closed-form, discriminative, and English-language tasks. As a result, they capture whether a model can identify or select a correct answer more directly than whether it can produce coherent, faithful, useful, and contextually appropriate text.

The second strand concerns multilingual and Portuguese-oriented evaluation. Although recent initiatives have expanded benchmark coverage beyond English through multilingual resources and translated datasets [Singh et al. 2025, OpenAI et al. 2024a, Dang et al. 2024], comprehensive evaluation resources specifically tailored to Portuguese remain more limited. In Brazil, initiatives such as PoETa [Pires et al. 2023], the Open Portuguese LLM Leaderboard (OPLL) [Garcia 2024], evaluations based on national examinations such as ENEM and OAB [Almeida et al. 2025], and Portuguese NLP datasets have advanced the assessment of Portuguese-oriented models [Corrêa et al. 2024b, Corrêa et al. 2024a, Abonizio et al. 2024, Almeida et al. 2024]. However, these resources remain largely aligned with the broader benchmarking tradition, emphasizing classification, natural language inference, multiple-choice question answering, sentiment analysis, hate speech detection, and examination-style tasks.

The third strand concerns evaluation methods for Natural Language Genera-

tion (NLG) tasks. In summarization, previous studies commonly rely on ROUGE and BERTScore, while also considering coherence, focus, non-redundancy, and grammaticality [Lin 2004, Zhang et al. 2020, Jorge et al. 2025, Souza et al. 2024]. In text simplification, metrics such as SARI, BLEU, and BERTScore are frequently combined with human judgments of fluency, adequacy, and simplicity [Xu et al. 2016, Papineni et al. 2002, Zhang et al. 2020, Kew et al. 2023, Scalercio et al. 2025]. In question answering, evaluation often combines lexical overlap, semantic similarity, and criteria such as factuality, relevance, completeness, and clarity [Srivastava and Memon 2024, Lavie and Agarwal 2007, Sellam et al. 2020, Zheng et al. 2024]. In controllable text generation, especially financial text generation, prior work highlights adherence to source information, usefulness, analytical depth, and domain-specific adequacy [Yan 2022, Kim et al. 2024, Assis et al. 2024b, Assis et al. 2025c].

Taken together, prior work shows that LLM evaluation has advanced considerably, yet important gaps remain. Existing benchmarks support controlled comparison, but provide limited insight into how models generate open-ended, context-sensitive, and domain-adequate text in PT-BR. Portuguese-oriented NLG research offers important foundations for the direction, yet these efforts are usually organized around individual tasks, datasets, metrics, and criteria. Consequently, they do not yet provide a unified perspective for comparing LLM-powered solutions across multiple generative tasks in Portuguese.

The dissertation addresses these gaps by evaluating LLM solutions across summarization, text simplification, generative question answering, and MF-CTG, using these tasks as complementary perspectives on meaning preservation, language manipulation, knowledge use, factual grounding, controlled generation, and domain-specific adequacy [Souza et al. 2024, Scalercio et al. 2025, Paschoal et al. 2021, Assis et al. 2024b]. It further extends Portuguese evaluation efforts by comparing Portuguese-oriented, multilingual open-weight, and proprietary models through a protocol that combines automatic metrics, LLM-as-a-judge assessment, human evaluation, and computational efficiency [Zheng et al. 2024, Budennyy et al. 2023]. In doing so, the dissertation contributes with a broader view of LLM performance in open-ended text generation scenarios.

3. Materials and Methods

The methodological design of the dissertation combines three components: review-driven model selection, a multidimensional evaluation framework, and an experimental setup covering four PT-BR text generation tasks. Together, they address a central gap in LLM evaluation: the lack of systematic, generation-oriented assessment for Brazilian Portuguese across Portuguese-oriented, multilingual open-weight, and proprietary models.

3.1. Model Review and Selection

The first methodological step consisted of reviewing the Brazilian ecosystem of generative LLMs, referred to in the dissertation as the *LLM Fauna*. As summarized in Figure 1, this ecosystem has evolved from early Portuguese language models, such as GPorTuguese-2 [Guillou 2020] and PTT5 [Carmo et al. 2020], to more recent families such as Sabiá, Bode, Boto, Cabra, Tucano, and Periquito [Pires et al. 2023, Paiola et al. 2025, Santa Brígida 2024b, BotBot AI 2024, Corrêa et al. 2024b, Gibaut 2023]. The review shows that most Brazilian initiatives rely on adapting open-weight foreign architectures to Portuguese, usually through instruction tuning or continued training, while

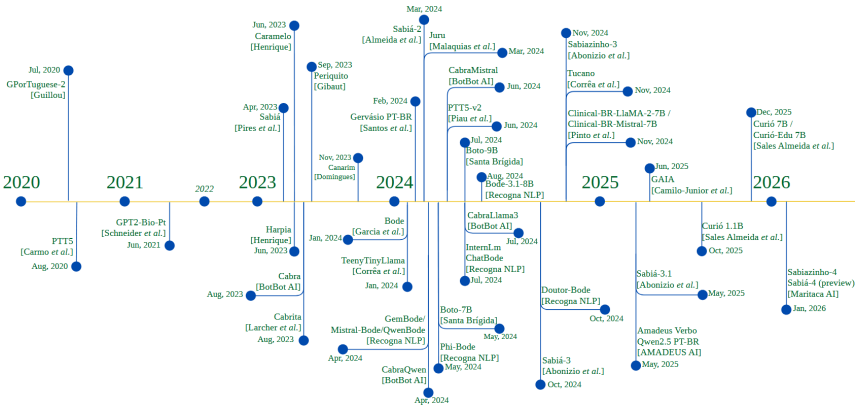


Figure 1. Timeline of initiatives for creating LLMs for PT-BR.

fewer efforts involve pre-training from scratch. It also highlights the role of open-source resources in sustaining local development, as well as the structural constraints that limit the scale of models developed for Brazil [Joshi et al. 2020, Lehdonvirta et al. 2025, Longpre et al. 2025].

This review informed the model selection strategy. The Open Portuguese LLM Leaderboard was used as a starting point because it provides a consolidated reference for Portuguese-capable models [Garcia 2024]. However, since the leaderboard is mostly based on discriminative and closed-form tasks, the dissertation uses it not as a final measure of generative quality, but as a principled entry point for selecting models to be evaluated in open-ended generation. The selection prioritized documented models explicitly adapted for Portuguese, accessible through remote APIs or local inference, and diverse enough to avoid overrepresenting a single model family.

Based on these criteria, the dissertation evaluates six Brazilian or Portuguese-oriented models across all four tasks: Sabiá-3 [Abonizio et al. 2024], Boto-9B [Santa Brígida 2024b], Bode-3.1-8B [Recogna-NLP 2024], CabraLlama-8B [BotBot AI 2024], Tucano [Corrêa et al. 2024b], and Periquito [Gibaut 2023]. GPT-4o [OpenAI et al. 2024b] is included as a proprietary reference. For the financial-domain task, the comparison is expanded with multilingual and proprietary models selected for their scale and relevance to analytical text generation: Mistral-7B [Jiang et al. 2023], Mistral-3-24B [Mistral-AI 2025], Gemma-3 12B and 27B [Gemma Team 2025], Llama-3-8B [Meta-AI 2024], Llama-4-Scout [Meta-AI 2025], and Command-R+ [Cohere AI 2024].

Model Selection Takeaways

The dissertation contributes to the Brazilian generative AI landscape by documenting the current stage of PT-BR LLM development. It shows that Brazil has made substantial progress, but also that most initiatives still rely on adapting foreign open-weight architectures to Portuguese. This finding highlights the importance of open-source ecosystems for local innovation, while also exposing the structural challenges faced by countries with fewer computational and financial resources. Thus, the dissertation connects model evaluation to the broader conditions shaping generative AI development in Brazil.

3.2. Evaluation Framework Designed in the Work

The dissertation designs a comprehensive evaluation framework to assess LLMs. Rather than treating evaluation as the application of a fixed set of metrics, the framework is orga-

nized around the main dimensions required to judge open-ended text generation in PT-BR. These dimensions include lexical overlap, semantic similarity, factual grounding, linguistic quality, communicative adequacy, human judgment, computational efficiency, and environmental impact. This design addresses a limitation of existing Portuguese evaluation efforts, which provide valuable task-specific foundations but often remain fragmented across isolated datasets and criteria[3].

The automatic layer was defined through a review of NLG evaluation practices, considering which metrics were meaningful for Portuguese text generation and for each task under analysis. Lexical metrics, including BLEU, ROUGE, METEOR, F1, and SARI, capture surface overlap, answer matching, and task-specific transformations [Papineni et al. 2002, Lin 2004, Lavie and Agarwal 2007, Xu et al. 2016]. Semantic metrics, such as BERTScore and BLANC, assess meaning preservation beyond exact wording [Zhang et al. 2020, Vasilyev et al. 2020]. Factuality metrics, including CTC-Score and AlignScore, examine whether generated outputs remain grounded in the source material and preserve event-level consistency [Deng et al. 2021, Zha et al. 2023]. NILC-Metrix indicators add a Portuguese-specific linguistic layer, covering textual properties that are not captured by standard NLG metrics [Leal et al. 2024].

These automatic signals are complemented by judgment-based evaluation. The dissertation draws on linguistic literature and task-specific evaluation practices to define natural-language criteria for LLM-as-a-judge assessment, covering dimensions such as adequacy, fluency, clarity, completeness, insightfulness, and domain-specific usefulness [Liu et al. 2023, Zheng et al. 2024]. Human assessment is then used as an additional adjudication layer[4], allowing qualitative inspection of cases in which automatic metrics and LLM-based judgments may fail to capture subtle aspects of generation quality. Finally, the framework incorporates inference time, estimated energy consumption, and equivalent CO_2 emissions using eco2AI [Budennyy et al. 2023].

Evaluation Framework Takeaways

The evaluation framework was designed from a review of NLG evaluation, linguistic quality, and task-specific assessment. Based on this review, the dissertation defines evaluation criteria and operationalizes them through automatic metrics, Portuguese linguistic indicators, and LLM-as-a-judge rubrics. These signals are complemented by a human review layer, allowing qualitative nuances to be examined beyond numerical scores.

3.3. Experimental Setup and Datasets

The experimental setup instantiates the framework across the four PT-BR text generation tasks. Summarization is evaluated with **RecognaSumm** [Paiola et al. 2024a], a Portuguese news summarization dataset covering diverse journalistic domains. Text simplification uses **PorSimplesSent** [Leal et al. 2018], adapted by treating the more complex sentence as input and the simpler aligned sentence as reference. Generative question answering is evaluated with the Portuguese portion of **Pirá 2.0** [Paschoal et al. 2021, Pirozelli et al. 2024], a specialized dataset centered on ocean-related topics, the Brazilian coast, and climate change. Together, these datasets allow the dissertation to examine content selection, meaning preservation, language manipulation, knowledge use, and factual response generation in PT-BR.

[3]Prompts, criteria, and rubrics are available in the dissertation and omitted here due to space constraints.
[4]In Portuguese, this procedure is referred to in the dissertation as *Humano-como-desembargador*.

For MF-CTG, the dissertation proposes and uses **LLM4FinComm** [Assis et al. 2025c], a PT-BR financial text generation dataset designed to reflect the Brazilian market. The dataset addresses the scarcity of financial NLG resources in Portuguese by pairing official corporate disclosures, specifically Material Facts collected from CVM sources, with expert human-written financial commentaries. In this task, the disclosure functions as the control document, while the expert commentary serves as the reference for evaluating the generated financial analysis.

The construction of LLM4FinComm followed a three-stage pipeline: collecting expert commentaries, retrieving potentially related corporate disclosures, and aligning commentaries with disclosures. The alignment procedure combined embedding-based filtering with LLM-based verification with majority voting. The final dataset contains 88 disclosure–commentary pairs, covering 41 companies and 24 sectors, thereby providing a specialized benchmark for assessing domain-specific generation in PT-BR finance.

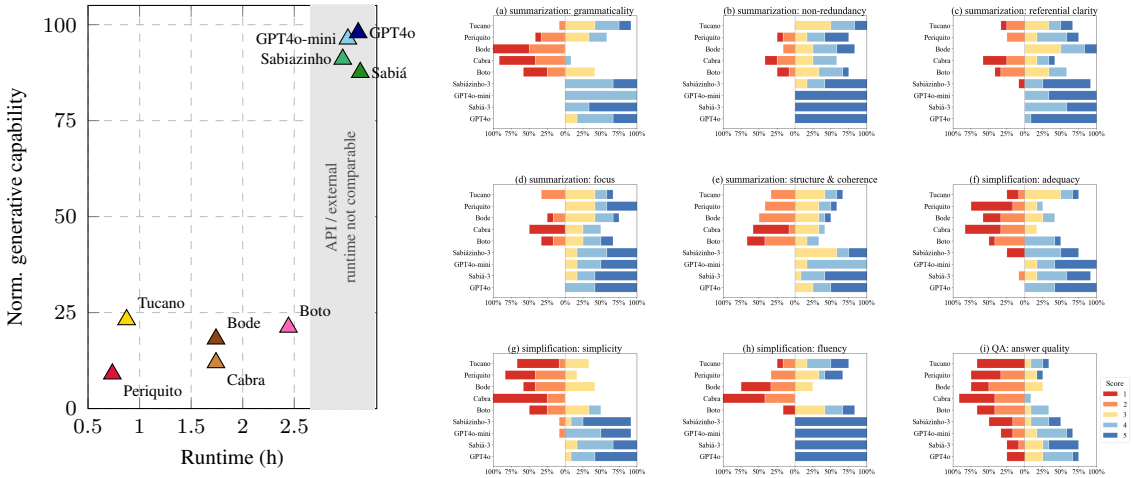
Methodological Contribution

The dissertation's methodology provides a structured framework for evaluating PT-BR LLMs in generative settings. It combines a review-driven selection of models, task-specific datasets, multidimensional evaluation criteria, and efficiency and sustainability measurements. By bringing these elements together, the dissertation offers a broader basis for analyzing generative quality, contextual adequacy, and practical viability in Brazilian Portuguese text generation.

4. Relevant Results and Contributions (Summary)

This section summarizes the dissertation's main findings and related publications.

4.1. Experimental Results



(a) Gen. perf. vs. time **(b) Human evaluation**

Figure 2. Comparison between automatic metrics and human evaluation. Subfigure 2a summarizes generative performance by independently min–max normalizing each metric and averaging the normalized scores, with cost proxied by inference time. Subfigure 2b reports human evaluation results.

The results in Figure 2 show that *Sabiá* models are generally close to the *GPT* family in automatic generative metrics, indicating strong PT-BR performance among proprietary Portuguese-oriented models. Among locally executed open-weight Brazilian models, *Tucano* stands out as the most relevant Green alternative, combining small

Table 1. Rank comparison across OPLL and downstream generative tasks.

Model	Overall		Generative ranks				OPLL
	rank-sum	avg-rank	Summ.	Simp.	QA	MF-CTG	rank
GPT-4o (#unk params)	5	1	1	1	1	1	1
Sabiá-3 (#unk params)	10	2	2	2	2	2	2
Boto (9.24B params)	21	4	4	4	5	5	3
Bode (8.03B params)	22	4	5	6	4	3	4
Cabra (8.03B params)	25	5	6	7	3	4	5
Tucano (2.44B params)	25	5	3	3	6	7	6
Periquito (3.55B params)	32	6	7	5	7	6	7

scale, lower computational cost, and competitive generative capability. *Boto* also reaches strong results in some settings, but it is larger and more expensive to run. In contrast, *Periquito*, *Bode*, and *Cabra* show that parameter count alone does not explain generative performance, pointing to the role of training data, adaptation, and model design.

Human evaluation broadly confirms the main tendencies observed in the automatic results. This suggests that the combination of linguistic, semantic, and factual metrics can approximate human perception at a general level, especially when identifying broad performance groups. Still, human assessment captures nuances that automatic metrics may miss, particularly among models with similar scores. The adjudication of the LLM-as-a-judge analyses also calls for caution: although useful as an additional evaluation signal, these judgments can be over-punitive and vary across instances. These differences indicate that multidimensional automatic evaluation is useful and necessary, but human judgment remains necessary for open-ended PT-BR generation.

Evaluation Takeaway

The dissertation concludes that combining multiple evaluation metrics helps approximate human perception more reliably than isolated scores. However, this combination is not enough to replace human evaluation, since qualitative nuances, task-specific expectations, and subtle preferences are still captured only by human judgment.

Finally, the Table 1 further reinforces that discriminative benchmark performance does not necessarily reflect generative performance. *Tucano*, for instance, performs poorly in closed-form rankings (OPLL), but becomes one of the most relevant open-weight models in the generative setting when efficiency, openness, and generative capability are considered together.

Green & Golden Synthesis

The dissertation shows that strong *Golden* performance in PT-BR text generation is achieved mainly by the *GPT* and *Sabiá* families. At the same time, it identifies *Tucano* as a promising *Green* alternative, offering a favorable balance between openness, efficiency, and generative capability. Overall, the results suggest that generative evaluation does not always follow the same patterns observed in discriminative benchmarks. Models with weaker closed-form rankings may still become relevant when open-ended generation, efficiency, and practical usability are considered together.

4.2. Related Publications and Collaborations

Table 2 summarizes publications and collaborations connected to the research conducted in the dissertation. Among the most directly related works, [Assis et al. 2025d] consolidates its broader evaluation perspective, while [Assis et al. 2025c] explores the proposed financial text generation task. Other works address related aspects in finance,

ESG-oriented question answering, sustainability-aware LLM development, and domain-specific analysis [Assis et al. 2025e, Assis et al. 2026b, Assis et al. 2026a]. Additional publications examine LLM prompting and generative capabilities in complementary settings [Assis et al. 2025a, Assis et al. 2025b], while [Assis et al. 2024a] provides an earlier basis for combining generative quality and computational impact. Beyond these direct connections, the table also reflects broader collaborations on language technologies, including diplomatic essay assessment [Cavalcanti et al. 2025], metaphor processing [Stellet et al. 2025], and Portuguese-centered computational tasks [Masid et al. 2026, Moura et al. 2026, Amorim et al. 2025b].

Table 2. Summary of publications related to the dissertation.

Publication	Venue	Year	Qualis	h5-index	h5 Stratum	Observation
[Assis et al. 2026b]	LREC	2026	A1	68	↑ 12,5%	–
[Assis et al. 2026a]	ICLR FinAI WS	2026	–	–	–	Best Poster Runner-Up Award
[Assis et al. 2025d]	JBCS	2025	A3	–	–	–
[Assis et al. 2025c]	ACM ICAIF	2025	–	30	↑ 25,0%	–
[Assis et al. 2025e]	ACM ICAIF	2025	–	30	↑ 25,0%	–
[Assis et al. 2025a]	ENIAC	2025	A4	9	↑ 75,0%	–
[Assis et al. 2025b]	SemEval	2025	A1	40	↑ 12,5%	–
[Assis et al. 2024a]	STIL	2024	B1	8	↑ 87,5%	–
[Ferreira et al. 2026]	ICLR FinAI WS	2026	–	–	–	–
[Masid et al. 2026]	PROPOR	2026	A4	13	↑ 62,5%	–
[Moura et al. 2026]	PROPOR	2026	A4	13	↑ 62,5%	–
[Stellet et al. 2025]	STIL	2025	B1	8	↑ 87,5%	–
[Cavalcanti et al. 2025]	STIL	2025	B1	8	↑ 87,5%	–
[Nepomuceno et al. 2025]	BRACIS	2025	A3	18	↑ 50,0%	–
[Paiva et al. 2025]	SBBD	2025	A4	8	↑ 87,5%	–
[Amorim et al. 2025a]	FUSION	2025	A3	23	↑ 37,5%	–
[Amorim et al. 2025b]	ACM SAC	2025	A1	37	↑ 12,5%	Best Paper Award
[Dutra et al. 2025]	NTCIR	2025	–	–	–	–
[Martinho et al. 2025]	KDMiLe	2025	B2	8	↑ 87,5%	Best Paper Award
[Romero et al. 2025]	KDMiLe	2025	B2	8	↑ 87,5%	Third Best Paper Award
[Amorim et al. 2024]	KDMiLe	2024	B2	8	↑ 87,5%	Second Best Paper Award

5. Final Remarks

The dissertation demonstrates that the evaluation of LLM-powered solutions for Brazilian Portuguese cannot be adequately captured by closed-form and predominantly discriminative benchmarks alone. Across summarization, text simplification, generative question answering, and financial-domain controllable generation, the results reveal substantial variation across tasks, models, and evaluation criteria. Proprietary systems, especially the *GPT* and *Sabiá* families, achieve the strongest overall generative performance, while open-weight models show more task-dependent behavior. Among them, *Tucano* emerges as a particularly relevant alternative by combining competitive generative capability with lower computational cost and environmental impact, whereas models such as *Bode* remain stronger in more knowledge-intensive settings.

More broadly, the dissertation contributes a multidimensional evaluation perspective for PT-BR generative AI, combining automatic metrics, Portuguese linguistic indicators, LLM-as-a-judge protocols, human assessment, and computational impact measurements. The findings reinforce that generative quality depends on dimensions that no single metric can fully capture, especially when contextual adequacy, usefulness, factual grounding, and human preference are considered. By connecting performance, efficiency, openness, and environmental responsibility, the dissertation advances a more situated and responsible path for evaluating LLMs in Brazilian Portuguese, *on the trail of a Green & Golden Generative AI*.

References

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabiá-3 Technical Report.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Étienne Goffinet, Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., and Penedo, G. (2023). The falcon series of open language models.

Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A New Generation of Portuguese Large Language Models.

Almeida, T. S., Pires, R., Abonizio, H., Nogueira, R., and Pedrini, H. (2025). PoETa v2: Toward more robust evaluation of large language models in portuguese.

Amorim, A., Assis, G., de Oliveira, D., and Paes, A. (2025a). Exploring language model fusion to improve generalization in portuguese hate speech detection. In *2025 28th International Conference on Information Fusion (FUSION)*, pages 1–8.

Amorim, A., Assis, G., Oliveira, D., and Paes, A. (2024). Multiple voices, greater power: A strategy for combining language models to combat hate speech. In *Anais do XII Symposium on Knowledge Discovery, Mining and Learning*, pages 121–128, Porto Alegre, RS, Brasil. SBC.

Amorim, A., Assis, G., Ribeiro, L., de Oliveira, D., and Paes, A. (2025b). Language flavors in the lusophone world: A BERT-based social media study of portuguese in brazil, portugal, and mozambique. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, SAC '25, page 1927–1934, New York, NY, USA. Association for Computing Machinery.

Assis, G., Alvarenga, L., Moraes, J., Azevedo, L., and Paes, A. (2025a). Investigating the Emergent Reasoning Abilities of Large Language Models in Sentiment Analysis of Codified Music. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*, pages 415–426, Porto Alegre, RS, Brasil. SBC.

Assis, G., de Azevedo, L., de Moraes, J. V., Alvarenga, L., and Paes, A. (2025b). Irapuarani at SemEval-2025 task 10: Evaluating strategies combining small and large language models for multilingual narrative detection. In Rosenthal, S., Rosá, A., Ghosh, D., and Zampieri, M., editors, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 38–48, Vienna, Austria. Association for Computational Linguistics.

Assis, G., Dutra, H., Vianna, D., Meira, Wagner, J., da Silva, A. S., and Paes, A. (2025c). Language Models for Automated Market Commentary from Corporate Disclosures. In *Proceedings of the 6th ACM International Conference on AI in Finance*, ICAIF '25, page 727–735, New York, NY, USA. Association for Computing Machinery.

Assis, G., Freitas, C., and Paes, A. (2025d). 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.

Assis, G., Surica, A., Kroll, P., Aires, G., Rabbani, D., Bollis, E., Pellicer, L., and Paes, A. (2026a). Developing an ESG-oriented large language model through ESG practices. In *The 2nd Workshop on Advances in Financial AI Workshop: Towards Agentic and Responsible Systems*.

Assis, G., Surica, A., Kroll, P., Mendes, G. A., Rabbani, D., Bollis, E., Pellicer, L. F. A. O., and Paes, A. (2026b). ESG-QA: Building a dataset for question answering on environmental, social, and governance pillars. In Piperidis, S., Bel, N., van den Heuvel, H., Ide, N., Krek, S., and Toral, A., editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 5377–5388, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Assis, G., Surica, A., Kroll, P., Munhoz, C., Rabbani, D., Bollis, E., Pellicer, L., and Paes, A. (2025e). On the potential of tool-enhanced small language models to match large models in finance. In *Proceedings of the 6th ACM International Conference on AI in Finance*, ICAIF '25, page 847–855, New York, NY, USA. Association for Computing Machinery.

Assis, G., Vasconcelos, A., de Azevedo, L., Ferro, M., and Paes, A. (2024a). Modestos e Sustentáveis: O Ajuste Eficiente Beneficia Modelos de Língua de Menor Escala em Português? In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 97–107, Porto Alegre, RS, Brasil. SBC.

Assis, G., Vianna, D., Pappa, G. L., Plastino, A., Meira Jr, W., da Silva, A. S., and Paes, A. (2024b). Analysis of material facts on financial assets: A generative AI approach. In Chen, C.-C., Liu, X., Hahn, U., Nourbakhsh, A., Ma, Z., Smiley, C., Hoste, V., Das, S. R., Li, M., Ghassemi, M., Huang, H.-H., Takamura, H., and Chen, H.-H., editors, *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 103–118, Torino, Italia. Association for Computational Linguistics.

BotBot AI (2024). Cabra Llama-3 8B. Available at https://huggingface.co/botbot-ai/CabraLlama3-8b.

Budennyy, S. A., Lazarev, V. D., Zakharenko, N. N., Korovin, A. N., Plosskaya, O. A., Dimitrov, D. V., Akhripkin, V. S., Pavlov, I. V., Oseledets, I. V., Barsola, I. S., Egorov, I. V., Kosterina, A. A., and Zhukov, L. E. (2023). eco2AI: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI. *Doklady Mathematics*.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data.

Cavalcanti, R., Casini, G., Assis, G., Real, L., Vianna, D., Mann, P., and Paes, A. (2025). Diplomatrix-BR: Um corpus paralelo de redações de autoria humana e de llms no concurso de diplomacia brasileira. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 192–205, Porto Alegre, RS, Brasil. SBC.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage,

M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. (2021). Evaluating large language models trained on code.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. I., Gonzalez, J. E., and Stoica, I. (2024). Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Cohere AI (2024). Introducing command r+: A scalable LLM built for business. https://cohere.com/blog/command-r-plus-microsoft-azure.

Corrêa, N. K., Falk, S., Fatimah, S., Sen, A., and De Oliveira, N. (2024a). TeenyTinyLlama: Open-source tiny language models trained in Brazilian Portuguese. *Machine Learning with Applications*, 16:100558.

Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2024b). Tucano: Advancing neural text generation for portuguese.

Dang, J., Singh, S., D'souza, D., Ahmadian, A., Salamanca, A., Smith, M., Peppin, A., Hong, S., Govindassamy, M., Zhao, T., Kublik, S., Amer, M., Aryabumi, V., Campos, J. A., Tan, Y.-C., Kocmi, T., Strub, F., Grinsztajn, N., Flet-Berliac, Y., Locatelli, A., Lin, H., Talupuru, D., Venkitesh, B., Cairuz, D., Yang, B., Chung, T., Ko, W.-Y., Shi, S. S., Shukayev, A., Bae, S., Piktus, A., Castagné, R., Cruz-Salinas, F., Kim, E., Crawhall-Stein, L., Morisot, A., Roy, S., Blunsom, P., Zhang, I., Gomez, A., Frosst, N., Fadaee, M., Ermis, B., Üstün, A., and Hooker, S. (2024). Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Deng, M., Tan, B., Liu, Z., Xing, E., and Hu, Z. (2021). Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. ACL.

Dutra, H., Martinho, L., Assis, G., Carvalho, J., and Paes, A. (2025). Aidavance at the ntcir-18 finarg-2 task: Making the most of small language models. NII Institutional Repository. This paper presents AIDAVANCE's approach to Subtask 2 (Detection of Argument Temporal References) of the NTCIR-18 FinArg-2 Task. We explored different classification strategies, including direct multi-class classification, a hierarchical cascade approach that first identifies the presence of a temporal reference before further categorization, and an LLM-based argument rewriting method. Our best model, a fine-tuned mDeBERTa using the multi-class approach, ranked fourth overall, achieving a Micro-F1 score of 0.6905 and a Macro-F1 score of 0.6711. Our findings reinforce that fine-tuning smaller encoder models remains an effective strategy for specialized classification tasks, even outperforming state-of-the-art LLMs.

Ferreira, T. C., Assis, G., and Paes, A. (2026). A financial agent for fundamental analysis: an empirical investigation in the brazilian stock market. In *The 2nd Workshop on Advances in Financial AI Workshop: Towards Agentic and Responsible Systems*.

Garcia, E. A. S. (2024). Open portuguese llm leaderboard. https://huggingface.co/spaces/eduagarcia/open_pt_llm_leaderboard.

Gemma Team (2025). Gemma 3 technical report.

Gibaut, W. (2023). Periquito-3B. Available at https://huggingface.co/wandgibaut/periquito-3B.

Government of Brazil (2025). Brazilian Artificial Intelligence Plan (PBIA) 2024–2028: AI for the Good of All. Available at: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2025/06/publicada-versao-final-do-plano-brasileiro-de-inteligencia-artificial-sob-coordenacao-do-mcti. Accessed on June 29, 2025.

Guha, N., Nyarko, J., Ho, D. E., Re, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J. J., Choi, J. H., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., and Li, Z. (2023). Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guillou, P. (2020). GPorTuguese-2: a Language Model for Portuguese Text Generation.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.*, 21(1).

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Hovy, D. and Spruit, S. L. (2016). The Social Impact of Natural Language Processing. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7B. *CoRR*, abs/2310.06825.

Jorge, G. A. Z., Bezerra, D. A., Xavier, C. C., and Pardo, T. A. S. (2025). Multilingual extractive summarization: Investigating state-of-the-art methods for english and brazilian portuguese. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 212–223, Cham. Springer Nature Switzerland.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Kew, T., Chi, A., Vásquez-Rodríguez, L., Agrawal, S., Aumiller, D., Alva-Manchego, F., and Shardlow, M. (2023). BLESS: Benchmarking large language models on sentence simplification. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

Kim, A. G., Kim, D. S., Muhn, M., Nikolaev, V. V., and So, E. C. (2024). AI, Investment Decisions, and Inequality. *Chicago Booth Accounting Research Center Research Paper, Fama-Miller Working Paper, MIT Sloan Research Paper*. Available at SSRN: https://ssrn.com/abstract=5075727 or http://dx.doi.org/10.2139/ssrn.5075727.

Lavie, A. and Agarwal, A. (2007). Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

Leal, S. E., Duran, M. S., and Aluíso, S. M. (2018). A nontrivial sentence corpus for the task of sentence readability assessment in portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 401–413, Santa Fe, New Mexico, USA.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2024). NILC-Metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Language Resources and Evaluation*, 58(1):73–110.

Lehdonvirta, V., Wú, B., and Hawkins, Z. (2025). *Compute North vs. Compute South: The Uneven Possibilities of Compute-Based AI Governance Around the Globe*, page 828–838. AAAI Press.

Li, P., Yang, J., Islam, M. A., and Ren, S. (2023). Making AI less "Thirsty': Uncovering and Addressing the Secret Water Footprint of AI models.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, S., Hilton, J., and Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Longpre, S., Singh, N., Cherep, M., Tiwary, K., Materzynska, J., Brannon, W., Mahari, R., Obeng-Marnu, N., Dey, M., Hamdy, M., Saxena, N., Anis, A. M., Alghamdi, E. A., Chien, V. M., Yin, D., Qian, K., LI, Y., Liang, M., Dinh, A., Mohanty, S., Mataciunas, D., South, T., Zhang, J., Lee, A. N., Lund, C. S., Klamm, C., Sileo, D., Misra, D., Shippole, E., Klyman, K., Miranda, L. J. V., Muennighoff, N., Ye, S., Kim, S., Gupta,

V., Sharma, V., Zhou, X., Xiong, C., Villa, L., Biderman, S., Pentland, A., Hooker, S., and Kabbara, J. (2025). Bridging the Data Provenance Gap Across Text, Speech, and Video. In *The Thirteenth International Conference on Learning Representations*.

Luccioni, S., Gamazaychikov, B., Strubell, E., Hooker, S., Jernite, Y., Wu, C.-J., and Mitchell, M. (2025). Ai energy score.

Malaquias-Junior, R., Pires, R., Romero, R., and Nogueira, R. (2024). Juru: Legal Brazilian Large Language Model from Reputable Sources.

Martinho, L., Dutra, H., Assis, G., Carvalho, J., and Paes, A. (2025). When modernity enhances tradition and specialization: A BERT ensemble for temporal financial argument detection. In *Anais do XIII Symposium on Knowledge Discovery, Mining and Learning*, pages 17–24, Porto Alegre, RS, Brasil. SBC.

Masid, M. R., Assis, G., Vianna, D., Paes, A., and Silva, A. S. d. (2026). Auditing the evaluators: How far can automatic evaluation go in assessing Portuguese financial texts? In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 222–233, Salvador, Brazil. Association for Computational Linguistics.

Meta-AI (2024). The Llama 3 Herd of Models.

Meta-AI (2025). The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.

Mistral-AI (2025). Mistral Small 3.1.

Moura, G. V. d., Assis, G., and Paes, A. (2026). Sintomas linguísticos: Geração aumentada por recuperação e raciocínio em LLMs sob a variação português-inglês em contextos médicos. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 467–477, Salvador, Brazil. Association for Computational Linguistics.

Nepomuceno, J., Rottschaefer, E., Assis, G., Ferreira, T., and Paes, A. (2025). Agentic AI meets fundamentalist financial analysis: Limits and possibilities. In *Anais da XXXV Brazilian Conference on Intelligent Systems*, pages 131–145, Porto Alegre, RS, Brasil. SBC.

OpenAI, :, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A., Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C. M., de Bourcy, C., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., and et al. (2024a). Openai o1 system card.

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A.,

Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A. T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., and et al. (2024b). Gpt-4o system card.

Paiola, P. H., Garcia, G. L., Correia, J. V. M., Manesco, J. R. R., Garcia, A. L. A., and Papa, J. P. (2025). The Bode Family of Large Language Models: Investigating the Frontiers of LLMs in Brazilian Portuguese. *Journal of the Brazilian Computer Society*, 31(1):918–939.

Paiola, P. H., Garcia, G. L., Jodas, D. S., Correia, J. V. M., Sugi, L. A., and Papa, J. P. (2024a). RecognaSumm: A novel Brazilian summarization dataset. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 575–579, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Paiola, P. H., Garcia, G. L., Manesco, J. R. R., Roder, M., Rodrigues, D., and Papa, J. P. (2024b). Adapting LLMs for the Medical Domain in Portuguese: A Study on Fine-Tuning and Model Evaluation.

Paiva, L., Assis, G., Amorim, A., Dias, L. G., Paes, A., and de Oliveira, D. (2025). Domínio delimitado, Ódio exposto: O uso de prompts para identificação de discurso de Ódio online com LLMs. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 493–506, Porto Alegre, RS, Brasil. SBC.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Paschoal, A. F. A., Pirozelli, P., Freire, V., Delgado, K. V., Peres, S. M., José, M. M., Nakasato, F., Oliveira, A. S., Brandão, A. A. F., Costa, A. H. R., and Cozman, F. G. (2021). Pirá: A Bilingual Portuguese-English Dataset for Question-Answering about the Ocean. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, CIKM '21, page 4544–4553. ACM.

Pawar, S., Park, J., Jin, J., Arora, A., Myung, J., Yadav, S., Haznitrama, F. G., Song, I., Oh, A., and Augenstein, I. (2025). Survey of Cultural Awareness in Language Models: Text and Beyond. *Computational Linguistics*, pages 1–96.

Pereira, F. V., Frazão, A., and Moreira, V. P. (2025). Automatic text simplification for the legal domain in brazilian portuguese. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 31–45, Cham. Springer Nature Switzerland.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland.

Pirozelli, P., José, M. M., Silveira, I., Nakasato, F., Peres, S. M., Brandão, A. A. F., Costa, A. H. R., and Cozman, F. G. (2024). Benchmarks for pirá 2.0, a reading comprehension dataset about the ocean, the brazilian coast, and climate change. *Data Intelligence*, 6(1):29–63.

Recogna-NLP (2024). Bode-3.1-8B-Instruct. Available at: https://huggingface.co/recogna-nlp/Bode-3.1-8B-Instruct-full. Accessed: 5 March 2025.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. (2024). GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Romero, V., Assis, G., Carvalho, J., Mann, P., and Paes, A. (2025). Opportunities vs. risks: Exploring automatic annotation of financial polarity biases via large language models. In *Anais do XIII Symposium on Knowledge Discovery, Mining and Learning*, pages 1–8, Porto Alegre, RS, Brasil. SBC.

Santa Brígida, L. (2024a). Boto-7B. Available at https://huggingface.co/lucianosb/boto-7B.

Santa Brígida, L. (2024b). Boto-9B. Available at https://huggingface.co/lucianosb/boto-9B.

Scalercio, A., de Souza, E. A., Finatto, M. J. B., and Paes, A. (2025). Evaluating LLMs for Portuguese Sentence Simplification with Linguistic Insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Vienna, Austria. Association for Computational Linguistics. To appear.

Schneider, E. T. R., de Souza, J. V. A., Gumiel, Y. B., Moro, C., and Paraiso, E. C. (2021). A GPT-2 Language Model for Biomedical Texts in Portuguese. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 474–479.

Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngui, J. G., Vila-Suero, D., Limkonchotiwat, P., Marchisio, K., Leong, W. Q., Susanto, Y., Ng, R., Longpre, S., Ko, W.-Y., Ruder, S., Smith, M., Bosselut, A., Oh, A., Martins, A. F. T., Choshen, L., Ippolito, D., Ferrante, E., Fadaee, M., Ermis, B., and Hooker, S. (2025). Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., and Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Souza, J. W. d. C., Cardoso, P. C. F., and Paixão, C. A. (2024). Sumarização automática. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 24. BPLN, 3 edition.

Srivastava, A. and Memon, A. (2024). Toward Robust Evaluation: A Comprehensive Taxonomy of Datasets and Metrics for Open Domain Question Answering in the Era of Large Language Models. *IEEE Access*, 12:117483–117503.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A. J., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A. M., La, A., Lampinen, A. K., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., and et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Stellet, L., Leite, I., Assis, G., and Paes, A. (2025). Meta4BR: Avaliando a fidelidade metafórica em traduções de metáforas para o português por llms. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 441–454, Porto Alegre, RS, Brasil. SBC.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Vasilyev, O., Dharnidharka, V., and Bohannon, J. (2020). Fill in the BLANC: Human-free quality estimation of document summaries. In Eger, S., Gao, Y., Peyrard, M., Zhao, W., and Hovy, E., editors, *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., Xu, Y., Kang, H., Kuang, Z., Yuan, C., Yang, K., Luo, Z., Zhang, T., Liu, Z., XIONG, G., Deng, Z., Jiang, Y., Yao, Z., Li, H., Yu, Y., Hu, G., Jiajia, H., Liu, X.-Y., Lopez-Lira, A., Wang, B., Lai, Y., Wang, H., Peng, M., Ananiadou, S., and Huang, J. (2024). Finben: An holistic financial benchmark for large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing Statistical Machine Translation for Text Simplification. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 401–415.

Yan, S. (2022). Disentangled variational topic inference for topic-accurate financial report generation. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 18–24, Abu Dhabi, United Arab Emirates (Hybrid). ACL.

Yuan, R., Lin, H., Wang, Y., Tian, Z., Wu, S., Shen, T., Zhang, G., Wu, Y., Liu, C., Zhou, Z., Xue, L., Ma, Z., Liu, Q., Zheng, T., Li, Y., Ma, Y., Liang, Y., Chi, X., Liu, R., Wang, Z., Lin, C., Liu, Q., Jiang, T., Huang, W., Chen, W., Fu, J., Benetos, E., Xia, G., Dannenberg, R., Xue, W., Kang, S., and Guo, Y. (2024). ChatMusician: Understanding and generating music intrinsically with LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6252–6271, Bangkok, Thailand. Association for Computational Linguistics.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Zha, Y., Yang, Y., Li, R., and Hu, Z. (2023). AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. ACL.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., Tang, J., and Huang, M. (2024). SafetyBench: Evaluating the safety of large language models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553, Bangkok, Thailand. Association for Computational Linguistics.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Appendix A: Research By-products

The research by-products associated with the dissertation are listed below.

1. **Software packages and research artifacts.** All source code, data processing scripts, evaluation routines, prompts, and related artifacts used in the dissertation and associated papers are publicly available as open resources.
   - Evaluation framework: https://github.com/MeLLL-UFF/brfauna-gen-eval
   - LLM4FinComm dataset and methods: https://github.com/AIDA-BR/LLM4FinComm
   - Additional code related to the other associated papers: https://github.com/MeLLL-UFF
   - Additional models related to the other associated papers: https://huggingface.co/melll-uff

2. **Journal publication.**
   (a) [Assis et al. 2025d]
   Venue: Journal of the Brazilian Computer Society (JBCS), 2025; Qualis: A3; h5-index: .
   This work consolidates the central evaluation perspective of the dissertation by comparing Brazilian and Portuguese-oriented LLMs in generative tasks. It provides the conceptual basis for the broader discussion on the Brazilian *LLM Fauna*, the contrast between discriminative and generative evaluation, and the need for assessment protocols designed specifically for open-ended PT-BR generation.
   Venue link: https://journals-sol.sbc.org.br/index.php/jbcs/index.
   Paper link: https://doi.org/10.5753/jbcs.2025.5814.

3. **Conference and workshop publications.**
   (a) [Assis et al. 2025c]
   Venue: ACM International Conference on AI in Finance (ACM ICAIF), 2025; Qualis: –; h5-index: 30.
   This work directly supports the financial-domain component of the dissertation by proposing the Material Fact Controllable Text Generation task. It connects the dissertation's generative evaluation framework to financial commentary generation, using Brazilian corporate disclosures as input and expert commentaries as references.
   Venue link: https://icaif25.org/.
   Paper link: https://doi.org/10.1145/3768292.3770438.

   (b) [Assis et al. 2026b]
   Venue: Language Resources and Evaluation Conference (LREC), 2026; Qualis: A1; h5-index: 68.
   This work brings the dissertation's concern with generative question answering into the ESG domain. By proposing ESG-QA, it connects the evaluation of LLMs to environmental, social, and governance dimensions, reinforcing the dissertation's interest in domain-specific resources, sustainability-oriented language technologies, and domain-grounded QA evaluation.
   Venue link: https://lrec2026.info/.
   Paper link: https://lrec.elra.info/lrec2026-main-420.

   (c) [Assis et al. 2026a]
   Venue: ICLR Workshop on Advances in Financial AI, 2026; Qualis: –; h5-index: –.
   This work operationalizes Green & Golden concerns by developing an ESG-oriented LLM through responsible AI and ESG practices. It relates to the dissertation by applying its generative QA evaluation framework to ESG-oriented financial question answering, combining domain adaptation, efficient model development, ESG-oriented evaluation, and sustainability-aware language technology. **Best Poster Runner-Up Award.**
   Venue link: https://sites.google.com/view/iclr2026finai/.
   Paper link: https://openreview.net/forum?id=EKp4ybPdmr.

   (d) [Ferreira et al. 2026]
   Venue: ICLR Workshop on Advances in Financial AI, 2026; Qualis: –; h5-index: –.
   This work relates to the dissertation's financial AI axis by investigating LLM-based agents for fundamental analysis in the Brazilian stock market. It complements the dissertation's focus on financial text generation by exploring how LLMs can support analytical reasoning over financial information in a local market context.
   Venue link: https://sites.google.com/view/iclr2026finai/.
   Paper link: https://openreview.net/forum?id=WqI93te1bQ.

   (e) [Masid et al. 2026]
   Venue: International Conf. on Computational Processing of Portuguese (PROPOR), 2026; Qualis: A4; h5-index: 13.
   This work is closely aligned with the dissertation's evaluation framework, as it examines how automatic NLG metrics and LLM-as-a-judge methods behave when assessing Portuguese financial texts. It directly dialogues with the dissertation's discussion on the limits of automatic evaluation, the role of human-informed assessment, and the difficulty of evaluating open-ended financial language generation.
   Venue link: https://propor2026.ufba.br/.
   Paper link: https://aclanthology.org/2026.propor-1.22/.

   (f) [Moura et al. 2026]
   Venue: International Conf. on Computational Processing of Portuguese (PROPOR), 2026; Qualis: A4; h5-index: 13.
   This work relates to the dissertation's emphasis on language-specific and domain-specific behavior in LLMs through a QA setting. By examining Portuguese–English variation in medical question answering and retrieval-augmented strategies, it applies a related evaluation perspective to a specialized domain, reinforcing the importance of assessing models in realistic linguistic contexts rather than relying only on general-purpose benchmarks.

Venue link: https://propor2026.ufba.br/.
Paper link: https://aclanthology.org/2026.propor-1.46/.

(g) [Nepomuceno et al. 2025]
Venue: Brazilian Conference on Intelligent Systems (BRACIS), 2025; Qualis: A3; h5-index: 18.
This work is connected to the financial-domain axis of the dissertation by investigating the limits and possibilities of LLM-based agents for fundamental financial analysis. It contributes to the broader context in which financial LLM applications are evaluated not only by textual output quality, but also by usefulness, reliability, and analytical value in practical settings.
Venue link: https://bracis.sbc.org.br/2025/.
Paper link: https://doi.org/10.1007/978-3-032-15987-8_9.

(h) [Assis et al. 2025e]
Venue: ACM International Conference on AI in Finance (ACM ICAIF), 2025; Qualis: –; h5-index: 30.
This work relates to the dissertation's Green perspective by examining whether tool-enhanced small language models can approach larger models in financial question answering. It reinforces the dissertation's argument that model usefulness should be analyzed together with efficiency, accessibility, openness, and task-specific adequacy.
Venue link: https://icaif25.org/.
Paper link: https://doi.org/10.1145/3768292.3770409.

(i) [Assis et al. 2025a]
Venue: National Meeting on Artificial and Computational Intelligence (ENIAC), 2025; Qualis: A4; h5-index: 9.
This work explores LLM reasoning and prompting beyond conventional text tasks through sentiment analysis over codified music representations. It contributes to the broader investigation of generative and reasoning capabilities developed during the dissertation, particularly in settings where language models must operate over non-standard symbolic inputs.
Venue link: https://bracis.sbc.org.br/2025/pt/eniac-2025/.
Paper link: https://doi.org/10.5753/eniac.2025.12548.

(j) [Assis et al. 2025b]
Venue: International Workshop on Semantic Evaluation (SemEval), 2025; Qualis: A1; h5-index: 40.
This work connects to the dissertation's evaluation concerns by comparing small and large language models in a multilingual benchmark setting. It contributes to the broader discussion on how different model families behave across languages, tasks, and evaluation protocols, including scenarios involving Portuguese and multilingual narrative texts.
Venue link: https://semeval.github.io/SemEval2025/.
Paper link: https://aclanthology.org/2025.semeval-1.7/.

(k) [Stellet et al. 2025]
Venue: Brazilian Symposium in Information and Human Language Technology (STIL), 2025; Qualis: B1; h5-index: 8.
This work connects the dissertation's broader generative evaluation perspective to another task, namely LLM-based metaphor translation into Portuguese. By focusing on meaning preservation and linguistic adequacy, it relates to the dissertation's concern with quality dimensions that are not fully captured by surface-level metrics, especially in culturally and semantically sensitive language tasks.
Venue link: https://bracis.sbc.org.br/2025/pt/stil-2025/.
Paper link: https://doi.org/10.5753/stil.2025.37845.

(l) [Cavalcanti et al. 2025]
Venue: Brazilian Symposium in Information and Human Language Technology (STIL), 2025; Qualis: B1; h5-index: 8.
This work contributes to Portuguese NLG evaluation by introducing Diplomatrix-BR, a resource for comparing human- and LLM-authored essays in the Brazilian diplomatic examination context. It relates to the dissertation by expanding the evaluation of generated Portuguese text to a high-stakes, culturally situated writing scenario.
Venue link: https://bracis.sbc.org.br/2025/pt/stil-2025/.
Paper link: https://doi.org/10.5753/stil.2025.37825.

(m) [Paiva et al. 2025]
Venue: Brazilian Symposium on Databases (SBBD), 2025; Qualis: A4; h5-index: 8.
This work relates to the dissertation's interest in contextual prompting by studying how event information can improve hate speech detection with LLMs. It contributes to the broader methodological context of prompt design, context-sensitive evaluation, and the use of LLMs in socially sensitive Portuguese-language tasks.
Venue link: https://sbbd.org.br/2025/.
Paper link: http://dx.doi.org/10.5753/sbbd.2025.247279.

(n) [Amorim et al. 2025a]
Venue: International Conference on Information Fusion (FUSION), 2025; Qualis: A3; h5-index: 23.
This work explores model fusion for Portuguese hate speech detection, contributing to the broader research context on efficient and robust language model use. It also relates to the dissertation's concern with combining signals, models, and criteria to improve performance in Portuguese NLP while considering practical constraints.
Venue link: https://fusion2025.org/.
Paper link: http://dx.doi.org/10.23919/FUSION65864.2025.11124084.

(o) [Amorim et al. 2025b]
Venue: ACM/SIGAPP Symposium on Applied Computing (ACM SAC), 2025; Qualis: A1; h5-index: 37.

This work connects to the dissertation's concern with linguistic context by studying how language models process Portuguese varieties across Brazil, Portugal, and Mozambique. It reinforces the importance of considering local and regional variation when evaluating language technologies, especially in multilingual and Lusophone contexts. **Best Paper Award.**
Venue link: https://www.sigapp.org/sac/sac2025/.
Paper link: https://doi.org/10.1145/3672608.3707931.

(p) [Dutra et al. 2025]
Venue: NTCIR Conference on Evaluation of Information Access Technologies (NTCIR), 2025; Qualis: –; h5-index: –.
This work supports the dissertation's financial NLP background by examining language models for temporal financial argument classification. It contributes to the broader set of financial-language tasks that informed the dissertation's focus on domain-specific evaluation and the analysis of language models in finance.
Venue link: https://research.nii.ac.jp/ntcir/ntcir-18/.
Paper link: http://dx.doi.org/10.20736/0002002045.

(q) [Martinho et al. 2025]
Venue: Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), 2025; Qualis: B2; h5-index: 8.
This work relates to the financial AI context of the dissertation by exploring ensemble strategies for temporal financial argument detection. It reinforces the relevance of specialized financial language processing as part of the broader research environment of the dissertation and complements the analysis of model behavior in financial tasks. **Best Paper Award.**
Venue link: https://bracis.sbc.org.br/2025/pt/kdmile/.
Paper link: https://doi.org/10.5753/kdmile.2025.247741.

(r) [Romero et al. 2025]
Venue: Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), 2025; Qualis: B2; h5-index: 8.
This work connects to the dissertation's discussion of LLMs as evaluators and annotators by examining automatic annotation of financial polarity biases. It complements the dissertation's analysis of LLM-based judgment and its limitations in specialized domains, especially when interpretive financial labels are involved. **Third Best Paper Award.**
Venue link: https://bracis.sbc.org.br/2025/pt/kdmile/.
Paper link: https://doi.org/10.5753/kdmile.2025.247742.

(s) [Amorim et al. 2024]
Venue: Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), 2024; Qualis: B2; h5-index: 8.
This work contributes to the broader research environment of the dissertation by investigating combinations of language models for Portuguese hate speech detection. It relates to the dissertation's interest in robust, task-sensitive, and efficient language model use in Portuguese. **Second Best Paper Award.**
Venue link: https://bracis.sbc.org.br/2024/kdmile-2024/.
Paper link: https://doi.org/10.5753/kdmile.2024.244731.

(t) [Assis et al. 2024a]
Venue: Brazilian Symposium in Information and Human Language Technology (STIL), 2024; Qualis: B1; h5-index: 8.
This work provides the early basis for the dissertation's Green & Golden perspective by combining Portuguese summarization, efficient adaptation, generative quality, and computational impact. It anticipates several elements later integrated into the dissertation's broader evaluation framework, especially the joint analysis of output quality and resource use.
Venue link: https://bracis.sbc.org.br/2024/15th-symposium-in-information-and-human-language-technology/.
Paper link: https://aclanthology.org/2024.stil-1.5/.

Appendix B: Acknowledgments

This dissertation research received institutional, financial, and computational support from several organizations and initiatives. CAPES and the Instituto de Ciência e Tecnologia Itaú (ICTi) provided scholarship support, while FAPERJ and CNPq contributed to the research environment, infrastructure, and scientific dissemination conditions.

Additional support was provided by the Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial (INCT IAIA) and, in particular, by the AIDA group, which formed part of the academic environment in which this research was conducted. The dissertation also benefited from initiatives of the Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT TILDIAR), the Maritaca AI Credits Grant, the Cohere Labs Grant Program, and the Google Cloud Credits Program, whose resources and computational support contributed to the execution of several experiments.