Journal of the Brazilian Computer Society, 2025, 31:1, doi: 10.5753/jbcs.2025.5793

Evaluating Large Language Models for Brazilian Portuguese Sentiment Analysis: A Comparative Study of Multilingual State-of-the-Art vs. Brazilian Portuguese Fine-Tuned LLMs

André da Fonseca Schuck [**São Paulo State University** | ***andre.schuck@unesp.br***]
Gabriel Lino Garcia **São Paulo State University** | ***gabriel.lino@unesp.br***]
João Renato Ribeiro Manesco [**São Paulo State University** | ***joao.r.manesco@unesp.br***]
Pedro Henrique Paiola [**São Paulo State University** | ***pedro.paiola@unesp.br***]
João Paulo Papa [**São Paulo State University** | ***joao.papa@unesp.br***]

Department of Computing, São Paulo State University , Av. Eng. Luiz Edmundo Carrijo Coube, 14-01, Vargem Limpa, Bauru, SP, 17033-360, Brazil.

Received: 31 March 2025 • **Accepted:** 09 July 2025 • **Published:** 08 October 2025

Abstract This study presents an extensive comparative analysis of Large Language Models (LLMs) for sentiment analysis in Brazilian Portuguese texts. We evaluated 23 LLMs—comprising 13 state-of-the-art multilingual models and 10 models specifically fine-tuned for Portuguese—across 12 public annotated datasets from diverse domains, employing the in-context learning paradigm. Our findings demonstrate that large-scale models such as Claude-3.5-Sonnet, GPT-4o, DeepSeek-V3, and Sabiá-3 delivered superior results with accuracies exceeding 92%, while smaller models (7-13B parameters) also showed compelling performance with top performers achieving accuracies above 90%. Notably, linguistic specialization through fine-tuning demonstrated mixed results—significantly reducing hallucination rates for some models but not consistently yielding performance improvements across all model types. We also observed that newer model generations frequently outperformed their predecessors, and in the one dataset where traditional machine learning methods were employed by the original authors for sentiment classification, all evaluated LLMs substantially surpassed these traditional approaches. Moreover, smaller-scale models exhibited a tendency toward overgeneration despite explicit instructions. These findings contribute valuable insights to the discourse on language-specific model optimization and establish empirical benchmarks for both multilingual and Portuguese-specialized LLMs in sentiment analysis tasks.

Keywords: Large Language Models, Sentiment Analysis, Brazilian Portuguese, In-context Learning, Comparative Evaluation, Natural Language Processing, Model Fine-tuning

1 Introduction

Large Language Models (LLMs) are advanced artificial intelligence systems capable of processing and generating coherent text through extensive pre-training on massive textual corpora [Naveed *et al.*, 2024]. These models, with parameters ranging from millions to billions, comprehend and process natural language through semantic and contextual modeling, as well as the probability estimation of associated with words within a given context [Yao *et al.*, 2024].

The rapid and recent development of LLMs such as GPT-4.0 [OpenAI *et al.*, 2024b], Gemini [Gemini Team *et al.*, 2023], and LLaMA-3 [Grattafiori *et al.*, 2024] has revolutionized Natural Language Processing (NLP) [Zhao *et al.*, 2023; Yang *et al.*, 2024b]. These state-of-the-art (SOTA) models demonstrate remarkable multilingual capabilities [Touvron *et al.*, 2023b; Gemini Team *et al.*, 2023; OpenAI *et al.*, 2024b], offering potential benefits for less common languages or those with limited corpora, such as Brazilian Portuguese [Souza *et al.*, 2020].

Despite their versatility, these models exhibit limitations when applied to underrepresented languages in their pre-training corpus [Larcher *et al.*, 2023]. In an effort to address these shortcomings, numerous researchers [Souza *et al.*, 2020; Larcher *et al.*, 2023; Pires *et al.*, 2023; Garcia *et al.*, 2024] have explored techniques to enhance the performance of LLMs initially trained predominantly on English data for use in other languages. These efforts aim to specialize LLMs in Portuguese through fine-tuning on monolingual datasets [Souza *et al.*, 2020; Pires *et al.*, 2023; Garcia *et al.*, 2024] or adapting tokenization mechanisms [Larcher *et al.*, 2023].

The results have been promising, as the achieved performance is comparable to that of SOTA LLMs when evaluated on tasks in Brazilian Portuguese, while offering the additional advantage of smaller model sizes and the integration of domain-specific knowledge relevant to Brazilian culture [Pires *et al.*, 2023].

Despite the numerous advantages, efforts toward the development of LLMs specialized in Brazilian Portuguese can still be considered nascent when compared to the extensive research conducted in other languages, such as Chinese [Zeng *et al.*, 2023; Cui *et al.*, 2024; Cui and Yao, 2024; Du *et al.*, 2024; Yang *et al.*, 2024a]. Furthermore, there is a noticeable lack of studies aimed at evaluating the performance of LLMs in Brazilian Portuguese across a range of specific tasks.

In an effort to help mitigate the identified gaps, this study aims to compare the predictive capabilities of various SOTA

LLMs with models fine-tuned for Portuguese, focusing the classic NLP task of sentiment classification. Sentiment analysis, or opinion mining, identifies and quantifies subjective information in textual data [Zhao *et al.*, 2016]. A fundamental subtask of sentiment analysis is sentiment classification, which determines the overall sentiment polarity of a text. This classification can be binary (e.g., positive and negative) or multi-class (e.g., positive, negative, and neutral) [Zhang *et al.*, 2023].

To achieve this objective, we conducted an extensive evaluation of 23 LLMs: 13 SOTA generalist models and 10 Portuguese-specialized models. The study also incorporated 12 public datasets in Brazilian Portuguese, annotated for sentiment polarity, providing a rich corpus for analysis.

The LLMs were rigorously evaluated on their sentiment analysis capacity for Brazilian Portuguese texts using the in-context learning (ICL) paradigm. This empirical comparative approach aimed to elucidate the potential advantages and limitations of language-specific model fine-tuning in sentiment analysis tasks.

2 Background

2.1 Brief History of LLMs

It is notorious that the capacity and performance of LLMs have been evolving rapidly in recent years, with each new release improving upon the state-of-the-art results obtained in various comparative tests [Brown *et al.*, 2020; Gemini Team *et al.*, 2023; OpenAI *et al.*, 2024b; Reid *et al.*, 2024]. Since the disclosure of the Transformers architecture [Vaswani *et al.*, 2017], a consensus has emerged in the literature regarding structural terms for LLMs [Devlin *et al.*, 2018; Radford *et al.*, 2019; Rae *et al.*, 2022; Touvron *et al.*, 2023b; Gemini Team *et al.*, 2023], with this architecture becoming a fundamental paradigm in the field [Zhao *et al.*, 2023].

The evolution of Language Models (LM) encompasses distinct developmental phases. Initially, LMs were grounded in statistical models with supervised learning, which critically depended on domain expertise for feature engineering and the provision of appropriate inductive bias. These early models were often constrained by limited datasets, yet found widespread application in information retrieval and NLP tasks [Liu *et al.*, 2021; Zhao *et al.*, 2023].

The second phase [Zhao *et al.*, 2023] marked a significant advancement through the introduction of neural networks (Multilayer Perceptron and Recurrent Networks). These networks revolutionized the field by learning representations, embeddings, and sequential modeling autonomously, shifting the learning paradigm from feature engineering to architecture engineering [Liu *et al.*, 2021].

The third phase introduced Pre-trained Language Models (PLM), predominantly implementing the Transformers architecture and trained on extensive data with generalist objectives, such as next-word prediction or masked word identification [Qiu *et al.*, 2020]. These models learn universal and contextualized linguistic representations through pre-training on massive textual corpora, incorporating broad knowledge into their embeddings [Liu *et al.*, 2021].

While PLMs demonstrated advanced capabilities in NLP, they initially lacked the specialized knowledge required for domain-specific tasks [Qiu *et al.*, 2020; Zhao *et al.*, 2023]. This limitation led to the emergence of the fine-tuning paradigm, where PLMs are adapted for specialized tasks through the introduction and adjustment of parameters using task-specific objective functions [Liu *et al.*, 2021]. The effectiveness of fine-tuning became particularly evident following the release of BERT [Devlin *et al.*, 2018] and GPT-2.0 [Radford *et al.*, 2019], establishing itself as a consensus approach in machine learning [Qiu *et al.*, 2020; Han *et al.*, 2021].

The fourth generation, characterized as Large Scale Language Models, represents a quantum leap in model scale, both in terms of parameters (billions/trillions) and pre-training data volume [Zhao *et al.*, 2023]. This unprecedented scaling revealed remarkable emergent capabilities, defined by [Wei *et al.*, 2022] as abilities that are absent in smaller models but manifest collectively in larger ones.

A striking example of these emergent capabilities is found in the work of [Brown *et al.*, 2020], which documented the emergence of ICL in GPT-3.0 (175 billion parameters), a capability notably absent in its predecessor GPT-2.0 (1.5 billion parameters) [Radford *et al.*, 2019].

Thus, in the fourth generation, the learning paradigm no longer requires model adaptation via fine-tuning, making it possible to reformulate the underlying task through the structuring and modulation of a textual prompt (prompt engineering) to manipulate the LLM's behavior, enabling it to make predictions and return the desired output [Liu *et al.*, 2021].

3 Related Work

3.1 Benchmark of LLMs on sentiment analysis tasks

As one of the principal tasks within NLP [Zhang *et al.*, 2023; Přibáň *et al.*, 2024], sentiment classification has emerged as a significant focus in LLM research [Simmering and Huoviala, 2023; Krugmann and Hartmann, 2024; Přibáň *et al.*, 2024; Buscemi and Proverbio, 2024], driven by the innovative capabilities these models bring to the field.

Initial comparative studies between LLMs and specialized PLMs revealed promising insights. Zhong *et al.* [2023] evaluated ChatGPT against various BERT-derived [Devlin *et al.*, 2018] task-specific PLMs using the GLUEbenchmark[Wang *et al.*, 2019], which includes sentiment classification on the SST2 dataset [Socher *et al.*, 2013]. Their findings demonstrated superior performance when combining ChatGPT with prompt engineering refinement. In a more extensive study, Wang *et al.* [2023] assessed ChatGPT (*gpt-3.5-turbo-0301*) as a potential universal sentiment analyzer for 7 sentiment analysis tasks and 17 different datasets, including SST2. While showing promising results, their research indicated that LLMs still marginally trail behind refined PLMs in sentiment classification tasks.

Further advancing this line of inquiry, Krugmann and Hartmann [2024] conducted a comprehensive evaluation of SOTA LLMs (GPT-3.5 and 4.0) against high performance transfer learning-based models such as BERT, RoBERTa [Liu *et al*., 2019] and SiEBERT [Hartmann *et al*., 2023]. Their findings revealed important correlations between classification performance and factors such as the number of classes and data characteristics (source, text length, among others), ultimately positioning LLMs as powerful tools for sentiment analysis.

While these initial studies [Krugmann and Hartmann, 2024; Zhong *et al*., 2023; Wang *et al*., 2023] demonstrated promising results for non-specialized LLMs compared to specialized PLMs, they primarily focused on English-language texts. Addressing this limitation, recent research has expanded into multilingual contexts. Přibáň *et al*. [2024] conducted a comparative analysis of various classification methods, including CNN, LSTM, multilingual Transformers, and LLMs (ChatGPT and LLaMA-2), evaluating their performance on English, Czech, and French texts using datasets such as SST2 [Socher *et al*., 2013] and IMDB [Maas *et al*., 2011]. Their results demonstrated LLMs' capability to effectively process multilingual data, often matching or surpassing specialized multilingual PLMs.

Similarly, Buscemi and Proverbio [2024] evaluated SOTA LLMs in a complex multilingual scenario, analyzing 20 texts with challenging sentiment nuances across 10 languages, including Brazilian Portuguese. Their comparison of ChatGPT (versions 3.5 and 4.0), Gemini-1.0-Pro [Gemini Team *et al*., 2023], and LLaMA-2-7B [Touvron *et al*., 2023b] revealed that while ChatGPT (4.0) and Gemini-1.0-Pro excelled in ambiguous scenarios, they struggled with more sophisticated patterns like irony.

Research specifically focusing on Brazilian Portuguese remains limited but significant. Several studies have contributed to the development of specialized models and the evaluation of their performance against SOTA models using Portuguese NLP benchmarks [Souza *et al*., 2020; Pires *et al*., 2023; Larcher *et al*., 2023; Garcia *et al*., 2024; Sales Almeida *et al*., 2024]. These works, introducing models such as Sabiá [Pires *et al*., 2023], Cabrita [Larcher *et al*., 2023], and Bode [Garcia *et al*., 2024], emphasize the importance of language-specific solutions in increasing the performance and comprehension of Brazilian Portuguese compared to predominantly English-trained SOTA models.

Based on these developments, Souza and Filho [2022] conducted a domain-specific comparative analysis of sentiment classification for Portuguese user reviews, utilizing embeddings from various BERT-based models, including BERTimbau [Souza *et al*., 2020], a Brazilian Portuguese-specialized BERT variant. Their results established BERTimbau as the superior BERT variant for Portuguese text classification tasks. More recently, de Araujo *et al*. [2024] evaluated GPT-3.5-Turbo's capabilities in Portuguese opinion mining tasks, including sentiment classification, concluding that the model demonstrates robust predictive performance without significant limitations.

4 Methodology

This study constitutes an empirical comparative research based on the analysis of 23 language models, comprising 13 SOTA models with multilingual capabilities and 10 with fine-tuning for the Portuguese language. The characterization of these models is presented in Section 4.1.

From a wide mapping of public Portuguese datasets for sentiment classification, 12 datasets were selected, described in Section 4.2. The in-context learning methodology and prompt engineering strategy are presented in Section 4.3 and Section 4.4, respectively. The criteria and processes for comparative evaluation of the models' predictive performance are detailed in Section 4.5.

4.1 Selected Models

The Table 1 summarizes the metadata of the models selected for this comparative study. These models are categorized along two main dimensions. The first concerns the parameter count: large-scale models contain over 70 billion parameters, while smaller-scale models range between 7 and 13 billion parameters. The second dimension relates to linguistic specialization, distinguishing between non-specialized (also known as generalist or multilingual) models and those fine-tuned in Brazilian Portuguese.

4.1.1 Generalist LLMs

Claude In early 2023, Anthropic released its closed-source LLM family, Claude, which has evolved to its current versions: Claude-3 and 3.5 [Anthropic, 2024b, 2023, 2024c]. The models are accessible through APIs and a chat interface[1] [Anthropic, 2023], with most technical specifications remaining proprietary.

These models were trained on a diverse dataset combining public internet information, third-party private data, and internally generated data, using word prediction techniques and human feedback reinforcement [Anthropic, 2024a]. The training approach focused on ensuring alignment with the company's guidelines while maintaining versatility across different domains.

Claude-3.5-Sonnet, the most advanced version, has demonstrated superior performance across reasoning, reading comprehension, mathematics, science, and coding benchmarks compared to its predecessors [Anthropic, 2024a]. Notably, in multilingual capabilities, the model achieved significant improvements in the Multilingual MMLU benchmark, making it particularly relevant for comparative studies in linguistic diversity [Anthropic, 2024a].

GPT The Generative Pre-trained Transformer (GPT) family comprises decoder-based LLMs developed by OpenAI [Minaee *et al*., 2024]. While the initial models GPT-1 [Radford *et al*., 2018] and GPT-2 [Radford *et al*., 2019] were open-source, subsequent versions GPT-3 [Brown *et al*., 2020] and GPT-4 [OpenAI *et al*., 2024b] are closed-source, accessible through APIs and the ChatGPT web application[2] [Minaee *et al*., 2024].

GPT-4, the latest and most capable model in the family, is a multimodal LLM based on the Transformer architecture. It was pre-trained on next-token prediction tasks and refined using reinforcement learning with human feedback [OpenAI

[1] https://claude.ai/
[2] https://chat.openai.com/

Table 1. Metadata of selected Language Models (LLMs) for the comparative benchmark study. The table organizes models by family, providing information about each model's characteristics, where PT-BR indicates Brazilian Portuguese fine-tuning. Both proprietary large-scale LLMs and open-source alternatives with varying parameters and specializations are included for comparison.

Family	Model	Version	Release Year	Base Model	Linguistic Fine-Tuning	# of Parameters	Open Source	Reference
Claude 3	Claude-3.5 Sonnet	claude-3-5-sonnet-20240620	2024	-	-	-	×	[Anthropic, 2024b]
GPT 4	GPT-4o	gpt-4o-2024-05-13	2024	-	-	-	×	[OpenAI *et al.*, 2024a]
Gemini	Gemini 1.5 Pro	gemini-1.5-pro-001	2024	-	-	-	×	[Reid *et al.*, 2024]
LLaMA 3	LLaMA 3-8B Instruct	llama-3-8b-it	2024	-	-	8 B	✓	[Meta, 2024]
	LLaMA 3.1-8B Instruct	llama-3.1-8b-it	2024	-	-	8 B	✓	
Gemma	Gemma-7B Instruct	gemma-7b-it	2024	-	-	7 B	✓	[Gemma Team *et al.*, 2024a]
Gemma 2	Gemma 2-9B Instruct	gemma-2-9b-it	2024	-	-	9 B	✓	[Gemma Team *et al.*, 2024b]
Qwen 2	Qwen 2-7B Instruct	qwen-2-7b-it	2024	-	-	7 B	✓	[Yang *et al.*, 2024a]
InternLM 2	InternLM 2-7B Chat	internlm2-chat-7b	2024	-	-	7B	✓	[Cai *et al.*, 2024]
DeepSeek	DeepSeek-V3	deepseek-v3	2025	DeepSeek V3 Base	-	671B	✓	[DeepSeek-AI *et al.*, 2025b]
	DeepSeek-R1 †	deepseek-r1	2025	DeepSeek V3 Base	-	671B	✓	[DeepSeek-AI *et al.*, 2025a]
	DeepSeek-R1-Distill-Qwen-7B	deepseek-r1-distill-qwen-7B	2025	Qwen2.5 Math 7B	-	7B	✓	
	DeepSeek-R1-Distill-Llama-8B	deepseek-r1-distill-llama-8B	2025	Llama 3.1 8B	-	8B	✓	
Sabiá	Sabiá-7B	sabia-7b	2023	LLaMA	PT-BR	7 B	✓	[Pires *et al.*, 2023]
	Sabiá-2 Medium	sabia-2-medium	2024	-	PT-BR	-	×	[Sales Almeida *et al.*, 2024]
	Sabiá-3	sabia-3	2024	-	PT-BR	-	×	[Abonizio *et al.*, 2024]
Bode	Bode-7B	bode-7b-alpaca-PT-BR	2023	LLaMA 2	PT-BR	7 B	✓	[Garcia *et al.*, 2024]
	Bode-13B	bode-13b-alpaca-PT-BR	2023		PT-BR	13 B	✓	
	Bode-3.1-8B-Instruct-lora	bode-3.1-8b-instruct-lora	2024	LLaMA 3	PT-BR	8 B	✓	
	InternLM-ChatBode-7B	internlm-chatbode-7b	2024	InternLM 2	PT-BR	7 B	✓	
	GemBode-7B-Instruct	gembode-7b-it	2024	Gemma	PT-BR	7 B	✓	[Garcia *et al.*, 2025]
Cabra	CabraLLaMA 3-8B	cabrallama-3-8b	2024	LLaMA 3	PT-BR	8 B	✓	-
	CabraMistral-v3-7b-32k	cabramistral-v3-7b-32k	2024	Mistral	PT-BR	7 B	✓	-

† To ensure benchmark parity, the DeepSeek-R1 model, being the only one among those evaluated with enhanced reasoning capabilities and with large parameters size, was selected as a strong reference classifier to contrast with the weak reference classifier (which always predicts the majority class from the training set), see Subsection 4.5.

et al., 2024b]. While its exact parameter count remains undisclosed, estimates suggest approximately 1.7 trillion parameters [Ding *et al.*, 2023; Yao *et al.*, 2024], significantly larger than its predecessor GPT-3's 175 billion parameters [Brown *et al.*, 2020].

The model has demonstrated human-comparable performance across various academic and professional tests, surpassing state-of-the-art results in traditional LLM benchmarks [OpenAI *et al.*, 2024b]. Its multilingual capabilities, evaluated through translated versions of the MMLU test [Hendrycks *et al.*, 2020], showed superior performance compared to competitors like Chinchilla [Hoffmann *et al.*, 2022] and PaLM [Chowdhery *et al.*, 2022]. These capabilities and multilingual proficiency make GPT-4 a crucial candidate for this comparative study.

Gemini The Gemini family, developed by Google, consists of Transformer decoder-based LLMs trained on multimodal data, including text, images, audio, and video [Gemini Team *et al.*, 2023]. While the first generation (Gemini-1.0) was available in three variants—Ultra, Pro, and Nano—only the Nano versions' parameters were officially disclosed, with Nano-1 containing 1.8 billion and Nano-2 containing 3.25 billion parameters [Gemini Team *et al.*, 2023].

Gemini-1.5-Pro, the latest iteration, introduced significant innovations, including a sparse mixture of expert Transformer models and an expanded context window of millions of tokens—substantially surpassing competitors like Claude-2.1 (200K tokens) and GPT-4 (128K tokens) [Reid *et al.*, 2024]. This version demonstrated a 22.3% improvement in multilingual capabilities over its 1.0 counterpart and performed 6.7% better than the 1.0 Ultra model [Reid *et al.*, 2024].

The selection of Gemini-1.5-Pro for this study is based on its advanced technical features, effective handling of complex contexts, and robust multilingual capabilities. Its performance in comparative tests, particularly in multilingual tasks, has shown significant improvements over its predecessor, achieving state-of-the-art results in benchmarks such as MMLU, where it demonstrated human expert-level performance.

Gemma Google also released Gemma, an open-source LLM family inspired by Gemini [Gemma Team *et al.*, 2024a]. The first generation included models with 2 and 7 billion parameters, along with their instruction-tuned variants. These Transformer decoder-based models were pre-trained primarily on English-language tokens from web documents, code, and mathematical content [Gemma Team *et al.*, 2024a]. In benchmark tests, the 7 billion parameter version outperformed comparable models like LLaMA-2-7B [Touvron *et al.*, 2023b] and Mistral-7B [Jiang *et al.*, 2023], as well as the slightly larger LLaMA-2-13B [Touvron *et al.*, 2023b].

The second generation, Gemma-2, features models with 2, 9, and 27 billion parameters, each with instruction-tuned variants [Gemma Team *et al.*, 2024b]. These models incorporate architectural improvements, including deeper neural networks, Grouped-Query Attention[Ainslie *et al.*, 2023], and alternating global-local attention layers [Beltagy *et al.*, 2020]. The Gemma-2-9B model demonstrated approximately 12% better average performance compared to its first-generation counterpart [Gemma Team *et al.*, 2024b].

While neither generation was specifically designed for multilingual tasks, both inherit Gemini's vocabulary architecture, featuring an extensive embedding parameter space capable of handling multiple languages [Gemma Team *et al.*, 2024b]. Given their promising performance and the opportunity to evaluate them in non-English scenarios, this study includes representatives from both generations: Gemma-7B and Gemma-2-9B.

LLaMA LLaMA [Touvron *et al.*, 2023a], Meta's open-source multilingual LLM family, has evolved to its third generation with versions 3 and 3.1, featuring Transformer decoder-based models ranging from 8 to 405 billion parameters [Grattafiori *et al.*, 2024]. Unlike Claude, GPT-4, and Gemini, LLaMA's open-source nature and non-commercial licensing [Minaee *et al.*, 2024] has facilitated its widespread adoption in research communities as a foundation for specialized LLMs.

While LLaMA-3 includes approximately 5% non-English training data across 30 languages [Meta, 2024], version 3.1 significantly enhanced its multilingual capabilities to support eight languages, including Portuguese [Grattafiori *et al.*, 2024]. This improvement was achieved through a specialized multilingual model that extracted high-quality annotations from non-English data sources, including human annotations, NLP tasks, and translated quantitative reasoning data for supervised fine-tuning [Grattafiori *et al.*, 2024]. The models were pre-trained on over 15 trillion high-quality tokens, primarily from public online sources [Meta, 2024].

Both LLaMA-3 and 3.1 8B models have achieved state-of-the-art results compared to similarly-sized LLMs [Grattafiori *et al.*, 2024]. Notably, the pre-trained LLaMA-3.1-8B outperformed competitors in five out of six evaluated categories, while its fine-tuned version excelled in multilingual tests, surpassing models like Mistral-7B [Jiang *et al.*, 2023] and Gemma-2-9B [Gemma Team *et al.*, 2024b]. Based on these achievements, their open-source nature, and the opportunity to compare similar-sized models with different multilingual capabilities, both LLaMA-3-8B Instruct and LLaMA-3.1-8B Instruct were selected for this comparative study.

Qwen The Qwen model family, developed by Alibaba, was initially released in 2023 with various Transformer-based versions, including open-source pre-trained models ranging from 1.8 to 14 billion parameters, along with specialized variants for instruction-following, coding, and mathematics [Bai *et al.*, 2023]. The 1.0 generation was pre-trained on trillions of tokens from diverse sources, including web documents, books, encyclopedias, and programming code, with content primarily in English and Chinese [Bai *et al.*, 2023].

In 2024, Qwen-2 was released with pre-trained and instruction-tuned models ranging from 0.5 to 72 billion parameters [Yang *et al.*, 2024a]. A key innovation of this generation was the expansion of training data to include 27 languages, including European Portuguese, significantly enhancing its multilingual capabilities.

The Qwen-2-7B Instruct model demonstrated improved performance across most benchmarks compared to both Qwen-1.5 and other state-of-the-art open-source LLMs, including LLaMA-3-70B and LLaMA-3-8B [Yang *et al.*, 2024a]. Given its significant performance in comparative tests and multilingual capabilities, the Qwen-2-7B Instruct version was selected for this study's evaluation.

InternLM The Intern series of foundation models was developed through collaboration between SenseTime corporation, the Shanghai Artificial Intelligence Laboratory, the Chinese University of Hong Kong, Fudan University, and Shanghai Jiaotong University [InternLM Team, 2023].

Following the initial InternLM release in 2023 [InternLM Team, 2023], the second generation InternLM-2 was made available in sizes ranging from 1.8 to 20 billion parameters [Cai *et al.*, 2024]. These models use a decoder-only transformer architecture and were pre-trained on over 2 trillion tokens predominantly from English and Chinese sources, followed by Supervised Fine-Tuning and Conditional Online Reinforcement Learning from Human Feedback, having the ability to handle large contexts (up to 200k tokens) [Cai *et al.*, 2024].

InternLM-2 models have demonstrated promising results across various benchmarks when compared with other open-source LLMs of up to 7 billion parameters [Cai *et al.*, 2024], such as LLaMA-2-7B [Touvron *et al.*, 2023b] and Qwen-7B [Bai *et al.*, 2023]. They performed particularly well in the FLORES 101 comparative examination [Goyal *et al.*, 2022], which tests translation capabilities across 101 languages including Brazilian Portuguese, establishing InternLM-2 as competitive for applications requiring robust language comprehension [Cai *et al.*, 2024].

The inclusion of InternLM-2 LLMs, represented by the InternLM-2-7B Chat version in the set of evaluated models, was based on their open-source nature and intermediate size (7 billion parameters), combined with this proven ability to understand and generate texts in multiple languages. InternLM-2-Chat fills an important gap, representing the category of smaller-scale multilingual open-source models, thus offering a valuable counterpoint between large proprietary models and models fine-tuned in Brazilian Portuguese.

DeepSeek The DeepSeek LLM project represents an initiative by the Chinese company DeepSeek aimed at the dissemination and development of open-source language models [DeepSeek-AI *et al.*, 2024a]. The first generation of these models, released in early 2024, comprises versions of 7 and 67 billion parameters, optimized or not for conversational interactions, pre-trained on approximately 2 trillion tokens, predominantly in English and Chinese languages, showing strong inspiration from the LLaMA model architecture [DeepSeek-AI *et al.*, 2024a].

DeepSeek-V3, the most recent version of the family, preserved characteristics introduced in the second generation [DeepSeek-AI *et al.*, 2024b], such as the Mixture-of-Experts architecture (DeepSeekMoE) and the Multi-head Latent Attention mechanism. This new generation presents significant scalability regarding the total number of parameters, reaching 671B with 37B active per token, and also in terms of pre-training token volume, totaling 14.8 trillion with enhanced multilingual coverage compared to previous generations, which primarily focused on English and Chinese language data [DeepSeek-AI *et al*., 2024b, 2025b].

Although it achieved superior performance among the evaluated open-source models and comparable performance to proprietary LLMs [DeepSeek-AI *et al*., 2025b], DeepSeek-V3 gained notoriety in academic and professional circles mainly due to its derivative model, DeepSeek-R1 [DeepSeek-AI *et al*., 2025a]. This represents the first generation of models with reasoning capabilities developed by DeepSeek, being built from DeepSeek-V3 and achieving performance comparable to the state-of-the-art in reasoning models, OpenAI-o1 (OpenAI-o1-1217 [3]). Initially, the versions DeepSeek-R1-Zero, DeepSeek-R1, and dense models between 1.5 and 70B parameters were made available, distilled from DeepSeek-R1 and based on the LLMs LLaMA-3.1-8B and Qwen-2.5-Math-7B [DeepSeek-AI *et al*., 2025a].

For the conduct of this study, 4 models from the DeepSeek family were selected: the DeepSeek-V3-671B model, representing large-scale multilingual models (>70B); the smaller-scale distilled versions DeepSeek-R1-Distill-LLaMA 3.1-8B and Qwen-7B, as representatives of smaller-dimension multilingual models (<13B); and DeepSeek-R1-671B, used as a strong reference classifier.

4.1.2 Brazilian Portugues Fine-tunned LLMs

Bode The Bode model family [Garcia *et al*., 2024] comprises various subsets of Brazilian Portuguese fine-tuned models derived from LLMs such as LLaMA-2 [Touvron *et al*., 2023b], Gemma [Gemma Team *et al*., 2024a], and InternLM [Cai *et al*., 2024]. These models, available on HuggingFace,[4] aim to enhance the capabilities of existing LLMs in Portuguese language processing.

The family is organized into distinct subsets based on their foundation models: the Bode subset derived from LLaMA models, GemBode [Garcia *et al*., 2025] from Google's Gemma, PhiBode [Garcia *et al*., 2025] from Microsoft's Phi [Gunasekar *et al*., 2023], and InternLM-ChatBode from InternLM-2. The fine-tuning process utilized translated versions of Alpaca and UltraAlpaca datasets, employing efficient methods such as Low-Rank Adaptation (LoRA) [Hu *et al*., 2021] and QLoRA [Dettmers *et al*., 2023] to incorporate Brazilian Portuguese linguistic and cultural nuances.

In binary sentiment analysis tasks, Bode-13B demonstrated superior performance, achieving 10% higher accuracy than LLaMA-2-7B[5] and 64% better than LLaMA-2-13B[6]. Based on these results and evaluations from the Open Portuguese LLM Leaderboard [Garcia, 2024], four models were selected for this comparative study: Bode-7B, Bode-13B, GemBode-7B-it, and InternLM-ChatBode-7B.

Cabra The Cabra family consists of open-source LLMs fine-tuned on proprietary Brazilian Portuguese datasets called "CabraSets", developed by BotBot [BotBot AI, 2024a]. These models aim to enhance linguistic understanding of Brazilian language and culture [BotBot AI, 2024b]. Available on HuggingFace[7], the family includes CabraLLaMA3 models [BotBot AI, 2024c] with 8 and 70 billion parameters, CabraMistral-v3-7B-32k derived from Mistral-7B [Mistral AI Team, 2023], and Cabra-72B based on Qwen-1.5-72B [Qwen Team, 2024].

All models were fine-tuned using the "Cabra" datasets, with CabraMistral-v3-7B-32k utilizing "Cabra12k" and the others employing "Cabra30k". In the Open Portuguese LLM Leaderboard [Garcia, 2024], particularly in sentiment analysis tasks using TweetSentBR [Brum and das Graças Volpe Nunes, 2018], the models achieved notable scores: CabraMistral-v3-7B-32k (65.71), CabraLLaMA3-8B (68.08), CabraLLaMA3-70B (73.85), and Cabra-72B (71.64).

For this study, CabraMistral-v3-7B-32k and CabraLLaMA3-8B were selected based on their favorable performance-to-size ratio, with CabraLLaMA3-8B showing competitive performance compared to larger variants. This selection also aligns with the parameter scale of other models considered in this study.

Sabiá Sabiá LLMs, developed by Maritaca AI, include both open-source models like Sabiá-7B [Pires *et al*., 2023] and closed-source versions such as Sabiá-65B, Sabiá-2 (Small and Medium variants), and the latest Sabiá-3 [Pires *et al*., 2023; Sales Almeida *et al*., 2024; Abonizio *et al*., 2024]. The first-generation models, Sabiá-7B and Sabiá-65B, were derived from LLaMA-7B and 65B respectively, fine-tuned on a quality-filtered Portuguese subset of the ClueWeb dataset [Overwijk *et al*., 2022a,b].

The models' performance was evaluated across 14 Portuguese datasets, collectively known as Portuguese Evaluation Tasks (Poeta) [Pires *et al*., 2023]. In sentiment analysis tasks, both models showed substantial improvements over their base LLaMA versions for native Portuguese content, though Sabiá-65B performed slightly below LLaMA-65B on translated datasets [Pires *et al*., 2023].

Sabiá-2-Medium demonstrated the effectiveness of language-specific specialization by matching or surpassing GPT-4's performance [Sales Almeida *et al*., 2024]. In professional certification, university admission, and high school exams, it was only outperformed by GPT-4-Turbo[8] and Claude-3-Opus[9], while being 10 to 22 times more cost-effective [Sales Almeida *et al*., 2024]. Given these capabilities, three models were selected for this study: the open-source Sabiá-7B, Sabiá-2-Medium, and the latest Sabiá-3.

4.2 Datasets

This study utilized 12 public datasets containing annotated Brazilian Portuguese texts for sentiment classification. The characteristics of these selected datasets are summarized in Table 2.

All datasets were standardized for binary sentiment classification, retaining only instances labeled as `Positive` and

[3] `https://platform.openai.com/docs/models#o1`
[4] `https://huggingface.co/recogna-nlp`
[5] `https://huggingface.co/meta-llama/Llama-2-7b`
[6] `https://huggingface.co/meta-llama/Llama-2-13b`
[7] `https://huggingface.co/botbot-ai`
[8] version gpt-4-0125-preview
[9] version claude-3-opus-20240229

Table 2. Mapped datasets. The selected datasets for comparative tests contain texts from different domains in Brazilian Portuguese, are labeled for sentiment polarity, and are published and available online.

Dataset	Translated/ Native	Content	Test Size	Training set Label Distribution	Reference
IMDB_PT	Translated	Movie Reviews	5.000	✓ 50% ✕ 50%	[Maas *et al*., 2011; Pires *et al*., 2023]
SST2_PT	Translated	Movie Reviews	872	✓ 56% ✕ 44%	[Socher *et al*., 2013; Pires *et al*., 2023]
TweetSentBr	Native	Social Media Posts	1.495	✓ 50% ✕ 50%	[Brum and das Graças Volpe Nunes, 2018]
ReLI	Native	Book Reviews	627	✓ 83% ✕ 17%	[Freitas *et al*., 2014]
Computer-BR	Native	Social Media Posts	128	✓ 30% ✕ 70%	[Moraes *et al*., 2016]
MTMSLA	Native	Social Media Posts	102	✓ 58% ✕ 42%	[Araujo *et al*., 2016]
CSP-Eletrônicos	Native	Product Reviews	38	✓ 70% ✕ 30%	[Belisário *et al*., 2019]
CSP-Livros	Native	Book Reviews	35	✓ 50% ✕ 50%	[Belisário *et al*., 2019]
4P Corpus	Native	Product Reviews	278	✓ 82% ✕ 18%	[Silva and Pardo, 2019]
RePro	Native	Product Reviews	1.516	✓ 54% ✕ 46%	[dos Santos Silva *et al*., 2024; Real *et al*., 2019]
OPCovidBR	Native	Social Media Posts	123	✓ 50% ✕ 50%	[Vargas *et al*., 2020]
TA-Restaurantes	Native	Restaurant Reviews	113	✓ 90% ✕ 10%	[Oliveira and de Melo, 2020]

`Negative`, with other labels such as `Neutral` being removed. The labels were encoded as integers: 1 for `Positive` and -1 for `Negative`. Unless originally partitioned by their authors, the datasets were split into training (80%) and test (20%) subsets while preserving the balance of the original labels.

IMDB_PT Is the Portuguese translation of the IMDB dataset [Maas *et al*., 2011], containing movie reviews labeled as `Positive` or `Negative`. This study utilized the version provided by Maritaca AI, which includes predefined training and test splits and is part of the Poeta benchmark [Pires *et al*., 2023].

SST2_PT Another Poeta benchmark dataset [Pires *et al*., 2023], is the machine-translated Portuguese version of SST2 [Socher *et al*., 2013]. It comprises approximately $67,000$ training and 872 validation instances, each labeled as `Positive` or `Negative`.

TweetSentBr Contains Brazilian Portuguese tweets annotated based on user reactions to the posts' main topics [Brum and das Graças Volpe Nunes, 2018]. Part of the Poeta evaluation [Pires *et al*., 2023], this study used a subset comprising 75 training and $2,000$ test instances.

ReLI The ReLI corpus [Freitas *et al*., 2014] consists of $1,600$ manually annotated book reviews in Portuguese, covering 14 different books with $12,470$ sentences. The corpus contains $2,883$ `Positive`, 596 `Negative`, and 212 dual-labeled sentences.

Computer-BR Contains $2,317$ manually annotated Portuguese tweets about computers [Moraes *et al*., 2016]. Following the authors' approach, tweets originally labeled as `Irony` were converted to `Negative`.

MTMSLA A subset of [Araujo *et al*., 2016], contains 774 Portuguese tweets with 297 `Positive`, 213 `Negative`, and 264 `Neutral` labels.

CSP-Eletrônicos Comprises 234 manually annotated electronic product reviews, containing 131 `Positive`, 59 `Negative`, and 43 `Neutral` reviews [Belisário *et al*., 2019].

CSP-Livros Contains 350 book reviews extracted from the ReLI corpus [Freitas *et al*., 2014], social media, and an online shopping platform, with 88 `Positive`, 87 `Negative`, and 175 `Neutral` labels [Belisário *et al*., 2019].

4P Corpus The 4P Corpus [Silva and Pardo, 2019] contains 642 Portuguese sentences from 542 Buscapé reviews covering four products (two digital cameras and two mobile phones), manually classified as `Positive` or `Negative`.

RePro Derived from B2W-Reviews01 [Real *et al*., 2019], RePro [dos Santos Silva *et al*., 2024] contains $10,000$ manually annotated reviews of e-Commerce products. For this study, only instances with single polarity labels (`['POSITIVE']` or `['NEGATIVE']`) were retained.

OPCovidBR Comprises $2,000$ Portuguese tweets about COVID-19 collected during the pandemic, annotated at both opinion and document polarity levels as `Positive` or `Negative` [Vargas *et al*., 2020].

TA-Restaurantes Contains Brazilian Portuguese reviews of restaurants from TripAdvisor[10] [Oliveira and de Melo, 2020] . The dataset includes 561 subjective sentences labeled as `Positive` or `Negative`, extracted from an original set of $1,049$ sentences.

4.3 In-context Learning

Large Language Models applied to Natural Language Processing stand out mainly through two paradigms: fine-tuning and In-Context Learning (ICL) [Han *et al*., 2021; Dong *et al*., 2023]. Fine-tuning consists of using pre-trained weights from

[10]`https://www.tripadvisor.com.br`

PLM/LLMs as a foundation for specialization in a specific task, utilizing a reduced dataset [Qiu *et al.*, 2020].

In this approach, the model parameters are refined for a specific objective, preserving the linguistic knowledge incorporated during pre-training [Han *et al.*, 2021]. Studies demonstrate that this methodology achieves SOTA results in various NLP tasks when compared to the direct use of pre-trained models [Brown *et al.*, 2020; Qiu *et al.*, 2020; Han *et al.*, 2021; Zhao *et al.*, 2023]. However, its implementation faces challenges such as the need for task-specific datasets [Brown *et al.*, 2020], significant computational costs, and commercial restrictions associated with restrictive licenses of advanced models like GPT 4.0 and Gemini [Brown *et al.*, 2020; Touvron *et al.*, 2023b].

In contrast, the ICL paradigm emerges as a promising alternative, leveraging the emergent capabilities [Wei *et al.*, 2022] of modern LLMs, which derive from their scale in terms of parameters and training corpus extension. According to Dong *et al.* [2023], ICL can be understood as learning by analogies through contextual examples, distinguishing itself from traditional learning by not requiring parameter updates via gradient backpropagation. In this approach, predictions are made directly by the pre-trained model, as illustrated in Figure 1.

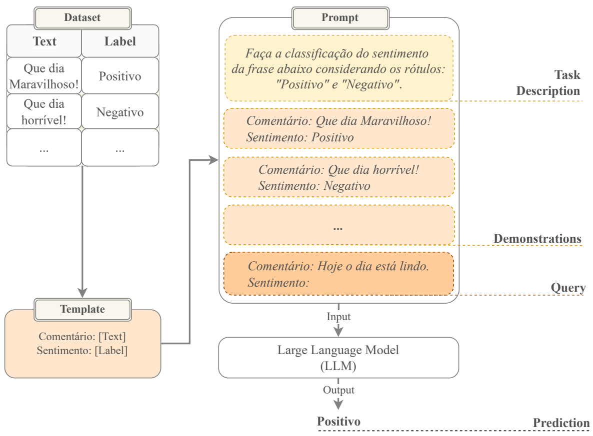


Figure 1. In-Context Learning Strategy. The process illustrates the transformation of tabular data into a structured template for LLM processing. The prompt is constructed by incorporating selected examples and task-specific instructions in natural language. The LLM processes this contextualized input and generates as output the label corresponding to the query. Adapted from [Dong *et al.*, 2023]

The ICL technique gained prominence following the publication by Brown *et al.* [2020], where the authors demonstrated a direct correlation between the number of language model parameters and their in-context learning capability. Using GPT 3.0, with 175 billion parameters, the research showed that model performance is enhanced by adding natural language instructions and target task demonstrations.

Dong *et al.* [2023] highlight the main advantages of ICL: an interpretable interface for communication with LLM through natural language, ease of incorporating human knowledge via adjustments in prompt and examples, decision-making process analogous to human reasoning, and computational efficiency as it doesn't require model adaptation. However, the approach presents limitations, including inferior performance compared to fine-tuning [Brown *et al.*, 2020; Mosbach *et al.*, 2023], restrictions on the number of examples due to LLMs' maximum input size, opaque operational mechanisms, and performance instability influenced by task and demonstration structuring [Lu *et al.*, 2022; Dong *et al.*, 2023; Mosbach *et al.*, 2023].

Considering ICL's adaptability [Krugmann and Hartmann, 2024], this method was selected to conduct comparative tests between LLMs. Based on the findings of Simmering and Huoviala [2023], which identified superior performance in sentiment classification using 6 demonstrations, the same number of examples was adopted. Detailed specifications regarding prompt structuring and demonstration selection will be presented subsequently.

4.4 Prompt Engineering

Prompt Engineering is a discipline focused on guiding LLM responses through systematic design and optimization of input instructions [Chen *et al.*, 2023]. It can be conceptualized as natural language programming, where human knowledge is adapted to address the specific requirements of language model interactions [Reynolds and McDonell, 2021].

The field gained prominence, as noted by Zhou *et al.* [2022], due to the frequent misalignment between natural language prompts and expected outputs, necessitating extensive experimentation to achieve desired behaviors given the limited understanding of instruction-model compatibility. This has led to research efforts aimed at understanding prompt dynamics, cataloging available knowledge [Dong *et al.*, 2023; Giray, 2023; White *et al.*, 2023], and developing efficient prompt generation methodologies, both manual [Reynolds and McDonell, 2021] and automated [Reynolds and McDonell, 2021; Zhou *et al.*, 2022; Wang *et al.*, 2022]. These studies have also explored optimal demonstration selection for ICL [Liu *et al.*, 2022; Rubin *et al.*, 2022; Ye *et al.*, 2023] and their sequencing [Lu *et al.*, 2022].

For this study, a manual prompt was developed, as shown in Figure 2, incorporating guidelines to enhance model responses. These guidelines include clear and objective instruction specification, structured output format definition, and strategic use of demonstrations [Reynolds and McDonell, 2021; Giray, 2023; Simmering and Huoviala, 2023].

The demonstration selection process involved randomly sampling 3 examples from each class (`Positive` and `Negative`) from the respective training subsets across all 12 datasets utilized in this study. The 6 demonstrations were organized in an interleaved fashion within the prompt, maintaining consistent structure up to the Query section (Figure 2) across all inferences performed on the corresponding test set. This procedure ensured that the same 6 demonstrations were systematically employed for all predictions within each dataset, providing methodological consistency and enabling direct comparability between the evaluated models while minimizing potential confounding variables related to example selection.

Despite acknowledging the implications of random demonstration selection and manual prompt engineering, these methodologies were adopted for the present study. The demonstrations selection method for ICL significantly influ-

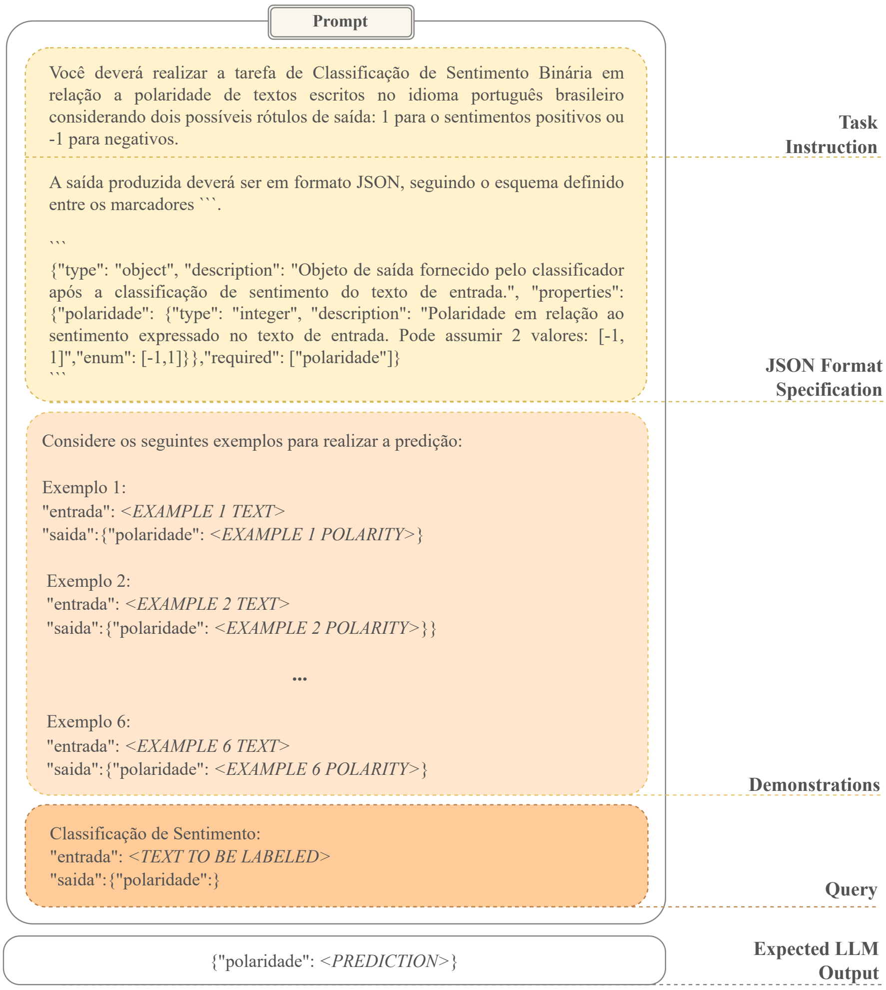


Figure 2. Prompt structure implemented for sentiment classification of Brazilian Portuguese texts. The prompt is organized into four main components, laterally identified as: (1) Task Instruction: specifies the binary classification task with polarity values of 1 (positive) and -1 (negative); (2) JSON Format Specification: defines the structured output schema in JSON format, specifying data types and allowed values; (3) Demonstrations: presents a series of numbered examples (1, 2, ..., 6) containing input-output pairs to guide the classification; and (4) Query: contains the text to be classified. The figure concludes with the Expected LLM Output, which illustrates the expected prediction format from the model.

ence model outputs, with research indicating that randomly chosen demonstration subsets tend to produce performance instabilities in LLMs [Lu *et al*., 2022; Li and Qiu, 2023].

As noted by Lu *et al*. [2022], there is no evidence of prompt performance transferability or label ordering effectiveness across different LLMs. To maximize predictive performance, prompt engineering, example selection, and demonstration ordering should be conducted using automated and systematic methods [Zhou *et al*., 2022; Liu *et al*., 2022; Lu *et al*., 2022] for each specific model.

However, given that the primary research objective is to compare LLMs' predictive capabilities in sentiment classification for Brazilian Portuguese texts, we opted to accept the risk of sub-optimal performance for feasibility and comparability reasons. This methodological choice is acknowledged as one of the study's limitations.

4.5 Evaluation

To evaluate LLMs' performance in binary sentiment classification of Brazilian Portuguese texts, this study employed the ICL strategy with 6 demonstrations. Each instance from the test subset was passed to the models as prompts, as shown in Figure 2. The LLM consumption method, along with the configurations and main parameters used by each model, are presented in Table 3.

All experiments were performed using the Google Colab[11] service with different hardware (GPUs), since the availability of specific hardware is not always guaranteed by the provider. The notebooks containing the experiment codes, as well as the test and demonstration datasets are available in this article's repository[12].

To maximize prediction accuracy and reduce the effect of inherent non-determinism of LLMs, generation randomness-related parameters were configured to their most conservative values, ensuring that model outputs frequently correspond to those tokens with the highest associated probabilities. The specific configuration of these parameters varied according to each model's consumption method and available settings.

[11] `https://colab.google/`

[12] `https://github.com/AndreSchuck/EvaluatingLargeLanguageModelsforBrazilianPortugueseSentimentAnalysis`

Table 3. Comparative analysis of hardware configurations, consumption methods and operational parameters for the selected language models.

Model	Consuption Method	Framework	Hardware	Generation Parameters
Claude-3.5-Sonnet	API	Proprietary API	CPU †	max_tokens= 20, temperature=0.0
GPT-4o	API⋆	Proprietary API	CPU	max_tokens=20, n=1, seed=4, temperature=0
Gemini-1.5-Pro	API	Proprietary API	CPU	max_output_tokens=20,temperature=0,
LLaMA-3-8B-Instruct	Local	HFTP *	L4 GPU‡	max_new_tokens=150, do_sample=False
LLaMA-3.1-8B-Instruct	Local	HFTP	L4 GPU	max_new_tokens=150, do_sample=False
Gemma-7B-Instruct §	Local	HFT $	L4 GPU	max_new_tokens=20, do_sample=False
Gemma-2-9B-Instruct §	Local	HFT	L4 GPU	max_new_tokens=20, do_sample=False
Qwen-2-7B-Instruct	Local	HFT	L4 GPU	max_new_tokens=20, do_sample=False
InternLM-2-7B-Chat	Local	HFTP	A100 GPU ††	max_new_tokens=20, do_sample=False
DeepSeek-V3	API	OpenRouter API	CPU	temperature=0, top_k=1, max_tokens=20
DeepSeek-R1	API	OpenRouter API	CPU	temperature=0, top_k=1
DeepSeek-R1-Distill-Qwen-7B	Local	HFTP	A100 GPU	max_new_tokens=20, do_sample=False
DeepSeek-R1-Distill-LLaMA-8B	Local	HFTP	A100 GPU	max_new_tokens=20, do_sample=False
Sabiá-7B	Local	HFTP	L4 GPU	max_new_tokens=20, do_sample=False
Sabiá-2-Medium	API	Proprietary API	CPU	temperature=0, max_tokens=20, do_sample=False
Sabiá-3	API	Proprietary API	CPU	temperature=0, max_tokens=20, do_sample=False
Bode-7B §§	Local	HFTP	L4 GPU	max_new_tokens=20, do_sample=False
Bode-13B§§	Local	HFTP	L4 GPU	max_new_tokens=20, do_sample=False
Bode-3.1-8B-Instruct-lora	Local	HFTP	A100 GPU	max_new_tokens=20, do_sample=False
InternLM-ChatBode-7B	Local	HFTP	L4 GPU	max_new_tokens=20, do_sample=False
GemBode-7B-Instruct	Local	HFT	A100 GPU	max_new_tokens=20, do_sample=False
CabraLLaMA-3-8B	Local	HFTP	A100 GPU	max_new_tokens=20, do_sample=False
CabraMistral-v3-7B-32k	Local	HFTP	A100 GPU	max_new_tokens=20, do_sample=False

⋆ Batch API.
§ 4 bits quantization.
§§ 8 bits quantization.
* HuggingFace Transformers Pipeline.
$ HuggingFace Transformers.
† CPU = Google Colab CPU with 12 GB of CPU RAM.
‡ L4 GPU = Google Colab L4 GPU with 22.5 GB of GPU RAM.
†† A100 GPU = Google Colab A100 GPU with 40 GB of GPU RAM.

Table 3 presents the complete set of inference parameters utilized for each evaluated model. For LLaMA-3-8B and LLaMA-3.1-8B models, a higher *Maximum number of new tokens* was required, as these models produced highly literal responses regarding the JSON structure specified in the prompt instruction, as demonstrated in Figure 3.

Ex 1: {"polaridade":1}

Ex 2 : {'polaridade': 1}

Ex 3: {
"polaridade": 1
}

Ex 4: 'polaridade': 1

(a) Other LLMs Outputs.

Ex: {'type': 'object', 'description': 'Objeto de saída fornecido pelo classificador após a classificação de sentimento do texto de entrada.', 'properties': {'polaridade': {'type': 'integer', 'description': 'Polaridade em relação ao sentimento expressado no texto de entrada. Pode assumir 2 valores: [-1, 1]','enum': [-1,1]}}, 'required': ['polaridade']}, 'saida': {'polaridade': 1}}

(b) LLaMA-3-8B and LLaMA-3.1-8B Outputs.

Figure 3. Comparison of output patterns produced by different LLMs for sentiment analysis tasks. (a) Typical response structures from most evaluated LLMs. (b) Distinctive output patterns from LLaMA-3-8B and LLaMA-3.1-8B models, highlighting their "literal" approach to structure the requested JSON output.

During the experiment, it was necessary to refine the response generation process for the Claude-3.5-Sonnet model. Initially, the model produced the correct JSON object structure but preceded it with a brief explanatory text about the task, followed by the word "JSON" before the actual JSON object containing the desired response. To optimize the output and reduce inference costs, a content restriction strategy was implemented [13]. In the communication turns between the user (who provided the prompt with instructions, demonstrations, and text to be classified) and the assistant (who generated the model's response), the word "JSON" was included in the assistant's function. This approach effectively limited the response content, eliminating the unwanted introductory text.

A distinctive behavioral pattern was observed in the Gemini-1.5-Pro model, where certain input instances triggered internal safety filters, resulting in content flagged as violating usage policies. In these cases, the API response returned empty values in the primary classification field, while populating additional fields with safety policy information. These instances, lacking the expected JSON structure for sentiment predictions, were systematically categorized as hallucinations (value 2) during the response parsing phase for evaluation purposes.

The generated responses were processed by an algorithm using a single regular expression pattern to verify the expected output format. This pattern was designed to identify a JSON-like structure containing a key named *"polaridade"* (which can be delimited by either single or double quotes) followed by a colon and a value that must be either 1 or -1, allowing for possible whitespace variations.

If the pattern is recognized in the model's response, the integer value (-1 or 1) associated with the *"polaridade"* key is extracted from the text. If the pattern is not identified, the value 2 is returned, indicating that the model produced a response outside the expected format, with this behavior being interpreted as a hallucination.

The evaluation phase involved comparing the models' outputs against the original test set labels. Accuracy (Acc) was chosen as the primary evaluation metric, following established practices in binary sentiment classification tasks [Larcher *et al.*, 2023; Pires *et al.*, 2023; Garcia *et al.*, 2024].

To address class imbalance effects on Accuracy, we also report the F_1 Score. The F_1 Score calculation employs the unweighted average approach, known as *Macro Average*, where individual scores are computed for each class and then averaged arithmetically.

In scenarios involving hallucinated responses, the *Macro F_1 Score* inherently penalizes performance due to the absence of True Positive instances in the hallucination category, resulting in a local score of zero for this class. While not contributing to the metric's numerator, this zero score increases the denominator count, effectively lowering the final *Macro F_1 Score* to reflect the presence of hallucinations in the evaluated responses.

The predictive performance of the models was evaluated through comparative analysis against two benchmark references established for each dataset. The first reference consists of a weak baseline classifier that consistently predicts the majority class identified in the training subset, representing the minimum acceptable performance threshold. The second reference establishes a strong baseline, represented by predictions generated from DeepSeek-R1-671B when subjected to identical prompts used across all evaluated LLMs.

The selection of DeepSeek-R1 as the strong baseline clas-

[13]`https://docs.anthropic.com/en/api/messages`

sifier was strategically determined based on its distinctive characteristics among the evaluated models. As documented in Table 1, DeepSeek-R1 stands as the only model in the study combining extensive parameter scale (671 billion parameters) with advanced reasoning capabilities, making it an optimal reference point for assessing the relative performance of other models under equivalent experimental conditions.

To ensure a robust methodology for comparing the results obtained from both evaluation metrics, the statistical significance of performance is assessed using the Wilcoxon signed-rank test for paired samples at a 5% significance level. This test compares LLMs results pairwise to evaluate 3 alternative hypotheses (H_1) against a common null hypothesis (H_0):

- **Test** 1 **- two-tailed**:
  H_0: The distribution of differences is symmetric around zero.
  H_1: The distribution underlying the differences is not symmetric about zero.

- **Test** 2 **- right-tailed**:
  H_0: The distribution of differences is symmetric around zero.
  H_1: The distribution underlying the differences is stochastically less than a distribution symmetric about zero.

- **Test** 3 **- left-tailed**:
  H_0: The distribution of differences is symmetric around zero.
  H_1: The distribution underlying the differences is stochastically greater than a distribution symmetric about zero.

Test 1 is performed for all possible model pair combinations, excluding pairs of identical models. If sufficient evidence exists to reject H_0 in Test 1, Tests 2 and 3 are then applied. Specifically, Test 1's H_0 indicates no significant difference between paired samples, while its H_1 examines any directional differences between groups. Test 2's H_1 evaluates whether group 1 has significantly higher values than group 2, whereas Test 3's H_1 assesses if group 1's values are significantly lower than group 2's.

Two methodological considerations warrant particular attention in this research context. The first concerns the intrinsic non-deterministic behavior of LLMs, which affects reproducibility not only in the present study but any research investigating LLM-generated outputs, particularly those employing single-run evaluations on benchmarks [Song *et al.*, 2024]. LLMs are fundamentally non-deterministic models, offering no guarantee that identical outputs will be generated across multiple executions, even when using the same input and instructions [Atil *et al.*, 2025].

This output instability directly impacts result reproducibility, a cornerstone of scientific research. While generation parameters such as temperature were configured to minimize randomness, other factors such as minimal variations in floating-point rounding, distributed computing utilization, and even the architectural essence of Transformer models themselves can influence their non-deterministic behavior [Yu, 2023; Atil *et al.*, 2025; Klishevich *et al.*, 2025].

Recognizing this challenge, several methodological approaches can mitigate the effects of non-determinism and increase consistency in LLMs. Recent research demonstrates that binary classification and sentiment analysis tasks can achieve near-perfect reproducibility [Wang and Wang, 2025]. Additionally, employing parsers for LLM responses amplifies consistency [Atil *et al.*, 2025], while Greedy Decoding, a technique employed whenever possible in this work ("*do_sample=False*"), demonstrates superior consistency compared to sampling approaches and LLMs tend to exhibit consistent performance on tasks with constrained output spaces [Song *et al.*, 2024]. These findings collectively provide methodological support for the approach adopted in this study.

The second consideration concerns the absence of data contamination assessment. Since LLMs are pre-trained on massive amounts of data, primarily sourced from the web, there is a risk that the LLMs examined in this study may have already been exposed to the evaluation datasets at some point during their training process. This potential exposure compromises the distinction between the models' generalization and memorization capabilities and could potentially overestimate the obtained results. This is therefore recognized as a methodological limitation of the present work.

5 Experimental Results and Discussions

5.1 Inference Costs, Duration Time and Carbon Emissions

Before discussing the experimental results, we present a comprehensive analysis of the computational costs associated with our experiments in terms of financial expenditure (in USD), inference duration, and estimated carbon emissions. These data are synthesized in Table 4, with carbon emissions estimated using the Machine Learning CO_2 Impact Calculator[14] Lacoste *et al.* [2019]. It is worth noting that, according to information provided by the Machine Learning CO_2 Impact Calculator, 100% of emissions generated by locally executed models were offset by the cloud provider.

Due to the proprietary nature of several LLMs (Claude-3.5-Sonnet, GPT-4o, Gemini-1.5-Pro, Sabiá-2-Medium, and Sabiá-3) and the computational requirements of others necessitating API consumption (DeepSeek-V3 and DeepSeek-R1), comprehensive inference time measurements and carbon emission estimations were not feasible for all models. This limitation stems from providers not supplying detailed operational metrics, particularly environmental impact data. The lack of carbon emissions data from API providers is further discussed in Section 6.

Another limitation relates to the cloud service provider selected for conducting the experiments. The Google Colab platform does not currently allow the selection of specific regions for server allocation, thereby precluding the choice of regions with enhanced energy efficiency for computational workloads.

[14] `https://mlco2.github.io/impact`

Table 4. Comparative analysis of inference time, cost and carbon footprint. The top group represents the models consumed via API, and the bottom group, the models deployed locally. Both groups are divided by a dashed line, which separates the generalist models from the fine-tuned models in PT-BR.

Model	Cloud Provider Region	Inferece Hours	Cost (USD)*	Carbon Emitted (kg CO_2eq)
Claude-3.5-Sonnet	-	4.60	$47.15	-
GPT-4o	-	9.63	$34.50	-
Gemini-1.5-Pro	-	-	$39.39	-
DeepSeek-V3	-	4.76	$25.15	-
DeepSeek-R1	-	73.37	$59.98	-
Sabiá-2-Medium	-	-	$15.18	-
Sabiá-3	-	-	$22.46	-
LLaMA-3-8B-Instruct	us-west4[§]	30.48	$6.08	0.53 - 0.66
LLaMA-3.1-8B-Instruct	us-west4	30.40	$6.06	0.53 - 0.66
Gemma-7B-Instruct	asia-southeast1	10.31	$2.06	0.31
Gemma-2-9B-Instruct	asia-southeast1	10.63	$2.12	0.32
Qwen-2-7B-Instruct	us-west4	3.00	$0.60	0.05 - 0.07
InternLM-2-7B-Chat	us-central1	7.16	$4.69	1.02
DeepSeek-R1-Distill-Qwen-7B	us-central1	4.68	$3.06	0.67
DeepSeek-R1-Distill-LLaMA-8B	asia-southeast1	5.11	$3.35	0.54
Sabiá-7B	us-west4	5.00	$1.00	0.09 - 0.11
Bode-7B	asia-southeast1	7.76	$1.55	0.23 [†]
Bode-13B	asia-southeast1	7.04	$1.40	0.21
Bode-3.1-8B-Instruct-lora	asia-southeast1	5.14	$3.36	0.54
InternLM-ChatBode-7B	us-west4	6.19	$1.23	0.11 - 0.14
GemBode-7B-Instruct	us-central1	6.41	$4.20	0.91
CabraLLaMA-3-8B	us-central1	4.78	$3.13	0.68
CabraMistral-v3-7B-32k	us-central1	5.76	$3.77	0.82
Total		**242.22**	**$291.46**	**7.56 - 7.89**

* 1 USD = 6.08 BRL
§ us-west4 region for GCP was not availabe at ML CO2 Impact Calculator, se we report the min and max values between regions us-west1, 2 and 3.
† Estimated using linear projection of Bode-13B and Gemma-7B-Instruct emissions values.

Regarding inference duration, we observed relatively balanced performance between API-consumed and locally deployed models. Notable exceptions include DeepSeek-R1, GPT-4o, LLaMA-3-8B-Instruct, and LLaMA-3.1-8B-Instruct. The substantially extended inference time for DeepSeek-R1 is attributable to its reasoning phase that precedes inference, considerably increasing execution duration.

To optimize financial resources, GPT-4o was accessed through its batch API, which reduces inference costs in exchange for extended response windows of up to 24 hours. The duration reported in Table 4 represents the total time from batch submission to complete processing. The extended inference times for LLaMA models primarily results from utilizing a configuration with a maximum number of new tokens set to 150, compared to 20 tokens for other models.

Beyond these configuration-specific factors, model-specific efficiency variations also significantly impact inference duration. An illustrative case involves the Gemma and Gemma-2 models, which, despite utilizing 4-bit quantized versions, maintained considerably elevated inference times (± 10 hours) compared to other models deployed on L4 hardware. Conversely, the Qwen-2 model, executed in full precision (Bfloat16) on identical L4 hardware, completed all inferences in merely 3 hours, establishing itself as one of the most computationally optimized models in our analysis. This comparison underscores the significant impact of model architecture and optimization on computational efficiency, independent of quantization strategies.

Concerning financial costs, a substantial disparity emerges between API-consumed and locally deployed models. Locally executed models incurred an average cost of $3.53, representing a 9.85-fold reduction compared to API-consumed models ($34.83). Among API-consumed models, the Sabiá family demonstrates cost-effectiveness, with proprietary models exhibiting lower costs than open-source alternatives accessed via API, exemplified by DeepSeek-V3.

Cost variations among locally deployed models are primarily attributable to hardware differences (L4 GPUs with 22.5GB RAM versus A100 with 40GB RAM) and total inference duration. On average, models running on A100 GPUs cost $0.65 per hour, approximately 3.25 times higher than those deployed on L4 GPUs ($0.20 per hour). However, this cost differential must be weighed against performance gains, as deploying models on A100 GPUs tends to reduce inference time to an average of 5.57 hours compared to 7.13 hours for L4 GPUs (excluding LLaMA-3 and LLaMA-3.1 models).

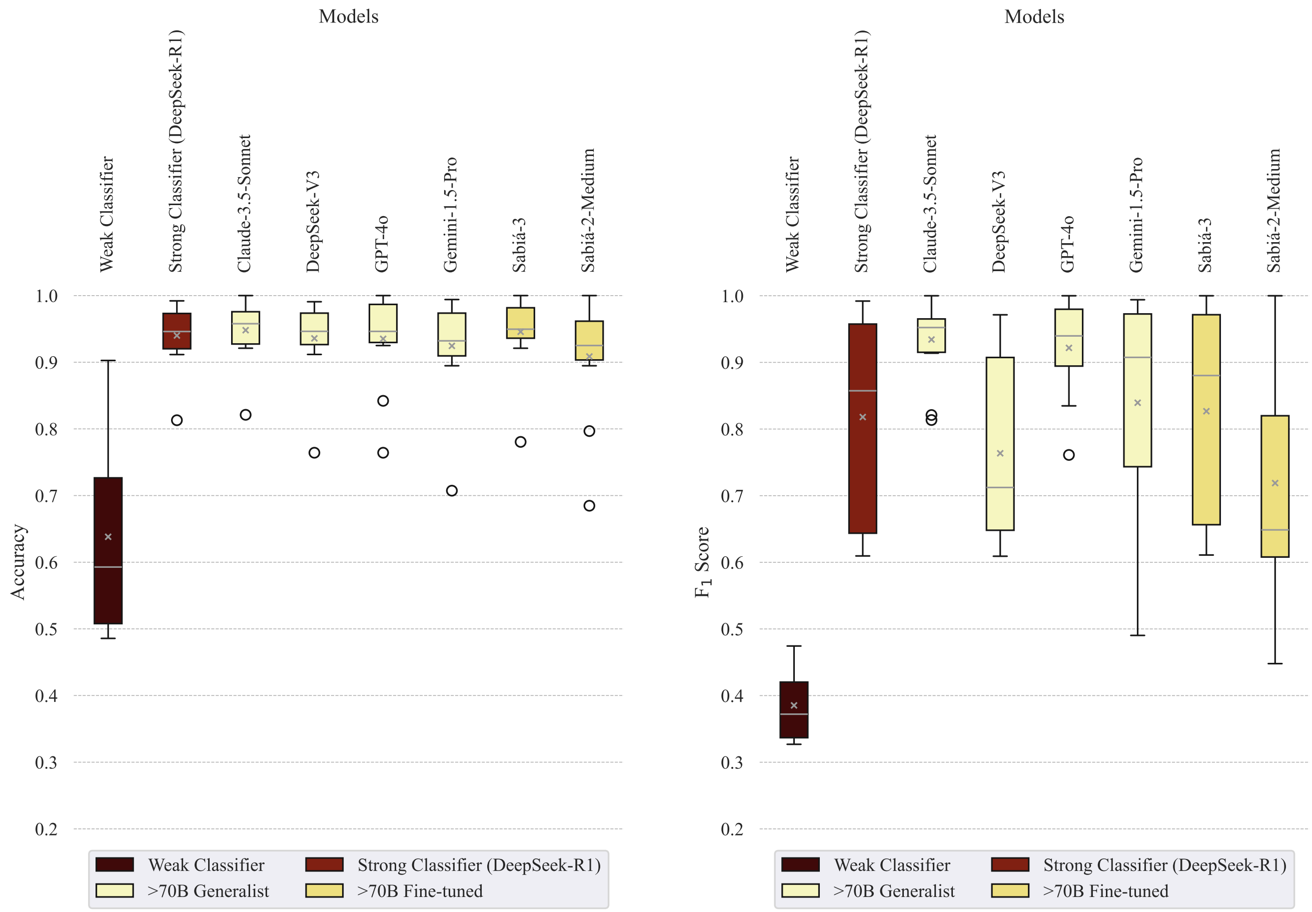
The environmental dimension of our computational analysis reveals equally significant patterns. Carbon emissions for locally deployed models were estimated to range between $7.56 - 7.89$ kg CO_2 equivalent (kg CO_2eq). These estimates were derived using IP addresses from Google Colab servers allocated during experimental execution to determine geographic regions and server locations. This geographic information was combined with per-model inference times and hardware specifications, subsequently input into the Machine Learning CO_2 Impact Calculator for emission calculations.

The carbon equivalent emissions of L4 machines were, on average, at least 2.45 times lower than A100 machines. The average emissions for A100 machines was 0.74 kg CO_2eq, while L4 computers ranged between 0.26 to 0.30 kg CO_2eq. To contextualize this comparison, we examine the LLaMA-3 models case, where inference time took approximately 30 hours while the average for other models was 6.36 hours. The estimated emissions range was 0.53 to 0.66 kg CO_2eq, which is comparable to 4 to 5 hours of A100 machine emissions, depending on the cloud provider's server region.

These findings collectively demonstrate that selecting appropriate hardware for LLM deployment involves a complex trade-off between computational efficiency, financial cost, and environmental impact. While higher-capacity GPUs GPUs offer superior computational performance with reduced inference times, specially for models that do not require intensive computational resources, they incur substantially higher environmental and financial costs. Although some models require more robust hardware configurations, deployment decisions merit careful consideration of these multifaceted implications.

5.2 Large-Scale Models Performance

The experimental evaluation for large LLMs encompassed Accuracy and F_1 Score metrics across 12 datasets. Performance statistics (mean and standard deviation) were aggregated per LLM, as presented in Table 5. The performance distribution across models is visualized in Figure 4a for Accuracy and Figure 4b for F_1 Score. Detailed dataset-specific results are available in Appendix A, with Accuracy and F_1 Score results presented in Table 8 and Table 9 respectively.



(a) Boxplot Accuracy by LLM. **(b)** Boxplot Macro F_1 Score by LLM.

Figure 4. Performance distribution of larger scale models across sentiment analysis tasks. The boxplots illustrate the statistical distribution of (a) Accuracy and (b) Macro F_1 Score, providing insights into performance variability across the evaluated datasets. The × marker represents the mean value for each evaluation metric, allowing comparison of central tendencies alongside the distribution spread, quartiles, and outliers.

Table 5. Performance comparison between small-scale LLMs (>70B parameters) in Brazilian Portuguese sentiment analysis tasks. Results present mean Accuracy and F_1 Score with standard deviations, stratified by linguistic specialization (Generalist vs. PT-BR fine-tuned) and ordered by decreasing Accuracy in each category. Two baseline classifiers are included as comparative reference.

Linguistic Fine-tuning	Model	Acc	F_1 Score
Baseline	Weak Classifier (Train Set Majority Class)	0.6382 ± 0.1458	0.3853 ± 0.0524
	Strong Classifier (DeepSeek-R1)	0.9401 ± 0.0488	0.8180 ± 0.1552
Generalist	Claude-3.5-Sonnet	**0.9481 ± 0.0482**	**0.9343 ± 0.0607**
	DeepSeek-V3	0.9358 ± 0.0600	0.7636 ± 0.1439
	GPT-4o	0.9351 ± 0.0687	0.9218 ± 0.0739
	Gemini-1.5-Pro	0.9245 ± 0.0760	0.8395 ± 0.1727
PT-BR	Sabiá-3	**0.9457 ± 0.0581**	**0.8267 ± 0.1588**
	Sabiá-2-Medium	0.9086 ± 0.0879	0.7189 ± 0.1676

Large-scale LLMs frequently achieve high-performance sentiment analysis in Brazilian Portuguese via in-context learning Analysis of Table 5 and Figure 4 reveals that all large-scale LLMs outperformed the Weak Classifier across both evaluation metrics. Furthermore, considerable parity is observed between the results obtained by the Strong Reference Classifier and the other large-scale LLMs, particularly regarding the primary evaluation metric, Accuracy. All large-scale models, whether with multilingual capabilities or fine-tuned for PT-BR, including the Strong Reference Classifier, produced mean accuracies exceeding 0.9. This finding provides strong evidence of the capability of large-scale LLMs to understand and execute the downstream task of binary sentiment classification in Brazilian Portuguese via the in-context learning paradigm.

The combined analysis of the Accuracy metric from Table 5 and Figure 4a indicates comparable performance among large-scale models, regardless of whether they are general-purpose or specifically optimized for Brazilian Portuguese. Claude-3.5-Sonnet, DeepSeek-V3, GPT-4, and Sabiá-3 exhibited lower variability (*Acc* standard deviation ranging from 0.0482 to 0.0687) and substantial performance overlap, while Gemini-1.5-Pro and Sabiá-2-Medium (0.076 and 0.0879) demonstrated marginally higher variability and slightly lower performance.

Regarding the F_1 score metric, Claude-3.5-Sonnet and GPT-4 models achieved consistently high values (average of 0.9343 and 0.9218, respectively) with minimal variability (0.0607 and 0.0739), whereas DeepSeek-V3, Sabiá-3, Gemini-1.5-Pro, and Sabiá-2 exhibited lower average values (0.7189 to 0.8395) with greater performance variability (standard deviation ranging from 0.1439 to 0.1727). This variability is primarily attributed to the occurrence of even a single response categorized as hallucination per dataset, which significantly impacts the F_1 Score calculation, reducing values by 33.33% and thereby increasing the overall variability for this metric compared to Accuracy.

Statistical tests indicate performance equivalence among large-scale LLMs in the downstream task. Statistical analysis using the Wilcoxon paired non-parametric test at 5% significance level confirmed significant differences between all LLMs and the Weak Classifier, as well as performance parity among the large-scale models themselves. Results for Accuracy and F_1 Score are presented in Table 7 and Table 8, respectively.

Among the large-scale models, the Wilcoxon signed-rank test results for Claude-3.5-Sonnet, DeepSeek-V3, GPT-4o, and Sabiá-3, when compared pairwise, do not support the idea that the distribution of differences between these groups is asymmetric. This suggests that the Accuracy scores obtained by these models across the 12 datasets are unlikely to be statistically different from each other.

Similarly, the results point in the same direction when comparing Claude-3.5-Sonnet and GPT-4o for the F_1 Score metric.However, when these two models are compared pairwise with the other large-scale LLMs, the results do not support the symmetry of distribution differences, indicating that their Macro F_1 Scores are likely superior. Both Claude-3.5-Sonnet and GPT-4o performed significantly better than Gemini-1.5-Pro and Sabiá-2-Medium, while the remaining LLMs (DeepSeek-V3, Sabiá-3, Gemini-1.5-Pro, and Sabiá-2-Medium) demonstrated statistically equivalent performance levels among themselves for this metric.

The comparison between general-purpose models and those fine-tuned for Brazilian Portuguese further supports conclusions favoring performance equivalence between these categories. Although statistical test results indicated significant differences between most general-purpose LLMs when compared to Sabiá-2-Medium, it is important to note that the latter essentially belongs to a previous generation. In contrast, the most recent LLM from the Sabiá family demonstrates statistical equivalence in comparison with all general-purpose large-scale LLMs.

Proprietary large-scale generalist models demonstrate superior reliability in prompt adherence with minimal hallucinations. Based on the results of the experiments reported in Table 11, Claude-3.5-Sonnet and GPT-4o were the only large-scale LLMs that followed the prompt specifications with complete consistency, generating responses that precisely matched the expected output format. However, it is important to note that, across all models, the percentage of responses categorized as hallucinations relative to the total number of responses produced (10, 372) was remarkably low, with an average of merely 0.25% for large-scale LLMs (including the strong reference model DeepSeek-R1).

Qualitative analysis of responses classified as hallucinations revealed distinct error patterns specific to each model architecture. The DeepSeek-R1 model, employed as the Strong Reference Classifier, exhibited consistent JSON syntax errors in all its hallucination cases, including omission of colons or quotation marks, and misspellings of the key term "polaridade". Similarly, the DeepSeek-V3 model produced some syntactically incorrect JSON structures, but its hallucinations were predominantly characterized by responses lacking the expected JSON format entirely, instead generating explanatory text describing its sentiment classification approach. This behavior likely stems from the knowledge distillation process from DeepSeek-R1, which enhances reasoning capabilities while significantly expanding average response length [DeepSeek-AI *et al.*, 2025b].

The hallucination patterns observed in other models revealed different underlying mechanisms. The Gemini-1.5-Pro model's hallucinations (100% of cases) were exclusively caused by its internal safety filters, which identified certain input texts as potentially violating usage policies, resulting in empty responses rather than sentiment classifications. In contrast, the Sabiá family of large-scale LLMs produced hallucinations primarily by assigning labels outside the specified binary range, particularly by generating correctly structured JSON objects containing a value of 0 (typically associated with neutral sentiment, which was not part of the task specification). Additionally, the Sabiá-2-Medium model frequently generated responses erroneously claiming defects or errors in the input text itself.

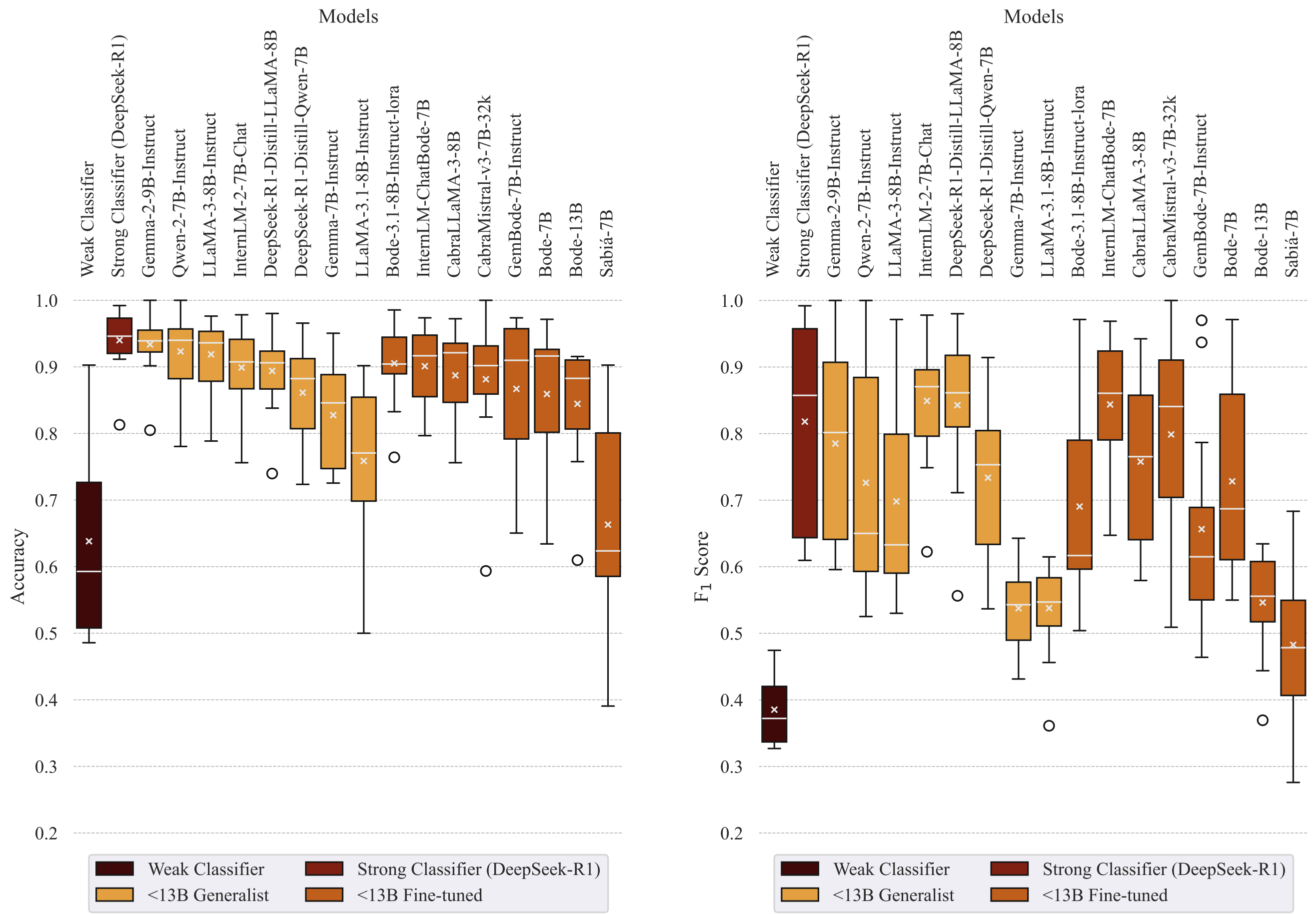
5.3 Small-Scale Models Performance

Similar to the approach taken with larger-scale models, descriptive statistics were also reported for smaller-scale LLMs (Table 6), along with evaluation metric variations summarized in boxplot graphs: Figure 5a for Accuracy and Figure 5b for F_1 Score.

Table 6. Performance comparison between small-scale LLMs (<13B parameters) in Brazilian Portuguese sentiment analysis tasks. Results present mean Accuracy and F_1 Score with standard deviations, stratified by linguistic specialization (Generalist vs. PT-BR fine-tuned) and ordered by decreasing Accuracy in each category. Two baseline classifiers are included as comparative reference.

Linguistic Fine-tuning	Model	Acc	F_1 Score
Baseline	Weak Classifier (Train Set Majority Class)	0.6382 ± 0.1458	0.3853 ± 0.0524
	Strong Classifier (DeepSeek-R1)	0.9401 ± 0.0488	0.7905 ± 0.1507
Generalist	Gemma-2-9B-Instruct	**0.9337 ± 0.0507**	0.7851 ± 0.1481
	Qwen-2-7B-Instruct	0.9232 ± 0.0604	0.7001 ± 0.1937
	LLaMA-3-8B-Instruct	0.9189 ± 0.0550	0.6472 ± 0.1943
	InternLM-2-7B-Chat	0.8990 ± 0.0657	0.8361 ± 0.1354
	DeepSeek-R1-Distill-LLaMA-8B	0.8939 ± 0.0640	**0.8427 ± 0.1219**
	DeepSeek-R1-Distill-Qwen-7B	0.8613 ± 0.0709	0.7215 ± 0.1446
	Gemma-7B-Instruct	0.8276 ± 0.0796	0.4915 ± 0.1212
	LLaMA-3.1-8B-Instruct	0.7587 ± 0.1250	0.4896 ± 0.0773
PT-BR	Bode-3.1-8B-Instruct-lora	**0.9054 ± 0.0613**	0.6778 ± 0.1722
	InternLM-ChatBode-7B	0.9010 ± 0.0622	**0.8438 ± 0.1040**
	CabraLLaMA-3-8B	0.8873 ± 0.0711	0.7252 ± 0.1854
	CabraMistral-v3-7B-32k	0.8814 ± 0.1046	0.7988 ± 0.1610
	GemBode-7B-Instruct	0.8670 ± 0.1056	0.6079 ± 0.2080
	Bode-7B	0.8593 ± 0.1095	0.7035 ± 0.1820
	Bode-13B	0.8445 ± 0.0911	0.5336 ± 0.0814
	Sabiá-7B	0.6630 ± 0.1558	0.4670 ± 0.1362

Small-scale LLMs prove to be efficient and viable alternatives for the underlying task. The evaluation of Table 6 and Figure 5 reveals that, similar to larger-scale models, small-scale LLMs achieved results consistently superior to the Weak Classifier in the vast majority of cases. Considering the results



(a) Accuracy Boxplot. **(b)** Macro F_1 Score Boxplot.

Figure 5. Performance distribution of smaller scale models across sentiment analysis tasks. The boxplots illustrate the statistical distribution of (a) Accuracy and (b) Macro F_1 Score, providing insights into performance variability across the evaluated datasets. The $\times$ marker represents the mean value for each evaluation metric, allowing comparison of central tendencies alongside the distribution spread, quartiles, and outliers.

obtained by the top 10 small-scale models (including both general-purpose and those fine-tuned for the target language), the average mean Accuracy was 0.8975, approximately 41% higher than the average of the weak reference classifier and only 4.5% lower than the strong reference classifier, despite the latter having, on average, $6.63 \cdot 10^{11}$ more parameters.

Regarding Accuracy, Table 6 and Figure 5a demonstrate that Gemma-2-9B-Instruct, Qwen-2-7B-Instruct, LLaMa-3-8B-Instruct, InternLM-2-7B-Chat and DeepSeek-R1-Distill-LLaMA-8B LLMs achieved results competitive with the strong classifier, as did the Brazilian Portuguese specialized models Bode-3.1-8B-Instruct-lora and InternLM ChatBode-7B, with their mean accuracies ranging between 0.8939 and 0.9337. Slightly behind are the PT-BR fine-tuned models CabraLLaMA-3-8B, CabraMistral-v3-7B-32k, GemBode-7B-Instruct, Bode-7B, and Bode-13B, with accuracies fluctuating between 0.8445 and 0.8882.

The models LLaMA-3.1-8B-Instruct and the Sabiá-7B LLM, based on first-generation LLaMA, demonstrated the lowest average performance in terms of the main metric, with the latter approaching the performance observed in the Weak classifier.

Lower values along with greater variability were also observed for the F_1 Score metric which, similar to larger-scale models, can be primarily explained through the lens of responses categorized as hallucinations. Most models achieved results close to the strong reference classifier, except for the general-purpose LLMs Gemma-7B-Instruct and LLaMA-3.1-8B-Instruct together with the Brazilian Portuguese fine-tuned models Bode-13B and Sabiá-7B, which exhibited the worst average performance for this metric, remaining relatively close to the weak reference classifier.

Overall, small-scale models generated a considerably low percentage of responses categorized as hallucinations, but they struggled with overgeneration. Excluding the general-purpose models Gemma-7B-Instruct and LLaMA-3.1-8B-Instruct and the PT-BR fine-tuned models Sabiá-7B and Bode-13B, which were most affected by hallucinations (average of 8.62%), the remaining small-scale LLMs produced a low percentage of responses outside the specified pattern. The average for the remaining general-purpose models was 0.20% and for those specialized in the target language was 0.26%, both relatively close to the value obtained by the strong reference model.

Despite being minimally affected by hallucinations, qualitative analysis of responses generated by small-scale models, both general-purpose and specialized, revealed a high rate of overgeneration. This means that small-scale models appeared to have greater difficulty recognizing when the requested task had been completed, generating numerous additional unnecessary/unwanted tokens beyond those used to compose the requested JSON object, typically continuing until reaching the maximum token limit established as a generation parameter.

This overgeneration behavior is clearly demonstrated in the raw outputs produced by smaller-scale models, with detailed examples available in Appendix C.2. Both generalist models (such as InternLM-2-7B-Chat and Gemma-7B-Instruct) and Portuguese fine-tuned variants (including InternLM-ChatBode-7B and CabraMistral-v3-7B-32k) exhibit similar patterns. Their outputs characteristically contain the correct JSON structure with sentiment classification, but are systematically contaminated by extraneous artifacts, mostly derived from prompt elements. For instance, typical responses include fragments such as "*{'polaridade': 1}\nExemplo:\n'entrada': 'O que é*" or "*{'polaridade': 1}\nClassificação de Sentimento:\n'entr*", illustrating this pervasive issue.

While this verbose behavior affects most small-scale models, a notable exception includes Qwen-2-7B-Instruct, which demonstrates markedly superior output consistency that more closely approximates the concise response patterns observed in large-scale models (Appendix C.1). A comprehensive qualitative analysis of output patterns across all evaluated models is provided in Appendix C.

To quantify the influence of overgeneration in small-scale models, the number of output tokens produced by each LLM was calculated using a common tokenizer (OpenAI Tiktoken *o200k_base*[15]). From this calculation, Figure 6 was produced, illustrating the distribution of tokens by model size, and Table 7, available in Appendix A, which consolidates the main descriptive statistics regarding the output tokens produced.

Analysis of Figure 6 reveals a higher concentration of output tokens in two distinct intervals: first, a peak near 7 tokens is identifiable, followed by a more dispersed distribution between 13 and 20 tokens for smaller-scale LLMs (< 13B), while larger-scale models also exhibit this concentration near 7 tokens with another peak around 11 tokens. Quantitatively (Table 7), a higher average number of tokens is observed for smaller-scale models (14.86), excluding LLaMA-3-8B-Instruct and LLaMA-3.1-8B-Instruct from the comparison, when compared to LLMs with more than 70B parameters (9.61). These two LLaMA models were excluded from this specific analysis because they demonstrated notably literal adherence to the JSON object structure as discussed in Subsection 4.5, necessitating a higher maximum output token limit (150) compared to the 20-token limit used for other models.

The overgeneration phenomenon observed predominantly in smaller-scale models presents significant implications for practical applications. While it did not substantially impede the workflow of this study—as regex pattern matching effectively extracted the required JSON objects from verbose outputs—it nevertheless constitutes an important consideration for real-world implementations. This behavior may implies on additional post-processing steps when integrating these models into production systems, creating overhead that could impact efficiency and resource utilization.

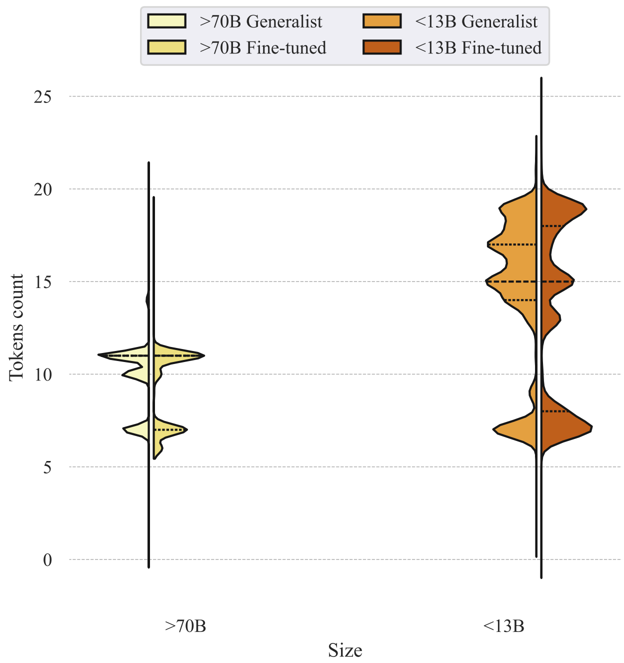


Figure 6. Distribution of output token counts across models by parameter size and scale. The analysis illustrates token generation patterns for both large and small-scale models, with comparative plots highlighting outputs token count distribution and variability. This visualization excludes data from LLaMA-3-8B-Instruct, LLaMA-3.1-8B-Instruct, and the Strong Reference Classifier (DeepSeek-R1) to maintain focus on models with comparable generation behaviors.

During analysis, we hypothesized that prompt complexity may have been a primary contributing factor to overgeneration, specifically the structural complexity of our prompt which contained, besides the task instructions and the text to be classified, a JSON schema definition and few-shot examples. Notably, we found very limited research addressing this specific phenomenon in the literature, suggesting that overgeneration might be a prompt-specific issue unique to this study's experimental design rather than a universal limitation of these models.

On the other hand, the fact that models generated superfluous tokens despite explicit instructions to produce only a JSON object highlights a potential limitation in their instruction-following capabilities when presented with our particular prompt structure. We leave this hypothesis open for future research, which might also explore whether architectural modifications, prompt engineering techniques, or specialized fine-tuning could mitigate this overgeneration tendency without compromising the models' performance on the primary sentiment analysis task.

Linguistic specialization of models in the target language, in some cases, tends to reduce hallucinations. In several cases, it was possible to directly compare the effect of linguistic specialization by examining LLMs alongside their base models, such as InternLM-2-7B-Chat versus InternLM-ChatBode-7B, Gemma-7B-Instruct versus GemBode-7B-Instruct, LLaMA-3-8B versus CabraLLaMA-3-8B and LLaMA-3.1-8B-Instruct versus Bode-3.1-8B-Instruct-lora.

This comparative analysis revealed that linguistic fine-tuning seems to reduce the number of hallucinations produced by the models by an average of 92% for Bode's Family

[15] `https://github.com/openai/tiktoken`

against their base versions. Specifically, there was a reduction of 942 hallucinations (-96.2%) from Gemma-7B-Instruct to GemBode-7B-Instruct, 827 (-96.5%) from LLaMA-3.1-8B-Instruct to Bode-3.1-8B-Instruct-lora, and 5 (-83.3%) from InternLM-2-7B-Chat to InternLM-ChatBode-7B.

On the other hand, the Cabra family model CabraLLaMA-3-8B showed an increase of 39 ($+125\%$) hallucinations compared to the results obtained from its base model. Despite the increase in the number of responses categorized as hallucinations, there was less dispersion regarding the hallucinations and the datasets in which they occurred, with a reduction of 50% (4 datasets) compared to LLaMA-3-8B (8 datasets).

For Bode-3.1-8B-Instruct-lora, a substantial improvement in average performance compared to its base model was observed across both evaluation metrics (19% in Accuracy and 38% in F_1 Score), with this improvement also reflected in the statistical tests (Figure 7 and Figure 8).

Improvements were also observed in the other two PT-BR fine-tuned models relative to their general-purpose versions, though these were less pronounced: a gain of 5% in mean Accuracy and 24% in mean F_1 Score between Gemma-7B-Instruct and GemBode-7B-Instruct, and a gain of 0.22% in mean Accuracy and 0.92% in mean F_1 Score from InternLM-2-7B-Chat to InternLM-ChatBode-7B. These smaller improvements did not constitute sufficient evidence to accept the alternative hypothesis of statistical superiority, except in the case of the F_1 Score metric between the Gemma-7B-Instruct and GemBode-7B-Instruct models.

Meanwhile, the CabraLLaMA-3-8B model showed an improvement of 8.57% in mean F_1 Score compared to the LLaMA-3-8B model, largely due to the reduction in the dispersion of the total number of datasets with at least one hallucination, while the mean Accuracy exhibited a slight reduction of 3.44% relative to its base model.

5.4 Cross-Scale Models Comparison

Comparative analysis with previous reported results highlights the potential of in-context learning with LLMs over conventional supervised learning approaches for this particular sentiment analysis task in Brazilian Portuguese. Evidence from this investigation substantiates that both large-scale and small-scale LLMs achieved promising performance in binary sentiment classification of Brazilian Portuguese texts using the ICL paradigm. This conclusion is supported not only by the magnitude of the obtained results but also through comparative evaluations with findings reported in prior research.

A pattern emerged regarding performance variations across datasets. For both large and small-scale models, the majority of the lowest accuracy values (19 out of 23 models) were observed on the same dataset: OPCovidBR [Vargas *et al.*, 2020]. This dataset accounts for 5 of the 9 outliers visible in Figure 4a for large-scale models, and for all 5 outliers (excluding the previously counted strong reference classifier) in Figure 5a for small-scale LLMs. Quantitatively, large-scale models (including the strong reference classifier) achieved a mean accuracy of 0.7782 on this dataset, while small-scale models averaged 0.7104, indicating greater difficulty in correctly interpreting and classifying texts from this particular corpus.

Despite these challenges, the Accuracy achieved by most of the LLMs significantly exceeded the results reported by Vargas *et al.* [2020], in which traditional classifiers including Naive Bayes, Decision Trees, and SVM—specifically trained on this dataset—achieved accuracy scores ranging between 0.48 and 0.63 (accuracy results were obtained from the supplementary materials available in the paper's repository).

Some recent-generation smaller-scale LLMs (between 7 and 13 billion parameters) demonstrate statistically equivalent performance to SOTA large-scale models (over 70 billion parameters) for Portuguese sentiment analysis. Considering the primary evaluation metric, Accuracy, the analysis showed statistical equivalence (Figure 7) between large-scale SOTA models such as DeepSeek-V3 (mean Accuracy: 0.9358) and GPT-4o (mean Accuracy: 0.9351) when compared with smaller-scale LLMs such as Gemma-2-9B-Instruct (mean Accuracy: 0.9337) and Qwen-2-7B-Instruct (mean Accuracy: 0.9232). This pattern of statistical equivalence was also observed in the comparison between GPT-4o and the smaller-scale Brazilian Portuguese specialized model InternLM-ChatBode-7B (mean Accuracy: 0.9010).

These findings represent a significant contribution to the debate on the relationship between scale and performance in NLP tasks for languages beyond English. Although models with a larger number of parameters, such as Claude-3.5-Sonnet and Sabiá-3, achieved the best absolute Accuracy averages (0.9481 and 0.9457, respectively), the ability of smaller and more recent models to achieve statistically comparable performance challenges the premise that larger scale necessarily results in better performance for specific tasks in languages not dominant in training datasets.

The demonstration that smaller-scale models can effectively compete with large LLMs has significant practical implications. These smaller models represent viable alternatives both from a performance perspective and computational efficiency, enabling their execution on more modest and economically accessible hardware infrastructures. This characteristic amplifies their potential application in contexts with computational resource constraints, particularly relevant for researchers and developers working with Brazilian Portuguese.

The experimental results reveal a consistent pattern of cross-generational improvements within language model families when evaluated on Brazilian Portuguese sentiment analysis. This evolution manifests in multiple performance dimensions including mean Accuracy, mean Macro F_1 Score, and hallucination rates. The comparative analysis demonstrates how newer generations of the same model family tend to exhibit enhanced capabilities in processing Portuguese text.

The results indicate a significant improvement in Brazilian Portuguese performance of Gemma-2-9B-Instruct compared to its previous version, Gemma-7B-Instruct. While Gemma-7B-Instruct achieved an average Accuracy of 0.8276 and an average F_1 Score of 0.4915 (Table 6), ranking lower among smaller-scale models, the Gemma-2-9B-Instruct version recorded an average Accuracy of 0.9337 and an F_1 Score of 0.7851, along with a 98.67% reduction in hallucinations, positioning it as the top general small-scale model. It is worth noting that the developers emphasize that neither the first nor the second generation of Gemma models have multilingual aspirations [Gemma Team *et al.*, 2024a,b].

The statistically significant difference in Accuracy (Figure 7) and Macro F_1 Score (Figure 8) between LLaMA-3-8B and earlier LLaMA-based models such as Bode-7B, Bode-13B, and Sabiá-7B demonstrates the evolution of LLaMA's capabilities for Brazilian Portuguese across generations. This is evidenced by results reported by Pires *et al.* [2023] and Garcia *et al.* [2024], which indicate the superiority of Sabiá-7B models compared to their base LLaMA-7B model, and of Bode7B and 13B compared to LLaMA-2-7B and LLaMA-2-13B.

The LLaMA family has been expanding its multilingual capabilities with each new version [Meta, 2024], with LLaMA-3.1-8B being explicitly designed to offer enhanced support for multiple languages including Portuguese. However, despite this targeted multilingual expansion, LLaMA-3.1-8B showed significantly lower performance than LLaMA-3-8B (which does not claim specific multilingual capabilities) [Grattafiori *et al.*, 2024] in both evaluation metrics. Furthermore, version 3.1 produced 27 times more responses categorized as hallucinations than version 3.

Qualitative analysis of the responses generated by LLaMA-3.1-8B revealed a pattern of task misinterpretation that warrants further examination. In responses classified as hallucinations, rather than producing the required JSON output object, the model frequently generated code snippets related to machine learning algorithms for sentiment classification. This behavior suggests potential limitations in prompt understanding or instruction following. One plausible explanation is, again, the relative complexity of the provided prompt structure, which may have exceeded the model's ability to accurately parse and respond to multi-part instructions in Portuguese.

This hypothesis is particularly noteworthy given that, according to the developers [Grattafiori *et al.*, 2024], LLaMA-3.1-8B was specifically designed with enhanced multilingual support including Portuguese, which theoretically should have resulted in superior performance compared to LLaMA-3-8B. These contradictory findings highlight the importance of prompt engineering and testing when deploying multilingual language models, as expanded language capabilities may not necessarily translate to improved task performance across all instruction contexts.

The evolution of large-scale models fine-tuned for Brazilian Portuguese was also evident in the experimental results. The comparative analysis between specialized PT-BR models, Sabiá-3 and Sabiá-2-Medium (as illustrated in Figure 7 and Figure 8), indicated a rejection of the null hypothesis (H_0) for both evaluated metrics. These findings suggest that Sabiá-3 significantly outperforms Sabiá-2-Medium in terms of both average Accuracy (0.9457 versus 0.9086) and average F_1 Score (0.8267 versus 0.7189). Additionally, Sabiá-3 demonstrated enhanced response reliability, exhibiting a substantially lower hallucination rate (0.11% compared to 0.66%), which represents an approximate 83% reduction in hallucinations.

6 Limitations

This research presents several methodological and scope limitations that warrant consideration. The experimental design decisions, while methodologically justified, introduce specific constraints that may influence the interpretation and generalizability of our findings.

As discussed in Section 4.4, the random selection of 6 demonstrations for In-Context Learning, while methodologically feasible, may introduce instabilities in language model performance. Recent studies [Liu *et al.*, 2022; Lu *et al.*, 2022; Rubin *et al.*, 2022; Ye *et al.*, 2023] demonstrate that systematic and automated selection and ordering of demonstrations can significantly enhance predictive performance.

Similarly, the manual construction of prompts used in the experiments, although following established guidelines for optimizing response effectiveness, may not have fully explored the optimization potential that other methods could provide, possibly resulting in sub-optimal model performance [Reynolds and McDonell, 2021; Zhou *et al.*, 2022; Wang *et al.*, 2022]. This limitation represents a methodological consideration that may have served as a potential confounder in our comparative analysis, as different models may exhibit varying degrees of sensitivity to prompt formulation and demonstration selection strategies.

Furthermore, the complexity and structure of the crafted prompt may differentially affect models performance, with smaller-scale language models potentially exhibiting greater sensitivity to prompt complexity compared to their larger counterparts. Future research should investigate how prompt complexity influences model performance across different architectural scales and explore diverse prompt engineering techniques to identify approaches that are both adequate and effective for the majority of evaluated models, thereby reducing the methodological bias introduced by manual prompt construction.

A fundamental methodological consideration that permeates this entire study concerns the inherent non-deterministic behavior of LLMs, which directly impacts the reproducibility of our findings. This output variability stems from multiple sources including algorithmic factors (sampling strategies and model architecture), implementation aspects (floating-point precision variations, distributed computing and optimizations), and system-level considerations [Yu, 2023; Song *et al.*, 2024; Atil *et al.*, 2025; Klishevich *et al.*, 2025].

Although we implemented conservative generation parameters and employed structured response parsing to enhance consistency, the single-run evaluation approach adopted, while methodologically justified by computational and financial constraints, limits the statistical robustness of our comparative conclusions. Future research should consider multi-run evaluations with appropriate statistical analysis to better characterize the variance inherent in LLM performance assessments.

Compounding these reproducibility challenges, our experimental setup utilized different hardware configurations (L4 GPUs and A100 GPUs) across models, which may introduce subtle variations in computational outcomes. Additionally, the consumption of proprietary APIs for several key models presents ongoing challenges, as these systems undergo continuous updates without public notification, potentially altering their behavior between evaluation periods and compromising long-term reproducibility.

Another methodological limitation concerns the absence of comprehensive evaluation regarding potential model contamination with respect to the datasets used in this study. Large language models are trained on vast corpora of text, which may include portions of public datasets similar or identical to those used in our evaluation.

This contamination could introduce biases in our analysis, potentially inflating performance metrics for certain models while providing an inaccurate representation of their actual generalization capabilities [Sainz *et al.*, 2023; Dong *et al.*, 2024]. Future work should implement rigorous contamination detection methods to ensure that performance evaluations reflect genuine model capabilities rather than memorization of previously encountered data [Elangovan *et al.*, 2021].

Furthermore, a relevant methodological and ethical limitation is the absence of a systematic investigation into biases, such as social and demographics. This limitation is primarily linked to an inherent challenge in the NLP field: the construction of datasets that are simultaneously comprehensive, high-quality, and annotated to permit the analysis of diverse biases. Creating datasets with these characteristics is a highly complex task, involving substantial costs in time and resources. As a result, it is common for developers of such resources to prioritize certain features over others.

Thus, conducting a bias analysis on existing datasets becomes an initiative as complex and costly as creating a new resource annotated specifically for this purpose. Consequently, our analysis could not determine whether the models exhibit differential performance across the diverse linguistic variations of Brazilian Portuguese or among different demographic groups. This gap is particularly relevant given that the training data of LLMs themselves may also not equitably represent all segments of speakers, specially in low resources languages, introducing another latent biases that our evaluation was unable to detect. Therefore, we encourage future work to consider methodologies that enable the analysis of these biases.

Beyond methodological constraints, the study's scope presents important limitations regarding the breadth of evaluation. The analysis focused on a restricted set of models (23) and exclusively evaluated binary sentiment classification tasks. This delimitation may restrict the generalization of results to other natural language processing tasks, limiting the applicability of findings in broader contexts where different linguistic phenomena, task complexities, or domain-specific requirements might reveal alternative performance patterns.

Finally, environmental and transparency considerations represent an emerging limitation that extends beyond this study to the broader field of LLM research. As discussed in Subsection 5.1, the absence of carbon footprint data from proprietary models and/or those consumed via API, or at minimum, information that would enable estimation of these values, represents a limitation not only of the present study but of all research utilizing these LLMs.

This lack of transparency, often concealed behind commercial justifications but also resulting from the absence of standardized guidelines for climate reporting [Hershcovich *et al.*, 2022], may obscure the real effects of the complex interaction between utility benefits and environmental costs [Strubell *et al.*, 2020; Bender *et al.*, 2021]. While open-source models allow for more precise environmental and computational cost assessments, the proprietary nature of leading commercial LLMs prevents comprehensive environmental impact evaluation across all evaluated models.

7 Conclusion

This study conducted an extensive comparative analysis of Large Language Models' capabilities in binary sentiment classification for Brazilian Portuguese texts. We evaluated 23 LLMs comprising 13 state-of-the-art multilingual models and 10 models fine-tuned specifically for the Portuguese language, testing their performance across 12 annotated datasets using the in-context learning paradigm.

Our findings demonstrate that both large-scale and small-scale LLMs exhibit significant effectiveness in sentiment analysis of Brazilian Portuguese texts. Large models such as Claude-3.5-Sonnet, DeepSeek-V3, GPT-4o, and Sabiá-3 achieved outstanding results, with average accuracies exceeding 93% and minimal hallucination rates. Notably, the specialized model Sabiá-3 performed comparably to leading multilingual models, indicating that high-quality language-specific optimization can match the capabilities of general-purpose large-scale LLMs.

Smaller models (7-13B parameters) also demonstrated competitive performance, with top performers like Gemma-2-9B-Instruct, Qwen-2-7B-Instruct, and LLaMA-3-8B-Instruct achieving accuracies above 91%. Among Portuguese-specialized smaller models, Bode-3.1-8B-Instruct-lora and InternLM-ChatBode-7B showed the most promising results. These findings suggest that smaller, more efficient models can serve as viable alternatives for practical applications in resource-constrained environments.

Our comparative analysis revealed several noteworthy patterns. First, newer generations within model families consistently outperformed their predecessors in Brazilian Portuguese sentiment analysis, highlighting the rapid advancement in LLM capabilities. Second, linguistic specialization through fine-tuning demonstrated mixed results—while substantially reducing hallucination rates for some models (particularly in the Bode family), it did not consistently yield significant performance improvements across all metrics and model types.

The study also uncovered interesting behavioral patterns among different model categories. Small-scale models exhibited a tendency toward overgeneration despite explicit instructions, producing additional unnecessary text beyond the requested format. This finding suggests that further research into prompting techniques and model adaptation may be beneficial for optimizing these models for structured output tasks.

In the broader context of sentiment analysis for Brazilian Portuguese, our experimental results significantly outperformed previously reported benchmarks that used traditional machine learning approaches specifically trained for this task. This demonstrates the considerable potential of in-context learning with LLMs as an efficient alternative to traditional supervised learning approaches for Portuguese NLP tasks.

Future research directions could address several limitations of the current study. First, developing systematic methodologies for demonstration selection and prompt optimization could further enhance models performance. Second, expanding the evaluation to include more complex NLP tasks beyond binary sentiment classification would provide a more comprehensive assessment of these models' capabilities in Portuguese. Finally, a deeper qualitative analysis of selected datasets and LLMs could yield important findings about biased performance across different demographic groups or linguistic variations within Brazilian Portuguese.

In conclusion, this study contributes to the growing body of research on multilingual and language-specialized LLMs by providing empirical evidence of their effectiveness in Portuguese natural language processing. The results demonstrate that both approaches—general-purpose multilingual models and Portuguese-specialized models—offer viable paths forward, with their relative advantages depending on specific use cases and deployment constraints.

Declarations

Acknowledgements

We would like to express our sincere gratitude to Maritaca AI for their generous provision of credits that made this study possible. The company's support was fundamental for the advancement of this research, allowing us to explore the capabilities of different language models in the specific context of Brazilian Portuguese. This collaboration exemplifies the company's commitment to scientific progress and the development of natural language processing technologies adapted to the linguistic particularities of Brazil.

Authors' Contributions

André da Fonseca Schuck contributed to the conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, software, resources, validation, visualization and writing of the original draft. Gabriel Lino Garcia, João Renato Riberito Manesco, Pedro Henrique Paiola and João Paulo Papa performed supervision and writing - review & editing. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors are grateful to the São Paulo Research Foundation (FAPESP) grants 2013/07375-0, 2023/14427-8, and 2024/00789-8, and also to the Brazilian National Council for Scientific and Technological Development (CNPq) grants 308529/2021-9 and 400756/2024-2. Partial support was received from Maritaca AI in the form of API credits that were essential for conducting the experiments with their language models.

Availability of data and materials

The datasets (and/or softwares) generated and/or analysed during the current study are available in the Github repository:
`https://github.com/AndreSchuck/EvaluatingLargeLanguageModelsforBrazilianPortugueseSentimentAnalysis`

References

Abonizio, H., Almeida, T. S., Laitz, T., *et al*. (2024). Sabiá-3 technical report. DOI: 10.48550/arXiv.2410.12049.

Ainslie, J., Lee-Thorp, J., de Jong, M., *et al*. (2023). Gqa: Training generalized multi-query transformer models from multi-head checkpoints. DOI: 10.48550/arXiv.2305.13245.

Anthropic (2023). Introducing claude. Available at:`https://www.anthropic.com/news/introducing-claude`.

Anthropic (2024a). Claude 3 model card. Available at:`https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf`.

Anthropic (2024b). Claude 3.5 sonnet. Available at:`https://www.anthropic.com/news/claude-3-5-sonnet`.

Anthropic (2024c). Introducing the next generation of claude. Available at:`https://www.anthropic.com/news/claude-3-family`.

Araujo, M., Reis, J., Pereira, A., *et al*. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, page 1140–1145. Association for Computing Machinery. DOI: 10.1145/2851613.2851817.

Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., Rajagopal, G. R., Sloan, A., Tudrej, T., Ture, F., Wu, Z., Xu, L., and Baldwin, B. (2025). Non-determinism of "deterministic" llm settings. DOI: 10.48550/arXiv.2408.04667.

Bai, J., Bai, S., Chu, Y., *et al*. (2023). Qwen technical report. arXiv.org. DOI: 10.48550/arXiv.2309.16609.

Belisário, L., Luiz G., F., and Thiago A. S., P. (2019). Classificação de subjetividade para o português: Métodos baseados em aprendizado de máquina e em léxico. In *27º Simpósio Internacional de Iniciação Científica e Tecnológica da USP (SIICUSP)*, pages 1–1. Available at:`https://repositorio.usp.br/bitstreams/0fd1a3e6-5182-496d-ad1b-245d33c3b424`.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. DOI: 10.48550/arXiv.2004.05150.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3442188.3445922.

BotBot AI (2024a). Botbot. Available at:https://www.botbot.bot/.

BotBot AI (2024b). botbot-ai/. https://huggingface.co/botbot-ai. Available at:https://huggingface.co/botbotrobotics.

BotBot AI (2024c). botbot-ai/cabrallama3-8b · hugging face. Available at:https://huggingface.co/botbot-ai/CabraLlama3-8b.

Brown, T. B., Mann, B., Ryder, N., *et al.* (2020). Language models are few-shot learners. *arxiv.org*, 33:1877–1901. DOI: 10.48550/arXiv.2005.14165.

Brum, H. and das Graças Volpe Nunes, M. (2018). Building a Sentiment Corpus of Tweets in Brazilian Portuguese. In chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). DOI: 10.48550/arxiv.1712.08917.

Buscemi, A. and Proverbio, D. (2024). Chatgpt vs gemini vs llama on multilingual sentiment analysis. arXiv.org. DOI: 10.48550/arXiv.2402.01715.

Cai, Z., Cao, M., Chen, H., *et al.* (2024). Internlm2 technical report. DOI: 10.48550/arXiv.2403.17297.

Chen, B., Zhang, Z., Langrené, N., *et al.* (2023). Unleashing the potential of prompt engineering in large language models: a comprehensive review. arXiv.org. DOI: 10.48550/arXiv.2310.14735.

Chowdhery, A., Narang, S., Devlin, J., *et al.* (2022). Palm: Scaling language modeling with pathways. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2204.02311.

Cui, Y., Yang, Z., and Yao, X. (2024). Efficient and effective text encoding for chinese llama and alpaca. DOI: 10.48550/arXiv.2304.08177.

Cui, Y. and Yao, X. (2024). Rethinking llm language adaptation: A case study on chinese mixtral. DOI: 10.48550/arXiv.2403.01851.

de Araujo, G., de Melo, T., and Figueiredo, C. M. S. (2024). Is chatgpt an effective solver of sentiment analysis tasks in portuguese? a preliminary study. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*. Available at:https://aclanthology.org/2024.propor-1.2.pdf.

DeepSeek-AI, :, Bi, X., *et al.* (2024a). Deepseek llm: Scaling open-source language models with longtermism. DOI: 10.48550/arxiv.2401.02954.

DeepSeek-AI, Guo, D., Yang, D., *et al.* (2025a). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. DOI: 10.48550/arXiv.2501.12948.

DeepSeek-AI, Liu, A., Feng, B., *et al.* (2024b). Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. DOI: 10.48550/arXiv.2405.04434.

DeepSeek-AI, Liu, A., Feng, B., *et al.* (2025b). Deepseek-v3 technical report. DOI: https://doi.org/10.48550/arxiv.2412.19437.

Dettmers, T., Pagnoni, A., Holtzman, A., *et al.* (2023). Qlora: Efficient finetuning of quantized llms.

Devlin, J., Chang, M.-W., Lee, K., *et al.* (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv.org*. DOI: https://doi.org/10.48550/arxiv.1810.04805.

Ding, X., Chen, L., Emani, M., *et al.* (2023). Hpc-gpt: Integrating large language model for high-performance computing. In *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*, SC-W '23, page 951–960, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3624062.3624172.

Dong, Q., Li, L., Dai, D., *et al.* (2023). A survey on in-context learning. In *arXiv.org*. DOI: 10.48550/arXiv.2301.00234.

Dong, Y., Jiang, X., Liu, H., *et al.* (2024). Generalization or memorization: Data contamination and trustworthy evaluation for large language models. DOI: 10.48550/arXiv.2402.15938.

dos Santos Silva, L. N., Real, L., Zandavalle, A. C. B., *et al.* (2024). Repro: a benchmark for opinion mining for brazilian portuguese. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *ACLWeb*, page 432–440. Association for Computational Lingustics. Available at:https://aclanthology.org/2024.propor-1.44.pdf.

Du, X., Yu, Z., Gao, S., *et al.* (2024). Chinese tiny llm: Pretraining a chinese-centric large language model. DOI: 10.48550/arXiv.2404.04167.

Elangovan, A., He, J., and Verspoor, K. (2021). Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-main.113.

Freitas, C., Motta, E., Milidiú, R., *et al.* (2014). Sparkling vampire... lol! annotating opinions in a book review corpus. *New language technologies and linguistic research: a two-way Road*, pages 128–146. Available at:https://www.researchgate.net/publication/271836545_Sparkling_Vampire_lol_Annotating_Opinions_in_a_Book_Review_Corpus.

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

Garcia, G. L., Paiola, P. H., Garcia, E., *et al.* (2025). Gembode and phibode: Adapting small language models to brazilian portuguese. In Hernández-García, R., Barrientos, R. J., and Velastin, S. A., editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 228–243, Cham. Springer Nature Switzerland. DOI: 10.1007/978-3-031-76607-7_17.

Garcia, G. L., Paiola, P. H., Morelli, L. H., *et al.* (2024). Introducing bode: A fine-tuned large language model for portuguese prompt-based task. In *arXiv.org*. DOI: 10.48550/arXiv.2401.02909.

Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., and et al (2023). Gemini: A family of highly capable multimodal models. In *arXiv.org*. DOI:

10.48550/arXiv.2312.11805.

Gemma Team, Mesnard, T., Hardin, C., *et al*. (2024a). Gemma: Open models based on gemini research and technology. arXiv.org. DOI: 10.48550/arXiv.2403.08295.

Gemma Team, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., and et al (2024b). Gemma 2: Improving open language models at a practical size. DOI: 10.48550/arXiv.2408.00118.

Giray, L. (2023). Prompt engineering with chatgpt: A guide for academic writers. *Annals of Biomedical Engineering*, 51:2629–2633. DOI: 10.1007/s10439-023-03272-4.

Goyal, N., Gao, C., Chaudhary, V., *et al*. (2022). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538. DOI: 10.1162/tacl_a_00474.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., and et al (2024). The llama 3 herd of models. DOI: 10.48550/arXiv.2407.21783.

Gunasekar, S., Zhang, Y., Aneja, J., *et al*. (2023). Textbooks are all you need. DOI: 10.48550/arXiv.2306.11644.

Han, X., Zhang, Z., Ding, N., *et al*. (2021). Pre-trained models: Past, present and future. In *arXiv.org*. DOI: 10.48550/arXiv.2106.07139.

Hartmann, J., Heitmann, M., Siebert, C., *et al*. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87. DOI: 10.1016/j.ijresmar.2022.05.005.

Hendrycks, D., Burns, C., Basart, S., *et al*. (2020). Measuring massive multitask language understanding. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2009.03300.

Hershcovich, D., Webersinke, N., Kraus, M., Bingler, J. A., and Leippold, M. (2022). Towards climate awareness in nlp research. DOI: 10.18653/v1/2022.emnlp-main.159.

Hoffmann, J., Borgeaud, S., Mensch, A., *et al*. (2022). Training compute-optimal large language models. In *arXiv.org*. DOI: 10.48550/arXiv.2203.15556.

Holmes, D. T. (2020). Chapter 2 - statistical methods in laboratory medicine. In Clarke, W. and Marzinke, M. A., editors, *Contemporary Practice in Clinical Chemistry (Fourth Edition)*, pages 15–35. Academic Press, fourth edition edition. DOI: 10.1016/B978-0-12-815499-1.00002-8.

Hu, E. J., Shen, Y., Wallis, P., *et al*. (2021). Lora: Low-rank adaptation of large language models. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2106.09685.

InternLM Team (2023). Internlm: A multilingual language model with progressively enhanced capabilities. Available at:`https://github.com/InternLM/InternLM-techreport`.

Jiang, A. Q., Sablayrolles, A., Mensch, A., *et al*. (2023). Mistral 7b. DOI: 10.48550/arXiv.2310.06825.

Klishevich, E., Denisov-Blanch, Y., Obstbaum, S., Ciobanu, I., and Kosinski, M. (2025). Measuring determinism in large language models for software code review.

Krugmann, J. O. and Hartmann, J. (2024). Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11:3. DOI: 10.1007/s40547-024-00143-4.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning. DOI: 10.48550/arxiv.1910.09700.

Larcher, C., Piau, M., Finardi, P., *et al*. (2023). Cabrita: closing the gap for foreign languages. In *arXiv.org*. DOI: 10.48550/arXiv.2308.11878.

Li, X. and Qiu, X. (2023). Finding support examples for in-context learning. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2302.13539.

Liu, J., Shen, D., Zhang, Y., *et al*. (2022). What makes good in-context examples for GPT-3? In Agirre, E., Apidianaki, M., and Vulić, I., editors, *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics. DOI: 10.18653/v1/2022.deelio-1.10.

Liu, P., Yuan, W., Fu, J., *et al*. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. DOI: 10.48550/arxiv.2107.13586.

Liu, Y., Ott, M., Goyal, N., *et al*. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.1907.11692.

Lu, Y., Bartolo, M., Moore, A., *et al*. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. 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 8086–8098, Dublin, Ireland. Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.556.

Maas, A. L., Daly, R. E., Pham, P. T., *et al*. (2011). Learning word vectors for sentiment analysis. Available at:`https://aclanthology.org/P11-1015`.

Meta (2024). Introducing meta llama 3: The most capable openly available llm to date. Available at:`https://ai.meta.com/blog/meta-llama-3/`.

Minaee, S., Mikolov, T., Nikzad, N., *et al*. (2024). Large language models: A survey. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2402.06196.

Mistral AI Team (2023). Mistral 7b in short. Available at:`https://mistral.ai/news/announcing-mistral-7b/`.

Moraes, S., Santos, A., Redecker, M., *et al*. (2016). Comparing approaches to subjectivity classification: A study on portuguese tweets. In Silva, J., Ribeiro, R., Quaresma, P., Adami, A., and Branco, A., editors, *Lecture Notes in Computer Science*, volume 9727, page 86–94. Springer International Publishing. DOI: 10.1007/978-3-319-41552-9_8.

Mosbach, M., Pimentel, T., Ravfogel, S., *et al*. (2023). Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2305.16938.

Naveed, H., Khan, A. U., Qiu, S., *et al*. (2024). A comprehensive overview of large language models. DOI: 10.48550/arXiv.2307.06435.

Oliveira, M. V. and de Melo, T. (2020). Investigating sets of linguistic features for two sentiment analysis tasks in brazilian portuguese web reviews. *Anais Estendidos* *do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*, pages 45–48. DOI: 10.5753/webmedia_estendido.2020.13060.

OpenAI, :, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., and et al (2024a). Gpt-4o system card. DOI: 10.48550/arxiv.2410.21276.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., and et al (2024b). Gpt-4 technical report. DOI: 10.48550/arxiv.2303.08774.

Overwijk, A., Xiong, C., and Callan, J. (2022a). Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3360–3362, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3477495.3536321.

Overwijk, A., Xiong, C., Liu, X., *et al.* (2022b). Clueweb22: 10 billion web documents with visual and semantic information. DOI: 10.48550/arxiv.2211.15848.

Pires, R., Abonizio, H., Almeida, T. S., *et al.* (2023). Sabiá: Portuguese large language models. In Naldi, M. C. and Bianchi, R. A. C., editors, *Lecture Notes in Computer Science*, page 226–240. Springer Nature Switzerland. DOI: 10.1007/978-3-031-45392-2_15.

Přibáň, P., Šmíd, J., Steinberger, J., *et al.* (2024). A comparative study of cross-lingual sentiment analysis. *Expert Systems with Applications*, 247:123247. DOI: 10.1016/j.eswa.2024.123247.

Qiu, X., Sun, T., Xu, Y., *et al.* (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63:1872–1897. DOI: 10.1007/s11431-020-1647-3.

Qwen Team (2024). Introducing qwen1.5. Available at:`https://qwenlm.github.io/blog/qwen1.5/`.

Radford, A., Narasimhan, K., Salimans, T., *et al.* (2018). Improving language understanding by generative pre-training. Available at:`https://www.mikecaptain.com/resources/pdf/GPT-1.pdf`.

Radford, A., Wu, J., Child, R., *et al.* (2019). Language models are unsupervised multitask learners. Available at:`https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf`.

Rae, J. W., Borgeaud, S., Cai, T., *et al.* (2022). Scaling language models: Methods, analysis & insights from training gopher. *arXiv:2112.11446 [cs]*, page 120. DOI: https://doi.org/10.48550/arXiv.2112.11446.

Real, L., Oshiro, M., and Mafra, A. (2019). B2w-reviews01: An open product reviews corpus. In *the Proceedings of the XII Symposium in Information and Human Language Technology.*, pages 200–208. SOCIEDADE BRASILEIRA DE COMPUTAÇÃO (SBC). Available at:`https://comissoes.sbc.org.br/ce-pln/stil2019/proceedings-stil-2019-Final-Publicacao.pdf`.

Reid, M., Savinov, N., Teplyashin, D., *et al.* (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2403.05530.

Reynolds, L. and McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2102.07350.

Rubin, O., Herzig, J., and Berant, J. (2022). Learning to retrieve prompts for in-context learning. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics. DOI: 10.18653/v1/2022.naacl-main.191.

Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., and Agirre, E. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-emnlp.722.

Sales Almeida, T., Abonizio, H., Nogueira, R., *et al.* (2024). Sabiá-2: A new generation of portuguese large language models. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2403.09887.

Scheff, S. W. (2016). Chapter 8 - nonparametric statistics. In Scheff, S. W., editor, *Fundamental Statistical Principles for the Neurobiologist*, pages 157–182. Academic Press. DOI: 10.1016/B978-0-12-804753-8.00008-7.

Silva, R. R. and Pardo, T. A. S. (2019). Corpus 4p: um córpus anotado de opiniões em português sobre produtos eletrônicos para fins de sumarização contrastiva de opinião. In *Proceedings of the 6a Jornada de Descrição do Português (JDP)*, pages 1–9. SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. Available at:`http://drive.google.com/file/d/1Nqu66l-z7eQenXEsvcnAEClt1LQzioJw/view`.

Simmering, P. F. and Huoviala, P. (2023). Large language models for aspect-based sentiment analysis. *arXiv (Cornell University)*, page 12. DOI: 10.48550/arxiv.2310.18025.

Socher, R., Perelygin, A., Wu, J., *et al.* (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S., editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics. DOI: 10.18653/v1/d13-1170.

Song, Y., Wang, G., Li, S., and Lin, B. Y. (2024). The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. DOI: 10.18653/v1/2025.naacl-long.211.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-61377-8_28.

Souza, F. D. and Filho, J. B. d. O. e. S. (2022). Bert for sentiment analysis: Pre-trained and fine-tuned alternatives. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages

209–218, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-98305-5_20.

Strubell, E., Ganesh, A., and McCallum, A. (2020). Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696. DOI: 10.1609/aaai.v34i09.7123.

Touvron, H., Lavril, T., Izacard, G., *et al*. (2023a). Llama: Open and efficient foundation language models. DOI: 10.48550/arXiv.2302.13971.

Touvron, H., Martin, L., Stone, K., *et al*. (2023b). Llama 2: Open foundation and fine-tuned chat models. In *arXiv.org*. DOI: 10.48550/arXiv.2307.09288.

Vargas, F. A., Sanches, R., and Rocha, P. R. (2020). Identifying fine-grained opinion and classifying polarity on coronavirus pandemic. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS 2020)*, page 511–520. Springer-Verlag. DOI: 10.1007/978-3-030-61377-8_35.

Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017). Attention is all you need. In *arXiv.org*. DOI: 10.48550/arXiv.1706.03762.

Wang, A., Singh, A., Michael, J., *et al*. (2019). Glue: A multi-task benchmark and analysis platform for natural language understanding. DOI: 10.18653/v1/w18-5446.

Wang, J. J. and Wang, V. X. (2025). Assessing consistency and reproducibility in the outputs of large language models: Evidence across diverse finance and accounting tasks.

Wang, Y., Kordi, Y., Mishra, S., *et al*. (2022). Self-instruct: Aligning language model with self generated instructions. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2212.10560.

Wang, Z., Xie, Q., Ding, Z., *et al*. (2023). Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv (Cornell Univesity)*. DOI: 10.48550/arxiv.2304.04339.

Wei, J., Tay, Y., Bommasani, R., *et al*. (2022). Emergent abilities of large language models. In *arXiv.org*. DOI: 10.48550/arXiv.2206.07682.

White, J., Fu, Q., Hays, S., *et al*. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *Arxiv (Cornell University)*. DOI: 10.48550/arxiv.2302.11382.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83. DOI: 10.2307/3001968.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., and et al (2024a). Qwen2 technical report. `https://arxiv.org/abs/2407.10671`.

Yang, J., Jin, H., Tang, R., *et al*. (2024b). Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, page 30. DOI: 10.1145/3649506.

Yao, Y., Duan, J., Xu, K., *et al*. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211. DOI: 10.1016/j.hcc.2024.100211.

Ye, J., Wu, Z., Feng, J., *et al*. (2023). Compositional exemplars for in-context learning. arXiv.org. DOI: 10.48550/arXiv.2302.05698.

Yu, B. (2023). Benchmarking large language model volatility. DOI: https://doi.org/10.48550/arxiv.2311.15180.

Zeng, A., Liu, X., Du, Z., *et al*. (2023). GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*. Available at:`https://openreview.net/forum?id=-Aw0rrrPUF`.

Zhang, W., Deng, Y., Liu, B., *et al*. (2023). Sentiment analysis in the era of large language models: A reality check. arXiv.org. DOI: 10.48550/arXiv.2305.15005.

Zhao, J., Liu, K., and Xu, L. (2016). Sentiment analysis: Mining opinions, sentiments, and emotions. *Computational Linguistics*, 42:595–598. DOI: 10.1162/COLI_r_00259.

Zhao, W. X., Zhou, K., Li, J., *et al*. (2023). A survey of large language models. `https://doi.org/10.48550/arXiv.2303.18223`.

Zhong, Q., Ding, L., Liu, J., *et al*. (2023). Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2302.10198.

Zhou, Y., Muresanu, A. I., Han, Z., *et al*. (2022). Large language models are human-level prompt engineers. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2211.01910.

A Consolidated Experimental Results

This appendix section presents a view of the experimental results obtained in the comparative analysis of generalist and Portuguese fine-tuned Language Models. The detailed tables showcase various aspects of model performance and behavior during the sentiment analysis task.

Table 7. Descriptive statistics of output token generation across models, quantified using the OpenAI Tiktoken o200k_base tokenizer.

Scale	Linguistic Fine-tuning	Model	Mean	Standard Deviation	Median	Min	Max
Baselines		Strong Classifier (DeepSeek-R1)	13.12	4.84	14	4	231
Large-scale (>70B)	Generalist	Claude-3.5-Sonnet	11.01	0.18	11	11	17
		DeepSeek-V3	11.33	1.09	11	10	21
		GPT-4o	10.00	0.02	10	10	11
		Gemini-1.5-Pro	6.95	0.56	7	0	7
	PT-BR	Sabiá-3	11.01	0.22	11	11	19
		Sabiá-2-Medium	7.34	1.30	7	6	16
Small-Scale (<13B)	Generalist	Gemma-2-9B-Instruct	14.18	2.29	15	9	22
		Qwen-2-7B-Instruct	7.15	0.86	7	7	16
		LLaMA-3-8B-Instruct	54.15	45.40	96	7	140
		InternLM-2-7B-Chat	16.87	0.81	17	15	18
		DeepSeek-R1-Distill-LLaMA-8B	17.15	0.78	17	16	19
		DeepSeek-R1-Distill-Qwen-7B	18.47	1.08	19	1	21
		Gemma-7B-Instruct	13.94	0.63	14	6	16
		LLaMA-3.1-8B-Instruct	102.78	20.27	96	7	171
	PT-BR	Bode-3.1-8B-Instruct-lora	16.01	4.63	18	0	21
		InternLM-ChatBode-7B	15.61	1.39	15	13	18
		CabraLLaMA-3-8B	15.16	5.32	19	0	20
		CabraMistral-v3-7B-32k	14.92	0.44	15	12	18
		GemBode-7B-Instruct	18.56	2.38	19	0	25
		Bode-7B	10.27	2.96	13	3	14
		Bode-13B	7.59	1.78	7	7	15
		Sabiá-7B	8.04	0.54	8	3	11

Table 7 presents descriptive statistics of output token generation across all evaluated models. For each LLM, the table quantifies mean, standard deviation, median, minimum, and maximum of output tokens produced during sentiment analysis. All token counts were calculated using the OpenAI Tiktoken o200k_base tokenizer for standardization purposes. It is worth noting that the reported counts may be slightly

Table 8. Accuracy obtained per model and dataset, stratified by scale and linguistic specialization.

Scale	Linguistic Fine-Tuning	Model	Datasets											
			IMDB_PT	SST2_PT	TweetSentBR	ReLI	Computer-BR	MTMSLA	CSP-Eletrônicos	CSP-Livros	4P Corpus	RePro	OPCovidBR	TA-Restaurantes
Baselines		Weak Classifier (Train set majority class)	0.5000	0.5092	0.6071	0.8262	0.6953	0.5784	0.6842	0.4857	0.8201	0.5449	0.5041	0.9027
		Strong Classifier (DeepSeek R1)	0.9544	0.9323	0.9177	0.9378	0.9609	0.9902	0.9211	0.9714	0.9784	0.9921	0.8130	0.9115
Large Sacle (70B)	Generalist	Claude-3.5-Sonnet	0.9548	0.9507	0.9217	0.9713	0.9609	0.9608	0.9211	1.0000	0.9892	0.9960	0.8211	0.9292
		DeepSeek-V3	0.9508	0.9117	0.9183	0.9346	0.9453	0.9804	0.9474	0.9714	0.9856	0.9908	0.7642	0.9292
		GPT-4o	0.9484	0.9312	0.9250	0.9442	0.9609	0.9902	0.8421	1.0000	0.9856	0.9914	0.7642	0.9381
		Gemini-1.5-Pro	0.9344	0.9037	0.9284	0.9298	0.9531	0.9804	0.8947	0.9714	0.9856	0.9941	0.7073	0.9115
	PT-BR	Sabiá-3	0.9518	0.9300	0.9210	0.9474	0.9453	0.9804	0.9737	1.0000	0.9856	0.9947	0.7805	0.9381
		Sabiá-2-Medium	0.9478	0.9128	0.6847	0.9298	0.9063	0.9608	0.8947	1.0000	0.9640	0.9855	0.7967	0.9204
Small Scale (13B)	Generalist	Gemma-2-9B-Instruct	0.9404	0.9278	0.9016	0.9378	0.9063	0.9412	0.9474	1.0000	0.9784	0.9901	0.8049	0.9292
		Qwen-2-7B-Instruct	0.9422	0.8807	0.8768	0.9378	0.8828	0.9510	0.9474	1.0000	0.9748	0.9842	0.7805	0.9204
		LLaMA-3-8B-Instruct	0.9376	0.8796	0.8742	0.9346	0.8750	0.9412	0.9474	0.9714	0.9712	0.9763	0.7886	0.9292
		InternLM-2-7B-Chat	0.9334	0.8670	0.8226	0.9362	0.8672	0.8824	0.9737	0.8857	0.9568	0.9782	0.7561	0.9292
		DeepSeek-R1-Distill-LLaMA-8B	0.9096	0.8658	0.8380	0.9187	0.8672	0.9314	0.9211	0.9714	0.8813	0.9802	0.7398	0.9027
		DeepSeek-R1-Distill-Qwen-7B	0.8780	0.8108	0.7965	0.8868	0.8281	0.9118	0.7895	0.9143	0.9281	0.9657	0.7236	0.9027
		Gemma-7B-Instruct	0.7452	0.8498	0.7557	0.8963	0.7422	0.7255	0.8421	0.8857	0.9137	0.9505	0.7480	0.8761
		LLaMA-3.1-8B-Instruct	0.9018	0.7041	0.8159	0.8931	0.8516	0.7255	0.5000	0.8571	0.6043	0.8536	0.7154	0.6814
	PT-BR	Bode-3.1-8B-Instruct-lora	0.9138	0.8865	0.8327	0.9378	0.8906	0.9510	0.8947	0.9714	0.9424	0.9855	0.7642	0.8938
		InternLM-ChatBode-7B	0.9396	0.8429	0.8112	0.9458	0.8594	0.8824	0.9737	0.9429	0.9532	0.9710	0.7967	0.8938
		CabraLLaMA-3-8B	0.9214	0.8681	0.7731	0.9330	0.8281	0.8529	0.9211	0.9429	0.9496	0.9723	0.7561	0.9292
		CabraMistral-v3-7B-32k	0.8896	0.8670	0.8246	0.9266	0.8359	0.8824	1.0000	0.9143	0.9460	0.9769	0.5935	0.9204
		GemBode-7B-Instruct	0.9228	0.8280	0.7764	0.9171	0.7969	0.7451	0.9737	0.9714	0.9640	0.9551	0.6504	0.9027
		Bode-7B	0.9208	0.8268	0.7222	0.9123	0.7266	0.8529	0.9211	0.9714	0.9604	0.9420	0.6341	0.9204
		Bode-13B	0.9092	0.8073	0.7577	0.8708	0.8047	0.8333	0.8947	0.9143	0.9137	0.9156	0.6098	0.9027
		Sabiá-7B	0.5970	0.5493	0.6198	0.8628	0.3906	0.6275	0.7895	0.6000	0.8345	0.6788	0.5041	0.9027

higher than the expected limits described in the methodology section (see Table 3) due to the use of a different tokenizer than those employed by the models during inference. This table provides insights into the verbosity characteristics and response consistency of each model when performing sentiment analysis tasks in Brazilian Portuguese.

Table 8 presents the raw experimental results for each model across all evaluated datasets, organized by Accuracy. Models are arranged according to their scale (LLMs with more than 70 billion parameters and LLMs with less than 13 billion parameters) and linguistic specialization (generalist versus Portuguese fine-tuned), then listed in descending order based on their mean Accuracy performance. The table includes two baseline references: a weak classifier representing the majority class in each training set, and a strong classifier implemented with DeepSeek-R1. This organization enables detailed analysis of how each model performs across different domains represented by the twelve Brazilian Portuguese sentiment analysis datasets.

Table 9 complements the accuracy analysis by presenting the Macro F_1 Score for each model-dataset combination. This metric is particularly valuable as it provides a more balanced assessment when dealing with class imbalance, which is common in several of the evaluated datasets. Unlike accuracy, which can be artificially inflated in imbalanced scenarios, the Macro F_1 Score gives equal weight to each class by calculating the harmonic mean of precision and recall independently for each class before averaging.

This approach reveals important nuances that might be obscured when relying solely on accuracy metrics. For instance, models with comparable accuracy values may exhibit substantial differences in their F_1 Scores, indicating variations in their ability to correctly identify both positive and negative sentiments with equal proficiency.

Understanding the relationship between model hallucinations and performance metrics is crucial for an extensive evaluation of LLMs. As discussed in Subsection 4.5, hallucinations significantly impact the calculation of Macro F_1 Score, as these instances receive a local score of zero for the hallucination class, which reduces the overall metric value despite not affecting the accuracy in the same way. This relationship explains some of the discrepancies observed between the Accuracy and F_1 Score results in the previous tables.

Table 11 provides a consolidated view of hallucination statistics across all evaluated language models, maintaining the same stratification by scale and linguistic specialization. The table quantifies the absolute count of hallucinations, the number of distinct datasets where hallucinations occurred, the percentage of total hallucinations attributed to each model, and the mean hallucination count per dataset.

Table 10 displays the raw results for hallucination occurrences across all experiments. This detailed breakdown allows for the identification of specific model-dataset combinations that are particularly prone to hallucinations, revealing patterns that may not be apparent in the consolidated statistics. For instance, some models demonstrate consistent hallucination behavior across multiple datasets, while others show pronounced vulnerability only with specific data types or domains. This granular view provides researchers and practitioners with insights into the reliability constraints of different LLMs when processing Brazilian Portuguese content for sentiment analysis tasks.

B Hypotesis Testing

This section presents the details of the hypothesis tests conducted to evaluate the statistical significance of performance differences between the language models. The Wilcoxon signed-rank test [Wilcoxon, 1945] was chosen due to several advantageous characteristics compatible with the experiments: it is robust for small sample sizes, makes no assumptions about the data distribution, and is a non-parametric alternative to the paired t-test [Scheff, 2016; Holmes, 2020].

The paired nature of this test is well-suited for the experimental design, where 23 different language models were compared against each other across the same set of 12 datasets. This approach is methodologically appropriate since all models processed identical test instances with the same prompts, creating naturally matched pairs of observations. The paired

Table 9. Macro F_1 Score obtained per model and dataset, stratified by scale and linguistic specialization.

Scale	Linguistic Fine-Tuning	Model	Datasets											
			IMDB_PT	SST2_PT	TweetSentBR	ReLI	Computer-BR	MTMSLA	CSP-Eletrônicos	CSP-Livros	4P Corpus	RePro	OPCovidBR	TA-Restaurantes
Baselines		Weak Classifier (Train set majority class)	0.3333	0.3374	0.3778	0.4524	0.4101	0.3665	0.4063	0.3269	0.4506	0.3527	0.3351	0.4744
		Strong Classifier (DeepSeek R1)	0.6365	0.6222	0.6096	0.9023	0.9529	0.9900	0.9138	0.9714	0.6461	0.9920	0.8126	0.7669
Large Sacle (>70B)	Generalist	Claude-3.5-Sonnet	0.9548	0.9507	0.9156	0.9518	0.9529	0.9595	0.9138	1.0000	0.9821	0.9960	0.8210	0.8135
		DeepSeek-V3	0.9508	0.6091	0.6114	0.8960	0.6303	0.6609	0.9415	0.9714	0.6540	0.6613	0.7632	0.8135
		GPT-4o	0.9484	0.9312	0.9208	0.9118	0.9529	0.9900	0.8348	1.0000	0.9763	0.9914	0.7611	0.8426
		Gemini-1.5-Pro	0.6257	0.6066	0.9250	0.8901	0.9447	0.9799	0.8869	0.9714	0.9763	0.9940	0.4903	0.7827
	PT-BR	Sabiá-3	0.6349	0.9300	0.6110	0.6110	0.9350	0.9799	0.9702	1.0000	0.9756	0.6636	0.7790	0.8306
		Sabiá-2-Medium	0.6336	0.6099	0.4480	0.5958	0.6020	0.9595	0.8869	1.0000	0.6377	0.6596	0.7963	0.7976
Small Scale (<13B)	Generalist	Gemma-2-9B-Instruct	0.6270	0.6196	0.8937	0.8965	0.5957	0.9388	0.9391	1.0000	0.6458	0.6616	0.8048	0.7986
		Qwen-2-7B-Instruct	0.6293	0.5877	0.8653	0.5947	0.5793	0.9496	0.9415	1.0000	0.6423	0.6578	0.5252	0.7414
		LLaMA-3-8B-Instruct	0.6255	0.5874	0.5763	0.5914	0.8545	0.6285	0.9415	0.9714	0.6373	0.6528	0.5301	0.7806
		InternLM-2-7B-Chat	0.6225	0.8656	0.8013	0.8856	0.8520	0.8755	0.9688	0.8833	0.9268	0.9780	0.7488	0.7806
		DeepSeek-R1-Distill-LLaMA-8B	0.9093	0.8658	0.8339	0.8569	0.8465	0.9289	0.9138	0.9714	0.5564	0.9801	0.7384	0.7113
		DeepSeek-R1-Distill-Qwen-7B	0.5862	0.8103	0.7834	0.5368	0.8026	0.9103	0.7841	0.9143	0.5964	0.6460	0.7231	0.7113
		Gemma-7B-Instruct	0.5405	0.5727	0.4828	0.5454	0.4920	0.4661	0.5832	0.6074	0.5747	0.6428	0.5127	0.4314
		LLaMA-3.1-8B-Instruct	0.6138	0.5330	0.5569	0.5741	0.5740	0.5371	0.4561	0.6148	0.5114	0.6116	0.5097	0.3611
	PT-BR	Bode-3.1-8B-Instruct-lora	0.6091	0.5921	0.5433	0.5978	0.8744	0.9499	0.6000	0.9714	0.6245	0.6574	0.7620	0.5040
		InternLM-ChatBode-7B	0.9395	0.8391	0.7852	0.8997	0.8458	0.8755	0.9688	0.9428	0.9188	0.6473	0.7923	0.6709
		CabraLLaMA-3-8B	0.6182	0.5795	0.7203	0.5794	0.8151	0.8400	0.9106	0.9424	0.9104	0.6481	0.7501	0.7806
		CabraMistral-v3-7B-32k	0.5924	0.8663	0.8078	0.5809	0.8151	0.8728	1.0000	0.9140	0.9093	0.9767	0.5091	0.7414
		GemBode-7B-Instruct	0.6164	0.5505	0.4886	0.5492	0.7867	0.4639	0.9702	0.6566	0.9370	0.6380	0.6135	0.6073
		Bode-7B	0.6142	0.5499	0.6551	0.5615	0.7190	0.8451	0.9013	0.9713	0.9302	0.6284	0.5999	0.7638
		Bode-13B	0.6069	0.5583	0.4826	0.5491	0.5289	0.5531	0.6107	0.6344	0.5945	0.6232	0.3694	0.4438
		Sabiá-7B	0.4716	0.4222	0.2760	0.6359	0.3600	0.4824	0.6833	0.5100	0.5283	0.6135	0.3351	0.4744

Table 10. Hallucination statistics across language models categorized by scale and linguistic specialization.

Scale	Linguistic Fine-Tuning	Model	Datasets											
			IMDB_PT	SST2_PT	TweetSentBR	ReLI	Computer-BR	MTMSLA	CSP-Eletrônicos	CSP-Livros	4P Corpus	RePro	OPCovidBR	TA-Restaurantes
Baselines		Weak Classifier (Train set majority class)												
		Strong Classifier (DeepSeek R1)	4	2	1						1			
Large Sacle (>70B)	Generalist	Claude-3.5-Sonnet												
		DeepSeek-V3		4	12		3	2			1	4		
		GPT-4o												
		Gemini-1.5-Pro	44	12									11	
	PT-BR	Sabiá-3	6		2	1						2		
		Sabiá-2-Medium	27	4	18	1	2				4	12		
Small Scale (<13B)	Generalist	Gemma-2-9B-Instruct	1	3			1				1	7		
		Qwen-2-7B-Instruct	18	4		2	1				2	9	3	
		LLaMA-3-8B-Instruct	7	4	5	1		1			1	10	2	
		InternLM-2-7B-Chat	6											
		DeepSeek-R1-Distill-LLaMA-8B									1			
		DeepSeek-R1-Distill-Qwen-7B	17			1					1	12		
		Gemma-7B-Instruct	842	19	19	18	7	5	3	2	8	44	7	5
		LLaMA-3.1-8B-Instruct	207	211	105	31	6	22	7	5	35	207	16	5
	PT-BR	Bode-3.1-8B-Instruct-lora	2	5	6	4			1		7	2		3
		InternLM-ChatBode-7B										1		
		CabraLLaMA-3-8B	63	5		1						1		
		CabraMistral-v3-7B-32k	1			1								
		GemBode-7B-Instruct	20	4	1	2		1		1		8		
		Bode-7B	5	5		5						6		
		Bode-13B	15	68	30	37	5	2	2	3	16	71	2	3
		Sabiá-7B	1.469		3									

Table 11. Hallucination counts per model and dataset, stratified by scale and linguistic specialization.

Scale	Linguistic Fine-Tuning	Model	Count	Distinct Datasets	% of Total	Mean
Baselines		Weak Classifier (Train set majority class)	0	0	0.00%	0
		Strong Classifier (DeepSeek R1)	8	4	0.08%	1
Large Sacle (>70B)	Generalist	Claude-3.5-Sonnet	0	0	0.00%	0
		DeepSeek-V3	26	6	0.25%	2
		GPT-4o	0	0	0.00%	0
		Gemini-1.5-Pro	67	3	0.65%	6
	PT-BR	Sabiá-3	11	4	0.11%	1
		Sabiá-2-Medium	68	7	0.66%	6
Small Scale (<13B)	Generalist	Gemma-2-9B-Instruct	13	5	0.13%	1
		Qwen-2-7B-Instruct	39	7	0.38%	3
		LLaMA-3-8B-Instruct	31	8	0.30%	3
		InternLM-2-7B-Chat	6	1	0.06%	1
		DeepSeek-R1-Distill-LLaMA-8B	1	1	0.01%	0
		DeepSeek-R1-Distill-Qwen-7B	31	4	0.30%	3
		Gemma-7B-Instruct	979	12	9.48%	82
		LLaMA-3.1-8B-Instruct	857	12	8.30%	71
	PT-BR	Bode-3.1-8B-Instruct-lora	30	8	0.29%	3
		InternLM-ChatBode-7B	1	1	0.01%	0
		CabraLLaMA-3-8B	70	4	0.68%	6
		CabraMistral-v3-7B-32k	2	2	0.02%	0
		GemBode-7B-Instruct	37	7	0.36%	3
		Bode-7B	21	4	0.20%	2
		Bode-13B	254	12	2.46%	21
		Sabiá-7B	1.472	2	14.25%	123

design accounts for inherent differences in difficulty levels, class distributions, and linguistic characteristics across datasets, enabling a more direct comparison of model capabilities by focusing on relative differences rather than absolute performance values.

Results of Wilcoxon tests for paired groups with 5% significance level for Accuracy metric are consolidated in Figure 7. The test evaluates the H_0 hypothesis that two related paired samples come from the same distribution, in other words, tests if the difference between paired observations in the population is zero. The Green circle symbol indicates sufficient evidence to reject H_0 in favor of H_1 : Model 1 $>$ Model 2. The Red circle symbol indicates sufficient evidence to reject H_0 in favor of H_1 : Model 1 $<$ Model 2 at the established significance level. Yellow circle indicates no sufficient evidence to reject H_0. White circle indicates that the evaluated models are identical, therefore the test was not applied.

Similarly, Figure 8 presents the results of the Wilcoxon signed-rank tests for the Macro F_1 Score metric, using the same significance level and visual encoding scheme than

Figure 7. Results of Wilcoxon tests for paired groups with 5% significance level for Accuracy metric.

Figure 7 .

C Qualitative Analysis

This section presents a comprehensive qualitative analysis of response patterns generated by all 23 evaluated models across the sentiment classification task. The analysis reveals distinct behavioral patterns strongly correlated with model scale. Large-scale models ($>$ 70B) demonstrated superior instruction adherence, producing highly concentrated response distributions with minimal variance from the requested JSON format.

Conversely, smaller-scale models (<13B) exhibited greater response fragmentation and systematic generation of structural artifacts, predominantly derived from prompt elements such as demonstration examples and task descriptions. Despite these formatting inconsistencies, the majority of models maintained high classification validity rates ($>$ 99.5%), indicating successful task execution even when accompanied by extraneous content.

The following subsections provide detailed model-by-model analysis, categorized by scale and linguistic specialization, examining response consistency, artifact patterns, and adherence to the specified JSON schema.

C.1 Large-scale models (>70B)

C.1.1 Generalist

DeepSeek-R1 The model demonstrated high consistency with 97.19% of responses concentrated in five variations (Table 12) of the requested JSON format, differing only in quotation marks (single/double) and spacing. The majority (94.57%) included markdown markers (“```*json*”). The remaining 2.81% comprised 8 verbose responses with explanations, 4 malformed JSONs, and 4 with line breaks. The highly predictable behavior indicates robustness for automated tasks, with inconsistencies representing rare events. The response validity rate was 99.92%.

Table 12. DeepSeek-R1 Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
46	1. ```json\n{\n ”polaridade”: 1\n}\n```	38.61 %
	2. ```json\n{\n ”polaridade”: -1\n}\n```	33.79 %
	3. ```json\n{”polaridade”: 1}\n```	12.46 %
	4. ```json\n{”polaridade”: -1}\n```	9.23 %
	5. {’polaridade’: 1}	3.07 %

Claude-3.5-Sonnet The model demonstrated exceptional performance with 99.86% of responses in the exact expected JSON format (“*{”polaridade”: 1}*” or “*{”polaridade”:*

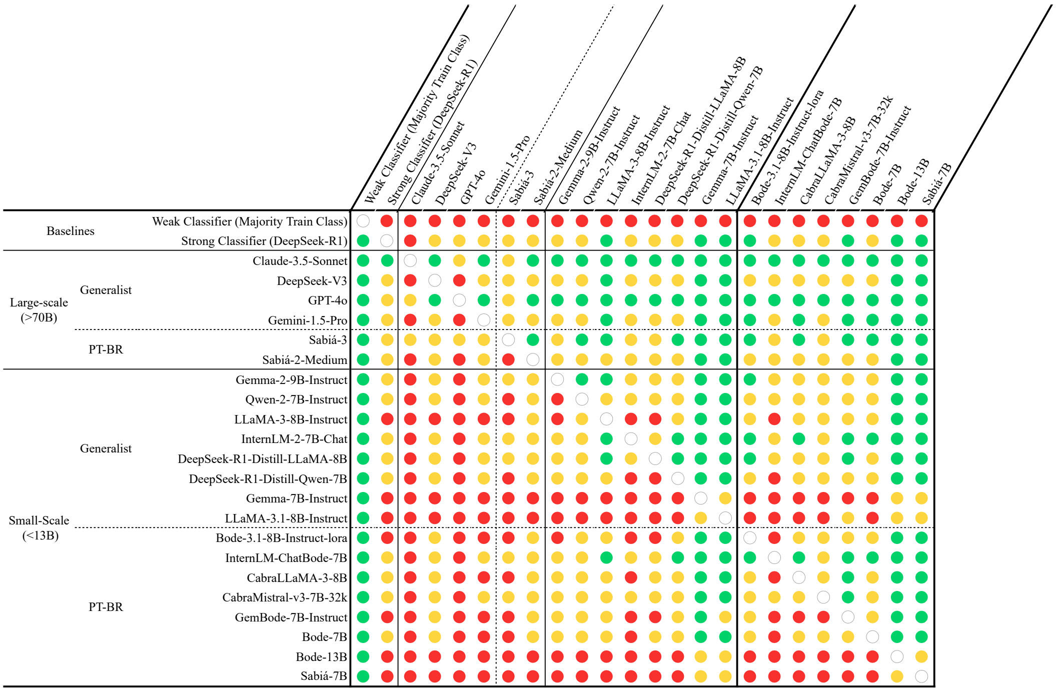


Figure 8. Results of Wilcoxon tests for paired groups with 5% significance level for Macro F_1 Score metric.

$-1\}$*"*). Only 0.14% of responses included additional justifications after the JSON, using markers such as "*Justificativa:*", "*Explicação:*" or "*Explanation:*". All 10,326 inferences maintained 100% compliance with the requested JSON schema, resulting in a validity rate of 100%. The model presented only 13 unique responses (Table 13), indicating high consistency and minimal variability in outputs.

Table 13. Claude-3.5-Sonnet Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
13	1. \n{\n "polaridade": 1\n}	56.01%
	2. \n{\n "polaridade": -1\n}	43.85 %
	3. \n{\n "polaridade": -1\n}\n\nEmbora o texto não exp	0.01 %
	4. \n{\n "polaridade": -1\n}\n\nJustificativa:	0.01 %
	5. \n{\n "polaridade": -1\n}\n\nJustificativa:	0.01 %

DeepSeek-V3 The model produced 61 distinct responses (Table 14), with 98.74% concentrated in 6 valid variations that alternated between single/double quotes and inline/multiline formatting. The remaining 1.26% (55 variations) included artifacts such as "*Agora, realize a classificação*" (0.51%), unsolicited explanations like "*Para classificar o sentimento*" or "*Para realizar a classificação*" (0.24%) and the fragment "*Classificação de Sentimento:*" (0.14%). The model presented a high response validity rate (99.74%)

Table 14. DeepSeek-V3 Generation Overview

Distinct Raw Responses	Top 5 Occurrences	%
61	1. ```json\n{'polaridade': 1}\n```	36.82 %
	2. ```json\n{'polaridade': -1}\n```	34.13 %
	3. ```json\n{"polaridade": 1}\n```	12.89 %
	4. ```json\n{"polaridade": -1}\n```	6.44 %
	5. ```json\n{\n "polaridade": 1\n}\n```	4.56 %

GPT-4o The model presented high consistency with only 4 distinct responses (Table 15), with a validity rate of 100%. All responses perfectly followed the requested JSON structure, containing exclusively the "*polaridade*" field with correct values (-1 or 1), without extra fields or verbosity. Variations were limited to minimal formatting differences: multiline indentation in the main responses and additional line breaks in minority variations.

Table 15. GPT-4o Generation Overview.

Distinct Raw Responses	Top 4 Occurrences	%
4	1. {\n "polaridade": 1\n}	53.62%
	2. {\n "polaridade": -1\n}	46.34 %
	3. \n{\n "polaridade": 1\n}	0.01 %
	4. \n{\n "polaridade": -1\n}	0.00 %

Gemini-1.5-Pro The model produced only 6 distinct responses (Table 16), demonstrating high consistency in model behavior and a validity rate of 99.35%. The 5 most frequent responses represent 99.99% of outputs, all maintaining perfect adherence to the requested JSON structure with the "*polaridade*" field and expected values (-1 or 1). The observed variations were limited to minimal formatting aspects: presence of one or two line breaks after JSON closure and a minority case without space after the colon. Notably, 0.65% of responses were blocked by the model's security filters, returning null values

Table 16. Gemini-1.5-Pro Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
6	1. {"polaridade": 1}\n	44.81 %
	2. {"polaridade": -1}\n	34.99 %
	3. {"polaridade": -1}\n\n	10.50 %
	4. {"polaridade": 1}\n\n	8.96 %
	5. NaN	0.65 %

C.1.2 PT-BR

Sabiá-3 The model generated 12 distinct responses (Table 17), with 99.89% concentrated in 4 main variations. These responses adhered to the requested JSON format, with consistent use of markdown code blocks and inclusion of the "*polaridade*" field. The remaining 0.11% presented anomalies: in 0.09% (9 occurrences), the returned value was "*{'polaridade': 0}*", outside the $[-1, 1]$ range, generally accompanied by explanatory notes (e.g., "*Note que a saída padrão...*"); in 0.02% (2 occurrences), there were error messages related to the input text (e.g., "*Parece que houve um erro na sua solicitação*"). The overall validity rate was 99.89%.

Table 17. Sabiá-3 Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
12	1. ```json\n{'polaridade': 1}\n```	51.81 %
	2. ```json\n{'polaridade': -1}\n```	41.77 %
	3. ```json\n{"polaridade": 1}\n```	3.83 %
	4. ```json\n{"polaridade": -1}\n```	2.46 %
	5. ```json\n{'polaridade': 0}\n```\n\n (Note que a saída padrão	0.02 %

Sabiá-2-Medium The 5 most frequent responses (Table 18) account for 91.29% of outputs and present good adherence to the instruction and specified format. There is a sharp drop from the 6th (6.22%) to the 7th position (0.53%). Between positions 6 to23 (8.36%), returns follow the JSON format but with greater variation in formatting and some cases of class 0. The remaining 0.34% correspond to error messages, generally attributed to problems in the input text. The validity rate of produced responses was 99.34%.

Table 18. Sabiá-2-Medium Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
54	1. {'polaridade': -1}	39.24 %
	2. {'polaridade': -1}	26.20 %
	3. 'polaridade': -1	10.71 %
	4. 'saida': {'polaridade': 1}	7.85 %
	5. 'saida': {'polaridade': -1}	7.28 %

C.2 Small-scale models (<13B)

C.2.1 Generalist

Gemma-2-9B-Instruct The model produced 31 distinct responses (Table 19), with 95.7% concentrated in the five main variations, all adhering to the requested JSON structure. The four most frequent differ only by formatting artifacts (e.g., , "```*json*"), without affecting content. The valid response rate was 99.87%, with prompt artifacts in 0.20% of cases and consistency in polarity (values -1 or 1). The remaining 4.25%, distributed across 26 smaller variations, exhibit small inconsistencies such as decimal values (1.0, -1.0), occasional JSON duplications, and rare cases of invalid polarity (0).

Table 19. Gemma-2-9B-Instruct Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
31	1. {'polaridade': 1}\n\n\n\n <end_of_turn><eos >	46.60 %
	2. {'polaridade': -1}\n\n\n\n <end_of_turn><eos >	35.71 %
	3. {'polaridade': 1}\n\n\n\n```json	8.03 %
	4. {'polaridade': -1}\n\n\n\n```json	4.60 %
	5. {'polaridade': 1}\n\n\n\n <end_of_turn>\n	0.79 %

Qwen-2-7B-Instruct The model generated 16 distinct responses (Table 20), with 96.1% concentrated in two main variations ("*{'polaridade': 1}*" and *"{'polaridade': -1}"*), faithfully adhering to the JSON format with single quotes and no artifacts. The remaining 3.9% were divided into 14 smaller variations: 2.6% with alternative formatting ("```*json*", double quotes), 1.2% with decimal values (1.0, -1.0), 0.35% with invalid polarity (0), and 0.02% with unsolicited explanatory responses. The valid response rate was 99.62%, evidencing excellent adherence to instructions and low incidence of artifacts.

Table 20. Qwen-2-7B-Instruct Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
16	1. {'polaridade': 1}	56.19 %
	2. {'polaridade': -1}	39.94 %
	3. ```json\n{"polaridade": -1}\n```	1.10 %
	4. {'polaridade': 1.0}	0.54 %
	5. ```json\n{"polaridade": 1}\n```	0.51 %

LLaMA-3-8B-Instruct The model presented bimodal behavior, with 83.28% of inferences concentrated in five main variations (Table 21). Two categories stood out: clean responses with only the polarity JSON (46.5%) and responses that partially or totally reproduce the original prompt schema (51.4%), including fields such as "*'type': 'object'*" and "*'description': 'Objeto de saída...'*". The top ten responses totaled 93.28%, while the remaining 396 formed a long tail (6.72%). Despite the high validity rate (99.7%), the model demonstrated a tendency toward literal prompt reproduction.

Table 21. LLaMA-3-8B-Instruct Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
406	1. {'polaridade': 1}	25.10 %
	2. {'type': 'object','description': Objeto de saída fornecido pelo classificador após a classificação de sentimento do texto de entrada.', 'properties': {'polaridade': {'type': 'integer','description': 'Polaridade em relação ao sentimento expressado no textode entrada. Pode assumir 2 valores: [-1, 1]','enum': [-1,1]}}, \n 'required': ['polaridade']}\n \n'saida': {'polaridade': 1}	20.92 %
	3. {'type': 'object','description': Objeto de saída fornecido pelo classificadorapós a classificação de sentimento do texto de entrada.', 'properties': {'polaridade': {'type': 'integer','description': 'Polaridade em relação ao sentimento expressado no textode entrada. Pode assumir 2 valores: [-1, 1]','enum': [-1,1]}}, \n 'required': ['polaridade']}\n \n'saida': {'polaridade': -1}	17.17 %
	4. {'polaridade': -1}	15.68 %
	5. {'polaridade': 1}	4.39 %

InternLM-2-7B-Chat The model presented highly fragmented behavior, with $1,359$ distinct responses and only 78.17% concentrated in the 20 most frequent variations. It demonstrated a strong tendency to reproduce demonstration artifacts, with 95.3% containing "*Exemplo:*" and 99.4% starting with "*'entrada':*", in addition to generating spurious fragments such as "*O que é*", "*O que eu g*" and "*Agora,*". In 4.7%, it also reproduced "*Classificação de Sentimento:*" from the main prompt. Despite contamination by artifacts and irrelevant text, the validity rate was high (99.94%). No response, however, presented the requested clean JSON, indicating failure to separate the demonstration structure from task execution. The overview of model response generation is presented in Table 22.

Table 22. InternLM-2-7B-Chat Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
1,359	1. {'polaridade': 1}\nExemplo:\n'entrada': 'O que é	14.72 %
	2. {'polaridade': 1}\nExemplo:\n'entrada': 'O que eu g	9.14 %
	3. {'polaridade': -1}\nExemplo:\n'entrada': 'Agora,	8.88 %
	4. {'polaridade': 1}\nExemplo:\n'entrada': 'Agora,	7.23 %
	5. {'polaridade': -1}\nExemplo:\n'entrada': 'O que eu g	6.43 %

DeepSeek-R1-Distill-LLaMA-8B The model presented verbose behavior, with 427 distinct responses and 74.1% concentrated in the ten main variations. In 72.58% of cases, it generated unsolicited explanations in Portuguese, initiated by "*Ok, eu preciso...*", indicating self-narration not induced by the prompts. Only 8.0% of responses included prompt artifacts ("*Classificação de Sentimento:*"), suggesting low literal reproduction. No response brought only the requested JSON. Despite this, the validity rate was excellent (99.99%), evidencing correct task execution combined with a systematic pattern of autonomous verbalization, possibly inherited from training. The overview of model response generation is presented in Table 23.

Table 23. DeepSeek-R1-Distill-LLaMA-8B Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
427	1. {'polaridade': 1}\nOk, eu preciso classificar o sentimento de um	15.06 %
	2. {'polaridade': 1}\nOk, eu preciso classificar a polaridade desse	12.35 %
	3. {'polaridade': 1}\nOk, eu preciso classificar a polaridade de um	8.89 %
	4. {'polaridade': -1}\nOk, eu preciso classificar a polaridade desse	8.82 %
	5. {'polaridade': -1}\nOk, vou analisar o texto de entrada para determin	6.45 %

DeepSeek-R1-Distill-Qwen-7B The model presented highly heterogeneous behavior (Table 24), with 733 distinct responses and 74.6% concentrated in 20 main variations. The leakage of internal reasoning tokens stood out, with 43.7% containing thinking (not present in prompts), followed by markdown JSON blocks (42.3%). Also frequent were the reproduction of prompt artifacts (41.0%) and the generation of spurious fragments such as "*O que você*", "*ninguem je*" and "*aqui está*" (24.2%). In 10.1%, unsolicited self-instructions emerged ("*Agora, considere...*"). The model combined correct task execution with reasoning leakage, literal reproduction, and autonomous generation, maintaining a high validity rate (99.66%).

Table 24. DeepSeek-R1-Distill-Qwen-7B Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
733	1. {'polaridade': -1}\n </think>\n\n ```json\n{'polaridade': -1	18.09 %
	2. {'polaridade': 1}\n</think>\n\n ```json\n{'polaridade': 1	10.78 %
	3. {'polaridade': 1}\nClassificação de Sentimento: ' entrada': 'O que você	7.09 %
	4. {'polaridade': -1}\nClassificação de Sentimento:' entrada': 'O que você	6.33 %
	5. {'polaridade': 1}\n</think>\n\n ```json\n{\n "polaridade":	6.13 %

Gemma-7B-Instruct The model presented dysfunctional behavior, with 401 unique responses (Table 25) and only 70.3% concentrated in the top 20. There was excessively verbose and out-of-scope generation, with 77.9% using markdown formatting ("```" or "**") and 47.8% containing elaborate and irrelevant explanations. Invented fragments were identified such as "*O objetivo deste trabalho é classificar*" (18.5%) and autonomous instructions initiated by "*Lembre-se*" (23.4%), absent from the prompts. The model generated 9.5% completely invalid responses and 23.3% with spurious text, including mentions of "*Python*" and random excerpts. No response followed the expected clean format (JSON only), resulting in a low validity rate (90.5%) and highlighting serious instruction-following failures.

Table 25. Gemma-7B-Instruct Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
401	1. {'polaridade': 1}\n\n **O objetivo deste trabalho é classific	8.46 %
	2. {'polaridade': -1}\n\n ** O objetivo deste trabalho é classific	7.94 %
	3. {'polaridade': 1}\n ```\n\n**Lembre-se	6.82 %
	4. {'polaridade': 1}\n\n **O objetivo da tarefa é classific	6.61 %
	5. {'polaridade': -1}\n ```\n\n**Requisitos:**\n\n*	5.49 %

LLaMA-3.1-8B-Instruct The model presented highly dysfunctional behavior, with 2,788 unique responses (Table 26) and a strong tendency toward literal reproduction of the JSON schema from the prompt: 85.4% included the complete fragment of the original structure ("*{'type': 'object','description': Objeto de saída fornecido pelo classificador após a classificação de sentimento do texto de entrada.', 'properties': {'polaridade': {'type': 'integer','description': 'Polaridade em relação ao sentimento expressado no texto de entrada. Pode assumir 2 valores: [-1, 1]','enum': [-1,1]}},\n 'required': ['polaridade']}\n\n*"). The two main responses, with 66.9% of inferences, consist almost exclusively of this repetition, while the remaining 33.1% form a long tail. In 10.89%, the model added the complete schema to the "*'entrada'*" key, followed by "*'saida'*"; 4.77% included autonomous generation of Python code with NLTK. Only 1.8% of responses presented the expected clean JSON, resulting in a low validity rate (90.65%) and evidencing instruction-following failures.

Table 26. LLaMA-3.1-8B-Instruct Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
2,788	1. {'type': 'object','description': Objeto de saída fornecido pelo classificadorapós a classificação de sentimento do texto de entrada.', 'properties': {'polaridade': {'type': 'integer','description': 'Polaridade em relação ao sentimento expressado no texto de entrada. Pode assumir 2 valores: [-1, 1]','enum': [-1,1]}}, \n 'required': ['polaridade']}\n \n'saida': {'polaridade': 1}	37.97 %
	2. {'type': 'object','description': Objeto de saída fornecido pelo classificadorapós a classificação de sentimento do texto de entrada.', 'properties': {'polaridade': {'type': 'integer','description': 'Polaridade em relação ao sentimento expressado no texto de entrada. Pode assumir 2 valores: [-1, 1]','enum': [-1,1]}}, \n 'required': ['polaridade']}\n \n'saida': {'polaridade': -1}	28.96 %
	3. {'polaridade': 1}	0.92 %
	4. {'polaridade': -1}	0.92 %
	5. Para realizar a classificação de sentimento, podemos utilizar uma abordagem baseada em técnicas de processamento de linguagem natural (NLP) e aprendizado de máquina. Aqui está um exemplo de como você pode fazer isso utilizando a biblioteca NLTK e scikit-learn em Python: \n\n```python \nimport nltk\nfrom nltk.sentiment import SentimentIntensityAnalyzer \nfrom nltk.tokenize import word_tokenize \nfrom nltk.corpus import stopwords \nfrom nltk.stem import WordNetLemmatizer \nfrom sklearn.feature_extraction.text import TfidfVectorizer \nfrom sklearn.model_selection import train_test_split \nfrom sklearn.linear_model import LogisticRegression \nfrom sklearn.metrics import accuracy_score \nimport json \n\n# Carregar o corpus de treinamento	0.50 %

C.3 PT-BR

Bode-3.1-8B-Instruct-lora The model presented hybrid behavior with high validity (99.7%) but low format fidelity, with only 20% clean responses (JSON only) and wide fragmentation, where only 67.5% of inferences are concentrated in the top 20 main variations. The analysis revealed systematic contamination by artifacts, 33.1% included anomalous code markers ("``````````````"), 13.3% containing verbose unsolicited explanations ("*Para realizar...*" or "*Para resolver...*"), and reproduction of demonstration elements (9.73%). The behavior characterizes a partially effective instruction-following pattern that correctly executes the classification task but fails to distinguish between demonstration structure and specific task execution, resulting in systematic contamination by structural elements of the provided examples. The overview of model response generation is presented in Table 27.

Table 27. Bode-3.1-8B-Instruct-lora Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
1,191	1. {'polaridade': -1}	13.46 %
	2. {'polaridade': 1}````````````````````````	9.55 %
	3. {'polaridade': -1}````````````````````````	7.51 %
	4. {'polaridade': 1}	6.46 %
	5. {'polaridade': 1}\nExemplo:\n'entrada': 'O filme é uma mist	4.66 %

InternLM-ChatBode-7B The model achieved a near-perfect classification validity rate (99.99%) but completely failed to follow the requested format, resulting in 0% clean responses, revealing paradoxical behavior. The model produced 952 distinct responses (Table 28), with outputs systematically contaminated by massive reproduction of prompt elements, with 50.0% of responses including the example structure (e.g.: "*Exemplo:\n'entrada':*") and 48.3% replicating the main instruction ("*Classificação de Sentimento:*"). Additionally, 12.22% of outputs contained truncated and literal fragments from example texts, such as "*A atuaç*" and "*Eu gostei*". This pattern characterizes instruction-following that executes the classification task with precision but is unable to distinguish the task from the prompt structure, making the model functional but inadequate for generating concise outputs.

Table 28. InternLM-ChatBode-7B Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
952	1. {'polaridade': 1}\nClassificação de Sentimento: 'entrada	13.22 %
	2. {'polaridade': -1}\nClassificação de Sentimento: 'entrada	11.34 %
	3. {'polaridade': -1}\nClassificação de Sentimento:	6.12 %
	4. {'polaridade': -1}\nClassificação de Sentimento:'entrada':	6.11 %
	5. {'polaridade': 1}\nClassificação de Sentimento:	5.93 %

CabraLLaMA-3-8B The model revealed extreme behavior with high diversity, generating 1,927 unique responses (Table 29). Only 28.9% of outputs corresponded solely to the requested JSON ("*{'polaridade': -1}*" with 17.0% and "*{'polaridade': 1}*" with 11.9%). The remaining 71.1% formed a long tail of 1,925 variations containing artifacts. These variations include massive reproduction of prompt elements, with 34.9% of responses containing "*Classificação de Sentimento:'entrada':*" and 22.0% reproducing "*Exemplo:\n'entrada':*", in addition to fragments of texts that refer to original examples such as "*O melhor filme de John*", "*Este filme é*" and "*O celular possui*", as well as content generation, such as "*eu odeio*" and "*Eu não entendo como*", evidencing capacity for contextually plausible but unsolicited text generation. Despite high dispersion, classification validity was high (99.3%), but response concentration was highly fragmented, and only 52.3% of inferences were concentrated in the top 20 main variations.

Table 29. CabraLLaMA-3-8B Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
1,927	1. {'polaridade': -1}	16.94 %
	2. {'polaridade': 1}	11.93 %
	3. {'polaridade': 1}\nClassificação de Sentimento: 'entrada': 'eu odeio	2.52 %
	4. {'polaridade': 1}\nClassificação de Sentimento: 'entrada': 'O filme é	2.28 %
	5. {'polaridade': 1}\nClassificação de Sentimento: 'entrada': 'O produto é	1.90 %

CabraMistral-v3-7B-32k The model presented relatively controlled behavior, with 287 unique responses (Table 30) and high concentration (92.3% in the top 20 main variations). The four main responses, representing 71.5% of the total. The model demonstrated systematic reproduction of prompt elements, such as "*Classificação de Sentimento:*" (present in 76.8% of responses) and "*'entrada':*" (53.6%). This pattern led to a total inability to generate clean outputs, with 0% of responses containing clean JSON, although classification validity was excellent (99.98%).

Table 30. CabraMistral-v3-7B-32k Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
287	1. {'polaridade': 1}\nClassificação de Sentimento:\n'entr	23.52 %
	2. {'polaridade': -1}\nClassificação de Sentimento:\n'entr	17.55 %
	3. {'polaridade': 1}\nClassificação de Sentimento:'entrada':	17.11 %
	4. {'polaridade': -1}\nClassificação de Sentimento:'entrada':	13.27 %
	5. {'polaridade': 1}\nExemplo:\n'entrada': 'E	4.59 %

GemBode-7B-Instruct The model demonstrated creative generation behavior with high dispersion, producing 1,726 unique responses (Table 31) with low concentration (43.2% in the top 20 main variations). The dominant pattern was output contamination: 94.0% of responses combined the prompt structure with autonomous and unsolicited text generation, such as "*Eu não sou um especialista*" (9.2%) and "*Eu não consigo entender por*" (4.4%). Consequently, only 3.8% of responses were clean, although classification validity remained excellent (99.68%).

Table 31. GemBode-7B-Instruct Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
1,726	1. {'polaridade': 1}\nExemplo:\n 'entrada': 'Eu não sou um especialista	4.88 %
	2. {'polaridade': -1}\nExemplo:\n 'entrada': 'Eu não sou um especialista	4.28 %
	3. {'polaridade': 1}\nExemplo:\n 'entrada': 'Eu não consigo entender por	3.57 %
	4. {'polaridade': -1}\nExemplo:\n 'entrada': 'Eu não sou um ci	3.49 %
	5. {'polaridade': 1}\nExemplo:\n 'entrada': 'não sei se o programa	2.52 %

Bode-7B The model presented bimodal and controlled behavior, with low response diversity (120 unique - Table 32) and high concentration (87.3% in the top 10 main variations). This pattern divided into two behaviors: 42.5% of outputs were the requested pure JSON (e.g.: "*{'polaridade': -1}*" and "*{'polaridade': 1}*"). In contrast, the remaining 57.5% contained structural artifacts, mainly the reproduction of "*Classificação de Sentimento:*" (present in 43.6% of responses). Despite the artifacts, the classification validity rate was excellent (99.8%), standing out as the model with the lowest diversity compared to other smaller-scale models with Portuguese fine-tuning.

Table 32. Bode-7B Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
120	1. {'polaridade': -1}	21.91 %
	2. {'polaridade': 1}	20.58 %
	3. {'polaridade': 1}\n Classificação de Sentimento:\n	16.86 %
	4. {'polaridade': 1}}\n Classificação de Sentimento:\n	6.68 %
	5. {'polaridade': 1}\n Classificação de Sentimento:\n	5.72 %

Bode-13B The model exhibited relatively controlled and clean behavior, with 219 (Table 33) unique responses and high concentration (93.7% in the top 10 main variations). Performance was excellent in generating clean outputs: the four main responses (87.1% of total) consisted of the requested JSON, differing only by initial space formatting. However, the main problem was the generation of the invalid value "*{'polaridade': 0}*" (2.2% of responses). The model maintained a validity rate of 97.5%, specifically impaired by the zero value problem, and presented minimal structural artifacts, characterizing behavior that almost perfectly executes the classification task but demonstrates occasional confusion about permitted valid values.

Table 33. Bode-13B Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
219	1. {'polaridade': 1}	36.29 %
	2. {'polaridade': -1}	23.29 %
	3. {'polaridade': 1}	15.77 %
	4. {'polaridade': -1}	11.71 %
	5. {'polaridade': 0}	2.18 %

Sabiá-7B The model demonstrated severely degraded behavior, producing $1,067$ unique responses (Table 34) with only 84.8% validity rate, the lowest observed. 81.8% of responses presented systematic truncation of prompt elements, evidenced by the three main responses that represent 76.0% of the total: "*{'polaridade': 1}\nClass*" (53.5%), "*{'polaridade': −1}\nClass*" (12.1%) and "*{'polaridade': 1}\nEx*" (10.4%), where "*Class*" and "*Ex*" seems to refer to truncation of the fragments "*Classificação de Sentimento:*" and "*Exemplo:*" present in the original prompts. Additionally, the model generated responses with spurious repetitive text such as "*de texto de texto de texto*", with severe structural deformations including patterns like "*1, 1, 1,*" and corrupted sequences, and only 2.3% of completely clean responses containing exclusively the requested JSON.

Table 34. Sabiá-7B Generation Overview.

Distinct Raw Responses	Top 5 Occurrences	%
1,067	1. {'polaridade': 1}\nClass	53.54 %
	2. {'polaridade': -1}\nClass	12.13 %
	3. {'polaridade': 1}\nEx	10.39 %
	4. {'polaridade': -1}\nEx	4.28 %
	5. {'polaridade': 1}\n\n	2.11 %