Portuguese Sentiment Analysis with Open-Source LLMs: Models, Prompts, and Efficient Deployment

João V R J Lima
Kunumi Lab - UFC
Universidade de Fortaleza
jvrodrigs@gmail.com

Vládia Pinheiro
Universidade de Fortaleza
vladiacelia@unifor.br

Carlos Caminha
Kunumi Lab - UFC
Universidade Federal do Ceará
caminha@ufc.br

Abstract

Robust sentiment analysis in Portuguese is central to applications across Lusophone contexts, yet systematic evaluations still focus predominantly on English and proprietary systems. This paper presents a comparative study of 29 open-source Large Language Models (LLMs) and two proprietary models on Portuguese sentiment classification under four prompting strategies: Zero-Shot, Few-Shot, Chain-of-Thought (CoT), and CoT with Few-Shot (CoT+FS). Experiments on a unified three-class benchmark built from three public review corpora (about 3,000 instances) comprise roughly 372,000 inferences, totaling approximately 150M input tokens and 65M output tokens. Results show that CoT+FS generally yields the best performance for larger models, while several compact open-source models obtain competitive F1-scores with substantially lower computational cost, making them suitable for real-world deployments. We identify concrete teacher–student configurations tailored for knowledge distillation in Portuguese sentiment analysis.

1 Introduction

The analysis of sentiments expressed in Portuguese is central to applications across Lusophone contexts, including customer feedback monitoring, reputation management, and public service evaluation. However, most systematic assessments of Large Language Models (LLMs) for sentiment analysis still concentrate on English and on proprietary systems, leaving the Portuguese NLP community with limited, fragmented evidence on how to effectively leverage recent open-source models and prompting strategies in real-world settings (Zhan et al., 2024; Hasan et al., 2024; Krugmann and Hartmann, 2024).

Traditionally, sentiment classification has relied on machine learning and statistical methods, such as bag-of-words, TF-IDF, and supervised models like Naive Bayes, SVMs, and decision trees (Medhat et al., 2014). With the advent of transformer-based architectures and contextual embeddings, models such as BERT and RoBERTa substantially improved the ability to capture linguistic nuances (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019). More recently, general-purpose LLMs with billions of parameters have demonstrated strong zero-shot and few-shot performance in multiple NLP tasks, including sentiment analysis, often without task-specific fine-tuning. Yet, their behavior for Portuguese remains less understood, particularly when compared systematically across different architectures, scales, and prompting techniques (de Araujo et al., 2024).

From the perspective of practitioners and institutions in Portuguese-speaking countries, several constraints intensify this gap. High-end proprietary LLMs are frequently accessible only via cloud APIs, raising concerns regarding cost, privacy, data sovereignty, and long-term dependency on external providers. In contrast, open-source LLMs, which can be deployed on-premise or in controlled environments, provide a promising alternative, especially when combined with quantization and hardware-aware optimization. Nonetheless, there is a lack of empirical evidence indicating which open-source models, and which inference-time prompting strategies, offer the best trade-off between accuracy and computational efficiency for sentiment analysis in Portuguese.

In this work, we address this gap by conducting a systematic evaluation of 29 open-source LLMs and two proprietary baselines on Portuguese sentiment classification. We frame the task as labeling user-generated review texts into three classes (positive, neutral, negative), using a unified benchmark built from multiple public Portuguese corpora. We compare four prompting strategies commonly adopted in practical deployments (Zero-Shot, Few-Shot, Chain-of-Thought, and Chain-of-Thought

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 212–221

with Few-Shot) to understand how model capacity and prompt design interact in this setting. Beyond reporting performance numbers, we focus on configurations that are feasible for local or low-cost deployment.

Our contributions are threefold: (1) We compiled and unified evaluation set Portuguese sentiment analysis benchmark for prompting-based evaluation, built from real user-generated reviews and mapped to a consistent three-class labeling scheme; (2) We systematically compare four prompting strategies across a diverse set of LLM families and sizes, identifying which combinations of model scale and prompt design are most effective and computationally viable for Portuguese sentiment classification; (3) We identify concrete teacher–student model pairs that are well-suited for future knowledge distillation targeting Portuguese, highlighting compact open-source models that achieve competitive performance.

2 Related Work

Early work on sentiment analysis was dominated by lexical resources, rule-based approaches, and traditional machine learning methods built on sparse representations such as bag-of-words and TF-IDF combined with classifiers like Naive Bayes, SVMs, and logistic regression (Medhat et al., 2014; Turney, 2002). These approaches played a central role in establishing sentiment classification as a practical task, but they typically rely on task-specific feature engineering and struggle with informal language, domain shift, and nuanced or implicit opinions.

For Portuguese, a series of studies explored sentiment classification using both classical machine learning and neural architectures on reviews, microblogs, and other user-generated content. Many of these works rely on supervised models trained on language-specific or multilingual embeddings and on corpora such as product reviews and social media datasets, often focusing on Brazilian Portuguese. With the advent of transformer-based language models, approaches based on models like BERTimbau and multilingual encoders (e.g., mBERT, XLM-R) became the de facto standard for Portuguese sentiment analysis (Souza et al., 2020; Devlin et al., 2019; Conneau et al., 2020). These models are usually fine-tuned for the target dataset, yielding strong results but requiring labeled data and model retraining for each domain.

Closer to our setting, Silva et al. introduced RePro, a 10k-review benchmark of Brazilian Portuguese e-commerce opinions annotated with both sentiment and coarse-grained topics, built from the B2W-Reviews01 corpus and accompanied by BERTimbau baselines for supervised sentiment and topic classification (dos Santos Silva et al., 2024). While RePro offers high-quality human labels and a valuable resource for aspect-aware opinion mining, it does not evaluate instruction-tuned LLMs or prompting strategies and is restricted to a single e-commerce source; our work is complementary, providing an inference-only comparison of dozens of LLMs and prompts on a unified three-class benchmark spanning three public review corpora.

More recently, instruction-tuned Large Language Models have been investigated for sentiment classification through prompting rather than fine-tuning. Several studies have shown that general-purpose LLMs (e.g., GPT-style and open-source instruction-tuned models) can achieve competitive or superior performance to traditional supervised baselines in zero-shot and few-shot regimes (Zhan et al., 2024; Hasan et al., 2024; Krugmann and Hartmann, 2024; Caminha et al., 2025a,b; Lasheras et al., 2025a,b). However, this line of work is predominantly centered on English and high-resource settings, with limited attention to Portuguese or to systematic comparisons across a broad set of open-source models. As a result, there is still little consolidated evidence on how different LLM families and sizes behave for sentiment analysis in Portuguese when used purely via prompting.

Prompt engineering strategies such as Zero-Shot, Few-Shot, and Chain-of-Thought (CoT) have been widely studied as mechanisms to steer LLM behavior and improve reasoning and classification quality (Kojima et al., 2022; Wei et al., 2022). While CoT and in-context learning have shown clear benefits for arithmetic and symbolic reasoning tasks, their impact on comparatively simpler classification problems, including sentiment analysis, is less consistent and often task-dependent. Existing works typically evaluate a small number of models or prompts, and there is, to the best of our knowledge, no large-scale study systematically comparing multiple prompting strategies for Portuguese sentiment classification across dozens of open-source and proprietary LLMs.

In parallel, a growing body of research has focused on efficient LLM deployment, exploring techniques such as quantization, pruning, parameter-efficient fine-tuning, knowledge distil-

lation, and Mixture-of-Experts (MoE) architectures (Hinton et al., 2015; Frantar et al., 2023). These approaches aim to reduce inference cost while preserving most of the predictive performance of larger teacher models. Prior work rarely provides language-specific empirical guidance—particularly for Portuguese—on selecting teacher and student models, leaving practitioners without clear evidence-based recommendations for compact open-source sentiment classifiers.

3 Methodology

This section describes the methodology adopted in this study. We cast the problem as a classification task whose goal is to assign a user-authored opinion text to one of the predefined sentiment classes: positive, negative, or neutral. For example, given the input text "O produto chegou rapidamente, mas a embalagem estava danificada.", the model should return the label "neutral".

To solve the classification task, we employed Large Language Models (LLMs) combined with different *prompt engineering* strategies, with the aim of steering model outputs and improving classification accuracy. We evaluated the effectiveness of the approach on the labeled dataset. Model predictions were compared with the original labels, using F1 score as the primary metric, appropriate for multi-class tasks with originally imbalanced classes.

Finally, we conducted a comparative analysis across different model configurations, varying LLM families, model sizes, and *prompt* strategies. The objective is to identify which combinations offer the best performance in terms of accuracy and computational cost.

3.1 Open Source Large Language Models

To conduct this study, we selected Large Language Models (LLMs) based on their aggregated performance as presented on the Chatbot Arena platform[1], considering data available up to August 15, 2025.

The selection followed three main criteria:

1. Model families with available versions up to 120 billion parameters;

2. Availability of pre-quantized versions in GPT-Generated Unified Format (GGUF), quantized at 4 bits;

3. Availability in the huggingface lmstudio-community [2].

Hugging Face Path	Size (B)
deepseek-r1-distill-llama-70b	70
deepseek-r1-distill-qwen-32b	32
deepseek-r1-distill-qwen-14b	14
deepseek-r1-distill-llama-8b	8
deepSeek-r1-distill-qwen-7b	7
deepseek-r1-distill-qwen-1.5b	1.5
gpt-oss-120b	120
gpt-oss-20b	20
gemma-3-27b-it	27
gemma-3-12b-it	12
gemma-3-4b-it	4
gemma-3-1b-it	1
llama-4-scout-17b-16e-instruct	109
llama-3.3-70b-instruct	70
meta-llama-3.1-8B-Instruct	8
llama-3.2-3b-instruct	3
llama-3.2-1b-instruct	1
qwen3-32b	32
qwen3-30b-a3b	30
qwen3-14b	14
qwen3-8b	8
qwen3-1.7b	1.7
qwen3-0.6b	0.6
qwen3-4b	4
phi-4	14
phi-4-mini-instruct	3.8
mistral-small-3.1-24b-instruct-2503	24
mistral-7b-instruct-v0.3	7
sabia-7b	7

Table 1: Overview of the open-source LLMs used in this study, detailing the source platform path and parameter count (in billions).

The first criterion ensures comparability between models and the feasibility of running the experiments. The second and third allows for optimized use of video memory (VRAM) during inference.

Table 1 provides an overview of the models used in this study, including the Hugging Face path and the number of parameters (in billions).

We added two proprietary OpenAI models to the experiment, specifically `gpt-5` and `gpt-5-nano`. The purpose of including these proprietary models in this study is to provide a reference point for comparison against state-of-the-art closed-source models.

3.2 Prompting Techniques

In this work, we evaluate four prompting strategies for each LLM: (i) Zero-Shot, (ii) Few-Shot, (iii) Chain-of-Thought (CoT), and (iv) CoT with Few-Shot (CoT+FS). In all cases, the prompt instructs the model to classify the sentiment expressed in the user's text and return a single sentiment label.

The strategies are defined as follows:

i **Zero-Shot:** A single generic prompt describing the task, with no examples or explicit reasoning.

[1] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

[2] https://huggingface.co/lmstudio-community

ii **Few-Shot:** The generic prompt augmented with a few labeled examples.

iii **Chain-of-Thought (CoT):** The prompt asks the model to reason step by step before answering.

iv **CoT + Few-Shot:** Combines labeled examples with explicit step-by-step reasoning instructions.

All prompts, the 3,000-instance test set, and the inference code are available in our GitHub repository.[3]. Figure 1 illustrates these prompts.

Common Prompt

```
You are an intelligent assistant capable of classifying sentiment based on the provided text. Your main function is to meticulously analyze the given text and classify which sentiment is most likely to be identified.
Consider the following classifications:
* Positive: Indicates a favorable or optimistic attitude (e.g., joy, satisfaction, enthusiasm).
* Negative: Refers to an unfavorable or pessimistic attitude (e.g., anger, frustration, sadness).
* Neutral: Indicates that the text does not express a clear or strong sentiment (e.g., factual information).

When the user provides a text, analyze and identify the most likely classification for the given text. Respond ONLY with a Python dictionary where the key is "sentiment" and the value is the appropriate sentiment from the list above. Do not add any text in your response, under any circumstances. *Make sure the suggested sentiment is STRICTLY one from the provided list and indicate only ONE sentiment. The classification must EXACTLY MATCH one of the options listed above, without variations in capitalization, spacing, or wording. Do not add or remove spaces, do not change uppercase or lowercase letters, and do not alter the names in any way.*
```

Chain of Thoughts

```
Before the final response, explain step-by-step how you interpreted the user's provided text. After explaining, based on your reasoning, give your final response in the expected format. Respond with only a Python dictionary where the key is "sentiment" and the value is the appropriate classification from the list above. There is no need to generate any additional text in your response, under any circumstance. *Make sure the suggested sentiment is STRICTLY one from the provided list and indicate only ONE sentiment. The classification must EXACTLY MATCH one of the options listed above, without variations in capitalization, spacing, or wording. Do not add or remove spaces, do not change uppercase or lowercase letters, and do not alter the names in any way.*
```

Few-shot

```
Here are some examples of user texts:
User text: "O produto chegou no prazo, mas a embalagem estava um pouco danificada.
Response: {{"sentiment": "Neutral"}}

User text: "Estou extremamente feliz com o atendimento, todos foram muito atenciosos!"
Response: {{"sentiment": "Positive"}}

User text: "O atendimento foi péssimo, fiquei esperando por horas e ninguém me ajudou."
Response: {{"sentiment": "Negative"}}

User text: "Péssimo!"
Response: {{"sentiment": "Negative"}}
```

Common Prompt

```
Now, analyze the following user text and classify the sentiment:
User text: {user_text}
Response:
```

Figure 1: Overview of prompting strategies.

3.3 Evaluation Metrics

We use F1-score as the main evaluation metric, as it balances Precision and Recall and is suitable for multi-class settings. For a single class, F1 combines Precision and Recall into a single value, $\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$, with Precision and Recall defined in the usual way from TP, FP, and FN.

We report the *weighted* F1-score over the three sentiment classes (negative, neutral, positive), where each per-class F1 is weighted by its number of instances, thus explicitly accounting for class imbalance.

Across all prompting strategies, models are instructed to output a Python dictionary with a single attribute `"sentiment"`. Its value (`Negative`, `Neutral`, or `Positive`) is mapped to the corresponding class label and used in the weighted F1 computation.

4 Experimental Evaluation

4.1 Dataset

For evaluation, we built a gold collection by combining three public datasets: Olist (Olist, 2018)—In 2018, the largest department store in the Brazilian market—released the "Brazilian E-commerce Public Dataset by Olist" on the Kaggle platform, a dataset with roughly 100,000 orders from 2016 to 2018 provided by various marketplaces in Brazil. B2W (Real et al., 2019)—In 2019, B2W Digital made available "B2W-Reviews01" an open corpus of product reviews containing about 130,000 e-commerce customer reviews collected from the Americanas.com website. Finally, UTLCorpus (Sousa et al., 2019), which is the most extensive set, with about 2 million reviews. It comprises two datasets: Movie Reviews, collected from Filmow, and App Reviews, collected from the Play Store.

In this work, because our focus was to build a benchmark, we deemed that constructing a sample by consolidating the three datasets would yield a more informative analysis. Sampling was stratified by dataset: we selected 1,000 examples from each (totaling 3,000 texts) via simple random sampling, with a fixed seed [seed=20] to ensure reproducibility. The three datasets annotated sentiment on a numeric 1–5 scale. We harmonized the labels to the ternary scheme adopted in this study via the function $\phi : \{1, 2, 3, 4, 5\} \rightarrow \{negative, neutral, positive\}$, such that $\phi(\{1, 2\})$ = negative, $\phi(3)$ = neutral, and $\phi(\{4, 5\})$ = positive. After harmonization, the resulting gold collection is substantially imbalanced, with 64.53% positive, 23.40% negative, and 12.07% neutral examples.

Sample-size planning studies for classification models show that hundreds of test cases are

[3] https://github.com/jvrodrigs/analisys-sentiment-ptbr-propor

typically required to estimate performance with confidence-interval widths below 0.05, and that at least about 75–100 test samples per class are needed even for coarser precision levels (Beleites et al., 2013). Our benchmark satisfies these conditions, providing several hundred examples in each sentiment class (exactly 1,936 positive, 702 negative, and 362 neutral instances).

4.2 Evaluation Setup

To evaluate the language models listed in Table 1, we ran all experiments locally using LM Studio with the `llama.cpp`[4] implementation. We employed pre-quantized 4-bit (INT4) checkpoints in the GPT-Generated Unified Format (GGUF), which substantially reduce GPU memory requirements while preserving model quality. All layers of each model were fully loaded into GPU memory to maximize inference throughput. The experiments were conducted on a workstation equipped with an NVIDIA RTX PRO 6000 GPU with 96 GB of VRAM.

Ensuring consistency in model outputs is crucial for evaluating LLM-based sentiment classification, since structured outputs enable automated processing and integration into classification pipelines. In our setup, all prompts instruct the model to return a Python dictionary with a single attribute named `"sentiment"`, whose value is the predicted sentiment label. If this attribute cannot be recovered from the model output, the corresponding instance is treated as an error and counted as a misclassification in the F1-score computation.

To extract the predicted sentiment from the generated text, we apply a regular expression that searches for a Python-style dictionary anywhere in the model output. This parser is tolerant to additional content, such as chain-of-thought explanations or other surrounding text, as long as the dictionary itself is well formed. Figure 2 illustrates this behavior with six representative examples: the first row, (a), (b), and (c), shows valid outputs in which the `"sentiment"` dictionary is correctly produced, either as a standalone JSON-like snippet or embedded in well-structured text; the second row, (d), (e), and (f), shows invalid outputs where the dictionary is missing, malformed, or otherwise not recoverable by our parser.

As a supervised reference, we fine-tuned BERTimbau-base (Souza et al., 2020) for three-way

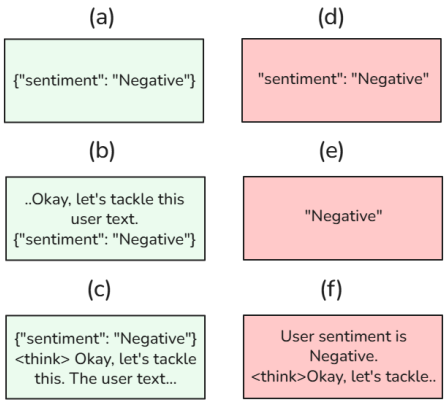

Figure 2: Examples of successful (a–c) and unsuccessful (d–f) outputs from the sentiment classification task using user texts.

sentiment classification (negative, neutral, positive). We sampled 30,000 labeled reviews (10,000 for each label) from the union of Olist, B2W and UTL after removing the 3,000 instances used in our test benchmark, and split them into 24,000 training and 6,000 validation examples with stratification over the ternary label space. The model was fine-tuned using the HuggingFace `transformers` library with standard hyperparameters for BERT-style encoders: maximum sequence length of 256 tokens, learning rate of $2 \cdot 10^{-5}$, batch size 16 for training and 32 for evaluation, weight decay of 0.01, and 3 epochs of training. We report the weighted F1-score of the resulting model on the same 3,000-instance test set used for the LLM-based experiments.

5 Results

Table 2 summarizes the F1-scores obtained by each model and prompting strategy. Figure 3 provides an aggregated view of the relationship between model size and performance for each strategy. Unless otherwise noted, differences discussed refer to weighted-F1 and should be interpreted in light of the reported standard error of means (SEM), estimated via bootstrap resampling over the test set.

We report weighted F1-scores over the three sentiment classes (negative, neutral, positive). Each model was evaluated under the four prompting strategies described in Section 3 (Zero-Shot, Few-Shot, CoT, and CoT+FS), using the same test set. In total, the experiments comprise roughly 372,000 model calls, processing exactly 150,610,700 input tokens and 65,679,508 output tokens across all combinations of models and prompts.

One finding that emerges from our compar-

[4] https://github.com/ggerganov/llama.cpp

Table 2: F1-Score results with standard deviation for the LLMs evaluated across four prompt strategies. The values in bold indicate the highest F1 scores for each model, considering the margin of error.

LLM	ZS	FS	CoT	CoT+FS
gpt-5	0.787 ± 0.008	**0.800 ± 0.007**	0.778 ± 0.008	**0.802 ± 0.007**
gpt-5-nano	0.783 ± 0.008	**0.788 ± 0.008**	0.743 ± 0.008	**0.794 ± 0.007**
gpt-oss-120b	0.783 ± 0.008	**0.797 ± 0.007**	0.783 ± 0.008	**0.798 ± 0.007**
gpt-oss-20b	0.786 ± 0.008	0.791 ± 0.008	0.774 ± 0.008	**0.799 ± 0.008**
DeepSeek-R1-Distill-Llama-70B	0.787 ± 0.008	0.789 ± 0.007	0.788 ± 0.008	0.783 ± 0.007
DeepSeek-R1-Distill-Qwen-32B	0.780 ± 0.007	0.777 ± 0.008	0.777 ± 0.008	**0.788 ± 0.007**
DeepSeek-R1-Distill-Qwen-14B	0.77 ± 0.008	0.769 ± 0.007	0.775 ± 0.008	0.775 ± 0.007
DeepSeek-R1-Distill-Llama-8B	**0.756 ± 0.008**	**0.763 ± 0.008**	0.745 ± 0.008	**0.760 ± 0.008**
DeepSeek-R1-Distill-Qwen-7B	**0.754 ± 0.008**	0.745 ± 0.008	0.745 ± 0.008	0.744 ± 0.008
DeepSeek-R1-Distill-Qwen-1.5B	**0.692 ± 0.008**	**0.695 ± 0.008**	0.676 ± 0.009	0.673 ± 0.008
gemma-3-27b-it	0.787 ± 0.008	**0.798 ± 0.007**	0.778 ± 0.008	**0.797 ± 0.007**
gemma-3-12b-it	0.782 ± 0.008	**0.787 ± 0.008**	0.780 ± 0.008	**0.792 ± 0.008**
gemma-3-4b-it	0.753 ± 0.008	**0.769 ± 0.008**	0.757 ± 0.008	**0.774 ± 0.007**
gemma-3-1b-it	0.490 ± 0.009	**0.757 ± 0.008**	0.739 ± 0.008	**0.754 ± 0.008**
llama-4-scout-17b-16e-instruct	0.799 ± 0.008	**0.817 ± 0.007**	0.780 ± 0.009	**0.811 ± 0.007**
llama-3.3-70B-Instruct	0.790 ± 0.008	**0.807 ± 0.008**	0.782 ± 0.008	**0.812 ± 0.007**
meta-llama-3.1-8B-Instruct	0.776 ± 0.008	**0.791 ± 0.008**	0.767 ± 0.008	**0.796 ± 0.007**
llama-3.2-3B-Instruct	0.717 ± 0.009	**0.784 ± 0.008**	0.705 ± 0.008	**0.790 ± 0.008**
llama-3.2-1B-Instruct	0.341 ± 0.010	**0.758 ± 0.008**	0.376 ± 0.009	**0.758 ± 0.008**
mistral-small-3.2-24B-Instruct-2506	0.774 ± 0.008	**0.796 ± 0.008**	0.778 ± 0.008	**0.797 ± 0.008**
mistral-7B-Instruct-v0.3	0.766 ± 0.008	**0.780 ± 0.008**	0.764 ± 0.008	**0.777 ± 0.008**
Phi-4	**0.795 ± 0.008**	**0.798 ± 0.007**	0.785 ± 0.008	**0.800 ± 0.007**
Phi-4-mini-instruct	0.660 ± 0.008	0.766 ± 0.008	0.747 ± 0.008	**0.781 ± 0.007**
Qwen3-32B	0.784 ± 0.008	0.781 ± 0.008	**0.793 ± 0.008**	0.784 ± 0.007
Qwen3-30B-A3B-Thinking-2507	**0.744 ± 0.008**	**0.738 ± 0.008**	0.735 ± 0.008	**0.745 ± 0.008**
Qwen3-14B	0.779 ± 0.008	**0.780 ± 0.008**	**0.782 ± 0.008**	**0.787 ± 0.007**
Qwen3-8B	0.771 ± 0.008	**0.774 ± 0.008**	0.770 ± 0.008	**0.779 ± 0.008**
Qwen3-4B	0.774 ± 0.008	**0.780 ± 0.008**	0.773 ± 0.008	**0.784 ± 0.007**
Qwen3-1.7B	**0.768 ± 0.008**	**0.774 ± 0.008**	0.762 ± 0.008	**0.775 ± 0.008**
Qwen3-0.6B	0.739 ± 0.009	**0.752 ± 0.008**	0.734 ± 0.009	**0.752 ± 0.008**
Sabia-7b	0.518 ± 0.011	0.490 ± 0.009	0.306 ± 0.009	**0.769 ± 0.009**

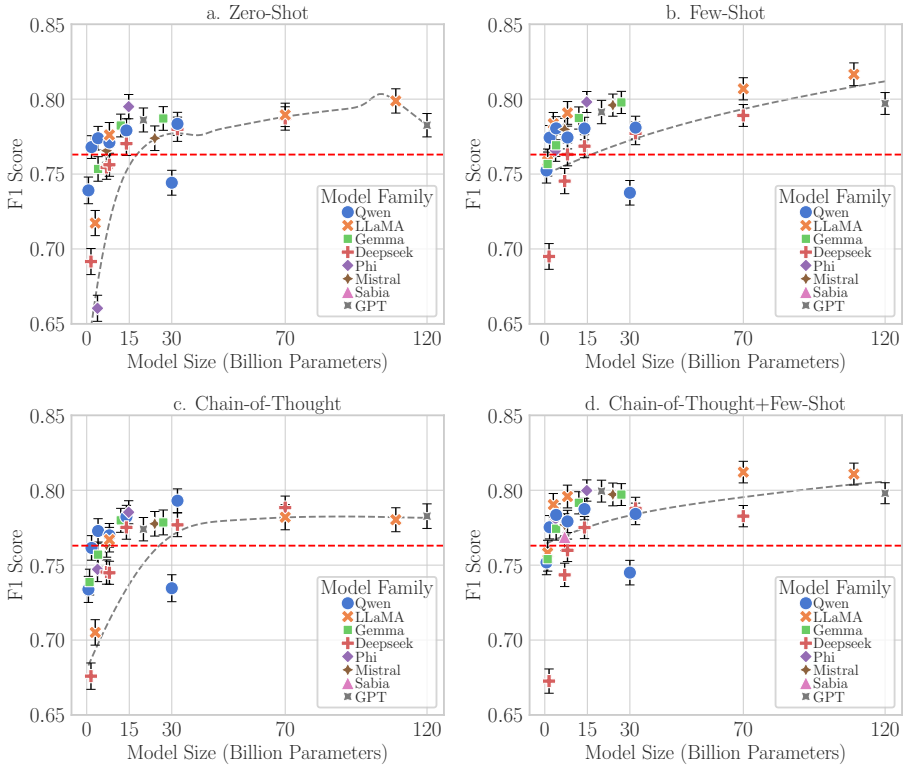

Figure 3: Comparison of F1-Scores across four prompting strategies as a function of model size (in billions of parameters). The dashed gray lines represent non-parametric regressions, and the dashed red lines represent BERTimbau. OpenAI's proprietary models were not included in this figure due to the lack of official disclosure of their size.

ative analysis is that the best-performing open-source models outperform proprietary. In particular, `llama-4-scout` achieves the highest weighted F1 (0.817), closely followed by `llama-3.3-70B` (0.812), both outperforming the strongest proprietary model evaluated, `gpt-5`, which achieves 0.802. This result highlights the maturity and competitiveness of current open-source LLMs, especially when combined with structured prompting strategies.

As a supervised reference, a BERTimbau-base model fine-tuned on 24k training examples from the same corpora reaches a weighted F1 of 0.7603 on our test set. This is below the best LLM configurations reported in Table 2, but comparable to several short-sized open-source models, reinforcing that prompting-based LLMs can match or surpass a strong Portuguese encoder baseline in this setting.

5.1 Prompting Strategies

Overall, Few-Shot and CoT+FS prompting tend to outperform plain Zero-Shot for most models, while pure CoT (without examples) rarely provides substantial gains for this task. CoT+FS emerges as the most robust strategy in the sense that it is frequently the best or statistically tied with the best configuration for a given model, and it is seldom the worst.

A notable exception is `sabia-7b`, which performs poorly under Zero-Shot (0.518) and Few-Shot (0.490) but rises sharply to 0.769 with CoT+FS—a jump far larger than that observed for any other model. This atypical behavior suggests that the model may have strong underlying language capabilities but weak instruction-following when guidance is minimal, with the structured reasoning in CoT+FS compensating for this limitation. Other exceptional behaviors can be observed with `gemma-3-1b-it`, `llama-3.2-1B-Instruct` and `Phi-4-mini-instruct`. All of them improve significantly when using FS or CoT.

5.2 Effect of Model Size

Figure 3 reveals a general trend in which weighted-F1 improves steadily with model size up to roughly 20–30B parameters, after which gains tend to flatten for most model families. Several architectures in the 30–70B range exhibit this plateau effect, achieving scores close to the best-performing models while requiring substantially fewer parameters.

An important exception to this pattern is the `llama-4-scout`, a 109B-parameter model that achieves the highest F1 in our benchmark. Its performance surpasses the 70B models by a non-trivial margin, indicating that extremely large models can still yield measurable improvements despite the overall diminishing-returns trend observed below this scale. In other words, while returns taper off beyond 30B for most families, the 109B model stands as an outlier that breaks the plateau.

For the specific task and benchmark considered, these results suggest that mid-sized models offer strong accuracy–efficiency trade-offs, but that very large models may still provide incremental gains at the extreme high end of the parameter spectrum. This distinction is relevant for practitioners: mid-sized models may suffice for practical deployments, whereas organizations prioritizing maximum accuracy above all else may still benefit from models in the 100B+ range.

5.3 Efficient Open-Source Models

A key outcome of our evaluation is the identification of compact open-source models that achieve competitive results under our prompting strategies. Several models in the 1–8B range reach weighted F1-scores close to those of substantially larger models, especially when guided by examples and strict output formatting.

These findings show that modern instruction-tuned open-source LLMs can serve as effective backbones for Portuguese sentiment analysis without task-specific fine-tuning, provided that prompts are carefully designed, and that the performance gap between compact and mid-sized models is often much smaller than the gap in parameter count. From a practical perspective, such models are strong candidates for on-premise or privacy-sensitive deployments, as they substantially reduce hardware requirements while preserving robust classification quality.

5.4 Teacher–Student Candidates and Pareto Frontier

To assess the trade-off between performance and model size, we analyze the CoT+FS setting—the most robust prompting strategy in our experiments—using a Pareto frontier defined over (model size, weighted-F1). Figure 4 presents this frontier: models positioned on or near it are those for which no other evaluated model achieves simultaneously higher F1 and fewer parameters.

The analysis reveals a set of Pareto-optimal compact and mid-sized architectures.

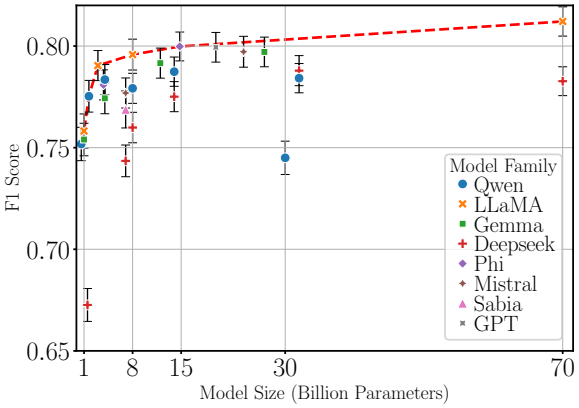

Figure 4: Pareto frontier applied to the results with the CoT + FS strategy.

Notably, `gemma-3-1b-it` (0.754 F1) and `llama-3.2-1B-Instruct` (0.758 F1) achieve performance little below to substantially larger models, such as `llama-3.3-70B` (0.812 F1), despite the large gap in parameter count. These models constitute natural "student" candidates for future distillation pipelines tailored to Portuguese sentiment analysis.

We also observe that the strongest performance below the 100B scale occurs around 70B parameters, where models such as `llama-3.3-70B` reach the highest F1 scores within this trade-off. Although `llama-4-scout` (109B) remains the overall best-performing model, its improvement over the 70B family is relatively small, indicating that 70B architectures represent a practical peak in the accuracy–efficiency balance. In this case, these models are natural candidates to be "teachers" in distillation

6 Conclusion

This work presented a systematic evaluation of Large Language Models for sentiment analysis in Portuguese, comparing 29 open-source models and two proprietary baselines under four prompting strategies on a unified three-class benchmark built from public review corpora. Using weighted F1-score as the main metric, we examined how model size, architecture, and prompt design interact in a realistic, inference-only scenario without task-specific fine-tuning.

Our results results three main findings. First, Few-Shot and CoT+FS prompting consistently improve performance over plain Zero-Shot for most models, while pure CoT (without examples) offers more limited and model-dependent gains. Second, performance tends to saturate around mid-sized models (approximately 20–30B parameters), with only modest improvements observed for larger models in our setting. Third, several compact open-source models in the 1–8B range achieve competitive weighted F1-scores when combined with well-designed prompts, offering favorable accuracy–efficiency trade-offs for deployment in environments with restricted computational resources or stricter data control requirements.

These findings have practical implications for the design of Portuguese sentiment analysis systems based on LLMs. They suggest that (i) careful prompt engineering is often more beneficial than relying solely on larger models, (ii) mid-sized open-source models are strong candidates when balancing performance and cost, and (iii) selected compact models can serve as viable backbones for locally deployable solutions. Moreover, by analyzing the accuracy–size Pareto frontier, we highlight concrete combinations of larger and smaller models that emerge as promising teacher–student configurations for future knowledge distillation pipelines tailored to Portuguese.

This study has limitations. Our benchmark is restricted to review-style opinion texts and may not fully capture other genres, domains, or varieties of Portuguese. Because it is built from publicly available corpora, there is a non-zero risk of overlap with the pre-training data of some evaluated models. Nonetheless, existing audits indicate that general-purpose LLMs are overwhelmingly trained on English (Bai et al., 2023; Team et al., 2024; DeepSeek-AI et al., 2024; Touvron et al., 2023), with Portuguese representing only a small fraction of their corpora, and our 3,000-instance test set is negligible at that scale; the heterogeneous performance patterns we observe across architectures and sizes are therefore more consistent with genuine generalization than with heavy memorization of this specific benchmark. As future work, we plan to broaden the evaluation to additional sentiment-related tasks and textual domains and to operationalize knowledge-distillation pipelines informed by the empirical landscape reported here.

7 Acknowledgements

The authors would like to thank **FUNCAP** (Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico) and the **Kunumi Institute** for their support of this research.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Claudia Beleites, Ute Neugebauer, Thomas Bocklitz, Christoph Krafft, and Jürgen Popp. 2013. Sample size planning for classification models. *Analytica Chimica Acta*, 760:25–33.

Carlos Caminha, Maria Silva, Iago Chaves, Felipe Brito, Victor Farias, and Javam Machado. 2025a. Evaluating llms and prompting strategies for automated hardware diagnosis from textual user-reports. In *Anais do LII Seminário Integrado de Software e Hardware*, pages 287–298, Porto Alegre, RS, Brasil. SBC.

Carlos Caminha, Maria de Lourdes M Silva, Iago C Chaves, Felipe T Brito, Victor AE Farias, and Javam C Machado. 2025b. Diaghw: A compact llm for hardware failure diagnosis via a novel knowledge distillation pipeline. In *Brazilian Conference on Intelligent Systems*, pages 393–407. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Gladson de Araujo, Tiago de Melo, and Carlos Mauricio S Figueiredo. 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*, pages 13–21.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, and 69 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Nildaimon dos Santos Silva, Livy Real, Ana Claudia Bianchini Zandavalle, Carolina Francisco Gadelha Rodrigues, Tatiana da Silva Gama, Fernando Guedes Souza, and Phillipe Derwich Silva Zaidan. 2024. Repro: a benchmark for opinion mining for brazilian portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 432–440.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers. *Preprint*, arXiv:2210.17323.

Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2024. Zero- and few-shot prompting with LLMs: A comparative study with fine-tuned models for Bangla sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17808–17818, Torino, Italia. ELRA and ICCL.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Johannes O. Krugmann and Julian Hartmann. 2024. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(3).

Uriel Lasheras, Elioenai Alves, Caio Ponte, Carlos Caminha, and Vládia Pinheiro. 2025a. Open llms meet causality in portuguese: A corpus-based fine-tuning approach. *Journal of the Brazilian Computer Society*, 31(1):1005–1030.

Uriel Lasheras, Elioenai Alves, Caio Ponte, Carlos Caminha, Diego Silva, and Vládia Pinheiro. 2025b. Ai agents powered by open-source language models for causal understanding in portuguese. In *Brazilian Conference on Intelligent Systems*, pages 146–161. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Olist. 2018. Brazilian e-commerce public dataset by olist. `https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce`. Kaggle.

Livy Real, Márcio Oshiro, and Alexandre Mafra. 2019. B2w-reviews01: An open product reviews corpus. In *Proceedings of the Symposium in Information*

and Human Language Technology (STIL). Sociedade Brasileira de Computação.

Rogério Figueredo Sousa, Henrico Bertini Brum, and Maria das Graças Volpe Nunes. 2019. A bunch of helpfulness and sentiment corpora in brazilian portuguese. In *Proceedings of Symposium in Information and Human Language Technology - STIL*. Sociedade Brasileira de Computação.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Tong Zhan, Chenxi Shi, Yadong Shi, Huixiang Li, and Yiyu Lin. 2024. Optimization techniques for sentiment analysis based on llm (gpt-3). *Preprint*, arXiv:2405.09770.