Journal of the Brazilian Computer Society, 2025, 31:1, doi: 10.5753/jbcs.2025.5667

Rating Prediction in Brazilian Portuguese: A Benchmark of Large Language Models

Emanuelle Marreira [**Amazonas State University** | ***erm.eng22@uea.edu.br***]
Tiago de Melo [**Amazonas State University** | ***tmelo@uea.edu.br***]
Miguel de Oliveira [**Federal University of Amazonas** | ***miguel.oliveira@icomp.ufam.edu.br***]
Carlos M. S. Figueiredo [**Amazonas State University** | ***cfigueiredo@uea.edu.br***]

Amazonas State University, Av. Darcy Vargas, 1.200 - Parque Dez de Novembro, Manaus - AM, 69050-020, Brazil.

Received: 11 March 2025 • **Accepted:** 27 May 2025 • **Published:** 03 October 2025

Abstract This study evaluates the performance of Large Language Models (LLMs) in predicting ratings for Brazilian Portuguese user reviews. We benchmark ten LLMs, including ChatGPT-3.5, ChatGPT-4o, DeepSeek, Mistral, LLaMA (3, 3.3), Gemma (1, 2), and the Brazilian Portuguese-specific models Sabiá-3 and Sabiazinho, using two prompting strategies: simple ($p1$) and detailed ($p2$). Results indicate that ChatGPT-4o and DeepSeek achieved the highest accuracy, particularly in predicting extreme ratings (1 and 5 stars). Sabiá-3 also performed competitively, highlighting the potential of language-specific models. Models performed better in objective categories such as food and baby products but struggled with more subjective domains like automotive and games. Cost analysis showed that DeepSeek is a more cost-effective alternative to ChatGPT-4o while maintaining similar accuracy. This study provides a systematic benchmark of LLMs for rating prediction in Brazilian Portuguese, offering insights into their effectiveness and limitations.

Keywords: Rating Prediction, Large Language Models, Natural Language Processing

1 Introduction

In recent years, digital platforms have undergone significant expansion, leading to a massive volume of user-generated textual content such as reviews on e-commerce websites and social media. In these environments, users share their reviews with the aim of assisting others in making informed decisions about products and services [de Melo *et al.*, 2019]. These contributions are recognized as valuable resources in the current digital market landscape. Automated extraction of sentiments from user-written opinions on specific topics is known as opinion mining or sentiment analysis [Liu and Zhang, 2012].

The rating assigned to online reviews serves as a key indicator for assessing the quality of products or services, typically expressed through a star-based system (e.g., a scale of 1 to 5) where higher ratings indicate positive perceptions of customers. Purchasing decisions are often influenced by helpful reviews, particularly those from reliable sources, with statistical correlations observed between higher ratings and favorable content [Wang *et al.*, 2022]. However, many digital platforms, such as social media and forums, lack explicit rating systems because of the impracticality of exhaustive evaluations or technical constraints that allow only free-text submissions. Given that ratings are highly relevant but not always available, the development of automated systems for rating prediction becomes essential. In such cases, machine learning models can be used to infer missing ratings based solely on textual input, a task widely recognized in the literature as *review rating prediction* [Ahmed and Ghabayen, 2022; Barman *et al.*, 2024]. Figure 1 shows an example in which the customer writes a review and assigns a rating to a product, and a machine learning model is expected to receive this review and predict the rating given by the customer.

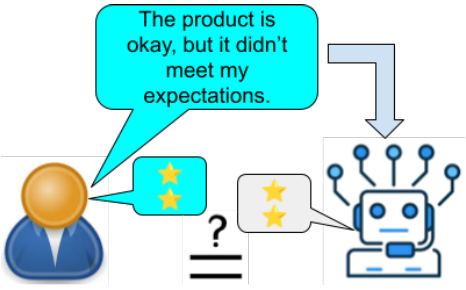


Figure 1. Rating prediction task.

The advent of *Large Language Models* (LLMs) has introduced new possibilities in text processing tasks, particularly through *prompt engineering*, which is an alternative to traditional parameter adjustments of the model. This technique enables models to be guided by textual instructions, allowing them to handle a wide range of tasks without requiring additional training. Despite significant advancements in LLM applications, most existing studies focus on the English language, making text processing in Portuguese more challenging due to the lack of linguistic resources available for this language [Pereira, 2021]. In this study, we propose an approach to predict user reviews ratings by utilizing a diverse set of LLMs capable of processing Portuguese text.

Our objective is to evaluate and compare the performance of these models while considering only the textual content of the reviews. To achieve this, we selected a diverse set of LLMs capable of processing Portuguese text. The chosen models include ChatGPT in versions 3.5 and 4o, DeepSeek, Mistral, LLaMA 3 and 3.3, and Gemma 1 and 2, which are general-purpose multilingual models, as well as a model specifically trained on Portuguese data, available in two versions: Sabiá and Sabiazinho.

We chose to evaluate the LLMs using a *zero-shot* approach because it allows us to assess their intrinsic capability to generalize and understand the task without relying on task-specific fine-tuning or additional training data. This approach is particularly relevant in scenarios where labeled data for rating prediction in Portuguese are scarce or where real-world applications require models to adapt to unseen texts dynamically. Additionally, the zero-shot setting enables us to analyze how well different LLMs leverage their pre-trained knowledge to infer ratings from textual reviews, thereby demonstrating their effectiveness across various domains and linguistic structures.

To the best of our knowledge, no prior research has focused on rating prediction for texts in Brazilian Portuguese. Given the motivations and implications of using LLMs for rating prediction, particularly in low-resource languages such as Portuguese, this research aims to address the following research questions (RQs):

- **RQ1**: Which LLM demonstrates the best performance in the rating prediction task?
- **RQ2**: Is there a specific rating category (stars) that is particularly challenging for the models to predict accurately?
- **RQ3**: Are there certain product types for which rating prediction is more difficult due to the nature of consumer reviews?
- **RQ4**: How does the performance of different prompts vary in the rating prediction task?
- **RQ5**: What are the computational costs associated with using these models?

The remainder of this paper is structured as follows. Section 2 presents related work on methods applicable to the *rating prediction* problem and the large language models commonly employed in Portuguese text analysis. Next, Section 3 details the research methodology and the proposed approach. The experimental results are discussed in Section 4. Finally, the limitations of this study are presented in 5, while the conclusions are presented in 6.

2 Related Work

2.1 Rating Prediction

Despite its wide recognition in the literature, rating prediction has been sparsely studied, with many different approaches applied to very language-specific datasets. These approaches often use different representations of knowledge, such as topics, extracted sentiments, and linguistic or semantic information to overcome data scarcity, as noted by Chambua and Niu [2021].

In Asghar [2016], the author apply a multiclass classification strategy to deal with the rating prediction problem in a Yelp review dataset in the English language. They experimented with n-gram representations and Latent Semantic Indexing to find topics in each review, reaching an accuracy of 64% with bigrams and trigrams combined with a Logistic Regression classifier.

Similarly, Hanić *et al.* [2024] explore a multiclass approach in a TripAdvisor restaurant review dataset, fully written in English. Their best results were achieved by combining the classic TF-IDF frequency-based text representation with a BERT model, showcasing the importance of using deep learning techniques for the task.

Beyond English-language studies, efforts have also been made to explore rating prediction in other languages, demonstrating the necessity of language-specific adaptations. Hossain *et al.* [2021] conducted a study on product reviews written in Bangla language to predict numerical ratings from text. They applied machine learning classification models, and the results showed that SVM achieved the highest accuracy at 90%, highlighting the importance of building rating prediction models to specific linguistic and cultural contexts.

2.2 Large Language Models

Large language models have significantly advanced natural language processing, particularly in text comprehension, sentiment analysis, and opinion mining. Their ability to generalize tasks through zero-shot and few-shot learning has enabled applications in several areas, including rating prediction [Kang *et al.*, 2023] and recommendation systems [Zhang *et al.*, 2024].

Sentiment analysis, a core NLP task, has seen the growing application of LLMs and enhanced contextual understanding and adaptability. Liu *et al.* [2025] demonstrated the effectiveness of LLMs in recommendation systems by leveraging their sentiment analysis capabilities to improve user preference modeling. Their study introduced a chain-based prompting strategy to extract semantic aspects-aware interactions, improving both interpretability and accuracy.

Kang *et al.* [2023] conducted an extensive evaluation of LLMs for user rating prediction, comparing their performance against traditional collaborative filtering (CF) models. They found that zero-shot LLMs, while capable of making reasonable predictions, underperformed compared to CF models that utilize user interaction data. However, fine-tuned LLMs achieved comparable or superior performance with only a small fraction of training data, demonstrating their potential for efficient user opinion understanding.

Despite the growing adoption of LLMs, research on their performance in Brazilian Portuguese is still limited. de Araujo *et al.* [2024] evaluated GPT-3.5 for zero-shot sentiment analysis in Brazilian Portuguese, analyzing its performance in identifying opinionated sentences, polarity classification, and comparative sentence detection. Their results showed that while GPT-3.5 performed well in polarity classification, it struggled with comparative and subjective sentiment identification, highlighting the need for further adjustments for this language.

Given the syntactic complexity and unique linguistic characteristics of Brazilian Portuguese, current LLMs require adaptation to perform effectively in this language. Although models like GPT-3.5 and GPT-4o offer strong multilingual capabilities, they are not specifically optimized for Brazilian Portuguese. Open-source alternatives such as DeepSeek [DeepSeek-AI, 2024], Gemma [Team *et al.*, 2024a,b], Llama [Roziere *et al.*, 2023; Dubey *et al.*, 2024], and Mistral [Jiang *et al.*, 2023] provide promising multilingual performance, but also lack the domain-specific training necessary for optimal results in Portuguese sentiment analysis and related tasks. A notable alternative is Sabiá-3 [Abonizio *et al.*, 2024], a Brazilian LLM designed specifically for Portuguese-language tasks and developed with a corpus emphasizing Brazilian culture, history, and linguistic nuances.

Our study aims to systematically assess the performance of these LLMs for the review rating prediction task in Brazilian Portuguese, comparing the effectiveness of general purpose multilingual models with Brazilian Portuguese-specific ones.

3 Materials and Methods

3.1 Data Collection

To build our dataset, we collected 344,723 user reviews from Amazon[1], covering 10 different product categories, and submitted between 2021 and 2024. The selected categories were chosen based on their popularity and include automotive, baby products, cellphones, food, games, laptops, books, fashion, pet supplies, and toys.

The dataset exhibits significant imbalances both in the distribution of ratings and across product categories. For example, 76.8% of the reviews received a 5-star rating from the customers, while the laptop category accounts for only 1.2% of the total reviews. To mitigate the issue of imbalance, which is not the focus of this work, we applied an undersampling technique. Specifically, we randomly selected 200 reviews for each category, ensuring an equal distribution across all rating levels.

3.2 Metrics

In this work, models were required to produce a value corresponding to the number of stars in reviews, ranging from 1 to 5. We approached this problem both as a classification task, where each rating is treated as a discrete class, and as a regression task, where ratings are considered as continuous numerical values. This dual approach allows us to compare how well different models capture the underlying rating distribution.

Thus, we used a comprehensive set of widely adopted metrics to assess the performance of classification models. Specifically, we used precision, recall, and the F1 score, which are commonly applied in text classification tasks. Furthermore, we used the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) to evaluate model performance, considering ratings greater than or equal to 4 as positive and the remaining ratings as negative. In this context, the AUC measures the model's ability to rank positive ratings higher than negative ones [Kang *et al.*, 2023].

Finally, we adopted the popular Root Mean Squared Error (RMSE) metric to quantify errors between predicted ratings, as adopted by Ahmed and Ghabayen [2022], which is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

where $\hat{y}_i$ represents the predicted rating, y_i the actual rating, and n the total number of samples.

We also evaluated the ability of the models to provide responses compatible with the format required in the prompt. For this, we introduced a metric here referred to as the Non-Response Rate (NR Rate). Responses that were outside the five predefined rating options or failed to provide a valid rating were considered invalid. Therefore, a high NR Rate indicates a lower ability of the model to generate responses in the expected format as specified in the prompt. For example, Figure 2 illustrates when a response is considered valid (left side) and when it is considered invalid (right side). Even if the model correctly identifies that the comment is positive, a response like "Great product!" would still be considered invalid.

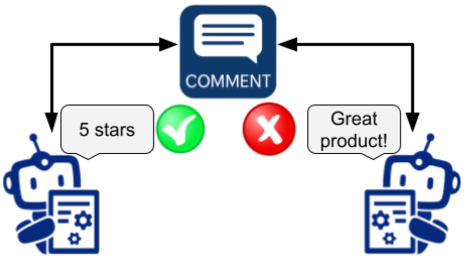


Figure 2. Response valid and invalid.

3.3 Large Language Models

In this study, we considered a diverse set of large language models (LLMs) capable of processing text in Portuguese and handling the task of rating prediction. To this end, we employed the ChatGPT 3.5 and ChatGPT 4o models from OpenAI, the Chinese-origin DeepSeek model, the Mistral model, the LLama models from Meta in versions 3 and 3.3, the Gemma models from Google in versions 1 and 2, and finally, the Sabiá and Sabiazinho models, which are uniquely designed and trained with a specific focus on Brazilian Portuguese. A brief description of each model is provided below:

- **ChatGPT 3.5**: Developed by OpenAI, ChatGPT 3.5 is a model optimized for natural language generation and specifically enhanced for conversational tasks. The model strikes a balance between performance and cost, offering a more efficient alternative to earlier versions. With robust capabilities across various NLP tasks, ChatGPT 3.5 serves as a reliable baseline for many applications. For our experiments, we utilized the ChatGPT-

[1] https://www.amazon.com.br

3.5 Turbo variant, which is known for its improved efficiency [Ye *et al.*, 2023].

- **ChatGPT 4o**: Built on the foundation of ChatGPT 4, ChatGPT 4o extends its capabilities in advanced reasoning across multimodal inputs, including text, audio, images, and video, while generating outputs in text, audio, and images. The model demonstrates exceptional reasoning capabilities, enabling it to analyze and synthesize complex information from diverse data types. It also shows improved performance in non-English languages, such as Portuguese, further highlighting its reasoning proficiency for multilingual contexts. Additionally, ChatGPT 4o introduces robust safety features, ensuring secure and reliable interactions [Al Nazi *et al.*, 2025].
- **DeepSeek**: Developed by independent researchers, DeepSeek is an open-source language model based on Transformer architecture, designed with a focus on scalability and versatility. The model was trained on a large-scale bilingual dataset containing 2 trillion tokens, primarily in Chinese and English. With advanced capabilities in reasoning, mathematics, and code generation. Additionally, the model incorporates techniques such as supervised fine-tuning (SFT) and direct preference optimization (DPO), enhancing its performance in dialogue tasks and alignment with user intent [DeepSeek-AI, 2024]. For our experiments, we used the V3 version of DeepSeek accessed via its API. The API exclusively supported the V3 version, and as a result, it was not possible to evaluate other versions of the model.
- **Gemma 1**: Gemma 1 is an open-language model family developed by Google DeepMind, based on the Gemini architecture. Trained on a large-scale dataset, Gemma 1 outperforms equivalent open models in various benchmarks for text comprehension and generation. Although the model includes fine-tuned versions for dialogue and safety, its training primarily focused on English-language data, which may affect its performance in other languages, including Portuguese [Team *et al.*, 2024a].
- **Gemma 2**: Gemma 2 is the second generation of the Gemma model family, developed by Google DeepMind. Designed for efficiency and high performance, the model is optimized for execution on GPUs and TPUs, making it accessible for a wide range of applications. Additionally, the model incorporates improvements in alignment and bias mitigation to enhance safety and reliability. Gemma 2 is a multilingual model capable of processing text in multiple languages, including Portuguese, although its performance may vary depending on the availability of training data for each language. According to its official documentation [Team *et al.*, 2024b], Gemma 2 demonstrates competitive performance in benchmarks for text comprehension and generation, surpassing equivalent models in various tasks.
- **LLaMA 3 8B**: Developed by Meta, LLaMA 3 8B is a foundational model renowned for its strong performance in coding, multilingual tasks, and reasoning. Capable of handling various modalities, including image, video, and speech processing, LLaMA 3 8B delivers competitive results across these domains. However, it remains a work in progress with limited public release. For our evaluation, we used the LLaMA 3 model with 8 billion parameters [Dubey *et al.*, 2024].
- **LLaMA 3.3 70B**: An advanced iteration of the LLaMA family, specifically fine-tuned to excel in both natural language processing and programming tasks. The model incorporates long-context capabilities, allowing it to handle sequences of up to 100,000 tokens, which is essential for complex tasks requiring extensive input data. Trained on an extensive dataset of 1 trillion tokens, LLaMA 3.3 70B demonstrates exceptional performance in multilingual settings and code-related benchmarks, surpassing its predecessor, LLaMA 3. Additionally, it supports infilling and instruction-tuning objectives, enabling more precise completion tasks and better alignment with user instructions. These features make it a versatile tool for a wide range of applications, including code generation and reasoning [Roziere *et al.*, 2023].
- **Mistral 7B**: Developed by Mistral AI, Mistral 7B is a 7-billion-parameter language model designed for high performance and efficiency. We used the fine-tuned version, Mistral 7B-Instruct, which demonstrates superior performance on instruction-following tasks, surpassing comparable models in both human and automated evaluations. The model was accessed via Hugging Face, where only the 7B version was available for evaluation. Mistral 7B is open-source and released under the Apache 2.0 license, making it a flexible and accessible tool for various applications [Jiang *et al.*, 2023].
- **Sabiá-3**: Developed by Maritaca AI, Sabiá-3 is a state-of-the-art language model specifically designed for Brazilian Portuguese. It was trained on a vast, high-quality corpus of Portuguese texts, with a strong focus on Brazil-centric resources, including cultural, historical, and academic content. This specialization enables the model to grasp the nuanced linguistic features, societal norms, and regional variations unique to the Brazilian context, making it highly effective for natural language processing tasks that demand accurate text interpretation. Sabiá-3 is accessible via an API named MariTalk, following a structure similar to OpenAI's deployment standards. The API ensures efficient interaction with the model while maintaining cost-effectiveness, as Sabiá-3 operates at three to four times lower cost per token compared to frontier models like GPT-4o, without compromising on performance in Brazil-specific tasks [Abonizio *et al.*, 2024].
- **Sabiazinho**: Sabiazinho is a more cost-effective model compared to Sabiá 3, designed for simpler tasks. The developers recommend this model for applications that prioritize speed and cost.

Table 1 presents a summary of the large language models investigated in this work.

Table 1. Main characteristics of the LLMs used in the study

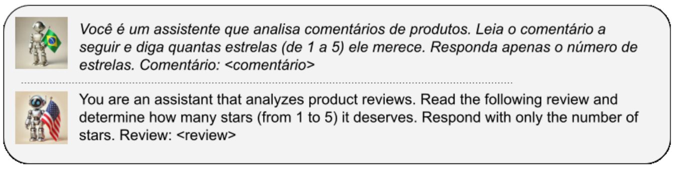
Model	Developer	Language Specialization	Pricing	Model Size	Notes
ChatGPT 3.5	OpenAI	Multilingual	Paid	Proprietary	Widely used for text-based tasks.
ChatGPT 4	OpenAI	Multilingual	Paid	Proprietary	Improved reasoning over ChatGPT 3.5.
DeepSeek	Chinese origin	Multilingual	Free	11B	Focused on precise text generation.
Mistral	Independent	Multilingual	Free	7B	Lightweight and optimized for efficiency.
LLama 3	Meta	Multilingual	Free	8B	High scalability for research purposes.
LLama 3.3	Meta	Multilingual	Free	70B	Enhanced architecture over LLama 3.
Gemma 1	Google	Multilingual	Paid	2B	Early version with wide domain coverage.
Gemma 2	Google	Multilingual	Paid	2B	Improved accuracy over Gemma 1.
Sabiá	Independent	Brazilian Portuguese	Free	13B	Trained specifically for Brazilian Portuguese.
Sabiazinho	Independent	Brazilian Portuguese	Free	7B	A lightweight version of Sabiá.

3.4 Prompts

The use of prompts in language model tasks has proven to be an essential approach to exploring the capabilities of large language models (LLMs). A well-structured prompt is crucial to clearly define the task's purpose, specific instructions, and the expected response format. In this study, we chose to investigate exclusively the zero-shot strategy, which involves task execution without explicit examples provided during the inference phase.

The choice of the zero-shot approach is justified by two main factors: simplicity and generalization. First, this strategy eliminates the need for additional examples, reducing the complexity and preparation time of the prompts. Second, zero-shot enables us to evaluate the intrinsic capability of LLMs to understand and follow the provided instructions without relying on specific adjustments for previously seen scenarios or data. Thus, this approach is ideal for testing the effectiveness of models in broad and under-defined scenarios.

In this work, we evaluated two types of prompts: a more straightforward one (Prompt 1) and another with more details (Prompt 2). The simpler prompt, illustrated in Figure 3, is structured as follows. Initially, we provide the purpose and context of the task, determining that the model must act as an assistant to evaluate product reviews. Then, we establish specific instructions, indicating that the model should determine the number of stars, ranging from 1 to 5. Finally, we clearly specify how the response should be formatted, requiring it to be just the number of stars.

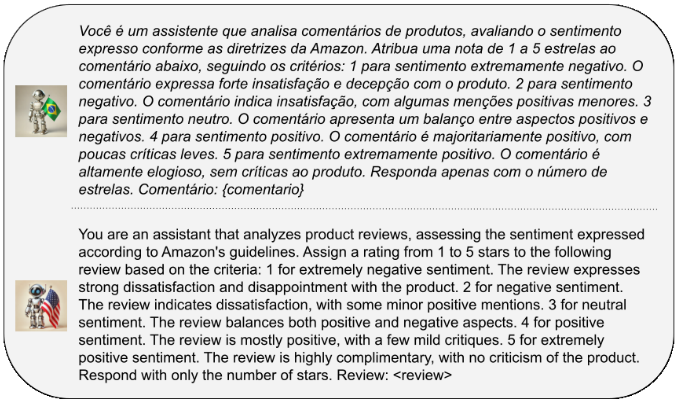

Você é um assistente que analisa comentários de produtos. Leia o comentário a seguir e diga quantas estrelas (de 1 a 5) ele merece. Responda apenas o número de estrelas. Comentário: <comentário>

You are an assistant that analyzes product reviews. Read the following review and determine how many stars (from 1 to 5) it deserves. Respond with only the number of stars. Review: <review>

Figure 3. Structure of Prompt 1.

In the more detailed prompt, represented in Figure 4, we add an explanation for each of the five rating criteria, ranging from 1 to 5 stars. These criteria were based directly on descriptions provided on Amazon's website, which define the meaning of each star level in terms of customer satisfaction and product quality expectations. This strategy aims to provide the model with a clearer understanding of the semantic boundaries between different ratings, particularly for intermediate scores where subjectivity may lead to ambiguity in classification. This approach follows common practices in review rating prediction tasks, where the rating definitions adopted by the platform itself are used to guide both model training and evaluation [Hanić *et al.*, 2024; Asghar, 2016]. Such descriptions reflect the way users perceive and apply the rating system in real-world contexts, even though they are not based on formal psychometric constructs. Therefore, while the detailed prompt does not rely on a theoretical sentiment scale, it remains aligned with practical user expectations and the operational semantics of e-commerce rating systems.

Additionally, we emphasized that the response must be strictly limited to the number of stars, ensuring consistency in the output format. This constraint is crucial to avoid outputs with explanatory text or alternative formats, which could compromise the automation pipeline for rating prediction. The combination of well-defined rating criteria and strict response formatting seeks to maximize both the accuracy and usability of the model outputs in practical applications.


Você é um assistente que analisa comentários de produtos, avaliando o sentimento expresso conforme as diretrizes da Amazon. Atribua uma nota de 1 a 5 estrelas ao comentário abaixo, seguindo os critérios: 1 para sentimento extremamente negativo. O comentário expressa forte insatisfação e decepção com o produto. 2 para sentimento negativo. O comentário indica insatisfação, com algumas menções positivas menores. 3 para sentimento neutro. O comentário apresenta um balanço entre aspectos positivos e negativos. 4 para sentimento positivo. O comentário é majoritariamente positivo, com poucas críticas leves. 5 para sentimento extremamente positivo. O comentário é altamente elogioso, sem críticas ao produto. Responda apenas com o número de estrelas. Comentário: {comentario}

You are an assistant that analyzes product reviews, assessing the sentiment expressed according to Amazon's guidelines. Assign a rating from 1 to 5 stars to the following review based on the criteria: 1 for extremely negative sentiment. The review expresses strong dissatisfaction and disappointment with the product. 2 for negative sentiment. The review indicates dissatisfaction, with some minor positive mentions. 3 for neutral sentiment. The review balances both positive and negative aspects. 4 for positive sentiment. The review is mostly positive, with a few mild critiques. 5 for extremely positive sentiment. The review is highly complimentary, with no criticism of the product. Respond with only the number of stars. Review: <review>

Figure 4. Structure of Prompt 2.

4 Results and Discussion

4.1 Performance of LLMs

To determine which LLM performs best on the rating prediction task (RQ1), we conducted experiments on several large language models (LLMs) using two different prompt formats ($p1$ and $p2$) and, where available, different versions

Table 2. Results of large language models for the rating prediction task (three decimals with bold highlights)

Model	Precision	Recall	F1 Score	AUC	RMSE	NR Rate
ChatGPT 3.5$_{(p1)}$	0.506 ± 0.002	0.487 ± 0.002	0.497 ± 0.002	0.914 ± 0.001	1.179 ± 0.019	3.7%
ChatGPT 3.5$_{(p2)}$	0.482 ± 0.002	0.457 ± 0.003	0.469 ± 0.002	0.909 ± 0.001	1.203 ± 0.018	5.3%
ChatGPT 4$_{(p1)}$	**0.535** ± 0.002	0.527 ± 0.003	0.531 ± 0.002	**0.922** ± 0.001	0.985 ± 0.015	1.4%
ChatGPT 4$_{(p2)}$	0.498 ± 0.004	0.496 ± 0.003	0.497 ± 0.003	0.920 ± 0.001	0.975 ± 0.005	0.4%
DeepSeek$_{(p1)}$	0.532 ± 0.002	**0.532** ± 0.002	**0.532** ± 0.002	0.912 ± 0.001	0.932 ± 0.004	**0.0%**
DeepSeek$_{(p2)}$	0.502 ± 0.002	0.502 ± 0.003	0.502 ± 0.002	0.914 ± 0.001	0.962 ± 0.005	**0.0%**
Mistral$_{(p1)}$	0.281 ± 0.030	0.236 ± 0.003	0.257 ± 0.002	0.759 ± 0.004	1.884 ± 0.034	15.8%
Mistral$_{(p2)}$	0.344 ± 0.005	0.066 ± 0.001	0.110 ± 0.002	0.806 ± 0.003	3.009 ± 0.006	81.0%
LLama 3 8B$_{(p1)}$	0.284 ± 0.001	0.273 ± 0.000	0.278 ± 0.000	0.683 ± 0.004	1.415 ± 0.019	4.1%
LLama 3 8B$_{(p2)}$	0.448 ± 0.003	0.328 ± 0.003	0.379 ± 0.003	0.879 ± 0.000	1.992 ± 0.004	26.7%
LLama 3.3 70B$_{(p1)}$	0.401 ± 0.004	0.395 ± 0.004	0.398 ± 0.004	0.836 ± 0.003	1.180 ± 0.011	1.7%
LLama 3.3 70B$_{(p2)}$	0.224 ± 0.004	0.214 ± 0.004	0.219 ± 0.004	0.481 ± 0.002	2.411 ± 0.014	4.5%
Gemma 1$_{(p1)}$	0.330 ± 0.001	0.316 ± 0.001	0.323 ± 0.001	0.734 ± 0.001	1.620 ± 0.002	4.4%
Gemma 1$_{(p2)}$	0.308 ± 0.001	0.308 ± 0.001	0.308 ± 0.001	0.672 ± 0.001	1.449 ± 0.003	**0.0%**
Gemma 2$_{(p1)}$	0.506 ± 0.000	0.505 ± 0.000	0.506 ± 0.000	0.897 ± 0.000	1.026 ± 0.000	0.2%
Gemma 2$_{(p2)}$	0.437 ± 0.001	0.437 ± 0.001	0.437 ± 0.001	0.877 ± 0.001	1.013 ± 0.001	**0.0%**
Sabiá-3$_{(p1)}$	0.508 ± 0.005	0.507 ± 0.005	0.507 ± 0.006	0.912 ± 0.001	**0.899** ± 0.007	0.1%
Sabiá-3$_{(p2)}$	0.470 ± 0.005	0.469 ± 0.005	0.470 ± 0.005	0.908 ± 0.001	0.900 ± 0.004	**0.0%**
Sabiazinho$_{(p1)}$	0.460 ± 0.006	0.460 ± 0.006	0.460 ± 0.006	0.906 ± 0.001	0.999 ± 0.008	0.1%
Sabiazinho$_{(p2)}$	0.487 ± 0.008	0.486 ± 0.008	0.487 ± 0.008	0.909 ± 0.002	1.002 ± 0.009	0.2%

of the models. The summary of results is shown in Table 2. ChatGPT 4$_{(p1)}$ and DeepSeek$_{(p1)}$ emerged as the most robust models for the task of rating prediction in Brazilian Portuguese. Both demonstrated consistent performance in key evaluation metrics, showing their ability to handle the linguistic complexity of Portuguese and to produce correctly formatted responses. Although DeepSeek$_{(p1)}$ achieved a slightly higher F1 score, a statistical analysis revealed that the difference was not significant ($p < 0.05$), suggesting that the two models can be considered equivalent in terms of overall reliability.

Beyond the F1 Score, an examination of the AUC, RMSE, and NR Rate metrics provides deeper understanding into the strengths of each model. ChatGPT 4$_{(p1)}$ achieved a slightly higher AUC compared to DeepSeek$_{(p1)}$, indicating a superior ability to accurately distinguish between positive and negative ratings. This suggests that ChatGPT 4$_{(p1)}$ is more effective in prioritizing correct classifications in binary contexts. However, DeepSeek$_{(p1)}$ demonstrated a lower RMSE, highlighting greater accuracy in predicting ratings as numerical values. This result reflects its ability to produce predictions closer to the actual values, which is particularly critical in tasks involving continuous variables.

Furthermore, DeepSeek$_{(p1)}$ achieved an NR Rate of 0%, indicating the complete absence of invalid responses, whereas ChatGPT 4$_{(p1)}$ exhibited a marginal NR Rate of 1.4%. This indicates that DeepSeek$_{(p1)}$ was highly robust in adhering to the required response format, while ChatGPT 4$_{(p1)}$, despite being very close, showed slight inconsistencies.

Based on these results, the two models exhibit complementary strengths. ChatGPT 4$_{(p1)}$ is slightly superior in scenarios where accurate class separation is critical, while DeepSeek$_{(p1)}$ excels in numerical accuracy and response format consistency, which is an essential aspect of the target task of rating prediction. Among the other models evaluated, it is worth mentioning that Sabiá-3$_{(p1)}$ and Gemma 2$_{(p1)}$ presented a slightly lower performance, both with F1 close to 50% and a low NR Rate. Mistral$_{(p2)}$ showed the poorest performance, with an extremely low F1 Score, reduced AUC, and the highest NR Rate (81%). These results suggest severe limitations in understanding the task and in generating responses in the required format. Another underperforming model was LLama 3.3 70B$_{(p2)}$, which exhibited low F1 Scores and AUC values along with a high RMSE, reflecting significant difficulties in accurately predicting ratings.

The analysis of these underperforming models shows that certain architectures struggle to handle the complexities of the rating prediction task, particularly when addressing the linguistic and cultural nuances of Brazilian Portuguese. This is particularly clear for Sabiazinho, which is a customized model for Portuguese, with a competitive performance even being small. This reinforces the importance of employing robust, well-trained models tailored to the specific requirements of such tasks. The analysis of the prompt's importance in the models' results will be discussed in Section 4.4.

4.2 Performance Analysis by Rating Category

To determine whether there is a rating category that is more challenging (RQ2), we evaluate the performance of the models with highest F1 score, ChatGPT 4$_{(p1)}$ and DeepSeek$_{(p1)}$, in the task of rating prediction, analyzing each of the five rating categories (from 1 to 5 stars) individually. To achieve this, we utilize confusion matrices, which present the distribution of the models' predictions in relation to the actual classifications. Each value displayed in Figure 5 corresponds to the accumulated values of the five executions performed for each model, allowing a more robust assessment of the consistency of the results.

We can note that the 5-star category exhibited the highest accuracy rates for both models, indicating their strong ability to recognize highly positive texts. This result suggests that when a user review expresses clear satisfaction, the models tend to classify it correctly. Similarly, the 1-star category also demonstrated a high number of correct predictions, likely due to the fact that extremely negative reviews often contain distinct linguistic patterns that are more easily recognizable by the models.

The 2-star, 3-star, and 4-star categories exhibited the highest misclassification rates, with frequent swaps between these ratings. This suggests that texts with intermediate evaluations pose a greater challenge for the models, possibly because they convey more ambiguous opinions, making accurate classification difficult. Although we did not perform an explicit sentiment–polarity experiment in this study, the pattern in Figure 5 is consistent with observations in prior work on review mining: texts with the highest rating (5 stars) tend to be short, enthusiastic and lexically homogeneous ("*Excelente*", "*Perfeito*", "*Recomendo muito*"), which provides strong cues for the models. In contrast, 1-star reviews are not simply the negative mirror of 5-star ones. Negative feedback is usually more heterogeneous: it can include detailed narratives, indirect complaints, sarcasm or irony, all of which increase linguistic variability and make prediction harder [Hanić *et al.*, 2024; Feng and Yan, 2024]. This asymmetry between positive and negative extremes helps explain why 5-star ratings are easier to learn, while 1-star ratings does not achieve the same margin of accuracy. Incorporating a lightweight sentiment-analysis layer or chain-of-thought prompting to disambiguate ironic and mixed-tone negatives is therefore an interesting avenue for future work.

An illustrative error that highlights the linguistic challenges faced by the models occurred with the following review from the dataset. The review reads: "*O produto é bom, funciona como descrito, mas o preço é completamente absurdo. Me sinto enganado por pagar tão caro.*" ("The product is good, it works as described, but the price is absolutely outrageous. I feel cheated for having paid so much."). In this case, ChatGPT-4o (prompt $p1$) incorrectly predicted a rating of 4, while the correct label was 2. This review exhibits a contrastive sentiment structure. The functional assessment of the product is clearly positive ("*funciona como descrito*"), whereas the emotional and evaluative component related to pricing is strongly negative ("*preço é completamente absurdo*", "*me sinto enganado*"). The coexistence of both positive and negative cues within the same text creates ambiguity for the model, making it difficult to accurately determine the overall sentiment and assign the correct rating. This type of misclassification illustrates how LLMs may overweight objective product descriptions while underestimating subjective dissatisfaction signals, especially when those signals are tied to external factors like price rather than product features themselves.

When comparing the performance of the models, we observe that DeepSeek$_{(p1)}$ demonstrated greater consistency in predictions, with a total of 5,319 correct classifications, compared to 5,277 for ChatGPT-4$_{(p1)}$. More specifically, DeepSeek$_{(p1)}$ exhibited a broader distribution of errors, particularly in the 1-star and 2-star categories, suggesting that the model struggles more to capture the nuances of neutral or slightly negative reviews.

4.3 Analysis by Product Category

In this section, we analyze the performance of the ChatGPT 4$_{(p1)}$ and DeepSeek$_{(p1)}$ models in the task of rating prediction applied to different product categories (RQ3). The objective of this experiment is to evaluate how each model performs across distinct semantic domains, identifying patterns of accuracy and difficulty.

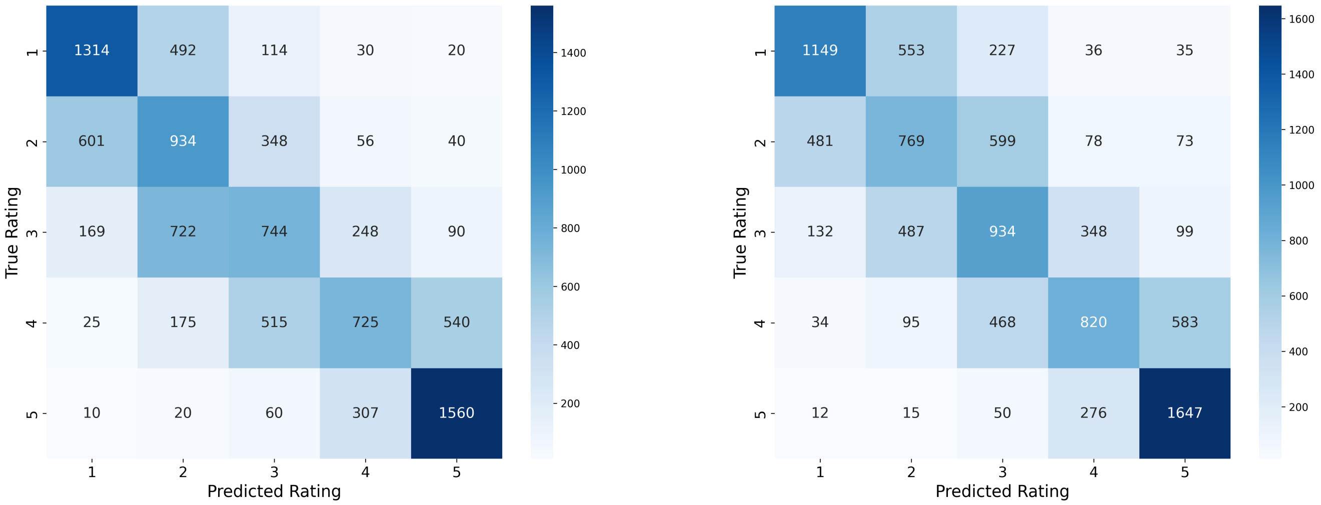
For this analysis, the models were evaluated on a dataset of ten product categories described in Section 3.1. Each model was executed five times to ensure robustness, and the metric used was the F1 Score, which balances precision and recall, providing a detailed view of the models' ability to correctly classify ratings. The analysis was conducted using boxplots, which allow visualization of the variation in F1 Score across runs for each product category. Figures 6a and 6b present individual results for the ChatGPT 4$_{(p1)}$ and DeepSeek$_{(p1)}$ models, respectively, allowing a detailed comparison of their performance.

To further investigate the reasons behind this difficulty, we conducted an extensive statistical analysis comparing key linguistic and structural characteristics of the comments across categories. We analyzed the Mean Number of Words per Comment (MNWC), the Mean Number of Sentences per Comment (MNSC), the Flesch Reading Ease (FRE), which reflects how readable a piece of content is, and the Type-Token Ratio (TTR), which indicates lexical diversity in the comment. Table 3 summarizes the results of these metrics for each category.

Table 3. Linguistic and structural analysis by category.

Category	MNWC	MNSC	FRE	TTR
Automotive	16.75	1.76	25.09	0.91
Baby	15.95	1.73	26.15	0.93
Books	28.11	2.25	22.93	0.87
Cell Phones	22.26	1.90	19.00	0.90
Fashion	15.73	1.84	20.68	0.93
Food	15.75	1.74	27.52	0.92
Games	22.6	1.79	23.70	0.91
Laptops	35.16	2.57	19.57	0.83
Pets	18.98	1.87	29.15	0.91
Toys	18.95	1.86	21.50	0.90

While these metrics provided understanding into textual complexity, they did not present a clear correlation with prediction difficulty for different categories. All results vary near the average F1 scores for both models. Thus, to suggest what could impact on category differences, we conducted a qualitative analysis focusing on the Games category. This analysis suggests that game-related comments often contain high linguistic complexity, incorporating gamer slang, subjective expressions, and specialized terms. For instance, the comment "*O balanceamento do multiplayer está horrível, mas a campanha é incrível.*" ("The multiplayer balancing is horrible, but the campaign is amazing.") demonstrates a



(a) ChatGPT $4_{(p1)}$ **(b)** $\text{DeepSeek}_{(p1)}$

Figure 5. Confusion Matrices Comparison Between ChatGPT and DeepSeek Models.

mix of contrasting sentiments, making rating prediction more challenging.

Another factor contributing to prediction difficulty is the discrepancy between user expectations and product reality. Unlike physical products such as laptops or fashion, where evaluations are based on tangible attributes (e.g., delivery time, build quality), game reviews often reflect subjective user experiences and expectations.

In conclusion, this analysis demonstrates that rating prediction difficulty varies across product categories. Although $\text{DeepSeek}_{(p1)}$ showed a marginally better overall performance, both models struggle in categories where comments exhibit greater subjectivity and linguistic complexity. These findings highlight the importance of adapting AI models to domain-specific challenges and leveraging specialized metrics to better assess model effectiveness in different application scenarios.

4.4 Evaluation of Prompt Effectiveness

In this section, we analyze the impact of prompt type on model performance (RQ4). To achieve this, we selected the variation in the F1 Score as the primary metric, as it provides a comprehensive assessment of predictive effectiveness.

Figure 7 presents the variation in the F1 Score for each model, calculated using the following formula:

$$\text{Variation \%} = \left(\frac{\text{F1 Score (P2)} - \text{F1 Score (P1)}}{\text{F1 Score (P1)}} \right) \times 100$$

Negative values in the graph indicate that prompt $p1$ yielded better performance than prompt $p2$, which was observed in models such as ChatGPT and DeepSeek. Conversely, positive values indicate that prompt $p2$ outperformed prompt $p1$, observed exclusively in LLaMA 3 8B and Sabiazinho.

Regarding the effect of prompt formulation, the results indicate that more robust models, such as ChatGPT 4 and DeepSeek, exhibited minor declines in F1 Score when using prompt $p2$ but showed improved response consistency, as evidenced by a lower NR Rate. In contrast, less robust models, such as Mistral and LLaMA 3.3 70B, experienced substantial decreases in performance with prompt $p2$. This suggests that more detailed and explanatory prompts may introduce confusion for models with lower processing capacity. In general, the prompt $p1$ proved to be more effective across most models, indicating that concise and direct prompts tend to yield better performance in rating prediction tasks.

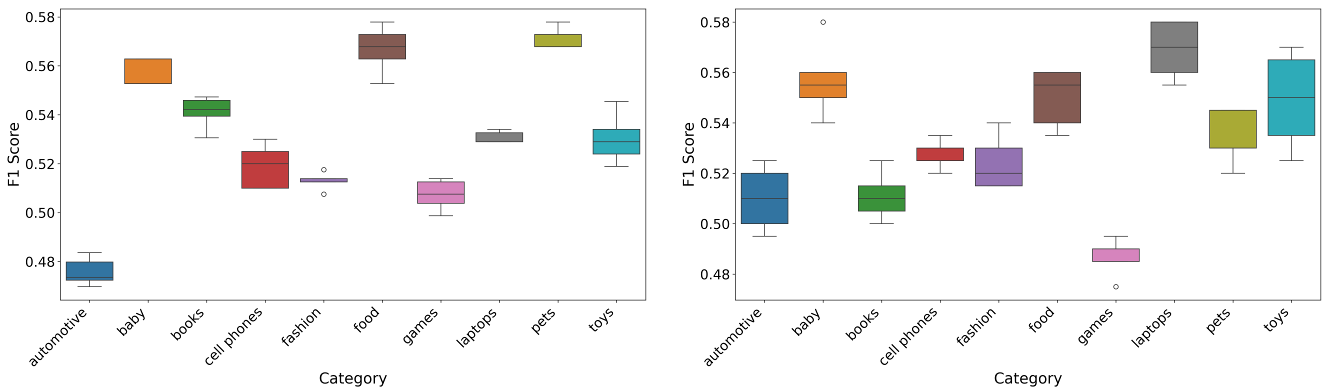
We also evaluated the influence of prompt type on the model's ability to generate valid responses. Figure 8 presents the absolute difference in the invalid response rate (NR Rate) between prompt $p1$ and prompt $p2$ across different language models. Mistral and LLaMA 3 8B exhibited a significant increase in invalid responses when using the more detailed prompt, indicating that prompt $p2$ severely impaired their ability to generate valid outputs. Conversely, more advanced models, such as ChatGPT-4, DeepSeek, and Sabiá-3, remained largely unaffected by the prompt variations.

These findings challenge the common assumption that a more detailed prompt inherently improves response validity in LLMs. Furthermore, the results emphasize the need for empirical evaluation when selecting an appropriate LLM for a specific application. This is consistent with the diverse range of prompt engineering techniques identified by [Sahoo *et al.*, 2024], which highlight the importance of adapting prompts to the specific requirements of the task. Finally, our analysis reinforces that prompt engineering should be *prompta secundum problemata eliguntur*, which means that prompts must be tailored to the problem at hand.

4.5 Evaluation of Model Costs

In this section, we evaluated the computational costs associated with using LLMs (RQ5). For this, we present a cost evaluation of the two best multilingual models, ChatGPT[2] and DeepSeek[3], discussed in the Section 4.1, along with Sabiá[4],

[2] https://openai.com/api/pricing
[3] https://api-docs.deepseek.com/quick_start/pricing
[4] https://www.maritaca.ai



(a) ChatGPT 4$_{(p1)}$

(b) DeepSeek$_{(p1)}$

Figure 6. Comparative results of ChatGPT and DeepSeek models by Product Category.

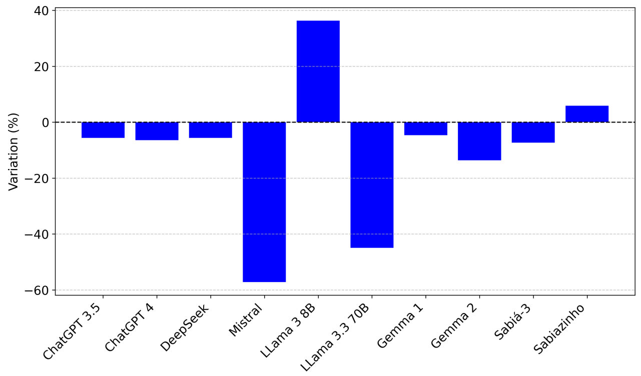


Figure 7. Variation in F1 Score: prompt $p2$ minus prompt $p1$.

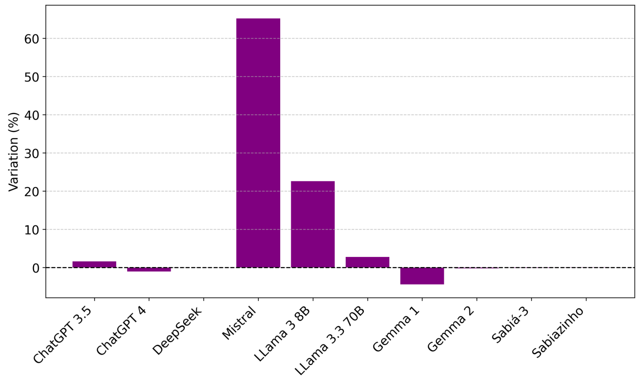


Figure 8. Absolute Change in Invalid Response Rate (NR Rate) (P2 - P1).

a model originally trained in Portuguese. Table 4 provides a summary of the cost values obtained from the official sources of each model[5]. The table presents the costs associated with input and output tokens, with values denominated in US dollars, calculated per 1 million tokens.

The column #Tokens in the table represents the number of tokens processed by each model. A key insight is that different models use distinct tokenization methods, leading to variations in token count even for the same input. As expected, models using prompt 2 ($p2$) generated more tokens than their equivalent versions using prompt 1 ($p1$). Interestingly, simpler models produced a higher token count than more recent models. This suggests that, in some cases, older models might have a higher total cost, despite their lower price per million tokens, due to their higher token usage.

Table 4. Cost comparison of evaluated language models.

Model	**Input (M)**	**Output (M)**	**#Tokens**	**Total ($)**
ChatGPT 3.5$_{(p1)}$	$3.00	$4.00	178,583	1,25
ChatGPT 3.5$_{(p2)}$	$3.00	$4.00	502,583	3,51
ChatGPT 4o$_{(p1)}$	$30.00	$60.00	148,835	13,39
ChatGPT 4o$_{(p2)}$	$30.00	$60.00	394,835	35,53
DeepSeek$_{(p1)}$	$2.00	$4.00	182,966	1,09
DeepSeek$_{(p2)}$	$2.00	$4.00	540,566	3,24
Sabiá-3$_{(p1)}$	$0.83	$1.67	175,575	0,43
Sabiá-3$_{(p2)}$	$0.83	$1.67	489,575	1,22
Sabiazinho$_{(p1)}$	$0.17	$0.50	213,530	0,14
Sabiazinho$_{(p2)}$	$0.17	$0.50	565,530	0,37

Finally, we note that the two models ChatGPT 4o$_{(p1)}$ and DeepSeek$_{(p1)}$ yielded similar results, but with significantly different costs. A notable observation is that the current version of ChatGPT 4o has a significantly higher cost compared to the other models, particularly in contrast to its recent competitor, DeepSeek. This highlights that DeepSeek stands out as the most cost-effective choice for performing rating prediction tasks in Portuguese, making it the preferred option for cost-efficiency.

5 Limitations and Improvements

We acknowledge several limitations that should be addressed in future research. First, due to computational constraints, we did not evaluate the full dataset of 344,723 reviews but relied on a sample of 2,000 comments. Although the dataset covers multiple product categories, it may not fully capture the diversity of linguistic expressions, sentiment nuances, and user expectations present in real-world e-commerce platforms. Future studies could consider expanding the dataset to include a broader range of review sources, such as social media and forums, where user-generated content follows different textual structures.

Second, while our benchmark includes a diverse set of LLMs, it is not feasible to evaluate every available model. The selection of models was based on accessibility and relevance to Brazilian Portuguese, but incorporating additional LLMs in future research could offer a more comprehensive perspective. In addition, exploring domain-specific fine-tuning strategies can improve performance in specialized

[5] Data collected on January 30, 2025.

contexts. It is also important to acknowledge that some of the top-performing models in our benchmark, particularly ChatGPT-4o and DeepSeek, are proprietary and subject to continuous updates that are not publicly documented. This dynamic nature may affect the reproducibility and longevity of our results, as future versions of these models might perform differently from what we observed. In this context, the use of open-source models becomes crucial to ensure stability, replicability, and transparency in academic research. Expanding benchmarks with a stronger emphasis on open models, alongside exploring domain-specific fine-tuning strategies, could contribute to more sustainable and generalizable advancements in rating prediction tasks for Brazilian Portuguese.

Another limitation concerns the evaluation metrics. While we adopted widely used classification and regression metrics such as Precision, Recall, F1, AUC, RMSE, and NR Rate, these metrics but RMSE do not fully capture the ordinal nature of the rating prediction task. In this context, errors like predicting 4 instead of 5 are less critical than predicting 1 instead of 5, a distinction that is not reflected in standard classification metrics. Incorporating ordinal-aware metrics, such as the Quadratic Weighted Kappa (QWK), would provide a more nuanced evaluation of model performance, as it penalizes errors proportionally to their ordinal distance. We consider this an important direction for future work to complement the current analysis with metrics better aligned with the practical implications of classification errors in rating prediction tasks.

Lastly, ethical considerations remain an important aspect of this study. Automated rating prediction raises concerns related to fairness, transparency, and potential biases in model predictions. Ensuring user privacy and data security is also a key challenge when processing large volumes of user-generated content. Future research should incorporate fairness assessments and mitigation strategies to address these ethical concerns, promoting the responsible and trustworthy use of AI in practical applications.

Additionally, future research could explore few-shot learning approaches to assess whether providing a small number of labeled examples enhances model performance, particularly in handling ambiguous or subjective reviews. Another promising direction involves investigating chain-of-thought prompting techniques, which guide models through intermediate reasoning steps. These strategies have shown improvements in complex reasoning tasks [Al Nazi *et al.*, 2025] and could potentially reduce errors related to rating subjectivity and improve classification reliability in Brazilian Portuguese.

6 Conclusions

In this study, we evaluated ten Large Language Models (LLMs), namely ChatGPT-3.5, ChatGPT-4o, DeepSeek, Mistral, LLaMA 3, LLaMA 3.3, Gemma 1, Gemma 2, Sabiá-3, and Sabiazinho, using two prompting techniques: a straightforward prompt ($p1$) and a detailed prompt ($p2$). The evaluation focused on rating prediction in Brazilian Portuguese, considering various product categories and assessing model performance in terms of accuracy, response validity, and computational costs. Among the evaluated models, ChatGPT-4o and DeepSeek demonstrated the best overall performance, particularly in correctly classifying extreme ratings (1 and 5 stars). However, models specifically trained in Brazilian Portuguese, such as Sabiá-3, also showed competitive results, indicating the potential of language-specific models for Portuguese NLP tasks.

The results suggest that the straightforward prompt $p1$ generally yielded better performance across most models, particularly those with lower computational capabilities, as prompt $p2$ occasionally introduced confusion. Cost analysis revealed that DeepSeek provided a cost-effective alternative to ChatGPT-4o while maintaining similar levels of accuracy. Furthermore, analysis by product category showed that the models performed better in objective categories such as food and baby products but struggled with subjective or highly technical categories such as automotive and games.

An important finding of this study is that the simple prompt ($p1$) consistently outperformed the detailed prompt ($p2$) across most models. This result indicates that, for rating prediction tasks in Brazilian Portuguese, concise prompts that focus directly on the target instruction are generally more effective in terms of predictive accuracy. However, it is worth noting that the detailed prompt ($p2$) contributed to slightly lower NR Rates in some models, suggesting better compliance with the expected output format. Therefore, we recommend using simple prompts ($p1$) when the primary goal is to maximize accuracy and model performance, and considering detailed prompts ($p2$) in scenarios where output consistency and adherence to formatting constraints are critical.

Although this research represents one of the first systematic benchmarks of LLMs for rating prediction in Brazilian Portuguese, future work will address the identified limitations by expanding the dataset beyond e-commerce reviews to include social media and forum data, allowing for a more comprehensive linguistic evaluation. Exploring fine-tuning strategies for open-source models like Sabiá-3 and DeepSeek could further enhance their performance, reducing the gap with closed-source models. Furthermore, incorporating adaptive prompt engineering techniques, such as dynamic prompt selection and reinforcement learning from human feedback (RLHF), could mitigate response inconsistencies and hallucination issues. Collaborations with cloud providers or research institutions may help overcome computational limitations, enabling broader evaluations of multilingual LLMs and their cross-lingual generalization capabilities in low-resource NLP tasks.

Declarations

Authors' Contributions

All authors were involved in every stage of the work and contributed equally to its completion.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This research is partially supported by the Amazonas State Research Support Foundation (FAPEAM) through the NeuralBond Project (UNIVERSAL 2023 Proc. 01.02.016301.04300/2023-04) and by Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (grant #408490/2024-1).

Availability of data and materials

The datasets (and/or softwares) generated and/or analyzed during the current study will be made available upon request.

References

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabiá-3 technical report. *arXiv preprint arXiv:2410.12049*. DOI: 10.48550/arXiv.2410.12049.

Ahmed, B. H. and Ghabayen, A. S. (2022). Review rating prediction framework using deep learning. *Journal of Ambient Intelligence and Humanized Computing*, 13(7):3423–3432. DOI: 10.1007/s12652-020-01807-4.

Al Nazi, Z., Hossain, M. R., and Al Mamun, F. (2025). Evaluation of open and closed-source llms for low-resource language with zero-shot, few-shot, and chain-of-thought prompting. *Natural Language Processing Journal*, page 100124. DOI: 10.1016/j.nlp.2024.100124.

Asghar, N. (2016). Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*. DOI: 10.48550/arXiv.1605.05362.

Barman, K. D., Bordoloi, B., Kumar, A., and Halder, A. (2024). Review rating predictions using improved deep learning architecture. In *2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 468–472. IEEE. DOI: 10.1109/cicn63059.2024.10847509.

Chambua, J. and Niu, Z. (2021). Review text based rating prediction approaches: preference knowledge learning, representation and utilization. *Artificial Intelligence Review*, 54:1171–1200. DOI: 10.1007/s10462-020-09873-y.

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*, pages 13–21. Available at:`https://aclanthology.org/2024.propor-1.2/`.

de Melo, T., da Silva, A. S., de Moura, E. S., and Calado, P. (2019). Opinionlink: Leveraging user opinions for product catalog enrichment. *Information Processing & Management*, 56(3):823–843. DOI: 10.1016/j.ipm.2019.01.004.

DeepSeek-AI (2024). Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*. Available at:`https://github.com/deepseek-ai/DeepSeek-LLM`.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., *et al.* (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. DOI: 10.48550/arXiv.2407.21783.

Feng, W. and Yan, J. (2024). Language abstraction in negative online customer reviews: The choice of corporate response strategy and voice. *SAGE Open*, 14(2):21582440241240561. DOI: 10.1177/21582440241240561.

Hanić, S., Bagić Babac, M., Gledec, G., and Horvat, M. (2024). Comparing machine learning models for sentiment analysis and rating prediction of vegan and vegetarian restaurant reviews. *Computers*, 13(10):248. DOI: 10.3390/computers13100248.

Hossain, M. I., Rahman, M., Ahmed, M. T., Rahman, M. S., and Islam, A. T. (2021). Rating prediction of product reviews of bangla language using machine learning algorithms. In *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, pages 1–6. IEEE. DOI: 10.1109/aims52415.2021.9466022.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., *et al.* (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*. DOI: 10.48550/arXiv.2310.06825.

Kang, W.-C., Ni, J., Mehta, N., Sathiamoorthy, M., Hong, L., Chi, E., and Cheng, D. Z. (2023). Do llms understand user preferences? evaluating llms on user rating prediction. *arXiv preprint arXiv:2305.06474*. DOI: 10.48550/arXiv.2305.06474.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer. DOI: 10.1007/978-1-4614-3223-4_13.

Liu, F., Liu, Y., Chen, H., Cheng, Z., Nie, L., and Kankanhalli, M. (2025). Understanding before recommendation: Semantic aspect-aware review exploitation via large language models. *ACM Transactions on Information Systems*, 43(2):1–26. DOI: 10.1145/3704999.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115. DOI: 10.1007/s10462-020-09870-1.

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., *et al.* (2023). Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*. DOI: 10.48550/arXiv.2308.12950.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*. DOI: 10.48550/arxiv.2402.07927.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., *et al.* (2024a). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

DOI: 10.48550/arxiv.2403.08295.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., *et al*. (2024b). Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*. DOI: 10.48550/arXiv.2408.00118.

Wang, Q., Zhang, W., Li, J., Mai, F., and Ma, Z. (2022). Effect of online review sentiment on product sales: The moderating role of review credibility perception. *Computers in Human Behavior*, 133:107272. DOI: 10.1016/j.chb.2022.107272.

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., *et al*. (2023). A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*. DOI: 10.48550/arxiv.2303.10420.

Zhang, X., Li, Y., Wang, J., Sun, B., Ma, W., Sun, P., and Zhang, M. (2024). Large language models as evaluators for recommendation explanations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 33–42. DOI: 10.1145/3640457.3688075.