Rating–Text Mismatch in Brazilian Portuguese Reviews: How Reliable Are Zero-Shot LLMs?

Emmanuelle Marreira and Carlos M. S. Figueiredo and Tiago de Melo
LSI - Intelligent Systems Laboratory
UEA - State University of Amazonas
Manaus-AM
{erm.eng22,cfigueiredo,tmelo}@uea.edu.br

Abstract

This study evaluates the ability of large language models (LLMs) to detect incoherence between the text of product reviews and their assigned rating (1 or 5 stars). Using popular LLMs such as GPT-5, Llama-4 and DeepSeek-3.2, and models optimized for Brazilian Portuguese, Sabiá-3.1 and Bode-3.1, we show that some are capable of detecting incoherence among texts and ratings (F1 > 90%) in a zero-shot protocol. Models also present a high agreement in the predictions, where several prediction rounds led to low variability (Fleiss' κ > 0.95). With the demonstrated incoherence present in all product categories (aprox. 10% of comments), the results suggest that LLMs are very promising to perform this high semantic interpretation task, and they can be used as valuable tools for online monitoring and recommendation systems.

1 Introduction

The popularization of e-commerce platforms has made user reviews a crucial resource for both consumers and companies (Hu et al., 2014). Although numeric ratings, often represented by stars (1 to 5), provide an immediate summary of customer perception, the accompanying text frequently contains nuances that the numeric score alone does not capture—or, in some cases, contradicts (Aljrees et al., 2024). Figure 1 illustrates this scenario, where the textual comment[1] is positive, but the rating is 1 star. Automatically detecting such contradictions is relevant for three main reasons: (i) increases the reliability of publicly displayed satisfaction metrics, (ii) helps merchants identify potential fraud or input errors, and (iii) offers a novel signal for recommender systems and sentiment analysis. Additionally, such inconsistency can increase the cognitive effort of consumers, lead to less accurate purchase decisions, and diminish the utility of the review platform (Mudambi et al., 2014).

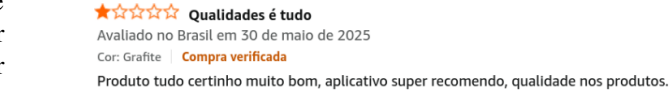

Figure 1: Discrepancy between text and rating.

Historically, verifying the coherence between rating and text has been treated as a subtype of sentiment analysis, requiring manually annotated datasets and task-specific supervised models. The advent of Large Language Models (LLMs) introduced the possibility of assessing inconsistencies in a zero-shot or few-shot regime, eliminating the need for feature engineering and extensive fine-tuning. Despite these advances, the literature lacks systematic studies that quantify (a) the accuracy of these models for this specific task and (b) the variability of their output when multiple independent runs of the same LLM are considered.

This work investigates whether LLMs can consistently identify semantic inconsistencies between rating (1 or 5 stars) and the textual content of reviews. We restricted our study to extreme ratings because they exhibit clearer polarity, which facilitates the analysis of explicit contradictions. Intermediate ratings, on the contrary, are more ambiguous and would complicate the validation. Each review was submitted to a set of LLMs in three independent runs, which allowed us to assess the stability of the models in the face of output variability.

The contributions of this article are as follows: (i) data curation: release of a balanced, annotated set of extreme-star reviews across ten e-commerce categories, in Brazilian Portuguese, containing both the numeric rating and the original text[2]; (ii) zero-shot evaluation protocol: description of a

[1] Comment translated, reflecting the original text: Product all right very good, app super recommend, quality in the products.

[2] The original dataset, along with the annotated data, are available in https://zenodo.org/records/18778352

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 959–967

generic, reproducible prompt to detect rating–text inconsistencies, applicable to any modern LLM; (iii) variability analysis: quantification of within-model agreement across three runs using classical reliability metrics (Fleiss' κ and Cohen's κ), which are still seldom explored in LLM research; and (iv) practical implications: a discussion showing how detected inconsistencies can support automatic content moderation and improve reputation systems in digital retail environments.

The remainder of this paper is organized as follows. Section 2 reviews related work; Section 3 presents the methodology; Section 4 reports the experiments and results; and Section 5 concludes and outlines directions for future work.

2 Related Works

2.1 Comments vs Ratings

Almansour et al. (2022) identify that the correlation between sentiments expressed in text and numerical scores tends to be low to moderate, indicating that relying exclusively on one of these sources may lead to inaccurate conclusions. The origin of the discrepancies between comments and scores was investigated by Geierhos et al. (2015), who point out that one of the factors contributing to this inconsistency is individual random errors. Furthermore, studies such as those by Mellinas et al. (2019) and Sharma et al. (2020) suggest that consumers tend to penalize negative experiences more severely than they reward positive experiences in the texts.

Fazzolari et al. (2017) analyzed the inconsistency in hotel reviews, identifying that 12% of low scores were classified as positive and 5% of high scores as negative. To reduce this ambiguity, Islam (2014) proposed a system that unifies the numerical score with the polarity extracted from the text. Similarly, Aljrees et al. (2024) used TextBlob to identify biases in application reviews, classifying 24.72% of reviews with a polarity below 0.5 and a rating above 3 as biased.

Collectively, these studies demonstrate that textual comments may be more representative of the user's actual sentiment than the scores themselves, reinforcing the need for approaches that integrate both sources of information.

2.2 *Large Language Models*

LLMs have significantly advanced Natural Language Processing (NLP), particularly in text comprehension and opinion mining. Their ability to generalize tasks through zero-shot predictions has enabled applications in various fields, including rating prediction (Marreira et al., 2025a; Kang et al., 2023) and recommendation systems (Zhang et al., 2024).

In the context of explainable recommendation systems, Liu et al. (2025) propose a solution that seeks to align the predicted score with the explanation generated by the system. The model utilizes an LLM to predict a product's score based on user and product information; subsequently, it generates a textual explanation coherent with that score. To evaluate the coherence between the score and the explanation, the authors use the GPT-4o model, highlighting the potential of LLMs in detecting inconsistencies between comments and assigned ratings.

While prior work has already investigated rating–comment discrepancies (Marreira et al., 2025b), this study extends this line of research by evaluating a broader set of LLMs with a different technique. Furthermore, although Portuguese is one of the five most widely used languages on the internet (Pereira, 2021), research on this topic focuses primarily on English. This gap highlights the need to investigate this analysis in Brazilian Portuguese, leveraging the potential of LLMs to enhance the understanding of inconsistencies between comments and review scores.

3 Methodology

3.1 Dataset

We developed a crawler to collect product reviews posted on the Amazon Brazil e-commerce website, spanning 2021 to 2024. Only reviews in Portuguese were considered. The research focuses exclusively on the review text and the numerical rating provided by the user. The core objective is to identify contradictions between the assigned rating and the textual content of the review, restricting the analysis to extreme ratings, that is, 1-star or 5-stars reviews.

The methodological decision to exclusively analyze extreme ratings, specifically 1 and 5 stars, was a strategy implemented to ensure the study concentrated on cases with unequivocal polarity expectations. Despite its limited real-world representativeness, this approach facilitated the identification of explicit contradictions between textual content and rating, where the presence of inconsistency is

critically significant.

Ratings of an intermediate nature, exemplified by 2 to 4 stars, frequently encompass a combination of mixed or neutral sentiments. This introduces a degree of ambiguity that has the potential to adversely affect the objectivity of the consistency analysis. Investigation into the characteristics of these intermediate ratings represents a valuable direction for future research endeavors.

To ensure class balance across ratings, we guaranteed that each category contains the same number of 1-star and 5-star reviews. The summary of the dataset is presented in the Table 1. In total, the dataset comprises 20,586 reviews. Larger volumes are observed in *Books* and *Fashion*, whereas *Computers* and *Pets* are less represented. This variation reflects relative popularity across categories, with *Books* standing out, possibly driven by the habit of readers to review their experiences.

Table 1: Distribution of reviews by category.

Category	1-star	5-stars	Total
Automotive	873	873	1,746
Baby	1,057	1,057	2,114
Cell Phones	867	867	1,734
Grocery	742	742	1,484
Games	1,217	1,217	2,434
Computers	185	185	370
Books	2,259	2,259	4,518
Fashion	1,443	1,443	2,886
Pets	445	445	890
Toys	1,205	1,205	2,410
Total	**10,293**	**10,293**	**20,586**

3.2 Models

In our experiments, we considered both multilingual and Portuguese-specialized models, varying in scale. Since generative models are inherently non-deterministic, each model was applied in three rounds to observe consensus. To mitigate this non-determinism, the temperature parameter was set to 0.0 (zero)[3].

3.2.1 GPT-5

Developed by OpenAI, ChatGPT-5 is a state-of-the-art multimodal language model built on the Transformer architecture (Vaswani et al., 2017). Trained on a diverse and rigorously filtered dataset, ChatGPT-5 employs reinforcement learning to refine internal reasoning before responding, enhancing accuracy, reflection, and safety alignment. The system introduces major improvements in reasoning, instruction following, hallucination reduction, and sycophancy mitigation, excelling particularly in writing, coding, and health tasks (OpenAI, 2025). The model was accessed through its OpenAI API.

3.2.2 DeepSeek-3.2

Developed by independent researchers, DeepSeek is an open-source language model built on the Transformer architecture, emphasizing scalability and versatility. It was trained on a large bilingual dataset containing around 2 trillion tokens, mainly in Chinese and English. The model demonstrates advanced abilities in reasoning, mathematics, and code generation. Moreover, it employs techniques such as supervised fine-tuning (SFT) and direct preference optimization (DPO), which enhance its performance in dialogue tasks and its alignment with user intent (Bi et al., 2024). In our experiments, we used version V3.2 of DeepSeek, accessed through OpenAI API.

3.2.3 LLaMA-4

Llama 4 is a cutting-edge, natively multimodal language model developed by Meta AI. We specifically used the LLaMA-4-Scout-17B-16E-Instruct model in our experiments, featuring 17 billion active parameters within a 16-expert mixture-of-experts (MoE) architecture (109 billion total parameters), trained on approximately 40 trillion tokens. It supports both text and image inputs, offers a long context window of up to 10 million tokens, and is optimized for efficient inference (including single NVIDIA H100 GPU deployment) while delivering strong performance across tasks such as dialogue, image reasoning, coding and summarisation (Meta, 2025). The model was accessed through Hugging Face Inference API.

3.2.4 Sabiá-3

Developed by Maritaca AI, Sabiá-3 is a language model tailored specifically for Brazilian Portuguese. It was trained on an extensive, high-quality corpus of Portuguese texts, emphasizing Brazil-focused sources such as cultural, historical, and academic materials. This specialization allows the model to capture the nuanced linguistic features,

[3]With the exception of the Bode-3.1 model, which was executed locally using the configuration $do\_sample = False$, to ensure deterministic outputs.

social conventions, and regional variations distinctive to the Brazilian context, making it highly effective for natural language processing tasks that require precise text understanding. Sabiá-3 is accessible through an API called MariTalk, which enables efficient interaction with the model while remaining cost-effective. Sabiá-3 operates at three to four times lower cost per token compared to frontier models, without sacrificing performance on Brazil-specific tasks (Abonizio et al., 2024). In our experiments, we used version V3.1 of Sabiá.

3.2.5 Bode-3.1

The Bode model family (Garcia et al., 2024) consists of fine-tuned large language models specialized in Brazilian Portuguese. These models, available on Hugging Face, were designed to enhance Portuguese-language understanding and generation by integrating cultural, linguistic, and contextual nuances specific to Brazil. The fine-tuning process employed translated versions of datasets such as Alpaca and UltraAlpaca. In this project, we used the model Bode-3.1-8B-Instruct-LoRA version, which is optimized for instruction-following tasks, maintaining robust performance in Portuguese dialogue and reasoning while remaining lightweight enough for efficient experiments. The model was loaded locally from the Hugging Face Hub.

3.3 Prompt

The prompt used in the experiments was structured into five blocks with specific functions. Figure 2 illustrates the written prompt in Portuguese and Figure 3 the translated prompt in English. The first block (1) instructs the model to verify the coherence between the rating and the Portuguese text (PT-BR). The second block (2) specifies the input into three different data: category text and rating. The third block (3) specifies the output into classes: *COERENTE* (COHERENT) and *INCOERENTE* (INCOHERENT). The fourth block (4) specifies the rules that must be adopted by the model. Finally, the fifth block (5) presents some orientations for the models. This clear and concise structure promotes uniform responses and robustness in zero-shot settings.

The combination of these five blocks results in a concise, deterministic, and operationally robust prompt, suitable for use in zero-shot contexts. The clear separation between the decision rule and the response format was crucial to ensuring uniformity across runs and enabling within-model consistency analysis. The prompt was defined through exploratory experiments in which different formulations were tested until a stable version was reached, i.e., a prompt that consistently yielded correct responses under the automated evaluation protocol.

1. Você é um verificador de coerência entre avaliação e texto (PT-BR).
2. Entrada: categoria, texto, avaliação (avaliação ∈ [1,5])
3. Saída: 'COERENTE' ou 'INCOERENTE' (apenas uma palavra, sem explicações).
4. Regra central:
   - Texto NEGATIVO ⇒ avaliação = 1.
   - Texto POSITIVO ⇒ avaliação = 5.

   Orientações:
   - Se a regra falhar ⇒ INCOERENTE.
   - Textos neutros ou irônicos sem contradição ⇒ COERENTE.
   - Se o texto estiver vazio, só emojis neutros ou sem opinião clara ⇒ COERENTE.
   - Se houver críticas e elogios, decida pelo sentimento dominante do texto.
   - Use 'categoria' apenas para entender gírias/termos do domínio; não altere a regra.
   - Analise ⇒ categoria: {categoria} — comentário: {comentario} — avaliação: {avaliacao}
   - Responda apenas com: COERENTE ou INCOERENTE
5. Resposta:

Figure 2: Prompt (PT-BR) used in the experiments.

3.4 Metrics

To evaluate the performance of our classification models, we relied on a broad set of well-established metrics. In particular, we employed precision, recall, and the F1 score, which are standard measures frequently used in text classification research. Also, in order to quantify the agreement among the three independent runs of the LLM, Fleiss' κ coefficient was adopted (Fleiss, 1971). It is a generalization of Cohen's κ for $k > 2$ annotators, suitable for nominal categories and the absence of missing data—requirements that correspond exactly to the setting of this work, in which each review received

🇺🇸

You are a coherence checker between evaluation and text (PT-BR).

Input: category, text, evaluation (evaluation ∈ [1,5])

Output: 'COHERENT' or 'INCOHERENT' (only one word, no explanations).

Central rule:

- NEGATIVE text ⇒ evaluation = 1.
- POSITIVE text ⇒ evaluation = 5.

Guidelines:

- If the rule fails ⇒ INCOHERENT.
- Neutral or ironic texts without contradiction ⇒ COHERENT.
- If the text is empty, only neutral emojis or without clear opinion ⇒ COHERENT.
- If there are both criticisms and compliments, decide by the dominant sentiment of the text.
- Use 'category' only to understand slang/domain terms; do not alter the rule.
- Analyze ⇒ category: {category} — comment: {comment} — evaluation: {evaluation}
- Answer only with: COHERENT or INCOHERENT

Answer:

Figure 3: Prompt (translated) used in the experiments.

a binary label (INCOHERENT or COHERENT) from three runs of our models investigated.

Formal definitions: Let N be the total number of items evaluated and k the number of annotators. For each item $i \in \{1, \ldots, N\}$ and category $c \in \{1, \ldots, C\}$, we denote by n_{ic} the count of annotators who selected category c. In the present work, $C = 2$ and $k = 3$. With these notations, the average proportion of observed agreements ($\bar{P}$) and the proportion of agreements expected by chance ($\bar{P}_e$) are given by

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{c=1}^{C} n_{ic}(n_{ic} - 1)}{k(k-1)}, \qquad \bar{P}_e = \sum_{c=1}^{C} p_c^2,$$

where $p_c = \frac{1}{N_k} \sum_{i=1}^{N} n_{ic}$ represents the marginal frequency of category c. Fleiss' coefficient is then defined by

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}.$$

The values of κ range between -1 and 1, where $\kappa = 1$ indicates perfect agreement, $\kappa = 0$ corresponds to the agreement of chance-level and $\kappa < 0$ suggests systematic disagreement. Following the scale of (Landis and Koch, 1977), the intervals 0–0.20 (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), and 0.81–1.00 (almost perfect) are considered.

4 Experiments

4.1 Inconsistency Identification

The internal consistency of the models was assessed through three independent runs (v1-v3) of 20.586 comments using the same prompt. Furthermore, considering the three simultaneous executions, Fleiss' Kappa coefficient was calculated for each model. The number of detections of incoherent comments per model for each round, their percentage of the corpus, and the Fleiss' Kappa are presented in Table 2.

The results show that all models achieved a low variance among the prediction rounds with Kappa values higher than 0.95 in all cases, except for Bode-3.1, which produced deterministic outputs due to local execution with the configuration $do\_sample = False$. These values indicate that all models presented run in almost perfect agreement. The highest consistency is shown by the Bode-3.1 model (Kappa = 1.0), with exact predictions in all rounds, followed by DeepSeek-3.2 (Kappa = 0.9883).

In Table 2, we also can observe that the number of incoherent comment detections varied widely between the models. While Llama-4 detected the maximum of 1,588 (aprox. 7.72%) comments as incoherent, GPT-5 only detected around 370 (aprox. 1.80%). These results show that all models are consistent in the predictions among the rounds, but the models differ substantially in the rate at which they flag reviews as incoherent, suggesting different detection thresholds or calibration behaviors.

4.2 Human Agreement

A conservative criterion was adopted, selecting a representative random sample of the dataset for detailed analysis of the models. Table 3 shows the distribution of this subset of reviews. To generate the ground truth of this test sample, two human reviewers, trained for the task, labeled the comments as COHERENT, INCOHERENT, MIXED or NEUTRAL by confronting their text with its respective rating. Mixed reviews contained both positive and

Table 2: Incoherent responses per execution round and Fleiss' Kappa.

Model	**v1**	**v2**	**v3**	**Fleiss' Kappa**
GPT-5	371	373	370	0.9534
Sabiá-3.1	550	550	540	0.9587
Deepseek-3.2	1,020	1,023	1,023	0.9883
Llama-4	1,587	1,588	1,588	0.9709
Bode-3.1	872	872	872	1.0000

negative information, whereas neutral reviews consisted of factual statements or user typos. Ironic or sarcastic reviews were identified by the annotators and classified according to their underlying sentiment. For instance, if a review expresses negative sentiment through irony while assigning a five-star rating, it was labeled as incoherent. The agreement between them was 79,7%, with a Cohen's Kappa coefficient of 0.67, indicating substantial agreement according to the Landis and Koch scale.

Table 3: Distribution of reviews subset by category.

Category	1-star	5-stars	Total
Automotive	22	18	40
Baby	24	16	40
Cell Phones	24	16	40
Grocery	23	17	40
Games	23	17	40
Computers	14	15	29
Books	24	16	40
Fashion	16	24	40
Pets	21	19	40
Toys	23	17	40
Total	**214**	**175**	**389**

The contradictions between the two annotators were solved by a third independent evaluator. Furthermore, mixed and neutral reviews were classified as incoherent to align with the binary classification adopted by the models. The model's final prediction labels were determined by the majority vote across the three rounds (i.e., agreement in at least two rounds). Table 4 shows the precision, recall, and F1-score metrics per model per class, according to the ground truth.

The results show that the best model is DeepSeek-3.2, with F1 of 97.91% and 97.98% for incoherent and coherent detections, respectively. It is followed by Llama-4, which achieved F1 of 96.68% and 96.64%, respectively. Although both models are multilingual, they still performed better than Sabiá-3.1, an optimized model for Brazilian Portuguese, which achieved a F1 of 93.82% and 94.79% for incoherent and coherent detections, respectively.

Both Sabiá-3.1 and GPT-5 models favored predictions as coherent, as can be seen when comparing precision and recall metrics. Low recalls for incoherent detections (88.36% for Sabiá, and 69.84% for GPT-5) indicate a high occurrence of false negatives, i.e. a weaker capacity for detecting incoherence in the comments. Llama-4 showed the opposite behavior, with a higher recall (100%), but with more false positives, which led to a lower precision (93.56%). All these results reflect on the number of incoherent detections in in Table 2, where both Sabiá-3.1 and GPT-5 detected the lowest occurrences and Llama-4 the highest.

The Bode-3.1 model, optimized for Brazilian Portuguese, achieved the lowest performance among all evaluated models. Its F1-scores for incoherent and coherent detection (33.61% and 70.74%, respectively) indicate frequent confusion between the two classes. Notably, Bode-3.1 is a small model with 8 billion parameters; thus, the proposed task may be particularly challenging, even when using the largest available variant of this version.

4.3 Analysis of Incoherence by Rating

In order to examine how ratings relate to textual incoherence, we analyzed the same test dataset of Table 3, where 214 comments (55.0%) had a 1-star rating, and 175 (45.0%) had a 5-star rating.

Out of the 389 reviews in the test subset, 152 (39.1%) were labeled INCOHERENT. When broken down by rating, incoherence affected 93 of 214 1-star reviews (43.5%), typically reflecting positive text paired with a low rating, and 59 of 175 5-star reviews (33.7%), typically reflecting negative text paired with a high rating. Although these rates are of similar magnitude, the higher share and larger absolute count of incoherent cases among 1-star reviews suggests that rating–text mismatches

Table 4: Metrics by model.

Model	Class	Precision	Recall	F1
GPT-5	INCOHERENT	1.0000	0.6984	0.8224
	COHERENT	0.7782	1.0000	0.8753
DeepSeek-V3.2	INCOHERENT	0.9689	0.9894	0.9791
	COHERENT	0.9898	0.9700	0.9798
LLaMA-4-Scout	INCOHERENT	0.9356	1.0000	0.9668
	COHERENT	1.0000	0.9350	0.9664
Sabiá-3.1	INCOHERENT	1.0000	0.8836	0.9382
	COHERENT	0.9009	1.0000	0.9479
Bode-3.1-8B	INCOHERENT	0.8163	0.2116	0.3361
	COHERENT	0.5618	0.9550	0.7074

are more prevalent at the negative extreme of the scale. Overall, these results indicate that coherence-checking models should consider not only textual polarity, but also rating priors and potential asymmetries in how users assign extreme ratings.

4.4 Analysis of Incoherence by Category

Given its best overall classification performance on the human-labeled subset (Table 4) and strong run-to-run stability, we use DeepSeek-3.2 to examine how rating–text incoherence varies across product categories. Figure 4 reports the category-level rates of incoherent reviews correctly identified with respect to our ground-truth annotations. We observe that incoherence is present across all categories, with most of them clustering around 10%. The *Computers* category exhibits the lowest rate (4.8%), whereas *Cell Phones* reaches the highest (12.9%). These differences suggest that mismatch patterns may be category-dependent, potentially reflecting factors such as review style, product complexity, or how users interpret extreme ratings. We leave a controlled analysis of these factors, for example, accounting for review length, sentiment intensity, and category sample size as future work.

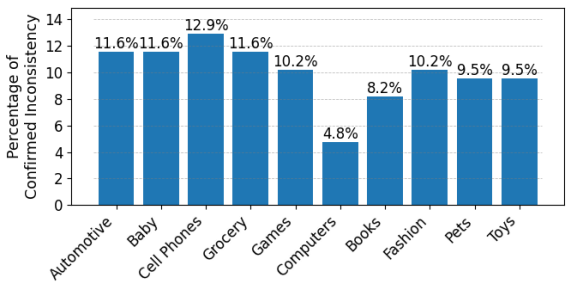

Figure 4: Reviews correctly classified as incoherent by DeepSeek-3.2 by category.

5 Conclusions

This study evaluated the effectiveness of Large Language Models (LLMs) in detecting semantic incoherence between review texts and their respective numeric ratings (1 or 5 stars) within a Brazilian e-commerce context. Using a curated dataset of over 20,000 reviews and a zero-shot protocol, the research compared the performance of state-of-the-art multilingual models (GPT-5, Llama-4, DeepSeek-3.2) with models optimized for Portuguese (Sabiá-3.1 and Bode-3.1). The methodology focused not only on classification accuracy but also on the stability of model responses across multiple independent runs, utilizing agreement metrics such as Fleiss' Kappa.

The results demonstrated that, in general, LLMs have high internal consistency, with very low variability between prediction rounds (Kappa > 0.95). The best overall performance was presented by the DeepSeek-3.2 model, which achieved an F1-score of 97.91% in detecting incoherence, outperforming both the other multilingual models and those specialized in Portuguese. Models such as GPT-5 and Sabiá-3.1 were observed to be conservative, favoring the "coherent" classification and generating more false negatives, while Llama-4 showed high recall but lower precision. Furthermore, the analysis revealed that incoherence is more frequent in 1-star reviews (43.5%) than in 5-star reviews, suggesting a greater fragility in the text-rating correlation when the user expresses dissatisfaction. In addition, incoherent evaluations are present in all product categories in similar rates, except for *Computers* category, with a very lower rate.

As future work, we propose to expand the analysis to other techniques, such as few-shot prompting, chain-of-thought prompting and fine-tuning. Be-

sides, we plan to explore reviews with intermediate ratings (2 to 4 stars), where ambiguity and mixed sentiments make the detection of contradictions more complex. It is also pertinent to investigate further the reasons why certain product categories, such as *Computers*, present lower inconsistency rates, hypothesizing that the consumer's technical profile might influence the rating accuracy. Finally, it is suggested to improve the understanding of how emotional influence affects the coherence between the written text and the rating assigned by human reviewers.

6 Acknowledgements

The authors acknowledge the support provided by the Universidade do Estado do Amazonas (UEA) through the Academic Productivity Grant (GPA) (Administrative Ordinance No. 1177/2025-GR/UEA). This work was also supported by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant no. 408490/2024-1.

References

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2024. Sabiá-3 technical report. *arXiv preprint arXiv:2410.12049*.

Turki Aljrees, Muhammad Umer, Oumaima Saidani, Latifah Almuqren, Abid Ishaq, Shtwai Alsubai, Imran Ashraf, and 1 others. 2024. Contradiction in text review and apps rating: prediction using textual features and transfer learning. *PeerJ Computer Science*, 10:e1722.

Amal Almansour, Reem Alotaibi, and Hajar Alharbi. 2022. Text-rating review discrepancy (trrd): an integrative review and implications for research. *Future Business Journal*, 8(1):3.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Michela Fazzolari, Vittoria Cozza, Marinella Petrocchi, and Angelo Spognardi. 2017. A study on text-score disagreement in online reviews. *Cognitive Computation*, 9(5):689–701.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gabriel Lino Garcia, Pedro Henrique Paiola, Luis Henrique Morelli, Giovani Candido, Arnaldo Cândido Júnior, Danilo Samuel Jodas, Luis Afonso, Ivan Rizzo Guilherme, Bruno Elias Penteado, and João Paulo Papa. 2024. Introducing bode: a fine-tuned large language model for portuguese prompt-based task. *arXiv preprint arXiv:2401.02909*.

Michaela Geierhos, Frederik Simon Bäumer, Sabine Schulze, and Valentina Stuß. 2015. " i grade what i get but write what i think." inconsistency analysis in patients' reviews. In *ECIS*.

Nan Hu, Noi Sian Koh, and Srinivas K Reddy. 2014. Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales. *Decision support systems*, 57:42–53.

Mir Riyanul Islam. 2014. Numeric rating of apps on google play store by sentiment analysis on user reviews. In *2014 international conference on electrical engineering and information & communication technology*, pages 1–4. IEEE.

Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2023. Do llms understand user preferences? evaluating llms on user rating prediction. *arXiv preprint arXiv:2305.06474*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Shijie Liu, Ruixin Ding, Weihai Lu, Jun Wang, Mo Yu, Xiaoming Shi, and Wei Zhang. 2025. Coherency improved explainable recommendation via large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12201–12209.

Emanuelle Marreira, Tiago de Melo, Miguel de Oliveira, and Carlos MS Figueiredo. 2025a. Rating prediction in brazilian portuguese: A benchmark of large language models. *Journal of the Brazilian Computer Society*, 31(1):827–838.

Emanuelle Marreira, Tiago de Melo, Miguel Oliveira, and Carlos Maurício. 2025b. Detectando incoerências avaliativas em e-commerce com llms-um estudo de caso na amazon brasil. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 535–539. SBC.

Juan Pedro Mellinas, Juan L Nicolau, and Sangwon Park. 2019. Inconsistent behavior in online consumer reviews: The effects of hotel attribute ratings on location. *Tourism Management*, 71:421–427.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed on: November 10, 2025.

Susan M Mudambi, David Schuff, and Zhewei Zhang. 2014. Why aren't the stars aligned? an analysis of online review content and star ratings. In *2014 47th Hawaii International conference on system sciences*, pages 3139–3147. IEEE.

OpenAI. 2025. Gpt-5 system card. https://openai.com/pt-BR/index/gpt-5-system-card/. Accessed on: November 10, 2025.

Denilson Alves Pereira. 2021. A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.

Abhinav Sharma, Sangwon Park, and Juan L Nicolau. 2020. Testing loss aversion and diminishing sensitivity in review sentiment. *Tourism Management*, 77:104020.

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.

Xiaoyu Zhang, Yishan Li, Jiayin Wang, Bowen Sun, Weizhi Ma, Peijie Sun, and Min Zhang. 2024. Large language models as evaluators for recommendation explanations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 33–42.