HuNeBR: A Multitask Benchmark to Evaluate LLMs' Understanding of Northeastern Brazilian Portuguese Humor

José Gama[1], **David Maia**[2], **Leandro Balby**[1], **Fábio Morais**[1], **João Brunet**[1]
[1] Federal University of Campina Grande (UFCG)
[2] Federal Institute of Paraíba (IFPB)
Campina Grande, Paraíba, Brazil
jose.gama@copin.ufcg.edu.br, david.maia@ifpb.edu.br, lbmarinho@dsc.ufcg.edu.br
fabio@computacao.ufcg.edu.br, joao.arthur@computacao.ufcg.edu.br

Abstract

Humor recognition is a major challenge in Natural Language Processing (NLP) due to its subtle and context-dependent nature. Despite advances, Large Language Models (LLMs) still struggle with this task, especially in Brazilian Portuguese, where no dedicated benchmarks exist. This paper presents HuNeBR, a new benchmark of 475 annotated humorous texts from Northeastern Brazilian comedians. The benchmark evaluates LLMs on three tasks: identifying punchlines, classifying texts into eight comic styles, and explaining humor. This is the first benchmark to evaluate LLMs on the in-depth interpretation of humorous texts in Brazilian Portuguese, going beyond the binary tasks of traditional humor benchmarks. Both general-purpose and Portuguese-specialized LLMs were evaluated under zero-shot and few-shot settings. The findings indicate that LLMs perform very well at identifying punchlines, show inconsistent results in classifying comic styles, and produce humor interpretations that mostly align with human judgments. Among the models assessed, general-purpose multilingual systems like GPT-4 and Gemini 2.5 Flash achieved the top overall performance, whereas Sabiá 3.1, a model specialized in Brazilian Portuguese, demonstrated competitive results across all three tasks, highlighting the value of locally trained models in capturing linguistic and cultural subtleties.

Keywords: large language models, brazilian humor benchmark, regional linguistic features, speech transcriptions, comic styles, joke punchlines, text interpretation

1. Introduction

Large Language Models (LLMs) have significantly advanced state-of-the-art performance in natural language processing (NLP) tasks, including translation, summarization, and information retrieval (Bharathi Mohan et al., 2024). Over the last few years, these models have consistently outperformed previous baselines across a wide range of benchmarks and real-world applications. For example, GPT-4 and Claude models have achieved near-human performance in standardized tests, such as the bar exam and Graduate Record Examinations (GRE) (Achiam et al., 2024). In machine translation, models such as Google's PaLM 2 and Meta's SeamlessM4T have outperformed previous neural systems by reducing error rates and enabling high-quality translation for low-resource languages (Anil et al., 2023; Anastasopoulos et al., 2023).

Similarly, in code generation tasks, LLMs such as CodeLlama and GPT-4 have shown the ability to solve competitive programming problems and generate production-ready code (Roziere et al., 2023). In educational contexts, LLM-powered tutoring systems have been shown to improve learning outcomes by providing personalized explanations and feedback (Kasneci et al., 2023). These advances have also been systematically evaluated using benchmarks such as HumanEval (Chen et al., 2021), MMLU-PRO (Wang et al., 2024), and MR-GSM8K (Zeng et al., 2025), proving that LLMs not only improve research benchmarks but also prove effective in practical, high-impact scenarios.

Despite advances, interpreting subjective and cultural dimensions - such as humor - remains a challenge (Cuskley et al., 2024). Humor is central to human communication, shaping social bonds and guiding attention, but it is notoriously difficult to model computationally. As Jentzsch and Kersting (2023) note, humor is often implicit, context-dependent, and relies on subtle linguistic or cultural cues, making it a long-standing AI challenge. This complexity extends to available resources (Kalloniatis and Adamidis, 2024). Most humor recognition datasets are in English and primarily target simple binary classification tasks, such as humor detection (Meaney et al., 2021), comparative funniness prediction (Hossain et al., 2020), sarcasm detection (Khodak et al., 2018), and irony detection (Van Hee et al., 2018). However, to date, no benchmark has been proposed for humor recognition in Brazilian Portuguese.

To address this gap, we introduce HuNeBR[1] - a benchmark to evaluate the ability of LLMs to understand humor in Brazilian Portuguese. We built the benchmark using YouTube shorts transcripts featuring comedians from Northeast Brazil. This region plays a significant role in national humor

[1] https://github.com/llm-pt-ibm/brazilian_northeast_humor_benchmark

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 299–311
11–12 May 2026.

and provides the cultural context for the authors' work, ensuring that the selection captured a diverse range of humor styles and regional expressions. This process resulted in a dataset of 475 annotated comedic texts. HuNeBR comprises three tasks: i) punchline identification; ii) classification into comic styles (fun, benevolent humor, wit, nonsense, irony, sarcasm, cynicism, and satire); and iii) explanation of the comic elements in each humorous text.

We evaluated both general-purpose and Brazilian Portuguese-specialized LLMs on HuNeBR under two scenarios: zero-shot, without task examples, and few-shot, with a few examples provided. This evaluation aimed to identify the models' strengths and weaknesses. The results indicate that LLMs achieve strong performance in punchline identification, unstable performance in comic style classification, and humor interpretations that are largely consistent with human judgments. Among the evaluated models, general-purpose multilingual systems such as GPT-4 and Gemini 2.5 Flash achieved the highest overall results. Also, Sabiá 3.1, a model specialized in Brazilian Portuguese, delivered competitive performance across all three tasks, highlighting the potential of locally trained models to capture linguistic and cultural nuances.

The remainder of this paper is organized as follows. Section 2 reviews related work, covering the main directions in Brazilian Portuguese LLM evaluation, as well as research in computational humor recognition. Section 3 introduces the HuNeBR benchmark, including dataset construction, task definitions, and evaluation metrics. Section 4 details the evaluation setup, describing the models used and the results obtained. Finally, Sections 5 and 6 discuss these results, draw conclusions, and outline directions for future research.

2. Related Works

Existing work on the evaluation of LLMs in Brazilian Portuguese has focused mainly on two directions: structured exam-based evaluations, such as ENEM-inspired benchmarks (Silveira and Mauá, 2017; Almeida et al., 2023) and the Brazilian Bar exam (Delfino et al., 2017); and social media tasks, including sentiment analysis (Brum and Volpe Nunes, 2018), and hate speech detection (Fortuna et al., 2019; Vargas et al., 2022). More recent contributions have expanded this landscape with small-scale QA datasets, such as Faquad (Sayama et al., 2019), and large-scale open-domain question answering on Brazilian events, such as Almeida et al. (2025b). Additionally, IberoBench (Baucells et al., 2025) has sought to address the evaluation gap for Iberian languages more broadly, while BRoverbs (Almeida et al., 2025a) measures models' understanding of Portuguese proverbs. Although these initiatives enrich the evaluation of LLMs in Brazilian Portuguese, none addresses the humor domain.

Research in computational humor recognition, spanning nearly three decades, has evolved from symbolic rule-based systems focused on linguistic incongruities (Taylor and Mazlack, 2004; Ritchie, 2001) to statistical approaches based on handcrafted features (Mihalcea and Strapparava, 2005). This evolution signaled a transition to neural and pre-trained language model–based methods (Kalloniatis and Adamidis, 2024).

Initially, the field was driven by supervised machine learning approaches (e.g. SVM and Naive Bayes), which heavily depended on feature engineering, including aspects such as alliteration, human centrality and ambiguity (Zhang and Liu, 2014; Cattle and Ma, 2018). With the advent of deep learning approaches, models such as LSTM networks and, more recently, Transformer-based architectures such as BERT and RoBERTa (Morishita et al., 2020; Hasan et al., 2021) have shown significant improvements in humor detection, generation, and explanation. Multi-modal approaches have further expanded the field by incorporating visual and acoustic cues to capture the multifaceted nature of humor (Choube and Soleymani, 2020).

Despite significant progress in recent years, the vast majority of benchmarks and datasets remain centered on English, with only a handful of exceptions in languages such as Spanish, Russian, and Hindi (Ortega-Bueno et al., 2018; Ermilov et al., 2018; Chauhan et al., 2021). To the best of our knowledge, there are no publicly available datasets or benchmarks specifically focused on humor in Brazilian Portuguese.

Addressing this gap is important: Brazilian Portuguese is spoken by over 200 million people, and humor plays a central role in everyday communication, online discourse, and cultural expression in Brazil. Developing resources that reflect these linguistic and cultural specificities is important to building NLP systems that are both inclusive and contextually accurate. To this end, our work introduces the first benchmark specifically designed to evaluate humor recognition in Brazilian Portuguese.

3. HuNeBR Benchmark

This section presents the main components of the HuNeBR benchmark: the dataset of humorous texts with annotations for their comic interpretation (Section 3.1), the benchmark tasks including their prompts and evaluation scenarios (Section 3.2), and the metrics employed to assess model predictions for each task (Section 3.3).

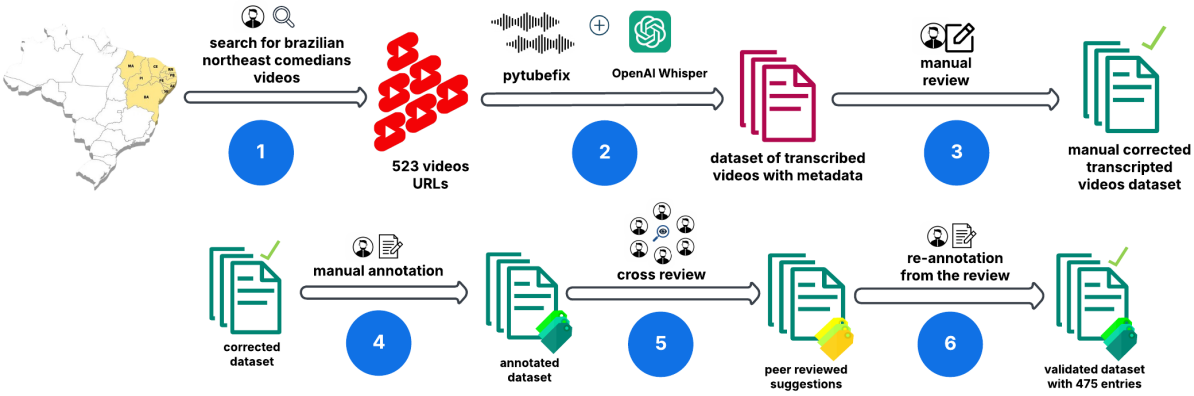

Figure 1: Overview of data collection and annotation steps.

3.1. Dataset

We chose to focus the dataset on Northeast Brazilian comedians for three key reasons: first, the authors' familiarity with this cultural context supports a more accurate interpretation of linguistic nuances and humor styles; second, Northeast comedians hold a prominent and influential place in the national comedy scene (Cavalcante et al., 2020; Universidade de Fortaleza (Unifor), 2021; Nascimento, 2019), ensuring relevance; finally, no existing benchmark specifically addresses this regional context, suggesting an area that remains relatively underexplored, which this work seeks to address.

The dataset was constructed in six main stages, as illustrated in Figure 1. First, we compiled a list of comedians from the Northeast region of Brazil, identified through comedy events, news articles, podcast appearances, and public recognition. This search continued until saturation was reached across all nine regional states. We performed manual searches on Google and YouTube using Portuguese queries (e.g. *"Humoristas (comedians) + [state name] + [comedy events]"*) to identify comedians and gather 523 YouTube Shorts URLs. YouTube Shorts were chosen due to their one-minute duration, simplifying both data processing and analytical consistency. The collected content included stand-up performances, podcast excerpts, scripted videos, and comedy shows.

The second step involved using the pytubefix[2] to extract the audio and metadata (e.g., titles, publication dates) from each video. The audio was then transcribed with OpenAI's Whisper[3] model, and transcripts were manually reviewed and corrected. For dialogues, the transcription was structured to clearly indicate each speaker. In the third step, after manual review and filtering—removing duplicates, strongly offensive content (e.g., hate speech or discriminatory expressions), and incomplete jokes—the dataset was finalized with 475 video transcriptions. Colloquial language and mild profanity were preserved, as they are common features of comedic discourse and are often intrinsic to this type of content. The videos were published between May 26, 2017, and November 5, 2024.

Next, the first author performed a three-dimensional interpretive annotation. First, all punchlines were identified as segments triggering incongruity detection and reinterpretation, in line with cognitive models of humor processing (Vaid et al., 2003).

At the stylistic level, each text was categorized into eight comic styles (fun, benevolent humor, nonsense, wit, irony, satire, sarcasm, and cynicism) following Schmidt-Hidding (1963) and empirically expanded by Ruch et al. (2018), who conceptualize them as distinct expressive modes of humor differing in social function, cognitive structure, and evaluative tone. Although originally discussed within differential psychology, we adopt these categories as functional descriptors of humor realization rather than as personality traits, since in oral comedic performance such styles manifest linguistically through target orientation, evaluative stance, pragmatic framing, and mechanisms of incongruity construction.

Briefly, fun refers to light-hearted amusement; benevolent humor to tolerant and socially positive joking; nonsense to absurd or logic-defying constructions; wit to clever verbal ingenuity; irony to implicit meaning reversal; satire to critical social commentary; sarcasm to sharp, often mocking remarks; and cynicism to contemptuous or distrustful humor. The co-occurrence matrix of such styles in the dataset (Figure 2) shows that fun, benevolent humor, wit, and irony frequently co-occur, while nonsense and sarcasm are less common. Satire and cynicism are the rarest styles, appearing with the lowest frequencies.

Additionally, each text was assigned an explanatory rationale describing the techniques employed

[2] https://pypi.org/project/pytubefix/
[3] https://github.com/openai/whisper

and the intended humorous effect. Such rationale, together with punchline segmentation and comic style attribution across eight stylistic dimensions, results in a high annotation density per instance, enabling fine-grained evaluation of humor interpretation.

Figure 2: Comic styles co-occurrence matrix.

In the fifth step, six additional reviewers independently examined randomized, non-overlapping subsets of the 475 entries, assessing 78, 79, 82, 79, 78, and 79 items, respectively. All reviewers were residents of Northeastern Brazil for a minimum of 21 years and included three undergraduate and three graduate students in Computer Science, each with prior experience in NLP tasks.

Before the review phase, reviewers participated in a structured training on punchline definitions, comic styles, and text explanations, using annotated examples to ensure conceptual alignment and minimize ambiguity. During the review, each item was assessed independently, and reviewers could suggest annotation modifications through dedicated feedback channels.

Limiting the dataset to comedians from this region allowed reviewers to leverage their cultural familiarity, ensuring accurate interpretation of regional expressions, local references, and nuanced humor. Suggestions were distributed as follows: 1.89% for humor reasoning, 1.05% for punchlines, 17.05% for comic style additions, and 8.84% for style removals. Changes in one task often affected others, as edits to a punchline or explanation could trigger a reassessment of the related comic styles.

Finally, in the sixth step, data were re-annotated based on the reviewers' feedback, resulting in measurable updates across annotation dimensions. Regarding comic styles, the percentage of annotations modified was as follows: 1.68% for fun, 2.32% for humor, 4.00% for nonsense, 4.84% for wit, 4.63% for irony, 4.42% for satire, 3.58% for sarcasm, and 2.95% for cynicism. At the textual level, word counts increased by 1.3% for identified punchlines, 0.82% for jokes explanations, and 1.79% for corrected transcriptions. The relatively low proportion of post-review modifications indicates substantial annotation stability prior to adjudication.

This expert review protocol, conducted between January and March 2025, follows a primary-annotation plus structured independent review design, with full version histories publicly available on Zenodo[4] for transparency and reproducibility.

3.2. Tasks Construction

The benchmark is composed of three tasks: punchline identification, comic style classification, and humor reasoning. These tasks were selected because they reflect complementary stages in how humans understand humor—from recognizing the punchline, to identifying stylistic intent, to reasoning about why something is comic. Moreover, previous studies have evaluated LLMs on similar tasks individually and in other languages, such as punchline identification (Romanowski et al., 2025a), joke explanation (He et al., 2025), and comic style classification (e.g., humor vs. sarcasm) (Choi et al., 2023). By integrating these three dimensions, HuNeBR provides a more comprehensive evaluation of LLMs' humor interpretation capabilities.

All tasks were implemented through structured prompts aligned with the annotation guidelines used in the dataset construction. Operational definitions were explicitly embedded in the prompts whenever applicable, and output formats were strictly constrained to ensure direct comparability with the manually annotated dataset, which we treat as the gold standard for evaluation. Complete prompt templates are publicly available in the benchmark repository.

The constructed tasks are as follows:

1. **Punchlines Identification:** The model receives the humorous text together with an explicit definition of punchlines as segments that trigger incongruity detection and reinterpretation. It must extract only the corresponding spans in a structured list format, without additional commentary, assessing its ability to identify structural resolution points in joke construction.

2. **Comic Style Classification:** For each comic style, a dedicated prompt presents its operational definition and decision criteria. The model evaluates whether the style is present in the text and produces a binary output (1 for

[4] https://zenodo.org/records/15473224

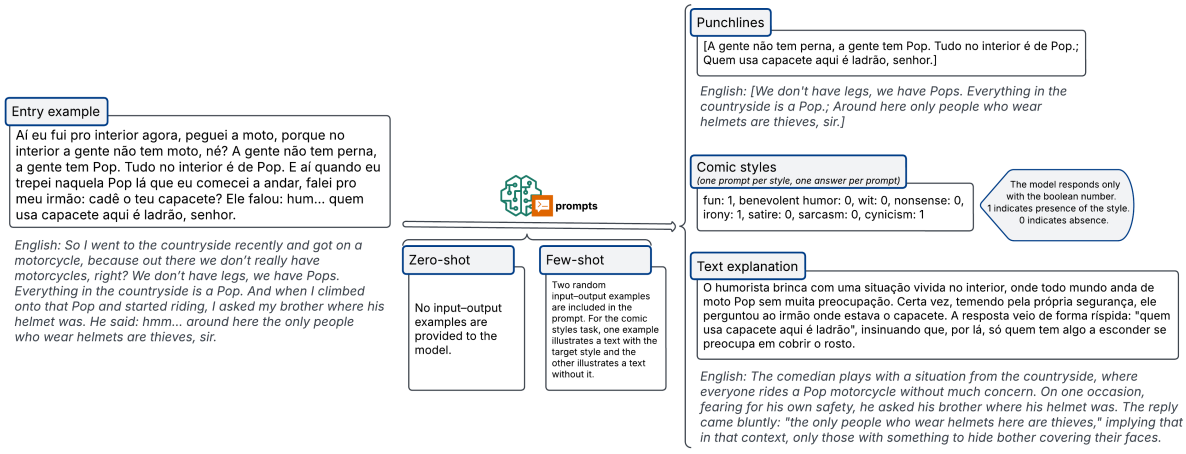

Figure 3: Example input and outputs for the three HuNeBR tasks.

presence, 0 for absence), measuring sensitivity to stylistic and formal cues while avoiding cross-category interference.

3. **Humor Reasoning:** The model is instructed to generate a concise explanation of the elements that produce the comic effect, assessing its capacity to articulate underlying humor mechanisms.

To assess the influence of in-context examples, the first two tasks were evaluated under both zero-shot and few-shot settings (Brown et al., 2020). In the few-shot scenario, two randomly selected examples were included using a fixed seed derived from the input text and task name; for comic style classification, one positive and one negative instance were provided per style. The humor reasoning task was evaluated only in the zero-shot setting to avoid bias. As humor explanation is inherently open-ended, often allowing multiple valid interpretations, few-shot examples could encourage pattern imitation rather than independent reasoning.

Figure 3 illustrates an example of input text with its corresponding outputs across all tasks. Particularly, it presents a text with two punchlines and three simultaneous comic styles, requiring specific knowledge for understanding — the brand of a popular motorcycle.

3.3. Metrics

For punchline identification, we used the Dice Similarity Coefficient, which measures the degree of overlap between two sets of words, ranging from 0 (no intersection) to 1 (identical). Specifically, the metric measures the similarity between the punchline segments identified by the models and the annotated ones, focusing on lexical overlap. The coefficient has also been applied in NLP tasks, such as query–document matching in chatbots (Prasetya and Priyatno, 2022).

For comic style classification, models produced binary labels for each of the eight styles, evaluated via macro-averaged F1-score. The F1-score is the harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives. Since the distribution of comic styles in the dataset is imbalanced, macro-averaging was chosen to ensure that each class contributes equally to the final score.

For each comic style, the classification task was formulated as an independent binary decision (presence vs. absence). While we report the macro-averaged F1-score across styles, we also analyze class-specific performance to better understand model behavior. In particular, we examine how models perform separately on instances assigned the positive label (presence of a style) and the negative label (absence of a style), which allows us to identify systematic biases such as over-predicting the presence of certain styles.

Finally, to evaluate model performance on the humorous text explanation task, we adopted a model-as-judge approach, relying on o3-mini (OpenAI, 2025), which achieved the best results in the judge model benchmark proposed by Tan et al. (2024). With a specialized prompt, the judge model rated the agreement between model-generated and human-annotated explanations on a five-point Likert scale: 1 for totally disagree, 2 for partially disagree, 3 for neutral or mixed, 4 for partially agree, and 5 for totally agree. The evaluation focused on alignment with human reasoning, factual accuracy, and clarity, while keeping the judge model unaware of the explanation sources to avoid bias.

To verify the judge model's reliability, we manually inspected a random 5% of the samples from each agreement level (1–5). Human validation confirmed the model's judgment in 73%, 71%, 80%,

79%, and 75% of cases, respectively. In the few cases of disagreement, the model's score differed by only one level above or below the human rating, indicating that discrepancies were minor. These results demonstrate substantial agreement between the model and human reviewers, reinforcing the robustness of the judge-based evaluation procedure.

4. Evaluation

We evaluated seven large language models on HuNeBR to establish reference performance and assess the impact of zero-shot and few-shot prompting. Each model was tested on HuNeBR's three tasks - punchline identification, comic style classification, and humor reasoning. Both prompting strategies were applied to the first two tasks.

Although outputs were freely generated, each task required a specific format. When models deviated, post-processing recovered valid outputs instead of discarding them, reducing penalties for minor formatting issues. For Punchline Identification, only the first list was extracted; cases without a list were excluded. For Comic Style Classification, the first digit was used; responses without numbers were discarded. Humor Reasoning accepted any free-text explanation. The following sections describe the evaluated models and results.

4.1. Models

Our evaluation on HuNeBR included both general-purpose and specialized Brazilian Portuguese LLMs. We selected five state-of-the-art models widely used in multilingual NLP: DeepSeek-R1-Qwen3-8B (AI, 2024a), Granite-3.3-8B-Instruct (Research, 2024), GPT-4 (Achiam et al., 2024), Gemini-2.5-Flash (DeepMind, 2024), and LLaMA-3-405B-Instruct (AI, 2024b). In addition, to account for models tailored to Brazilian Portuguese, we included Sabiá-3.1 (Abonizio et al., 2025), the largest Brazilian Portuguese language model to date, and Gemma-3-Gaia-PT-BR-4b-it (CEIA-UFG, 2024), a compact model designed for efficient experimentation and instruction-following in Portuguese. This combination allows for a fair comparison between frontier LLMs and systems optimized for the linguistic and cultural nuances of Brazilian Portuguese. Hereafter, we refer to these models using their respective acronyms: DeepSeek, Granite, GPT, Gemini, LLaMA, Sabiá, and Gaia.

4.2. Results

The following sections provide a detailed analysis of the results for each of the three tasks evaluated.

Model	Zero-shot	Few-shot
GPT	**0.73**[0.71;0.75]	**0.73** [0.71;0.75]
Gemini	0.68 [0.66;0.70]	**0.72** [0.70;0.74]
Sabiá	**0.63** [0.61;0.65]	0.61 [0.59;0.64]
LLaMA	0.59 [0.57;0.62]	**0.65** [0.63;0.68]
Granite	**0.56** [0.54;0.58]	0.54 [0.42;0.67]
Gaia	**0.49** [0.46;0.51]	0.47 [0.44;0.50]
DeepSeek	0.19 [0.13;0.25]	**0.47** [0.41;0.52]

Table 1: Mean Dice Similarity per model across both prompting settings in the Punchline Identification task, with 95% confidence intervals.

4.2.1. Punchlines Identification

As shown in Table 1, the models can be grouped according to their overall performance in the punchline identification task. GPT and Gemini tended to achieve the highest Dice scores under both prompting strategies, while the Portuguese-specialized models Sabiá and Gaia generally showed moderate results. DeepSeek had the lowest score with zero-shot, with a wide confidence interval suggesting high variability and limited generalization. An example of an error in punchline identification, performed by DeepSeek, is its selection of a text that contains some humor (*"You crazy, pothead, he unites people, it's an inexplicable union."*), but that is not the correct punchline: *"It looks like a prayer, damn it! Because everyone is silent..."*.

Among the general-purpose models, DeepSeek, Gemini and LLaMA appeared to improve from zero-shot to few-shot. DeepSeek showed the largest relative gain, nearly 28%, suggesting that in-context examples significantly aided task adaptation. LLaMA also showed a notable increase of around 6%. However, GPT maintained similar scores in both settings. Finally, Granite, despite reaching a moderate few-shot score of 0.54, exhibited a less consistent performance, evidenced by a broad confidence interval from 0.42 to 0.67.

For the Brazilian Portuguese-specialized models, Sabiá and Gaia exhibited stable performance across both zero-shot and few-shot strategies, as reflected by their overlapping confidence intervals.

4.2.2. Comic Style Classification

The comic style classification task poses a significant challenge for the evaluated models, as described by the F1-macro scores in Table 2. Gemini and GPT achieved the highest scores, both reaching an F1 score of only 0.25 in the zero-shot configuration, while Gemini maintained the same value under few-shot prompting and GPT showed a marginal decrease to 0.24. LLaMA followed closely with scores of 0.24 and 0.22. The lowest results were obtained by DeepSeek and Gaia, both remaining at 0.12 across the two prompting strategies.

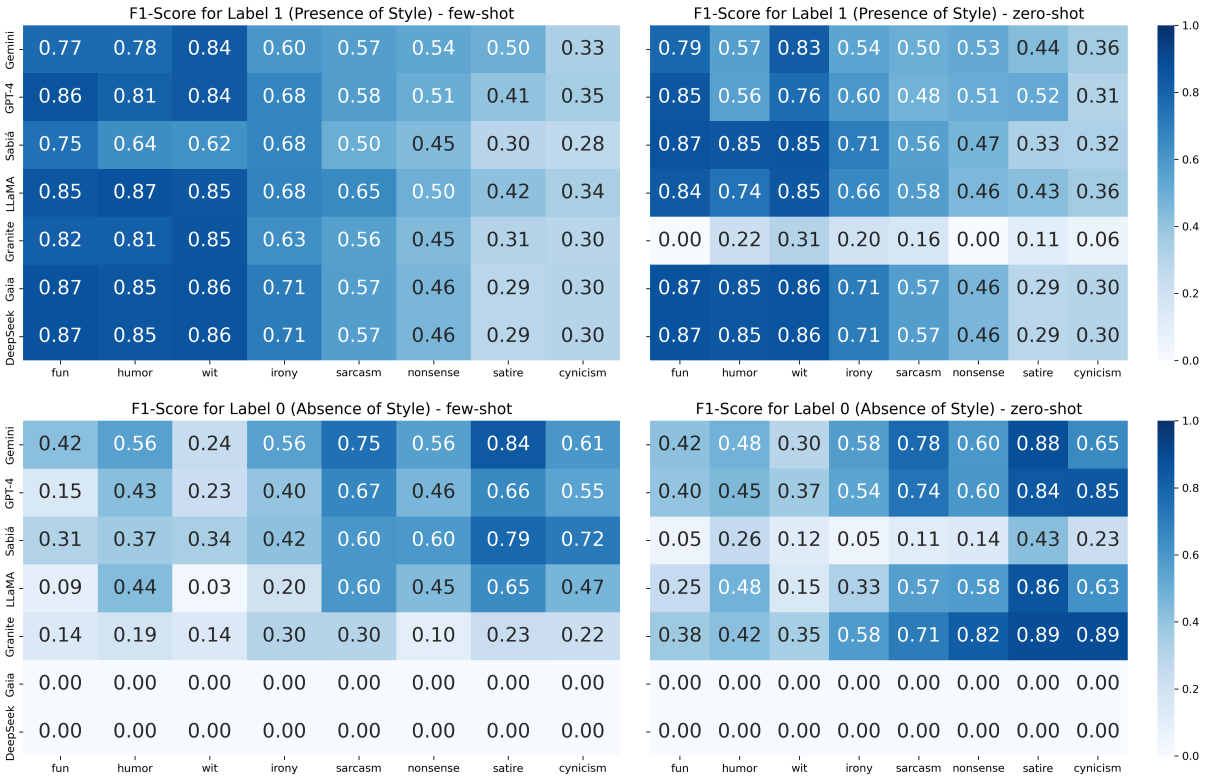

Figure 4: F1-Scores for each model across the labels of all comic styles, comparing the zero-shot and few-shot prompting strategies on the Comic Style Classification task.

Granite and Sabiá exhibited modest gains when switching from zero-shot to few-shot, increasing from 0.18 to 0.21 and 0.20 to 0.25, respectively. Overall, scores ranged from 0.12 to 0.25, indicating narrow performance among the models in this task.

Model	Zero-shot	Few-shot
Gemini	**0.25**	**0.25**
GPT	**0.25**	0.24
Sabiá	0.20	**0.25**
LLaMA	**0.24**	0.22
Granite	0.18	**0.21**
DeepSeek	0.12	0.12
Gaia	0.12	0.12

Table 2: F1-Macro per model and prompting setting on the Comic Style Classification task.

Figure 4 details the performance of detecting the presence versus the absence of each comic style. Gaia and DeepSeek exhibited an extreme asymmetry between classes, achieving F1-scores of 0.0 for the absence of all styles while maintaining high scores around 0.8–0.9 for its presence. This pattern indicates a decision collapse, where the models systematically predict the presence of a style for nearly all instances, failing to discriminate between positive and negative cases. For example, both models achieved F1-scores around 0.86–0.87 for fun and wit, but 0.00 for the corresponding absence of these styles. Granite showed a similar pattern, scoring 0.0 on presence for fun and nonsense styles with zero-shot. Such behavior points to systematic biases toward surface-level patterns rather than balanced style discrimination.

Comparing the zero- and few-shot scenarios, Gemini showed strong performance in detecting fun, humor, and wit across both strategies, with a notable 21% improvement in humor under few-shot. Results were moderate for irony, sarcasm, and nonsense, showing only a small few-shot gains. For more challenging styles, satire reached 0.50, while cynicism remained the hardest style (0.36). In detecting absent styles, the model performed best for sarcasm and satire in zero-shot (0.78 and 0.88), with moderate results for others styles.

On the other hand, GPT achieved the highest F1-scores overall for identifying the presence of fun, humor, wit, and irony (0.86, 0.81, 0.84, and 0.68 in few-shot), showing stable or improved performance compared to zero-shot. It performed moderately for sarcasm, nonsense, and satire, but poorly for cynicism. When identifying the absence of styles, GPT excelled in cynicism, satire, and sarcasm (0.85, 0.84, and 0.74 in zero-shot), though it struggled with fun, humor, and wit. Similarly, LLaMA performed strongly on the presence of fun, humor, and wit (0.85–0.87 in few-shot), with moderate scores for irony (0.68) and sarcasm (0.65).It performed

well in detecting the absence of satire (0.86 in zero-shot), moderately on sarcasm and cynicism, and poorly on wit (0.15) and fun (0.25) in zero-shot.

Finally, Sabiá performed strongly in detecting fun, humor, wit, and irony, with its best results reaching 0.87, 0.85, 0.85, and 0.71 in the zero-shot setting. For the remaining styles, performance ranged from low (cynicism and satire, around 0.30) to moderate (sarcasm and nonsense, around 0.45–0.50). In identifying the absence of styles, it performed well for satire (0.79) and cynicism (0.72) in few-shot, but remained weak for others, not exceeding 0.60.

Overall, the models struggled to accurately classify certain comic styles, as illustrated in the following excerpt: *"Look at us, human beings. Imagine God, God himself. God sends his only son to die for a humanity that pushes a door that says PULL..."* In this example, GPT, Gemini, and Granite failed to detect the sarcasm, whereas Sabiá and DeepSeek incorrectly labeled the passage as ironic.

4.2.3. Humor Reasoning

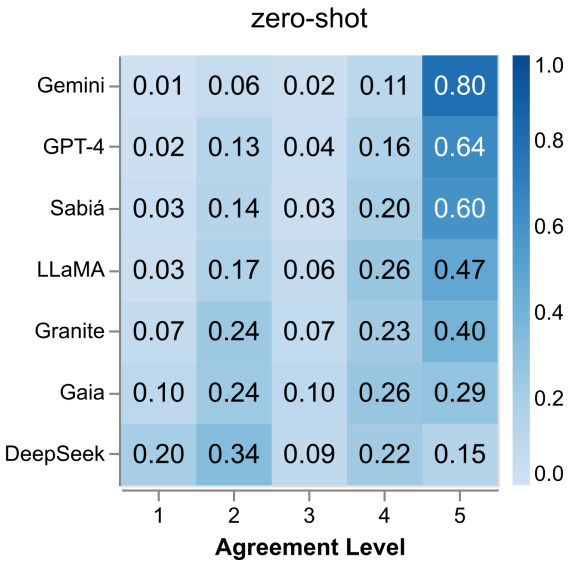

Figure 5: Distribution of agreement levels assigned by the judge model (o3-mini) in the Humor Reasoning task, ranging from strong disagreement (1) to full agreement (5) with annotated explanations.

For the Humor Reasoning task, the models were instructed to use the context of the joke to generate a concise explanation of the elements that produce the comic effect. Figure 5 shows the agreement between model explanations and the ground-truth, as rated by the judge model. Gemini and GPT achieved the highest agreement scores among the general models, followed by Sabiá. When considering the combined proportions of agreement levels 4 and 5, Gemini led with 0.91. Similarly, GPT reached 0.80 for the combined levels. Sabiá achieved a positive agreement with a proportion of 0.80. By contrast, DeepSeek performed worst with 0.54 at levels 1 and 2.

Some examples of incorrect explanations returned by the models illustrate common errors. DeepSeek sometimes repeated parts of the joke instead of providing an explanation, e.g., *"His friend looked over and asked, 'What do you think, Flávio?' I said, 'I don't think she wants to kiss me.' His friend looked over..."*.

Similarly, Gaia occasionally failed to capture the humorous elements. Consider the following excerpt, translated from Brazilian Portuguese: "*Comedian: Whoever has gone through some very funny situations with cachaça, a Brazilian sugarcane spirit, in their life, shout "me"! Audience member: Me! Comedian: What was the drink? Audience member: Beer! I was fishing in Araguaia, drinking Crystal. Comedian: Crystal? Oh, sure! And he has the nerve to say it was beer! Audience member: Well, since it was Crystal, it caused quite a commotion. Comedian: Yeah, it got things a bit messy. That's normal. Audience member: Then you go into the bushes, right, since there's no bathroom, you squat and fall forward. Comedian: Are you giving a tutorial on how to... Audience member: Fell forward. Comedian: You fell forward? Audience member: Forward, drunk! Comedian: Just imagine him there... I love it, I'm loving it, I'm loving it, I'm loving it. For me, this could already be on TV, I'm loving it. Take your time. Audience member: Yeah! That was it!*".

In this case, the humor arises from a combination of incongruity and cultural reference: the spectator initially frames the story as involving "cachaça" but then reveals that the drink was actually "Crystal", a beer often stereotypically associated with low quality. The comedian exploits this mismatch to mock the situation, which may not be fully accessible to models lacking cultural grounding in Brazilian contexts.

This limitation is reflected in Gaia's explanation, which states that *"The humor in the text primarily resides in the repetition and exaggeration of the comedian's reactions to the spectator's comments, creating an absurd contrast between the simple description of the event and the emotional intensity expressed. The inversion of expectation, in which the comedian emphasizes the banality of the spectator's experience, and the spectator's own admission of having confused the drink, contribute to the comedic effect. The repetition of the phrase "Me!" and the comedian's exaggeration of the situation amplify the irony and absurdity of the scene."*, but fails to identify the culturally grounded reference that drive the joke.

5. Discussion

The models show limited benefits from few-shot examples in punchline identification. GPT-4, Gemini and Gaia achieved robust and stable performance in joke segmentation across zero-shot and few-shot settings, suggesting a well-established pre-trained grasp of the canonical joke structure. This stability across prompt variations aligns with Romanowski et al. (2025a), who found comparable prompt-independent performance in ChatGPT for punchline extraction. In contrast, DeepSeek showed a 28% improvement under few-shot prompting, indicating stronger dependence on examples and weaker prior knowledge. A few observed errors echo findings from Bertero and Fung (2016), as models occasionally marked segments they judged as humorous, but that do not resolve any incongruity, showing structural understanding but limited pragmatic humor comprehension.

Classifying comic styles is the most challenging task for LLMs. Overall, models handle explicit and positive comic styles (e.g., fun, humor, wit) reasonably well, but struggle with complex or critical styles such as sarcasm, cynicism, and satire, which convey negative or evaluative tones, highlight flaws in individuals or societal norms, and require nuanced social, philosophical, or political reasoning. Small models often collapse predictions into a single label, reflecting limited discriminative ability or difficulty processing large prompts. Challenges may also arise from insufficient exposure to culturally or regionally specific humor, limiting interpretation of subtle cues. Stronger performance on positive styles appears driven by surface linguistic markers (e.g., lightness, joy, puns) rather than deep understanding, leading to overgeneralization when multiple comic tones co-occur. These patterns echo prior humor benchmarks (Choi et al., 2023; Yi et al., 2025), underscoring that recognition of critical or regionally nuanced humor still requires broader contextual and social reasoning.

LLMs can explain why texts are comic, but with limitations. Most models identify comic features in texts, often agreeing with human annotations, suggesting that reasoning capabilities vary across architectures — strong in general-purpose models like Gemini and GPT-4, moderate in Granite and Gaia, and limited in DeepSeek. Manual evaluation of a subset of explanations revealed that smaller models tend to associate humor primarily with colloquial language or profanity, which alone does not capture what makes the texts comic. Larger models, when failing, struggle with phonetic wordplays or subtle cultural references, reflecting limitations in multi-step reasoning and cultural knowledge, consistent with findings from Narad et al. (2025).

6. Conclusion

This work introduced HuNeBR, a benchmark for evaluating Large Language Models on humorous texts in Brazilian Portuguese, based on authentic transcriptions of Northeastern Brazilian comedians and reflecting culturally rich oral humor. By integrating punchline identification, comic style classification, and humor reasoning, the benchmark enables a multidimensional evaluation of humor comprehension. Experimental results indicate that models perform moderately to strongly on punchline detection and explanation tasks but struggle with fine-grained comic style classification.

Future work may extend the dataset to additional Brazilian regions and platforms, explore reasoning-oriented prompting strategies, and investigate linguistic and contextual factors underlying model errors. HuNeBR provides a resource for studying humor in culturally grounded and dialect-specific contexts, particularly in Northeastern Brazilian Portuguese, and offers a replicable framework for developing similar benchmarks in other underrepresented dialectal or cultural settings.

7. Acknowledgments

We thank the six participants who reviewed the dataset transcriptions and annotations. This work was supported by IBM and the Parque Tecnológico da Paraíba (PaqTcPB).

8. Ethical Considerations and Limitations

This benchmark comprises textual transcriptions of publicly available YouTube videos, curated exclusively for academic and non-commercial research purposes. The dataset constitutes a transformative use of the original material, treating the content as linguistic data rather than entertainment, and includes attribution metadata (e.g., video title and URL) to ensure transparency and respect for authorship in line with international fair use principles.

Nonetheless, some limitations must be acknowledged: comedian selection was constrained by public visibility and availability on YouTube Shorts, potentially introducing platform and popularity bias; manual annotation inherently involves interpretive subjectivity despite cross-review procedures among six reviewers; and the relatively small size of the dataset (475 examples) may limit large-scale generalization or cross-regional comparisons.

9. Bibliographical References

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2025. Sabiá-3 technical report.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

DeepSeek AI. 2024a. Deepseek r1-qwen3-8b model card. Hugging Face model card.

Meta AI. 2024b. The llama 3 herd of models. ArXiv preprint arXiv:2407.21783.

Thales Sales Almeida, Giovana Kerche Bonás, and João Guilherme Alves Santos. 2025a. BRoverbs–Measuring how much LLMs understand Portuguese proverbs. *Journal of the Brazilian Computer Society*, 31(1).

Thales Sales Almeida, Giovana Kerche Bonás, João Guilherme Alves Santos, Hugo Abonizio, and Rodrigo Nogueira. 2025b. Tiebe: Tracking language model recall of notable worldwide events through time. *arXiv preprint arXiv:2501.07482*.

Thales Sales Almeida, Thiago Laitz, Giovana K Bonás, and Rodrigo Nogueira. 2023. Bluex: A benchmark based on brazilian leading universities entrance exams. In *Brazilian Conference on Intelligent Systems*, pages 337–347. Springer.

Antonios Anastasopoulos et al. 2023. Seamlessm4t—massively multilingual multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Rohan Anil et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, José Javier Saiz, Robiert Sepúlveda-Torres, et al. 2025. Iberobench: A benchmark for llm evaluation in iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519.

Dario Bertero and Pascale Fung. 2016. Deep learning of audio and language features for humor prediction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 496–501.

G Bharathi Mohan, R Prasanna Kumar, P Vishal Krishh, A Keerthinathan, G Lavanya, Meka Kavya Uma Meghana, Sheba Sulthana, and Srinath Doss. 2024. An analysis of large language models: their impact and potential applications. *Knowledge and Information Systems*, 66(9):5047–5070.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Henrico Brum and Maria das Graças Volpe Nunes. 2018. Building a sentiment corpus of tweets in Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Andrew Cattle and Xiaojuan Ma. 2018. Recognizing humour using word associations and humour anchor extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1849–1858, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Diego Frank Marques Cavalcante, Jamille Ipiranga de Lima, and Germana Cabral Goiana. 2020. A linguagem do humor regional do suricate seboso como forma de valorização da cultura nordestina/the language of regional humor of seboso suricate as a way to value northeast culture. *Brazilian Journal of Development*, 6(1):3209–3224.

CEIA-UFG. 2024. Gemma-3 gaia pt-br 4b it model card. Hugging Face model card.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Navonil Majumder, Amir Zadeh, Asif Ekbal, Pushpak Bhattacharyya, Louis-philippe Morency, and Soujanya Poria. 2021. M2h2: A multimodal multiparty hindi dataset for humor recognition in conversations. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 773–777.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand

social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Akshat Choube and Mohammad Soleymani. 2020. Punchline detection using context-aware hierarchical multimodal fusion. In *Proceedings of the 2020 international conference on multimodal interaction*, pages 675–679.

Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083.

Google DeepMind. 2024. Gemini 2.5 flash. Official model page.

Pedro Delfino, Bruno Cuconato, Edward Hermann Haeusler, and Alexandre Rademaker. 2017. Passing the brazilian oab exam: data preparation and some experiments. *arXiv preprint arXiv:1712.05128*.

Anton Ermilov, Natasha Murashkina, Valeria Goryacheva, and Pavel Braslavski. 2018. Stierlitz meets svm: humor detection in russian. In *Conference on artificial intelligence and natural language*, pages 178–184. Springer.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12972–12980.

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Rada Mihalcea, and Naihao Deng. 2025. Chumor 2.0: Towards better benchmarking Chinese humor understanding from (ruo zhi ba). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21799–21818, Vienna, Austria. Association for Computational Linguistics.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 Task 7: Assessing Humor in Edited News Headlines. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, pages 746–758, Barcelona, Spain (Online). Association for Computational Linguistics.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 325–340, Toronto, Canada. Association for Computational Linguistics.

Antonios Kalloniatis and Panagiotis Adamidis. 2024. Computational humor recognition: a systematic literature review. *Artificial Intelligence Review*, 58(2):43.

Enkelejda Kasneci et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Julie-Anne Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 531–538.

Terufumi Morishita, Gaku Morio, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at semeval-2020 task 7: Stacking at scale with heterogeneous language models for humor recognition. In *Proceedings of the fourteenth workshop on semantic evaluation*, pages 791–803.

Reuben Narad, Siddharth Suresh, Jiayi Chen, Pine S. L. Dysart-Bricken, Bob Mankoff, Robert Nowak, Jifan Zhang, and Lalit Jain. 2025. Which llms get the joke? probing non-stem reasoning abilities with humorbench.

Luana Matos do Nascimento. 2019. A representação dos nordestinos através dos influenciadores whindersson nunes e thaynara og. In *Anais do ENECULT – Encontro de Estudos Multidisciplinares em Cultura*, Salvador, BA. Mestranda no PPGMC/UFF. Acessado em: 19 fev. 2026.

OpenAI. 2025. Openai o3-mini: A compact reasoning model. Accessed: 2025-10-15.

Reynier Ortega-Bueno, Carlos E Muniz-Cuza, José E Medina Pagola, and Paolo Rosso. 2018. Uo upv: Deep linguistic humor detection in spanish social media. In *Proceedings of the third workshop on evaluation of human language technologies for Iberian languages (IberEval 2018) co-located with 34th conference of the Spanish society for natural language processing (SEPLN 2018)*, pages 204–213.

Muhammad Riko Anshori Prasetya and Arif Mudi Priyatno. 2022. Dice similarity and tf-idf for new student admissions chatbot. *RIGGS: Journal of Artificial Intelligence and Digital Business*, 1(1):13–18.

IBM Research. 2024. Granite 3.3-8b instruct model card. Hugging Face model card.

Graeme Ritchie. 2001. Current directions in computational humour. *Artificial Intelligence Review*, 16(2):119–135.

Adrianna Romanowski, Pedro H. V. Valois, and Kazuhiro Fukui. 2025a. From punchlines to predictions: A metric to assess LLM performance in identifying humor in stand-up comedy. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–46, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Adrianna Romanowski, Pedro H. V. Valois, and Kazuhiro Fukui. 2025b. From punchlines to predictions: A metric to assess LLM performance in identifying humor in stand-up comedy. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–46, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Baptiste Roziere et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Willibald Ruch, Sonja Heintz, Tracey Platt, Lisa Wagner, and René T Proyer. 2018. Broadening humor: Comic styles differentially tap into temperament, character, and ability. *Frontiers in psychology*, 9:6.

Hélio Fonseca Sayama, Anderson Viçoso Araujo, and Eraldo Rezende Fernandes. 2019. Faquad: Reading comprehension dataset in the domain of brazilian higher education. In *2019 8th Brazilian conference on intelligent systems (BRACIS)*, pages 443–448. IEEE.

Douglas C Schmidt, Jesse Spencer-Smith, Quchen Fu, and Jules White. 2023. Towards a catalog of prompt patterns to enhance the discipline of prompt engineering. *A Publication of SIGAda, the ACM Special Interest Group on Ada*, page 43.

Wilhelm Schmidt-Hidding. 1963. *Europäische Schlüsselwörter: Humor und Witz, Band I [European Keywords: Humor and Wit, Vol. 1]*. Huber, Munich.

Igor Cataneo Silveira and Denis Deratani Mauá. 2017. University entrance exam as a guiding test for artificial intelligence. In *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 426–431. IEEE.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Julia M Taylor and Lawrence J Mazlack. 2004. Computationally recognizing wordplay in jokes. In *Proceedings of the annual meeting of the cognitive science society*, volume 26.

Universidade de Fortaleza (Unifor). 2021. Pesquisadores propõem que o humor cearense seja considerado patrimônio cultural da humanidade. *Unifor – Pós-Graduação*. Acessado em: 19 fev. 2026.

Jyotsna Vaid, Rachel Hull, Roberto Heredia, David Gerkens, and Francisco Martinez. 2003. Getting a joke: The time course of meaning activation in verbal humor. *Journal of Pragmatics*, 35(9):1431–1449.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 39–50.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark.

Advances in Neural Information Processing Systems, 37:95266–95290.

Peiling Yi, Yuhan Xia, and Yunfei Long. 2025. Irony detection, reasoning and understanding in zero-shot learning. *IEEE Transactions on Artificial Intelligence*.

Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. 2025. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. In *Proceedings of the International Conference on Learning Representations (ICLR), 2025*. Poster.

Renxian Zhang and Naishi Liu. 2014. Recognizing humor on twitter. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 889–898.

10. Language Resource References

Maritaca AI. 2025a. Sabia 3.1. https://docs.maritaca.ai/api/pt/list-models. Accessed at: 2025-10-17.

Meta AI. 2025b. Llama 3.1 405b instruct. https://huggingface.co/meta-llama/Llama-3.1-405B-Instruct. Accessed at: 2025-10-17.

S. S. T.; PEREIRA L. A.; AMADEUS M.; SCOTTI R.; FAZZIONI D.; NOVAIS A. M. A.; JORDÃO S. A. A. CAMILO-JUNIOR, C. G.; OLIVEIRA. 2025. Gaia: An open language model for brazilian portuguese. https://huggingface.co/CEIA-UFG/Gemma-3-Gaia-PT-BR-4b-it. Accessed at: 2025-10-17.

Community Contributors. 2025. pytubefix: A fixed and maintained fork of pytube for youtube video downloading. https://pypi.org/project/pytubefix/. Accessed at: 2025-10-24.

Google DeepMind. 2025. Gemini 2.5 flash. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash. Accessed at: 2025-10-17.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. https://huggingface.co/deepseek-ai/DeepSeek-R1-0528-Qwen3-8B. Accessed at: 2025-10-17.

IBM Granite Team. 2025. Granite-3.3-8b-instruct. https://huggingface.co/ibm-granite/granite-3.3-8b-instruct. Accessed at: 2025-10-17.

OpenAI. 2022. Whisper: Openai's speech recognition model. https://github.com/openai/whisper. Accessed at: 2025-10-24.

OpenAI. 2025a. Gpt-4. https://platform.openai.com/docs/models/gpt-4. Accessed at: 2025-10-17.

OpenAI. 2025b. o3-mini. https://platform.openai.com/docs/models/o3-mini. Accessed at: 2025-10-17.

The following language resources were created in this work. These include a dataset and a benchmark.

- **HuNeBR Dataset**: A Dataset of Annotated Humorous Transcriptions from YouTube Shorts by Northeastern Brazilian Comedians. Available at: https://zenodo.org/records/15473224
- **HuNeBR Benchmark**: a benchmark to evaluate LLMs regarding the understanding of humor in texts by comedians from Northeastern Brazil. Available at: https://github.com/llm-pt-ibm/brazilian_northeast_humor_benchmark