The Inadequacy of Automatic Evaluation Metrics in Question Answering: A Case-Study in Portuguese

Júlia da Rocha Junqueira and Viviane P. Moreira
Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil
{julia.junqueira, viviane}@inf.ufrgs.br

Abstract

Questions and answers are among the most fundamental forms of human communication. Question Answering (QA) is the task of correctly generating answers based on a context. To assess the success of the task, the answers are typically evaluated using traditional metrics such as BLEU, ROUGE, and METEOR. However, these metrics often fail to reflect the actual quality of the outputs. More recently, new evaluation metrics and the LLM-as-a-judge paradigm have also been applied to the evaluation of QA. To gain a deeper understanding of the capabilities and limitations of QA metrics, this work performs a comparative analysis of traditional and more recent approaches for QA evaluation. Experiments were conducted on the Pirá dataset (in Portuguese) using four LLMs to generate answers. Additionally, human evaluation was conducted to assess the correctness, completeness, clarity, and relevance of the generated content. We demonstrate that lexical metrics are limited in evaluating QA. We also observed that human evaluators favor models that provide higher information density, even when this contradicts prompt constraints, whereas lexical metrics penalize this verbosity. This divergence confirms that traditional metrics are insufficient for capturing the trade-off between adherence to instructions and the semantic richness valued by native speakers.

1 Introduction

Interaction through questions and answers is one of the most fundamental forms of human communication, playing a vital role in both everyday language and educational settings. In the field of Natural Language Processing (NLP), the task of Question Answering (QA) involves providing accurate and contextually relevant answers to queries based on a given text (Allam and Haggag, 2012).

Despite the recent advancements in Large Language Models (LLMs), the evaluation of their performance in QA tasks still relies heavily on traditional automatic metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). These metrics, however, are primarily based on lexical overlap and often fail to capture deeper semantic relationships. They show limited sensitivity to synonyms, paraphrasing, and the overall semantic appropriateness of the generated answers.

While evaluation metrics such as BERTScore (Zhang et al., 2019) and the LLM-as-a-judge paradigm (Zheng et al., 2023; Chiang and Lee, 2023; Wang et al., 2023) seek to overcome these limitations, they still lack consistent validation, particularly in languages other than English. BERTScore often fails to penalize candidate sentences that are lexically or stylistically similar to the reference, even when they contain subtle errors. As a result, translations or generated answers that are fluent but incorrect may still receive high scores if they present significant lexical overlap with the reference, reducing the ability of the metric to distinguish between outputs that are superficially similar but semantically flawed and those that are factually accurate (Hanna and Bojar, 2021).

Given these challenges, it becomes necessary to examine how different evaluation metrics assess the quality of generated data and how they correlate with human judgments. This study assesses various evaluation metrics used for the QA task, with a specific focus on their limitations. To contribute to the empirical evaluation of less-resourced languages, we conducted our experiments using a Portuguese dataset. Our methodology was designed to stress-test the alignment between automatic metrics and human judgment. For this, we conducted a comparative analysis using the Pirá dataset (Pirozelli et al., 2024). We evaluated the performance of four LLMs (Sabiá-3, GPT-4o, Gemini 2.0 Flash, and DeepSeek Chat) across metrics ranging from tradi-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 551–561

tional n-gram overlap (BLEU, ROUGE, METEOR) to semantic approaches (BERTScore, LLM-as-a-judge). Additionally, we compared these automatic scores with human evaluation, focusing on correctness, completeness, clarity, and relevance. This evaluation framework enabled us to identify significant discrepancies between metric-based rankings and human preferences, particularly in terms of how models handle prompt constraints versus informational density.

The key contributions of this work include a systematic analysis of the reliability of automated evaluation metrics by contrasting their scores with human judgments. Additionally, we present a case study on prompt adherence versus the perceived quality in Portuguese QA. By analyzing behavioral differences between multilingual and native models, we demonstrate that 'better' human ratings do not necessarily imply superior reasoning, but rather reflect differences in instruction-following. We identified that Sabiá-3 tends to generate verbose responses, despite explicit prompts for succinctness. Yet, this led to higher human scores, exposing a bias in which evaluators prefer detail over compliant brevity.

2 Background

This section reviews the foundations of evaluation methodologies commonly used for the QA task. It discusses the limitations of surface-level metrics for capturing factual correctness, as well as recent advancements in model-based evaluation that aim to better approximate human judgment.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) compares the reference text with the generated text, assigning a high score when there is considerable similarity in vocabulary between the generated and reference texts (Mohammadshahi et al., 2022). There are different variants of ROUGE. ROUGE-L is based on the idea of the longest common subsequence between the generated question and the reference question. This means that the metric checks the longest ordered sequence of words that appears in both questions, even if the words are not exactly consecutive. However, this evaluation approach may, in some cases, penalize a generated question even if it is valid, simply because it does not present high lexical similarity to the reference questions (Mohammadshahi et al., 2022).

BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) revolutionized (at the time) the way translation systems are evaluated, offering a fast, cost-effective alternative to extensive human evaluation. BLEU is based on the central premise that the closer a machine translation is to a professional human translation, the better it is. The metric operates by comparing n-grams between the candidate translation and one or more human reference translations, counting the number of matches found. These matches are position-independent and focus on content overlap between translations.

METEOR (Banerjee and Lavie, 2005) is an alternative to BLEU, aiming to offer word-order-sensitive evaluation in machine translations. This approach was developed to provide a stronger correlation with human judgments of translation quality, particularly at the sentence level. METEOR creates an explicit word-to-word alignment between the candidate translation and a reference text. To further refine its evaluation, METEOR introduces a penalty for fragmentation. This penalty assesses how well-ordered the matched words in machine translation are relative to the reference, rewarding translations that are not just lexically similar but also structurally coherent.

BERTScore (Zhang et al., 2019) is an automatic evaluation metric used to assess the quality of text generated by natural language models. It was developed to overcome weaknesses present in traditional n-gram-based metrics. Instead of relying on exact word matches, BERTScore uses contextual embeddings from pre-trained models, such as BERT (Devlin et al., 2019), to compare a candidate sentence with a reference sentence on a token-by-token basis. BERTScore also has significant drawbacks; a qualitative analysis revealed that it struggles to detect factual inaccuracies. For instance, it can assign a high similarity score to a translation that incorrectly substitutes "German language" for "English" or changes "Monday" to "Tuesday". It can also fail to recognize the equivalence of different notations for the same concept, such as "5ft 11in" and "1.80 metres". The performance of BERTScore can be inconsistent, with no single configuration being universally superior (Zhang et al., 2019).

The concept of *LLM-as-a-Judge* (Zheng et al., 2023; Chiang and Lee, 2023; Wang et al., 2023) leverages LLMs as automated evaluators for complex and often subjective tasks, offering a scalable alternative to traditional human experts or automatic metrics (Gu et al., 2025). This paradigm addresses the challenge of balancing context-aware

assessment with the need for efficiency, as it combines the nuanced reasoning capabilities of human evaluators with the speed of automated systems. The core process involves providing an LLM with input data, such as text or images, along with a specific context or prompt that defines the evaluation criteria. Despite its potential, the adoption of this approach is hindered by significant challenges such as a preference for longer answers or certain response positions, and defending against adversarial manipulation.

3 Related Work

A persistent bottleneck in NLP research is the alignment gap between automated scoring and human preference. This is particularly acute in QA, where valid responses can vary significantly in length, style, and lexical choice. Unlike classification tasks with binary outcomes, text generation requires metrics that can quantify semantic equivalence rather than just surface-level form.

Studies have conducted comparative analyses of evaluation metrics across different tasks and languages. They consistently demonstrated that no single metric provides a comprehensive picture of text quality, and that the choice of evaluation method significantly influences the perceived performance of different systems. The work by Davoodijam and Alambardar Meybodi (2024) emphasizes the importance of selecting appropriate evaluation metrics that can capture various aspects of text quality, including content accuracy, structural coherence, readability, and semantic relevance. Their analysis of summarization evaluation metrics provides valuable views that extend beyond summarization to other text generation tasks, including QA.

Blagec et al. (2020) examined the performance metrics commonly used to benchmark machine learning and AI models, concluding that reliance on single or overly simple metrics may not reflect true model performance, especially with imbalanced datasets, potentially hindering progress in the field. Sai et al. (2022); Novikova et al. (2017); Blagec et al. (2022) addressed the challenge that, while models have advanced significantly, the methods for evaluating them have struggled to keep pace. Traditional metrics, such as BLEU, ROUGE, and METEOR, are often inadequate for newer, more complex tasks like text generation and frequently do not correlate with human judgment. Nevertheless, they remain widely popular in the field.

Recent work has expanded faithfulness evaluation to Portuguese. Paula et al. (2025) investigated the applicability of Natural Language Inference (NLI) methods for evaluating summary faithfulness. Their analysis revealed that even NLI-based metrics rely on lexical overlap heuristics, such as ROUGE, rather than on semantic understanding. Bencke et al. (2024) investigated the reliability of using LLMs to replace human annotators in the evaluation of information retrieval systems. The major problem addressed is the high cost and time required to obtain human-relevance judgments (determining whether a document is relevant to a specific search query), and the conclusion is that, while they may not perfectly replicate human reasoning in every instance, LLMs are a reliable alternative.

Despite advances in evaluation methodology, significant limitations remain. The correlation between automatic metrics and human judgment varies considerably across tasks, domains, and languages (Chen et al., 2019). The computational cost of neural evaluation methods can be prohibitive for large-scale evaluation, while traditional metrics may not capture the semantic richness required for effective QA evaluation.

The inadequacy of traditional metrics is not unique to QA but has been extensively documented in other generation tasks. Studies in machine translation (Chen et al., 2022; Specia et al., 2010; Chauhan and Daniel, 2023; Vogel, 2004), text summarization (de Camargo et al., 2025; Nimah et al., 2023; Bhandari et al., 2020; Nguyen et al., 2024), question generation (Ehlert et al., 2025; Nema and Khapra, 2018), dialogue response generation (Liu et al., 2016) have consistently pointed out the limitations of n-gram-based evaluation.

Furthermore, a gap in the literature remains regarding the specific behavior and efficacy of evaluation metrics for QA in Portuguese. While previous work has explored faithfulness in summarization (Paula et al., 2025), relevance in information retrieval (Bencke et al., 2024), and hate speech detection (Oliveira et al., 2023), the specific challenges of evaluating QA have not been adequately addressed in this context. Most existing comparative studies on QA metrics are heavily skewed towards English, potentially overlooking language-specific nuances in morphology and syntax. This paper addresses this disparity by providing a comprehensive analysis of evaluation metrics for Portuguese QA.

4 Methodology

To achieve the primary goal of this study – benchmarking the reliability of evaluation metrics for Portuguese QA –we designed a framework focused on the *alignment gap* between automated scoring and human judgment. Unlike traditional approaches that prioritize model performance, our methodology isolates the evaluation process itself. We generated responses under strict conciseness constraints to specifically analyze the impact of *verbosity bias* and *instruction adherence* across different metric paradigms.

The Pirá dataset (Pirozelli et al., 2024) was selected for its status as a Portuguese, human-verified QA resource, and it covers the domains of oceanography, the Brazilian coast, and climate change. Unlike other datasets available in Portuguese (which rely solely on automatic translation from English), Pirá involves a curation process in which translations are explicitly checked by human annotators. Each instance in the dataset includes a paraphrased question and a validation answer, which ensures high data reliability by confirming that questions are unambiguously answerable by native speakers. This ensures that our experiments are not influenced by words that have suffered from the problem of "Translationese", where text is translated so literally that it respects the grammar and dictionary definitions of the target language, but fails to capture the cultural nuance or natural flow of the original meaning. Among the different benchmarks in the dataset, we specifically selected the standard benchmark split, which consists of open-ended questions paired with scientific supporting texts. This format forces the models to generate free-form answers based on context, thereby testing their ability to adhere to factual constraints rather than the simplified multiple-choice format. Its total size is 2.2K instances, with 1.8K for training, 225 for validation, and 227 for testing. Each instance contains the context in English, the context in Portuguese, the question in English, the question translated into Portuguese, the question paraphrased in Portuguese (a few instances), the answer in English, the answer translated into Portuguese, and the answer validated in both languages, among other data.

The experimental framework used to answer the research questions is depicted in Figure 1. The methodology steps are described next.

The **Pre-processing** step involved automatically replacing translated questions with their paraphrased versions, aiming to standardize and reduce ambiguity. Additionally, the original dataset contains some unanswerable questions, for which the answer is set to "null"; we removed these instances from our set because the goal of the experiment is to assess how well language models can generate answers based on a given context. If a question has no reference answer in the dataset, there is no way to meaningfully compare the model's output. For the experimental phase of this work, a sample of 50 instances was selected from the test set.

In the **Generation** phase, four language models were used to generate the answers: Sabiá-3 (Pires et al., 2023), GPT-4o (OpenAI et al., 2024), Gemini 2.0 Flash (Team et al., 2025), and DeepSeek (Liu et al., 2024). Considering that the Pirá dataset consists of "semi-factual" questions, which require answers based strictly on the provided context, a more restrictive generation configuration was chosen to minimize creativity and the introduction of external information. The prompt used to instruct the models was: *"Answer with no more than one short sentence, using only contextual information. Do not add any external data."* This guideline was designed to ensure that the answers were concise and factually anchored in the reference text.

The **Metric Calculation** phase employed a multi-dimensional assessment strategy that combines traditional automated metrics, model-based evaluation (LLM-as-a-judge), and human validation.

To complement and validate the automatic metrics, a human evaluation was conducted. The generated responses were evaluated by native Portuguese speakers. The 50 questions were divided into two sets of 25, with annotators split into two groups of three. Each group was assigned to one set, meaning every question in a set was independently evaluated by three annotators. For each question, annotators rated the outputs generated by all four models across four dimensions — correctness, completeness, relevance, and clarity — on a Likert scale from 1 to 5. This resulted in each annotator evaluating 100 generated answers (25 questions $*$ 4 model outputs), totaling 300 evaluated answers per set. This approach ensured a balanced distribution of the annotation workload.

Inter-rater agreement was assessed using pairwise weighted Cohen's Kappa (quadratic), which accounts for the ordinal nature of the Likert scale. Agreement ranged from slight to moderate across

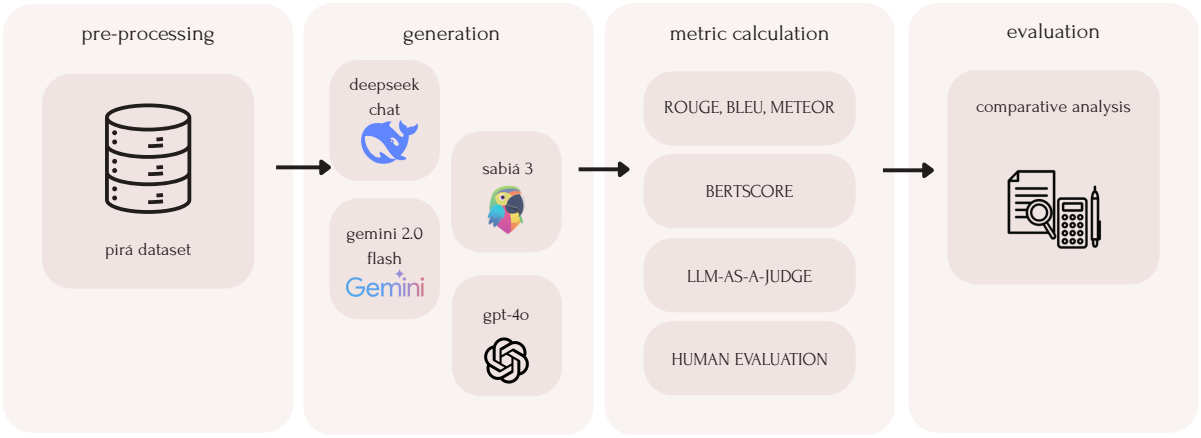

Figure 1: Methodological workflow proposed to analyze QA evaluation metrics, detailing the models used for generation and the metrics evaluated.

both groups (κ = 0.13–0.50), which is consistent with the inherently subjective nature of open-ended QA evaluation. For the correlation analysis, we used the mean rating across the three annotators in each group as the human judgment score.

To ensure consistency among annotators during evaluation, strict guidelines were established for the boundaries of the Likert scale (Likert, 1932), varying from 1 (worst) to 5 (best), considering four dimensions:

(i) *Correctness* (Corr): Is the answer factually correct according to the context? Often referred to in the literature as *Faithfulness* or *Factual Consistency*), it assesses whether the answer is factually correct in relation to the context. A correct response must not contain hallucinations or unverified information that contradicts the source text (Maynez et al., 2020).

(ii) *Completeness* (Compl): Does the answer cover all relevant information? This criterion measures whether the generated text encapsulates all relevant information from the prompt that is present in the source context.

(iii) *Relevance* (Rel): Does the answer actually address the question asked, without straying from the topic? Measures the alignment between the generated response and the specific query (Maynez et al., 2020).

(iv) *Clarity* (Clar): Is the answer well-written, cohesive, and easy to understand? This criterion evaluates the grammatical accuracy and readability of the text, ensuring the response is intelligible and well-structured, independent of its factual content.

Our goal was subdivided into the following research questions:

- **RQ1: Metric Alignment.** To what extent do lexical overlap metrics (*e.g.,* ROUGE, BLEU) and semantic evaluators (*e.g.,*, BERTScore, LLM-as-a-judge) correlate with human evaluations? To address RQ1, we computed the Spearman correlation coefficients between automated metrics and human judgments. Response quality was measured using a set of established metrics (ROUGE, BLEU, METEOR, and BERTScore as described in Section 2). The generated text quality was also assessed using the LLM-as-a-Judge approach. For this study, the GPT-4o model (OpenAI et al., 2024) was used as a "judge" to analyze and score the responses generated by the other models, providing a scalable evaluation based on complex semantic criteria. This choice was based on work by Liu et al. (2023) who used GPT-4 as a judge and achieved a Spearman correlation of 0.514 on summarization tasks. The judge received the question, the supporting context, and the generated answer, and was prompted to assign a score from 1 to 5 for each dimension. To reduce output variability, the prompt constrained the response format to four scores separated by commas in a fixed order (e.g., 5,4,5,5). The criteria definitions provided to the LLM judge were deliberately aligned with those given to human annotators, so that both evaluation modalities operated under the same conceptual framework. The full prompt is provided in Appendix A.
- **RQ2: The Adherence-Utility Trade-off.** How does the trade-off between strict prompt adherence and information density affect

model rankings? For RQ2, we evaluated the impact of verbosity bias on model rankings. We calculate the **Length Ratio (LR)**, defined as the quotient of the generated response length and the reference length (Len_{gen}/Len_{ref}) of each model. LR is contrasted with the other metrics and human evaluation scores.

- **RQ3: Distinguishing Elaboration from Error.** To what extent do current evaluation metrics conflate informational density with factual errors? To address RQ3, we quantified the extent of model elaboration using LR. Since the reference represents the ideal "short sentence" explicitly requested in the prompt, an LR significantly greater than 1.0 indicates *verbosity*, a trait that overlap-based metrics typically penalize as noise, but which models may employ to provide necessary context.

- **RQ4: Evaluator Bias.** Does the LLM-as-a-judge exhibit a "self-preference bias"? To answer RQ4, we investigate whether the LLM judge aligns more closely with the model outputs generated by the same model used for the judge, potentially inflating scores for text that mirrors its own training distribution. For this, we calculate the **Delta Bias (DB)**, defined as the arithmetic difference between the LLM-judge score and the human score ($S_{\text{LLM}} - S_{\text{Human}}$). We also experimented with disclosing the names of the models along with the predictions, to see if the judge demonstrates a bias in favor of themselves.

5 Results

The results obtained in this study provide insights into the evaluation of language models in QA tasks for Portuguese, and address our research questions defined in Section 4.

5.1 *RQ1: To what extent do lexical overlap metrics and semantic evaluators correlate with human evaluations?*

Table 1 presents the results of the automatic metrics. One of the most significant findings of this work is the substantial divergence between automatic metrics and human evaluation shown in Table 2. While GPT-4o showed clear superiority in ROUGE and BERTScore metrics, Sabiá-3 consistently received higher human evaluations across all qualitative dimensions, as shown in Table 3.

Model	R1	R2	RL	BL	MT	BP	BR	BF
Sabiá-3	.43	.27	.39	.11	.46	.75	.80	.78
GPT-4o	**.53**	**.36**	**.50**	.12	.45	**.82**	.80	**.81**
Gemini	.48	.32	.45	**.13**	**.48**	.78	.80	.79
Deepseek	.50	.33	.46	.11	**.48**	.79	**.81**	.80

Table 1: Results obtained on the QA task. The evaluated metrics were ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BLEU (BL), METEOR (MT), BERTScore Precision (BP), Recall (BR), and F1 (BF).

Metric	Corr	Compl	Rel	Clar
LLM	***0.15**	****0.38**	**0.12**	-
R1	0.05	0.01	-0.00	0.00
R2	-0.01	0.05	-0.06	-0.05
RL	0.04	0.01	-0.03	-0.01
BL	-0.04	-0.04	-0.06	-0.11
MT	0.04	0.11	-0.05	0.01
BP	-0.00	-0.15	-0.01	-0.08
BR	0.07	0.08	0.04	**0.05**
BF	0.03	-0.06	0.01	-0.02

Table 2: Spearman's ρ correlation amid Human Evaluation and LLM-as-a-judge (LLM), ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BLEU (BL), METEOR (MT), BERTScore Precision (BP), Recall (BR), and F1 (BF). Significance levels are indicated as * $p < 0.05$ and *** $p < 0.001$; all remaining correlations are non-significant ($p > 0.05$). Clarity could not be assessed for the LLM judge, as it assigned a constant score of 5 to all instances.

This discrepancy suggests that automatic metrics, although widely used in the literature, may not adequately capture crucial qualitative aspects of answers. Table 2 reveals a critical lack of alignment between lexical metrics and human judgment. Specifically, metrics such as ROUGE-L, BLEU, METEOR, and even BERTScore exhibit correlations close to zero. Statistically, these values suggest that an increase in lexical overlap or topical similarity with the reference answer has virtually no predictive power regarding the answer's factual correctness or clarity as perceived by humans. This challenges the validity of using these metrics as a standard for quality in Portuguese open-ended QA, as a high score in these metrics is just as likely to correspond to a low-quality answer as a high-quality one, from the human perspective.

Notably, the only moderate correlation observed in our experiments was between the LLM-as-a-Judge and human ratings for Completeness. This

suggests that the LLM judge is better aligned with humans in detecting information density and coverage than in assessing factual precision (Correctness) or Clarity. This finding aligns with recent literature suggesting that LLM judges, like humans, may possess a bias toward longer, more detailed responses (Hu et al., 2024), creating a natural correlation in the 'Completeness' criteria. Clarity could not be evaluated because the LLM judge attributed a score of 5 to all instances.

It was also observed that the LLM-as-a-Judge tends to be more lenient than the human evaluators. While human annotators utilized the full range of the Likert scale (1-5) to penalize inaccuracies, the GPT-4o judge exhibited a distinct tendency toward leniency, concentrating scores above three. Consequently, the average correctness score assigned by the LLM (μ=4.93) was higher than that of human evaluators (μ=4.63). In contrast, when assessing Completeness, the average scores assigned by the GPT-4o judge were slightly lower than those given by humans; however, this difference was considerably less pronounced than the leniency observed in the other evaluation criteria.

Judge	Model	Corr	Compl	Rel	Clar
Human	**Sabiá-3**	**4.74**	**4.70**	**4.84**	**4.87**
	GPT-4o	4.57	3.95	4.63	4.65
	Gemini	4.46	4.14	4.64	4.65
	Deepseek	4.71	4.37	4.71	4.75
LLM	**Sabiá-3**	**5.00**	**4.20**	**5.00**	5.00
	GPT-4o	4.96	3.90	4.96	5.00
	Gemini	4.98	3.98	4.98	5.00
	Deepseek	4.94	4.04	4.98	5.00

Table 3: Results obtained from human QA evaluation of the outputs generated by the LLMs. The reported values correspond to the simple average across the two evaluation sets for each model.

5.2 *RQ2: How does the trade-off between strict prompt adherence and information density affect model rankings?*

The trade-off between strict prompt adherence and information density substantially alters model rankings. The correlations in Table 2 show that automatic metrics tend to have weak, inconsistent, or even negative correlations with the criteria. It suggests that these metrics implicitly favor lexically aligned outputs, even when those answers are less informative or poorly communicable. In contrast, as shown in Figure 2, there is a correlation indicating that human judges are more sensitive to information density and semantic adequacy, rewarding answers that go beyond prompt compliance when the additional information improves relevance or completeness. As a result, such models may be penalized by automatic metrics but favored by humans.

The superior performance of Sabiá-3 in human evaluation, particularly in terms of completeness and clarity, suggests that models specifically developed for Portuguese can offer significant qualitative advantages in QA tasks. However, the discrepancy between automatic metrics and human evaluation may also stem from differences in how the models interpreted and followed the prompt: while GPT-4o may have aligned more closely with the expected format and structure defined by the prompt, resulting in higher scores from automatic metrics based on lexical overlap, Sabiá-3 often generated more robust and semantically enriched answers that, although potentially less aligned in surface form, were better received by human annotators for their informativeness and contextual adequacy.

The qualitative analysis illustrated in Table 4 shows that Sabiá-3 tends to produce more elaborate and contextually rich answers ("*The way to avoid undesired flow during offshore drilling is by using safety equipment placed inside the blowout preventer (BOP)*"), whereas other models provide more concise yet potentially less informative responses. This characteristic may explain the higher scores observed in the human evaluation. Additionally, GPT-4o maintained its lead in automatic metrics in Table 1, but with narrower margins, while human evaluation showed more homogeneous performance across models.

5.3 *RQ3: To what extent do current evaluation metrics conflate informational density with factual errors?*

N-gram-based metrics exhibit a structural limitation when evaluating information-dense responses, as they implicitly conflate semantic enrichment with hallucination. As these metrics are calculated based strictly on lexical overlap, models that attempt to provide additional contextualization detail may receive lower scores, even when the added information is factually correct and beneficial for human understanding.

The example below illustrates an error observed in the qualitative analysis:

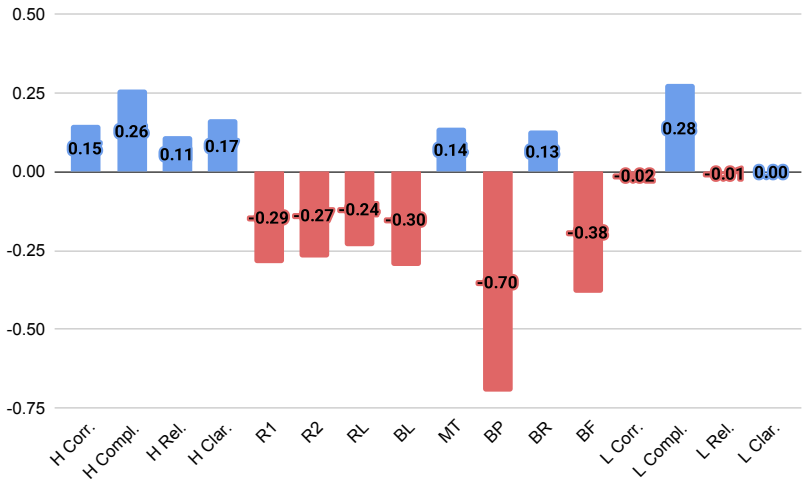

Figure 2: Spearman correlation between the Length Ratio (density) and the metrics, Human evaluation (H), ROUGE-1 (R1), ROUGE-2 (R2), ROUGE-L (RL), BLEU (BL), METEOR (MT), BERTScore Precision (BP), Recall (BR), and F1 (BF), and LLM-as-a-Judge (L)). Negative values indicate penalization of long responses; positive values indicate reward.

Question: *Where is the well located that received the first fully electric offshore installation of a smart completion system?*
Generated answer: *The text does not specify where the first fully electric offshore installation was located.*
Ground truth answer: *In Brazil, in Mossoró, in the state of Rio Grande do Norte.*

Human annotators consistently assigned low scores, correctly penalizing the factual error. However, semantic similarity metrics such as BERTScore assigned comparatively high scores (F1 $\approx$ 0.59), reflecting lexical and semantic alignment rather than factual accuracy. Moreover, the LLM-as-a-Judge assigns maximum scores across all criteria, including correctness. This example demonstrates that both automatic metrics and model-based evaluation may reward "well-formed" responses even when they contain critical factual errors, reinforcing the quantitative findings reported in Figure 2.

5.4 *RQ4: Does the LLM-as-a-judge exhibit a "self-preference bias"?*

Quantitative analysis of the DB reveals a systematic discrepancy between automated and human evaluation, corroborating the hypothesis of a "leniency bias" in the artificial judge rather than specific self-preference. We observed that global models, Gemini and GPT-4o, accumulated the highest bias sums (25.67, and 17.33, respectively). In contrast, Sabiá-3 and Deepseek (9.00, and 10.32, respectively) presented the lowest accumulated delta, demonstrating a higher calibration between human and artificial judgments for this model. These findings suggest that the disparity is not driven by the judge favoring its own identity ('self-preference bias'), but by a general tendency of the LLM to overestimate the quality of outputs—a leniency that human judges do not share.

This interpretation is further supported by the secondary experiment, where model names were explicitly provided. The resulting scores remained consistent with the original setup, indicating that model identification does not substantially alter the judge's preferences. Notably, this condition led to a broader use of the Likert scale, now including lower scores. Overall, these findings indicate that, under the evaluated conditions, the LLM-as-a-Judge does not exhibit a measurable self-preference bias.

6 Conclusion

This work investigated the reliability of automatic evaluation metrics and LLM-based judges in open-ended QA for Portuguese, focusing on their alignment with human judgment. As we evaluated lexical and semantic similarity metrics, human evaluation, and an LLM-as-a-judge, we demonstrated that widely adopted automatic metrics remain weak for human-perceived quality in this context.

Our results revealed a substantial alignment gap between automatic metrics and human evaluation. Metrics such as ROUGE, BLEU, METEOR, and BERTScore exhibited near-zero correlation with human judgments across qualitative dimensions.

Model	Generated Answer
Sabiá-3	A maneira de evitar o fluxo indesejado durante a perfuração em alto mar é utilizando equipamento de segurança colocado dentro da pilha de impedimento (BOP).
GPT-4o	Colocar equipamento de segurança dentro da pilha de impedimento (BOP).
Gemini 2FL	Equipamento de segurança é colocado dentro da pilha de impedimento (BOP).
Deepseek	Usar equipamento de segurança colocado dentro da pilha de impedimento (BOP) para manter o poço sob controle.
GT	requer equipamento de segurança colocado dentro da pilha do preventor de explosão (BOP) para manter o poço sob controle, evitar fluxo indesejado e proteger o meio ambiente e o pessoal.

Table 4: Answers generated by the language models, and the Ground Truth (GT).

While these metrics offer scalability, they may overlook essential qualitative dimensions that are critical to assessing the quality of generated answers.

Future research could explore the development of automatic metrics that are more sensitive to the characteristics of language. Additionally, the analysis could be expanded to include different domains, as well as exploring ensemble techniques that combine the strengths of different models.

Limitations

This study has a few limitations that should be considered when interpreting the results. First, the human evaluation was conducted on a sample of 50 instances, constrained by the cost of annotation. While this sample size is sufficient to reveal systematic patterns such as the near-zero correlations between automatic metrics and human judgment, it limits the statistical power of the analysis. Future work should replicate the study at larger scale to further consolidate these findings.

Second, the experiments were conducted on a single dataset covering a specific domain. Although domain-specific vocabulary could favor lexical metrics, they still failed to align with human judgments, suggesting the observed limitations are not an artifact of the domain choice. Nevertheless, replication across different domains and text genres would strengthen the generalizability of the conclusions.

Furthermore, although guidelines were provided to annotators, the intermediate Likert scale values (2, 3, 4) were not explicitly anchored, which may have introduced subjectivity in how individual annotators interpreted the scale boundaries.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, CNPq, and Cenpes/Petrobras.

References

Ali Mohamed Nabil Allam and Mohamed Hassan Haggag. 2012. The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3).

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Luciana Bencke, Felipe Paula, Bruno dos Santos, and Viviane P. Moreira. 2024. Can we trust llms as relevance judges? In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 600–612, Porto Alegre, RS, Brasil. SBC.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Kathrin Blagec, Georg Dorffner, Milad Moradi, Simon Ott, and Matthias Samwald. 2022. A global analysis of metrics used for measuring performance in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 52–63.

Kathrin Blagec, Georg Dorffner, Milad Moradi, and Matthias Samwald. 2020. A critical analysis of metrics used for measuring progress in artificial intelligence. *arXiv preprint arXiv:2008.02577*.

Shweta Chauhan and Philemon Daniel. 2023. A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Processing Letters*, 55(9):12663–12717.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao

Yang, and Ying Qin. 2022. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into automatic evaluation using large language models.

Ensieh Davoodijam and Mohsen Alambardar Meybodi. 2024. Evaluation metrics on text summarization: comprehensive survey. *Knowledge and Information Systems*, 66(12):7717–7738.

Hugo APG de Camargo, Pedro Henrique Paiola, Gabriel Lino Garcia, and João Paulo Papa. 2025. Abstractive summarization with llms for texts in brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):1030–1048.

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

Allan Duarte Ehlert, Júlia da Rocha Junqueira, Larissa Astrogildo de Freitas, and Ulisses Brisolara Corrêa. 2025. A review of recent advances on automatic question generation for the portuguese language. In *The International FLAIRS Conference Proceedings*, volume 38.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. 2024. Explaining length bias in llm-based preference evaluations. *arXiv preprint arXiv:2407.01085*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2122–2132.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2022. Rquge: Reference-free metric for evaluating question generation by answering the question. *arXiv preprint arXiv:2211.01482*.

Preksha Nema and Mitesh M Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959.

Huyen Nguyen, Haihua Chen, Lavanya Pobbathi, and Junhua Ding. 2024. A comparative study of quality evaluation methods for text summarization. *CoRR*.

Iftitahu Nimah, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1240–1266, Toronto, Canada. Association for Computational Linguistics.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *2017 Conference on Empirical Methods in Natural Language Processing*, pages 2231–2242. Association for Computational Linguistics.

Amanda Oliveira, Thiago Cecote, Pedro Silva, Jadson Gertrudes, Vander Freitas, and Eduardo Luz. 2023. How good is chatgpt for detecting hate speech in portuguese? In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 94–103, Porto Alegre, RS, Brasil. SBC.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Felipe Paula, Matheus Westhelle, Maria Corrêa, Luciana Bencke, and Viviane Moreira. 2025. How faithful are your summaries? a study of nli-based verification in portuguese. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 307–322, Porto Alegre, RS, Brasil. SBC.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.

Paulo Pirozelli, Marcos M José, Igor Silveira, Flávio Nakasato, Sarajane M Peres, Anarosa AF Brandão, Anna HR Costa, and Fabio G Cozman. 2024. Benchmarks for pirá 2.0, a reading comprehension dataset about the ocean, the brazilian coast, and climate change. *Data Intelligence*, 6(1):29–63.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Stephan Vogel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system. In *Proceedings of LREC*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Appendix

Context:
`{context}`
Question:
`{question}`
Answer:
`{answer}`

Evaluate the answer from 1 to 5 for each criterion below, responding only with the numbers separated by commas, in the order: Correctness, Completeness, Relevance, Clarity.

Criteria:

- **Correctness:** Is the answer factually correct according to the context? False or misleading information should be penalized.
- **Completeness:** Does the answer cover all relevant information? Partial or omitted answers may lose points.
- **Relevance:** Does the answer truly address the question asked, without deviating from the topic? Answers that sidestep the subject or are too generic should be penalized.
- **Clarity:** Is the answer easy to understand, without ambiguity or vague terms? Grammar and organization of the answer may be considered.

Example of response:
`5,4,5,5`