A Comparison of Commonly Used Automatic Evaluation Metrics for Open-Ended Tasks in Portuguese

Eduardo D. Faé, Dennis Giovani Balreira, Viviane Pereira Moreira

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

{edfae,dgbalreira,viviane}@inf.ufrgs.br

Abstract. *Open-ended tasks in Natural Language processing are characterized by the absence of a fixed set of outputs or short, predetermined responses. Typical examples include open-domain question answering and summarization. The automatic evaluation of these tasks is challenging, and existing metrics suffer from important limitations. Thus, many works have already been proposed to evaluate the performance of such metrics. However, the great majority of these works were conducted in English, leaving a gap for analyses in other languages, such as Portuguese. This work investigates the performance of some of the most popular automated intrinsic evaluation metrics for open-ended tasks by analyzing their correlation with human judgment. Using both the Summarization and Question Answering tasks, this study compares traditional n-gram-based metrics with metrics based on contextual embeddings generated by Pre-trained Language Models (PLMs), performing all analyses using models and datasets available for Portuguese. Results indicate that while PLM-based metrics generally show slightly higher correlation with human perception, their performance is highly reliant on the quality of the models used. Still, none of the evaluated metrics displayed a strong correlation with human judgments, suggesting the need for more robust metrics.*

1. Introduction

Open-ended tasks are problems designed so that they may contain more than one correct solution or be solved in more than one unique way. These tasks typically allow for a diversity of valid outputs and interpretations and are commonly present in areas that require creativity and critical thinking. They are extremely common in Natural Language Processing (NLP), as natural languages are inherently dynamic, volatile, diverse, and context-dependent.

These tasks can be found everywhere, from translating a sentence to writing an article. In NLP, two common open-ended tasks are summarization and question answering (QA). Both tasks recently reached new levels of importance, with the diffusion of Large Language Models (LLMs) and popularization of newer Information Retrieval (IR) techniques, such as RAG [Lewis et al. 2020]. However, like many other open-ended tasks, they still share common problems in terms of evaluation.

In general, open-ended tasks are significantly more complex to verify and evaluate than closed-ended ones. This is a result of their unique ability to produce multiple semantically correct outputs for a single input, which poses challenges when analyzing model performance. Many metrics have been proposed to address this issue, some of which have

become widely used as comparison tools in the NLP field. Traditional metrics generally follow the concept of n-grams [Shannon 1948] as a means of evaluating an output given a ground-truth, while newer metrics tend to rely on Pre-trained Language Models (PLMs) to try capturing the semantic similarity between a generated text and an expected answer.

For such, many works have sought to compare the performance of different metrics and examine their correlation with human evaluation [Liu et al. 2016, Deutsch et al. 2022, Liu et al. 2023]. With the translation task in particular having a great start, given the promotion of multiple large-scale evaluation campaigns aimed to measure the performance of early-adopted metrics, such as [NIST 2006, Koehn and Monz 2006]. However, when it comes to summarization, and especially QA, the number of works becomes a lot smaller [Fabbri et al. 2021, Junqueira and Moreira 2026]. Furthermore, most existing studies focus primarily on English datasets and models, leaving other languages, such as Portuguese, comparatively underexplored.

This work aims to perform these correlation analyses with models and datasets available for Portuguese. Still, this work focuses on two distinct types of open-ended tasks, summarization and QA, the latter being a task for which the literature remains comparatively limited. Additionally, we integrate real human evaluators to avoid introducing bias via tools like LLM-as-a-judge.

To achieve this, we collected evaluations from real human judges. With each of the two tasks having 40 generated answers, evaluated by four distinct judges, in four quality dimensions. This accounted for 160 evaluations per quality dimension for each task, totaling 1,280 human evaluations. The results showed that PLM-based metrics present slightly higher correlations with human judgments than n-gram-based ones. However, they appear to be extremely reliant on the quality of the PLM used. Still, both types of metrics failed to maintain high degrees of correlation with human judgments.

2. Background

Abstractive Approach. When it comes to summarization and QA, approaches generally fall into two categories: extractive and abstractive. While extractive approaches select and reproduce portions of the original text, abstractive approaches generate new text that may paraphrase, reorganize, or synthesize information from the source. In general, the abstractive approach is the more open-ended of the two, as it allows for a more diverse set of answers. Thus, will be the preferred approach for this work.

Intrinsic Evaluation. Evaluation methods in NLP can generally be divided into intrinsic and extrinsic approaches. Intrinsic evaluation assesses a model directly on its intended task using task-specific metrics, whereas extrinsic evaluation assesses the model by measuring its impact on the performance of a downstream application or end task. Performing intrinsic evaluation in open-ended tasks is particularly challenging, given the nature of such tasks, in which an output can be partially correct, and many distinct outputs can be correct, allowing for variations in writing styles, lexical choices, semantic paraphrases, textual structures, or levels of detail. Which lead to the proposal of a plethora of automated metrics, such as ROUGE [Lin 2004] and BERTScore [Zhang et al. 2019].

N-gram-based Metrics. n-gram-based metrics evaluate generated texts by measuring the overlap of contiguous word sequences between candidate and reference outputs. These

methods assume that higher lexical similarity indicates better generation quality, making them computationally efficient and easy to reproduce. However, in open-ended tasks, semantically correct answers may be expressed using different vocabulary or sentence structures, which limits the ability of n-gram-based approaches to fully capture meaning and contextual adequacy. Despite these limitations, such metrics remain widely adopted for evaluating NLP systems, with BLEU [Papineni et al. 2002], ROUGE [Lin 2004], and METEOR [Banerjee and Lavie 2005] among the most commonly used approaches.

PLM-based Metrics. PLM-based metrics evaluate generated texts by leveraging contextualized semantic representations obtained from language models. These metrics can capture semantic similarity even when the candidate and reference texts use different lexical choices or syntactic structures. As a result, PLM-based metrics generally achieve stronger correlations with human judgments in open-ended tasks. Nevertheless, they incur higher computational costs and still struggle to assess factual consistency or reasoning quality. Among the most prominent PLM-based evaluation metrics are MoverScore [Zhao et al. 2019], BERTScore [Zhang et al. 2019], and BARTScore [Yuan et al. 2021].

3. Related Work

Several studies have investigated the correlation between automatic evaluation metrics and human judgments for open-ended tasks. Back in 2005, the creators of METEOR [Banerjee and Lavie 2005] demonstrated that their technique achieved higher correlation with human judgments than BLEU [Papineni et al. 2002]. In the following years, large-scale evaluation campaigns, such as those conducted by NIST [NIST 2006] and WMT [Koehn and Monz 2006], began systematically comparing the most prominent metrics in machine translation against collected human annotations. This led to the popularization of these metrics, promoting their adoption for other open-ended NLP tasks as better automated evaluation methods.

Around 2015, analyses of these metrics in other open-ended tasks, such as summarization and dialogue generation, began to emerge. However, studies from this time, such as [Liu et al. 2016], still demonstrated limited correlation between traditional metrics and human evaluation, reporting extremely low results. Recent studies still show medium to low correlations between automated metrics and human judgments [Deutsch et al. 2021, Fabbri et al. 2021, Deutsch et al. 2022, Liu et al. 2023, Junqueira and Moreira 2026].

For the Summarization field, one of the most important studies was SummEval [Fabbri et al. 2021], which introduced a benchmark designed to investigate the correlation between automatic summary evaluation metrics and human judgment. In their analysis, they divided the evaluation of each summary into four quality dimensions: coherence, fluency, relevance, and consistency, aiming to capture different aspects that affect the quality of a summary. The authors tested a wide variety of metrics and pointed out that metrics based on n-grams, such as BLEU and ROUGE [Lin 2004], show low correlation with human perception, whereas PLM-based metrics present better results but still fall short of fully capturing the nuances and aspects present in human evaluation.

In contrast, comparatively fewer studies have investigated automatic evaluation for QA. Notable examples include: [Krishna et al. 2021], who demonstrated that the ROUGE-L metric provides limited insight for this type of task, as it can be artificially optimized without reflecting true answer quality; [Farea et al. 2025], which presents a broader

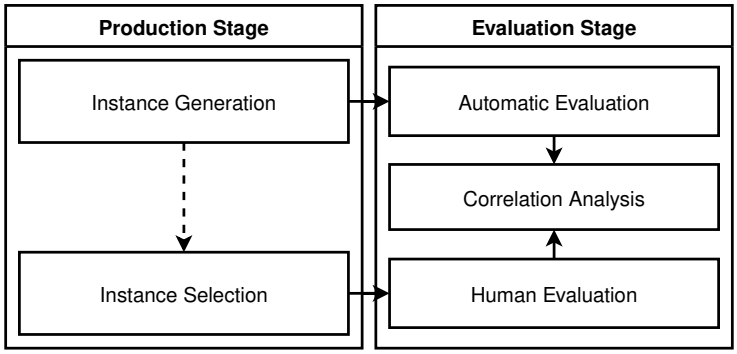
analysis of existing QA systems; and [Junqueira and Moreira 2026], which demonstrated how lexical metrics are limited in evaluating QA for Portuguese. The latter of which is among the few studies that analyzed the correlation between automatic metrics and human judgments in QA for non-English languages.

This work aims to expand the literature by addressing two main limitations: the lack of Portuguese-focused studies on the reliability of automatic intrinsic evaluation metrics for open-ended tasks and the lack of QA-focused analyses of metric correlations. Still, unlike recent studies such as [Junqueira and Moreira 2026], we focus on introducing a unified multi-task methodology spanning both summarization and QA, which enables a better analysis of the current capacities of automatic intrinsic evaluation metrics for open-ended tasks in Portuguese.

4. Methodology

This work is divided into five stages to better organize and present the methodology and its execution. Further explanations about each stage are provided throughout this section. Figure 1 provides an overview of this methodology. During the production stage, solutions are generated and selected; in the evaluation stage, solutions are evaluated by automated metrics and human judges, enabling a correlation analysis.

Figure 1. Overview of the methodology.



4.1. Instance Generation

To facilitate test replication and avoid bias in the use of an authorial dataset, both models and datasets were sourced from public repositories. Described below are the models and datasets chosen for the purposes of this work.

Models. When analyzing the correlation between a metric and human evaluation, it is essential to have a wide range of good and bad outputs. For this, we opted for smaller models, as they tend to be more unstable, better encompassing the desired quality range.

For Summarization, the model used was the Portuguese T5 for Abstractive Summarization [Paiola et al. 2022], with the fine-tuned version on the XL-Sum dataset [Hasan et al. 2021] being chosen. It was selected after a round of testing for being a small model that originated in an article published at a renowned conference.

In our search for small Portuguese models for the QA task, we failed to find open models that achieved a minimum standard and came from reliable sources. Thus,

for QA, the model selected was the smallest instruction-tuned version of Gemma 3 [Team et al. 2025]. This fulfills our goal of being a small model and, while not being primarily developed with Portuguese in mind, still has native support for Portuguese.

Datasets. The predominant language in NLP datasets is English; therefore, finding well-cataloged and curated datasets for languages other than English is a difficult task. It was no different for this work; with most of the tested datasets presenting problems, whether due to errors stemming from direct translation or poorly designed references. After several tests, the datasets selected are described below.

For Summarization, the dataset used was the CNN Dailymail Azure Pt[1]. This is a translated version of the CNN/DailyMail dataset [Nallapati et al. 2016], which provides a catalog of news articles from the media outlet, accompanied by written summaries used as references for calculating metrics. This dataset was chosen because it is a translation of one of the datasets used by SummEval [Fabbri et al. 2021], enabling a better comparison of the results in this work with those reported by SummEval.

For QA, the dataset used was FairytaleQA Translated Pt [Leite et al. 2024]. This dataset was created from the automatic translation of the open-source FairytaleQA dataset [Xu et al. 2022], which contains a compilation of short fairy tales and questions related to each tale, accompanied by their expected answers. The dataset was chosen for its ludic tone and focus on the abstractive approach, which led to a decrease in directly extracted answers when compared to most others, which tend to follow the SQuAD [Rajpurkar et al. 2016] pattern and focus on the extractive approach.

4.2. Automatic Evaluation

As a means of automatic evaluation, three of the most influential metrics from two of the most well-established approaches used for automatic intrinsic evaluation were selected. Table 1 presents each metric, the approach they are based on, and their year of creation.

Table 1. Description of the metrics evaluated by this work.

Metric	*Based on*	*Created in*
BLEU	n-grams	2002
ROUGE	n-grams	2004
METEOR	n-grams	2005
BERTScore	PLMs	2019
MoverScore	PLMs	2019
BARTScore	PLMs	2021

PLM-based metrics rely on strong Portuguese models to achieve good results. For this, the large version of BERTimbau [Souza et al. 2020] was the selected model for both BERTScore and MoverScore, as it was the most powerful Portuguese BERT model readily available. As for BARTScore, there are far fewer models designed for Portuguese. We opted for BART Base Portuguese[2], as it is pre-trained in Portuguese.

In addition to standard BLEU, BLEU-1 and BLEU-2, which use unigrams and bigrams, respectively, were analyzed. As for ROUGE, all ROUGE-1, ROUGE-2, and

[1]Dataset: `huggingface.co/datasets/arubenruben/cnn_dailymail_azure_pt_pt`
[2]Model: `https://huggingface.co/adalbertojunior/bart-base-portuguese`

ROUGE-L were used. For BERTScore, precision, recall, and f1-score values were provided. For more information about each metric, refer to our Github repository[3].

4.3. Instance Selection

The selection of instances for human evaluation was performed at random to avoid bias. For each task, a sample of 40 distinct instances was selected and presented to participants in the human evaluation, as described in the next section.

4.4. Human Evaluation

We used human evaluations to assess metric performance, correlating each score obtained through an automated metric with the scores from the human judgment. During data collection, the four quality dimensions used by SummEval [Fabbri et al. 2021] were employed as to better encompass output quality.

1. Coherence: General quality of the text.
2. Consistency: Alignment between the output and the source text.
3. Fluency: Quality of the generated sentences.
4. Relevance: Key points selected for the text.

However, in the scope of this work, the fluency dimension was replaced by a naturalness one. This change was made because the human evaluations were not conducted by specialists, which impairs the precise analysis of text fluency, as a great amount of linguistic knowledge is required for its evaluation. Naturalness was chosen for its similarity and relative compatibility with fluency, while being easier for laypeople to evaluate.

There are no consolidated quality dimensions designed for the QA evaluation. For this reason, we adopted the same ones that were used for Summarization. We understand that not all dimensions translate directly to this task, but we needed a baseline for the evaluations and considered that maintaining a standard would be a better approach than proposing a distinction without the specific knowledge of an expert.

Data Collection. Human evaluations were collected through two online forms, each focused on one of the tasks. The 40 outputs of each task were divided into two groups and then presented to eight people, leaving four judges for each of the 80 generated texts. The judges were asked to evaluate each of the outputs in the four aforementioned quality dimensions by scoring them on a Likert scale from 1 to 5. To calculate the final score of an output, a simple average of the collected scores was performed.

Participant Profiles. The eight participants were all Computer Science students, aged 21 to 30, and of mixed gender. This makes them knowledgeable in the general functioning of LLMs and familiar with the structure of online forms. However, as the judges were not Linguistic specialists, some nuances may have been lost during data collection.

4.5. Correlation Analysis

To evaluate the alignment between the automated metrics and human evaluation, we adopted both Pearson and Spearman correlations, as they are suitable for continuous data and are widely employed by the literature. Thus, being reliable tools for measuring the

[3]Github: `https://github.com/eduardofae/OET-Metric-Comparison-PT`

correspondence between the metric scores and the collected human data. The discretization of the values adopted for the analysis of this work followed [Schober et al. 2018]. To analyze inter-rater agreement, we calculated Krippendorff's α [Krippendorff 2011] between reviewers, both following the analyzed dimensions and aggregating all scores for each annotator, allowing for dimension-specific and global inter-rater agreement analysis.

5. Results

Table 2 presents the average scores obtained by each metric in each task. The final four rows correspond to the average scores assigned by human evaluation (HE), divided into the four aforementioned quality dimensions. Furthermore, as a means of inter-rater agreement analysis, Krippendorff's Alpha [Krippendorff 2011] between judges is also provided.

Table 2. Average value assigned by each form of evaluation for each task.

	Summarization		QA	
Metric	Avg Score	Krippendorff's α	Avg Score	Krippendorff's α
ROUGE-1	0.22	-	0.34	-
ROUGE-2	0.05	-	0.22	-
ROUGE-L	0.19	-	0.33	-
BLEU	0.01	-	0.18	-
BLEU-1	0.23	-	0.34	-
BLEU-2	0.09	-	0.25	-
METEOR	0.18	-	0.41	-
BERTScore-p	0.66	-	0.75	-
BERTScore-r	0.65	-	0.76	-
BERTScore-f1	0.65	-	0.76	-
BARTScore	-4.25	-	-5.48	-
MoverScore	0.54	-	0.63	-
Consistency (HE)	2.39	0.46	3.44	0.56
Naturalness (HE)	3.58	0.44	4.06	0.43
Relevance (HE)	2.46	0.37	3.48	0.58
Coherence (HE)	2.59	0.37	3.55	0.56

The results indicate substantially lower human evaluation scores for consistency, relevance, and coherence in the summarization task when compared to QA. This behavior may be associated with differences in task complexity, dataset quality, or model capability. However, as this work focuses on the effectiveness of evaluation metrics rather than model performance itself, lower-performing models may even be advantageous, as they produce a wider diversity of both high and low-quality outputs, enabling a better view of metric behavior across different quality levels.

When evaluating the performance of the automatic metrics, although average scores may provide an overview of their raw evaluations, they are insufficient to determine which metrics better reflect human judgment, as each metric operates under distinct scoring scales and evaluation principles. Therefore, correlation analyses are required for a more reliable assessment of the alignment between each automatic metric and human evaluation. These analyses are presented in the following subsections, with results for each task being provided individually.

When it comes to inter-rater agreement, we see relatively low α values for both tasks, with combined values of 0.47 for Summarization and 0.54 for QA. This is a consequence of the open-ended nature of both tasks, as well as the reliance on non-specialized annotators. QA, the more objective of the two tasks, consistently yields higher values than Summarization, which is the broader task. Furthermore, the Naturalness dimension remained constant across tasks, as it is a task-independent dimension, while still exhibiting low values, as it is nevertheless highly subjective. The α values obtained for summarization closely match the ones reported by [Fabbri et al. 2021] from crowd-source workers, signaling the importance of expert reviewers.

5.1. Summarization

Table 3 presents Pearson (ρ) and Spearman (ρ_s) correlation values between each metric and human judgment for the summarization task. The results show relatively weak correlations between metrics and human evaluation. Contrary to our expectations, PLM-based metrics did not outperform n-gram-based ones, with BARTScore [Yuan et al. 2021] particularly underperforming. However, the metric that achieved the highest overall alignment was still a PLM-based one, being it MoverScore [Zhao et al. 2019].

Table 3. Pearson (ρ) and Spearman (ρ_s) correlations between human perception and scores of automatic metrics for the Summarization task.

	Consistency		Naturalness		Relevance		Coherence	
Metric	ρ	ρ_s	ρ	ρ_s	ρ	ρ_s	ρ	ρ_s
ROUGE-1	0.21	0.19	0.29	0.24	0.29	0.27	0.24	0.21
ROUGE-2	0.18	0.07	0.27	0.18	0.23	0.16	0.24	0.13
ROUGE-L	0.28	0.24	0.36	0.31	**0.38**	**0.35**	0.35	0.32
BLEU	0.28	0.23	0.35	0.33	0.33	0.34	0.35	0.33
BLEU-1	0.31	0.28	0.28	0.22	0.28	0.23	0.27	0.22
BLEU-2	0.27	0.22	0.27	0.20	0.25	0.22	0.27	0.20
METEOR	0.32	0.32	**0.37**	**0.35**	0.27	0.31	0.33	0.30
BERTScore-p	0.26	0.21	**0.37**	0.30	0.29	0.23	0.36	0.26
BERTScore-r	0.31	0.24	0.32	0.26	0.29	0.23	0.32	0.22
BERTScore-f1	0.30	0.22	**0.37**	0.30	0.30	0.24	0.36	0.25
BARTScore	0.05	0.04	0.30	0.34	0.15	0.16	0.24	0.25
MoverScore	**0.38**	**0.37**	0.31	0.24	0.34	0.34	**0.39**	**0.36**

Similar findings were previously reported by SummEval [Fabbri et al. 2021] and related studies, reinforcing the difficulty of intrinsically evaluating abstractive summarization systems. Still, the correlations obtained in this work were generally lower than those reported by previous studies. Which may be associated with limitations in the dataset quality, as automatic metrics fundamentally depend on reliable reference texts.

Table 4 shows limitations of the automatic metrics observed in the summarization task. In the first example, the generated summary conveys the central meaning of the source content, resulting in a high human score, yet all metrics, especially n-gram-based ones, assigned relatively low scores. In contrast, the second example entirely cuts out important details, such as the death of a band member and how this is the band's last tour, which leads to a low human score. However, it was better evaluated by all automatic

metrics, except MoverScore, which still scored them close to equally. Indicating that metrics may capture topical similarity more effectively than factual faithfulness.

Table 4. Examples illustrating failures of automatic metrics in summarization.

Example 1 (High-quality Summary)	**Example 2 (Low-quality Summary)**
Summary: Iraqi forces are being sent to central Iraq to help fight the extremist group calling itself the Islamic State (IS). **Human Score:** 4.5/5	**Summary:** The band Twisted Sister disclosed on Wednesday that they'll be leaving the band in the next two months, after a three-year tour. **Human Score:** 1.6/5

5.2. Question Answering

Unlike the behavior observed in the summarization task, the evaluated metrics achieved marginally higher correlation scores in the QA task, as seen in Table 5. This behavior may be explained by the task's underlying characteristics, which, even following an abstractive approach, remain closer to closed-ended settings than summarization, as the outputs are typically shorter and frequently contain terms directly extracted from the source context. Consequently, metrics based on text comparison tend to perform better, as reference texts and outputs are more inclined to share a similar structure and terms.

Table 5. Pearson (ρ) and Spearman (ρ_s) correlations between human perception and scores of automatic metrics for the QA task.

	Consistency		Naturalness		Relevance		Coherence	
Metric	ρ	ρ_s	ρ	ρ_s	ρ	ρ_s	ρ	ρ_s
ROUGE-1	0.58	0.59	0.45	0.54	0.56	0.56	0.55	0.54
ROUGE-2	0.50	0.57	0.33	0.52	0.51	0.53	0.49	0.54
ROUGE-L	0.56	0.58	0.44	0.53	0.55	0.55	0.54	0.52
BLEU	0.43	0.49	0.24	0.31	0.44	0.48	0.43	0.48
BLEU-1	0.51	0.54	0.36	0.50	0.50	0.51	0.49	0.48
BLEU-2	0.53	0.59	0.37	0.55	0.53	0.56	0.52	0.57
METEOR	0.64	0.68	**0.53**	**0.65**	0.63	0.65	0.62	0.61
BERTScore-p	0.63	0.65	0.42	0.47	0.62	0.64	0.60	0.58
BERTScore-r	**0.70**	**0.76**	0.51	0.62	**0.69**	**0.74**	**0.67**	**0.71**
BERTScore-f1	0.68	0.73	0.47	0.58	0.67	0.71	0.65	0.67
BARTScore	0.47	0.41	0.35	0.39	0.47	0.39	0.47	0.41
MoverScore	0.47	0.67	0.28	0.51	0.48	0.65	0.46	0.61

Despite the comparatively better performances, metrics only achieved an overall medium correlation, with BERTScore [Zhang et al. 2019] scraping the lower bounds of a high correlation in some variant/dimension pairs. This result reinforces the limitations of current intrinsic evaluation methods, even in tasks with shorter and more lexically constrained outputs. Similar findings were also reported by [Junqueira and Moreira 2026], which demonstrated that widely adopted automatic metrics remain weakly correlated with human judgments in the QA task.

Table 6 highlights limitations of automatic metrics observed in the QA tasks. In the first example, the generated answer is semantically equivalent to the reference, leading

to high human scores. Contrarily, in the second example, the generated answer is factually incorrect and only loosely related to the reference, resulting in a low human score. However, most metrics, with the sole exception of BERTScore and MoverScore, attributed higher scores to the second instance. These examples demonstrate how current automatic metrics may fail to accurately capture semantic equivalence and answer validity.

Table 6. Examples illustrating failures of automatic metrics in QA.

Example 3 (High-quality Answer)	**Example 4 (Low-quality Answer)**
Question: What did the Happy Hunter discover when he pulled the line for the last time?	**Question:** What strange thing did the prince do?
Answer: The fishing hook was lost.	**Answer:** He was carrying a dagger in his belt.
Reference: He discovered he lost the fishing hook without knowing when he dropped it.	**Reference:** He didn't use the dagger during the fight.
Human Score: 4.7/5	**Human Score:** 1.5/5

5.3. Discussion

The results reinforce the need for the development of evaluation metrics that better reflect human perception. Although PLM-based metrics achieved slightly higher correlations with human judgments, especially in QA, the observed correlations maintained medium to low values across all evaluated metrics, with outliers scraping the high correlation bound. Additionally, results suggest that n-gram-based metrics may be capable of achieving competitive or even superior performances.

Among the evaluated PLM-based approaches, BARTScore [Yuan et al. 2021] presented notably lower performance than expected. A possible explanation is the limited availability of large BART models pre-trained in Portuguese, since PLM-based metrics strongly depend on the quality of the underlying model.

Among n-gram-based strategies, METEOR [Banerjee and Lavie 2005] showed the highest alignment with human judgment, reinforcing the benefits of a synonym dictionary. While ROUGE-1 and ROUGE-2 [Lin 2004] underperformed in the Summarization task, highlighting their sensitivity to variations in lexical choices and syntactic structures.

The results also highlight the scarcity of well-structured and carefully curated datasets in Portuguese, forcing studies to rely on translated or adapted datasets, introducing additional sources of noise and inconsistency. Since automatic metrics fundamentally depend on reference texts, problems such as translation errors, cultural mismatches, and poorly formulated references directly compromise the reliability of intrinsic evaluation.

6. Conclusion

This work investigated the correlation between automatic evaluation metrics and human judgments in open-ended tasks, with a particular focus on summarization and QA in Portuguese. The obtained results demonstrate that widely adopted automatic metrics still present medium to low correlations with human evaluation. The study also identified important limitations, including the scarcity of high-quality datasets and language models for Portuguese. These factors negatively impact both output generation and the reliability of automatic metrics. Overall, the results reinforce the need for more robust evaluation methods capable of better reflecting human judgment.

Limitations

While this study provides a unified multi-task methodology for analyzing automatic evaluation metrics in Portuguese, several limitations must be acknowledged.

Sample Size and Scale of Human Evaluation. Due to resource constraints, the human-evaluation benchmark was limited to 40 instances per task. Expanding the evaluation to a larger and more diverse set of models and datasets would provide a more comprehensive assessment of metric stability across different quality levels.

Non-Expert Evaluator Profiles. The eight human judges selected for this study were Computer Science students. Although they are knowledgeable in the general functioning of LLMs, they lack formal linguistic training, meaning some subtle nuances may have been lost during data collection.

Dependence on Machine-Translated Datasets. This study relied on machine-translated benchmarks, which inherently introduce specific artifacts, translation errors, and semantic distortions. Since automatic metrics are bound to the quality of their reference texts, these limitations directly constrain the reliability of the calculated correlations.

Constraints of Under-Resourced Language Models. The performance of PLM-based metrics is highly dependent on the quality of the underlying models. Any non-English language, such as Portuguese, suffers from the scarcity of large, natively pre-trained models. This limitation was particularly evident in BARTScore's results, which yielded correlation scores that were drastically lower than expected.

Cross-Task Adaptation of Quality Dimensions. In the absence of firmly consolidated, standardized evaluation dimensions designed specifically for QA, this work mirrored the four dimensions used in summarization. Although this choice maintained the baseline across both tasks, certain dimensions may not perfectly translate to the QA task.

Ethics Statement

All human evaluations were collected with the participants' written consent; they were fully informed about the study's purpose, their participation was entirely voluntary, and they were allowed to withdraw at any time without penalty. No personal data was collected from the participants, apart from which group they belonged to. The participants were all closely related to the authors, but were advised that trying to inflate the results would only lead to a worse performance for the metrics.

AI Usage

Generative AI was used as an auxiliary tool during research and revision of this paper. For research, the generative models' ability to conduct in-depth research was used to scan the literature and find works that support specific claims or findings. Generative AI was not the primary source for research, being used in limited cases and always with caution, as we understand information obtained through this means is often unreliable and may be impacted by hallucinations. For revision, Generative AI was used to correct grammar, suggest phrasing, and analyze weaknesses.

Acknowledgments

This work has been partially funded by CNPq and Capes Finance Code 001.

References

Banerjee, S. and Lavie, A. (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*, pages 65–72.

Deutsch, D., Dror, R., and Roth, D. (2021). A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Deutsch, D., Dror, R., and Roth, D. (2022). Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052.

Fabbri, A. R., Kryściński, W., McCann, B., et al. (2021). SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Farea, A., Yang, Z., Duong, K., et al. (2025). Evaluation of question answering systems: Complexity of judging a natural language. *ACM Comput. Surv.*

Hasan, T., Bhattacharjee, A., Islam, M. S., et al. (2021). Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Junqueira, J. d. R. and Moreira, V. P. (2026). The inadequacy of automatic evaluation metrics in question answering: A case-study in Portuguese. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 551–561, Salvador, Brazil. Association for Computational Linguistics.

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Krishna, K., Roy, A., and Iyyer, M. (2021). Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the ACL*, pages 4940–4957.

Leite, B., Osório, T. F., and Cardoso, H. L. (2024). Fairytaleqa translated: Enabling educational question and answer generation in less-resourced languages. In *Technology Enhanced Learning for Inclusive and Equitable Quality Education*, pages 222–236.

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Liu, C.-W., Lowe, R., Serban, I., et al. (2016). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response

generation. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Liu, Y., Iter, D., Xu, Y., et al. (2023). G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Nallapati, R., Zhou, B., dos Santos, C., et al. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

NIST, N. (2006). Nist 2006 machine translation evaluation official results.

Paiola, P. H., de Rosa, G. H., and Papa, J. P. (2022). Deep learning-based abstractive summarization for brazilian portuguese texts. In *BRACIS 2022: Intelligent Systems*, pages 479–493.

Papineni, K., Roukos, S., Ward, T., et al. (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.

Rajpurkar, P., Zhang, J., Lopyrev, K., et al. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020*, page 403–417.

Team, G., Kamath, A., Ferret, J., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Xu, Y., Wang, D., Yu, M., et al. (2022). Fantastic questions and where to find them: Fairytaleqa – an authentic dataset for narrative comprehension.

Yuan, W., Neubig, G., and Liu, P. (2021). Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277.

Zhang, T., Kishore, V., Wu, F., et al. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhao, W., Peyrard, M., Liu, F., et al. (2019). Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1902.02622*.