# Evaluating Reference-Free Summarization Quality Metrics for Portuguese: A Study with Human Judgments in Financial News

**João Victor Assaoka Ribeiro and Thomas Pires Correia and José Vitor Sousa Cardoso Requena and Lilian Berton**
Instituto de Ciência e Tecnologia - Universidade Federal de São Paulo
Unidade Parque Tecnológico - Avenida Cesare Mansueto Giulio Lattes, n° 1201
Eugênio de Mello, CEP: 12247-014
joao.assaoka@unifesp.br, lberton@unifesp.br

## Abstract

Automatic summarization of financial news in Portuguese lacks reliable reference-free evaluation metrics. While LLM-as-a-Judge approaches are gaining traction, their correlation with human perception in specialized domains remains under-explored. This work evaluates the efficacy of Question Answering (QA) based metrics against a direct LLM-as-a-Judge baseline for Portuguese financial news. We propose a pipeline comparing Lexical, Binary, and Semantic (LLM-based) QA scoring methods, validated against a human ground truth of 50 news items annotated for Faithfulness and Completeness. Our results show that granular QA metrics significantly outperform the monolithic LLM-Judge in evaluating Completeness, with QA-Binary achieving the highest rank correlation ($\rho \approx 0.49$ with pessimistic human aggregation). For Faithfulness, we observe a strong ceiling effect in human evaluation, yet the Semantic QA metric demonstrated a "superhuman" ability to detect subtle hallucinations (e.g., temporal shifts) missed by annotators. We conclude that decomposing evaluation into atomic QA pairs is superior to holistic judging for the Portuguese financial domain.

## 1 Introduction

The digitalization of the financial sector has generated a huge volume of unstructured data, making automatic summarization a critical asset for investment decision-making (Nassirtoussi et al., 2014). While Large Language Models (LLMs) have achieved remarkable fluency in Portuguese summarization, their deployment in high-stakes domains is constrained by the risk of hallucinations—fluent but factually incorrect statements (Maynez et al., 2020). Furthermore, the practical application of these models faces a fundamental challenge often overlooked in academic benchmarks: the absence of gold standards. In real-world production pipelines, reference summaries written by experts are unavailable, rendering traditional reference-based metrics (such as ROUGE or METEOR) inapplicable.

To address this reference-free constraint, the industry is shifting towards the "LLM-as-a-Judge" paradigm, where a powerful model scores generated text based on a holistic prompt (Zheng et al., 2023). However, relying on a monolithic LLM score creates a "black box" evaluation. It remains unclear whether these models can accurately penalize subtle semantic errors in specialized domains like Portuguese financial news, or if they are merely biasing scores towards highly fluent outputs (Wu and Aji, 2025). We hypothesize that for the financial domain, where precision is paramount, evaluation must be decomposed into atomic verifications rather than relying on holistic judgments.

In this work, we evaluate the efficacy of Question Answering (QA)-based metrics (Wang et al., 2020a) as a granular, reference-free alternative for Portuguese. Unlike monolithic judging, QA metrics assess quality by verifying if specific facts (answers) from the source are preserved in the summary (Completeness) and if facts in the summary are supported by the source (Faithfulness). We implement a pipeline comparing three answer-scoring strategies: Lexical (overlap), Binary (exact retrieval), and Semantic (LLM-based reasoning), contrasting them against a direct LLM-as-a-Judge baseline.

Our study offers two significant contributions to the evaluation of Portuguese financial summarization:

1. **Granularity outperforms Holistic Judging for Completeness:** We demonstrate that granular QA metrics correlate significantly better with human judgments than the direct LLM-Judge. Notably, a strict *QA-Binary* approach achieved the highest rank correlation ($\rho \approx 0.49$) with a pessimistic human ag-



*Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*
*Vol. 1*, pages 899–907
April 13-16, 2026 

gregation, suggesting that capturing the presence/absence of key facts is more reliable than vague quality scores.

2. **Super-Human Faithfulness Auditing:** We identify a strong "ceiling effect" in human evaluation, where annotators fail to notice subtle hallucinations in fluent text. Our qualitative analysis reveals that the *Semantic QA* metric acts as a "super-human" auditor, successfully detecting temporal shifts and inference errors that human evaluators missed.

By validating these metrics against a human-annotated corpus of 50 financial news items, this work establishes a reliable, scalable, and transparent methodology for auditing financial summaries in Portuguese without the need for reference texts.

## 2 Related Work

The evaluation of automatic summarization has historically relied on lexical overlap metrics such as ROUGE (Lin, 2004). While ROUGE remains a standard for measuring content recall, research has consistently shown its limitations in capturing factual consistency and faithfulness, especially in abstractive summarization where the model may use synonyms or paraphrase information incorrectly (Maynez et al., 2020). In the financial domain, where a single misinterpreted digit or entity can change the entire meaning of a report, the failure of lexical metrics to detect hallucinations is a critical vulnerability (Ji et al., 2023).

To mitigate these limitations, recent literature has proposed Question Answering (QA)-based metrics as a more granular approach to faithfulness. The foundational frameworks such as FEQA (Durmus et al., 2020) and QAGS (Wang et al., 2020b) operate on the principle that a summary is faithful if a QA model can find the same answers in both the summary and the source text. These methods have demonstrated a significantly higher correlation with human judgment regarding factual consistency compared to ROUGE. Subsequent works like QuestEval (Scialom et al., 2021) extended this approach to measure recall, while QAFactEval (Fabbri et al., 2022) optimized the QA components for better consistency. However, the performance of QA-based metrics is heavily influenced by the quality of the question generation (QG) and answer extraction (AE) components, which may introduce their own errors into the evaluation pipeline (Zhang et al., 2024).

Parallel to the development of QA metrics, the emergence of Large Language Models has fostered the "LLM-as-a-Judge" paradigm. Works such as G-Eval (Liu et al., 2023) demonstrate that high-capacity models like GPT-4 can outperform traditional reference-based metrics by following complex evaluation instructions (Chain-of-Thought). This approach has been further refined by benchmarks like MT-Bench (Zheng et al., 2023), which validate the use of LLMs to grade other models' outputs. More recently, (Song et al., 2024) proposed fine-grained evaluations using LLMs to address the lack of interpretability in holistic scores. Despite these advances, applying LLM-as-a-Judge to specialized languages like Portuguese and niche domains like finance remains an open area of research, as models may exhibit cultural or domain-specific biases during the judging process.

Our work builds upon these two lineages, QA-based metrics and LLM evaluation, by specifically examining their performance in the Portuguese financial news context. Unlike previous studies that often treat QA scoring as a black box, we isolate and compare Lexical, Binary, and Semantic scoring methods, validating them against a curated human ground truth to determine the most reliable strategy for the financial domain.

## 3 Data and Annotation Protocol

This study is conducted within the scope of a broader research project aimed at developing a sentiment analysis dataset for Portuguese financial news. The primary corpus consists of 4,841 financial news articles collected from the Brazilian finance portal "Exame". The choice of a single source was a deliberate decision to mitigate content duplication and avoid data leakage between different portals that republish the same news.

### 3.1 LLM-Assisted Summarization Strategy

The overarching goal of the dataset we are building is sentiment analysis. Financial news articles often contain excessive contextual information or irrelevant facts that may dilute the predominant sentiment. Recognizing this challenge, we opted not to use the full text for annotation, as text reduction also helps make the sentiment annotation process more viable. Instead, we implemented a summarization step assisted by a Large Language Model (LLM). Our underlying hypothesis was that a structured summary would reduce the incidence

of “Neutral” labels and focus the annotator’s attention on central events. This work focuses specifically on how we are evaluating the quality of these summaries.

We employed the `gpt-4.1-nano` model through the OpenAI API to process each article at the time of dataset creation, as it was the most recent model available. The "nano" version was selected because it is the most cost-effective option according to the provider. For the subsequent quality evaluation phase, which is the core of this study, the newly released `gpt-5-nano` was used. The model was instructed to generate exactly three sentences representing the beginning, middle, and end of the narrative, focusing strictly on the main facts.

Table 1 presents the descriptive statistics of the summarization process. The average compression rate is approximately 5.41, indicating a significant reduction in text volume while aiming to preserve core information.

Table 1: Descriptive statistics of news summarization (Words and Compression Rate).

| | Document | Summary | Compression Rate |
|---|---|---|---|
| Mean | 495.92 | 91.32 | 5.41 |
| Std | 299.85 | 12.10 | 3.16 |
| Min | 86.00 | 58.00 | 1.01 |
| 25% | 299.00 | 83.00 | 3.35 |
| 50% | 425.00 | 91.00 | 4.62 |
| 75% | 606.00 | 99.00 | 6.53 |
| Max | 3793.00 | 149.00 | 41.81 |

### 3.2 Human Evaluation Protocol

Our primary objective is to assess the quality of automatic metrics for Faithfulness and Completeness. Initial automated experiments indicated divergences between different metric approaches. To establish a reliable ground truth for comparison, we required human validation. Given the absence of reference summaries (gold standard) in our proprietary corpus, we established a reference-free human evaluation protocol using a random sample of 50 news items ($n = 50$).

#### 3.2.1 Annotator Profiling and Qualification

The annotation process was conducted by four members of the research group, ensuring a controlled environment with strict adherence to the guidelines. The team consists of Computer Science researchers stratified between undergraduate and graduate levels. Crucially, despite not being professional financial analysts, the annotators possess high domain familiarity: 75% of the group are active investors with a consolidated habit of consuming financial news. This background is pivotal for the task, as it ensures the annotators understand the specific vocabulary (e.g., "EBITDA", "Selic", "Fiscal Surplus") and the implicit sentiment of market movements, mitigating the cognitive load and potential misinterpretations common to layperson annotators.

#### 3.2.2 Task Guidelines

Each news item was evaluated by two different annotators, resulting in a total of 100 human annotations distributed across all six possible combinations of annotator pairs. The annotators assessed the summaries on two dimensions using a Likert scale from 1 to 5:

- **Completeness:** Does the summary cover the key events of the original article?
- **Faithfulness:** Does the summary contain hallucinations or false information compared to the source?

#### 3.2.3 Inter-Annotator Agreement

Table 2 details the absolute disagreement between annotators. Faithfulness annotations showed high agreement (43 perfect matches), suggesting that the presence of hallucinations is a more objective task or that the summaries were generally of high quality. Conversely, Completeness showed higher divergence (21 cases with disagreement $\geq 1$), reflecting the subjective nature of determining "key events" in financial narratives.

Table 2: Annotator Disagreement Distribution (Absolute difference between scores).

| Difference | Completeness | Faithfulness |
|---|---|---|
| 0 | 29 | 43 |
| 1 | 17 | 6 |
| 2 | 4 | 0 |

Figures 1 and 2 illustrate the distribution of Human scores against the automatic QA metrics, demonstrating that the sample covers a representative range of the dataset’s quality spectrum.

## 4 Methodology

To evaluate the quality of the summaries without human references, we compare two distinct approaches: a granular Question-Answering (QA) framework and a Direct LLM-as-a-Judge baseline.

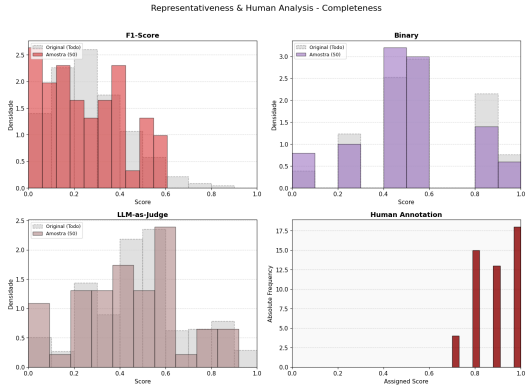


Figure 1: Distribution of Human Judgments compared to QA Metrics - Completeness

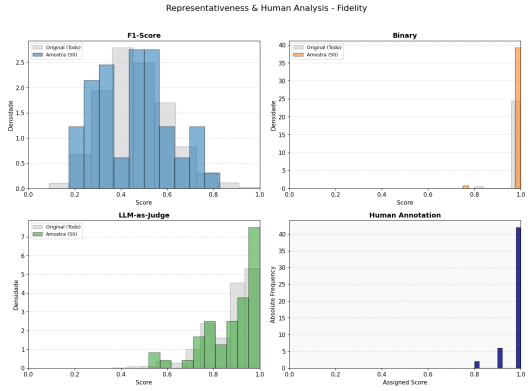


Figure 2: Distribution of Human Judgments compared to QA Metrics - Faithfulness

Due to space limitations, the specific prompts utilized in both approaches are omitted from this manuscript. However, all prompts, along with the full evaluation pipeline, are publicly available in our GitHub repository[1].

### 4.1 Model

For the evaluation tasks described below, we employed the `gpt-5-nano` model. The "nano" series was selected due to its optimized inference speed and cost-efficiency, while maintaining competitive performance on summarization and classification tasks according to OpenAI technical reports.

### 4.2 Baseline: Direct LLM-as-a-Judge

As a baseline, we employed a monolithic evaluation method where the `gpt-5-nano` model directly scores the summary without the intermediate QA steps.

We utilized a zero-shot persona-based prompt instructing the model to act as a Senior Economic Journalism Editor. The model receives the original article and the generated summary and is tasked with assigning a score from 0 to 10 on two specific dimensions:

1. **Faithfulness:** Assessing strictly for the presence of hallucinations or information not supported by the source.
2. **Completeness:** Evaluating if the summary captures the main ideas and critical numbers of the original financial report.

These scores are normalized to $[0, 1]$ to allow for direct comparison with the QA-based metrics.

### 4.3 QA-Based Evaluation Framework

Our primary proposal is a reference-free evaluation pipeline grounded in the intuition that a good summary should verify the same key facts as the source document. This framework adapts the principles of previous works to the financial domain.

The evaluation pipeline consists of five stages:

#### 4.3.1 Question Generation (QG)

The model generates $N \leq 5$ pairs of questions ($Q$) and expected answers ($A_{exp}$) focusing on the main facts of a source text ($T_{source}$). The directionality of generation determines the metric being evaluated:

- **Completeness ($D \rightarrow S$):** Questions are generated from the original Document ($D$). If the Summary ($S$) can answer these questions, it preserves key information.
- **Faithfulness ($S \rightarrow D$):** Questions are generated from the Summary ($S$). If the original Document ($D$) confirms these answers, the summary is hallucination-free.

#### 4.3.2 Financial Risk-Aware Weighting

Standard QA metrics often treat all facts as equal. However, in the financial domain, the severity of an error varies significantly.

To capture this nuance, we implemented a zero-shot prompting strategy where the LLM adopts the persona of a Financial Data Quality Auditor. The model assigns an importance weight $w_i \in [1, 10]$ to each QA pair based on the potential impact of misinformation on an investor's decision-making process.

The prompt framing differs slightly depending on the metric:

[1] https://github.com/Assaoka/Evaluating_Reference-Free_Summarization_Quality_Metrics_for_Portuguese_in_Financial_News

- **For Completeness:** The model assesses the severity of *omission*. A weight of 10 ("Devastating") implies that missing this fact renders the summary useless for investment decisions, while a weight of 1 ("Acceptable") indicates trivial details.
- **For Faithfulness:** The model assesses the danger of *hallucination*. It evaluates how dangerous it would be if the summary invented or distorted the specific fact represented by the question.

This risk-based weighting helps ensure that the final score reflects the functional utility of the summary.

#### 4.3.3 Question Answering (QA)

We probe the opposing text ($T_{target}$) to obtain a candidate answer ($A_{cand}$) for each question $Q_i$.

$$A_{cand,i} = \text{Model}(Q_i, T_{target}) \tag{1}$$

If the information is not present in $T_{target}$, the model is instructed to return "Information not found".

#### 4.3.4 Answer Scoring Metrics

We evaluate the quality of $A_{cand}$ against $A_{exp}$ using three distinct methods:

- **QA-F1 (Token Overlap):** Calculates the F1-Score of token overlap between the expected and candidate answers. While computationally efficient, this metric penalizes semantically correct answers that use different lexical choices (e.g., "profit rose" vs. "earnings increased").
- **QA-Binary (Answerability):** A strict binary metric that evaluates answerability. It assigns a score of 1 if $A_{cand}$ contains a factual answer, and 0 if the model returned "Information not found". This metric ignores the semantic correctness relative to the expectation, focusing solely on retrieval.
- **QA-LLM (Semantic Equivalence):** An "LLM-as-a-Judge" approach designed to handle semantic variability. The model compares $A_{cand}$ and $A_{exp}$ and assigns a score $s \in [0, 5]$ based on factual equivalence, which is then normalized to $[0, 1]$. This metric addresses the limitations of QA-F1 and QA-Binary.

#### 4.3.5 Aggregation Strategies

Finally, the document-level score is computed using either an arithmetic mean ($\mu_A$) or a weighted mean ($\mu_P$) based on the importance weights $w_i$:

$$\text{Score}_{\text{Arithmetic}} = \frac{1}{N}\sum_{i=1}^{N} S_i \tag{2}$$

$$\text{Score}_{\text{Weighted}} = \frac{\sum_{i=1}^{N}(S_i \cdot w_i)}{\sum_{i=1}^{N} w_i} \tag{3}$$

## 5 Experiments and Results

### 5.1 Completeness Analysis

#### 5.1.1 Descriptive Statistics

Table 3 presents the descriptive statistics for the entire sample ($n = 50$). We observe that the weighted versions of the metrics ($\text{QA}_\text{W}$) consistently yield higher mean scores compared to their arithmetic counterparts ($\text{QA}_\text{A}$). Specifically, the weighted score was higher in 40 cases for QA-F1, 36 for QA-Binary, and 39 for QA-LLM.

This trend suggests that the summarization model effectively prioritized information with higher importance weights ($w_i$), preserving critical financial facts even when omitting peripheral details. The baseline LLM-Judge, however, presents a notably higher mean (0.67) and lower standard deviation, indicating a tendency towards more generous and less discriminative scoring compared to the granular QA approaches.

Table 3: Descriptive Statistics of Completeness Metrics.

| | QA-F1 | | QA-Binary | | QA-LLM | | Judge | Human |
|---|---|---|---|---|---|---|---|---|
| | W | A | W | A | W | A | | |
| Mean | 0.29 | 0.25 | 0.57 | 0.50 | 0.50 | 0.44 | 0.67 | 0.89 |
| Std | 0.19 | 0.18 | 0.26 | 0.25 | 0.25 | 0.24 | 0.14 | 0.10 |
| Min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.30 | 0.70 |
| 25% | 0.17 | 0.11 | 0.42 | 0.40 | 0.36 | 0.29 | 0.60 | 0.80 |
| 50% | 0.24 | 0.22 | 0.55 | 0.50 | 0.51 | 0.40 | 0.70 | 0.90 |
| 75% | 0.41 | 0.39 | 0.79 | 0.60 | 0.69 | 0.60 | 0.80 | 1.00 |
| Max | 0.65 | 0.61 | 1.00 | 1.00 | 0.95 | 0.92 | 0.90 | 1.00 |

#### 5.1.2 Correlation with Human Judgments

To validate the metrics, we initially compared them against the arithmetic mean of the two human annotators. Table 4 breaks down the metric performance across bins of human scores.

We observe a general upward trend in QA metric scores as human satisfaction increases, particularly in the transition to the highest score bucket (1.0).

Table 4: Metric Means and Standard Deviations across Human Score Bins (Mean Aggregation).

| Human | N | QA-F1 | | QA-Binary | | QA-LLM | | Judge |
|---|---|---|---|---|---|---|---|---|
| | | W | A | W | A | W | A | |
| 0.7 | 4 | $0.29 \pm 0.09$ | $0.25 \pm 0.10$ | $0.48 \pm 0.11$ | $0.40 \pm 0.00$ | $0.44 \pm 0.14$ | $0.35 \pm 0.10$ | $0.78 \pm 0.05$ |
| 0.8 | 15 | $0.23 \pm 0.19$ | $0.20 \pm 0.18$ | $0.49 \pm 0.32$ | $0.43 \pm 0.31$ | $0.44 \pm 0.29$ | $0.38 \pm 0.29$ | $0.66 \pm 0.14$ |
| 0.9 | 13 | $0.22 \pm 0.19$ | $0.20 \pm 0.17$ | $0.52 \pm 0.28$ | $0.45 \pm 0.26$ | $0.42 \pm 0.27$ | $0.37 \pm 0.23$ | $0.59 \pm 0.10$ |
| 1.0 | 18 | $0.37 \pm 0.18$ | $0.34 \pm 0.17$ | $0.68 \pm 0.17$ | $0.62 \pm 0.18$ | $0.63 \pm 0.19$ | $0.57 \pm 0.19$ | $0.72 \pm 0.15$ |

The slight dip observed in the 0.9 bin (typically representing a divergence between scores 4 and 5) may be attributed to the subjectivity in distinguishing "excellent" from "perfect" summaries, or simply sample noise.

Figure 3 illustrates these distributions. The QA-Binary_A metric demonstrates a clearer monotonicity across classes compared to the baseline LLM-Judge, which exhibits a flatter distribution and fails to penalize lower-quality summaries effectively.

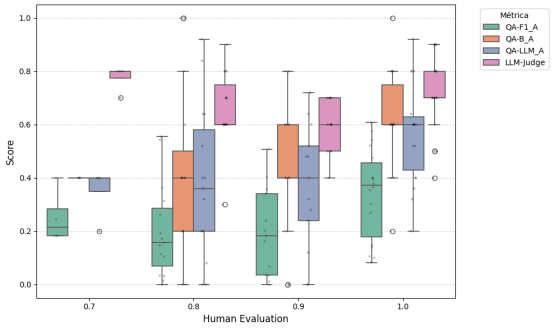


Figure 3: Distribution of QA Metrics (Arithmetic) and LLM-Judge scores grouped by Human Evaluation (Mean Aggregation).

Table 5 details the Pearson ($r$) and Spearman ($\rho$) correlations.

Table 5: Correlation with Human Judgments (Mean Aggregation). Bold indicates best performance.

| Metric | QA-F1 | | QA-Binary | | QA-LLM | | Judge |
|---|---|---|---|---|---|---|---|
| | W | A | W | A | W | A | |
| Pearson ($r$) | 0.26 | 0.28 | 0.31 | 0.33 | 0.31 | **0.34** | 0.04 |
| *p-value* | *0.07* | *0.05* | *0.03* | *0.02* | *0.03* | *0.02* | *0.78* |
| Spearman ($\rho$) | 0.26 | 0.27 | 0.40 | **0.41** | 0.33 | 0.35 | 0.07 |
| *p-value* | *0.07* | *0.06* | *0.00* | *0.00* | *0.02* | *0.01* | *0.64* |

Results indicate a clear superiority of QA-based approaches over the direct LLM-Judge, which showed no significant correlation ($p > 0.05$). Surprisingly, the arithmetic aggregation ($\_A$) outperformed the importance-weighted aggregation ($\_W$) across most metrics. This suggests that while the generated weights ($w_i$) capture factual importance (as seen in Table 3), they may introduce noise when correlating with the holistic human perception of completeness in this specific domain.

Furthermore, QA-Binary achieved the highest rank correlation ($\rho = 0.41$). We hypothesize that the binary nature of this metric aligns better with the discrete steps of the human Likert scale, whereas the continuous variance in QA-LLM scores might introduce granularity that does not necessarily reflect human-perceived quality differences in high-compression scenarios.

### 5.1.3 Impact of Pessimistic Strategy

To rigorously test the metrics' robustness, we repeated the analysis using the *minimum* score between the two annotators. This approach simulates a "pessimistic" ground truth, prioritizing the detection of missing information identified by the stricter annotator.

Table 6: Correlation with Human Judgments (Minimum Aggregation).

| Metric | QA-F1 | | QA-Binary | | QA-LLM | | Judge |
|---|---|---|---|---|---|---|---|
| | W | A | W | A | W | A | |
| Pearson ($r$) | 0.36 | 0.38 | 0.43 | 0.44 | 0.42 | **0.44** | 0.22 |
| *p-value* | *0.01* | *0.01* | *0.00* | *0.00* | *0.00* | *0.00* | *0.12* |
| Spearman ($\rho$) | 0.36 | 0.36 | 0.49 | **0.49** | 0.44 | 0.44 | 0.23 |
| *p-value* | *0.01* | *0.01* | *0.00* | *0.00* | *0.00* | *0.00* | *0.11* |

As shown in Table 6 and Figure 4, correlations improved significantly across all QA metrics, with QA-Binary reaching $\rho \approx 0.49$. This reinforcement suggests that our reference-free metrics are particularly effective at identifying completeness failures, aligning closely with the more rigorous human judgments. The direct LLM-Judge improved slightly but remained statistically insignificant.

### 5.1.4 Completeness vs. Compression Rate

Finally, we analyzed the relationship between the Compression Rate and Completeness scores.

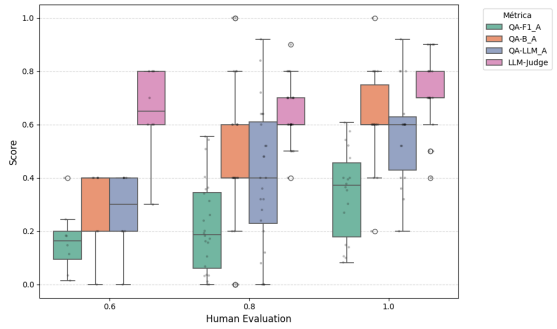


Figure 4: Distribution of QA Metrics vs. Minimum Human Score. The metrics show clearer separation between quality bins.

Table 7: Correlation between Compression Rate and Completeness Metrics (Arithmetic Mean). QA-B denotes QA-Binary.

| Metric | QA-F1 | QA-B | QA-LLM | Judge | Human |
|---|---|---|---|---|---|
| $r$ | -0.48 | -0.61 | -0.59 | -0.35 | -0.44 |
| $\rho$ | -0.51 | -0.61 | -0.58 | -0.40 | -0.42 |

Table 7 confirms a consistent negative correlation across all evaluators, validating the intuition that higher compression naturally imposes a trade-off in information preservation. Notably, the granular QA metrics (QA-Binary and QA-LLM) exhibit a higher sensitivity to compression ($r \approx -0.60$) compared to the Human ground truth ($r = -0.44$). Conversely, the direct LLM-Judge displays the weakest correlation ($r = -0.35$), suggesting it may underestimate the impact of aggressive summarization on content coverage.

## 5.2 Faithfulness Analysis

### 5.2.1 Descriptive Statistics and The Ceiling Effect

The summarization model demonstrated exceptional performance in generating hallucination-free summaries. As detailed in the Descriptive Statistics (Table 8), the **QA-Binary** metric achieved a near-perfect mean score of $0.99$, while human annotators assigned the maximum score (5/5) to 84% of the sample $(42/50)$.

This distribution creates a significant "Ceiling Effect," effectively nullifying the utility of standard correlation analysis. As shown in Table 9, the lack of variance in the human ground truth resulted in statistically insignificant $p$-values ($p > 0.05$) across almost all metrics. The metrics could not be linearly correlated with a variable that acts almost as a constant.

Table 8: Descriptive Statistics of Faithfulness Metrics.

| | QA-F1 | | QA-Binary | | QA-LLM | | Judge | Human |
|---|---|---|---|---|---|---|---|---|
| | W | A | W | A | W | A | | |
| Mean | 0.46 | 0.46 | 0.99 | 0.99 | 0.89 | 0.88 | 0.79 | 0.97 |
| Std | 0.16 | 0.16 | 0.04 | 0.04 | 0.12 | 0.12 | 0.13 | 0.07 |
| Min | 0.17 | 0.17 | 0.74 | 0.75 | 0.54 | 0.52 | 0.40 | 0.80 |
| 25% | 0.32 | 0.32 | 1.00 | 1.00 | 0.82 | 0.80 | 0.70 | 1.00 |
| 50% | 0.49 | 0.47 | 1.00 | 1.00 | 0.92 | 0.92 | 0.80 | 1.00 |
| 75% | 0.57 | 0.55 | 1.00 | 1.00 | 1.00 | 1.00 | 0.90 | 1.00 |
| Max | 0.83 | 0.83 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 9: Correlation with Human Judgments (Faithfulness). Note the high p-values due to the ceiling effect.

| Metric | QA-F1 | | QA-Binary | | QA-LLM | | Judge |
|---|---|---|---|---|---|---|---|
| | W | A | W | A | W | A | |
| Pearson ($r$) | 0.21 | 0.17 | -0.06 | -0.06 | -0.05 | -0.06 | 0.11 |
| *p-value* | *0.14* | *0.23* | *0.67* | *0.67* | *0.72* | *0.67* | *0.44* |
| Spearman ($\rho$) | 0.23 | 0.18 | -0.06 | -0.06 | -0.01 | -0.01 | 0.09 |
| *p-value* | *0.10* | *0.22* | *0.67* | *0.67* | *0.93* | *0.94* | *0.54* |

### 5.2.2 Qualitative Error Analysis

Given the limitations of quantitative analysis caused by the ceiling effect, we conducted a qualitative inspection of the outliers. We selected the bottom-3 scoring samples from each metric and analyzed the specific Question-Answer pairs to understand the divergences between human and automatic scoring.

Our analysis revealed three key behaviors:

- **Sensitivity to Temporal and Extrinsic Hallucinations:** The automated metrics successfully identified cases where humans awarded perfect scores (5/5) but the summary contained subtle factual errors. For instance, in Sample #3501, the summary introduced a temporal hallucination by explicitly stating "2024" as a target date. The source text, anchored in 2022, mentioned "next year" (implying 2023). While human annotators missed this timeline shift, the **QA-LLM** metric correctly penalized the inconsistency. Similarly, in Sample #3180, the summary fabricated a "resumption of the agreement" which was never mentioned in the source. Again, humans missed this extrinsic hallucination, but the QA framework flagged it.

- **The "False Positive" of Style (QA-F1):** The token-based **QA-F1** metric proved unreliable for faithfulness validation, often penalizing correct answers due to verbosity differences.

In Sample #599, both the reference text and the summary answer were factually aligned regarding the "2030 oil production targets". However, the generated answers had significantly different lengths, causing the F1 score to drop disproportionately. A similar pattern was observed in Sample #1275. This confirms that token overlap metrics are unsuitable for detecting hallucinations in scenarios where the model paraphrases or summarizes the answer.

- **Granular Verification (QA-LLM vs. QA-Binary):** The comparison between QA-Binary and QA-LLM highlights the need for semantic verification. In Sample #2372, the summary confused the "approval of urgency" with the "approval of the proposal itself". The **QA-Binary** metric failed to catch this nuance (likely matching keywords), whereas the **QA-LLM** identified the semantic divergence. Furthermore, Sample #599 demonstrated the rigor of the proposed method: the summary claimed the oil sector would "maintain economic relevance until 2030", while the source only stated production targets for that year. The weighted QA-LLM correctly identified this inference error, acting as a strict auditor even when human reviewers were more lenient.

In conclusion, the qualitative analysis suggests that QA-based metrics (specifically QA-LLM) function as effective "super-human" auditors, capable of detecting granular factual inconsistencies, such as temporal shifts and subtle inference errors, that human readers often overlook.

## 6 Conclusion

In this work, we proposed and validated a reference-free evaluation framework tailored for Portuguese financial news summarization. This approach addresses the dependency on scarce human gold standards and mitigates the fragility of monolithic "LLM-as-a-Judge" metrics. Our experiments demonstrate that granular Question-Answering (QA) approaches are superior to holistic judging in capturing quality nuances.

Regarding Completeness, the QA metrics exhibited a stronger correlation with human judgments compared to the direct Judge baseline. Breaking down the evaluation into atomic questions effectively reduces subjectivity. Furthermore, the "pessimistic" aggregation strategy and qualitative analysis indicate that QA metrics are robust, aligning automatic evaluation with the rigor required for the financial domain.

For Faithfulness, the observed "Ceiling Effect" suggests that current models generate highly fluent text that can mislead human evaluators. Qualitative analysis revealed that the QA-LLM metric acted as a "super-human" auditor, detecting subtle hallucinations regarding dates and logical inferences that passed unnoticed by humans. Although QA-Binary performed slightly better in Completeness statistics, we hypothesize that a more granular human scoring scale could reveal the superiority of QA-LLM in identifying these semantic subtleties.

Regarding the importance weighting mechanism, our results indicate that the simple arithmetic mean performed on par with, or even surpassed, the risk-aware weighted approach. We attribute this to the high compression nature of the task: in concise summaries, essentially every remaining piece of information becomes critical, potentially saturating the utility of differential weights.

## 7 Future Work

Building upon these findings, we outline three main directions for future research.

First, we intend to expand the sample size ($N$) to ensure greater statistical significance in subtle scenarios. Future annotation rounds will also incorporate more granular scales and a larger pool of annotators to better capture semantic subtleties.

Second, we plan to refine the weighting strategy, specifically investigating methods to align the automated weights more closely with human perception of error severity. We aim to explore whether different weighting paradigms—such as non-linear scales or dynamic mechanisms—can better approximate the human judgment of financial risk, going beyond simple linear aggregation.

Finally, and crucial to the broader scope of this project, we will evaluate how summary quality impacts downstream tasks, specifically Sentiment Analysis. We aim to test the hypothesis that low completeness (e.g., omitting negative data) leads to incorrect sentiment classification (False Neutral/Positive). This validates the necessity of high-quality summarization not just as an end product, but as reliable input for subsequent financial modeling.

## 8 Acknowledgments

The authors acknowledge financial support from FAPESP (Grants 2024/17511-2 and 2025/21207-0), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Serasa Experian. We also thank Exame for providing the dataset used in this analysis.

## References

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 2587–2601.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. 2014. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6594–6604.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020b. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Minghao Wu and Alham Fikri Aji. 2025. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 297–312.

Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-Hong Huang, and Evangelos Kanoulas. 2024. Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics. *arXiv preprint arXiv:2406.15264*.

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.