Auditing the Evaluators: How Far Can Automatic Evaluation Go in Assessing Portuguese Financial Texts?

Marina Ramalhete Masid[1], **Gabriel Assis**[2], **Daniela Vianna**[1], **Aline Paes**[2] and **Altigran Soares da Silva**[3]

[1]Jusbrasil, Salvador, BA, Brazil,
[2]Instituto de Computação, Universidade Federal Fluminense, Niterói, RJ, Brazil,
[3]Instituto de Computação, Universidade Federal do Amazonas, Manaus, AM, Brazil
{marina.ramalhete,daniela.vianna}@jusbrasil.com.br, {assisgabriel,alinepaes}@id.uff.br, alti@icomp.ufam.edu.br

Abstract

Automatic metrics are widely used to evaluate text quality across various natural language processing tasks. Despite their convenience and scalability, the extent to which these metrics reliably reflect textual quality remains an open challenge. The LLM-as-a-judge paradigm has recently emerged, aligning more closely with human judgments by using LLMs themselves as evaluators. However, there is still a gap in such evaluations across specific domains and languages, as most prior work focuses on generic task benchmarks in English. In this paper, we examine the robustness of both traditional automatic metrics and the LLM-as-a-judge approach for assessing the quality of financial commentaries in Portuguese, an underexplored task and language that has been neglected in previous work. We introduce fine-grained perturbations into the texts generated by specialists to analyze which types of noise most significantly affect evaluation outcomes, using noise-free counterparts as references. The results highlight the weaknesses of classical metrics in this specific task and the limitations of even recent evaluation paradigms, underscoring the need to develop context- and domain-sensitive evaluators.

1 Introduction

Artificial Intelligence is increasingly transforming the financial sector, especially through the automation of language-intensive tasks enabled by Large Language Models (LLMs) (Li et al., 2023; Huang et al., 2023; Wu et al., 2023). One key application is the generation of commentaries from mandatory corporate disclosures, a crucial step in financial communication that helps translate formal announcements into insights for investors, analysts, and the public (Zhu et al., 2023; Assis et al., 2025).

Assessing the quality, accuracy, and reliability of financial communication remains a challenging task, given the complexity and volume of information involved (Sai et al., 2022; Schmidtová et al., 2024). In addition to general linguistic requirements such as fluency and morphosyntactic correctness, financial texts must ensure the factual accuracy of entities, numerical data, and described events. Manual evaluation by domain experts remains the most trustworthy source of judgments; however, it is expensive, time-consuming, and subject to variability, which limits its scalability and reproducibility.

In this context, the widespread adoption of LLMs has made automatic text quality metrics increasingly vital (Sai et al., 2022; Schmidtová et al., 2024). Established metrics, such as ROUGE, BLEU, and METEOR (Caglayan et al., 2020), originally developed for tasks like summarization and machine translation, provide a way to quickly and scalably compare the quality and fidelity of texts. Nevertheless, several previous works reveal that excessive reliance on these metrics can be insufficient for evaluating the quality, accuracy, usefulness, and reliability of texts (Casola et al., 2025), and this is not different in the financial domain (Assis et al., 2024). Metrics based predominantly on lexical overlap or superficial n-gram similarity often fail to capture the semantic depth, contextual nuances, and, crucially, the factual veracity of the presented data (Maynez et al., 2020; Fabbri et al., 2021). In this way, some of those metrics can give high scores in texts for using phrases similar to a reference yet still contain misinterpretations of complex data or critical omissions of relevant entities (Maynez et al., 2020; Pagnoni et al., 2021).

This paper investigates the reliability of automatic evaluation metrics for assessing financial commentaries in Portuguese. Our approach systematically introduces textual noise into *expert-authored* texts and compares the perturbed version with the original using automatic metrics. We implement seven distinct perturbations that are

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 222–233

broadly applicable but particularly relevant to the financial domain. These include alterations that distort the content's meaning, such as entity substitutions, which are expected to degrade evaluation scores. Conversely, we also introduce modifications to preserve semantic integrity, such as replacing nouns with synonyms. While some changes are designed to maintain the perceived quality of the text, our evaluation reveals that certain substitutions can inadvertently shift the intended meaning.

Beyond traditional metrics, we emphasize recent advances in using LLMs as evaluators (Gu et al., 2025), prompting them to assess perturbed texts for factual correctness, numerical and entity accuracy, fluency, and overall writing quality.

Although noise-free texts may still lack higher-level qualities such as analytical depth or argumentative complexity, the absence of noise represents a baseline requirement for textual quality. Our goal is to assess whether existing evaluation approaches can reliably detect fundamental degradations even at a basic level of textual quality. If they fail to identify elementary forms of deterioration, this signals the need for caution when applying them to more sophisticated texts, particularly in domains like finance, where precision and clarity are essential.

Our main contributions are as follows: **(i.)** We investigate a critical yet underexplored domain for automatic evaluation metrics, their use to assess financial commentaries; **(ii.)** we focus on the Portuguese language within the financial domain, a combination that presents unique linguistic challenges such as grammatical gender and verb conjugation; **(iii.)** we evaluate the robustness and reliability of commonly used automatic metrics in capturing true text quality under these controlled perturbations, both with the systematic evaluation of metrics and using LLMs-as-judges as a framework. Lastly, (iv) our results show that neither classical metrics nor current LLM-based evaluation approaches can be safely relied upon as proxies for correctness in financial text evaluation, particularly in capturing structural disruptions, factual accuracy, and domain-specific nuances.

2 Related Work

As Natural Language Generation (NLG) capabilities have significantly advanced with LLMs, the importance of accurate evaluation methods has likewise increased. Several studies indicate that commonly used metrics can be misleading and require careful scrutiny (Freitag et al., 2023, 2024), including the emerging LLM-as-judge paradigm, which itself has notable limitations (Gao et al., 2025).

Consequently, a series of shared tasks (Freitag et al., 2023, 2024) has sought to critically examine NLG evaluation metrics, aiming to improve methodologies while frequently highlighting their limitations in assessing high-quality generated texts. Even evaluations employing LLMs-as-judges exhibit weaknesses. A comprehensive survey by Gao et al. (2025) identifies issues including biases related to text length, sensitivity to input order, variability across evaluation criteria, and inadequacies in assessing factual accuracy.

In line with our research, noise-based approaches to evaluate NLG have also been explored. Dušek et al. (2019, 2020) investigated semantic noise by analyzing errors in short sentences from the E2E dataset. Additionally, Xie et al. (2023) proposed DeltaScore, a noise-based metric specifically designed for storytelling narratives. Similarly, Thomson and Reiter (2020) examined errors involving incorrectly named entities, numerical inaccuracies, and contextual inconsistencies, proposing an annotation-based methodology to establish a gold standard for evaluating text accuracy.

Our work stands out by addressing two distinct aspects. While most evaluations focus on English, we address Portuguese, an underexplored language in NLG evaluation. Furthermore, we examine the financial domain, which demands high sensitivity from evaluation metrics, and investigate the performance of both traditional metrics and LLM-based methods in this context.

3 An Evaluation Framework Based on Textual Noise in Financial Commentaries

This section presents our evaluation framework for analyzing the effect of textual noise on the assessment of financial commentaries by automatic metrics, as illustrated in Figure 1. The process begins with a curated dataset of expert-written financial commentaries — human-authored texts that interpret or summarize the implications of official corporate disclosures likely to influence investor decisions. These texts are then subjected to carefully designed perturbations. We then analyze how standard evaluation metrics respond to the differences between original and noisy texts to examine the

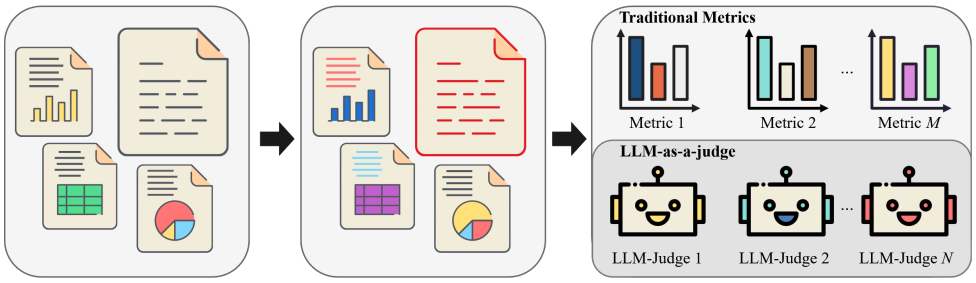

Figure 1: Overview of the pipeline: We first collect commentaries, then introduce controlled perturbations, and finally assess the outputs using automatic evaluation.

impact of these modifications. We also incorporate LLM-based evaluations, prompting the models to identify and characterize textual discrepancies.

Together, these steps enable us to evaluate both the stability of automatic metrics under controlled distortions and the feasibility of using LLMs to detect more subtle textual shifts. The following subsections detail each framework component.

3.1 Problem Statement

We consider a pair (x, t_x), where x denotes an official corporate disclosure (also known as a material fact), and t_x is a financial commentary derived from x by a domain expert. Let $p(t_x) \rightarrow \{t'_x\}$ be a perturbation function that receives a commentary t_x and produces one or more modified versions, forming a set $Y_p = \{t_{x_1}, \ldots, t_{x_k}\}$, with $k \geq 1$.

A metric function is used to compare two texts according to specific properties (e.g., morphosyntactic similarity, semantic equivalence, or overall quality). Such a function can either return a real-valued score $m(t_x, t'_x) \rightarrow \mathbb{R}$ or a categorial outcome $m(t_x, t'_x) \rightarrow \mathbb{C}$, depending on the nature of the evaluation (e.g., numerical similarity index vs. discrete quality class). Given a set of metrics $M = \{m_1, \ldots, m_w\}$, our goal is to identify which metrics in M exhibit stability or sensitivity with respect to each perturbation function, i.e., whether they maintain consistent outputs under minor controlled variations or respond appropriately to meaningful changes in the input text.

3.2 Data

Following Assis et al. (2025), our evaluation includes financial commentaries in Portuguese, collected directly from the websites of financial analysis firms, using a keyword-based filtering strategy to identify relevant events. We selected 72 commentaries from 35 companies, enabling coverage of diverse factors beyond the events themselves, such as geographical context, company size, sector, and monetary scope. The companies to which these analytical commentaries refer are Brasken, GPA, Magazine Luiza, Hidrovias do Brasil, LOG, GOL, Construtora Tenda, SLC Agrícola, Varejo Farma, Méliuz, Bradesco, Banco do Brasil, Agro Galaxy, Americanas, Natura, CEMIG, Oncoclínicas, ALLOS, Mater Dei, Sabesp, Casas Bahia, Rumo, Cogna, Petrobras, Vale, Hypera Pharma, Vibra, BRF, VBI Logística, CSHG Logística, Cosan, XP Malls, Telefônica Brasil S.A., Bresco Logística, and BTG Pactual. This selection allows us to capture a broad range of factors beyond the specific events and companies involved, including geographical context, company size, sector, and monetary dimensions.

3.3 Noise Introduction

We construct a synthetic dataset by applying controlled textual perturbations to expert-authored financial commentaries from our corpus. Specifically, we introduce seven types of noise: entity replacement, sentence reordering, typo insertion, synonym substitution, random insertion, random swapping, and random deletion. These noise types were selected to reflect both common errors across domains and risks specific to financial commentary. Typo insertion and synonym substitution simulate everyday mistakes that may occur during drafting or automated generation, allowing us to assess how effectively automatic metrics and LLMs-based evaluation handle surface-level distortions. In contrast, sentence reordering, entity replacement, and random insertion, deletion, and swapping introduce more consequential semantic shifts, particularly in finance, where misnaming a company or omitting

key information can lead to serious misreading.

Typo Insertion (TI) simulates orthographic errors by injecting typos using the spaCy `Augmenty`[1] library. Texts are processed with the Portuguese spaCy model, and typos are inserted in a controlled proportion of characters and samples. This noise typically preserves the original semantics.

Sentence Reordering (SR) splits the text into individual sentences and shuffles their order using spaCy for Portuguese sentence segmentation. While this often preserves the core meaning, it can alter the emphasis. In financial commentaries, however, disrupting the temporal and causal flow may mislead readers or obscure key recommendations.

Synonym Substitution (SS) selects N words from the sentence at random, excluding stop words. Each of these selected words is replaced with a randomly chosen synonym. This modification should not alter the linguistic and semantic evaluation.

Random Insertion (RI) refers to the process of inserting words into a sentence at random positions. This is achieved by randomly selecting N words in a sentence, ensuring that they are not stop words. Then, synonyms of these words are inserted at random positions in the sentence, all while keeping the original words untouched. This noise should alter the syntax and semantics of the text.

Random Swapping (RS) randomly selects two words and switches their positions in the text, ensuring that all other elements remain unaffected.

Entity Replacement (ER) identifies organization entities ("ORG") using spaCy's optimized Portuguese model[2], then replaces them with others of the same type via the Faker library[3]. These preserve structure while altering content, ideally signaling a loss of meaning, though some substitutions may appear plausible.

Random Deletion (RD) This noise operates by iterating through each word in a sentence and removing it with a fixed probability p.

Table 1 presents examples of modified versions of a commentary to better illustrate each noise type.

Regarding the volume of noise, we maintained a consistent noise level across all perturbations: 50% of sentences per commentary were modified, with 5% of words perturbed within each altered sentence.

[1] https://spacy.io/universe/project/augmenty
[2] https://spacy.io/models/pt
[3] https://pypi.org/project/Faker/

3.4 Metrics Implementation

We investigate the extent to which different metrics can reliably assess synthetically altered texts. Our framework encompasses lexical, semantic, alignment, and automatic evaluation scenarios. Regarding **lexical overlap evaluation**, we adopt ROUGE (Lin, 2004) to measure content similarity based on n-gram co-occurrence statistics. ROUGE comprises a set of automatic evaluation metrics originally developed for summarization tasks. We used the ROUGE-1, ROUGE-2, and ROUGE-L variants, which quantify the overlap between candidate and reference texts based on unigrams, bigrams, and longest common subsequences, respectively. We also include BLEU (Papineni et al., 2002), another classical metric based on n-gram precision. Unlike ROUGE, which emphasizes recall by measuring how much of the reference is captured by the candidate, BLEU focuses on precision, assessing how much of the candidate matches the reference. They remain widely adopted due to their computational efficiency and ease of interpretation. We chose to retain BLEU and ROUGE as baseline metrics due to their continued use in financial NLP evaluations, despite their known limitations (Yan, 2022; Wang et al., 2024; Assis et al., 2024).

To evaluate the **semantic similarity** metrics, we adopt BERTScore (Zhang et al., 2019), which relies on contextualized embeddings and token probability analysis. Our implementation relies on mBERT (Devlin et al., 2019) given that the current implementation of BERTScore does not provide direct support for models trained in Portuguese.

Our framework further incorporates **interpretable alignment-based metrics** via the CTC framework (Deng et al., 2021), offering a unified evaluation paradigm for natural language generation grounded in information alignment. We used two key metrics from this framework: consistency, which checks how well the generated content aligns with the original input, and groundedness, which assess whether the output is grounded in external knowledge or context. These measures are particularly valuable for assessing factual accuracy and relevance, especially in cases where staying true to the source material is crucial, as is the case with commentaries based on material facts.

Finally, we also rely on **LLMs-as-a-judges**. This paradigm is gaining traction due to its scalability and potential to simulate manual evalua-

Noise type	Noisy output	Noisy type	Noisy output
ER	The Petrobras recorded higher revenue in the quarter, supported by stronger international sales. Operating expenses declined slightly due to cost-control initiatives. As a result, net profit rose 8% compared to the previous period.	SR	Operating expenses declined slightly due to cost-control initiatives. The Braskem reported higher revenue in the quarter, supported by stronger international sales. As a result, net profit rose 8% compared to the previous period.
TI	The Braskem reported higher revenue in the quarter, supported by #tronger international sales. Operating expenses declined slightly du1 to cost-control initiatives. As a result, net profit rose 8% compared to the previous period.	SS	The Braskem recorded higher income in the quarter, supported by stronger international sales. Operating expenses declined slightly due to cost-control programs. As a result, net profit rose 8% compared to the previous period.
RI	The Braskem reported higher earnings revenue in the quarter, supported by stronger international sales. Operating expenses declined slightly due to cost-control initiatives. As a result, gain net profit rose 8% compared to the previous period.	RS	The Braskem higher reported revenue in the quarter, supported by stronger international sales. Operating declined expenses slightly due to cost-control initiatives. As a result, profit net rose 8% compared to the previous period.
RD	The <> reported higher revenue in the quarter, supported by stronger international sales. Operating expenses <> slightly due to cost-control initiatives. As a result, net profit rose 8% compared to the previous period.		

Table 1: English-translated examples of noise injection in financial commentaries. The original text is: "*The Braskem reported higher revenue in the quarter, supported by stronger international sales. Operating expenses declined slightly due to cost-control initiatives. As a result, net profit rose 8% compared to the previous period.*"

tion in complex and open tasks (Gu et al., 2025). LLMs are instructed[4] to provide an overall classification of each text as *good* or *bad*, and to indicate whether any inconsistencies were present concerning specific aspects such as writing quality, fluency, numerical accuracy, named entities, and factual consistency. Our goal is to assess the models' ability to detect textual noise and deliver consistent quality judgments, thereby evaluating their reliability as automated reviewers. We incorporate six open-source language models into our framework with varying architectures and sizes: Llama3.1 (8B) (Grattafiori et al., 2024), DeepSeek-R1 (8B, 14B, 32B) (DeepSeek-AI, 2025), Phi-4 (14B) (Abdin et al., 2024), and Gemma-3 (27B) (Gemma-Team, 2025). Reproducibility and transparency drive the choice of open-source models. Each LLM is expected to evaluate whether a financial commentary is good (high quality) or bad (low quality) based on five criteria explained next. Failures or unexpected outputs are marked as "na".

We evaluate five factors: (i) **Factual accuracy (Facts)**: correctness of information relative to the original commentary; (ii) **Entity accuracy (Entity)**: consistency of named entities across versions; (iii) **Numerical value accuracy (Values)**: correctness of financial figures, independent of context; (iv) **Fluency**: clarity and coherence of the text; (v) **Writing quality (Writing)**: presence of grammatical errors.

4 Results and Discussion

This section discusses the results obtained from the classical and LLM-as-a-judge paradigms.

[4]The prompt is available at A.

4.1 Automatic Metrics

This subsection analyzes the results of BLEU, BERTScore, ROUGE, and CTC, which capture different aspects of semantic, lexical, and factual consistency. Higher scores indicate closer alignment with the original, noise-free texts. Table 2 reports the outcomes of applying these metrics, treating the noisy versions as targets and the expert-written texts as ground truth.

The BLEU scores reveal significant variation across noise types. *Typo Insertion (TI), Entity Replacement (ER)*, and *Sentence Reordering (SR)* maintain relatively high BLEU scores of 0.78, 0.92, and 0.97, respectively, indicating minimal disruption to lexical similarity. In contrast, *Random Deletion (RD)* and *Random Swapping (RS)* yield the lowest BLEU scores (0.05), underscoring their substantial impact on the text's coherence and fluency. This is due to its primary nature of capturing lexical and word-order similarity, not meaning, grammar, or readability. This way, we can conclude that BLEU is stable under lexical disruptions but not sensitive to semantic or structural distortions.

BERTScore, which evaluates semantic similarity, shows that *Entity Replacement (ER)* achieves the highest precision, recall, and F1 scores (0.96, 0.97, and 0.97, respectively). This suggests that replacing entities has a limited effect on the metric's perception of the overall semantic structure. Conversely, *Random Deletion (RD), Random Insertion (RI)* and *Random Swapping (RS)* yield the lowest scores, with F1 values of 0.68, 0.66, and 0.63, respectively, reflecting a significant degradation in semantic alignment. This shows that BERTScore effectively distinguishes between mild and severe

Noise	BLEU	BERTScore			ROUGE				CTC			
		prec.	rec.	f1	R1	R2	RL	RLsum	ground.	ground.$_{ref}$	fact.	fact.$_{ref}$
Typo Insertion (TI)	0.78	0.90	0.93	0.92	0.90	0.81	0.90	0.90	301.88	417.16	0.66	0.91
Random Deletion (RD)	0.05	0.69	0.66	0.68	0.63	0.35	0.63	0.60	231.91	275.05	0.63	0.74
Random Insertion (RI)	0.21	0.65	0.68	0.66	0.74	0.51	0.74	0.74	277.44	333.50	0.60	0.72
Random Swapping (RS)	0.05	0.63	0.64	0.63	0.96	0.26	0.40	0.57	276.50	335.33	0.62	0.75
Synonym Substitution (SS)	0.19	0.72	0.72	0.72	0.70	0.46	0.68	0.68	276.39	342.62	0.61	0.76
Entity Replacement (ER)	0.92	0.96	0.97	0.97	0.96	0.94	0.96	0.96	301.60	431.22	0.66	0.94
Sentence Reordering (SR)	0.97	0.85	0.86	0.85	1.00	0.97	0.45	0.97	308.03	366.74	0.68	0.81

Table 2: Semantic, lexical, and factual metrics by noise type (higher values suggest better quality).

distortions (captures semantic degradation), but it is insensitive to entity-level factual changes and role reversals. This way, for tasks where factual accuracy or entity identity is crucial (such as financial texts), BERTScore alone is insufficient.

The ROUGE metrics, which measure overlap in n-grams and sequences, further highlight how noise can go unnoticed by traditional metrics. *Entity Replacement (ER)* consistently achieved high scores across all ROUGE variants used. Meanwhile, *Sentence Reordering (SR)* showed high results for all variants except ROUGE-L, reflecting the disruption of longer sequences caused by noise. *Random Deletion and Insertion (RD and RI)*, along with *Synonym Substitution (SS)*, led to the most consistent drops across the ROUGE variants, as somewhat expected. This means that ROUGE behaves similarly to BLEU, mainly capturing surface similarity but not deeper semantic or structural coherence.

The CTC metric evaluates aspects of groundedness and factuality based on alignment. Here, '`ground.`$_{\texttt{ref}}$' and '`fact.`$_{\texttt{ref}}$' indicate that the ground truth provided to the metric was the noiseless financial commentary, while '`ground`' and '`ref`' were calculated using the corporate document mentioned in the texts. Overall, the noisy texts exhibit stronger alignment with the noise-free commentary, although differences are observed across error types. *Sentence Reordering (SR)*, *Entity Replacement (ER)*, and even *Typo Insertion (TI)* maintain high scores on the metric, whereas the others exhibit more pronounced drops — particularly *Random Deletion (RD)*, which stands out negatively. These results indicate that CTC captures both semantic and factual alignment, not just surface similarity. High scores for ER and TI imply that factual grounding remains largely intact, while low scores for RD and similar perturbations suggest that deleting content compromises factual completeness and coherence.

Overall, deletion, insertion, and word-level swapping errors are more easily captured by automatic metrics, which align with their design, whereas full-sentence reordering shows limited effects. Notably, *Entity Replacement (ER)* consistently yields high scores, suggesting minimal impact on text quality when interpreted strictly by the metric outcomes. However, this overlooks a critical issue: in financial analyses, changes to entities can significantly affect accuracy and validity, as entities carry specific financial contexts and implications. Consequently, high ER scores highlight a limitation of these metrics to potentially assess automatically generated financial texts.

4.2 LLM-as-a-judge

This section presents a comparative evaluation of six language models, each serving as a judge in assessing financial commentaries. The models are evaluated based on their ability to identify and assess the distortions introduced in Section 3.3 when compared against the original, unaltered financial commentaries. All reported results represent the average outcomes across three independent runs.

In Section 4.2.1, we discuss the overall quality of financial commentaries after the introduction of each type of noise. Next, we extend the evaluation in Section 4.2.2 by analyzing how each LLM perceives the impact of textual noise, focusing on the five key factors introduced in Section 3.4.

4.2.1 Overall results

Table 3 summarizes the overall performance of each model in evaluating the 72 financial commentaries, each subjected to one of the seven categories of textual noise. Examining the results in Table 3, we observe a notable divergence between deepseek-r1:8b and the other models in assessing commentary quality. Regardless of the noise type, deepseek-r1:8b tends to classify most texts as high quality, particularly under *Typo Insertion (TI)* and *Sentence Reordering (SR)*, which appear to have no adverse effect on its evaluations. Although deepseek-r1:8b

Judge	TI			RD			RI			RS			SS			ER			SR		
	✓	✗	na	✓	✗	na	✓	✗	na	✓	✗	na	✓	✗	na	✓	✗	na	✓	✗	na
llama3.1:8b	31	40	1	4	67	1	1	69	2	4	67	1	2	69	1	24	48	0	43	29	0
gemma-3:27b	33	39	0	0	72	0	0	72	0	0	72	0	0	72	0	7	65	0	15	57	0
deepseek-r1:8b	60	11	1	48	24	0	54	18	0	49	23	0	48	23	1	47	25	0	61	10	0
deepseek-r1:14b	38	32	2	10	61	1	9	61	2	14	55	2	8	62	2	27	42	3	40	29	4
deepseek-r1:32b	34	36	2	3	69	1	6	65	1	6	66	0	5	65	2	22	48	3	30	41	1
phi4:14b	32	40	0	0	72	0	0	72	0	0	72	0	1	71	0	16	56	0	45	27	0

Table 3: LLM-as-a-judge results. Symbols indicate model judgments: ✓ for "good", ✗ for "bad", and na for no response. Values represent averages over three runs.

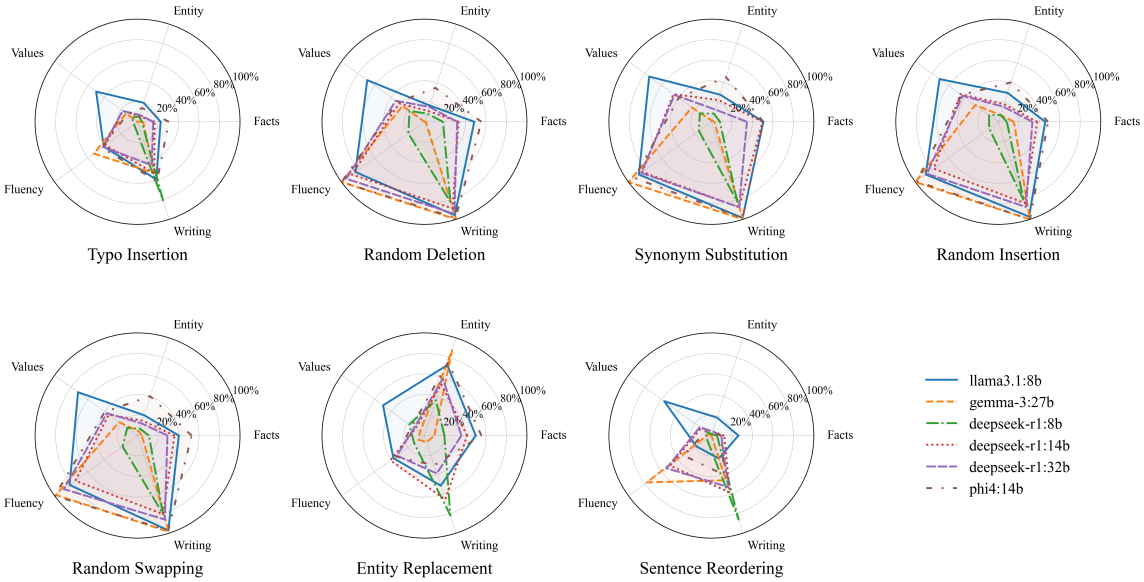

Figure 2: Radar charts of the impact of each noise type on five key aspects of financial commentaries, as judged by LLMs. Each axis indicates the proportion of detected issues per dimension.

is a distilled version of llama-3.1:8b (DeepSeek-AI, 2025), it showed to be less reliable in detecting textual disruptions. In contrast, llama-3.1:8b shows greater sensitivity to noise.

Among the DeepSeek-R1 family (8b, 14b, 32b), performance improves with model size, especially in identifying low-quality analyses. These models are most responsive to *Random Deletion (RD)*, *Random Insertion (RI)*, *Random Swapping (RS)*, and *Synonym Substitution (SS)*.

Overall, gemma-3:27b, deepseek-r1:32b, and phi-4:14b exhibit more consistent performance across noise types, but still fail to provide stable and comprehensive judgments across all evaluation dimensions. Notably, gemma-3:27b is the most sensitive to *Entity Replacement (ER)*, whereas deepseek-r1:32b struggles more to detect *Typo Insertion (TI)* as problematic. In contrast, phi-4:14b is less sensitive to *Sentence Reordering (SR)* errors. Interestingly, all models, except the previously mentioned deepseek-r1:8b, perceive *Synonym Substitution (SS)* as harmful.

4.2.2 Impact of Noise on Performance

Here, we expand our evaluation by examining how each LLM responds to the presence of textual noise. Figure 2 presents a radar chart for each type of inserted noise. Its dimensions show the cases in which the LLMs indicated that the noise may affect one of the five key factors outlined in Section 3.4, assessing its impact across them. Specifically, the chart shows the answer "no" for most key factors and "yes" for *Writing*, as can be verified in the prompt in appendix A.

Regarding **Typo Insertion (TI)**, although any key factor may be affected, *Fluency* and *Writing* are expected to be most impacted. This held for most models on *Fluency*, except deepseek-r1:8b, which often claimed the text remained "*fluid and coher-*

ent". For *Writing*, most models correctly flagged errors, but the Deepseek family was inconsistent: "yes" sometimes meant errors were found, and other times that quality was preserved, making their answers unreliable for this factor.

The llama3.1:8b showed an unexpected pattern for *Values*: while it sometimes acknowledged typos (e.g., "*The financial values... are the same... However, typos and formatting errors were observed*"), it often failed to compare values accurately, suggesting confusion rather than noise sensitivity.

The **Random Deletion (RD)** chart shows that this noise mainly affected *Fluency* and *Writing*, as expected, since omissions disrupt coherence and grammar. Among models, deepseek-r1:8b detected the fewest inconsistencies, while its 32b version was more critical, citing fragmentation and discontinuity. *Values* again emerged as a sensitive dimension, especially for llama3.1:8b, which flagged omissions in the financial texts with detailed justifications. Nonetheless, most models did not penalize deletions unless numeric discrepancies were explicit, suggesting a limited ability of llama3.1:8b to infer missing values as inaccuracies. For *Facts*, llama3.1:8b and phi-4:14b flagged more errors, citing missing or incomplete content. On *Entities*, phi-4:14b was notably context-aware, penalizing cases where framing was lost. For instance, it noted: "*The entities mentioned are partially referenced, but in an incorrect and confusing manner*", while llama3.1:8b simply stated: "*The target and golden analyses mention the same main entities...*"

The impact of **Synonym Substitution (SS)** closely resembled that of Random Deletion (RD). For *Facts*, models gave mixed responses. Some flagged semantic drift, e.g., "*The target analysis alters the meaning... introducing misinterpretations regarding the sale of Bankly, the financial terms, and the proposed actions*". Others accepted paraphrasing, stating "*The target analysis reports the same facts... albeit with different wording*". In some cases, substitutions went entirely undetected.

Entity-level substitutions using synonyms revealed weaknesses in the LLMs' evaluation. For example, replacing "*Vale*" (a company) with "*Baixada*" (a geographical term that matches the meaning if "Vale" were interpreted as a common noun rather than a proper name) was flagged only by phi-4:14b and llama3.1:8b, revealing inconsistent model sensitivity. *Values* were also affected, particularly when substitutions involved named locations or project identifiers. For instance, changing "Platform" to "Pier" led to justifications citing unit inconsistencies and unclear quantitative references.

Both **Random Insertion (RI)** and **Random Swapping (RS)** mirror the behavior observed in previous noise types. *Fluency* and *Writing* remain the most affected dimensions. Notably, gemma-3:27b shows a contrasting profile: it is less effective at detecting issues in *Facts*, *Entities*, and *Values*, but consistently sensitive to disruptions that may affect *Fluency* and *Writing*. In these cases, it frequently returns comments such as "*The text in the target commentary is extremely confusing, with unnecessary repetitions and words out of context*" and "*The target commentary contains numerous grammatical errors, repetitions, and meaningless words*", highlighting its responsiveness to syntactic and lexical noise. As with other dimensions, deepseek-r1:8b detects more issues in *Writing*, but struggles to capture impacts on other aspects.

Entity Replacement (ER) impacts a distinct set of dimensions. As expected, *Entity* is most affected, followed by *Facts*, given ER's potential to undermine truthfulness and relational integrity. gemma-3:27b shows particular sensitivity to *Entity*, while phi-4:14b is more responsive to *Facts*. Nonetheless, detection rates across models remain well below 100%, suggesting limited coverage.

Lastly, **Sentence Reordering (SR)** reveals a distinct pattern. Most models do not perceive sentence reordering as highly disruptive. This is expected, as SR might make the texts less fluid, while maintaining their overall meaning. An exception is gemma-3:27b, which penalizes *Fluency*, along with previously noted deviations, llama3.1:8b on *Values* and deepseek-r1:14b on *Writing*.

In general, deepseek-r1:32b and phi-4:14b show the most balanced performance, with phi-4:14b standing out despite its smaller size. While gemma-3:27b occasionally leads to error detection, it fails markedly in other areas. Both llama3.1:8b and deepseek-r1:8b fail to provide consistent responses.

5 Conclusion

This paper examined the robustness of automatic text evaluation by introducing fine-grained perturbations into reference texts in a novel task of financial commentary generation in Portuguese. We assess the impact of seven types of textual noise using both traditional metrics and the LLM-as-a-judge paradigm.

Classical metrics reveal important limitations:

BLEU and ROUGE detect surface disruptions but struggle with structural changes, while BERTScore and CTC often overlook critical domain-specific alterations such as entity replacements.

LLM-as-a-judge evaluations show notable variation across models. Smaller models, such as deepseek-r1:8b, frequently overrate noisy texts, whereas larger models, like phi-4:14b and deepseek-r1:32b, demonstrate more balanced, context-aware assessments. Gemma-3:27b is particularly sensitive to fluency issues but less reliable on factual dimensions. Challenges persist across models in handling entity substitutions and implicit value inconsistencies, highlighting limitations in current evaluation approaches.

Future work will explore the inclusion of LLM-generated texts, the design of a rigorous qualitative assessment framework with domain experts in the loop, and the development of context- and domain-sensitive evaluation strategies, potentially leveraging ensembles of cooperative LLMs (Verga et al., 2024).

Limitations

Although we aimed to investigate the robustness of automatic evaluation in Portuguese financial commentaries, we recognize that our conclusions may not generalize to other languages, domains, or tasks. While additional metrics could have been considered, many recent approaches require model training and are not readily applicable to Portuguese. Finally, we acknowledge that a more robust and systematic human evaluation, particularly involving domain and linguistics experts, could offer further insights into the observed results.

Ethical Considerations: The financial texts used in this study cannot be broadly redistributed due to authorship and distribution policies established by the original financial research firms. However, the texts are publicly available and can be accessed through the data collection procedure described in the paper. This way, this study respects the intellectual property and distribution policies of the original financial research firms. All data were used in accordance with the terms of access and solely for research purposes, ensuring compliance with ethical and legal standards.

Acknowledgments

This research was partially financed by CNPq (National Council for Scientific and Technological Development), FAPERJ - *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*, processes SEI-260003/002930/2024, SEI-260003/000614/2023, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We also thank the support of the CNPq National Institutes of Science and Technology, IAIA (grant 406417/2022-9) and TILD-IAR (grant 408490/2024-1).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report.

Gabriel Assis, Hugo Dutra, Daniela Vianna, Júnior Meira, Wagner, Altigran Soares da Silva, and Aline Paes. 2025. Language Models for Automated Market Commentary from Corporate Disclosures. In *Proceedings of the 6th ACM International Conference on AI in Finance*, ICAIF '25, page 727–735, New York, NY, USA. Association for Computing Machinery.

Gabriel Assis, Daniela Vianna, Gisele L. Pappa, Alexandre Plastino, Wagner Meira Jr, Altigran Soares da Silva, and Aline Paes. 2024. Analysis of material facts on financial assets: A generative AI approach. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 103–118, Torino, Italia. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Silvia Casola, Yang Janet Liu, Siyao Peng, Oliver Kraus, Albert Gatt, and Barbara Plank. 2025. Evaluation should not ignore variation: On the impact of reference set choice on summarization metrics. *arXiv preprint arXiv:2506.14335*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the*

2021 Conference on Empirical Methods in Natural Language Processing, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426, Tokyo, Japan. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Computer Speech and Language*, 59:123–156.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs Breaking MT Metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. LLM-based NLG Evaluation: Current Status and Challenges. *Computational Linguistics*, 51(2):661–687.

Gemma-Team. 2025. Gemma 3 Technical Report.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 21 others. 2024. The Llama 3 Herd of Models.

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.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 374–382, New York, NY, USA. Association for Computing Machinery.

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.

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, Online. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Patrícia Schmidtová, Saad Mahamood, Simone Balloccu, Ondřej Dušek, Albert Gatt, Dimitra Gkatzia, David M Howcroft, Ondřej Plátek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. *Preprint*, arXiv:2011.03992.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models.

Ziao Wang, Yunpeng Ren, Xiaofeng Zhang, and Yiyuan Wang. 2024. Generating long financial report using conditional variational autoencoders with knowledge distillation. *IEEE Transactions on Artificial Intelligence*, 5(4):1669–1680.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Zhuohan Xie, Miao Li, Trevor Cohn, and Jey Lau. 2023. DeltaScore: Fine-grained story evaluation with perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5317–5331, Singapore. Association for Computational Linguistics.

Sixing Yan. 2022. Disentangled variational topic inference for topic-accurate financial report generation. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 18–24, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Di Zhu, Theodoros Lappas, and Thami Rachidi. 2023. Commentary generation for financial markets. *Expert Systems with Applications*, 211:118364.

A Prompt for LLM as Evaluators

Figure 3 shows the system prompt used when employing LLMs as evaluators.

Prompt

```
You are a financial expert. Your role is to evaluate the quality of a financial
commentary (target commentary) compared with a commentary authored by domain
specialists (golden commentary).

Consider the following points:

Facts: Check whether the facts presented in the target commentary are correct
with respect to the golden commentary. Answer "yes" when the facts are correct;
otherwise, answer "no."

Entities: Check whether the entities in the golden commentary are preserved in
the target commentary. Answer "yes" if the entities are the same in both
analyses; otherwise, answer "no."

Values: Verify whether the financial figures are identical in both analyses.
Consider only the numerical values, regardless of any surrounding context.
Answer "yes" if the values match; otherwise, answer "no."

Fluency: Is the target commentary text fluent and coherent? Answer "yes" if the
text is fluent and coherent; otherwise, answer "no."

Writing: Does the target commentary contain grammatical errors? Answer "yes"
if there are grammatical errors in the target commentary; otherwise, answer
"no."

Conclusion: Finally, classify the target commentary as "good" or "bad" with
respect to the golden commentary.

Items: Based on the five criteria above—Facts, Entities, Values, Fluency, and
Writing—list which of these items are incorrect in the target commentary.

Your answer must be in JSON format following this structure:

{
  "Facts": {
    "Comment": <string>,
    "Answer": "yes" or "no"
  },
  "Entities": {
    "Comment": <string>,
    "Answer": "yes" or "no"
  },
  "Values": {
    "Comment": <string>,
    "Answer": "yes" or "no"
  },
  "Fluency": {
    "Comment": <string>,
    "Answer": "yes" or "no"
  },
  "Writing": {
    "Comment": <string>,
    "Answer": "yes" or "no"
  },
  "Conclusion": "good" or "bad",
  "Items": ["Facts", "Entities", "Values", "Fluency", "Writing"]
}
```

Figure 3: System prompt provided to the LLMs-as-judges evaluation.