Do LLM Judges Agree with Humans? Evaluating Financial Commentaries from Material Facts

Gabriel Assis[1], **Daniela Vianna**[2], **Livy Real**[3,4], **Marina Ramalhete Masid**[2], **João Nepomuceno**[1], **Eduardo Rottschaefer**[1], **Altigran Soares da Silva**[3], **Aline Paes**[1]

[1] Universidade Federal Fluminense, Niterói, RJ, Brasil
{assisgabriel@id,joaonepomuceno@id,eduoliveira@id,alinepaes@ic}.uff.br,
[2]Jusbrasil, Salvador, BA, Brasil
{daniela.vianna,marina.ramalhete}@jusbrasil.com.br,
[3]Universidade Federal do Amazonas, Manaus, AM, Brasil
{livy,alti}@icomp.ufam.edu.br,
[4]Instituto Kunumi, Belo Horizonte, MG, Brasil

Abstract. *The rapid adoption of large language models (LLMs) in finance has enabled automated generation of in-domain texts such as earnings summaries, market analyses, and commentary on regulated disclosures. Generating accurate and accessible financial commentary from material facts poses challenges related to domain-specific language, strict factual faithfulness, and readability. Evaluating such outputs is difficult: traditional automatic metrics overlook financial correctness and adequacy, while human expert assessment is reliable but costly and subjective. This paper proposes a multidimensional evaluation protocol for generating financial commentary in Portuguese and investigates the use of LLMs as evaluators ("LLMs-as-judges") in this high-stakes, low-resource setting. We systematically compare human expert judgments and LLM-based evaluation preferences, while also investigating complementary dimensions such as writing quality, factuality, usefulness, and simplicity. To the best of our knowledge, this is the first proposal of an evaluation protocol for this task in Portuguese. Furthermore, we analyze where LLM judgments align with or diverge from human assessments, providing practical insights and recommendations for evaluation methodologies in financial AI applications.*

1. Introduction

Financial Commentary Generation from Material Facts is the task of producing analytical narratives from regulatory disclosures that publicly listed companies are legally required to release [Assis et al. 2025a]. "Financial commentary" refers to interpretive texts that explain, contextualize, and analyze such disclosures or related market events, ultimately helping investors make informed decisions [Zhu et al. 2023]. In practice, investors often rely on expert-written reports from financial analysts and investment research firms, motivating recent efforts to automate parts of this analytical work with language models.

Recently, the rise of large language models (LLMs) has created opportunities for financial text generation in multiple languages, including Portuguese, with applications ranging from earnings summaries [Mukherjee et al. 2022] to market analysis [Xiao et al. 2025] and investment decision support [Assis et al. 2025a]. In particular, the Brazilian financial market presents important linguistic and domain-specific challenges that remain underexplored in predominantly English-centric research. At the same

time, finance is a high-stakes domain where errors can lead to substantial financial consequences. As a result, generated commentaries must balance readability and coherence with factual consistency, faithfulness, and domain-specific precision [Masid et al. 2026].

These requirements make the evaluation of financial commentary generation particularly challenging, as quality is multi-dimensional and often depends on specialized financial interpretation. Although traditional metrics such as BLEU [Papineni et al. 2002] and ROUGE [Lin 2004] have been previously adopted in financial text generation and related NLG tasks, their evaluation is primarily based on lexical overlap, limiting their ability to capture factual accuracy, financial validity, and deeper semantic aspects of generated commentaries. Semantic metrics such as BERTScore [Zhang et al. 2020] provide a richer semantic comparison between texts than lexical-overlap metrics, yet they may still overlook domain-specific inaccuracies or financially meaningful errors [Masid et al. 2026].

As a result, recent work has increasingly investigated the use of LLMs as evaluators [Masid et al. 2026]. The *LLM-as-a-judge* paradigm [Zheng et al. 2023, Liu et al. 2023] relies on the ability of these models to approximate human preferences while supporting scalable evaluation across multiple dimensions. In this work, we employ a *"sequential knockout pairwise evaluation"* protocol [Sandan et al. 2025], in which outputs are compared pairwise across successive evaluation rounds. This approach enables scalable analyses of evaluator agreement and robustness through comparative assessments by elimination, reducing the cost and effort associated with extensive human evaluation. However, its robustness remains unexplored in high-quality evaluation settings for Portuguese financial commentaries, where distinguishing among strong candidate outputs is especially challenging [Assis et al. 2025a, Masid et al. 2026]. In addition, LLM-based evaluators also present important limitations, such as susceptibility to fluent but factually inaccurate generations, sensitivity to prompt design, and biases toward specific linguistic styles, verbosity levels, or response structures [Chen et al. 2024, Gao et al. 2025].

In this work, we compare automatic metrics, human judgments, and LLM-based evaluations to investigate the following research questions: **RQ1.** *Under the proposed knockout tournament protocol, which LLMs emerge as the strongest generators of financial commentary?*; **RQ2.** *How do automatic metrics, LLM-based evaluations, and human assessments correlate?*; and **RQ3.** *Where do systematic disagreements arise across evaluations?*. We also analyze human annotations independently to better understand agreement patterns and the dimensions that pose the greatest challenge for expert evaluation.

To investigate these questions, our sequential knockout evaluation is inspired by prior work on pairwise preference evaluation [Sandan et al. 2025, Thakur et al. 2024] and arena-style comparisons [Zheng et al. 2023]. In this setting, commentaries are compared pairwise, with the winner advancing to the next round. Evaluators may optionally consider dimensions such as writing quality, factuality, usefulness, and simplicity when eliciting their preferences.

The strategy is instantiated using LLM4FinComm [Assis et al. 2025a], a dataset of Portuguese financial commentaries generated by both humans and LLMs. Human judgments are collected through a web-based evaluation platform[1] developed for this study, with annotations from three evaluators: a linguist and two financial-domain

[1] GitHub: https://github.com/AIDA-BR/human_and_judge_eval

experts. The same approach is further reproduced with three flagship LLMs — Qwen3-235B [Yang et al. 2025], Claude Opus 4.6 [Anthropic 2026], and Gemini 3.1 Pro [Google DeepMind 2026].

Our main contributions are threefold: (i) we propose a sequential knockout evaluation protocol for financial commentary assessment; (ii) we introduce the first benchmark for Portuguese financial commentary evaluation comparing human experts and LLM-based judges; and (iii) we provide an analysis of systematic disagreements between human and LLM evaluation preferences. The results reveal three main findings. First, GPT-4o and Llama-4-Scout emerge as strong candidates in our evaluation. Second, automatic metrics show moderate alignment with the aggregated rankings and can reliably identify strong candidates but fail to discriminate among top-performing models. Third, the LLM judges and human experts employed as evaluators in this experiment differed in meaningful ways, capturing different aspects of commentary quality.

2. Related Work

Financial text generation and evaluation have largely relied on traditional automatic metrics, particularly lexical-overlap measures such as ROUGE and its variants [Celikyilmaz et al. 2020]. Prior work spans financial report generation [Yan 2022], earnings call summarization [Mukherjee et al. 2022], and equity research report generation [Jin et al. 2026], with evaluation signals ranging from BLEU and ROUGE to semantic similarity and LLM-based components.

Recent studies also explore LLMs as judges of financial text generation [Xie et al. 2024]. Particularly, [Goldsack et al. 2025] show that such approaches correlate with expert assessments across multiple dimensions, though no single LLM consistently achieves the best correlation, and systematic biases such as positional bias and self-preference [Wataoka et al. 2024] effects remain a concern. Complementing this, evidence from the Earnings2Insights shared task [Takayanagi et al. 2025] reveals a divergence between systems that perform well under automated evaluation and those that better support human decision-making.

Research on Portuguese financial texts highlights additional challenges in transferring evaluation methodologies beyond English. [Masid et al. 2026] show that both traditional metrics and LLM-as-a-judge approaches struggle under controlled perturbations, often failing to detect subtle semantic and factual degradations. Similar findings are reported by [Assis et al. 2025a], who argue that automatic metrics lack sensitivity to fine-grained differences among outputs of comparable quality. Alternative QA-style approaches [Ribeiro et al. 2026] decompose evaluation into atomic factual checks, offering a complementary perspective for Portuguese financial summarization.

Our work complements this line of research by examining both human and LLM-based evaluation in the context of Portuguese financial commentary generation. Rather than relying on direct scoring, we investigate a successive pairwise comparison setup in which outputs are evaluated through iterative preference judgments.

3. A Multi-Round Pairwise Strategy for Evaluating Financial Commentaries

We frame evaluation as a sequence of pairwise preference judgments, rather than assigning absolute scores to individual outputs. This choice is motivated by the subjective

and context-dependent nature of financial commentary evaluation: different outputs may satisfy the same communicative goal in different ways, making absolute score calibration difficult across examples, annotators, and models. Pairwise comparison, in contrast, asks annotators to make a localized decision between two commentaries generated for the same material fact, allowing the evaluation to focus on concrete differences in writing quality, factuality, usefulness, and simplicity. Prior work further supports this design, showing that pairwise LLM-as-a-judge evaluations tend to better align with human judgments [Thakur et al. 2024], and that sequential pairwise evaluation strategies can reduce annotation costs while preserving correlation [Sandan et al. 2025].

For each material fact, N generated commentaries are organized into $N-1$ comparisons. In each comparison, the annotator, either human or LLM-based, observes the material fact and two candidate commentaries, selects the preferred one, and may optionally mark the criteria underlying the decision (writing quality, factuality, usefulness, and simplicity). The winner is carried forward as the current holder and compared with the next challenger until all commentaries have been assessed. Figure 1 provides an overview of this strategy. In the following subsections, we describe the data used (§ 3.1), the comparison criteria (§ 3.2), the human annotation process (§ 3.3), the annotation interface (§ 3.4), and the LLM-based evaluation setup (§ 3.5).

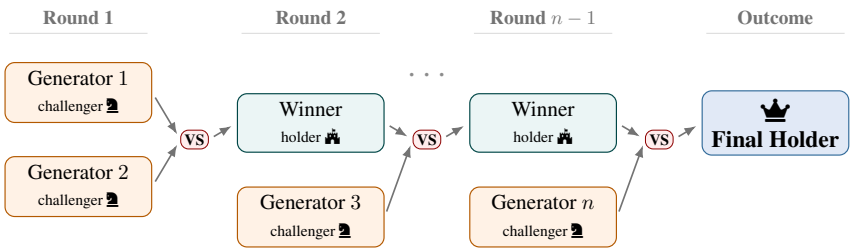

Figure 1. Multi-round pairwise evaluation under a sequential knockout strategy. Winners remain as holders and face the next challenger.

3.1. Data Selection

We use LLM4FinComm [Assis et al. 2025a] to evaluate our approach on Portuguese financial commentaries. The dataset contains 88 material fact–human-written commentary pairs. For each material fact, 9 LLM-generated commentaries are available, resulting in 792 generated texts and 880 texts in total when human references are included.

Since evaluating the full dataset would be impractical with the available human annotation resources (as discussed in § 3.3), we select a smaller subset focused on high-performing candidates. This choice aligns with RQ1, as our goal is to identify which generators are most effective for generating financial commentary. Let $\mathcal{F}$ be the set of material facts and $\mathcal{G}$ the set of LLM generators. For each pair (f, g), LLM4FinComm provides automatic scores against the human reference, including BLEU, ROUGE, BERTScore and CTC [Deng et al. 2021], a factuality metric. We select the 19 material facts with the highest-scoring associated generations. For each selected material fact, we retain the top three generated commentaries and the corresponding human-written reference, yielding four commentaries per material fact and 76 commentaries in total.

This selection strategy is also motivated by prior analyses of LLM4FinComm [Assis et al. 2025a], which show that automatic metrics have lim-

ited ability to distinguish among top-performing generators. We therefore use these metrics for pre-selection to keep the human evaluation feasible, while still enabling a later comparison between metric-based rankings and preference-based judgments. The final subset includes commentaries produced by Llama-4-Scout [Meta AI 2025]; Llama3-8B [Meta AI 2024]; Sabiá-3 [Abonizio et al. 2024]; Command-R+ [Cohere Labs 2024]; Gemma3-12B and 27B [Gemma Team 2025]; Mistral-3-24B [Mistral AI 2025]; Mistral-7B [Jiang et al. 2023]; GPT-4o [OpenAI 2024] and their human references.

3.2. Evaluation Dimensions

Annotators are presented with a set of preference dimensions that may justify their decision to support the pairwise comparison. The selection of a criterion was not obligatory, but should be done if there was an evident dimension that drove the annotator's judgment. The dimensions were defined based on prior literature on text generation evaluation and financial text assessment [Assis et al. 2025a], and then refined with the support of a linguist to improve their clarity, consistency, and adequacy for the annotation task. A pilot round of annotation was conducted to assess the usefulness of these criteria for the task. These dimensions are not used as independent scores, but as optional indicators of the aspects in which the selected commentary stands out. Thus, annotators could choose a preferred commentary whether or not they marked any criteria.

The criteria are defined as follows: (a) **Writing Quality**, the commentary is well written, clear, and remains focused on the topic introduced by the material fact; (b) **Factuality**, the commentary is faithful to the information provided in the material fact and does not introduce unsupported or misleading claims; (c) **Usefulness**, the commentary is informative and contributes concretely to the reader's understanding of the event. It may also provide relevant information about actors, entities, or contextual elements beyond those explicitly described in the material fact, when such information helps the global interpretation; and (d) **Simplicity**, the commentary uses accessible and clear language, making it understandable to non-technical readers.

In addition to these criteria, annotators are instructed to use an optional free-text field whenever they wish to explain their judgment, point out specific strengths or weaknesses, or describe relevant aspects not covered by the predefined dimensions.

3.3. Human Preference Annotation

Two annotators with complementary expertise evaluated all pairwise comparisons for each material fact, one with experience in financial contexts, and one senior linguist. Both annotators received the same comparison rounds, ensuring that disagreements could be identified directly at the level of each pairwise decision. The identity of the generators was concealed throughout the annotation process to reduce potential author-related bias and ensure that preferences were based solely on the material facts. In practical terms, this also means annotators were unaware whether each commentary came from an LLM or a human author. Additionally, in the human annotation setting, the human-written commentary was always introduced only in the final round. We adopted this strategy because, if the human commentary had been included in earlier pairwise comparisons and systematically preferred, the tournament would provide little information about which LLM-generated commentary was strongest (RQ1).

After this stage, we identified the final-round comparisons in which the annotators disagreed. These cases were adjudicated by a third senior annotator with expertise in business and finance. This adjudication step was restricted to final-round disagreements, where outcomes are most relevant for identifying the strongest candidates. Overall, this setup combines financial and linguistic perspectives while preserving blind comparisons.

3.4. Annotation Interface Design

The human annotation was conducted through a dedicated web interface developed with Streamlit[2] and designed to support the pairwise preference protocol. For each comparison round, the interface presents the material fact and two candidate commentaries side by side. After selecting a preferred commentary, annotators may optionally indicate which criteria motivated their decision using a criteria form.

3.5. LLM-as-a-Judge Annotation

The LLM-as-a-judge evaluation follows the same sequential knockout pairwise protocol used in the human annotation protocol described in § 3.3. In each comparison round, the model received the material fact and the two candidate commentaries. The prompts[3] asked the model to select the preferred commentary, indicate the criteria supporting its decision through a predefined JSON output, and provide a brief justification. As in the human setup, the identity of the generators was hidden. From the second round onward, the prompt informed the model that the current holder had won the previous comparison, mirroring the strategy's sequential structure.

Concretely, we selected three frontier models as LLM judges: Claude Opus 4.6 [Anthropic 2026], Gemini 3.1 Pro [Google DeepMind 2026], and Qwen3-235B [Yang et al. 2025]. This choice was motivated by the stronger instruction-following capabilities of recent frontier models and by their expected lower susceptibility to positional bias [Goldsack et al. 2025]. Moreover, none of these models was included among the generators evaluated, as described in § 3.1, reducing the risk of self-preference bias.

Each model was run three times using the same round configuration presented to human annotators. To assess the effect of candidate ordering, we performed three additional runs with randomized commentary order.

3.6. Evaluation and Ranking Methodology

We analyze the annotation results from two complementary perspectives. First, we adopt an Elo-based [Elo 1978] ranking strategy, following its use in Arena AI [Zheng et al. 2023], and in competitive ranking settings such as chess. Elo scores estimate a participant's relative strength by updating their rating after each comparison, giving more weight to victories over stronger opponents than to victories over weaker ones. For this analysis, we exclude the human-written commentaries, since they were always shown in the final round of the human annotation protocol, which introduces a positional bias that could distort direct Elo comparisons. After computing Elo scores for each judge, we convert them to ranks, aggregate them using the rank-sum method, and analyze the resulting correlations across judges.

[2] https://fin-eval.streamlit.app/ [3] Temperature was set to 0. The complete prompts and additional details are available at https://github.com/AIDA-BR/human_and_judge_eval

Second, we analyze agreement and disagreement patterns among LLM judges, human annotators, and the selected evaluation attributes. Rather than focusing solely on the final ranking, this perspective examines how consistently judges prefer the same outputs, how their answers correlate, and which criteria are associated with their preferences.

4. Results

This section reports and discusses the main evaluation results, covering model rankings, judge agreement, preference criteria, and comparisons with automatic metrics.

4.1. Elo-Based Ranking Analysis

Table 1 reports the rankings obtained from judge-specific Elo scores and the highest-agreement aggregated rankings, using Kendall's W as a multi-annotator agreement coefficient. Aggregated scores are computed as the rank-sum of the individual judge rankings. Considering Kendall's W for the aggregated rankings, in connection with RQ2, the strongest agreement is observed for the aggregation of the three LLM judges, followed by the combination of Gemini, the linguist annotator, and the financial expert. Among model–human pairs, Gemini is closest to the linguist annotator in the Elo ranking, suggesting that they may prioritize more similar aspects of the generated commentaries. Combinations involving LLMs and the linguist annotator also show substantial agreement in general, indicating that LLM judges tend to align more closely with the linguist annotator than with the financial expert, in line with RQ3. These results indicate that agreement patterns are not simply separated between human and LLM judges; rather, some human–LLM combinations remain consistent.

Table 1. Individual and aggregated rankings derived from Elo across judges.

Model	Individual judges					Aggregated rank					
	Q	C	G	L	F	QCG $W = .807$	GLF $W = .726$	CGL $W = .667$	QGL $W = .619$	QCL $W = .559$	All $W = .499$
GPT-4o	6 (992.75)	2 (1038.68)	**1** (1136.31)	**1** (1024.91)	3 (1002.14)	2	**1**	**1**	**1**	2	**1**
Llama 4 Scout	4 (1013.47)	6 (996.90)	2 (997.73)	2 (1009.43)	2 (1019.02)	5	2	2	2	4	2
Gemma 3 27B	**1** (1068.36)	**1** (1040.27)	4 (993.70)	7 (994.23)	8 (990.28)	**1**	7	3	3	**1**	4
Sabiá-3	3 (1023.74)	3 (1018.20)	3 (997.37)	6 (995.98)	5 (996.09)	3	5	4	5	5	3
Mistral 3 24B	2 (1051.49)	4 (1011.57)	5 (990.18)	5 (996.42)	7 (992.43)	4	6	5	4	3	5
Gemma 3 12B	5 (1000.81)	5 (998.03)	6 (988.13)	4 (997.99)	4 (998.00)	6	4	6	6	6	6
Command R+	8 (937.49)	9 (951.77)	8 (969.74)	3 (998.39)	**1** (1019.26)	8	3	7	7	7	7
Llama 3 8B	7 (977.28)	7 (988.12)	7 (976.40)	8 (994.17)	6 (994.17)	7	8	8	8	8	8
Mistral 7B	9 (934.86)	8 (956.75)	9 (951.48)	9 (988.31)	9 (988.37)	9	9	9	9	9	9

Notes. Q = Qwen, C = Claude, G = Gemini, L = linguist, and F = finance expert. Individual cells report the bootstrap rank, with the corresponding bootstrap Elo median in parentheses. Aggregated columns report ranks obtained by summing bootstrap ranks across the indicated judges; lower ranks are better.

When all judges are combined, Kendall's W is moderate (0.499), confirming disagreement in the overall ranking composition. For this reason, we do not treat any aggregated ranking as definitive. Nevertheless, with respect to RQ1, GPT-4o ranks among the top models across judges, and Llama-4-Scout ranks highly in most aggregated settings. Gemma-3 27B and Sabiá-3 also remain competitive across several settings, while Command-R+ is ranked highly by the financial expert, although this is not a general pattern across all judge types. Overall, these findings provide evidence, related to RQ1 and RQ3, that model quality depends on the evaluation perspective, with different judges emphasizing different aspects of financial commentary generation.

Table 2. Correlation between automatic metrics and the all-judge Elo ranking.

Metric combination	Spearman	Kendall τ	Top-3	Top-5
BLEU + $\text{CTC}_{\text{factual-ref}}$	**0.483**	**0.333**	0.333	**0.800**
$\text{BERTScore}_{\text{recall}}$ + $\text{BERTScore}_{\text{f1}}$ + ROUGE-1	**0.483**	**0.333**	**0.667**	**0.800**
$\text{BERTScore}_{\text{recall}}$ + ROUGE-1	0.467	0.278	**0.667**	**0.800**
BLEU + ROUGE-L + $\text{CTC}_{\text{factual-ref}}$	0.450	0.278	**0.667**	**0.800**

Addressing RQ2, Table 2 reports the ranking correlations between the automatic metrics available in LLM4FinComm and the all-judge Elo ranking. We computed all possible metric combinations and, due to space constraints, report only the four combinations with the highest correlations. Overall, the results show moderate Spearman correlations and weak Kendall correlations, suggesting that automatic metrics do not recover a full ranking that closely matches the Elo-based ordering. However, we also report the overlap with the top-3 and top-5 models in the reference ranking. Some of the best-performing metric combinations achieve reasonable coverage, reaching 0.667 overlap for the top-3 and 0.800 for the top-5. This suggests that automatic metrics may be useful for identifying strong candidates, but are less reliable for fine-grained ranking among models with similar performance [Assis et al. 2025b]. We also observe that the best-performing combinations include lexical, semantic, and factuality-oriented metrics, such as BLEU, ROUGE, BERTScore components, and CTC-based factuality metrics. This indicates that financial commentary evaluation involves complementary dimensions that no single family of automatic metrics can fully capture.

4.2. Human and LLM Judge Comparison

We use consensus across judges runs and human evaluation to assess the reliability of the sequential knockout pairwise protocol. Because the protocol is order-sensitive, we first analyze positional effects from reordering. The starting commentary wins in 28.1% of the rounds, only slightly above the random expectation of 25% for four candidates per round, suggesting that the initial position has at most a minor influence on the outcome. However, the three-run design mitigates this effect, with LLMs reaching unanimous decisions on 65–90% of items.

Humans comparison The linguist (L) and the financial expert (F) agreed on 42.1% of items (Cohen's $\kappa = 0.30$), a level of agreement that, while modest, sits well above the 17.5% expected by chance and reflects distinct evaluation stances rather than annotation noise. The third annotator, a senior financial expert, then adjudicated the remaining cases and sided with F in 72.7% of them. This divergence suggests distinct evaluation stances in our sample, with the linguist giving greater weight to communicative effectiveness and the financial experts placing more emphasis on practical utility. For example, the adjudicator sometimes cited neutrality when justifying their choice, suggesting a concern with more impartial interpretations, whereas L valued the addition of contextual information that made the text more informative. This is reflected in their choices: annotators with a financial background generally preferred LLM-generated commentaries, while L more often favored the human-written references.

The criteria profiles further support this interpretation. When F marked criteria, the selections were more often associated with usefulness (88%) and writing quality (80%) than with factuality (46%). This suggests that, for F, factual correctness may have functioned less as an explicit comparative dimension than as a prerequisite, with prefer-

ences then shaped mainly by utility and presentation quality. It is also relevant to recap that marking criteria was optional for all human annotators.

LLM Judges comparison Pairwise agreement among LLM judges, computed on the consolidated per-item winner, ranges from 36.8% to 47.4% (averaging 42.1%); on individual runs, it is somewhat higher (42.9%–49.1%, averaging 46.1%), since consolidating three runs absorbs within-judge variation. All three judges converge on the same winner in 31.6% of cases. These item-level agreement figures differ from the rank-level agreement reported in Table 1, since judges may share a similar overall ordering while still disagreeing on many individual pairwise decisions. Criterion profiles further differentiate the judges: Qwen emphasizes factuality (96%) but rarely flags writing quality (40%); Gemini often selects factuality (97%) and simplicity (73%); and Claude is more balanced, flagging factuality (96%) and utility (85%).

Human vs. LLM Judges comparison Cross-group agreement averages 27.2%, below the within-group agreement observed among both LLMs and humans (42.1% in each group). This indicates that LLM judges are not more internally consistent than human experts. At the *pairwise-decision level*, the highest human–LLM agreement is 36.8%, observed for both Claude–linguist and Gemini–financial expert. However, Gemini and the linguist show the closest Elo-based rankings (Table 1), suggesting that agreement rates and ranking similarity capture complementary aspects of alignment. In contrast, Claude shows the lowest agreement with the financial expert (21.1%).

Model preferences also differ across groups. LLM judges favor GPT-4o most often (31.6% of wins), followed by Mistral-3 and Llama-4 (15.8% each), and human texts (14.0%). Although these model outputs appeared with different frequencies due to metric-based pre-selection (§3.1), participation frequency alone does not explain preference, as Llama-4 and human texts win only 25.0% and 14.0% of their battles, respectively. Human annotators instead favor Llama-4 (28.9%) and human texts (26.3%), while also assigning votes to Command-R+ (10.5%), which receives no votes from LLM judges. Conversely, Gemma-3-27B accounts for 12.3% of LLM preferences but receives no human votes. These divergences suggest that LLM judges and human experts capture partly different aspects of commentary quality, including textual quality and decision-support value.

The variation in winners limits the use of win frequency as evidence for a single best model. Still, Llama-4-Scout emerges as a strong candidate, with consistent selections across groups, while human commentaries receive many votes, particularly from human annotators and Claude.

4.3. Automatic Metrics and Judge Preferences

The automatic metrics available in LLM4FinComm are mostly reference-based. In this sense, we can infer that the linguist (L) tends to prefer outputs with greater lexical proximity to the human reference, as reflected by the alignment with lexical automatic metrics in 81.8% of their LLM preferences. This preference may also be related to the contextual information present in human-written commentaries, which is often absent from LLM-generated outputs. The CTC-based factuality metrics also reveal an interesting dynamic. Final winners have lower CTC scores than their al- ternatives in 57.1% of comparisons overall, with this effect concentrated among LLM judges (Qwen3: 73.7%, Opus: 63.6%, Gemini: 57.9%). Human annotators do not show the same inverse pattern, with L's win-

ners having higher CTC scores in 63.6% of cases, while the financial expert (F) is roughly neutral (47.1% lower vs. 52.9% higher).

Command-R+ illustrates this dynamic. Its generations receive no votes from LLM judges, yet F ranks the model first overall, with Command-R+ accounting for 24.6% of F's preferences. F's justifications emphasize structured financial reasoning and actionable insight, dimensions that LLM judges and automatic metrics may fail to capture and that reference-based metrics may even penalize when outputs diverge from the reference. This is consistent with prior work showing that systems favored by automatic or LLM-based scores do not necessarily better support human decision-making [Takayanagi et al. 2025]. At the same time, it contrasts with some comments from the senior financial adjudicator, who often emphasized neutrality as a desirable property. Still, the adjudicator sided with F in 72.7% of disputed items, suggesting some degree of domain-aligned preference, although differences in seniority and evaluation stance may also play a role. A broader pool of annotators would be needed to better separate and understand these effects.

5. Conclusions

This paper proposes a sequential knockout protocol for evaluating financial commentaries and examines whether LLM judges approximate human expert preferences under the same evaluation framework. Our results provide evidence on three fronts:

For RQ1, on model performance, GPT-4o and Llama-4-Scout are consistently well positioned across evaluation perspectives, while Gemma-3-27B performs better under LLM judges and Command-R+ is distinctively preferred by the financial expert, despite receiving no LLM votes. For RQ2, on evaluation alignment, automatic metrics do not reproduce judge-based rankings, either human or LLM-based, contrasting with [Thakur et al. 2024]. We attribute this difference to our candidate selection strategy, which focuses on already high-scoring commentaries and makes fine-grained ranking more difficult for traditional metrics; still, automatic metrics remain useful for candidate pre-selection, covering up to 80% of top-ranked models, though not in the same order. For RQ3, on systematic disagreements, within-group agreement is comparable for LLM judges and human annotators (42.1%), whereas cross-group agreement drops to 27.2%, suggesting that disagreements are driven mainly by differences between evaluation perspectives rather than by greater instability in either group.

Finally, the sequential knockout protocol offers a practical alternative by reducing the number of comparisons required from annotators [Sandan et al. 2025]. However, deriving stable conclusions from this setup remains non-trivial in this domain. We therefore combine Elo-based ranking with agreement analysis to extract complementary signals about model performance and judge alignment. Future work should expand the annotator pool and compare this protocol with alternative designs, such as non-retained-winner or exhaustive pairwise evaluation.

Acknowledgments

This work received support from CNPq (National Council for Scientific and Technological Development) under grants 307088/2023-5, held by Aline Paes, and 301925/2025, held by Altigran Soares da Silva, and through the AIDEn Project (Proc. 400936/2025-9). From FAPERJ (Proc. SEI-260003/002930/2024 and SEI-260003/000614/2023), from

FAPEAM through the POSGRAD 2025 Program and the NeuralBond Project (grant 01.02.016301.04300/2023-04), and from CAPES – Finance Code 001. Additional support was provided by the CNPq Brazilian National Institutes of Science and Technology (INCTs), IAIA (grant #406417/2022-9), and TILD-IAR (grant #408490/2024-1). We thank Raquel da Silva Carlim for her contribution to the annotation process.

Limitations

Our analysis is based on 19 material facts, reflecting both the feasibility of annotation [Sandan et al. 2025] and our focus on the highest-scoring candidate set. This limited sample size may affect statistical precision. In addition, material facts were preselected using automated metrics, which may bias the subset toward cases with smaller metric-judge gaps than in the full dataset. In the human annotation protocol, the human-written commentary was always introduced in the final round, creating a positional confound that we partially address through Elo-based ranking and repeated LLM runs with reordered candidates. We adopt Elo as an interpretable first step for ranking pairwise preferences. Although sparse comparisons can affect Elo stability, our bootstrap analysis mitigates this limitation through bootstrap resampling. Future work may complement this analysis with order-invariant models such as Bradley–Terry–Luce or Thurstone models [Lipovetsky and Conklin 2003]. We used the same guidelines for human and LLM judges, stating that choosing a criterion for the preference was optional, but LLMs tend to point to at least one criterion, whereas humans do not. This is another point to consider when comparing LLM and human evaluation: even when using the same evaluation guidelines, they may produce different outputs that are difficult to compare. The Cohen's κ among annotators further reflects the difficulty and subjectivity of the task, as different annotators may emphasize different dimensions of commentary quality [Sharma et al. 2026]. Larger samples, more annotators, and alternative ordering strategies would strengthen generalizability, although balancing this demand with the practical constraints of expert human evaluation remains challenging, especially for senior financial expertise.

Ethics Statement

Annotation was conducted by three collaborators (one linguist and two financial-domain experts) who volunteered and were informed about the study's purpose. The dataset (LLM4FinComm) consists of publicly disclosed material facts and generated commentaries. By documenting systematic divergences between automatic, LLM-based, and expert evaluation, our work cautions against the naive use of automated metrics in financial NLP, where unreliable quality assessment may mislead downstream users, including investors.

Disclosure of GenAI usage

The LLMs evaluated in this study are described in §3. GenAI tools (Claude Code and GPT Codex) were used to support auxiliary data-processing tasks via Python scripts, including data preparation, file manipulation, and CSV formatting. GenAI tools were also used for language polishing, translation between Portuguese and English, and LaTeX coding. All generated or tool-assisted content was reviewed by the authors, who remain responsible for the analyses, interpretations, and conclusions presented in this work.

References

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabiá-3 technical report.

Anthropic (2026). Claude opus 4.6 system card. https://www-cdn.anthropic.com/14e4fb01875d2a69f646fa5e574dea2b1c0ff7b5.pdf.

Assis, G., Dutra, H., Vianna, D., Meira, Wagner, J., da Silva, A. S., and Paes, A. (2025a). 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.

Assis, G., Freitas, C., and Paes, A. (2025b). Exploring Brazil's LLM Fauna: Investigating the generative performance of large language models in portuguese. *Journal of the Brazilian Computer Society*, 31(1):939–971.

Celikyilmaz, A., Clark, E., and Gao, J. (2020). Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. (2024). Humans or LLMs as the judge? a study on judgement bias. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Cohere Labs (2024). Introducing command r+: A scalable llm built for business. https://cohere.com/blog/command-r-plus-microsoft-azure.

Deng, M., Tan, B., Liu, Z., Xing, E., and Hu, Z. (2021). Compression, transduction, and creation: A unified framework for evaluating natural language generation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *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.

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. Arco Publishing.

Gao, M., Hu, X., Yin, X., Ruan, J., Pu, X., and Wan, X. (2025). LLM-based NLG Evaluation: Current Status and Challenges. *Computational Linguistics*, 51(2):661–687.

Gemma Team (2025). Gemma 3 Technical Report.

Goldsack, T., Wang, Y., Lin, C., and Chen, C.-C. (2025). From facts to insights: A study on the generation and evaluation of analytical reports for deciphering earnings calls. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10576–10593, Abu Dhabi, UAE. Association for Computational Linguistics.

Google DeepMind (2026). Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-

A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b.

Jin, S., Li, S., Zhang, S., and Yan, R. (2026). Finrpt: Dataset, evaluation system and LLM-based multi-agent framework for equity research report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(1):507–515.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. ACL.

Lipovetsky, S. and Conklin, M. (2003). Priority estimations by pair comparisons: Ahp, thurstone scaling, bradley-terry-luce, and markov stochastic modeling. In *Proceedings of the ASA Joint Statistical Meeting, JSM*.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (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.

Masid, M. R., Assis, G., Vianna, D., Paes, A., and Silva, A. S. d. (2026). Auditing the evaluators: How far can automatic evaluation go in assessing Portuguese financial texts? 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 222–233, Salvador, Brazil. Association for Computational Linguistics.

Meta AI (2024). The llama 3 herd of models.

Meta AI (2025). The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.

Mistral AI (2025). Mistral Small 3.1.

Mukherjee, R., Bohra, A., Banerjee, A., Sharma, S., Hegde, M., Shaikh, A., Shrivastava, S., Dasgupta, K., Ganguly, N., Ghosh, S., and Goyal, P. (2022). ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI (2024). Gpt-4o system card.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on ACL*, ACL '02, page 311–318, USA. ACL.

Ribeiro, J. V. A., Correia, T. P., Requena, J. V. S. C., and Berton, L. (2026). Evaluating reference-free summarization quality metrics for Portuguese: A study with human judgments in financial news. 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 899–907, Salvador, Brazil. Association for Computational Linguistics.

Sandan, I. B., Dinh, T. A., and Niehues, J. (2025). Knockout LLM assessment: Using large language models for evaluations through iterative pairwise comparisons. In Ar-

viv, O., Clinciu, M., Dhole, K., Dror, R., Gehrmann, S., Habba, E., Itzhak, I., Mille, S., Perlitz, Y., Santus, E., Sedoc, J., Shmueli Scheuer, M., Stanovsky, G., and Tafjord, O., editors, *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 121–128, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Sharma, N., Agarwal, N., and Sirts, K. (2026). Towards consistent detection of cognitive distortions: Llm-based annotation and dataset-agnostic evaluation. In Piperidis, S., Bel, N., van den Heuvel, H., Ide, N., Krek, S., and Toral, A., editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10866–10882, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Takayanagi, T., Goldsack, T., Izumi, K., Lin, C., Takamura, H., and Chen, C.-C. (2025). Earnings2Insights: Analyst report generation for investment guidance. In Chen, C.-C., Winata, G. I., Rawls, S., Das, A., Chen, H.-H., and Takamura, H., editors, *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 246–251, Suzhou, China. Association for Computational Linguistics.

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., and Hupkes, D. (2024). Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics*.

Wataoka, K., Takahashi, T., and Ri, R. (2024). Self-preference bias in LLM-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*.

Xiao, Y., Sun, E., Luo, D., and Wang, W. (2025). Tradingagents: Multi-agents LLM financial trading framework. In *The First MARW: Multi-Agent AI in the Real World Workshop at AAAI 2025*.

Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., et al. (2024). Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.

Yan, S. (2022). Disentangled variational topic inference for topic-accurate financial report generation. In Chen, C.-C., Huang, H.-H., Takamura, H., and Chen, H.-H., editors, *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.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. (2025). Qwen3 technical report.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Zhu, D., Lappas, T., and Rachidi, T. (2023). Commentary generation for financial markets. *Expert Systems with Applications*, 211:118364.