# Effective Performance Measurement: Challenges and Opportunities in KPI Extraction from Earnings Calls

**Rasmus T. Aavang[1,2], Rasmus Tjalk-Bøggild[2], Alexandre Iolov[2], Giovanni Rizzi[2], Mike Zhang[3], Johannes Bjerva[1]**

[1]Department of Computer Science, Aalborg University, Denmark
[2]ALIPES ApS, Denmark
[3]University of Copenhagen
**Correspondence:** rtaj@cs.aau.dk

## Abstract

Earnings calls are a key source of financial information about public companies. However, extracting information from these calls is difficult. Unlike the templatic filings required by the U.S. Securities and Exchange Commission (SEC) to report a company's financial situation, earnings conference calls have no built-in labels, are unstructured, and feature conversational language. We explore this challenging domain by assessing the information captured by models trained on SEC filings and in-context learning methods. To establish a baseline, we first evaluate the generalization capabilities of SEC-trained models across established SEC datasets. To support our investigation, we introduce three novel benchmarks: (1) SEC Filings Benchmark (**SECB**), (2) Earnings Calls Benchmark (**ECB**), and **ECB-A**, a subset with 2,460 expert annotation groups to support our qualitative analysis. We find that encoder-based models struggle with the domain shift. Finally, we propose a system utilizing LLMs to perform open-ended extraction from unstructured call transcripts, verified by human evaluation (79.7% precision), providing a baseline for this valuable domain through the consistent tracking of emergent KPIs.

## 1 Introduction

The efficient market hypothesis (Fama, 1970) states that, given current prices already incorporate all public information, the primary driver of significant price change is new information. Consequently, industry investors are in a constant *race for new information*, as a company's valuation can swing 40% in seconds (Figure 2). Two of the most critical public disclosures in this race are SEC filings and corresponding earnings conference calls. Earnings calls comprise two key components: (i) Senior management's presentation of *key performance indicators* (KPIs) and (ii) Q&A session with analysts and investors. As a result, calls play a crucial role for industry investors in assessing a company's value. However, extracting information from earnings calls is difficult; neither gold annotations nor a method for automatic KPI extraction exists. We address key challenges in the emerging task of information extraction from earnings calls by evaluating both state-of-the-art methods, developed initially for SEC filings, and in-context learning as illustrated in Figure 1. Our guiding RQs:

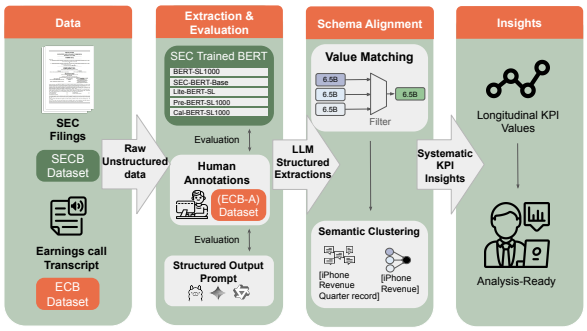


Figure 1: **Analysis and Pipeline.** We ground our analysis in the established SEC filings domain. To capture the open-ended set of KPIs in earnings calls, we adopt a relation extraction strategy to benchmark encoders and in-context learning against expert annotations. Finally, we aggregate structured outputs to generate consistent, longitudinal KPI tracking suitable for financial analysis.

**RQ1** How well do encoder-based models, fine-tuned on structured SEC filings, generalize across the linguistic shift to the unstructured domain of earnings conference calls?

**RQ2** How well can current State-of-the-Art (SOTA) Large Language Models perform structured KPI extraction, and what actionable impact can be derived from the current performance?

## 2 Background

**Automated KPI Extraction** Financial NLP has shifted from encoders like FinBERT (Araci, 2019; Yang et al., 2020) to focus on larger models such as FinMA (Xie et al., 2023) and InvestLM (Yang et al.,



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1434–1460
July 2-7, 2026 

2023b), proprietary tools like Bloomberg GPT (Wu et al., 2023), and RAG-based FinGPT (Yang et al., 2023a). Despite this progress, work targeting KPI extraction is limited and usually relies on predefined schemas. KPI-BERT (Hillebrand et al., 2022) tries to link values to descriptions in German financial documents. FiNER-139 (Loukas et al., 2022) and HiFi-KPI (Aavang et al., 2025) classify entities into fixed taxonomies. Less investigated is KPI extraction from earnings calls, despite its impact on investment returns (Chen et al., 2018; Qin and Yang, 2019; Ma et al., 2020; Barahona Diaz and Hu, 2024), likely because of the free-flowing, unstructured, and unlabeled nature of the calls. To address this unlabeled nature, we adopt a relation extraction approach (Etzioni et al., 2008). Inspired by recent advancements in open-domain extraction like ODKE+ (Khorshidi et al., 2025), we utilize state-of-the-art LLMs: Llama-3.3, Qwen3-30B-A3B, Gemma-3-27B-it and Gemini 3 pro (Grattafiori et al., 2024; Yang et al., 2025; Team et al., 2025; Google DeepMind, 2025) to dynamically extract KPIs without prior schemas.

**Financial Theory of Performance Measurement** Various methods of performance measurement have been explored – see Khan and Shah (2011) for an overview. We adopt the definitions of Ghalayini and Noble (1996), distinguishing between *traditional* (retrospective, financial) and *non-traditional* (operational, forward-looking) metrics (Appendix Table 12). While prior work focuses on the extraction of traditional KPIs. We also consider non-traditional KPIs, more present in earnings calls and deemed crucial in the assessment of business performance (Ghalayini and Noble, 1996).

**SEC Filings** Publicly traded American companies follow the regulations of the Securities and Exchange Commission (SEC) and file 10-Qs (Quarterly updates) and 10-Ks (Annual updates). (U.S. Securities and Exchange Commission, 2024). 10-Qs and 10-Ks are highly templatic and focus on traditional performance measures.

**Earnings Calls** During these calls, management and investors discuss financial results (Investopedia, 2023). Figure 2 shows the race for new information in practice. An error in the reported value for the KPI "EBITDA margin expansion" increased the company's valuation by over $3 billion, before a correction in the call prompted a $3 billion drop. Conference calls are not legally mandated, but most companies do them with the structure:

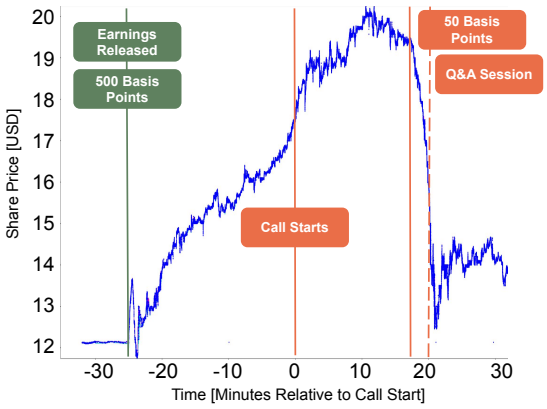


Figure 2: **Lyft's share price** from the release of its earnings report to the end of the earnings call. When the incorrect value is presented in the earnings release the price rises quickly. However, once the error is corrected during the earnings call, the price rapidly drops.

| | # Entries | Period | Entities |
|---|---|---|---|
| **FiNER-139** | 1.1M | 2016-2020 | 387K |
| **HiFi-KPI** (Lite) | 1.9M(8.0K) | 2017-06/2024 | 5,300K |
| **SECB** | 41K | 2023-2024 | 78K |
| **ECB** (ECB-A) | 10.5K (587) | 2023-2024 | (2.5K) |

Table 1: **Dataset Statistics:** Comparison of earlier datasets and our benchmarks (**SECB**, **ECB**, **ECB-A**).

1. **Presentation and Discussion**

   Management, usually the CFO and CEO, discuss the financial results and present KPIs (Corporate Finance Institute, 2024).

2. **Q&A Session**

   Usually, earnings calls end with a Q&A session, where investors and analysts – and today even retail investors – can vote on potential questions to ask (Markov and Yezegel, 2023).

**LM-Based Metrics** Language model-based metrics like BERTScore (Zhang et al., 2019) and cross-encoders (e.g., STSB-RoBERTa-large) surpass rule-based methods (Reimers and Gurevych, 2019; Ebrahim and Joy, 2024). LLMs demonstrate robust judgment capabilities (Zheng et al., 2023), though prone to self-preference bias (Wataoka et al., 2024). Consequently, we employ DeepSeek-V3.2 (DeepSeek-AI et al., 2025) as a distinct evaluator, complemented by a RoBERTa-based score.

## 3 Experimental Setup

We construct three novel datasets from 20 S&P 500 companies[1]. We obtain SEC filings from U.S. Securities and Exchange Commission (2026) and earnings call transcripts from Financial Modeling Prep (2024).

1. SEC Filings Benchmark (**SECB**)
2. Earnings Call Benchmark (**ECB**)
3. Earnings Call Benchmark Annotated (**ECB-A**)

Table 1 compares these to existing resources also included in our analysis. To bridge the gap between templated filings and conversational calls, we optimize **SECB** to preserve broader context, than HiFi-KPI (Aavang et al., 2025), which prioritizes KPI density. Within this expanded "free text," we introduce the pseudo-tags ***regex dollar*** and ***regex percentage*** to capture unannotated KPIs. **ECB** has 10,477 chunks from 2023–2024 earnings calls across 20 companies, segmented by speaker turns (uninterrupted speech by the same speaker). To ensure reliability, we establish **ECB-A**, an evaluation subset consisting of 10 randomly selected full transcripts. A domain expert performed a two-stage annotation process separated by a 3-month break: an initial pass to identify KPI descriptors and values, followed by a confirmation pass to verify entities and connect entity relations. **ECB-A** contains 587 chunks (avg. 147 words) annotated with 2,460 entities (4.19/chunk) and 934 relational groups (1.59/chunk). (Annotation and parsing details in **Appendix A and D**).

**Experiments** We investigate KPI extraction from earnings call transcripts using two distinct paradigms: domain-specific encoders (SEC-based BERT models) and in-context learning via few-shot prompting. We assess the generalization ability across SEC filings-based dataset, by standardizing the labelspace and evaluating SEC-BERT-BASE and BERT-SL1000 on **SECB**, as well as on the FiNER-139 (Loukas et al., 2022) and HiFi-KPI (Aavang et al., 2025) test sets as a sequence labeling task[2]. Following this baseline analysis, we evaluate performance on unstructured earnings calls using the annotated **ECB-A** dataset. We also test LLMs with a structured few-shot prompting scheme (details in **Appendix G**). This prompt mirrors our expert annotator guidelines: models must first identify KPI entity spans and then aggregate them into relational groups to generate a descriptive label. To evaluate the SEC-based encoders against our open-ended relation schema, we map their predicted token-level classes directly to the 'Label' field in our grouping structure.

[1]The selected tickers are: AAPL, JNJ, JPM, AMZN, BA, PG, XOM, NEE, GOOGL, DOW, PLD, MSFT, PFE, BAC, HD, CAT, KO, CVX, DUK, and SHW.

[2]Code and datasets are available at https://github.com/aaunlp/effective-performance-measurement.

| | μ-F1 | | M-F1 | |
|---|---|---|---|---|
| Dataset | SB | SL1000 | SB | SL1000 |
| **SECB** | 0.057 | **0.143** | 0.032 | **0.192** |
| FiNER-139 | **0.842** | 0.662 | **0.859** | 0.624 |
| HiFi-KPI | 0.240 | **0.501** | 0.001 | **0.012** |

Table 2: **SEC-BASED BERT models on SEC Filings Performance.** SEC-BERT-BASE (SB) and BERT-SL1000, evaluated on FiNER-139, HiFi-KPI and **SECB**. Ignoring the "O" label.

**ECB-A Metrics** We base our evaluation on the following metrics with semantic scoring from cross-encoder *STSB-Roberta-Large* (Reimers and Gurevych, 2019).

1. **Exact F1:** Standard F1 requiring exact string matches for both value and label.
2. **Semantic F1:** Derived from the mean maximum similarity of predictions to ground truths (precision) and vice-versa (recall), allowing for many-to-one mapping.
3. **Match F1:** A soft F1 score derived from the label similarity of predictions strictly aligned to ground truths by value (precision) and vice-versa (recall), treating unaligned items as zero.
4. **LLM Judge:** The percentage of value-grouped ground truths found and evaluated as equivalent by DeepSeek-V3.2

Finally, we utilize **ECB** to show a practical system employing semantic clustering to identify KPIs.

## 4 Empirical Results

Our experiments reveal a stark contrast in model performance between structured SEC filings and more conversational earnings calls. Table 2 shows that while absolute scores vary, neither model experiences a catastrophic drop in performance when shifting datasets. Although the thousands of labels and regex-based labels in **SECB** inherently limit absolute performance metrics, the reasonable Micro-F1 scores confirm the models' utility. In Figure 3 we simplify gold labels into: *XBRL*, *regex*, and

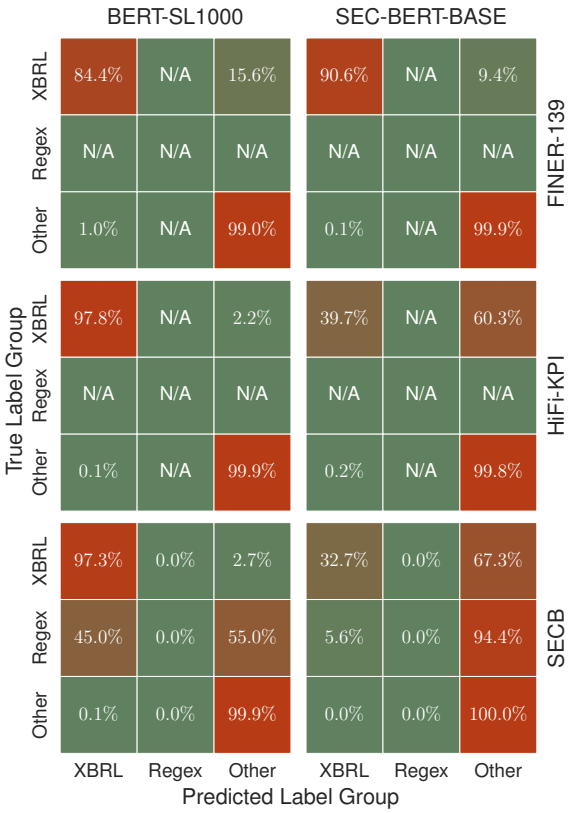


Figure 3: Confusion matrices for SEC-BERT-BASE and SL1000 on HiFi-KPI, FiNER-139, and **SECB**.

| Model | Scores (%) | | | |
|---|---|---|---|---|
| | Exact | Semantic | Match | LLM Judge |
| SEC-BERT-BASE | 0.0 | 0.0 | 0.0 | 0.0 |
| Lite-BERT-SL | 0.0 | 7.1 | 1.6 | 0.0 |
| Pre-BERT-SL1000 | 0.0 | 5.6 | 1.3 | 0.8 |
| Cal-BERT-SL1000 | 0.0 | 4.8 | 1.5 | 0.9 |
| BERT-SL1000 | 0.0 | 4.7 | 1.1 | 0.4 |
| Gemma-3-27B | 3.2 | 40.0 | 11.6 | 8.8 |
| Qwen3-30B-A3B | 3.5 | 38.1 | 26.2 | 33.9 |
| Llama-3.3-70B | 3.4 | 51.5 | 25.8 | 26.6 |
| Gemini 3 Pro | **11.5** | **61.6** | **39.2** | **45.5** |

Table 3: **Performance Comparison on ECB-A.**

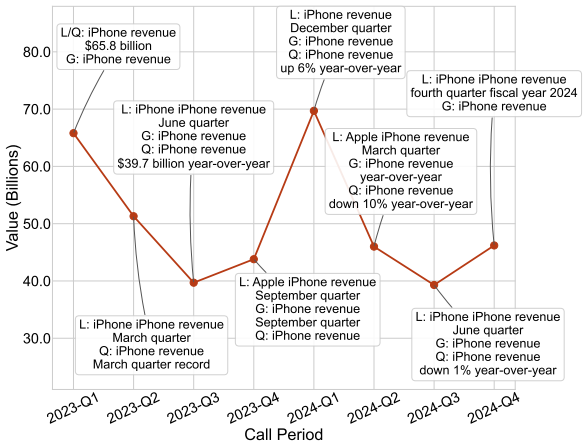


Figure 4: **Longitudinal KPI Tracking.** Our system automatically identifies Centroids (e.g., 'iPhone revenue'). Callouts display the raw predictions from L (Llama-3.3), G (Gemma-3), and Q (Qwen3).

*other*. We find that BERT-SL1000 is significantly more aggressive, classifying 45% of *Regex* spans as *XBRL*, compared to only 5.6% for SEC-BERT-BASE. This suggests BERT-SL1000 may generalize better to the unlabelled financial data found in earnings calls. However, as shown in Table 3, SEC-trained models fail to generalize to earnings calls; even though BERT-SL1000 extracts some correct numbers, it fails to predict compatible labels. In contrast, the generative models show promise and significantly outperform the SEC-trained models, while exact match performance is still low, the semantic-based metrics show promise for the generative models. Performance is especially impressive considering the models operate in a fully unconstrained setting, identifying entities directly from text without reliance on a closed taxonomy.

Qwen-3 and Llama-70B find 26.55% and 33.94% of the expert annotations according to the LLM judgment, with Gemini 3 pro achieving the best performance with an exact F1 of 11.5% and extraction of 45.5% of the annotated KPIs according to the LLM judgment. While exact F1 is low, semantic F1 is high, especially for Llama-3.3 and Gemini 3 pro. Despite a high semantic score, Gemma-3's low match rate indicates frequent value-relation misalignment, likely causing its poor LLM-as-a-judge performance. The significantly higher Semantic and LLM-Judge scores demonstrate that while LLMs capture the underlying financial concepts, they struggle with the strict lexical boundaries of Exact Match extraction. The continued performance improvement with model scaling, even at the state-of-the-art level, highlights the task's inherent complexity.

## 5 Industry Application & ECB

We scale our experiments to the ECB dataset, comprising two years of longitudinal data across 20 companies. Figure 4 shows our system enabling KPI discovery, meaning we can consistently discover and track KPIs across time without any defined ontologies. Since our prompt enforces a strict entity schema, we implement a post-hoc aggregation pipeline mirroring the match metric logic. Due to compute constraints and reasonable performance, we opt for using *Gemma*, *Llama*, and *Qwen*. First, we focus on value alignment; we re-

| Model | Share of Pred. (%) | Centroid (%) | Overlap (%) |
|---|---|---|---|
| Llama-3.3-70B | 77.63 | 63.19 | 31.26 |
| Qwen3-30B-A3B | 78.68 | 61.86 | 25.84 |
| Gemma-3-27B | 77.70 | 59.53 | 30.25 |

Table 4: **Model Contributions.** All models contribute comparably to the final extraction. Llama has the most common centroid, while Qwen has the lowest chance of using the same label as the other models.

| Metric | Score |
|---|---|
| Krippendorff's $\alpha$ | 0.429 |
| Precision | 79.67% (478/600) |

Table 5: **Human Evaluation.** Though Krippendorff's $\alpha$ (0.429) indicates moderate agreement, the final extraction shows high precision of 79.67% (478/600).

quire numerical values to be within a $1\%$ tolerance. We then grouped extracted KPI entities based on semantic similarity. We apply a similarity threshold of $0.85$ between all entities' KPI names in the same cluster. For each identified cluster, we assign a canonical label corresponding to the cluster centroid—defined as the KPI name with the minimum aggregate distance to all other variations in the group. Finally, we limit the tracking of KPIs to the centroids found in 4 different periods for the same company. We default the period to be the current call unless the date entity extracted explicitly mentioned another quarter or year. Further, we observe a perfect match comparing the values extracted for "iPhone Revenue" by our system with the actual reported values in Apple's SEC filings. Our system finds 1,323 KPIs that can be consistently tracked across at least 4 periods for the 20 companies in two years of calls. All 3 models agree on 4.16% of these KPI extractions. Table 4 shows all models contributing to the final results, even though Gemma-27B showed worse performance in isolation.

**Human Evaluation** To verify our final extractions via the post-hoc aggregation, we employ three evaluators to verify 200 extractions each (100 overlapping; see Appendix E). Evaluators assessed the KPI label and value correctness. Krippendorff's $\alpha$ (Hayes and Krippendorff, 2007) of 0.43 and average Cohen's $\kappa$ (Cohen, 1960) of 0.39 imply moderate agreement (Wong et al., 2021); however, raw agreement remains high (69%). This suggests the low $\alpha$ is dampened by the high positive label prevalence. Our high system precision of 79.67% demonstrates promising extraction capabilities in this domain. Which can, however, definitely be improved by future more sophisticated methods.

| Model | Valid KPI (%) | # Unmatched Predictions |
|---|---|---|
| Gemma-3-27B-it | 18% | 318 |
| Qwen3-30B-A3B-Instruct-2507 | 18% | 701 |
| Llama-3.3-70B-Instruct | 32% | 588 |
| Gemini 3 pro Preview | 26% | 771 |

Table 6: **Unmatched Predictions.** Llama-3.3 and Gemini 3 Pro yield the highest validity. Llama achieves the peak validity (32%), likely because of its more conservative extraction volume compared to Gemini.

## 6 Analysis

We begin our analysis with a systematic manual comparison between **ECB-A** and model predictions. We randomly sample 100 unmatched extractions from each model to determine if these discrepancies stem from model error or omissions in the expert annotation. Table 6 shows a strong tendency towards over-extraction, but also that some KPIs have been missed by the annotator. These unannotated but valid predictions demonstrate that ECB-A serves as a high-quality but inherently partial gold benchmark.

### 6.1 Error Analysis

Table 7 shows the extraction most commonly evaluated as wrong. The top 2 ("azure ai VAL" and "1 billion") are meaningless labels. More interestingly 100% of the extractions of cash flow are evaluated as wrong. It seems from there the error rate drops quickly.

| KPI (Centroid Label) | Total | Wrong | Err (%) |
|---|---|---|---|
| azure ai VAL azure ai customers | 6 | 6 | 100.0 |
| 1 billion | 5 | 5 | 100.0 |
| cash flow | 4 | 4 | 100.0 |
| rotcce | 3 | 3 | 100.0 |
| electric utilities infrastructure up VAL | 4 | 3 | 75.0 |
| international segment international segment revenue | 4 | 3 | 75.0 |
| nii | 8 | 4 | 50.0 |
| google service google service revenues | 9 | 4 | 44.4 |
| organic sales growth | 15 | 4 | 26.7 |
| gross margin | 17 | 4 | 23.5 |

Table 7: Top 10 KPIs flagged by annotators as wrongly extracted, sorted by error rate.

**Differences between Calls and Filings** Why is there such a discrepancy between results on Earnings calls and SEC filings? In this section, we highlight concrete differences between these two sources of information, with a thorough analysis of the difficulties for the KPI-extraction models. We use Green for KPI label and Blue for Value.

> As you know, free cash flow has been our primary financial metric through this recovery, and based on our performance year-to-date, we still plan to be in the guidance range for the year as well as the $10 billion target by 2025 and 2026. (The Boeing Co., 2023)

This statement by Boeing's CEO Dave Calhoun exemplifies the linguistic complexity of earnings calls, presenting challenges for natural language processing systems. The *free cash flow* metric serves as the anchor for the "$10 billion target", yet this relationship is obscured because this figure points to future performance (2025-2026) rather than the current period. Meanwhile, current performance is described only vaguely through indirect reference to an unspecified *guidance range*. This linguistic structure creates an ambiguity pattern typical of earnings calls, in which optimistic numerical projections receive prominence while potentially damaging current KPIs remain underspecified. With respect to performance, *Gemma-3* is able to do this perfectly, relating them with the label "free cash flow 2025 2026", the same is *Gemini-3-pro* with the label "plan target free cash flow 2025 and 2026". *Llama-3.3* extracts both $10 billion and free cash flow as entities; however, it does not relate them to each other in a group with a label. Finally, *Qwen* relates "2025", "2026", "year-to-date", altogether, and ends up using the label "still plan free cash flow guidance range year-to-date" with the $10 billion as value. This non-agreement between any of the models of course also means that our final system predicts nothing for cash flow guidance. The omission of 'free cash flow' from SEC filings—despite its status as a primary metric—exemplifies the unique value and challenge of the under-investigated earnings call domain. Our system successfully identified this as a longitudinal KPI across eight periods; notably, all four models correctly extracted the current '$310 million' value with only minor label variations. However, longitudinal consistency in other quarters was occasionally disrupted by temporal ambiguity arising especially from the Q&A session.

**The High Variability in Call Culture** The next quote highlights how much call culture varies across companies. JP MORGAN CHASE almost exclusively uses traditional performance measures in their calls. They follow a strict format, where they read aloud their earnings material, resulting in more extractions by the SEC-based models. Given the variability in call cultures, future studies should focus on scaling this benchmark to more companies.

> Starting on page 1, the Firm reported net income of $ 12.9 billion, EPS of $ 4.37 on revenue of $ 43.3 billion with an ROTCE of 19% (JPMorgan Chase & Co., 2024)

**Q&A Session** Earnings calls usually end with a more informal Q&A session. We examine the same example call featured in Figure 2, which demonstrates the huge value of mastering this domain. In the following quotes, we highlight modeling requirements unique to the Q&A.

1. Lyft does not make clear that it is a correction in their presentation, showing the need to keep track of *inconsistencies*.
2. Understanding that multiple people are speaking, and how they are related to the call.
3. Detection of negation, as Nikhil's mention of 500 basis points reflects what he believes the figure is *not*.
4. Detection of levels of abstraction, as the same metric can be referenced with different levels of specificity - e.g. 'margin expansion' vs. 'EBITDA margin expansion'.

> **Question from:** Nikhil Devnani (Analyst)
> "Can we just please clarify the EBITDA margin expansion? I think the slide says 500 basis points. ... But Erin, you mentioned 50. So, I think it is 50, but if you could just clarify that again, please?" (The Motley Fool Staff, 2024)

> **Answer from:** Erin Brewer – CFO
> "Thanks, Nikhil. This is Erin. And this is actually a correction from the press release. You're correct in my prepared remarks, I referenced 50 basis points of margin expansion" (The Motley Fool Staff, 2024)

## 7 Discussion

Our empirical results and qualitative analysis reveal the high variability between earnings calls. While market reaction to earnings reports is immediate and automated (Figure 2), the conversational nature and complexity of earnings calls have so far prevented similar high-speed, autonomous absorption. It is clear that earnings calls, down to the level of specific KPIs, influence the stock price. BERT-based financial KPI extraction models function well for SEC filings, as they are standardized by strict auditing and close to devoid of cultural variation. However, they fail to generalize to the subjective and promotional language used in earnings calls, posing challenges for current NLP methods. This complexity is compounded by variability in company-specific styles and cultural factors inherent to each organization or, indeed, each speaker. Our analysis and system provide insights into the challenges and opportunities in earnings calls, and while our human evaluation reveals an error rate necessitating human oversight, it is a step towards more efficient and faster processing of new information. Our exploration of KPIs present in earnings calls lays the groundwork for fine-tuning large language models for extractions.

## 8 Conclusion

This work characterizes the unique challenges and opportunities in automated KPI extraction from earnings calls. We introduce three novel benchmarks: the SEC filings benchmark (**SECB**), the earnings calls benchmark (**ECB**) with a smaller annotated subsample (**ECB-A**). Our empirical evaluation demonstrates that, while current KPI extraction methods show generalization capabilities across SEC filings datasets, they do not generalize to the more unstructured nature of earnings calls. Our qualitative analysis reveals why earnings calls present unique challenges, complicating KPI extraction due to subjective phrasing, company-specific terminology, and varying levels of formality that contrast sharply with structured SEC filings. Finally, validated by human evaluation, our work provides a robust baseline for the emerging task of automated KPI extraction from this valuable data source, with experiments and analysis laying the groundwork for future advances in real-time financial decision-making.

## Limitations

**Data Scale and Annotation** Due to the scarcity of experts in this domain and the compensation such experts usually demand, we were only able to recruit one annotator for the ECB-A dataset. We try to mitigate this by having the expert go over the annotations twice. Even though we attempt to get as diverse a sample as possible by randomly sampling 10 distinct companies in various industries, as mentioned in the paper, cultural differences between companies mean that we do not necessarily know how well these results generalize, especially outside of major US companies.

**Methods and Evaluation** The selection of a cutoff of 0.85 for semantic similarity was empirically derived, even though we provide a sensitivity analysis in the appendix C showing that our results are stable to some degree of change in parameters. Ideally, future work should aim to have this dynamically tuned by the actual model using clustering, e.g., K-means or, likely better-suited for this task, DBSCAN. There is a potential issue with data leakage, which could be and most likely is part of the training data for some of these LLMs.

**Evaluation** Some of the evaluation relies on an LLM as a judge, where it is important to note that an LLM as a judge is not always reliable; we try to mitigate this by using other automatic metrics as well as human evaluators. Because it is significantly easier to verify a correct result than annotate a ground truth, we utilize 3 human judges, who, however, must be noted as not experts in the field, though with a basic understanding.

## Ethics statement

There are risks with automated system especially in a financial context. There is a system risk that wrong extraction could result in the wrong financial decision temporaily pricing a stock at the wrong price. Which could lead to financial losses and gains for other actors in the market as well as the system user. There is a risk that systems like these put institutional investors even further in front of retail investors with less sophisticated setups for investment; however, faster, accurate pricing of securities also has the advantage of less volatility in the markets, as well as fairer prices. Our work is based on readily available data and adheres to the ACL Code of Ethics.

## Acknowledgments

We would like to thank the AAU-NLP group for helpful discussions and feedback on an earlier version of this article. We would like to give a special acknowledgement to Ernests Lavrinovics, for helping with evaluation of our final system. We want to also thank Alipes ApS for their support in facilitating and funding this research and as well as useful discussions with their Quant team. Rasmus Aavang is supported by the Industrial Ph.D. programme from Innovation Fund Denmark (grant code 4297-00016B). MZ and JB, were supported by the research grant (VIL57392) from VILLUM FONDEN. MZ also received funding from the Danish Government to Danish Foundation Models (4378-00001B).

## References

Rasmus T. Aavang, Giovanni Rizzi, Rasmus Bøggild, Alexandre Iolov, Mike Zhang, and Johannes Bjerva. 2025. Hifi-kpi: A dataset for hierarchical kpi extraction from earnings filings. *arXiv preprint arXiv:2502.15411*.

D Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Genesis Barahona Diaz and Yawen Hu. 2024. The impact of earnings call sentiment on stock market returns. Master's thesis, UIS.

Jason V. Chen, Venky Nagar, and Jordan Schoenfeld. 2018. Manager-analyst conversations in earnings conference calls. *Review of Accounting Studies*, 23(4):1315–1354.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Corporate Finance Institute. 2024. Earnings Call: Definition, Example, and What to Look For. Accessed: 2024-11-11.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Erhang Li, Fangqi Zhou, Fangyun Lin, Fucong Dai, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoran Wei, Haowei Zhang, Haowen Luo, Haozhe Ji, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, Jialiang Huang, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jingchang Chen, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexin Huang, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Liang Zhao, Liangsheng Yin, Lihua Guo, Lingxiao Luo, Linwang Ma, Litong Wang, Liyue Zhang, M. S. Di, M. Y Xu, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Panpan Huang, Peixin Cong, Peiyi Wang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Songyang Zhou, Tao Ni, Tao Yun, Tian Pei, Tian Ye, Tianyuan Yue, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjun Gao, Wentao Zhang, Xi Gao, Xiangwen Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyuan Li, Xu Chen, Xuecheng Su, Xuehai Pan, Xuheng Lin, Xuwei Fu, Y. Q. Wang, Yang Zhang, Yanhong Xu, Yanru Ma, Yao Li, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yiliang Xiong, Ying He, Ying Zhou, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuduan Wang, Yue Gong, Yuhan Wu, Yuheng Zou, Yukun Li, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehua Zhao, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhiyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Zizheng Pan, Zongqing Yao, Bei Feng, Hui Li, J. L. Cai, Jiaqi Ni, Lei Xu, Meng Li, Ning Tian, R. J. Chen, R. L. Jin, S. S. Li, Shuang Zhou, Tianyu Sun, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xinnan Song, Xinyi Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Dongjie Ji, Jian Liang, Jianzhong Guo, Jin Chen, Leyi Xia, Miaojun Wang, Mingming Li, Peng Zhang, Ruyi Chen, Shangmian Sun, Shaoqing Wu, Shengfeng Ye, T. Wang, W. L. Xiao, Wei An, Xianzu Wang, Xiaowen Sun, Xiaoxiang Wang, Ying Tang, Yukun Zha, Zekai Zhang, Zhe Ju, Zhen Zhang, and Zihua Qu. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Fahad Ebrahim and Mike Joy. 2024. Warwicknlp at semeval-2024 task 1: Low-rank cross-encoders for efficient semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Eval-*

*uation (SemEval-2024)*, pages 246–252. Association for Computational Linguistics.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Eugene F Fama. 1970. Efficient capital markets. *Journal of finance*, 25(2):383–417.

Financial Modeling Prep. 2024. Financial Modeling Prep. Accessed: 2024-10-16.

Alaa M Ghalayini and James S Noble. 1996. The changing basis of performance measurement. *International journal of operations & production management*, 16(8):63–80.

Google DeepMind. 2025. Gemini 3: A new era of agentic intelligence. Technical report, Google. Accessed: 2026-02-09.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant

Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Lars Hillebrand, Tobias Deußer, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. 2022. KPI-BERT: A joint named entity recognition and relation extraction model for financial reports. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 606–612. IEEE.

Inc. Hugging Face. 2024. Hugging face transformers: State-of-the-art natural language processing for pytorch, tensorflow, and jax. Python library for natural language processing and machine learning.

Investopedia. 2023. Earnings call definition. Accessed: 2024-11-11.

JPMorgan Chase & Co. 2024. Jpmorgan chase & co. q3 2024 earnings call transcript. Accessed: 2024-11-30.

Khurram Khan and Attaullah Shah. 2011. Understanding performance measurement through the literature. *African journal of business management*, 5(35):13410–13418.

Samira Khorshidi, Azadeh Nikfarjam, Suprita Shankar, Yisi Sang, Yash Govind, Hyun Jang, Ali Kasgari, Alexis McClimans, Mohamed Soliman, Vishnu Konda, Ahmed Fakhry, and Xiaoguang Qi. 2025. Odke+: Ontology-guided open-domain knowledge extraction with llms. *Preprint*, arXiv:2509.04696.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431, Dublin, Ireland. Association for Computational Linguistics.

Zhiqiang Ma, Grace Bang, Chong Wang, and Xiaomo Liu. 2020. Towards earnings call and stock price movement. *arXiv preprint arXiv:2009.01317*.

Stanimir Markov and Ari Yezegel. 2023. Giving retail investors a say in disclosure. *Available at SSRN 4836378*.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 3982. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

The Boeing Co. 2023. Q3 2023 earnings call. Accessed: 2025-13-03.

The Motley Fool Staff. 2024. Lyft (LYFT) Q4 2023 Earnings Call Transcript. `https://www.fool.com/earnings/call-transcripts/2024/02/13/lyft-lyft-q4-2023-earnings-call-transcript/`. Accessed: 2024-11-22.

U.S. Securities and Exchange Commission. 2024. Exchange act reporting and registration. Accessed: 2024-11-11.

U.S. Securities and Exchange Commission. 2026. U.S. Securities and Exchange Commission. Accessed: 2026-02-09.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication reliability–an empirical approach to interpreting inter-rater reliability. *arXiv preprint arXiv:2106.07393*.

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*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *Preprint*, arXiv:2306.05443.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang,

Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023a. Fingpt: Open-source financial large language models. *Preprint*, arXiv:2306.06031.

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023b. Investlm: A large language model for investment using financial domain instruction tuning. *Preprint*, arXiv:2309.13064.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *Preprint*, arXiv:2006.08097.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

| Dataset | Unique Labels |
|---|---|
| FiNER-139 | 140 |
| HiFi-KPI | 198K |
| **SECB** | 1,615 |

Table 8: Unique labels in each dataset.

| Dataset | Count |
|---|---|
| SECB | 21,258 |
| HiFi-KPI | 159,481 |
| FiNER-139 | 40,569 |

Table 9: BERT-SL1000 use of the special-OOS.

# A Detailed Experimental Setup

## A.1 Model setup

Following the explanation in (Loukas et al., 2022), we download their SEC-BERT-BASE model from Hugging face (Hugging Face, 2024), we use Hugging face to download their dataset and then train the model on the FiNER-139 train set, with the validation set as evaluation on a GTX 1080 TI for 10 epochs with early stopping patience of 2 with a batch-size of 32 (not specified in the (Loukas et al., 2022) paper), and set it to truncation at max length of 512 tokens (also not specified) to match the limit for the input size of the SEC-BERT-BASE model. The early stopping performance improvement was also based on the validation set of FiNER-139, the best model was achieved after all 10 epochs had run.

## A.2 Parser

The parser drops all tables, then tries to grab text inside ['p', 'div', 'span', 'section'], then, having done that, we deduplicate based on which extraction leads to the longest substring. After this, we do postprocessing, cleaning out malformed snippets in the same way as (Aavang et al., 2025), where we check if it starts with a "." and clean any potential leading whitespace, lastly dropping snippets not starting with a capital letter. Furthermore, we drop all snippets more than 3 std. deviations longer than the mean, meaning snippets longer than 4513 characters.

## A.3 Standardizing the Label Space between SEC-based models.

One of the core issues we had to figure out to compare the different SEC filings datasets across different datasets is how they are majorly finetuned to a certain dataset. Therefore, the labelset varies significantly in the number of unique labels present. We therefore have to convert between these sets we do this in the following way. First since the formatting of the labels are slightly different we cut out the "us-gaap" part of the SL1000 model predictions if predicting on the FiNER-139 dataset. Then we check if a label with the same name is part of the set and if not we give the pseudo label "UNK" that we can then use to track if it is at least right in the broader context of something being a fincial key figure or not. When it comes to the conversion the other way around where SEC-BERT-BASE does not have the label that is actually present if it predicts something for a label that then has a label not part of the SEC-BERT-BASE label space we then use the placeholder "UNK" as well. An interesting thing for this is that it means two things are true. None of the models can ever correctly predict the regex_label / regex_percentage labels. The special-OOS label that the SL1000 based models all can also never be completely correct.

## A.4 ECB & ECB-A

We segment the calls by looking for the occurrence of newlines in connection with a speaker's name. If a newline occurs before a speaker's name, we assume it to be a new speaker. Furthere we drop all transcriptions of what the operator says in the call as they will not say anything interesting or containing KPIs anyway. Finally, to manage context size for the bert based models, we further split and reconstruct the chunk with spaCy if any chunk contains more than 10 sentences.

# B Matching logic for automatic evaluation of ECB-A

To calculate automatic metrics, we take any direct supersets of both the model predictions and the ground truths for a certain chunk, and discard if there is anyway is any set tagged that is just the more elaborate label with less information. We then take all the candidate extractions and ground truth values then we compare these against each other. The best scoring extraction then consume the ground truth match. Meaning that there is a one-to-one mapping, this is such that a model can not artificially enhance its score by predicting the same correct label many times. The value match

is successful if ground truth and values either have the same value or, if they match the value, are multiplied by a different multiple of 1000s and have a cross-encoder score over 0.75. We count it as a value match as well. If the model has extracted something with the is_range parameter as true, then we allow it to potentially consume multiple ground truths, such that if, e.g. "4-5 unit a month" it can match both 4 and 5. Finally, for non-numeric values (e.g., 'Record'), we consider these a match if the Gestalt pattern matching similarity ratio between the extraction and ground truth strings are greater than 0.8.

## C Sensitivity in Threshold for the Semantic Clustering

| Model | Share (%) | Centroid (%) | Overlap (%) |
|---|---|---|---|
| Llama-3.3-70B | $77.50 \pm 0.66$ | $62.51 \pm 9.56$ | $31.03 \pm 1.21$ |
| Qwen3-30B-A3B | $78.70 \pm 0.03$ | $61.61 \pm 8.54$ | $25.49 \pm 1.56$ |
| Gemma-3-27B | $77.58 \pm 0.33$ | $58.93 \pm 8.59$ | $30.10 \pm 1.11$ |

Table 10: **Variance in model contributions.** Mean and standard deviation across the 4 different parameters setting. The average total number of extractions (dataset size) was $1354.75 \pm 193.34$.

Table 10 show the robustness of our threshold for the clustering cutoff by trying [0.75, 0.80, 0.85, 0.9] as well. We see that it doesn't have a big impact on the final clusters; however, there are, of course less clusters the higher you set the threshold.

## D ECB-A Annotation Setup

The annotation resembles the idea behind (Hillebrand et al., 2022) The expert annotator was put in a setting where they had to review at least the whole earnings transcript in one sitting to try to balance for fatigue in the mundane task across companies. They were given the possibility of using 3 options "Traditional", "Non-traditional", and "Value". They could then click on the interface in figure 5.

| | |
|---|---|
| One click | Traditional |
| Two click | Non-Traditional |
| Third click | Value |
| Fourth click | Reset |

Table 11: Interface for manual annotator

One click meant Traditional, another click "non-traditional" yet another "value" and a fourth click resets the annotation of the token. They were free to go back and forth by themselves during the tagging process and correct their annotations.
Then, at a later date, the annotations were confirmed by the same annotator in another interface, where one would be able to model relations between entities. Here, the annotator was instructed to annotate related entities, enabling later relation extraction. This was done by the interface in Figure 6 where you click on each entity and then create a relation.

### D.1 Traditional vs Non-Traditional Performance metrics

| Traditional Performance Measures | Non-traditional Performance Measures |
|---|---|
| Based on outdated traditional accounting system | Based on company strategy |
| Mainly financial measures | Mainly non-financial measures |
| Intended for middle and high managers | Intended for all employees |
| Lagging metrics (weekly or monthly) | On-time metrics (hourly, or daily) |
| Difficult, confusing, and misleading | Simple, accurate, and easy to use |
| Lead to employee frustration | Lead to employee satisfaction |
| Neglected at the shopfloor | Frequently used at the shopfloor |
| Have a fixed format | Have no fixed format (depends on needs) |
| Do not vary between locations | Vary between locations |
| Do not change over time | Change over time as the need changes |
| Intended mainly for monitoring performance | Intended to improve performance |
| Not applicable for JIT, TQM, CIM, FMS, RPR, OPT, etc. | Applicable |
| Hinders continuous improvement | Helps in achieving continuous improvement |

Table 12: Comparison of Traditional and Non-traditional Performance Measures from (Ghalayini and Noble, 1996)

Table 16 shows the most common labels annotated as either tradtional or non-traditional in ECB-A.

## E Evaluation

The annotation setup for the evaluation of the final system consists of 3 annotators, each were tasked with annotating 200 extractions each, with 100 extractions overlapping between the annotators. The setup was a command-line tool built in Python, that presented the annotators with the extracted KPI Label and value; they were then to evaluate if this extraction was correct or not from the corresponding chunk the system had extracted the KPI from. They could then either input yes or no, and if they selected no, they could present a short reasoning for why they did not think the extraction was correct. Full annotator guidelines for evaluation in 1 and the accompanying table specifying fiscal years in Table 13. The Cohen's kappa between annotators can be seen in Figure 7

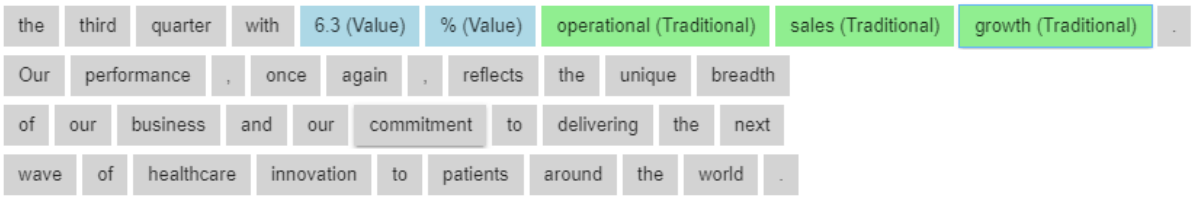


Figure 5: **Tagging Interface During Annotation** Example : 2024 Q3 JNJ earnings transcript

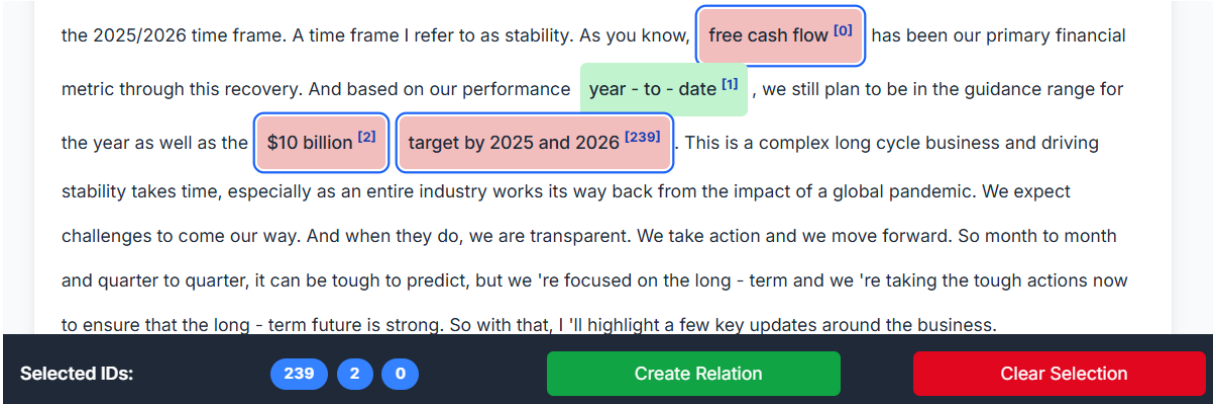


Figure 6: **Relation Extraction Annotation Interface.**

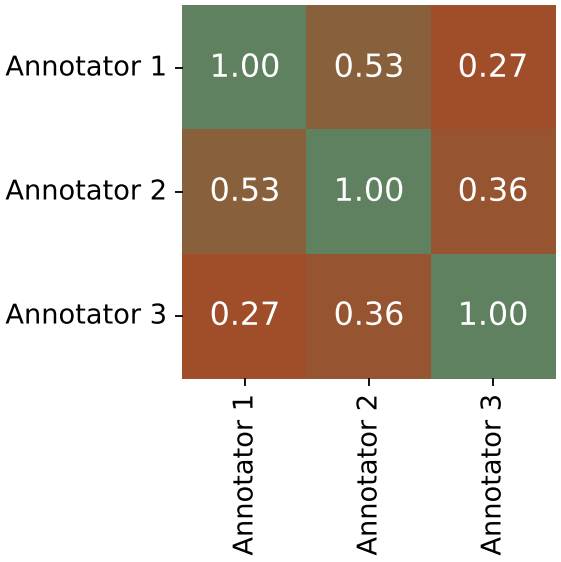


Figure 7: **Inter-Annotator Agreement (Cohen's Kappa).** Annotators 1 and 2 exhibit strong alignment, whereas Annotator 3 demonstrates notably lower agreement with the other evaluators.

| Ticker | Company | FY End | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|---|---|
| AAPL | Apple Inc. | Late Sep | Oct–Dec | Jan–Mar | Apr–Jun | Jul–Sep |
| HD | Home Depot | Late Jan | Feb–Apr | May–Jul | Aug–Oct | Nov–Jan |
| MSFT | Microsoft Corp. | Jun 30 | Jul–Sep | Oct–Dec | Jan–Mar | Apr–Jun |
| PG | Procter & Gamble | Jun 30 | Jul–Sep | Oct–Dec | Jan–Mar | Apr–Jun |
| AMZN | Amazon.com Inc. | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| BA | Boeing Co. | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| BAC | Bank of America | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| CAT | Caterpillar Inc. | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| CVX | Chevron Corp. | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| DOW | Dow Inc. | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| GOOGL | Alphabet Inc. | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| JNJ | Johnson & Johnson | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| JPM | JPMorgan Chase | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| KO | Coca-Cola Co. | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| NEE | NextEra Energy | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| PFE | Pfizer Inc. | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| PLD | Prologis Inc. | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |
| XOM | Exxon Mobil | Dec 31 | Jan–Mar | Apr–Jun | Jul–Sep | Oct–Dec |

Table 13: Fiscal year end dates and quarterly periods.

### E.1 Evaluator Guidelines

```
1. Task Overview
The goal of this task is to accurately identify and verify financial data points within company transcripts.
2. General Labeling Principles
Strict Extraction: All components of a label must be present in the actual text. Do not paraphrase.
Range Handling: If a value falls in the middle of a provided range in the text, it should be marked as
    correct.
Abbreviations: Be aware that standard financial abbreviations are frequently used.
e.g.:
Opex: Operating Expenses
Capex: Capital Expenditure
3. Financial Terminology & Unit Conversions
Basis Points (bps).
Conversion Table:
1 Basis Point | 1 bp | 0.01% | 0.0001 | 1/10,000
100 Basis Points | 100 bps | 1.00% | 0.01 | 1/100
4. Metadata & Period Verification
You will be provided with metadata at the top of the task to identify the source.
Example Header: Company: AAPL | Year: 2024 | Quarter: 4
Model/system Date: Date: 2024-Q4 (This is a system-generated field).
Fiscal Year vs. Calendar Year: please verify that the transcript references the correct period in the data.
Do not assume that Q1 is always January to March
Many companies have Fiscal Years (FY) that do not align with the Calendar Year.
Example: A company's Q1 might run from July to September.
Action: you can refer to the provided Fiscal Calendar Table to confirm the specific reporting period for
    the company in question.
5. Error Handling & Rejections
If you identify a label as incorrect, you must provide a justification.
Requirement: When rejecting a label, write a clear, concise reason explaining why it is invalid (e.g.,
    "Wrong time period," "Value not in text," "Hallucinated number").
```

Listing 1: The annotator guidelines we followed in annotation of the transcripts.

#### E.1.1 Full SECB Example


```
{
    "form_type": "10-K",
    "accession_number": "0000012927-23-000007",
    "filing_date": "20230127142633",
    "quarter_ending": "20221231",
    "company_name": "BOEING CO",
    "text": "The Companys deferred income tax assets of $12,301 can be
    used in future years to offset taxable income and reduce income
    taxes payable. The Companys deferred income tax liabilities of
    $9,306 will partially offset deferred income tax assets and result
    in higher taxable income in future years and increase income taxes
    payable. Tax law determines whether future reversals of temporary
    differences will result in taxable and deductible amounts that
    offset each other in future years. The particular years in which
    temporary differences result in taxable or deductible amounts
    generally are determined by the timing of the recovery of the
    related asset or settlement of the related liability. The deferred
    income tax assets and liabilities relate primarily to U.S. federal
    and state tax jurisdictions. From a U.S. federal tax perspective
    , the Company generated a tax NOL in 2020 that was carried back to
    prior years when the tax rate was 35% due to the CARES Act benefit
    as described above. The Company generated tax NOL in 2021 and
    interest carryovers in 2021 and 2022 that can be carried forward
    indefinitely and federal research and development credits that can
    be carried forward 20 years.",
"entities": [
 [
```

```
    45,
    51,
    "us-gaap:DeferredTaxAssetsGross",
    "instant",
    "2022-12-31",
    "2022-12-31",
    "iso4217:USD",
    12301000000.0
  ],
  [
    188,
    193,
    "us-gaap:DeferredIncomeTaxLiabilities",
    "instant",
    "2022-12-31",
    "2022-12-31",
    "iso4217:USD",
    9306000000.0
  ],
  [
    938,
    940,
    "us-gaap:EffectiveIncomeTaxRateReconciliation
    AtFederalStatutoryIncomeTaxRate",
    "duration",
    "2020-03-27",
    "2020-03-27",
    "xbrli:pure",
    0.35
  ]
]
}
```


### E.2 Annotator Guidelines

```
Basic guidelines:
We want to find KPI_FACTS in the text. To find these, you are going to do two things.
Tag entities in the text
Relate entities from (1) to each other.

The entities are broadly the following:
KPI
Scope
Value
Temporal Context or Modifier
Modifiers

Note that there are likely gonna be a lot of cases where not all categories are present
Follow these steps for every sentence:
Read for Understanding: First, read the entire text snippet to understand its full meaning.
Tag All Entities: Identify and label the five entity types in the sentence. Be precise with your
    highlighting (span selection).
Draw All Relations: Connect the entities according to the rules in Section 5. A KPI is the central hub for
    all relations.
Numerical Values:
Annotate all the number values that appear; these can be in the form of dollar figures, percentages,
    records, or the number of products produced.
Please annotate the $ sign as well as any order of magnitude specifier
E.g. , $5 million, five million dollars
Soft (Vague or qualitative) Values:
 For the year, we saw top-line growth from rate cases and riders across our jurisdictions.

Please also annotate more soft forms of KPIs, such as thousands of units, record number,
     strong growth, stable rate etc.
```

```
 strong  growth in YouTube subscriptions
 will  reflect the increases in depreciation and expenses
Increase -> value
depreciation and expenses ->  KPI.

KPI descriptions:
Having tagged the value, you should tag the related description.
The name of the metric. This is the core concept you will be annotating.
Rule: Tag the complete noun phrase that defines the metric. This must include essential modifiers that
     change the metric's definition.
Include: Adjectives like net, gross, adjusted, quarterly, annual, monthly, recurring.
Exclude: Determiners (a, the, our), verbs (reached, was), and descriptive but non-essential words (strong,
     record) if a KPI_VALUE is present

This is for example:
 quarterly  net revenue -> whole company description
 Google  Services operating margins -> that subdivision
 operating  margins -> whole company

Example sentence:
   [retention] is dropping down to [70%]
 Retention     -> KPI
70%->  value
Modifiers:
Do include Essential modifiers, e.g.,    net   ,    gross   ,    adjusted   ,    quarterly   ,    annual

Example  Our  [quarterly net revenue], driven by strong performance in the [cloud division], reached a
     [record] [$10B]

See clarification for the  cloud  division  in the following.
Scope - Subcomponent or Product:
If some KPI has to do with some specific subcomponent of the company, please be sure to tag that
     accordingly as well. Here, it is common for companies to go through a whole subdivision at once,
     meaning you often have to relate the subdivision to many KPIs.

Business division A formal part of the company's structure e.g.,  Cloud  Division ,  Services  Arm ,
      Global  Operations
A Product or Service: A specific offering from the company (e.g.,    iPhone   ,  Windows  11   ,  Model
     3   ).
A Geographical Market: A region where the company operates (e.g.,  North  American ,    Asia   ).

Temporal context / Modifier:
Please also connect with the temporal description, this can be    2025   ,    2026   ,  next  year ,
      in  the current quarter ,  year  to date  and many more

For temporal context, please make sure to also annotate if something is    expected   ,    target   ,
       projected   ,    expected
Relation:
You should use the annotation tool to relate the KPI description to its values; you have the    value_of
     relation
Such that each group describes one direct connection between value and KPI and potential modifiers.
Examples:
Our [quarterly net revenue], driven by strong performance in the [cloud division], reached a record [$10B]

[quarterly net revenue]<->[cloud division]<-> [$10B] => (value_of)

 Our  [guidance] for [Q1 2026] is [projected] [quarterly net revenue] from the [Cloud Division] of [$12B]."

Tagged Entities
[guidance] - MODALITY
[Q1 2026] - TEMPORAL_CONTEXT
[projected] - MODALITY
[quarterly net revenue] - KPI
[Cloud Division] - SCOPE
[$12B] - KPI_VALUE

Relation These would all be related in the same group:
[guidance]    [Q1 2026]   [projected]   [quarterly net revenue]   [Cloud Division]    [$12B]

Multiple entities:

"[Cloud Division] [revenue] was [$10B] in [2024], while [Services Arm] [profit] is [expected] to be [$3B]."

In this case, there are two distinct facts about two different KPIs. You would create two separate KPI
     relations groups
KPI_Fact 1:
KPI: [revenue]
KPI_VALUE: [$10B]
SCOPE: [Cloud Division]
TEMPORAL_CONTEXT: [2024]
KPI_Fact 2:
KPI: [profit]
KPI_VALUE: [$3B]
SCOPE: [Services Arm]
MODALITY: [expected]
```

Listing 2: The annotator guidelines we followed in annotation of the transcripts.

## F Cost & Runtime

Table 14 shows both the runtime and the cost of running these large LLM-based models of running the different LLMs on the Apple Q1 2023 earnings call transcript, which ends up being 54 chunks. Gemma-3-27B is actually free (for us) because we use Google's free endpoint. The runtime also shows that from a financial perspective there is still gains to be made just from being faster. It is also clear that the slower the model the better performance.

## G Prompt Setup Details

We use Openrouter[3] to run Llama-70B, Gemini-3 pro and Qwen-3 model, we use Google Cloud[4] to run Gemma model and we use the OpenAI Python API library. For the models that support an extraction schema, we utilize the extraction schema in 4, and for the models that do not, we only utilize the prompt in 5

[3] See https://openrouter.ai/.
[4] See https://cloud.google.com/

| | Time Metrics (s) | | | Cost Metrics ($) | | |
|---|---|---|---|---|---|---|
| **Model** | **Mean** | **Std Dev** | **Total** | **Mean** | **Std Dev** | **Total** |
| Gemma-27b-it | 15.06 | 22.32 | 813.04 | 0.000000 | 0.000000 | 0.000000 |
| Qwen3-30b-a3b | 5.94 | 18.18 | 320.67 | 0.000418 | 0.000584 | 0.022553 |
| Llama-3.3-70b-instruct | 5.72 | 17.07 | 308.65 | 0.002075 | 0.001187 | 0.112056 |
| Gemini-3-pro preview | 48.22 | 37.29 | 2603.93 | 0.049013 | 0.036039 | 2.646720 |

Table 14: Comparison of execution time and API costs across evaluated models.

| KPI Label | Unique Companies |
|---|---|
| revenue | 5 |
| free cash flow | 4 |
| net income | 4 |
| capex | 3 |
| cash flow | 2 |

Table 15: Top 5 most common KPIs consistently reported across unique companies.

| Rank | Traditional KPI | Occurrences | Non-Traditional KPI | Occurrences |
|---|---|---|---|---|
| 1 | Revenue | 10 | Active Devices | 2 |
| 2 | EPS | 6 | Apple Pay Available | 2 |
| 3 | Operational Sales Growth | 6 | 737S Production Deliveries | 2 |
| 4 | Operating Margin | 5 | FDA Approval | 2 |
| 5 | Revenues | 5 | Freeform, A Brand-New App | 1 |

Table 16: Top 5 most common Traditional and Non-Traditional KPIs.

```
You are an expert financial entity extractor. Your sole task is to read `### TEXT TO ANALYZE ###` section
     and extract all entities according to the JSON schema

    ## Entity Extraction Instructions ##
    Entities could be:

    * `kpi_name` The name of the metric.
        * **Rule:** Tag the complete noun phrase, including essential modifiers like "net", "gross",
            "adjusted", "quarterly", or "annual".
        * **Rule:** EXCLUDE non-essential fluff ("strong," "record") and determiners ("a," "the," "our").
        * Examples: "quarterly net revenue", "Google Services operating margins", "retention rate".

    * `kpi_value` The *quantifiable* value of the KPI.
        * **Rule:** This must be a numerical value.
        * **Math Rule:** If a range is provided (e.g., "$10-20M"), calculate the arithmetic average for the
            `Value` field, but record the bounds in the range fields.
        * Examples: "$10B", "70%", "five million dollars", "thousands of units".

    * `qualitative_desc`  A *subjective* or *non-numerical* description of the KPI's performance or trend.
        * **Rule:** This must highlight a specific qualitative milestone.
        * Examples: "strong growth", "stable rate", "increase", "dropping down", "disappointing results",
            "record number".

    * `scope` The specific business unit, product, or market the KPI refers to.
        * Examples: "Cloud Division", "Services Arm", "iPhone", "North American", "Services", "Boeing
            Commercial", "Boeing Defense and Space"

    * `date` The temporal context for the KPI.
        * **Rule:** This should include any relevant time frames or specific dates. Including if something
            is a future projection or historical fact.
        * Examples: "2024," "Q1 2026," "next year," "in the current quarter," "year to date", "end of
            year", "expect", "project"*

    * `modality` The certainty or context of the fact (e.g., if it's a projection vs. a reported fact).
        * **Rule:** Forward of backwards looking (e.g., "guidance" and "projected").
        * Examples: "projected", "expected", "target", "guidance".

    Having identified the entities, you should structure them into groups relating relevant entities
         together.

    ### 2. Field Definitions ###
    * **Source:** The exact text span from the input from which the metric was derived.
    * **Entities:** A list of all relevant entities in the text.
    * **Source Value:** The original text value of the metric.
    * **Label:** Construct a concise label from the entities in the text this could be `scope`, `kpi_name`,
         `date` and `modality`. Use only entities present in the source text.
    * **Value:** The numerical value as a float. If a range, use the average.
    * **Value_NonNumeric:** If the value is a non numerical highlight of performance.
    * **Is_Range:** Boolean indicating if the value comes from a range.
    * **Top_of_range / Bottom_of_range:** The specific upper/lower bounds if Is_Range is True.

    **Label Construction Rule:**
    You must generate a standard `Label` for each group to serve as a unique ID.
    * **Source:** Use ONLY the text of the entities found in that specific group.
    * **Order:** Construct the string in this exact precedence:
        1.  `scope`
        2.  `modality`
        3.  `kpi_name`
        4.  `date`
    * **Formatting:** Separate parts with a single space.
    * **Example:** If you find Scope="Cloud", KPI="Revenue", Date="Q1", the Label is "Cloud Revenue Q1".

    Example 1:
    "Quarterly revenues crossed the $10 billion mark for the first time"
    Extracted as
    {
    "Entities": [
            {"text": "Quarterly", "category": "date"},
            {"text": "revenues", "category": "kpi_name"},
            {"text": "$10 billion", "category": "kpi_value"}
        ],
    "Groups" : [{
        "Source": "Quarterly revenues crossed the $10 billion mark for the first time",
        "Entities": [
                {"text": "Quarterly", "category": "date"},
                {"text": "revenues", "category": "kpi_name"},
                {"text": "$10 billion", "category": "kpi_value"}
            ],
        "Source Value": "$10 billion",
        "Is_Range": false,
        "Top_of_range": null,
        "Bottom_of_range": null,
        "Value": 10000000000.0,
        "Value_NonNumeric": null,
        "Label": "revenues Quarterly"
```

```
}]
}

Example 2:
"Boeing Defense and Space. BDS booked $6 billion in orders during the quarter. Revenue was $5.5 billion"
Extracted as
{
"Entities": [
    {"text": "Boeing Defense and Space", "category": "scope"},
    {"text": "BDS", "category": "scope"},
    {"text": "$6 billion", "category": "kpi_value"},
    {"text": "orders", "category": "kpi_name"},
    {"text": "during the quarter", "category": "date"},
    {"text": "Revenue", "category": "kpi_name"},
    {"text": "$5.5 billion", "category": "kpi_value"}
],
"Groups" : [{
    "Source": "Boeing Defense and Space. BDS booked $6 billion in orders during the quarter",
    "Entities": [
                    {"text": "Boeing Defense and Space", "category": "scope"},
                    {"text": "BDS", "category": "scope"},
                    {"text": "orders", "category": "kpi_name"},
                    {"text": "during the quarter", "category": "date"},
                    {"text": "$6 billion", "category": "kpi_value"}
                ],
    "Source Value": "$6 billion",
    "Is_Range": false,
    "Top_of_range": null,
    "Bottom_of_range": null,
    "Value": 6000000000.0,
    "Value_NonNumeric": null,
    "Label": "Boeing Defense and Space BDS orders during the Quarter"}
,
{
    "Source": "Boeing Defense and Space. BDS booked $6 billion in orders during the quarter. Revenue
         was $5.5 billion",
    "Entities": [
                    {"text": "Boeing Defense and Space", "category": "scope"},
                    {"text": "BDS", "category": "scope"},
                    {"text": "Revenue", "category": "kpi_name"},
                    {"text": "$5.5 billion", "category": "kpi_value"},
                    {"text": "during the quarter", "category": "date"}
                ],
    "Source Value": "$5.5 billion",
    "Is_Range": false,
    "Top_of_range": null,
    "Bottom_of_range": null,
    "Value": 5500000000.0,
    "Value_NonNumeric": null,
    "Label": "Boeing Defense and Space BDS Revenue during the Quarter"
}]
}

Example 3:
"We expect net income to be in the range of $1.2 billion to $1.4 billion for the fiscal year 2026."
Extracted as
{
    "Entities": [
        {"text": "expect", "category": "modality"},
        {"text": "net income", "category": "kpi_name"},
        {"text": "$1.2 billion", "category": "kpi_value"},
        {"text": "$1.4 billion", "category": "kpi_value"},
      {"text": "fiscal year 2026", "category": "date"}
    ],
    "Groups": [{
        "Source": "We expect net income to be in the range of $1.2 billion to $1.4 billion for the
             fiscal year 2026.",
        "Entities": [
            {"text": "expect", "category": "modality"},
            {"text": "net income", "category": "kpi_name"},
            {"text": "fiscal year 2026", "category": "date"},
            {"text": "$1.2 billion", "category": "kpi_value"},
            {"text": "$1.4 billion", "category": "kpi_value"}
        ],
        "Label": "expect net income fiscal year 2026",
        "Source Value": "$1.2 billion to $1.4 billion",
        "Value": 1300000000.0,
        "Value_NonNumeric": null,
        "Is_Range": true,
        "Top_of_range": 1400000000.0,
        "Bottom_of_range": 1200000000.0
    }]
}

Example 4:
"We have seen record high use of our AI cloud tool."
Extracted as
```

```
    {
        "Entities": [
            {"text": "record high", "category": "qualitative_desc"},
            {"text": "use", "category": "kpi_name"},
            {"text": "AI cloud tool", "category": "scope"}
        ],
        "Groups": [{
            "Source": "We have seen record high use of our AI cloud tool.",
            "Entities": [
                {"text": "record high", "category": "qualitative_desc"},
                {"text": "use", "category": "kpi_name"},
                {"text": "AI cloud tool", "category": "scope"}
            ],
            "Label": "AI cloud tool use",
            "Source Value": "record high",
            "Value": null,
            "Value_NonNumeric": "record high",
            "Is_Range": false,
            "Top_of_range": null,
            "Bottom_of_range": null
        }]
    }

    ### 3. Context ###
    - **Stock Ticker:** $tickr
    - **Fiscal Period:** $fiscal_period
    - **Time of Report:** $time_of_report

    ### 4. TASK ###
    Analyze the following text and generate the JSON output.
    If no metrics are found, output the structure with empty lists.
    Output ONLY the valid JSON object and nothing else.
    ### TEXT TO ANALYZE ###
    <text> $target_text </text>
```

Listing 3: The few-shot prompt used.

```
{
    "name": "financial_entity_extraction",
    "strict": True,
    "schema": {
        "type": "object",
        "properties": {
            "Entities": {
                "type": "array",
                "description": "A comprehensive list of all financial entities found in the text,
                     classified by type.",
                "items": {
                    "type": "object",
                    "properties": {
                        "text": {
                            "type": "string",
                            "description": "The exact substring extracted from the source text."
                        },
                        "category": {
                            "type": "string",
                            "description": "The classification of the entity.",
                            "enum": [
                                "kpi_name",
                                "kpi_value",
                                "qualitative_desc",
                                "scope",
                                "date",
                                "modality"
                            ]
                        }
                    },
                    "required": ["text", "category"],
                    "additionalProperties": False
                }
            },
            "Groups": {
                "type": "array",
                "description": "Logical groupings of entities that form a single financial fact.",
                "items": {
                    "type": "object",
                    "properties": {
                        "Source": {
                            "type": "string",
                            "description": "The full text span containing the fact. Must include previous
                                 sentences if context (like Scope) is needed."
                        },
                        "Entities": {
                            "type": "array",
                            "description": "The subset of entities that belong to this specific fact.",
                            "items": {
```

```
                                "type": "object",
                                "properties": {
                                    "text": {"type": "string"},
                                    "category": {"type": "string"}
                                },
                                "required": ["text", "category"],
                                "additionalProperties": False
                            }
                        },
                        "Source Value": {
                            "type": "string",
                            "description": "The raw string representation of the value (e.g., '$10-12
                                million')."
                        },
                        "Label": {
                            "type": "string",
                            "description": "Unique ID. Strict Order: [Scope] [Modality] [KPI Name] [Date].
                                Use ONLY entities present in this group."
                        },
                        "Value": {
                            "type": ["number", "null"],
                            "description": "The numeric representation. If a range, this is the average."
                        },
                        "Value_NonNumeric": {
                            "type": ["string", "null"],
                            "description": "The qualitative description if no number exists (e.g. 'record
                                high')."
                        },
                        "Is_Range": {
                            "type": "boolean",
                            "description": "True if the source mentions a lower and upper bound."
                        },
                        "Top_of_range": {
                            "type": ["number", "null"],
                            "description": "The upper bound of the range."
                        },
                        "Bottom_of_range": {
                            "type": ["number", "null"],
                            "description": "The lower bound of the range."
                        }
                    },
                    "required": [
                        "Source", "Entities", "Source Value", "Label",
                        "Value", "Value_NonNumeric",
                        "Is_Range", "Top_of_range", "Bottom_of_range"
                    ],
                    "additionalProperties": False
                }
            }
        },
        "required": ["Entities", "Groups"],
        "additionalProperties": False
    }
}
```

Listing 4: Extraction Schema Used

## H LLM-as-a-Judge setup

For the LLM as a judge setup, we use DeepSeek-V3.2 (DeepSeek-AI et al., 2025) we access through the deepseek platform[5] We use the following prompt for the LLM-as-a-judge setup

```
You are a strict financial auditor evaluating an
     Information Extraction system.

TASK:
Determine if the 'Model Prediction' refers to the
     same financial concept as the 'Ground Truth',
     given the context.

CONTEXT TEXT:
"{context_text}"

SHARED VALUE: {value_str}

COMPARISON:
1. Ground Truth Label: "{gt_label}"
2. Model Prediction Label: "{pred_label}"

INSTRUCTIONS:
- If the Model Prediction is a valid synonym or a
     reasonable extraction of the Ground Truth
     concept, say YES.
- If the Model Prediction captures a DIFFERENT
     concept (e.g., "Gross Profit" vs "Net
     Profit"), say NO.

OUTPUT FORMAT:
Return ONLY a JSON object:
{{
  "reasoning": "Brief explanation of your decision",
  "is_equivalent": true or false
}}
```

Listing 5: The LLM as a judge prompt used.

[5] https://platform.deepseek.com/