HiFi-KPI: A Dataset for Hierarchical KPI Extraction from Earnings Filings

Rasmus T. Aavang[1,2], **Giovanni Rizzi**[2], **Rasmus Bøggild**[2]
Alexandre Iolov[2], **Mike Zhang**[3,1,4], **Johannes Bjerva**[1]
[1]Department of Computer Science, Aalborg University, Denmark
[2]ALIPES ApS, Denmark
[3]University of Copenhagen, Denmark
[4]Pioneer Centre for AI, Denmark
rtaj@cs.aau.dk

Abstract

Accurate tagging of earnings reports can yield significant short-term returns for stakeholders. The machine-readable inline eXtensible Business Reporting Language (iXBRL) is mandated for public financial filings. Yet, its complex, fine-grained taxonomy limits the cross-company transferability of tagged Key Performance Indicators (KPIs). To address this, we introduce the **Hi**erarchical **Fi**nancial **K**ey **P**erformance **I**ndicator (HiFi-KPI) dataset, a large-scale corpus of 1.65M paragraphs and 198k unique, hierarchically organized labels linked to iXBRL taxonomies. HiFi-KPI supports multiple tasks and we evaluate three: KPI classification, KPI extraction, and structured KPI extraction. For rapid evaluation, we also release **HiFi-KPI-Lite**, a manually curated 8K paragraph subset. Baselines on HiFi-KPI-Lite show that encoder-based models achieve over 0.906 macro-F1 on classification, while Large Language Models (LLMs) reach 0.440 F1 on structured extraction. Finally, a qualitative analysis reveals that extraction errors primarily relate to dates. We open-source all code and data at https://github.com/aaunlp/HiFi-KPI.

1. Introduction

As unstructured data in finance grows (Lewis and Young, 2019), there is a strong interest in developing NLP benchmarks and methods. Financial reporting standards in the EU, UK, USA, and others require public companies to file financial summaries with iXBRL. iXBRL provides machine-readable tags for key information. However, these tags are highly fine-grained (up to 198K labels), limiting generalization (see Figure 1). Making annotation difficult, time-consuming, and expensive, even for expert taggers, with over 34% of documents containing errors (Bricker, 2020). Although several datasets have been created for the financial domain (Chen et al., 2022b; Jørgensen et al., 2023) and even for financial reports (Loukas et al., 2021, 2022; Sharma et al., 2023; Lai et al., 2024), the rich structural information embedded by iXBRL remains an untapped resource for enhancing contextual understanding and enabling cross-company generalization.

We present the first resource that provides context for the KPI tags present in SEC filings. iXBRL contains several taxonomies. We focus on two: the Presentation taxonomy that dictates the layout in reports and the Calculation taxonomy that defines arithmetic relationships. We provide context for the KPI tags by creating a unified Presentation taxonomy across companies, and the same for the Calculation taxonomy. Further, we are the first to provide KPI tag context by preserving the relationships with time periods, numerical values, and currencies. Our research into leveraging this rich structure is guided by the following questions:

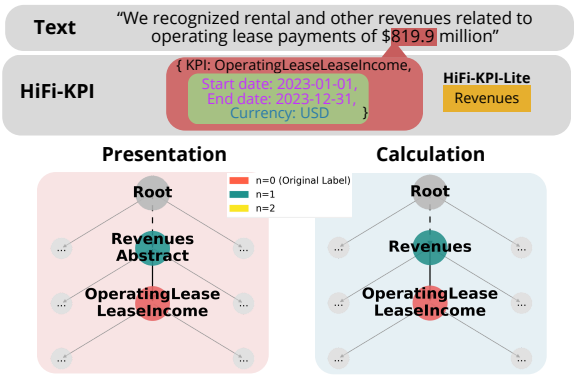

Figure 1: **HiFi-KPI** provides valuable contextual information for key performance indicators, including associated time periods and currencies. Additionally, HiFi-KPI offers hierarchical context, demonstrated by the hyper-specific OperatingLeaseLeaseIncome mapping to RevenuesAbstract (n=1) in the Presentation taxonomy, and Revenues (n=1) in the Calculation taxonomy. Finally, **HiFi-KPI-Lite** provides expert mapping to Revenues.

RQ1. To what extent can the structural hierarchies of iXBRL taxonomies be leveraged to generalize hyper-specific financial labels for automated information extraction?

RQ1.1. Which iXBRL taxonomy, Presentation or Calculation, provides the most effective structural basis, and how does model performance vary across different levels of label granularity?

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 441–455
11–16 May 2026.

Resource	Samples	Unique Labels	Context	Taxonomy
Loukas et al. (2022)	1.12M	139	×	×
Sharma et al. (2023)	79K	2.8K	×	×
HiFi-KPI (Ours)	1.65M	198K	✓	✓
HiFi-KPI-Lite (Ours)	8K	5	✓	N/A

Table 1: Comparison of HiFi-KPI & HiFi-KPI-Lite and previous SEC filings based resources.

RQ1.2. How well do SOTA Large Language Models extract information from expert-defined labels?

Accurate information extraction from iXBRL has great value as accurate analysis of earnings reports can yield significant short-term returns for investors (Ke and Ramalingegowda, 2005), besides helping corporate compliance. We investigate this by contributing:

- HiFi-KPI: An iXBRL-based dataset with 1.65M paragraphs and 4.5M entities, and a smaller HiFi-KPI-Lite with four expert-defined labels.
- Two cleaned and unified taxonomies enabling user-specified levels of granularity with our granularity selection method that overcomes the hyper-specificity of the original labels.
- Baselines for text classification, sequence labeling, and LLM-based structured extraction for HiFi-KPI-Lite.

2. Related Work

Financial NLP Datasets. Financial NLP datasets cover sentiment analysis (Gupta et al., 2020), named entity recognition (Alvarado et al., 2015; Shah et al., 2022), and numerical reasoning (Chen et al., 2022a). While early approaches used rule-based methods (Cong et al., 2007; Sheikh and Conlon, 2012; Hutto and Gilbert, 2014), modern efforts leverage large corpora e.g. from SEC Filings (Loukas et al., 2021). KPI-EDGAR (Deußer et al., 2022) that defines an NER and relation extraction task. Additionally, FiNER-139 (Loukas et al., 2022) and Sharma et al. (2023), which both leverage SEC filings to create a sequence labeling task with a limited label set. FiNER-139 limiting itself to the 139 most common labels, resulting in high sparsity (80.42% untagged entries) and Sharma et al. (2023) using US-GAAP metrics only; HiFi-KPI introduces valuable context, enabling more complex generative tasks and taxonomies, which in turn facilitate the analysis of the full label set. This makes HiFi-KPI the first resource to utilize the full labelset, resulting in 0% untagged entries, and a density of 2.77 tags/entry.

Language Models in Finance. Transformer-based models (Vaswani et al., 2017) has been a huge influence on the NLP field, leading to several specialized models for the finance domain, a notable example is the proprietary LLM BloombergGPT (Wu et al., 2023), and commonly used open-source financial language models are FinBERT variants (Yang et al., 2020; Araci, 2019). To move from sentiment analysis to granular data extraction, we focus on sequence labeling. While pretrained models like SEC-BERT (Loukas et al., 2022) exist, a fine-tuned token classification version is not public. We therefore use google-bert/bert-base-uncased (Devlin et al., 2019) to establish a reproducible baseline. Sentence transformers (Reimers and Gurevych, 2019) create semantically informative sentence embeddings orders of magnitude faster than BERT, making them ideal for quickly creating static representations to be able to iterate over many different label spaces for our hierarchical task. The introduction of GPT-3 (Brown et al., 2020) popularized decoder-only architectures. This new paradigme lead to state-of-the-art open-source models such as gemma-3-27B (Team et al., 2025), DeepSeek-V3.1 (DeepSeek-AI et al., 2024), Qwen3-30B-A3B (Qwen Team, 2025) and mistral-Small-3.2-24B (Mistral AI, 2025). HiFi-KPI can benchmark these generative models not only on extracting the correct label but also on extracting important contextual information for these labels.

SEC Filings U.S. public companies must file quarterly (10-Q) and annual (10-K) reports (U.S. SEC, 2000, 2024a), which contain standardized financial statements. Since June 15, 2020, SEC filings follow the inline eXtensible Business Reporting Language (iXBRL) open standard for accounting data.(SEC, 2018; Caltuna, 2020). iXBRL is a standard that enables automatic computational extraction of tagged KPIs. SEC filings in iXBRL contain detailed financial statements that often follow standard templates. (see Figure 2) HiFi-KPI contains the contextual data available in iXBRL. This context is paramount for actually assessing companies' financial situation, as one will often see KPIs reported in relation to the previous quarter or year. A common formulation is like the following

> $0.52 and $0.34 per basic share, or $0.49 and $0.34 per diluted share, respectively, for the three months ended March 31, 2020 and 2019, respectively.

The word "respectively" appears over 650k times in the dataset, highlighting the need to link figures like "earnings per share" to their correct time period.

> The 2019 facilities consist (...) (i) a $675.0 million United States dollar-denominated revolving credit facility, (ii) a CAD $70.0

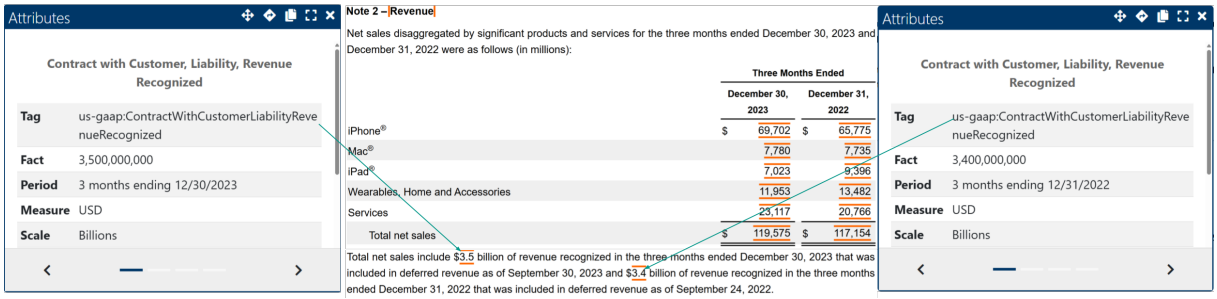

Figure 2: **Example HiFi-KPI.** The dataset is created by scraping the contextual snippets often following tables in SEC filings; Example is from the SEC's Inline XBRL Viewer for the 10Q for AAPL Q4 2023.

million Canadian dollar-denominated (..)
(iii) a €200.0 million euro-denominated ...

Since the data includes figures in USD, EUR, and CAD, conflating these currencies would be a fundamental error due to their vastly different value.

iXBRL Taxonomies iXBRL is highly complex, just the US-GAAP (Generally Accepted Accounting Principles) taxonomy, that describes the accounting standards of the U.S. Securities and Exchange Commission (SEC), contains about 17,000 unique tags (Caltuna, 2020). iXBRL guidelines require annotators to choose the most specific tag. Lastly, each company can also extend the taxonomy. This leads to companies using iXBRL differently, with highly specific and company-based tags. e.g.

- `aapl:EquitySecuritiesFVNIAccumulatedGrossUnrealizedGainBeforeTax`
- `tsla:LeasedAssetsNet`

Because company-specific tags are not standardized across firms, iXBRL provides a framework of relations and taxonomies to aid comparisons. This paper focuses on two of them as they define key financial meaning

`.cal` (Calculation) Describes the arithmetic relations between concepts. (e.g., "fact A + fact B = fact C") (XBRL International, 2025b).

For example, Tesla's tag direct parent, `tsla:LeasedAssetsNet` ↪ `us-gaap:Assets`. for Apple, we see `aapl:Equity...GainBeforeTax` ↪ `us-gaap:EquitySecuritiesFvNiCost`, both aggregating to `us-gaap:AssetsNet` at the root.

`.pre` (Presentation) Which arranges tags in a structure that is appropriate to represent the hierarchical relationships in business data. (Open Risk, 2023; XBRL International, 2025a).

For example, Tesla's tag has a direct parent: `tsla:LeasedAssetsNet` ↪ `us-gaap:AssetsNoncurrentAbstract`. Apple's tag have a different direct parent, but both share the parent `us-gaap:StatementOfFinancialPositionAbstract`.

	Train (Lite)		Dev. (Lite)		Test (Lite)	
Cutoff Date	2023.10.31		2024.05.31		2024.06.01	
# Paragraphs	1.33M	(847)	149K	(755)	168K	(847)
# Entities	3.64M	(2,399)	418K	(2,545)	448K	(2,399)
Avg. Entities	2.73	(2.83)	2.80	(3.37)	2.67	(2.83)
Avg. Words	87.31	(77.20)	87.69	(83.08)	86.26	(77.20)
Avg. Length (Chars)	556.00	(498.06)	558.46	(538.65)	549.68	(498.06)

Table 2: **Dataset Statistics.** We show the full dataset and the lite version statistics in brackets.

HiFi-KPI enables viewing these company-specific tags, as represented by their parents in the taxonomy, at a desired granularity. We release the code for the implementation of our bottom-up granularity selection algorithm for easy use.

3. HiFi-KPI

HiFi-KPI supports multiple downstream tasks such as text classification, sequence labeling, structured information extraction, multi-label classification, and financial question answering. Table 2 summarizes key statistics for both the full dataset and the lite subset with expert labels made for fast inference. We collected all quarterly and annual reports published between 2017-01-01 and 2024-06-01, yielding 42,018 quarterly and 14,389 annual reports. We parsed each iXBRL document using `beautifulsoup` (Richardson, 2007) and regular expressions to extract text spans along with every iXBRL tag, associated time period, and numeric values. Any unparsed spans were discarded. The extracted iXBRL tags show high data integrity with no snippets that wrongly include a preceding dollar sign ($) or a succeeding magnitude specifier. From these steps, we obtained 1.9M paragraphs (1.11M from quarterly reports; 0.74M from annual) and 5.3M tagged entities. The long-tailed label distribution (Figure 3) approximates a power law, though the most frequent labels occur even less often than suggested by the fit.

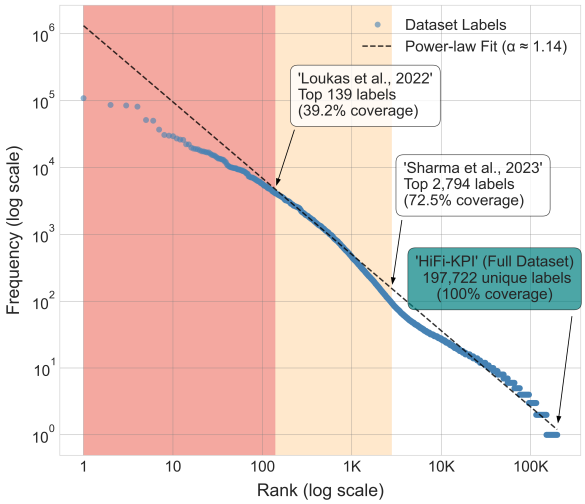

Figure 3: **Rank-frequency Distribution.** iXBRL labels on a log-log scale follow a power-law distribution, indicated by the fitted dashed line. Shaded regions highlight the cumulative coverage of all labels in HiFi-KPI by the most frequent label count from (Loukas et al., 2022) and (Sharma et al., 2023)

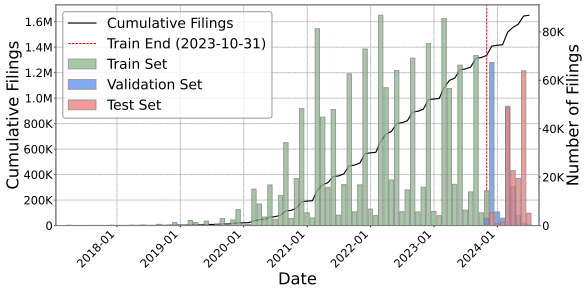

Figure 4: **Temporal Data Split** HiFi-KPI is split temporally into a train, validation, and test set. HiFi-KPI is evenly spread in time, with quarterly peaks.

Dataset Refinement and Quality Evaluation We evaluate the original company annotators' accuracy as well as the parsed entries' quality by manual verification of samples. We randomly sampled 100 entries from each data split (train, validation, test). We mitigate annotator fatigue by presenting half of the entries in the order of train, validation, and test, while the remaining half were presented in the reverse order. We ask an industry professional to locate errors with respect to character span (Span), temporal period (Date), currency (Currency), and value (Value). Further, we asked the annotator to note which entries seemed to be incorrectly parsed. This could be that the start or end of the text snippet suggests that there is more relevant context. Since the iXBRL standard is highly complex, we were not able to assess the quality of the iXBRL tag chosen by the company annotator. The results of this test can be seen in Table 3. Although the test revealed a low overall error rate, we used the reported errors to further refine the dataset, creating regular expressions to filter out snippets with "false starts" (e.g., those beginning with whitespace or a non-capitalized letter). After filtering out the errors, we ran the same experiment again with 100 newly sampled snippets from the train, validation, and test splits for a total of 600 checked samples. We find under 1% of entities having any errors, resulting in the final cleaned dataset having 1.65M snippets and 4.5M entities.

Temporal Split. We split the dataset temporally to eliminate forward-looking bias, so the model is not able to infer previous extractions based on things learned from the future. The temporal split means we have a short time window for the test and validation split. Therefore, most companies only report a 10-K in either the validation split or the test split timeframe. Therefore, we make the validation split more representative by using company-specific cutoff dates. We still follow a temporal split, setting the company-specific cutoff date, resulting in something as close to a 50/50 split as possible. If only one filing is present, the choice is random. This approach leads to 89.88% of companies in the validation set also appearing in the test set. Finally, all companies with their first filing after 2023-10-31 are assigned to the test set to better evaluate generalization to previously completely unseen domains. The final temporal split of entries in the dataset can be seen in Figure 4.

Taxonomy Creation. Each iXBRL report has several file attachments describing parent–child relationships. We focus on the relationships in two of the complex formats: `.cal` and `.pre`. We download attachments and parse them with Arelle (Arelle Development Team, 2025) to JSON. Since different companies use the taxonomy in different ways, children can have multiple parents. Therefore, we aggregate the per-document hierarchies to build two unified taxonomies. We do this by the following formula, using the most common parent as the parent. P_{master}:

$$P_{\text{master}}(t) = \arg\max_{p \in \mathcal{P}(t)} \text{count}(p, t),$$

where t is a tag, $\mathcal{P}(t)$ its possible parents, and $\text{count}(p, t)$ the frequency of documents with the parent–child relation. In the very rare case of ties, we pick randomly. Figure 5 shows statistics for the unified Presentation (1.57M edges, depth=21) and Calculation (202K edges, depth=12) taxonomies. Figure 5 shows that the Calculation and Presentation taxonomies both provide an even spread of nodes across all depths in the taxonomy, of course, with fewer nodes at the very top of the tree. Further Figure 5 also shows how the Presentation taxon-

Error Category	Train		Validation		Test		Total	
	Before	After	Before	After	Before	After	Before	After
Sample	277 (100)	287 (100)	280 (100)	285 (100)	257 (100)	242 (100)	814 (300)	814 (300)
Errors - entity count (entry count)								
Date	1 (1)	0 (0)	2 (2)	0 (0)	0 (0)	1 (1)	3 (3)	1 (1)
Currency	7 (7)	0 (0)	7 (4)	0 (0)	8 (6)	0 (0)	22 (17)	0 (0)
Span	5 (4)	2 (2)	5 (4)	3 (3)	10 (5)	0 (0)	20 (13)	5 (5)
Value	0 (0)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)
Parsing Artifacts (entry count)								
Malformed	- (11)	- (2)	- (3)	- (1)	- (8)	- (4)	- (22)	- (7)
Errors	13 (23)	3 (5)	14 (13)	3 (4)	18 (19)	1 (5)	45 (55)	7 (14)
Error Rate (%)	4.7 (23.0)	1.0 (5.0)	5.0 (13.0)	3.7 (4.0)	7.0 (19.0)	0.4 (5.0)	5.5 (18.3)	0.8 (9.3)

Table 3: **Error analysis of HIFI-KPI.** The table compares error counts from a manual audit of 300 random entries (100 per split) *before* and *after* applying a regex-based cleaning filter. We report both the entity-level metrics and entry-level metrics in parentheses. The final entity-level error rate is less than 1%.

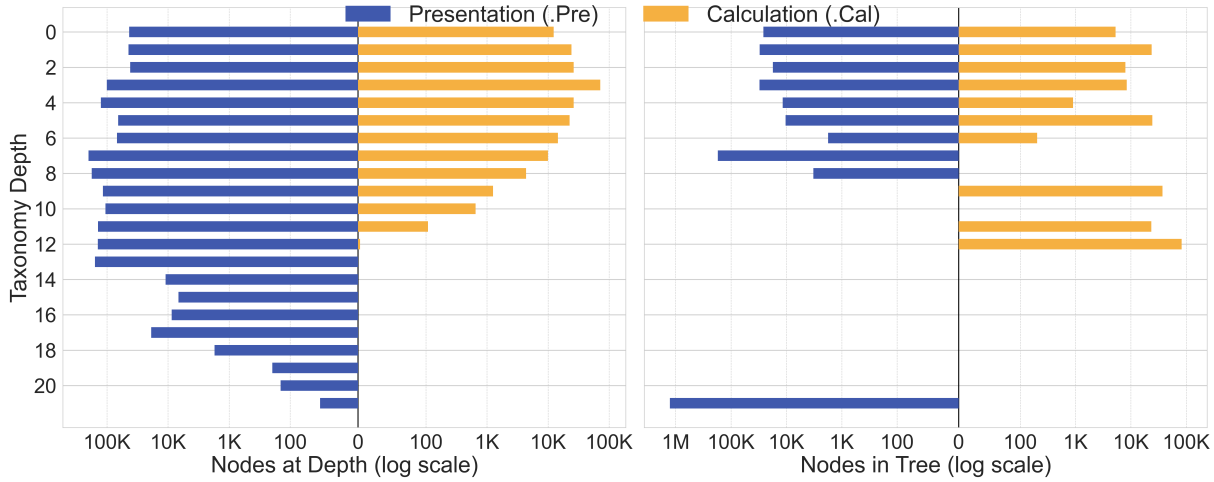

Figure 5: **Taxonomy Statistics** The plot shows how many nodes are at a given depth (left), and the size of a tree at a given depth (right). The Presentation taxonomy is deeper, though both taxonomies distribute nodes without pronounced clustering at any particular level.

omy allows for significantly more depth and nuance in its layers, with a large percentage of nodes belonging to a tree that goes all the way to depth 21, compared to the significantly less deep Calculation taxonomy.

Bottom-up Aggregation Algorithm. We employ a bottom-up hierarchical aggregation method to address the sparsity of our initial $\sim$ 200,000 fine-grained labels. Starting from the leaf nodes, we iteratively map each label to its immediate parent, which systematically reduces specificity while increasing the cardinality of each class. This bottom-up approach is particularly effective as our taxonomy is narrow at the uppermost layers, meaning a top-down approach would change nothing for a lot of labels. The procedure is detailed in Algorithm 1.

Algorithm 1: Taxonomy-Based Grouping via Bottom-Up Selection

Input:

- A hierarchical taxonomy T,
- Number of collapse steps n

Iterative Collapsing.
for $i = 1$ *to* n **do**

1. Let L be the set of all leaf nodes in T.
2. **foreach** *leaf* $l \in L$ **do**
   Given l is not a root
   Let p be the parent of l.
   Replace l with p in T.

return T

Label (HiFi-KPI)	Concept (HiFi-KPI-Lite)
us-gaap:IncomeLossFromContinuingOperations	Earnings
us-gaap:NetIncomeLoss	Earnings
...	
us-gaap:OperatingIncomeLoss	EBIT
cmtl:WeightedAveragePerformanceSharesOutstandingDuringThePeriodThatAreExcludedfromEPSCalculation	EPS
enb:WeightedAverageInterestInOwnCommonShares	EPS
...	
us-gaap:FeeIncome	Revenues
us-gaap:InsuranceCommissionsAndFees	Revenues
..	

Table 4: **Industry Expert Mappings Excerpt.** HiFi-KPI- is made by using mappings selected by an industry expert to generalizable financial Concepts. Two of each category were selected. The full set is available in appendix C.

HiFi-KPI-Lite. We collaborated with a senior domain expert with over a decade of experience, leading a department at a top quantitative finance firm. We investigate the possibilities of a dataset based on manual concept linking. To create HiFi-KPI-Lite, we mapped selected iXBRL terms to their corresponding general finance concepts as shown in Table 6. The financial expert selected four key financial figures for evaluating financial reports: *Earnings*, *EBIT*, *EPS*, and *Revenues*, and then identified relevant iXBRL tags. HiFi-KPI-Lite is created from HiFi-KPI by applying the expert mapping to convert XBRL labels. To make a very curated and relevant subset, we only retain snippets with more than half the entities being part of the expert mappings.

4. Experiments

We demonstrate the usefulness of our context-rich dataset and granularity selection method by establishing three baselines. Text classification, Sequence labeling, and structured information extraction with LLMs. To better describe the performance of our models on the large label set. We report the macro-F1 score as a function of cumulative support, where tags are incrementally included in the calculation, ordered from most to least frequent. (Details about metric calculations in Appendix A.)

Text Classification. We define a simple classification task: Predict the first entity's label from the entry. We embed each entry with `EmbeddingGemma-300M` (Choi et al., 2025) and then fine-tune only the classification head using the Adam optimizer (Kingma and Ba, 2017) with a learning rate of 1×10^{-5} for 20 epochs on the training set. We do this for iterations $n = \{1 - 10\}$ of our bottom up selection Algorithm 1.

Sequence Labeling. The task is to identify and classify each token with a label. To establish strong initial baselines while managing computational resources, the sequence labeling experiments focus on the 1,000 most frequent tags for a given label set. Tags outside this set are mapped to a *OOS* (out-of-scope) label. We use `bert-base-uncased` (Devlin et al., 2019) and `FLANG-BERT` (Shah et al., 2022), a BERT model adapted for the financial domain, with a standard token classification head, we fine-tune full models using the Adam optimizer (Kingma and Ba, 2017), a learning rate of 1×10^{-5} and max. 50 epochs with early stopping patience of 2. We show the utility of the taxonomy, while being mindful of compute resources, by training a model on the $n = 1$ of the `.cal` and `.pre` taxonomies.

LLM-Based Structured Data Extraction. Finally, we use HiFi-KPI-Lite to evaluate LLM-based structured extraction, including tags, dates, currency, and numeric values. We compare four LLMs: Gemma-3-27B (Team et al., 2025), Qwen3-30B-A3B (Qwen Team, 2025) and mistral-Small-3.2-24B (Mistral AI, 2025), DeepSeek-V3.1 (DeepSeek-AI et al., 2024). We use a 1-shot prompt, where we describe how we want to extract the data and give an example of a single extraction(Full prompt avalible in appendix B). We evaluate the models on how well they extract fully identical entities with the gold standard. Further, we evaluate it as a label extraction task, where the goal is to extract the correct labels independent of the contextual information. We do 5 runs of each LLM on the Lite set, except for DeepSeek-V3.1.

5. Results and Analysis

5.1. Quantitative Results and Analysis

Text Classification. Figure 6 shows that more coarse-grained labels of the `.pre` taxonomy generally boosts macro-F1, as the bottom-up approach reduces label sparsity. For `.cal`, performance saturates quickly, possibly because its hierarchy is less deep. The model struggles with infrequent tags, showing opportunities to develop stronger

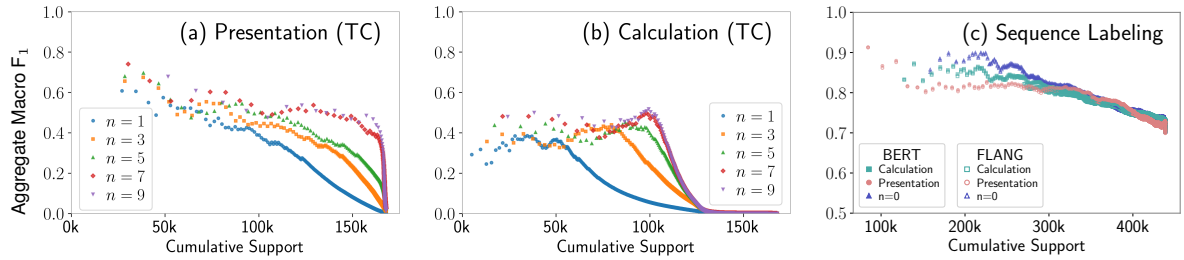

Figure 6: **Results HiFi-KPI.** Plots show aggregate macro-F_1 scores as a function of cumulative label support, on the labelset derived from n iterations of our bottom-up method. Labels are ordered from the most to the least common tag (left to right). Models achieve strong performance on well-represented labels across all levels of granularity, with F_1 scores decreasing for less frequent labels. Subplots correspond to: Presentation taxonomy (left), Calculation taxonomy (middle), and Sequence Labeling task (right).

	Full Entity Extraction				Label Extraction
Model	**F1**	**Precision**	**Recall**	**Jaccard**	**Macro-F1**
BERT (SL)	-	-	-	-	**0.914**
FLANG-BERT (SL)	-	-	-	-	0.907
EmbeddingGemma-300m (TC)	-	-	-	-	0.521
Mistral-small-3.2-24B	$0.422_{\pm 0.008}$	$0.423_{\pm 0.008}$	$0.422_{\pm 0.008}$	$0.268_{\pm 0.006}$	$0.470_{\pm 0.015}$
Gemma-3-27B-it	$0.356_{\pm 0.008}$	$0.358_{\pm 0.008}$	$0.355_{\pm 0.008}$	$0.217_{\pm 0.006}$	$0.543_{\pm 0.030}$
Qwen3-30B-A3B	$\mathbf{0.440}_{\pm 0.008}$	$0.448_{\pm 0.008}$	$0.433_{\pm 0.009}$	$0.282_{\pm 0.007}$	$0.543_{\pm 0.005}$
DeepSeek-V3.1	0.436	0.463	0.413	0.279	0.464

Table 5: **Results HiFi-KPI-Lite**, We report the F1, Precision, Recall, and Jaccard similarity as well as the SD based on 5 runs, except for DeepSeek-V3.1. We report their match with gold entities, as well as their performance on the label extraction task. The encoder-based text classification sequence labelling models can not extract full entities. The table shows the best performance by Qwen3 on entity extraction and BERT (Sequence Labelling) for the label extraction.

methods.

Sequence Labeling. Sequence labeling shows higher performance than text classification. We observe the effectiveness of the `.pre` taxonomy, as the most common label the special (out-of-scope) label has significantly lower support than in the Calculation or the ungrouped. The most common ungrouped tags achieve the highest macro F1, followed by the most common in the `.cal`, and then `.pre`, before all representations converge to the same macro F1 for long-tail labels. We observe that the difference between the FLANG and BERT-based models is close to non-existent. The performance for this task is high, especially considering the 1000 different labels.

LLM-based Extraction. Table 5 shows very consistent performance across models. Overall, performance demonstrates the difficulty of extracting full, well-structured records from financial text. One common theme is failures in normalizing values correctly to match the gold standard. Interestingly, DeepSeek-V3.1 demonstrates a more conservative extraction strategy, achieving the highest precision among all models at the expense of lower recall. opposite of Qwen that seems to have the greatest tendency to predict a label even with low confidence, instead of using the (out-of-scope) token. However, most errors performed by the LLMs seem to be inconsistent and have variety in their form.

5.2. Qualitative Analysis

Text Classification. Inspecting the classification results, the model predictions have granularity errors, for example, the second most common error for both $n = 1$ and $n = 3$ is respectively `us-gaap:AccountsNotesAndLoansReceivableLineItems` predicted as `us-gaap:DebtInstrumentLineItems` and `us-gaap:StatementOfCashFlowsAbstract` as `us-gaap:BusinessAcquisitionLineItems`. Both cases highlight that, likely due to more training samples, the model has a tendency to favor the more common label. This is consistent with Figure 6, which shows that higher support correlates with higher performance. This highlights the potential of future work investigating hierarchical classification methods like Zhou et al. (2020); Agrawal et al. (2025); Jain et al. (2024)

Sequence Labeling. Manual review shows that both FLANG and BERT-based models have close to perfect performance on the task. An interesting pattern is, for example, in the 10-K from First BanCorp on December 31, 2023.

> As of December 31, 2023, the Company's securities portfolio held 657 securities of which 632 securities were in an unrealized loss position.

The long-tailed nature of the data results in neither the calculation nor the $n = 0$ model having the tag `fbnc:DebtSecuritiesAvailableForSaleAndHeldToMaturityNumberOfPositions` because it falls outside the 1000 most common tags, whereas the `.pre` based model uses the more general tag `us-gaap:ScheduleOfTradingSecuritiesAndOtherTradingAssetsLineItems`, correctly classifying with higher abstraction and robustness. The $n = 0$ and `.cal` based model correctly realizes that this tag is outside their dataset's label set. This correct realization shows potential for a future system that conditionally uses the different models. This again highlights the flexibility of the `.pre` compared to the `.cal`. This also shows that HiFi-KPI can be used to successfully train models that use these hierarchies to go from a more specific tag with lower confidence to a more abstract tag with higher confidence and robustness. To answer **RQ1**, our qualitative and quantitative analysis shows how HiFi-KPI can be used to correctly classify a larger portion of the dataset compared to using more specific tags. Further for **RQ1.1**, we find that the `.pre` provides the most flexible representation and best enables training of models for a specific level of granularity. This is likely because the `.pre` is deeper than the `.cal` and has a more varied distribution of node depths, as illustrated in Figure 5.

LLM-based Extraction For the LLM-based extraction on HiFi-KPI-Lite, one theme is that the `Gemma-27B` does not understand "three months ended". (Example in Figure 7.)
`DeepSeek-V3.1` is less prone to extraction, clear from its lower recall but highest of all precision. Most LLM extraction errors are inconsistent and have more variety. One example of errors is that the Qwen model has a greater tendency than other models to guess on a label instead of using the (out-of-scope) label. One other consistent thing is that the models struggle to normalize values correctly to match the gold standard. Lastly, it is important to note the quite strict criteria for this task, which requires it to extract a totally correct entity, especially considering the simple rules HiFi-KPI-Lite is created from, where multiple interpretations can be correct.

Text

"The Reciprocal Exchanges generated $61 million of earned premiums for the three months ended March 31, 2024."

Gold Standard

```
{ 'label': 'revenues',
  'start_date_for_period':
   '2024-01-01',
  'end_date_for_period':
   '2024-03-31',
  'currency_unit': 'USD',
  'value': 61000000.0 }
```

Gemma-27B Prediction

```
{ 'label': 'revenues',
  'start_date_for_period': '
   2024-03-31',
  'end_date_for_period':
   '2024-03-31',
  'currency_unit': 'USD',
  'value': 61000000.0 }
```

Figure 7: Gemma-27B incorrectly predicts start dates, using the same `start_date_for_period` as `end_date_for_period`. All other models predict the right Gold date.

To answer **RQ1.2**, we find that state-of-the-art LLMs show promise on HiFi-KPI-Lite; however, due to the very specific domain and complexity of the task, they struggle with accurately performing the specified task, especially in understanding magnitude classifiers, domain-specific terms like "basis points". Finally, our fine-tuned embedding models, especially for sequence labeling, outperform LLM-based models at label classification, which again highlights the value of domain-specific datasets. HiFi-KPI and our initial models represent a significant step towards the automatic extraction of financial KPIs and iXBRL tagging. For investors, this automation fulfills the original promise of iXBRL by delivering truly structured and analyzable data. This helps prevent major investment mistakes based on flawed premises, such as an incorrect date being used for a tag. Finally, the system equips legal and regulatory bodies with a more efficient tool for verifying compliance.

6. Conclusion

In this paper, we introduce HiFi-KPI consisting of 1.65M paragraphs with 4.5M annotations derived from SEC filings. HiFi-KPI takes a significant step towards building an automated system for KPI tag extraction, by creating a more generalizable financial NLP resource than previously available based

on financial reports. HiFi-KPI has two unified taxonomies `.pre` and `.cal` that structure the iXBRL label set, making it possible to select a specified granularity. Further HiFi-KPI is the first set to provide valuable contextual details for labels, including temporal, currency, and numeric values. We report initial baselines for this new resource with encoder-based classification models as well as LLM-based structured extraction of these new contextual details. Finally, to facilitate rapid prototyping and evaluation, as well as hint at what is possible, we also introduce HiFi-KPI-Lite with expert mappings, showing the potential for even better aggregation algorithms. Our analysis shows that fine-tuned encoder-based models achieve strong performance on the label extraction task, while SOTA LLMs show potential with entity extraction from HiFi-KPI-Lite, further improvement is certainly possible.

Limitations

A limitation of our experiments is that the annotation quality may vary throughout our dataset, evidenced by the fact that the SEC regularly publishes Data Quality Reminders. For instance, the SEC has noted that some filers use different labels for the same element on income statements across periods (U.S. SEC, 2023) or don't report the most fundamental key figure, earnings per share, correctly (U.S. SEC, 2024b). Lastly, there is a bias in the dataset created as the text snippets in the dataset consist only of the spans in these documents, and we only include snippets that match our simple parser methodology.

Ethical Considerations

HiFi-KPI is built from public 10-K and 10-Q reports filed with the U.S. Securities and Exchange Commission (SEC) intended for public disclosure. Our scraping methodology adheres to the guidelines and terms laid out by the SEC.

The dataset is subject to two potential sources of bias. The bias towards bigger companies, as they are more likely to be publicly traded, and required to file with the SEC. Second, a geographical bias towards the US. Therefore, findings may not generalize to companies operating under different international financial reporting standards.

All human labor involved in the creation of the dataset was compensated fairly. The senior domain expert who assisted in the creation of HiFi-KPI-Lite was compensated at his usual hourly rate. The manual data quality checks were also performed by a compensated professional.

Training large-scale models can have a significant environmental impact. To mitigate this, we have limited our experiments to what is necessary for establishing strong baselines. Further, to address this, we create the HiFi-KPI-Lite subset. This smaller, curated resource allows for rapid and efficient model evaluation, significantly lowering the barrier to entry for researchers and reducing the overall environmental impact of working with our data.

Acknowledgments

We would like to thank the AAU-NLP group for helpful discussions and feedback on an earlier version of this article. We want to also thank Alipes ApS for their support in facilitating and funding this research and the useful discussions with their Quant NLP 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).

Bibliographical References

Neeraj Agrawal, Saurabh Kumar, Priyanka Bhatt, and Tanishka Agarwal. 2025. Hierarchical text classification using contrastive learning informed path guided hierarchy. *arXiv preprint arXiv:2506.04381*.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Arelle Development Team. 2025. Arelle: Open source xbrl platform. https://github.com/Arelle/Arelle. Accessed: 2025-01-20.

W. Bricker. 2020. Why xbrl should be on your radar. https://www.financialexecutives.org/FEI-Daily/September-2020/Why-XBRL-Should-Be-on-Your-Radar.aspx.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse,

Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Chelsea Caltuna. 2020. Understanding sec xbrl filings: A primer on sec data.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2022a. Finqa: A dataset of numerical reasoning over financial data.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022b. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv:2210.03849*.

Min Choi, Sahil Dua, and Alice Lisak. 2025. Introducing embeddinggemma: The best-in-class open model for on-device embeddings. Accessed: 2025-09-24.

Yu Cong, Alexander Kogan, and Miklos A. Vasarhelyi. 2007. Extraction of structure and content from the edgar database: A template-based approach. *Journal of Emerging Technologies in Accounting*, 4(1):69–86.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. Deepseek-v3 technical report.

Tobias Deußer, Syed Musharraf Ali, Lars Hillebrand, Desiana Nurchalifah, Basil Jacob, Christian Bauckhage, and Rafet Sifa. 2022. Kpi-edgar: A novel dataset and accompanying metric for relation extraction from financial documents. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1654–1659.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Aaryan Gupta, Vinya Dengre, Hamza Abubakar Kheruwala, and Manan Shah. 2020. Comprehensive review of text-mining applications in finance. *Financial Innovation*, 6:1–25.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Vidit Jain, Mukund Rungta, Yuchen Zhuang, Yue Yu, Zeyu Wang, Mu Gao, Jeffrey Skolnick, and Chao Zhang. 2024. Higen: Hierarchy-aware sequence generation for hierarchical text classification. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2024, page 1354.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A dataset for multilingual financial NLP. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 894–909, Dubrovnik, Croatia. Association for Computational Linguistics.

Bin Ke and Santhosh Ramalingegowda. 2005. Do institutional investors exploit the post-earnings announcement drift? *Journal of Accounting and Economics*, 39(1):25–53.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Viet Dac Lai, Michael Krumdick, Charles Lovering, Varshini Reddy, Craig Schmidt, and Chris Tanner. 2024. Sec-qa: A systematic evaluation corpus for financial qa. *arXiv preprint arXiv:2406.14394*.

Craig Lewis and Steven Young. 2019. Fad or future? automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research*, 49(5):587–615.

Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Edgar-corpus: Billions of tokens make the world go round. *arXiv preprint arXiv:2109.14394*.

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.

Mistral AI. 2025. Mistral small 3.1. https://mistral.ai/news/mistral-small-3-1. Accessed: 2025-09-24.

Open Risk. 2023. Xbrl presentation linkbase. https://www.openriskmanual.org/wiki/XBRL_Presentation_Linkbase. Accessed on 2023-10-26.

Qwen Team. 2025. Qwen3 technical report.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Leonard Richardson. 2007. Beautiful soup documentation. *April*.

SEC. 2018. SEC adopts inline XBRL for tagged data. *U.S. Securities and Exchange Commission*.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Soumya Sharma, Subhendu Khatuya, Manjunath Hegde, Afreen Shaikh, Koustuv Dasgupta, Pawan Goyal, and Niloy Ganguly. 2023. Financial numeric extreme labelling: A dataset and benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3550–3561, Toronto, Canada. Association for Computational Linguistics.

Mahmudul Sheikh and Sumali Conlon. 2012. A rule-based system to extract financial information. *Journal of Computer Information Systems*, 52(4):10–19.

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.

U.S. SEC. 2000. Final Rule: International Disclosure Standards. Accessed: 2024-11-11.

U.S. SEC. 2023. Changing labels for the same reported item on the income statement over multiple periods. Accessed: 2025-02-10.

U.S. SEC. 2024a. Exchange act reporting and registration. Accessed: 2024-11-11.

U.S. SEC. 2024b. Incorrect tagging for earnings per share data. Accessed: 2025-02-10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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*.

XBRL International. 2025a. Presentation. https://specifications.xbrl.org/presentation.html. Accessed: 2025-01-20.

XBRL International. 2025b. Taxonomies. https://www.xbrl.org/the-standard/what/key-concepts-in-xbrl/taxonomies/. Accessed: 2025-01-20.

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

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1106–1117.

7. Language Resource References

A. Elaboration on Metric Calculation, Defining Precision, Recall, Micro F1 and Macro F1

For the HiFi-KPI Lite set, we define precision, recall, micro F1, and macro F1 using an adapted approach, as generative LLM predictions are unrestricted. A misclassified prediction is counted as a false negative for the true label and a false positive for the predicted label. A correct prediction is counted as a true positive. Using these definitions, we compute micro F1 as in any standard classification task. For macro F1, we take the average F1 score of only the ground truth labels, excluding labels that appear solely in the predicted set.

For the Figure 6, we compute the cumulative sum by iterating over the label distribution from the most frequent to the least frequent label in the test set. We then calculate the macro-average F1 score for the top x included labels.

B. System prompt

System Prompt

```
###################
### System Prompt ###
###################

You are an expert data extraction assistant. Your task is to read a given text and extract financial or entity-related information. For each entity found in the text, extract:

- **value**: numerical representation;
- **currency_/_unit**: currency most often USD, shares, EUR, CAD, etc.;
- **label**: revenues, earnings, eps, ebit, or XBRL-OOS (if it is none of the others);
- **start_date_for_period**: (if available)
- **end_date_for_period**: (if available)


If no relevant data is found, return an empty list. Otherwise, return a json line of one or more dictionaries after the "entities" key, each containing these fields exactly:

[
    {
        "entities": [
            {
            "label": "<extracted label>",
            "start_date_for_period": "<YYYY-MM-DD>",
            "end_date_for_period": "<YYYY-MM-DD>",
            "currency_/_unit": "<unit or currency>",
            "value": <numeric value>
            }
        ]
    }
]

No additional commentary or text should be included, only valid JSON.

Example:

Text: The Company has incurred losses since 2008 resulting from a combination of: declining net interest income, as our loan portfolio decreased from $109.8 million at December 31,
      2008 to $62.3 million at December 31, 2016; increased provisions for loan losses between 2009 and 2012; and increasing non-interest expense related to professional fees
      and repossessed asset write-downs and costs. The Company recently incurred net losses of $866 for the nine months ended September 30, 2017 and $1,260 during the year
      ended December 31, 2016. Our interest income for the nine months ended September 30, 2017 has increased with the increase in the balance of our loan portfolio, however,
      this growth has also resulted in an increase to our provision for loan losses. Our non-interest expense has also increased for compensation and occupancy cost and includes
      costs for problem asset resolution at the beginning of the year. The loss for 2016 was largely a result of our net interest income reflecting the low balance of our loan portfolio,
      increasing professional fees for problem asset resolution and additional costs associated with operating as a public company. Non-interest expense for 2016 was also
      impacted by an operational loss not reimbursable from our insurance.

[
    {
        "entities": [
            {
                "label": "XBRL-OOS",
                "start_date_for_period": "2008-12-31",
                "end_date_for_period": "2008-12-31",
                "currency_/_unit": "USD",
                "value": 109800000.0
            },
            {
                "label": "XBRL-OOS",
                "start_date_for_period": "2016-12-31",
                "end_date_for_period": "2016-12-31",
                "currency_/_unit": "USD",
                "value": 62300000.0
            },
            {
                "label": "earnings",
                "start_date_for_period": "2017-01-01",
                "end_date_for_period": "2017-09-30",
                "currency_/_unit": "USD",
                "value": -866000.0
            },
            {
                "label": "earnings",
                "start_date_for_period": "2016-01-01",
                "end_date_for_period": "2016-12-31",
                "currency_/_unit": "USD",
                "value": -1260000.0
            }
        ]
    }
]
```

C. Finance Expert Handpicked Labels and Their Meaning

Label	Category
us-gaap:IncomeLossAttributableToParent	Earnings
us-gaap:IncomeLossFromContinuingOperations	Earnings
us-gaap:IncomeLossFromContinuingOperationsBeforeIncomeTaxesExtraordinaryItemsNoncontrollingInterest	Earnings
us-gaap:IncomeLossFromContinuingOperationsBeforeIncomeTaxesMinorityInterestAndIncomeLossFromEquityMethodInvestments	Earnings
us-gaap:NetIncomeLoss	Earnings
us-gaap:NetIncomeLossAvailableToCommonStockholdersBasic	Earnings
us-gaap:OperatingIncomeLoss	EBIT
bw:IncrementalCommonSharesAttributableToDilutiveEffectOfNetIncome	EPS
cmtl:WeightedAveragePerformanceSharesOutstandingDuringThePeriodThatAreExcludedfromEPSCalculation	EPS
enb:WeightedAverageInterestInOwnCommonShares	EPS
fcx:DilutiveSecuritiesExcludedfromComputationofEPSAmount	EPS
gpmt:AntidilutiveSecuritiesExcludedfromComputationofEarningsPerShareInterestExpense	EPS
gs:ImpactOfUnvestedShareBasedPaymentAwardsAsSeparateClassOfSecuritiesOnEarningsPerShareBasic	EPS
land:WeightedAverageNumberOfOperatingPartnershipUnitsHeldByNoncontrollingInterest	EPS
pcg:PlanOfReorganizationBackstopCommitmentPremiumCommonStockShares	EPS
us-gaap:DistributedEarnings	EPS
us-gaap:DividendsAndInterestPaid	EPS
us-gaap:EarningsPerShareBasic	EPS
us-gaap:EarningsPerShareBasicAndDiluted	EPS
us-gaap:IncrementalCommonSharesAttributableToConversionOfDebtSecurities	EPS
us-gaap:IncrementalCommonSharesAttributableToParticipatingNonvestedSharesWithNonForfeitableDividendRights	EPS
us-gaap:IncrementalCommonSharesAttributableToShareBasedPaymentArrangements	EPS
us-gaap:ParticipatingSecuritiesDistributedAndUndistributedEarningsLossBasic	EPS
us-gaap:UndistributedEarnings	EPS
us-gaap:WeightedAverageNumberOfSharesContingentlyIssuable	EPS
us-gaap:WeightedAverageNumberOfSharesRestrictedStock	EPS
us-gaap:DirectFinancingLeaseRevenue	Revenues
us-gaap:FeeIncome	Revenues
us-gaap:InsuranceCommissionsAndFees	Revenues
us-gaap:OperatingLeaseLeaseIncome	Revenues
us-gaap:PremiumsEarnedNet	Revenues
us-gaap:Revenues	Revenues
us-gaap:UnregulatedOperatingRevenue	Revenues
us-gaap-supplement:FeeIncome	Revenues
us-gaap-supplement:InterestIncomeOperatingPaidInKind	Revenues

Table 6: Mapping of XBRL labels to expert labels.