FIRE: A Dataset for FInancial Relation Extraction

Anonymous ACL submission

Abstract

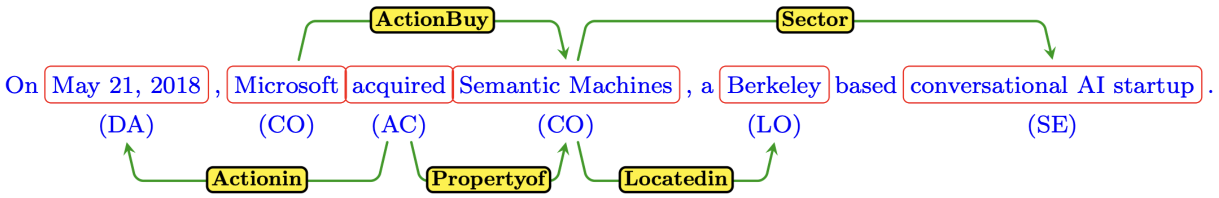
This paper introduces FIRE (**FI**nancial **R**elation **E**xtraction), a sentence-level dataset of named entities and relations within the financial sector. Comprising 3,025 instances, the dataset encapsulates 13 named entity types along with 18 relation types. The textual data was collected from public financial reports and financial news articles, effectively capturing a wide array of financial information about a business including, but not limited to, corporate structure, business model, revenue streams, and market activities such as acquisitions. The full dataset was labeled by a single annotator to minimize labeling noise. Detailed annotation guidelines are provided, as well as an open-source, web-based text labeling tool aimed at streamlining annotation. The labeling time for each sentence was recorded during the labeling process. We show how this feature, along with curriculum learning techniques, can be used to improved a model's performance. The FIRE dataset is designed to serve as a valuable resource for training and evaluating machine learning algorithms in the domain of financial information extraction, as well as a resource for financial analysts to automatically and efficiently extract critical information from financial documents. The dataset and the code to reproduce our experimental results are available at https://github.com/blinded_for_review. The repository for the labeling tool can be found at https://github.com/blinded_for_review.

1 Introduction

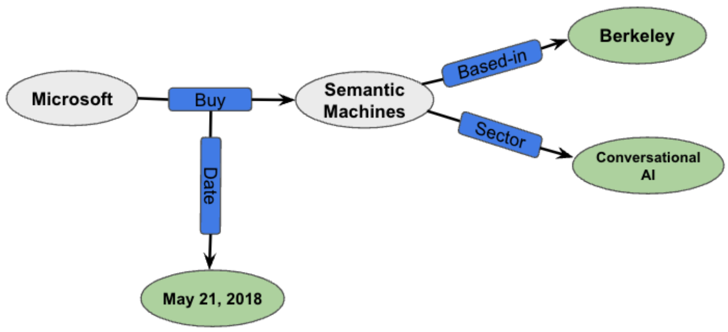
The proliferation of textual data in the financial domain presents a unique opportunity for the application of machine learning and Natural Language Processing (NLP) techniques. The extraction of named entities and their relations from unstructured financial texts, such as Security and Exchange Commission (SEC) filings (U.S. Securities and Exchange Commission) and financial news articles (Bloomberg - Financial news, analysis, and data), is a crucial task with significant implications for financial analysis and decision-making.

Joint Named Entity Recognition (NER) and Relation Extraction (RE) is a complex yet crucial task in NLP, particularly within the financial domain. It involves the identification and classification of named entities in text, and concurrently categorizing the relationships between these entities. This dual task requires a deep understanding of the textual content, demanding high levels of linguistic and domain-specific knowledge. The creation of labeled datasets for such tasks is a costly and time-consuming process. The complexity of financial terminologies and the subtlety of relations often necessitate multiple rounds of review and refinement, contributing to the high cost of dataset creation. This complexity has led to instances where previously hand-labeled and published RE datasets have undergone subsequent corrections post-publication. Examples of such non-financial datasets include TACRED (Zhang et al., 2017b) and its revised counterpart, TACRED Revisited (Alt et al., 2020), as well as DocRED (Yao et al., 2019) and its updated version, Re-DocRED (Tan et al., 2022).

The lack of a comprehensive, well-annotated dataset in the financial domain hampers the development and evaluation of algorithms for these tasks. In response to this identified gap, we present FIRE, a dataset specifically constructed for joint NER and RE within the financial domain. Drawn from both financial documents, mainly SEC filings, and financial news articles, FIRE provides a diverse range of linguistic constructs and financial terminologies. The dataset is constituted of 3,025 instances, all hand-labeled according to comprehensive annotation guidelines. Note that an instance (or an example) refers a labeled object, consisting of a single sentence or multiple sentences with associated entity and relation information. Figure 1a



(a) A sentence and it's labels from the FInancial Relation Extraction (FIRE) dataset. Entity terms are surrounded by a red box, with the entity type abbreviation annotated below the box. An edge between a pair of entities indicates a relation. (DA), (CO), (AC), (LO) and (SE) stand for *Date*, *Company*, *Action*, *Location* and *Sector*, respectively.



(b) An example of constructing a Knowledge Graph (KG) using the labels from the sentence. All sentences in a dataset can be combined to create a KG that summarizes all the collected information.

Figure 1: A labeled sentence from the FIRE dataset and an example of how a Knowledge Graph can be built using the collected labels.

presents a labeled sentence from the dataset while figure 1b is one example of how the labeled data can be used to create a knowledge graph. More examples can be found in the annotation guidelines document which is provided with the dataset. The dataset incorporates 13 named entity categories and 18 relation types, effectively capturing vital details about businesses, including aspects such as their organizational structure, income streams, business strategies, and market maneuvers, including acquisitions.

The dataset also serves as a substantial resource for training, evaluating, and comparing the performance of models specialized in the finance sector. With the recent surge in the development of specialized Large Language Model (LLM)s, there is a critical need for domain-specific datasets to benchmark their performance. An open-source project on github, called 10-KGPT (Smiley, 2023), leverages GPT to analyze and summarize 10-Q and 10-K filings. BloombergGPT (Wu et al., 2023) is a 50 billion parameter LLM specialized for the financial domain. BloombergGPT was assessed across various financial tasks such as sentiment analysis, NER, and Question Answering datasets. However, it did not undergo evaluation on any standalone financial RE dataset. FIRE, with its focus on financial NER and RE, offers a suitable platform for rigorous evaluation. The diversity and complexity of instances in FIRE ensure a comprehensive assessment of these models, taking into account a wide array of financial terminologies and relations. By providing this high-quality, manually annotated dataset, we aim to further the progress in the field of financial NLP, particularly in the application and improvement of state-of-the-art LLMs.

An additional feature of FIRE is the inclusion of a *labeling time* data field for each record in the dataset. This feature may provide researchers with additional granularity when analyzing performance. Labeling time can serve as an implicit indicator of example difficulty, offering potential applications for the implementation of curriculum learning strategies (Bengio et al., 2009). By leveraging this feature, researchers can explore and develop methods that dynamically adjust the learning process based on the difficulty of the examples, potentially leading to more efficient learning and improved model performance. In our experiment results section, we provide an initial result of incorporating the *labeling time* feature into the training process. To the best of our knowledge, this has not been studied yet in the literature.

The paper contributions are summarized as follows:

	FinRED	KPI-EDGAR	**FIRE (This Work)**
Hand-Labeled	✗	✓	✓
No. of Instances	7,775	1,355	3,025
No. of Entity Types	N/A	12	13
No. of Entity Mentions	16,780	4,522	15,334
No. of Relation Types	29	1	18
No. of Relation Mentions	11,121	3,841	8,366

Table 1: Comparison of FinRED, KPI-EDGAR, and FIRE datasets. FIRE has the advantage over FinRED in that it is hand-annotated and over KPI-EDGAR in that it is larger, has diverse relations and is more comprehensive in terms of covering financial aspects over a business. Note that FinRED statistics for entity and relation mentions were not readily available. The figures included below were manually computed after a review of the FinRED data files.

- We introduce FIRE, a novel dataset for joint NER and RE within the financial context. FIRE is accompanied by comprehensive annotation guidelines and is hand-annotated by a single annotator to minimize labeling noise.
- We provide an open-source web-based labeling tool, designed to facilitate efficient and precise annotation for NER and RE tasks.
- We demonstrate that utilizing the labeling time of each example can enhance model performance through curriculum learning strategies

The rest of this paper is organized as follows: Section 2 goes over some previous general-purpose and domain-specific NER and RE datasets and compares FIRE to existing datasets in finance. Section 3 provides a detailed description of the FIRE dataset, including the composition, data collection and annotation processes. Section 4 presents an evaluation of selected state-of-the-art models on the FIRE dataset, discussing the associated performances and implications. Finally, section 5 concludes the paper and outlines potential directions for future work.

2 Related Work

Sentence vs. Document Level RE: Sentence-level RE identifies relationships between entities in a single sentence, while document-level RE captures relationships across multiple sentences or entire documents. Document-level RE offers a broader understanding of entity relationships, but sentence-level RE can pinpoint specific relationships more quickly. Document-level datasets include BC5CDR (Li et al., 2016), DWIE (Zaporojets et al., 2020), DocRED (Yao et al., 2019), and Re-DocRED (Tan et al., 2022). Some popular sentence-level RE-datasets include TACRED (Zhang et al., 2017b), FB-NTY (Hoffmann et al., 2011), and WebNLG (Gardent et al., 2017). While many of these are general-purpose, there are domain-specific datasets too (Luan et al., 2018; Perera et al., 2020). FIRE, despite having some multi-sentence instances, is mainly a sentence-level RE dataset.

Relation Extraction Datasets and Distant Supervision. Creating RE datasets is costly due to labeling. One common technique to deal with this problem is distant supervision which relies on a knowledge base to automatically label text data (Mintz et al., 2009). In particular, sentences that mention two entities connected by a relation in the knowledge base are assumed to be expressing that same relation. This strong assumption leads to a large number of noisy samples. To address this issue, researchers have developed methods that relax the distant supervision assumptions(Riedel et al., 2010; Bengio et al., 2009). Despite its limitations, distant supervision remains a popular and effective method for generating large-scale datasets for relation extraction tasks. Several relation extraction datasets have been developed using distant supervision, including FB-NYT (Hoffmann et al., 2011), a dataset constructed by aligning Freebase (Bollacker et al., 2008) relations with The New York Times articles, and WebNLG (Gardent et al., 2017), a text generation dataset created from DBPedia (Bizer et al., 2009), among others. Such datasets have been widely used for training and evaluating relation extraction models. Conversely, FIRE is a supervised dataset in which every instance has been annotated manually following extensive annotation guidelines. While this approach elevates the cost of labeling and poses scalability challenges, it guarantees a high level of precision in the labels.

Financial Relation Extraction. Several NER

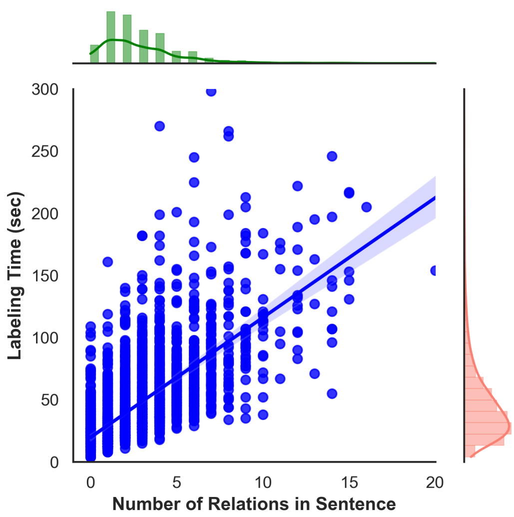


Figure 2: Scatter plot of labeling time (in seconds) versus the number of relations in the sentence. The marginal distributions and histograms are displayed at the edges of the plot. For sentences with the same number of relations, there is a wide distribution of labeling times, showing how the two quantities are correlated but still provide different information.

and/or RE datasets in the financial domain have been previously proposed. FiNER-ORD (Shah et al., 2023) is an NER dataset automatically collected by applying pattern-matching heuristics on financial news articles. Unlike FIRE, this is an NER-only dataset with only three entity types. Another related work is (Wu et al., 2020), which established a Chinese corpus for relation extraction from financial news. However, this work focuses on relation extraction in the Chinese language, while our dataset targets relation extraction in the English language. Two datasets that most closely resemble ours are FinRED, an RE dataset introduced in (Sharma et al., 2022), and KPI-EDGAR, a joint NER and RE dataset introduced in (Deußer et al., 2022). Both are specialized in the financial domain. FinRED contains 7,775 instances covering 29 relation types and was collected from earning call transcripts and financial news articles. However, FinRED was labeled using the distant supervision technique, which can lead to a large number of noisy samples as outlined previously. In contrast, all instances in FIRE were hand-annotated by a human annotator. Similar to FIRE, the KPI-EDGAR dataset is also hand-annotated. The focus of this dataset is on extracting Key Performance Indicators (KPIs) from financial documents and link them to their numerical values and other attributes. It supports 12 entity types but in terms of relations, a binary link either exists between two entities or not. In contrast, FIRE supports an extensively diverse set of relations and its entities extend to broader business aspects, not being exclusively centered on KPIs. Table 1 compares the statistics of FIRE with both FinRED and KPI-EDGAR.

Labeling Time and Curriculum Learning. In FIRE, we've included a 'labeling time' attribute for each instance. This data, representing the time it took the annotator to label that particular instance from the dataset, was gathered during the annotation stage without additional cost. This could be useful to researchers examining annotation complexities or considering strategies like curriculum learning - a method inspired by progressive human learning, where models are exposed to easier samples first, gradually moving onto complex ones (Bengio et al., 2009). This method has been extensively applied in a variety of machine learning tasks (Zhang et al., 2017a; Kocmi and Bojar, 2017; Narvekar et al., 2020). A difficulty metric is required to apply curriculum learning. For example, a simple static (known a priori) difficulty metric for textual data can be the length of sentence in tokens. More sophisticated metrics are data-driven and adjust based on model feedback (Ma et al., 2017; Kumar et al., 2010). In this context, we suggest that 'labeling time' may act as an proxy for the difficulty of an example. As illustrated in Figure 2 we observe a positive correlation between the labeling time of a sentence and the number of relations it contains. Despite this correlation, the labeling time can vary significantly for a fixed number of relations, indicating that it is not a redundant feature. Qualitatively similar results are observed when comparing labeling time to sentence length or number of entities in a sentence. In section 4, we provide an initial result of how incorporating the labeling time feature into the training process can improve the performance of trained models.

3 FIRE Dataset

3.1 License and Intended Use

License. The dataset and its associated resources are provided under the Creative Commons Attribution 4.0 International License (CC 4.0) (Creative Commons, 2023).

The labeling tool developed in conjunction with the dataset is licensed under the MIT open-source license, see the LICENSE file for details.

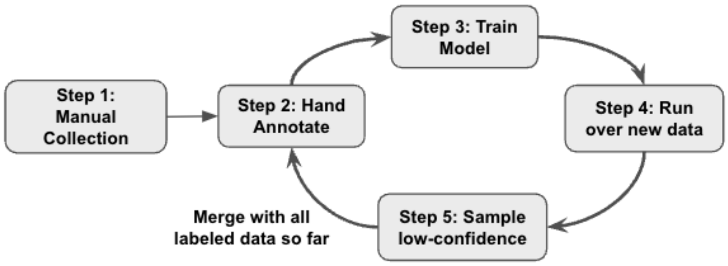


Figure 3: Stages of data collection: 1) Manually gather relevant sentences. 2) Hand-label them to create a "seed" dataset. 3) Train an RE-specific model on this dataset. 4) Use the model on new financial content to identify entities and relations. 5) From the model's output, select sentences with low-confidence predictions to reduce confirmation bias. Remove existing labels from these sentences, manually annotate them, and merge with prior data. Repeat until the desired dataset size is achieved.

Intended Use. The intended use of the FIRE dataset is two-fold: First, to advance the research in the area of joint NER and RE, specifically within the financial domain. It is designed to serve as a benchmark for evaluating the performance of existing models, as well as a training resource for the development of new models. Second, the FIRE dataset can serve as a valuable resource for financial analysts and auditors, enabling them to harness automated algorithms for expedient and efficient extraction of critical information from financial documents.

3.2 Data Splits and Statistics

In Table 1, some basic statistics of the FIRE dataset are displayed. The different entity and relation types as well as their distribution in the dataset can be found in appendix A.

The dataset was initially partitioned randomly into training, development (validation), and testing sets following a 70%, 15%, 15% split, respectively. Because financial reports, by their nature, often exhibit repetitive patterns in their language and structure, extra care was taken in creating the test set. Specifically, the Jaccard similarity score was computed for each pair of sentences from train and test sets. Jaccard similarity is defined as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, where A and B are sets of tokens in two instances. It measures the degree of similarity between two sets. Any sentence in the test set exhibiting a Jaccard similarity score above 50% with any sentence in the training set was replaced by a different sentence from the train set. This approach helps to reduce data leakage and ensures that the test set provides a robust and unbiased evaluation of model performance.

3.3 Data Collection and Annotation

Data Sources and Pre-Processing. Approximately 25% of the dataset's records were sourced from publicly accessible financial news articles (Bloomberg - Financial news, analysis, and data; Yahoo Finance, 2023; CNBC, 2023; The Economic Times, 2023; The Financial Express, 2023), while the remaining 75% were extracted from publicly available SEC filings such as 10-K and 10-Q financial reports. For the SEC filings, we used the dataset of *Cleaned and Raw 10-X Files* spanning the years 1993-2021 (McDonald, 2023). This dataset contains all 10-K variants, e.g., 10-Q, 10-K/A, 10-K405. Every report in this dataset has already been cleaned and parsed to remove all non-textual related objects. For the financial news pieces, we obtained the original articles directly from their respective sources and manually conducted the cleaning process to extract the raw text.

Data Collection and Labeling. The process, shown in Figure 3, began with selecting a subset of financial reports and articles. An annotator identified and labeled key sentences with relevant entities and relations, creating a "seed" dataset. This dataset trained a joint NER and RE model (refer to 4.1), which then scanned new documents to suggest potential sentences. However, only the sentence selection was automated; actual labeling was always done manually. To mitigate confirmation bias, selections were deliberately made from low-confidence predictions generated by the model. Also, to reduce bias, the annotator was not shown the model's predictions. This cycle continued until we achieved the desired dataset size, with all annotations done by a single non-domain expert human annotator.

Annotation Guidelines. For the FIRE dataset,

Annotator Pair	Entity F1 (%)	Relation F1 (%)
Main Annotator and A	78.29	59.72
Main Annotator and B	70.57	49.19
Main Annotator and C	50.46	16.05
A **and** B	69.73	48.46
A **and** C	46.72	14.19
B **and** C	49.52	17.49

Table 2: Inter-annotator micro F1 scores. Annotators A and B are engineers familiar with the NER/RE task. Annotator C had no prior familiarity with the NER/RE task nor any expertise in engineering, finance, or linguistics.

a comprehensive set of labeling rules was established, incorporating both general entity and term annotation guidelines based on the ACL RD-TEC guidelines (QasemiZadeh and Schumann, 2016), as well as domain-specific rules tailored to each entity and relation present in the dataset. The guidelines also provide guidance for resolving ambiguous or conflicting edge cases.

Inter-Annotator Agreement. To assess difficulty of the annotation task, a subset of 150 samples was randomly selected and provided to three independent annotators. Annotators A and B were engineers with familiarity with the NER/RE task and annotator C was a professor with expertise outside of finance, engineering, and linguistics. Annotator A underwent several iterations of training to improve the quality of their annotations. In contrast, Annotators B and C were instructed to familiarize themselves with the annotation guidelines for 1-2 hours before starting the labeling task, without any prior training. The agreement between the annotators, including the main annotator of the dataset, was measured using the pair-wise entity and relations micro F1 score, as detailed in Table 2. This score was computed by treating one set of annotations as the ground truth labels and the other as predictions. Note that the result is the same regardless of which annotations were designated as ground truth. Although Cohen's Kappa is usually the preferred metric for inter-annotator agreement, it is not suitable for the NER/RE task (Deléger et al., 2012; Hripcsak and Rothschild, 2005). The highest agreement was found with the annotator who received additional training. There was also greater agreement between the main annotator and annotator B as compared to annotator C, likely due to the annotator's technical background and familiarity with the NER/RE task. These results suggest that the task has a high level of technical complexity and that, even with the detailed annotation guidelines, training of new annotators requires an iterative education process. Furthermore, even with some iteration in annotator training, as was the case for annotator A, the inter-annotator agreement indicates significant room for improvement. For this reason, the entire FIRE dataset is labeled by a single annotator who wrote the annotation guidelines and invested significant time and effort to ensure consistency. The consistent labeling of the FIRE dataset is confirmed by the results in section 4.3, where the F1 scores for trained models are much higher than the figures in Table 2.

3.4 Labeling Tool

We introduce an open-source, web-based text annotation tool alongside the FIRE dataset (blind Review, 2023). Tailored for entity and relation labeling, the tool offers features for efficient annotation and error minimization. It supports shortcuts for quick labeling and an optional *rules file* upload to set constraints on permissible relations between entity types, inspired by the work of (Lyu and Chen, 2021). For example, in FIRE, a rule might dictate that the *ActionSell* relation is exclusive to the *Company* entity type. This ensures accurate annotations by preventing incompatible entity-relation combinations. The tool also logs the annotation time for each instance, as detailed in section 2.

4 Experimental Results

4.1 Models

To benchmark the performance of state-of-the-art models on FIRE, two family of models were selected for evaluation: RE-specific LLMs and general-purpose generative (causal) LLMs. RE-specific LLMs are LLMs that were designed specifically to solve the NER and RE task, by modifying the architecture of the base model as well as the training procedure. On the other hand, general-

Algorithm 1: A Simple Curriculum Learning Algorithm

Data: Dataset D, Difficulty metric M, Number of tiers N, Number of fine-tuning epochs E

Result: Trained Model Θ

1 Divide D into N tiers $(T_1, T_2, \ldots, T_N)$ in increasing order of difficulty based on metric M;
2 $D_{\text{current}} = \emptyset$;
3 **for** $i = 1$ *to* N **do**
4     $D_{\text{current}} = D_{\text{current}} \cup T_i$;
5   Train on D_{current} for one epoch;
6 Fine-tune on entire dataset D for E epochs;
7 **return** *Trained Model* Θ

purpose causal LLMs are designed with the language modeling objective and have no direct connection to the RE task. Because of this difference, causal LLMs require a larger model size to compete with the custom designed ones.

Two prominent joint NER and RE models were selected for fine-tuning to evaluate the performance on the FIRE dataset: SpERT (Eberts and Ulges, 2020) and PL-Marker (Ye et al., 2021). **SpERT** effectively applies the Transformer architecture, complemented by a robust negative sampling strategy. It thus serves as a good starting point for evaluation. **PL-Marker** employs a unique marker mechanism to mark entity boundaries in sentences. This approach enables the model to capture entity-relation structures more efficiently. Both models are built upon the BERT (Devlin et al., 2019) framework.

For general purpose generative models, we opted for **Llama 2** (Touvron et al., 2023) and **GPT-3.5** (Brown et al., 2020), evaluating them in both few-shot and fine-tuned settings.

Together, these models provide a reasonably comprehensive assessment of the FIRE dataset's performance and potential.

4.2 Setup and Evaluation

Standard Fine-Tuning SpERT and PL-Marker were allotted 24 hours on an NVIDIA GEFORCE RTX 2080 Ti GPU for hyper-parameter tuning, with results detailed in appendix B. Llama 2 and GPT-3.5 were fine-tuned using a specific data format (appendix C), without hyper-parameter tuning due to computational constraints.

Few-Shot Prompting For Llama 2 and GPT 3.5, a custom designed prompt was designed to evaluate the models in a few-shot setting. The prompt includes a definition and description of each relation type. The models are then prompted to extract relations only, i.e. no entity evaluation is performed. Prompt details are in Appendix C.

Curriculum Learning In addition to the standard training setup, another experiment was performed by training both SpERT and PL-Marker models according to a curriculum determined by the labeling time information. A very simple curriculum learning algorithm is used as described in algorithm 1. The training set is first divided into N tiers in increasing order of difficulty according to a metric M. Then, the model is trained successively for one epoch on each tier, as well as all previous tiers. Finally, the model is fine-tuned on the entire dataset for number of epochs E. In our experiment, we set $N = 10$ and $E = 20$ for both models. A compute budget of 24 hours is again given for both models to search for the best learning rate and batch size.

The difficulty metric M was computed as follows: given a sentence's labeling time t, we consider features such as the number of entities n_{ent}, the number of relations n_{rel}, and boolean variables indicating the length of the sentence as either short or medium, with large sentences encoded by setting both short and medium variables to zero. Using these features, we fit a simple linear regression model to predict t as:

$$\hat{t} = \beta_0 + \beta_1 \cdot n_{ent} + \beta_2 \cdot n_{rel} \quad (1)$$
$$+ \beta_3 \cdot short + \beta_4 \cdot medium \quad (2)$$

The difficulty metric M is then defined as the normalized residual of the actual and predicted labeling time:

$$M = \frac{t - \hat{t}}{\max(t) - \min(t)} \quad (3)$$

This metric gives us a sense of how much harder (or easier) a sentence is to label compared to what we'd expect based solely on its features. The reason M is not chosen to be the labeling time t is because a sentence with large t is not always "more difficult" to label than a sentence with small t. The difference could be due to the features discussed above, e.g. the sentence with large t could simply contain more entities but is actually easier to label. This is why proper normalization is required to choose M.

Model	Evaluation	Entity F1 (%)	Relation F1 (%)
SpERT	*Standard Fine-Tuning*	$84.63^{\pm 0.25}$	$67.41^{\pm 0.92}$
	Curriculum Learning	$\mathbf{85.39}^{\pm 0.33}$	$\mathbf{68.11}^{\pm 0.53}$
PL-Marker	*Standard Fine-Tuning*	$83.78^{\pm 0.18}$	$67.01^{\pm 0.67}$
	Curriculum Learning	$84.65^{\pm 0.54}$	$67.67^{\pm 0.82}$

Table 3: Performance of the three models on the FIRE test data. Mean and standard deviation (in superscript) are reported for micro F1 score for both entities and relations.

Evaluation. For the fine-tuning and curriculum learning experiments of both SpERT and PL-Marker, three independent training runs were performed. The mean and standard deviation of the micro F1 score are reported. Llama 2 and GPT 3.5 are evaluated on a single run only. The exact match micro F1 score was used as the evaluation metric for relations, i.e. entity boundaries, entity types, as well as the relation label must exactly match the ground truth labels to be considered correct. We use the train/eval/test splits for FIRE as reported in section 3.2. All experiments were ran on a single NVIDIA GEFORCE RTX 2080 Ti GPU.

4.3 Results

Table 3 presents the performance of the RE-specific models. SpERT and PL-Marker show similar results. In comparison to other NER and RE datasets, SpERT's top F1 score on the ConLL04 dataset (Roth and tau Yih, 2004), a general-purpose dataset with fewer entity and relation types, is comparable to its performance on FIRE. This indicates consistent annotations in the FIRE dataset, further supported by the models outperforming the inter-annotator agreement scores in Table 2.

Curriculum learning enhanced the performance of both SpERT and PL-Marker compared to standard training. This confirms our assumption that the labeling time is an informative feature that can be used to improve the generalization capabilities of the models. Note that while we employed a very simple curriculum learning algorithm, more advanced and sophisticated techniques have been proposed in the literature that can potentially achieve even higher improvements. Nevertheless, our primary contribution focuses on the dataset, and a thorough evaluation of all curriculum learning techniques can be explored in future research.

Table 4 showcases the results for general-purpose generative LLMs. Fine-tuning outperforms few-shot learning significantly. GPT-3.5 surpasses Llama 2, especially when fine-tuned.

Model	Evaluation	Relation F1 (%)
Llama 2	*Few-Shot*	11.10
	Fine-Tuned	34.63
GPT 3.5	*Few-Shot*	15.95
	Fine-Tuned	**54.50**

Table 4: Performance of GPT 3.5 and Llama 2 models on the FIRE test data. Result is over one run only.

However, these models still lag behind RE-specific models. Our findings are consistent with a recent study (Han et al., 2023) that also identified a significant performance gap between ChatGPT (OpenAI, 2023) and state-of-the-art methods, particularly in more complex tasks. This can be explained by multiple factors, mainly the difficulty in doing strict evaluation of generative models which lack a fixed output format. This underscores the need for further research on using untrained causal LLMs for relation extraction, especially on datasets with diverse entity and relation types.

5 Conclusion

In this paper, we introduced FIRE, a dataset carefully curated for the task of joint named entity and relation extraction in the financial domain. The comprehensive annotation guidelines and the open-source labeling tool accompanying the dataset further contribute to its robustness and usability. Our evaluations with RE-specific and generative LLMs highlight FIRE's challenges and potential. We also explored the benefits of incorporating labeling time in training. It is evident that the development of more refined models capable of understanding the complexities of financial domain-specific data is required. Looking forward, we anticipate that FIRE will serve as a valuable resource for researchers and practitioners in the fields of natural language processing, financial analysis, and auditing.

6 Limitations

The primary limitation of this dataset lies in its domain-specific nature; the dataset is curated specifically for the financial domain, which may limit its applicability and generalization to other fields. Additionally, the dataset is sourced solely from English language documents, which restricts its utility in multi-lingual or cross-lingual studies. Furthermore, the dataset is thoroughly annotated by a single human who is not a finance domain expert nor a linguist. Thus, the inherent subjectivity and possible biases or lack of domain-knowledge in manual annotation cannot be completely ruled out. Finally, the dataset is not meant to be an all-encompassing solution. Due to the complex and nuanced language often used in financial reports and news articles, certain entities and relations may not be captured by the existing entity and relation categories in the dataset. Finally, all entities in FIRE are extracted verbatim from the text. If an entity is implied but not explicitly stated, it would not be captured in FIRE as well as any relation relating to it. Future iterations of FIRE would benefit from addressing these limitations, expanding both its domain knowledge and linguistic diversity.

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Tacred revisited: A thorough evaluation of the tacred relation extraction task. In *Proceedings of ACL*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *International Conference on Machine Learning*.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165. The Web of Data.

Blinded blind Review. 2023. Blinded for review.

Bloomberg - Financial news, analysis, and data. 2023. Bloomberg - Financial news, analysis, and data. https://www.bloomberg.com/.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

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, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff 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. *ArXiv*, abs/2005.14165.

CNBC. 2023. https://www.cnbc.com.

Creative Commons. 2023. Creative commons attribution 4.0 international license. https://creativecommons.org/licenses/by/4.0/. Accessed: 2023.

Louise Deléger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnár, Laura Stoutenborough, Michal Kouril, Keith A. Marsolo, and Imre Solti. 2012. Building gold standard corpora for medical natural language processing tasks. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2012:144–53.

Tobias Deußer, Syed Musharraf Ali, Lars Patrick 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. *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. *ArXiv*, abs/1810.04805.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *Proceedings of the 24th European Conference on Artificial Intelligence*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *ArXiv*, abs/2305.14450.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Annual Meeting of the Association for Computational Linguistics*.

George Hripcsak and Adam S. Rothschild. 2005. Technical brief: Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 12 3:296–8.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 379–386, Varna, Bulgaria. INCOMA Ltd.

M. Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016. Baw068.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreferencefor scientific knowledge graph construction. In *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*.

Shengfei Lyu and Huanhuan Chen. 2021. Relation classification with entity type restriction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 390–395, Online. Association for Computational Linguistics.

Fan Ma, Deyu Meng, Qi Xie, Zina Li, and Xuanyi Dong. 2017. Self-paced co-training. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2275–2284. PMLR.

Bill McDonald. 2023. Cleaned and raw 10-x files, software repository for accounting and finance. University of Notre Dame, Mendoza College of Business.

Mike D. Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Annual Meeting of the Association for Computational Linguistics*.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *J. Mach. Learn. Res.*, 21(1).

OpenAI. 2023. Introducing chatgpt. *OpenAI Blog*.

N. Perera, M. Dehmer, and F. Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in Cell and Developmental Biology*, 8:673.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1862–1868, Portorož, Slovenia. European Language Resources Association (ELRA).

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *ECML/PKDD*.

Dan Roth and Wen tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Conference on Computational Natural Language Learning*.

Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and Sudheer Chava. 2023. Finer: Financial named entity recognition dataset and weak-supervision model. *ArXiv*, abs/2302.11157.

Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. 2022. Finred: A dataset for relation extraction in financial domain. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 595–597, New York, NY, USA. Association for Computing Machinery.

Dorian Smiley. 2023. 10-kgpt: Analyze 10-q and 10-k fillings with gpt. `https://github.com/doriansmiley/10-kGPT`.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting docred – addressing the false negative problem in relation extraction. In *Proceedings of EMNLP*.

The Economic Times. 2023. `https://economictimes.indiatimes.com`.

The Financial Express. 2023. `https://www.financialexpress.com`.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and

Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

U.S. Securities and Exchange Commission. 2023. U.S. Securities and Exchange Commission (SEC). `https://www.sec.gov/`.

Haoyu Wu, Qing Lei, Xinyue Zhang, and Zhengqian Luo. 2020. Creating a large-scale financial news corpus for relation extraction. *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 259–263.

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

Yahoo Finance. 2023. `https://finance.yahoo.com`.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2021. Packed levitated marker for entity and relation extraction. In *Annual Meeting of the Association for Computational Linguistics*.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2020. Dwie: an entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.*, 58:102563.

Dingwen Zhang, Le Yang, Deyu Meng, Dong Xu, and Junwei Han. 2017a. Spftn: A self-paced fine-tuning network for segmenting objects in weakly labelled videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5340–5348.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017b. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

A Distribution of entity and relation types in FIRE

Table 5 breaks down the quantity of each entity type in the dataset while Table 6 displays the same information but for relations. For a detailed description of each entity and relation type, see the annotation guidelines document accompanying the dataset.

Number of Entity Mentions		**15,334**
Average number of entities per instance		**5.29**
Amount of each entity	Company	22.41%
	FinancialEntity	15.60%
	Date	15.37%
	Designation	8.08%
	Money	7.78%
	Action	5.57%
	Quantity	5.27%
	Product	4.39%
	Sector	3.90%
	Location	3.74%
	Person	3.41%
	BusinessUnit	2.71%
	GeopoliticalEntity	1.70%

Table 5: FIRE Dataset Entity Statistics

B Hyper-parameter Selection

For our experiments, we allocated a tuning budget of 24 hours on an NVIDIA GEFORCE RTX 2080 Ti GPU for each model to search for the optimal hyper-parameters on the validation set.

Table 7 displays the selected hyper-parameters for SpERT and PL-Marker in the standard fine-tuning experiments.

Table 8 presents the hyper-parameters for the curriculum learning experiments for SpERT and PL-Marker. To reduce the search space, instead of searching for one learning rate for each data tier, we select a fixed learning rate for tiers 1 to 3, 4 to 6 and 7 to 9. Thus we search for only three learning rates for all tiers, in addition to the final learning rate for the whole dataset.

C Llama 2 and GPT 3.5 Prompts

C.0.1 Few-Shot Learning Prompts

For few-shot learning, the following 1-shot prompt was used:

> *Find the relation between the entities given in the context and produce a list of triplets containing two entities and their relations.*
>
> *Only find out the following relations ActionBuy, Actionin, ActionSell, ActionMerge, Actionto, Constituentof , Designation, Employeeof, Locatedin, Productof, Propertyof, Quantity, Sector,*

Number of Relation Mentions		8,366
Average number of relations per instance		2.92
Amount of each relation	Valuein	11.17%
	Value	9.98%
	Designation	9.95%
	Actionto	8.55%
	Actionin	6.35%
	Propertyof	6.33%
	Locatedin	6.06%
	Sector	5.76%
	Productof	5.71%
	Constituentof	5.27%
	Employeeof	4.67%
	ValueChangeIncreaseby	4.31%
	ActionBuy	3.87%
	ValueChangeDecreaseby	3.64%
	Subsidiaryof	3.16%
	Quantity	3.08%
	ActionSell	1.66%
	ActionMerge	0.40%

Table 6: FIRE Dataset Relation Statistics

Subsidiaryof, Value, ValueChangeDecreaseby, ValueChangeIncreaseby and Valuein

ActionMerge indicate two company or organizations enters into merger agreements to form a single entity.

ActionBuy represents the action of purchasing/acquiring a Company, FinancialEntity, Product, or BusinessUnit by a Company or a Person.

Actionto represents the relation between the action entity and the entity on which the action has taken.

Constituentof relation denotes one financial entity is part of another financial entity.

Actionin indicates the Date associated with an Action entity, signifying the time of occurrence of the action.

ActionSell represents the action of selling a Company, FinancialEntity, Product, or BusinessUnit by a Company or a Person.

Employeeof denotes the past, present or future employment relationship between a Person and a Company.

Designation indicates the job title or position of a Person, or the Designation of a Company in the financial context, providing information about the role or responsibility of the entity.

Locatedin indicates the geographical location or country associated with an entity, specifying the place or region where the entity is located. Money and Quantity can be in the place where they were generated, lost, profited, etc. Note that a Company is only Located in a place if it based in that place.

Productof indicates a Product is manufactured, sold, offered, or marketed by a Company, establishing a relationship between the Company and the Product.

Propertyof serves as an umbrella relation" that indicates a general association between two entities,mainly representing ownership or part-of/composition relationships. This relation is used to connecttwo entities when a more specific relation is not yet defined.

Quantity represents the countable quantity a FinancialEntity, BusinessUnit or Product.

Sector indicates the economic sector or industry to which a Company belongs, providing information about the broad business area or category of the Company's operations.

Subsidiaryof indicates that a Company is a subsidiary of a parent Company, either wholly or majority owned. Note that "brands" are always considered subsidiaries of their parent Company. A highly occurring pattern is a parent company selling its subsidiary company, in which case the Subsidiaryof relation is not annotated.

Value represents a non-countable value of a FinancialEntity, BusinessUnit or Product such as a monetary value or a percentage. A Company can also have a Value relation, but only for monetary values such as indicating the net worth of a company or the sale price in an acquisition.

Model	Learning Rate (NER)	Batch Size (NER)	Learning Rate (RE)	Batch Size (RE)
SpERT	—	—	5e-5	2
PL-Marker	7e-5	2	4e-6	2

Table 7: Selected hyper-parameters for standard fine-tuning. Note that PL-Marker has a separate training run for its NER module. Therefore, we search for the learning rate and batch size of this module as well.

Model	Learning Rate				Batch Size
	Tier 1-3	Tier 4-6	Tier 7-9	Final	
SpERT	8e-6	5e-5	3e-5	5e-5	8
PL-Marker	7e-6	4e-5	4e-5	1e-6	4

Table 8: Hyper-parameters for curriculum learning experiments. Note that for PL-Marker, we apply curriculum learning on the RE module only. For the NER module, we fix the learning rate to $5e-5$ and the batch size to 4.

ValueChangeDecreaseby indicates the decrease in monetary value or quantity of a FinancialEntity. An additional more rare use-case is the Quantity of a BusinessUnit decreasing, such as number of employees or number of offices.

ValueChangeIncreaseby indicates the increase in value or quantity of a FinancialEntity. An additional more rare use-case is the Quantity of a BusinessUnit increasing, such as number of employees or number of offices.

Valuein indicates the Date associated with a Money or Quantity entity, providing information about the specific time period to which the Money or Quantity value is related.

Please find few examples below

Context : Bank of America to Buy Merrill Lynch for $50 Billion

Answer : [['Bank of America', 'Merrill Lynch', 'ActionBuy'], ['Buy', 'Merrill Lynch', 'Actionto'], ['Merrill Lynch', '$50 Billion', 'Value']]

C.1 Fine-Tuning Prompts

For fine-tuning, the dataset examples were transformed to the following prompt which was used to train the models:

Question: Find the relation between the entities given in the context and produce a list of triplets containing two entities and their relations. Only find out the following relations: ActionBuy, Actionin, ActionSell, ActionMerge, Actionto, Constituentof, Designation, Employeeof, Locatedin, Productof, Propertyof, Quantity, Sector, Subsidiaryof, Value, ValueChangeDecreaseby, ValueChangeIncreaseby, and Valuein.

Context: Bank of America to Buy Merrill Lynch for $50 Billion

Answer: [['Bank of America', 'Merrill Lynch', 'ActionBuy'], ['Buy', 'Merrill Lynch', 'Actionto'], ['Merrill Lynch', '$50 Billion', 'Value']]