EFSA: Towards Event-Level Financial Sentiment Analysis

Tianyu Chen[1,2,3*], **Yiming Zhang**[4*], **Guoxin Yu**[1,2,3], **Dapeng Zhang**[5], **Li Zeng**[6†], **Qing He**[1,2,3,4], **Xiang Ao**[1,2,3†]
[1]Key Laboratory of AI Safety, Chinese Academy of Sciences (CAS), Beijing, China.
[2]Key Lab of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, China.
[3]University of Chinese Academy of Sciencs, Beijing, China.
[4]Henan Institute of Advanced Technology, Zhengzhou University, Zhengzhou, China.
[5]School of IoT Engineering, Jiangsu Vocational College of Information Technology, Wuxi, China.
[6]Information Technology Department I, Shenzhen Stock Exchange.
{chentianyu22s,aoxiang}@ict.ac.cn

Abstract

In this paper, we extend financial sentiment analysis (FSA) to event-level since events usually serve as the subject of the sentiment in financial text. Though extracting events from the financial text may be conducive to accurate sentiment predictions, it has specialized challenges due to the lengthy and discontinuity of events in a financial text. To this end, we reconceptualize the event extraction as a classification task by designing a categorization comprising coarse-grained and fine-grained event categories. Under this setting, we formulate the **E**vent-Level **F**inancial **S**entiment **A**nalysis (**EFSA** for short) task that outputs quintuples consisting of (company, industry, coarse-grained event, fine-grained event, sentiment) from financial text. A large-scale Chinese dataset containing 12,160 news articles and 13,725 quintuples is publicized as a brand new testbed for our task. A four-hop Chain-of-Thought LLM-based approach is devised for this task. Systematically investigations are conducted on our dataset, and the empirical results demonstrate the benchmarking scores of existing methods and our proposed method can reach the current state-of-the-art. Our dataset and framework implementation are available at https://github.com/cty1934/EFSA.

1 Introduction

Since Robert F. Engle was awarded the Nobel Prize because he researched the influence of news on financial market volatility (Engle and Ng, 1993), it catalyzed a surge in academic interest in Financial Sentiment Analysis (FSA) (Luo et al., 2018; Xing et al., 2020; Du et al., 2024). FSA holds significant importance within the application domain of sentiment analysis (Du et al., 2024), encompassing the study of financial textual sentiment (Kearney and Liu, 2014) within news to forecast financial market dynamics. Recall that it is the events described in the financial texts and the related sentiments that dominate the impact of financial news on market volatility (Xing et al., 2020), the primary focus of FSA should sit on events extraction and related sentiment analysis.

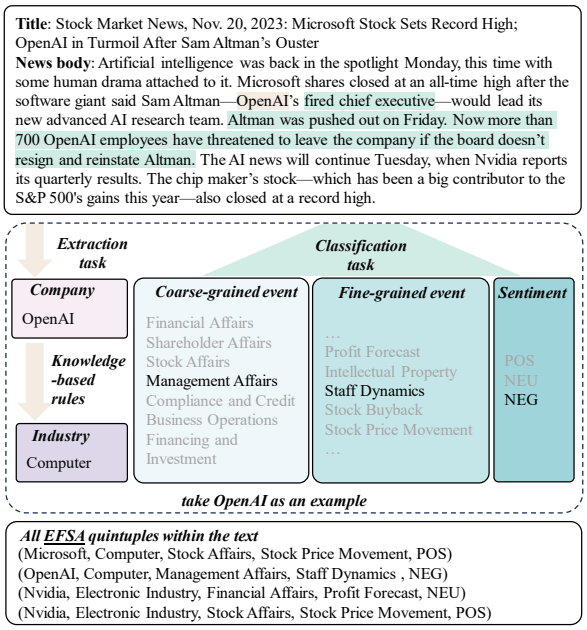

Figure 1: An example of Event-level Financial Sentiment Analysis from financial news. There could be multiple entities associated with their events with different sentiments.

However, most existing FSA studies focus on predicting entities and sentiments while neglecting the analysis of events within financial texts. To name some, FiQA (Maia et al., 2018), an open challenge financial news dataset, is designed to analyze sentiment corresponding to a certain entity. SEntFiN (Sinha et al., 2022) is a news headline dataset for entity analysis, including entity recognition from a predefined list and related sentiment analysis. A most recent FSA benchmark, namely Fin-Entiy (Tang et al., 2023), aims to jointly predict the entities and the associated sentiments from

*These authors contributed equally to this work.
†Correspondence to Xiang Ao and Li Zeng.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7455–7467

financial news. Nevertheless, financial text's sentiments frequently link to particular events. For example, in Figure 1, the same entity *Nvidia* exhibits distinctly opposite sentiments due to two different events: *stock price movement* and *profit forecast*. From this example, we can be aware that the event usually serves as the subject of the sentiment in financial text, while the entity is the target of the emotional impact. Extracting events from financial text may be conducive to accurate sentiment predictions.

To this end, we aim to discover the events in financial text to provide an easier financial sentiment prediction. An intuitive way to identify an event from financial text is extracting a specific text span from the original text just like existing aspect-based sentiment analysis (ABSA) tasks that find both aspect and opinion terms from customer's comments (Zhang et al., 2022b; Yu et al., 2023, 2021b). However, direct adapting approaches of ABSA for this task could be ineffective even infeasible due to the events in the financial text being overlong and discontinuous (c.f. Figure 1). We will provide a further discussion about the similarities and differences between ABSA and our task in Section 2.3.

In this paper, based on our observations, we provide an alternative setting for FSA and propose a novel task, named **E**vent-Level **F**inancial **S**entiment **A**nalysis (**EFSA**), involving the prediction of quintuples (company, industry, coarse-grained event, fine-grained event, sentiment). Here we have enhanced the FSA tasks in two facets. First, to overcome the difficulties associated with extracting events from financial texts, we reconceptualize the event extraction task as a classification task. We design a categorization system that comprises both coarse-grained and fine-grained event categories, specifically tailored for various event types in financial news. Besides, we construct a knowledge-based rule to classify companies by industry, enabling FSA to elevate to higher dimensions, such as indices or sector markets. Our task setting can offer substantial practical value in financial applications, including stock trading, stock market anomaly attribution analysis, enterprise risk management, etc. Figure 1 presents example quintuples in our EFSA task.

To support this task, we annotated a large-scale dataset from Chinese financial news. The dataset includes $12,160$ news articles, selected from an initial set of over $50,000$ articles collected from mainstream Chinese financial news websites. Detailed annotations were conducted based on the above task settings. To the best of our knowledge, this dataset is a large-scale fine-grained annotated dataset for FSA and the largest Chinese dataset in the event-level FSA domain.

We conducted comprehensive experiments to benchmark our dataset. Empirical results demonstrate the EFSA task presents a significant challenge, even for advanced large language models (LLMs) such as GPT-4, primarily due to the complexity of simultaneously predicting two categories and the fact that the sentiments in financial texts are primarily implicit. Recent studies on implicit sentiments underscore the efficacy of the Chain-of-Thought (CoT) approach in reasoning implicit sentiments in fine-grained sentiment analysis (Fei et al., 2023). Consequently, we devise a framework utilizing a 4-hop Chain of Thought (CoT) prompt based on LLMs, achieving the highest level of performance recorded for our task.

The main contributions of our work can be summarized as follows:

- We propose a novel Event-Level Financial Sentiment Analysis task for financial sentiment analysis, named EFSA, including the prediction of quintuples consisting of (company, industry, coarse-grained event, fine-grained event, and sentiment).
- We publicize the largest-scale Chinese financial corpus to support the EFSA task.
- We conduct systematically benchmark experiments to investigate the efficacy of existing methods for the EFSA task and introduce a novel LLM-based framework reaching the current state-of-the-art.

2 Problem Formulation and Discussion

In this section, we formulate the EFSA task and differentiate it from the traditional ABSA task.

2.1 Problem Formulation

The problem of EFSA is formulated as follows: Given an input financial news context x contains n words, EFSA aims to predict a set of quintuples (c, i, e^1, e^2, s), corresponding to (*company, industry, coarse-grained event, fine-grained event, sentiment polarity*). Here, *company* c is a text span within the sentence. Each *company* c is categorized

Task	Output
EFSA	(c, i, e^1, e^2, s)
Coarse-grained EFSA	(c, i, e^1, s)
Fine-grained EFSA	(c, i, e^2, s)
Entity-Level FSA	(c, i, s)

Table 1: The sub-tasks of EFSA.

into a specific *industry* i by knowledge-based rules. *Coarse-grained event* e^1 and *fine-grained event* e^2 belong to two distinct predefined event type sets E^1 and E^2, where E^2 is a finer-grained division of E^1. $s \in$ {positive, negative, neutral} denotes the sentiment polarity.

2.2 The Sub-tasks of EFSA

The EFSA task could be further divided into two event-level sub-tasks, namely **C**oarse-grained EFSA (C-EFSA) and **F**ine-grained EFSA (F-EFSA), along with an entity-level FSA task. These tasks are outlined in Table 1. Due to the granularity difference of events categorization within different tasks, the difficulty of the three tasks, C-EFSA, F-EFSA, and complete EFSA, exhibits an increasing trend. This enables FSA researchers to conduct experiments on tasks of varying difficulty levels and financial practitioners to analyze market conditions at different event granularities. Additionally, our dataset can also support existing FSA tasks. By omitting event classification, our task can be simplified to a purely entity-level FSA task.

2.3 Relatedness to ABSA Task

Here, we discuss the similarities and differences between our EFSA task and the ABSA task.

Recall that ABSA aims to identify sentiment elements related to a specific text, which could either be single or multiple elements, including the dependency relations among them (Zhang et al., 2022b; Yu et al., 2021a). These sentiment elements comprise aspect terms, sometimes aspect categories, opinion terms, and aspect-level sentiment. Among them, aspect and opinion terms are specific text spans within a sentence, and identifying them is an extraction task. Determining aspect categories and sentiments are usually classification tasks.

Although EFSA and ABSA appear to be similar in form, as they both involve extracting the sentiment elements from the original text, i.e. companies and aspect terms, and identifying sentiment-related phrases (events or opinion terms), and finally determining sentiment, EFSA presents two main differences. Firstly, the events in EFSA may be extremely long and discontinuous, which results in our formulation simplifying the event extraction to a classification task. Few existing ABSA approaches can address such change without any modification. Secondly, the financial context makes its mapping from events to sentiments unique in EFSA. The mapping is constructed on domain expertise rather than subjective emotional expressions. These two reasons essentially differentiate EFSA from ABSA.

The sub-tasks of EFSA exhibit a certain degree of similarity to the ABSA task. For instance, the entity-level FSA can be addressed by most current ABSA approaches. Hence, in subsequent experiments in Section 4.1, we benchmark several applicable ABSA baselines on our dataset.

3 Dataset

Financial sentiment indicators are categorized into market-derived sentiment and human-annotated sentiments (Luo et al., 2018). The market-derived sentiment is estimated based on market dynamics such as stock price changes and trading volume, potentially incorporating noise from other sources (Fei et al., 2023). Therefore, we employed manual labeling conducted by financial experts to ensure the creation of a high-quality dataset. Specifically, the data construction process is elaborated in detail from the following three sections: data collection, data annotation, and data distribution.

3.1 Data Collection

We collected over $50,000$ financial news articles from various reputable Chinese news outlets from their publicly accessible websites. Each collected article in the dataset includes several elements: URL, publish time, title, and news body. To control the task's complexity, we limited our selection to a news body comprising no more than 300 Chinese characters. Based on the original data, we manually conducted data cleaning and filtering, retaining only high-quality data. Consequently, a total of $12,160$ articles were earmarked for annotation.

3.2 Data Annotation

Labeling System. To address the complex spectrum of event types in financial news, we developed a detailed event taxonomy by professional

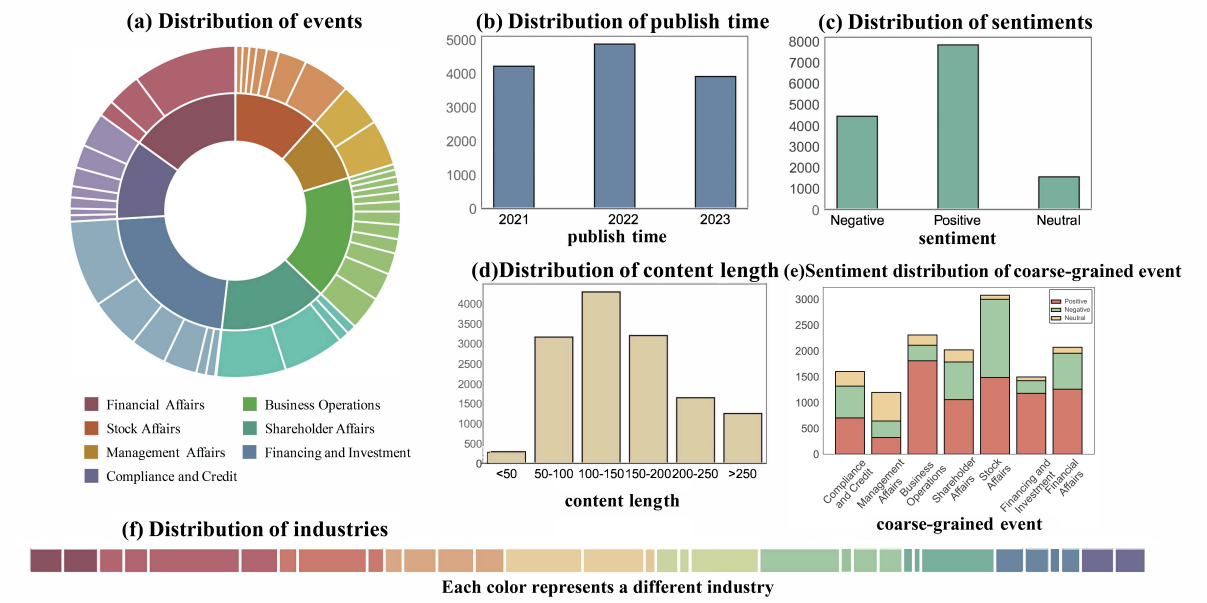

Figure 2: The statistics of data distribution of our dataset. (a) presents a nested pie chart that illustrates the distribution of event labels. The inner layer and the outer layer respectively represent the distribution of coarse- and fine-grained events; the dark and light shades of the same color represent the coarse-grained event and its subdivided fine-grained events. (b), (c), (d), and (e) present the distribution among different publish times, overall sentiments, news body length, and sentiment distribution of each coarse-grained event. (f) presents the distribution of industries, where various colors denote 32 distinct industries. The size of the color blocks in (a) and (f) represents the data size.

financial practitioners. This taxonomy comprises seven coarse-grained categories: *Financial Affairs, Shareholder Affairs, Stock Affairs, Compliance and Credit, Management Affairs, Business Operations* and *Financing and Investment*, and further extends to 43 fine-grained categories. The complete event taxonomy is presented in Appendix A, serving as a labeling reference for researchers. For sentiment labels, we adopted a three-category sentiment polarity classification consisting of positive, neutral, and negative. Each company was classified into a industry referencing the Shenwan Industry Classification Standard*.

Annotation Platform. To streamline the data annotation process, we developed a specialized annotation platform tailored to our task requirements. This platform presents news bodies to annotators as input, enabling them to choose specific text spans within the news article for company labels and to directly assign event labels, sentiment labels, and industry labels within the system. The screenshot of the annotation platform is shown in Appendix B.

The annotation platform enables collaborative annotation through a two-step process: labeling and reviewing. Each news article is first labeled by an annotator and then reviewed by a reviewer. If errors are found, the data is reassigned for re-annotation. In cases of disagreement, a third annotator is tasked with resolution. This ensures the accuracy of each data is rigorously assessed by at least two individuals.

Labeling Evaluation Metrics. To ensure consistency across various annotations, we tasked multiple annotators with independently labeling the same news article set. Following previous work (Hripcsak and Rothschild, 2005; Barnes et al., 2018), we use different metrics to measure the inter-annotator agreement of different annotation tasks. We employ the *AvgAgr* metric (Wiebe et al., 2005) to evaluate span extraction annotation consistency. The *AvgAgr* score of *Company* is 0.65. For classification annotation consistency score, we employ *Fleiss' Kappa* (Fleiss, 1971) to evaluate. The *Fleiss' Kappa* scores of classification annotation are as follows: *fine-grained event* (0.62), *sentiment* (0.67), and *industry* (0.70). The results, which fall between 0.61 and 0.8, demonstrate a substantial agreement (Landis JRKoch, 1977) among different annotators.

*The Shenwan Industry Classification is a widely used industry classification standard consisting of 32 industry categories. It was proposed by Shenwan Hongyuan Securities for financial investment management and research. https://www.swsresearch.com/institute_sw/allIndex/downloadCenter/industryType

3.3 Data Distribution

We balanced the distribution of different elements in our dataset to ensure a proportional data distribution. The specific data distribution is reported as follows:

Figure 2 (a) demonstrates the substantial uniformity of coarse- and fine-grained event distribution. For each fine-grained event, there is a sufficient amount of data to support, ensuring no instances no data shortage exists for any particular category. As shown in Figure 2 (b), We balanced the distribution of years to reflect market dynamics evenly across different periods. Most of our data's publication times span from 2021 to 2023. Figure 2 (d) demonstrates the distribution of context length. We restricted the length of the news context to 300 Chinese characters. Besides, we balanced the distribution of context length to manage the complexity of the task, with the average length of the news body in our dataset being 148.5. Figure 2 (c) and (e) demonstrate that our dataset achieves a balanced overall sentiment distribution. Additionally, it maintains a balance between positive and negative sentiments within specific coarse-grained events, since financial texts inherently tend to exhibit a sentiment bias rather than being neutral (Cortis et al., 2017; de França Costa and da Silva, 2018). For fine-grained events, given that certain events have specific sentiment tendencies, such as *Legal Affairs* often corresponding to negative emotions, we did not balance the sentiment distribution of each fine-grained event. We also calculated the distribution of industries (c.f. Figure 2 (f)) to ensure our dataset comprehensively and uniformly covers a wide range of sectors.

4 Experimental Evaluation

In this section, we benchmark EFSA with some widely used language models. In the following part, we will introduce the benchmark methods first, then detail our proposed framework, and finally present the evaluation metrics.

4.1 Benchmark Methods

We mainly benchmark two groups of pre-trained language models, including Large Language Models (LLMs) and Small Language Models (SLMs).

LLMs. We prioritized models with strong Chinese language capabilities, selecting general domain LLMs and financial domain-specific LLMs. Particularly, we fine-tuned the open-source, deployable general domain LLMs using LoRA (Hu et al., 2021) on our dataset. We interacted with the LLMs by constructing prompts to instruct LLMs to generate structured outputs. To ensure fairness in our benchmark tests, we use the same prompts for every LLMs under different settings, which are detailed in Appendix C.

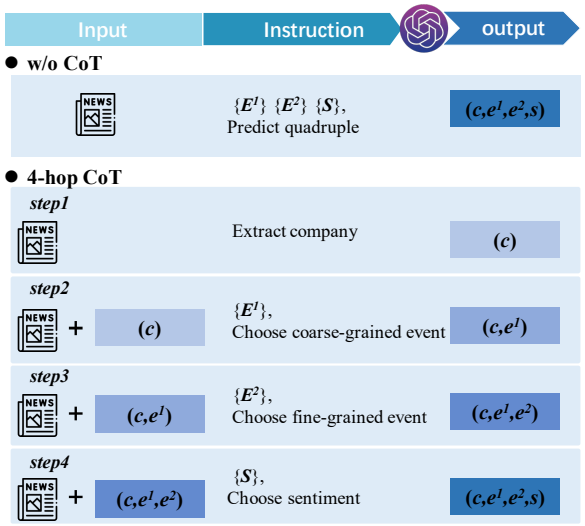

Figure 3: An illustration of the four-hop CoT framework. E^1, E^2, S respectively denotes the set of coarse-grained events, fine-grained events and sentiment polarities, respectively.

(1) **ChatGPT** (Brown et al., 2020). We interacted with ChatGPT by querying the API interface. Due to the cost constraint, we randomly selected 2,000 entries from our dataset for evaluation. We evaluated gpt-3.5-turbo and gpt-4-turbo-preview[†] on this subset under both zero- and 3-shot settings.

(2) **ChatGLM** (Du et al., 2021) is a Chinese and English bilingual language model, constructed utilizing the General Language Model (GLM) framework. ChatGLM-3 demonstrated superior performance on the Chinese LLM evaluation benchmark C-EVAL (Huang et al., 2023). Moreover, we extended our analysis to the latest iteration, ChatGLM-4, by querying the API interface glm-4 provided by ZHIPU AI open platform. ChatGLM-4 stands out as one of the LLMs known for its robust Chinese language alignment capabilities.

(3) **Llama2-Chinese** employs a Chinese instruction set for LoRA fine-tuning on Llama2-Chinese-7b (Touvron et al., 2023) to enhance its alignment with Chinese and capabilities for Chinese dialogue.

(4) **Baichuan2** (Yang et al., 2023) is a Chinese

[†]The latest model interface provided by OpenAI currently points to gpt-3.5-turbo-0613 and gpt-4-0125-preview.

and English bilingual language model. It achieved the best performance among models of the same size on standard benchmarks (C-Eval (Huang et al., 2023), MMLU (Hendrycks et al., 2020), etc).

(5) **QwenLM** (Bai et al., 2023) is a Chinese and English bilingual language model. It achieved better performance than LLaMA2-70B on all tasks and outperforms GPT-3.5 on 7 out of 10 benchmarks (Bai et al., 2023).

(6) **DISC-FinLLM** (Chen et al., 2023) is a financial domain-specific LLM fine-tuned by Baichuan2-13B-Chat with LoRA on using various financial task open datasets.

(7) **Xuanyuan** (Zhang and Yang, 2023) is a financial domain-specific LLM derived through incremental pre-training based on Llama2. It exhibits significant enhancements in both Chinese language proficiency and financial capabilities.

SLMs. Since Zhang et al. (2021b)'s pioneering application of generative methods to ABSA through the construction of generative paradigms, the SOTA leaderboard of ABSA has been consistently dominated by generative methods (Gou et al., 2023; Wang et al., 2022). Given the similarities between our sub-tasks and the ABSA task, we benchmarked several advanced SLMs that achieved promising performance on ABSA.

Leveraging the GAS (Generative ABSA)'s framework (Zhang et al., 2021b), we benchmarked the performance of the **mT5-large** and **BART-large-chinese** (Zhang et al., 2021a) models on entity-level sub-task. We also benchmarked **E2E-BERT** (Li et al., 2019) on this sub-task. We randomly split 80% data for training and the left 20% part is for testing. It is noted that the original training hyperparameters are used for each respective model.

4.2 Our Framework

Given Chain of Thought (CoT)'s success in analyzing implicit sentiment (Fei et al., 2023) and the inherent chained progression relationship between coarse- and fine-grained events in the EFSA task, we devised a four-hop CoT framework for our task. It involves four steps. **Step 1.** Utilize a news article as input and instruct the LLM to identify the company mentioned within the text. **Step 2.** Subsequently, use the news article and the identified companies as input. Instruct the LLM to choose a corresponding event from a predefined set of coarse-grained events. **Step 3.** Use the news article and the (company, coarse-grained-event) tuple as input. Instruct the LLM to choose a corresponding event from a predefined set of fine-grained events. **Step 4.** Use the news article and the (company, coarse-grained event, fine-grained event) triplet as input. Instruct the LLM to choose a corresponding sentiment of the triplet.

Figure 3 illustrates our four-hop CoT framework. The details of the four-hop CoT prompt are displayed in Appendix C. Based on this framework, we employed dialogue fine-tuning on deployable LLMs to evaluate effectiveness.

4.3 Evaluation Metrics

We combine the proposed framework with the aforementioned LMs on different tasks and compute the F1 scores for evaluation. Notably, due to the presence of geographic names, stock codes, and other identifiers in the news text's company labels, a company prediction is considered correct if it is included within the gold label. Predictions of events and sentiments must be exactly matched with the gold label to be considered correct.

5 Experimental Results

5.1 Main Results and Analysis

The results of different EFSA sub-tasks are reported in Table 2. There are two notable observations. First, the performance across sub-tasks to the complete EFSA task exhibits a trend of declining scores, consistent with the escalating difficulty of the tasks. Second, the overall scores of the entity-level task significantly surpass event-level tasks, which reveals the inherent complexity and challenges of our event-level tasks. We make a more comprehensive comparison of various methods below.

General Domain LLMs. Our prior experiments demonstrate that smaller-parameter LLMs (ChatGLM, Llama2-Chinese, Baichuan2, and QwenLM) perform poorly in zero-shot and few-shot settings (c.f. Appendix D). This can be attributed to the inherent complexity of the EFSA task and the limited capacity of smaller models to produce structured outputs, which leads to their failure in generating the specified quadruple responce format, culminating in incorrect outputs. Fine-tuning these LLMs can significantly improve their performance scores by enhancing both domain-specific ability and capability for structured output.

Larger-parameter LLMs (ChatGPT, GPT-4, and GLM-4) under zero-shot settings also fail to pro-

Settings	Methods	Entity-Level FSA	C-EFSA	F-EFSA	EFSA
	• ***General Domain LLMs***				
zero-shot	ChatGPT	58.47	37.92	36.96	26.17
	GPT-4	60.72	48.48	45.45	36.10
	GLM-4	71.25	57.31	52.72	49.41
3-shot	ChatGPT	58.92	39.24	37.13	27.57
	GPT-4	61.38	51.43	49.53	39.24
	GLM-4	70.39	56.97	54.90	50.68
LoRa fine-tune	ChatGLM3-6B-Chat	76.83	62.26	53.11	51.18
	Baichuan2-13B-Chat	**86.41**	71.82	67.79	67.06
	Qwen-7B-Chat	86.14	73.22	67.32	67.28
	Llama2-Chinese-7B-Chat	61.55	43.84	36.71	36.28
	• ***Financial Domain LLMs***				
zero-shot	DISC-FinLLM	63.74	32.43	22.45	19.25
	Xuanyuan	64.15	22.39	16.41	7.85
3-shot	DISC-FinLLM	65.23	38.67	27.07	24.68
	Xuanyuan	63.58	26.70	17.16	12.39
LoRa fine-tune	DISC-FinLLM	79.19	56.08	49.16	46.98
	Xuanyuan	-	-	-	-
	• ***SLMs***				
-	E2E-BERT	73.77	-	-	-
	GAS-T5	69.75	-	-	-
	GAS-BART	50.89	-	-	-
	• ***Our Framework***				
4-hop CoT	GPT-4	63.28	55.64	53.24	53.24
	ChatGLM3-6B-Chat	78.93	70.61	65.19	65.19
	Baichuan2-13B-Chat	81.74	75.66	69.52	69.52
	Qwen-7B-Chat	83.28	**76.03**	**71.43**	**71.43**
	Llama2-Chinese-7b-Chat	61.47	51.62	23.44	23.44

Table 2: Comparison results on different EFSA tasks. The reported scores are F1 scores over one run. '-' denotes that the corresponding results are not available. The best results are bolded. We employ dialogue LoRa fine-tuning based on our framework, which enables the original model to achieve significant score improvements. With the 4-hop CoT framework, the F-EFSA's score is identical to the EFSA's. This is because the accurate prediction of fine-grained events is built upon the correct prediction of coarse-grained events.

duce satisfactory results. Few-shot settings can enhance their performance, with particularly noticeable improvements for event classification tasks and slight improvements for sentiment analysis tasks. This may be attributed to the inherent capabilities of larger-parameter LLMs in sentiment analysis tasks. Few-shot demonstrations primarily boost the LLMs' proficiency in our specialized event classification tasks.

Financial Domain-Specific LLMs. Financial domain-specific LLMs are fine-tuned or pre-trained on open-source financial domain datasets based on general domain LLMs. Our benchmark results exhibit the enhanced performance of domain-specific LLMs relative to their original base models under the zero-shot settings (DISC vs. Baichuan, Xuanyuan vs. Llama), suggesting that domain-specific customization can significantly enhance performance on previously unseen tasks within the same domain. Furthermore, domain-specific LLMs demonstrate a commendable capability for structured output, which may be attributed to other structured tasks within their training datasets.

However, financial domain-specific LLMs do not outperform LLMs that are fine-tuned on our dataset. So we further fine-tuned financial domain-specific LLMs on our dataset. The DISC model itself is based on LoRA fine-tuning. We utilized a blend ratio of 7:3, integrating the fine-tuning weights derived from our dataset with DISC's original weights. LoRA fine-tuning is not applicable to XuanYuan, as XuanYuan is re-pretrained. The results show that further fine-tuning can enhance the overall capabilities. Yet, it still cannot surpass the performance of their base models fine-tuned solely on our dataset. Our dataset can serve as a rich resource to facilitate the advancement of financial domain-specific LLMs.

SLMs. The benchmark scores of SLMs underscore that SLMs outstrip the performance of zero-shot LLMs, affirming the conclusions presented in Zhang et al. (2023)'s study: while LLMs

have shown proficiency in many sentiment analysis tasks, they fall short when it comes to extracting structured sentiment and opinion information. Smaller models have learned better structured output capabilities through fine-tuning. However, due to the constraints imposed by task difficulty, the number of parameters has limited the upper potential. Consequently, the performance of SLMs does not exceed that of fine-tuned LLMs.

Our Framework. Our Chain of Thought (CoT) framework augments the efficacy of non-open-source LLMs through prompt-based adjustments alone. For open-source LLMs, the dialogue fine-tuning empowers the foundational model to realize more substantial score enhancements, particularly notable in models with initially subpar performance (e.g., Llama). Notably, the CoT framework hurts entity-level FSA tasks. This is because the score of the (company, sentiment) tuple is directly calculated based on the tuple part of the quadruple outputted by the 4-hop CoT, the process is susceptible to error accumulation. Inaccuracies in event prediction can lead to erroneous sentiment predictions, adversely affecting the accuracy of Entity-Level FSA.

5.2 Case Study

To better demonstrate the effectiveness of our framework, we perform a case study on GPT-4, as shown in Figure 4. As shown in Example 1, GPT-4 correctly predicted fine-grained events but made mistakes in predicting coarse-grained events, confusing the two difficult-to-distinguish categories of *shareholder affairs* and *stock affairs*. It is possibly due to the multiple occurrences of *stock* in the text leading to misdirection. Under our CoT framework, in the second hop of reasoning from company to coarse-grained events, GPT-4 focuses more on events directly related to the company itself, thereby making the correct prediction. The CoT framework can prevent a situation where the predictions for fine-grained events are correct, yet those for coarse-grained events are not. For sentiment prediction, an inexperienced LLM can easily be misled by *the company receives a notice of share reduction*. However, under the CoT framework, GPT-4 pays more attention to the sentiment of the fine-grained event (*Stock Holding Adjustment*) itself and can deduce the correct sentiment from the subsequent information that there was no actual share reduction action.

Error Analysis. We observed that LLMs sometimes produce outputs beyond the defined label sets, specifically illustrated in Example-2 and -3. LLMs may alter the fixed label instead of outputting as instructed. This observation aligns with Zhang et al. (2021a)'s previous research, highlighting the nature of the generation modeling since it does not perform "extraction" in the given sentence. This phenomenon is more pronounced in LLMs with smaller parameter sizes and can be mitigated through fine-tuning. Similar to instructing LLMs toward structured outputs, fine-tuning significantly enhances the ability of smaller-parameter LLMs to output as required.

6 Related Work

Previous research on the FSA datasets has concentrated on sentence or document level (Takala et al., 2014; Malo et al., 2014; Cortis et al., 2017; Sinha et al., 2022). This is based on an assumption that the given text conveys a single sentiment towards a certain topic. Recent FSA datasets exhibit a trend towards progressively finer granularity. However, fine-grained datasets (Maia et al., 2018; Tang et al., 2023) mostly focus on entities and sentiments, neglecting the concern of events. Furthermore, the resources for fine-grained FSA datasets are still limited (Du et al., 2024). Following Du et al. (2024), we summarize the most widely used and recent FSA benchmark datasets in Appendix E.

Event-level sentiment analysis is designed to identify user emotions on social platforms regarding current events (Zhang et al., 2022a). In this paper, we broaden the scope of event-level sentiment analysis by applying it to FSA, enhancing its relevance and utility in financial contexts.

7 Conclusion

In this paper, we present EFSA, a novel task for Financial Sentiment Analysis (FSA), which deepens FSA to the event level. To support this task, we constructed the largest Chinese dataset annotated for event-level FSA from a large-scale financial news corpus. We evaluated this task on widely used language models to present benchmark scores on the proposed dataset. Additionally, we designed a 4-hop reasoning prompting framework based on existing LLMs to resolve this task. Our experiments demonstrate the challenge of EFSA and the effectiveness of our proposed approach. Our work opens a new avenue in FSA and offers significant value to the entire financial domain.

Case study

Example-1

News body: Aim Pharm(002826.SZ) announced on January 17, 2022, that the company has received a "Notice of Progress on Share Reduction" from Mr. Zhou, a member of the board. The share reduction plan has been fully implemented, and Mr. Zhou did not reduce his holdings in the company's shares in any form during the period of this reduction plan.

Gold Label: (Aim Pharm, Health Care, Shareholder Affairs, Stock Holding Adjustment ,NEU)

GPT-4 zero-shot: (Aim Pharm, Health Care, Stock Affairs, Stock Holding Adjustment ,NEG) ✗

GPT-4 + CoT: (Aim Pharm, Health Care, Shareholder Affairs, Stock Holding Adjustment ,NEU) ✓

Error analysis

Example-2

News body: JPMorgan lowers the target price for Mengniu Dairy from HK$51.8 to HK$49, with an "overweight" rating.

Gold Label: (Mengniu Dairy, Food & Beverage, Management Affairs, Rating Adjustment, NEG)

Qwen-7B-Chat + CoT: (Mengniu Dairy, Food & Beverage, Management Affairs, **Stock** Rating Adjustment, NEG) ✗

Example-3

News body: iFlytek plunges in the afternoon session, hitting the limit down, with the current turnover exceeding 3.6 billion yuan

Gold Label: (iFlytek, Information Service, Stock Affairs, Stock Price Movement, NEG)

ChatGLM3-6B-Chat: (iFlytek, Information Service, Stock **Events**, Stock Price Movement, NEG) ✗

Figure 4: Examples provided include the input news text, the corresponding gold labels, and the predicted quantities. The red font denotes the incorrect part of the prediction.

Limitations

Due to budget constraints, we only conducted experiments on a subset of our dataset using LLMs that have not been open-sourced yet (Chatgpt, GPT-4, GLM-4), on a single run. This may result in bias in our evaluation scores. Although CoT's gradual processing approach leads to more accurate and reliable results, this may increase some additional computation cost and time in solving complex FSA problems.

Ethics Statement

The collected data originated entirely from publicly accessible websites. Our datasets do not disseminate personal information and are devoid of content that could potentially harm any individual or community. The annotators are undergraduate students, which are credited as authors of the paper, and compensated in accordance with regulations. Please note that our data is intended solely as an open-source dataset resource to foster the advancement of FSA research. Any illegal use of this data is strictly prohibited.

Acknowledgements

The research work is supported by the National Key R&D Plan No. 2022YFC3303305. Xiang Ao is also supported by the Project of Youth Innovation Promotion Association CAS, and Beijing Nova Program 20230484430.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jeremy Barnes, Patrik Lambert, and Toni Badia. 2018. Multibooked: A corpus of basque and catalan hotel reviews annotated for aspect-level sentiment classification. *arXiv preprint arXiv:1803.08614*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, et al. 2023. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. *arXiv preprint arXiv:2310.15205*.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 519–535.

Dayan de França Costa and Nadia Felix Felipe da Silva. 2018. Inf-ufg at fiqa 2018 task 1: predicting sentiments and aspects on financial tweets and news headlines. In *Companion Proceedings of the The Web Conference 2018*, pages 1967–1971.

Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria.

2024. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Robert F Engle and Victor K Ng. 1993. Measuring and testing the impact of news on volatility. *The journal of finance*, 48(5):1749–1778.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. Mvp: Multi-view prompting improves aspect sentiment tuple prediction. *arXiv preprint arXiv:2305.12627*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Colm Kearney and Sha Liu. 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185.

GG Landis JRKoch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159174.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting bert for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1910.00883*.

Ling Luo, Xiang Ao, Feiyang Pan, Jin Wang, Tong Zhao, Ningzi Yu, and Qing He. 2018. Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention. In *IJCAI*, pages 4244–4250.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Ankur Sinha, Satishwar Kedas, Rishu Kumar, and Pekka Malo. 2022. Sentfin 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology*, 73(9):1314–1335.

Pyry Takala, Pekka Malo, Ankur Sinha, and Oskar Ahlgren. 2014. Gold-standard for topic-specific sentiment analysis of economic texts. In *LREC*, volume 2014, pages 2152–2157.

Yixuan Tang, Yi Yang, Allen H Huang, Andy Tam, and Justin Z Tang. 2023. Finentity: Entity-level sentiment classification for financial texts. *arXiv preprint arXiv:2310.12406*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zengzhi Wang, Rui Xia, and Jianfei Yu. 2022. Unified-absa: A unified absa framework based on multi-task instruction tuning. *arXiv preprint arXiv:2211.10986*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: an investigation into common mistakes and silver bullets. In *Proceedings of the 28th international conference on computational linguistics*, pages 978–987.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Guoxin Yu, Xiang Ao, Ling Luo, Min Yang, Xiaofei Sun, Jiwei Li, and Qing He. 2021a. Making flexible use of subtasks: A multiplex interaction network for unified aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2695–2705.

Guoxin Yu, Jiwei Li, Ling Luo, Yuxian Meng, Xiang Ao, and Qing He. 2021b. Self question-answering:

Aspect-based sentiment analysis by role flipped machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1331–1342.

Guoxin Yu, Lemao Liu, Haiyun Jiang, Shuming Shi, and Xiang Ao. 2023. Making better use of training corpus: Retrieval-based aspect sentiment triplet extraction via label interpolation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4914–4927.

Qi Zhang, Jie Zhou, Qin Chen, Qingchun Bai, and Liang He. 2022a. Enhancing event-level sentiment analysis with structured arguments. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1944–1949.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. *arXiv preprint arXiv:2110.00796*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022b. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.

Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4435–4439.

A Event Taxonomy

The event taxonomy is translated from Chinese; for more accurate and detailed information, please refer to our open-source website.

Financial Affairs
- Profit Announcement
- Profit Forecast
- Other Financial Affairs

Shareholder Affairs
- Stock Holding Adjustment
- Shareholder Pledge
- Release of Pledge
- Other Shareholder Affairs

Stock Affairs
- Stock Price Movement
- Equity Incentives & Employee Stock Ownership Plans
- Stock Dividend
- Stock Buyback
- Stock Status
- Restricted Shares Release
- Other Stock Affairs

Business Operations
- Product Dynamics
- Capacity Changes
- Initiating Cooperation
- Technical Quality Control, Qualification Changes
- Government Subsidies
- New Company Establishment
- Institutional Research
- Intellectual Property
- Sales, Market Share Changes
- Project Bidding
- Project Dynamics
- Other Business Operations Affairs

Compliance and Credit
- Company Litigation
- Rating Adjustment
- Legal Affairs
- Clarification Announcements
- Regulatory Inquiries
- Case Investigations
- Administrative Penalties
- Other Compliance and Credit Affairs

Management Affairs
- Employee Dynamics
- Directors, Supervisors, and Senior Executives Dynamics

Financing and Investment
- Company Listing
- Mergers and Acquisitions
- Investment Events
- Stock Issuance
- Financing and Margin Trading
- Capital Flows
- Other Financing and Investment Affairs

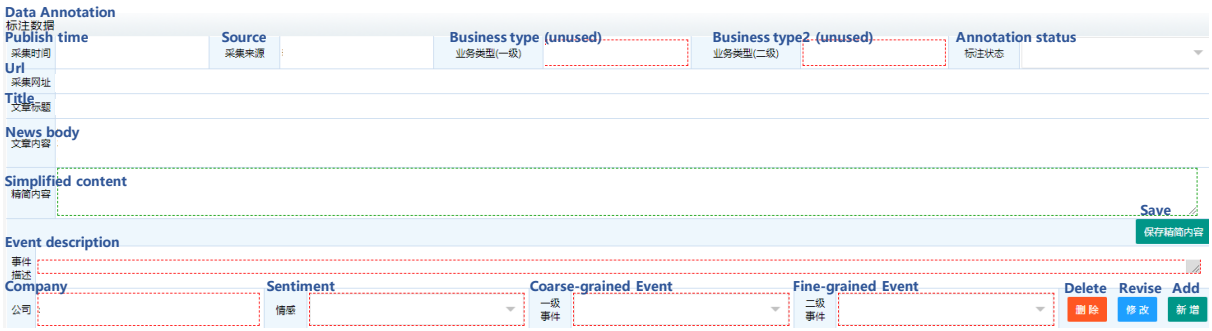

Figure 5: Screenshot of the annotation system.

Settings	Methods	Entity-Level FSA	C-EFSA	F-EFSA	EFSA
zero-shot	ChatGLM3-6B-Chat	62.07	19.43	10.14	8.04
	Baichuan2-13B-Chat	60.81	23.21	8.69	7.20
	Qwen-7B-Chat	59.36	20.58	8.15	5.77
	Llama2-Chinese-7B-Chat	12.62	1.99	0.43	0.38
3-shot	ChatGLM3-6B-Chat	66.66	27.87	16.90	13.86
	Baichuan2-13B-Chat	69.16	27.28	11.83	10.09
	Qwen-7B-Chat	63.16	24.55	16.76	14.60
	Llama2-Chinese-7B-Chat	41.59	15.15	6.22	4.96

Table 3: Results of smaller-parameter LLMs in zero-shot and 3-shot setting

B Annotation Platform

Figure 5 presents a screenshot of the annotation system, illustrating its primary functions.

C Prompt

The prompts are translated from Chinese; for more accurate and detailed information please refer to our open-source website.

Quintuple Instruction Prompt

Assuming you are a fine-grained sentiment analysis model in the finance domain, I will give you a list of primary events, a list of secondary events, a list of sentiment polarities, and some related financial news. Please analyze which company's event is mentioned in this financial news, then determine which primary event this event belongs to, further determine the secondary event based on the primary event, and finally identify the event's sentiment polarity.

Primary Event List: ['Financial Affairs', 'Shareholder Affairs', 'Stock Affairs', 'Management Affairs', 'Compliance and Credit', 'Business Operations', 'Financing and Investment'].

Secondary Event List: ['Profit Announcement', 'Profit Forecast', 'Other Financial Affairs', 'Stock Holding Adjustment', 'Shareholder Pledge', 'Release of Pledge', 'Other Shareholder Affairs', 'Stock Price Movement', 'Stock Status', 'Restricted Shares Release', 'Stock Buyback', 'Equity Incentives & Employee Stock Ownership Plans', 'Restricted Stock Release', 'Stock Dividend', 'Other Stock Affairs', 'Directors, Supervisors, and Senior Executives Dynamics', 'Employee Dynamics', 'Regulatory Inquiries', 'Company Litigation', 'Case Investigations', 'Administrative Penalties', 'Clarification Announcements', 'Legal Affairs', 'Rating Adjustment', 'Other Compliance and Credit Affairs', 'Project Bidding', 'Other Business Operations Affairs', 'Initiating Cooperation', 'New Company Establishment', 'Sales, Market Share Changes', 'Intellectual Property', 'Technical Quality Control, Qualification Changes', 'Government Subsidies', 'Institutional Research', 'Capacity Changes', 'Project Dynamics', 'Product Dynamics', 'Capital Flows', 'Investment Events', 'Financing and Margin Trading', 'Company Listing', 'Mergers and Acquisitions', 'Stock Issuance', 'Other Financing and Investment Affairs'].

Sentiment Polarity List: ['Positive', 'Negative', 'Neutral'].

Please answer in the form of a list of quadruples [(Company Name, Primary Event, Secondary Event, Sentiment Polarity)].

CoT Instruction Prompt

Step 1 Assuming you are a fine-grained sentiment analysis model in the finance domain, I will give you a piece of financial news, and you will determine the events of which companies are men-

Benchmark Dataset	Entries number	Annotation Type	Data source
Topic-Specific Sentiment Analysis (Takala et al., 2014)	297	Document-level (Topic)	News
PhraseBank (Malo et al., 2014)	4,846	Sentence-Level Polarity (sentiment polarity)	News Headlines
SemEval 2017 Task 5 (Cortis et al., 2017)	2,836	Sentence-Level (entity, sentiment score)	News headlines and Posts
SEntFiN (Sinha et al., 2022)	10,753	Sentence-Level (sentiment polarity)	News headlines
FiQA Task 1 (Maia et al., 2018)	1,173	Aspect-Level (entity, aspect, sentiment score)	News
FinEntity (Tang et al., 2023)	979	Aspect-Level (entity, sentiment polarity)	News
Ours	12,160	Event-Level (company, industry, coarse-grained event, fine-grained event, sentiment polarity)	News

Table 4: FSA Benchmark Datasets.

tioned.

Financial news as follows: [context]

Which company's event is described in the above financial news? Only answer with the company name, if there are multiple company names, separate them with commas. Do not add extra information.

Step 2 Assuming you are a fine-grained sentiment analysis model in the finance domain, I will give you a piece of financial news and the company name, and you will determine the primary event in the financial news for this company.

Financial news as follows: [context]

What is the primary event occurring to [response1] in the above financial news? Please select from the following list of primary events: ['Financial Affairs', 'Shareholder Affairs', 'Stock Affairs', 'Management Affairs', 'Compliance and Credit', 'Business Operations', 'Financing and Investment']. You must choose from the given list of primary events, output in the form of a tuple (Company Name, Primary Event). Do not add extra information.

Step 3 Assuming you are a fine-grained sentiment analysis model in the finance domain, I will give you a piece of financial news, the company name, and the primary event, and you will determine the secondary event for the company.

Financial news as follows: [context]

The primary event occurring to [response1] in the above financial news is [response2], please select the corresponding secondary event from the following list of secondary events. [Appropriate Secondary Event List] You must choose from the given list of secondary events, output in the form of a tuple (Company Name, Primary Event, Secondary Event). Do not add extra information.

Step 4 Assuming you are a fine-grained sentiment analysis model in the finance domain, I will give you a piece of financial news, the mentioned company, the primary and secondary events, and you will determine the sentiment polarity of this financial news event.

Financial news as follows: [context]

The primary event occurring to [response1] in the above financial news is [response2], and the secondary event is [response3].

Please select the appropriate sentiment from ['Positive', 'Negative', 'Neutral'], output in the form of a quadruple (Company Name, Primary Event, Secondary Event, Sentiment Polarity).

D Detailed Experimental Results

We conducted evaluations of all LLMs under zero-shot and three-shot settings. The performance of smaller-parameter LLMs counts was found to be suboptimal. Consequently, we excluded these results from the primary experimental results Table 2. These detailed results are reported in Table 3.

E FSA Benchmark Datasets

Table 4 summarizes the most widely used and recent FSA benchmark datasets.