EFSA-TD: Event-Level Financial Sentiment Analysis with Trigger Detection

Shanhong Liu, Hongde Liu, Xingren Wang, Feiyang Meng, Senbin Zhu, Chenyuan He, Jiayang Luo, Yuxiang Jia*
School of Computer Science and Artificial Intelligence
Zhengzhou University
Zhengzhou, China
{lsh2024, lhd_1013, wangxingren, fy_meng, nlpbin, hechenyuan_nlp, jyluo}@gs.zzu.edu.cn, ieyxjia@zzu.edu.cn

***Abstract*—Event-level financial sentiment analysis aims to identify sentiment polarities associated with financial events mentioned in text. Existing approaches typically represent events as abstract labels without explicitly modeling the textual evidence that signals event occurrences, which limits interpretability and robustness, especially in texts containing multiple events. To address this limitation, we propose EFSA-TD, a joint modeling framework for event-level financial sentiment analysis with trigger detection. EFSA-TD simultaneously detects financial events and their corresponding trigger spans while performing event-level sentiment analysis. By grounding event recognition and sentiment prediction in explicit textual evidence, EFSA-TD enables more accurate, robust, and explainable sentiment modeling. We further construct FinEvent-Trigger, a newly annotated dataset consisting of 9,120 financial news articles, in which 18,158 events are annotated with explicit event trigger spans. Experimental results show that the joint modeling of events and triggers by EFSA-TD improves both event classification and sentiment classification performance. We open source our data and code at https://github.com/forw30000/EFSA-TD.**

***Index Terms*—Financial sentiment analysis, Event-level sentiment analysis, Financial event extraction, Event trigger detection, Interpretability**

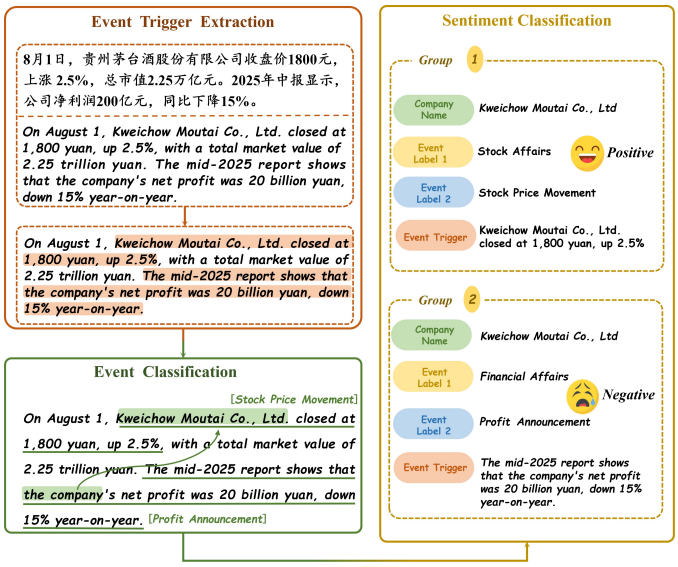

Fig. 1. An example of financial text sentiment analysis based on event trigger

I. Introduction

Financial sentiment analysis plays a crucial role in understanding market dynamics, investor behavior, and risk perception by extracting affective signals from financial texts such as news articles, earnings reports, and analytical commentaries [1]. With the rapid growth of online financial information, large volumes of financial news are generated daily, conveying complex sentiment signals that can significantly influence investment decisions and market movements [2]. Accurately identifying sentiment information from such texts has therefore become an important problem in both academic research and real-world financial applications [3] [4].

Early studies in financial sentiment analysis mainly focus on coarse-grained sentiment classification, where a single sentiment label is assigned to an entire document or sentence [5] [6]. While effective in simple scenarios, such approaches struggle to handle real-world financial texts that often describe multiple financial events with different sentiment polarities within the same context [7]. To address this limitation, event-level financial sentiment analysis associates sentiment polarity with specific financial events mentioned in text, enabling more fine-grained and actionable sentiment interpretation [8].

Despite the advantages of event-level modeling, existing approaches typically represent events as abstract categorical labels without explicitly modeling the textual evidence that signals event occurrences [9]. In complex financial narratives, sentiment expressions are often driven by concrete event descriptions, such as stock price movements, earnings announcements, or regulatory actions [10]. Ignoring the textual segments that trigger these events makes it difficult to explain why a particular event is identified or why a specific sentiment is assigned, thereby limiting model interpretability and robustness, especially in texts containing multiple overlapping events [11] .

To address the above issues, we propose an event-level financial sentiment analysis framework based on event trigger spans. As shown in Fig. 1. By jointly modeling events and their trigger spans, the framework enables event recognition and sentiment prediction grounded in textual evidence. Experimental results show that this approach effectively improves performance in both event classification and sentiment classi-

*Corresponding author: ieyxjia@zzu.edu.cn

fication tasks.

The main contributions of this paper are summarized as follows:

- We present **FinEvent-Trigger**, the first human-annotated benchmark dataset for event-level financial sentiment analysis with trigger detection, consisting of 9,120 financial news articles and 18,158 events. For each financial event, both sentiment annotations and explicit event trigger spans are provided.
- We introduce **EFSA-TD**, an event-level financial sentiment analysis framework that explicitly leverages event trigger information to ground event recognition and sentiment prediction in textual evidence.
- We conduct comprehensive experimental evaluations on the proposed dataset, demonstrating the effectiveness of event trigger spans and that EFSA-TD achieves state-of-the-art performance in both event classification and sentiment classification.

II. Related Work

A. Financial Sentiment Analysis

Financial Sentiment Analysis (FSA) is a core research task in financial natural language processing [6] [12], aiming to identify and model sentiment information related to market expectations, investor attitudes, and risk perceptions from unstructured texts such as financial news, analyst reports, and corporate disclosures [13]. Existing FSA approaches mainly focus on sentiment classification at the event or text level, emphasizing the prediction outcome while overlooking the explicit modeling of the underlying event factors or semantic cues that drive sentiment, which leads to limited interpretability [14]–[16]. Although some studies have begun to explore more fine-grained sentiment modeling paradigms [17], their expressive power remains constrained in complex financial contexts due to limitations in modeling capacity and application scope.

B. Financial Sentiment Analysis Dataset

Early financial sentiment analysis datasets were primarily annotated at the sentence or document level. Malo et al. [18] introduced the Financial Phrase Bank, and Cortis et al. [19] developed SemEval-2017, where sentiment labels are assigned to entire sentences or documents without explicitly considering the financial entities involved. Subsequently, research shifted toward entity-centric datasets for financial news analysis. Sinha et al. [20] proposed SEntFiN, which provides entity recognition and corresponding sentiment annotations based on a predefined set of entities. More recently, Tang et al. [21] introduced the financial sentiment analysis benchmark FinEntity, aiming to jointly predict entities and their associated sentiments from financial news, partially addressing the lack of entity-level annotations in earlier datasets.

Nevertheless, sentiment expressions in financial texts are often not determined solely by entities themselves but are closely tied to specific financial events [22]. To this end, Chen et al. [23] introduced the task of Event-Level Financial Sentiment Analysis (EFSA),which has attracted considerable attention and represents an important step toward modeling sentiment at the event level rather than the entity level. Despite this progress, existing event-level financial sentiment datasets typically annotate only the overall sentiment polarity of events, without providing fine-grained semantic grounding for the sources of sentiment, resulting in limited interpretability [24] [25].

Interpretability is widely regarded as an important research direction in financial sentiment analysis [14] [26]. If sentiment in financial news is only assigned abstract classifications, the resulting event information is often overly simplified, which makes it difficult to analyze and trace events and their associated sentiment. Moreover, large language models (LLMs) may produce hallucinations and generate inaccurate information, further affecting analysis reliability. We argue that the key sentences triggering events in financial news contain highly important information, such as the magnitude of stock price changes or investment amounts, which are crucial for supporting rational decision-making by investors. Therefore, our work aims to bridge this gap.

III. Dataset Construction

Table I provides an overview of our dataset in comparison with existing financial sentiment analysis datasets. Unlike previous datasets, ours not only supports entity-level financial sentiment analysis but also enables event-level sentiment analysis, with explicit event trigger spans provided to support semantic grounding. In this section, we describe the data sources and the detailed process of dataset construction.

TABLE I
COMPARISON OF FINANCIAL SENTIMENT ANALYSIS TASKS AND DATASETS.

Dataset	Number	Entity	Sentiment	Event	Trigger
FinEntity	979	✓	✓	×	×
EFSA	12,160	✓	✓	✓	×
Ours	**9,120**	✓	✓	✓	✓

A. Data Collection and Preprocessing

The data used in this study were collected from multiple mainstream Chinese financial news and information platforms, including the financial information provider JRJ[1], listed company announcements, and financial databases including Infobank[2]. The collected texts cover companies across a wide range of industries and scales, and involve diverse types of financial events, providing a realistic representation of event distributions in real-world financial news.

In the initial stage, approximately 22,000 news articles were collected. During preprocessing, we conducted systematic filtering and cleaning procedures to improve data quality. Specifically, advertisements, company profile introductions,

[1]https://www.jrj.com/
[2]http://bjinfobank.com

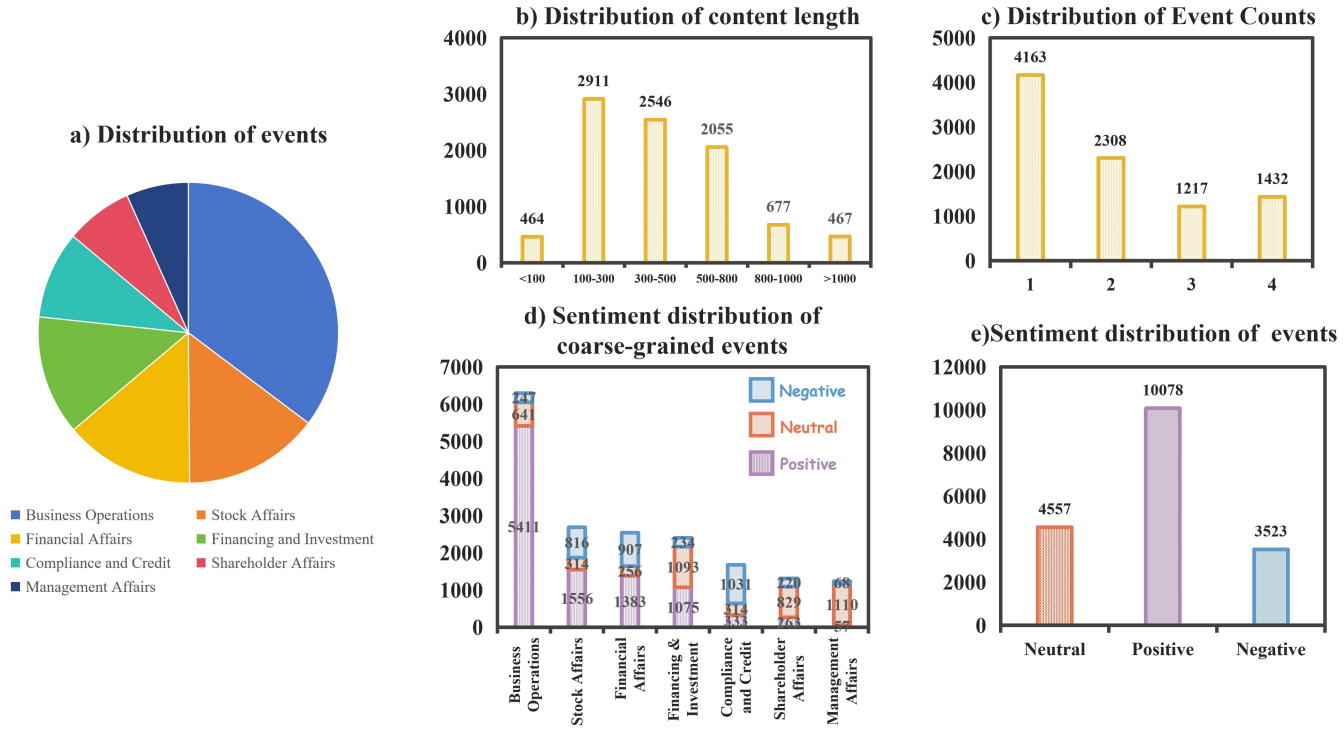

Fig. 2. Dataset statistics for event-level financial sentiment analysis.(a)Distribution of event categories.(b) Distribution of text length and event trigger spans length.(c) Distribution of the number of events per news instance.(d)–(e) Distribution of event-level sentiment polarity.

and news articles that did not contain explicitly described financial events were removed. Only articles with complete metadata (including title, content, and source) and clear, well-defined financial event descriptions were retained for subsequent annotation.

B. Event Annotation Scheme and Design

In the event annotation process, we adopt the event taxonomy proposed by Chen et al. in the Event-Level Financial Sentiment Analysis (EFSA) framework [23]. The EFSA taxonomy is hierarchically structured, consisting of seven coarse-grained event categories and 43 fine-grained event types. The seven top-level categories cover Financial Affairs, Shareholder Affairs, Stock Affairs, Compliance and Credit, Management Affairs, Business Operations, and Financing and Investment.

At the sentiment annotation level, we categorize the sentiment polarity of each event, reflecting its potential impact on the target entity, into positive, negative, and neutral classes.

Building upon the EFSA annotation framework, we further introduce event trigger spans as an additional annotation field. Event trigger spans are the core sentences that indicate the occurrence of an event and serve as the key elements for our joint modeling.

C. Annotation Platform and Process

To improve annotation efficiency and ensure consistency, we developed a dedicated annotation platform tailored for event-level financial sentiment analysis. The platform supports news text visualization, event type selection, sentiment polarity assignment, and event trigger spans highlighting, providing annotators with a unified and user-friendly interface.

The annotation process was conducted by five domain experts with experience in financial text analysis. Each annotator initially labeled 1,000 data samples. During this initial phase, we collected statistics on inconsistencies and ambiguities in their annotations. Based on these observations, we refined a comprehensive annotation guideline addressing events and contexts prone to ambiguity or disagreement, which was then applied to the full annotation process.

For each news article, two annotators independently performed event identification, event type classification, sentiment polarity labeling, and event trigger spans annotation. In cases of disagreement on any annotation item, a third annotator reviewed the instance and made the final decision. Through this annotation process, we obtained 9,120 high-quality annotated financial news articles from the initial pool, covering a total of 18,158 annotated events.

D. Inter-Annotator Agreement

To assess the reliability of the annotations, we computed inter-annotator agreement using Fleiss' Kappa [27] for both event-level sentiment polarity and event trigger spans annotations.

For event-level sentiment polarity, the overall Fleiss' κ score is 0.7881, indicating a substantial level of agreement among annotators. This result suggests that the annotators shared a consistent understanding of sentiment judgment at the event level.

For event trigger spans annotation, since trigger spans are span-based annotations and may involve minor boundary variations across annotators, we employed a token-level agreement analysis. This approach captures partial agreement when annotators identify the same trigger with slightly different span boundaries (e.g., "significantly increased" vs. "increased"). Under this evaluation setting, the Fleiss' κ score for event trigger spans reaches 0.81, indicating almost perfect agreement.

These results demonstrate that the dataset exhibits strong and reliable annotation consistency for both sentiment polarity and event trigger identification, providing a trustworthy basis for subsequent model training and evaluation.

E. Dataset Analysis

We conducted a statistical analysis of the annotated data, as shown in Fig. 2. Fig. 2(a) illustrates that the dataset covers all predefined event categories in the event-level financial sentiment analysis framework, reflecting the diversity of financial events in real-world financial news. Among them, Business Operations events account for the largest proportion, approximately 34%, which is consistent with the characteristics of real-world financial news.

Fig. 2(b) presents the distribution of textual length characteristics. The average length of a news article is approximately 437 characters, while the average length of an event trigger span is around 55 characters, providing sufficient contextual information for event understanding and sentiment polarity assessment.

Fig. 2(c) shows the distribution of the number of events contained in each news article. The dataset includes both single-event instances and multi-event instances containing up to four events, providing strong support for studying event recognition and event-level sentiment analysis in complex financial texts where multiple events coexist.

Fig. 2(d) and (e) show the distribution of event-level sentiment polarity. Positive events mainly involve improved business performance, stock price increases, and intellectual property–related activities; negative events are primarily associated with operational risks and compliance issues; neutral events are mostly related to information disclosure and executive changes.

IV. Experiments

In this section, we benchmark FinEvent-Trigger with some widely used language models. We first introduce the task and experimental setup, then describe the benchmark methods and detail our proposed framework EFSA-TD, and finally present the evaluation metrics.

A. Tasks and Settings

We evaluate our dataset and method on four financial sentiment analysis tasks with increasing granularity:

- **Entity-Level FSA:** Predicts a pair **(entity, sentiment)**, assigning an overall sentiment label to each financial entity mentioned in the text, without distinguishing underlying events.
- **C-EFSA:** Predicts a triplet **(entity, coarse-grained event type, sentiment)**, introducing event-level modeling with coarse-grained financial event categories.
- **F-EFSA:** Predicts a quadruple **(entity, coarse-grained event type, fine-grained event type, sentiment)**, requiring deeper understanding of event categories and specific event attributes.
- **T-EFSA:** Predicts a quintuple **(entity, coarse-grained event type, fine-grained event type, sentiment, event trigger span)**, grounding sentiment predictions in explicit evidence. This is the most fine-grained task and the main focus of our dataset and method.

Across all tasks, sentiment analysis is uniformly formulated as a three-class classification problem, with sentiment labels defined as positive, negative, and neutral.

B. Baseline Models

In our experiments, we select a diverse set of representative LLMs as baseline methods, covering both closed-source and open-source categories of models. This design aims to systematically compare the modeling capacity and adaptability of different model families for financial sentiment analysis tasks.

For closed-source models, we evaluate GPT-5.1[3] [28], GLM-4.7[4] [29], Qwen3-Max[5] [30], and DeepSeek-R1[6] [31]. All experiments are conducted by querying the official API interfaces provided by their respective vendors [32] [33]. Under this setting, model parameters are kept frozen, and all predictions are generated based on carefully designed prompts. This allows us to assess the generalization ability of closed-source LLMs under zero-shot or few-shot conditions for financial sentiment analysis.

For open-source models, we adopt LLaMA-3.1-8B-Instruct [34], ChatGLM3-6B-Chat [35], Baichuan2-13B-Chat [36] and Qwen2.5-7B-Instruct [37]. Experiments are conducted on 4 RTX 3090 and 2 RTX 4090 GPUs. Fine-tuning uses the AdamW optimizer with a learning rate of 2×10^{-5} for 3 epochs, incorporating gradient clipping and dynamic learning rate scheduling. We randomly split the data into 8:2 for training and testing.

C. EFSA-TD Framework

We propose EFSA-TD (Event-Level Financial Sentiment Analysis with Trigger Detection), a structured prompting framework that guides LLMs to jointly detect event triggers and perform financial event classification and sentiment analysis. Our framework leverages step-by-step reasoning and self-verification mechanisms to ensure the validity, consistency, and completeness of extracted event triggers, event categories, and sentiment predictions.

To systematically guide the model, we design a structured prompt P that enforces explicit constraints on company-related

[3]https://platform.openai.com
[4]https://open.bigmodel.cn/dev/api
[5]https://qwen.ai/home
[6]https://api-docs.deepseek.com

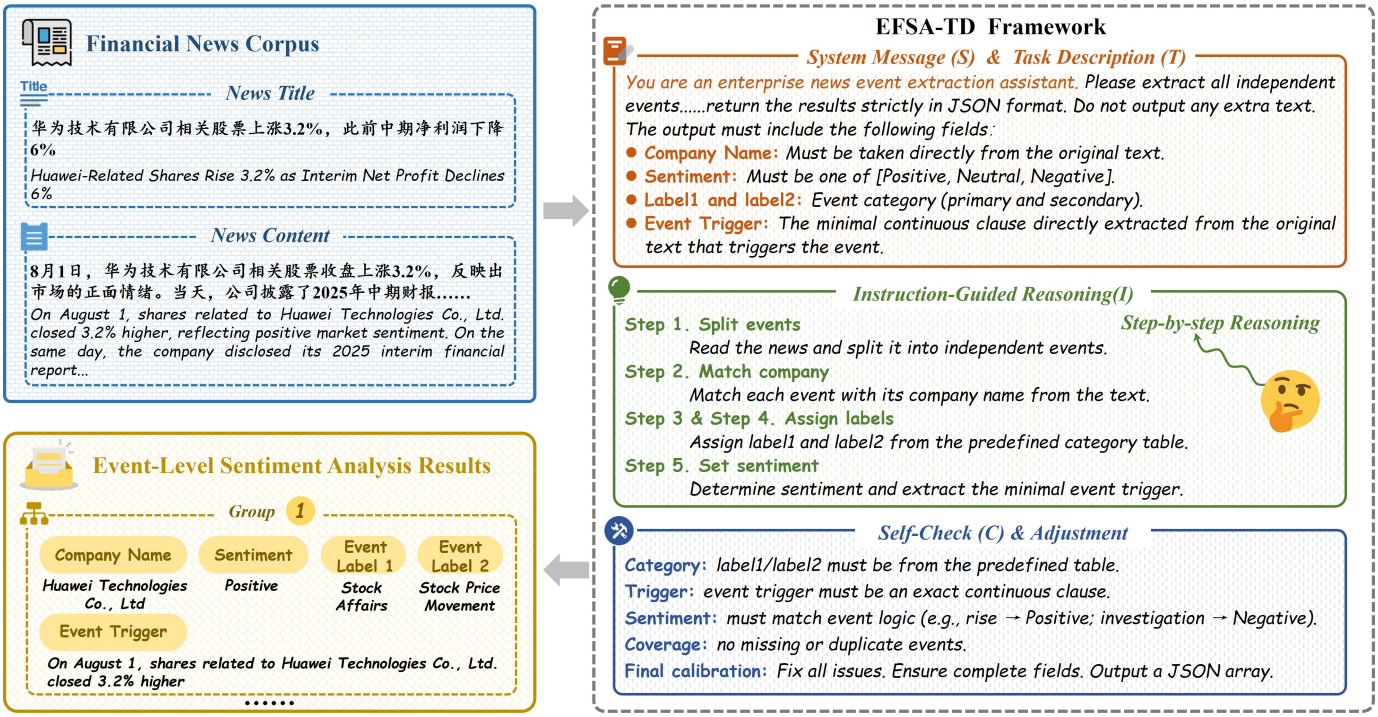

Fig. 3. The overall workflow of the proposed framework EFSA-TD.

event detection, event classification, and sentiment prediction. Formally, the framework is defined as a four-tuple:

$$P = \langle S, T, I, C \rangle.$$

1) System Message (S): The system message S defines the model's expert role as a financial event sentiment analysis assistant and specifies that outputs should be produced in a predefined structured format.

2) Task Description (T): Given a financial news document x containing a set of financial entities $e \in E$, we formulate event-level financial sentiment analysis as a structured joint prediction task. For each entity-related event, the model predicts a five-tuple

$$y = (e, c, f, s, t),$$

where c denotes the coarse-grained event type, f the fine-grained event type, $s \in \{\text{Positive}, \text{Negative}, \text{Neutral}\}$ the sentiment polarity, and t the event trigger span that provides explicit textual evidence for the event.

3) Instructions (I): The instructions I consist of a carefully designed five-step reasoning sequence that guides the LLMs through the analysis. Formally,

$$I = \langle I_1, I_2, I_3, I_4, I_5 \rangle,$$

where each step is defined as follows:

- I_1: Instruct the model to read the news document x and decompose it into all independent company-related events by selecting the trigger span t. This step corresponds to the factor $P(t \mid x)$.
- I_2: For each extracted trigger span t, match it to the corresponding company e involved in the event. This step begins the conditional prediction of event attributes given the trigger, i.e., $P(e \mid t, x)$.
- I_3: Given the trigger t and company e, determine the coarse-grained event category c. This corresponds to the factor $P(c \mid t, e, x)$ in the conditional prediction.
- I_4: Based on the trigger t, company e, and coarse-grained category c, identify the fine-grained event subtype f, i.e., $P(f \mid t, e, c, x)$.
- I_5: Finally, conditioned on the trigger t, company e, and event types (c, f), predict the sentiment polarity s, corresponding to $P(s \mid t, e, c, f, x)$. This step completes the joint prediction according to

$$P(e, c, f, s, t \mid x) = P(t \mid x)\, P(e, c, f, s \mid t, x).$$

4) Self-Check (C): A key component of our framework is the final self-check mechanism, which asks the model to verify four critical conditions before producing its final predictions: (1) **Event Category Consistency Check**: Both coarse-grained (c) and fine-grained (f) event types must strictly belong to the predefined taxonomy, and f must be semantically consistent with c. (2) **Trigger Span Integrity Check**: Every predicted trigger span (t) must be a contiguous substring of the original document x. (3) **Sentiment Alignment Check**: The predicted sentiment polarity (s) must align with the event expressed by t (e.g., a stock price increase is positive, a regulatory investigation is negative). (4) **Event Coverage Check**: For each entity (e), all independent events must be fully extracted, without omission or duplication.

TABLE II
THE RESULTS OF EFSA-TD AND BASELINE MODELS. THE BEST RESULTS ARE IN BOLD; SECOND-BEST RESULTS ARE UNDERLINED.

Settings	Models	Entity-Level FSA	C-EFSA	F-EFSA	T-EFSA
• *Closed-Source Models*					
zero-shot	GPT-5.1	74.09	65.54	61.26	49.73
	GLM-4.7	67.36	55.65	52.51	43.24
	Qwen3-Max	76.55	63.48	60.87	49.94
	DeepSeek-R1	63.19	49.20	44.58	37.43
3-shot	GPT-5.1	75.87	68.96	64.01	51.33
	GLM-4.7	69.12	58.26	56.84	45.95
	Qwen3-Max	78.24	66.83	<u>64.13</u>	53.21
	DeepSeek-R1	63.98	51.85	46.15	40.07
• *Open-Source Models*					
LoRA fine-tune	Llama-3.1-8B-Instruct	69.78	56.83	48.18	44.21
	ChatGLM3-6B-Chat	70.73	62.48	57.11	47.18
	Baichuan2-13B-Chat	84.47	67.97	61.79	51.06
	Qwen2.5-7B-Instruct	83.03	67.65	61.38	54.57
CoT	Llama-3.1-8B-Instruct	70.42	57.51	48.96	44.85
	ChatGLM3-6B-Chat	71.21	63.05	57.74	47.92
	Baichuan2-13B-Chat	<u>85.04</u>	68.61	62.47	51.69
	Qwen2.5-7B-Instruct	83.78	68.32	63.09	<u>55.23</u>
EFSA-TD	Llama-3.1-8B-Instruct	71.88	58.18	49.64	42.47
	ChatGLM3-6B-Chat	73.43	63.47	59.23	49.71
	Baichuan2-13B-Chat	**85.42**	<u>69.13</u>	63.52	52.63
	Qwen2.5-7B-Instruct	84.08	**70.26**	**64.56**	**55.34**

D. Evaluation Metrics

We use the Micro-F1 score as the primary evaluation metric to assess model performance across different financial sentiment analysis tasks. For each event, a prediction is considered correct if and only if all fields required by the task are correctly predicted.

In the event trigger span prediction task for event-level tasks, we adopt a Partial Matching (PM) strategy to handle boundary variations. Specifically, a predicted trigger span is considered correct if its character-level overlap with the gold trigger span exceeds 0.6.

V. RESULTS

Table II presents the performance comparison of different models across four financial sentiment analysis tasks with increasing granularity: Entity-Level FSA, C-EFSA, F-EFSA, and T-EFSA. Overall, model performance consistently decreases as task granularity becomes finer, reflecting the higher difficulty of fine-grained event-level financial sentiment analysis and event trigger detection.

A. Model Family Comparison

Different model families exhibit distinct performance characteristics. The LLaMA-3.1-8B-Instruct performs the worst across all tasks, while Qwen2.5-7B-Instruct performs consistently well, achieving significant performance improvements, especially under the enhanced EFSA-TD framework proposed in this work. The Baichuan2-13B-Chat and ChatGLM3-6B-Chat show similar trends, indicating that the intrinsic characteristics of different models and their pre-training data have a significant impact on performance in event-level financial sentiment analysis tasks.

B. Technique Comparison

In experiments on open-source models, the proposed EFSA-TD framework consistently outperforms direct LoRA fine-tuning and standard Chain-of-Thought (CoT) methods. For example, the Qwen2.5-7B-Instruct model equipped with the EFSA-TD framework achieves the best results on the C-EFSA, F-EFSA, and T-EFSA tasks, with F1 scores of 70.26, 64.56, and 55.34, respectively. These results confirm that the proposed structured prompting framework with joint event trigger spans can more effectively guide the model to identify the sentiment of financial events.

C. Closed-Source vs. Open-Source Models

Closed-source models (e.g., GPT-5.1 and Qwen3-Max) achieve strong performance across all evaluation metrics under zero-shot and few-shot settings. Nevertheless, the best-performing open-source models (e.g., Qwen2.5-7B-Instruct and Baichuan2-13B-Chat equipped with the EFSA-TD framework) outperform them, indicating that task-specific fine-tuning on dedicated datasets enables open-source methods to surpass commercial systems in financial sentiment analysis tasks [38].

D. Ablation Study

To evaluate the contributions of the event trigger and different components in the EFSA-TD framework, we conducted two ablation studies.

Module Ablation (Table III): Removing both instruction-driven reasoning and the final self-check module (w/o I & C) leads to consistent performance drops, e.g., Entity-Level FSA decreases from 84.08 to 83.06, and T-EFSA from 55.34 to 54.61. Removing only the self-check module (w/o C) also results in a performance drop, indicating that both the instruction-driven reasoning and the self-check module are equally important for the model's performance.

TABLE III
ABLATION STUDY OF THE EFSA-TD FRAMEWORK ON **QWEN2.5-7B-INSTRUCT**.

Settings	Entity-Level FSA	C-EFSA	F-EFSA	T-EFSA
Full	**84.08**	**70.26**	**64.56**	**55.34**
w/o C	83.50 (-0.58)	69.05 (-1.21)	62.75 (-1.81)	54.84 (-0.50)
w/o I & C	83.06 (-1.02)	67.70 (-2.56)	61.40 (-3.16)	54.61 (-0.73)

Event Trigger Ablation (Table IV): In the LoRA fine-tuning setup, we fine-tune open-source models without introducing any additional methods, and find that removing event trigger spans (Without Trigger) leads to performance drops across all EFSA sub-tasks. Adding trigger spans (With Trigger) significantly improves results; for example, Qwen2.5 gains 2.36 points on Entity-Level FSA and 2.56 points on C-EFSA. The improvement varies across models, with Baichuan2 and Qwen2.5 benefiting the most, while Llama3.1 shows relatively smaller gains. These results indicate that event triggers help enhance the model's ability to capture event-related information, not only improving the interpretability of financial sentiment but also significantly benefiting the performance of financial sentiment analysis.

TABLE IV
ABLATION STUDY ON EVENT TRIGGER SPANS

Models	Entity-Level FSA	C-EFSA	F-EFSA	T-EFSA
		• *Without Trigger*		
Llama3.1	68.81	55.58	47.12	*
ChatGLM3	69.36	60.91	55.95	*
Baichuan2	82.52	65.81	59.74	*
Qwen2.5	80.67	65.09	59.11	*
		• *With Trigger*		
Llama-3.1	69.78 (+0.97)	56.83 (+1.25)	48.18 (+1.06)	44.21
ChatGLM3	70.73 (+1.37)	62.48 (+1.57)	57.11 (+1.16)	47.18
Baichuan2	**84.47** (+1.95)	**67.97** (+2.16)	**61.79** (+2.05)	51.06
Qwen2.5	83.03 (+2.36)	67.65 (+2.56)	61.38 (+2.27)	**54.57**

E. Error Analysis

Through detailed analysis of model predictions, we identify several representative error patterns, as shown in Table V.

- **Event Entity Identification Errors:** When multiple related companies appear in the same article, the model may confuse the attribution of events. For example, in Case 1, the model incorrectly assigns an event belonging to a subsidiary to its parent company.
- **Sentiment Misclassification:** When an event contains multiple sentiment cues, the model may capture only a single sentiment. For example, in Case 2, the model overlooks the primary sentiment, leading to incorrect sentiment classification.
- **Trigger Boundary Errors:** Models often struggle to precisely identify trigger boundaries in complex sentences. For example, in Case 3, when a single event is expressed across multiple connected clauses, the model may sometimes overextend or underextend the trigger span.

TABLE V
CASE STUDIES OF ERROR EXAMPLES

Case study		
Case#1	Content	华为集团近日发布称，其子公司华为终端有限公司获得一项新型折叠屏手机的发明专利。 Huawei Group recently announced that its subsidiary, Huawei Device Co., Ltd., has obtained a patent for a new type of foldable smartphone.
	Ground Truth	{company: 华为终端有限公司, label1: 经营, label2: 知识产权, sentiment: Positive}
	Prediction	{company: 华为集团, label1: 经营, label2: 知识产权, sentiment: Positive} ✗
Case#2	Content	山西汾酒实现营业收入165.23亿元，同比7.72%；净利润66.48亿元，同比-3.3%。 Shanxi Fenjiu achieved operating revenue of CNY 16.523 billion, up 7.72% year-on-year, and a net profit of CNY 6.648 billion, down 3.3% year-on-year.
	Ground Truth	{company: 山西汾酒, label1: 财务, label2: 利润公布, sentiment: Negative}
	Prediction	{company: 山西汾酒, label1: 财务, label2: 利润公布, sentiment: Positive} ✗
Case#3	Content	长安汽车近日发布公告称，公司下半年将加大产能，预计生产新能源车销量50万辆，而公司旗下物流中心正在进行年度安全检查。 Changan Automobile recently announced that it will expand production capacity in the second half of the year, with an expected production of 500,000 new energy vehicles, while the company's logistics center is undergoing its annual safety inspection.
	Ground Truth	{event_trigger: 公司下半年将加大产能，预计生产新能源车销量50万辆}
	Prediction	{event_trigger: 公司下半年将加大产能，预计生产新能源车销量50万辆，而公司旗下物流中心正在进行年度安全检查} ✗

VI. CONCLUSION

This paper introduces FinEvent-Trigger, a benchmark dataset for fine-grained event-level financial sentiment analysis with trigger spans, and the EFSA-TD framework, which grounds sentiment predictions in explicit textual evidence. Experiments with multiple LLMs show that jointly modeling trigger spans consistently improves the performance of sentiment analysis. Ablation studies confirm that trigger supervision provides essential semantic support rather than merely auxiliary signals. Our work aims to provide a foundation for understanding both the sentiment polarities of financial events and the reasons behind them, making event-level financial

sentiment analysis more transparent and traceable. We hope this work advances methods for financial sentiment analysis.

ACKNOWLEDGMENT

This work is mainly supported by Henan Provincial Science and Technology Research Project (No.262102211057) and the Key Program of the National Natural Science Foundation of China (NSFC) (No.U23A20316).

REFERENCES

[1] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *The Journal of finance*, vol. 62, no. 3, pp. 1139–1168, 2007.

[2] B. Weng, M. A. Ahmed, and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources," *Expert systems with applications*, vol. 79, pp. 153–163, 2017.

[3] C.-J. Wang, M.-F. Tsai, T. Liu, and C.-T. Chang, "Financial sentiment analysis for risk prediction," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013, pp. 802–808.

[4] J. Pei, S. Vadlamannati, L.-K. Huang, D. Preoţiuc-Pietro, and X. Hua, "Modeling and detecting company risks from news," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, 2024, pp. 63–72.

[5] A. Huang, H. Wang, and Y. Yang, "Finbert : A large language model for extracting information from financial text," *Contemporary Accounting Research*, 2022.

[6] C. Kearney and S. Liu, "Textual sentiment in finance: A survey of methods and models," *International Review of Financial Analysis*, vol. 33, pp. 171–185, 2014.

[7] L. Deng and J. Wiebe, "Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 179–189.

[8] A. Mahajan, L. Dey, and S. M. Haque, "Mining financial news for major events and their impacts on the market," in *2008 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, vol. 1. IEEE, 2008, pp. 423–426.

[9] A. Sinha and T. Khandait, "Impact of news on the commodity market: Dataset and results," in *Future of Information and Communication Conference*. Springer, 2021, pp. 589–601.

[10] M. Ebrahimi, A. H. Yazdavar, and A. Sheth, "Challenges of sentiment analysis for dynamic events," *IEEE Intelligent Systems*, vol. 32, no. 5, pp. 70–75, 2017.

[11] A. Sorescu, N. L. Warren, and L. Ertekin, "Event study methodology in the marketing literature: an overview," *Journal of the Academy of Marketing Science*, vol. 45, no. 2, pp. 186–207, 2017.

[12] M. Baker and J. Wurgler, "Investor sentiment and the cross-section of stock returns," *The journal of Finance*, vol. 61, no. 4, pp. 1645–1680, 2006.

[13] S. W. Chan and M. W. Chong, "Sentiment analysis in financial texts," *Decision Support Systems*, vol. 94, pp. 53–64, 2017.

[14] K. Du, F. Xing, R. Mao, and E. Cambria, "Financial sentiment analysis: Techniques and applications," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–42, 2024.

[15] J. Zhao, J. Tu, H. Du, and N. Xue, "Media attitude detection via framing analysis with events and their relations," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 17 197–17 210.

[16] S. Zhu, C. He, H. Liu, P. Dong, H. Zhao, Y. Yan, Y. Jia, H. Zan, and M. Peng, "Silc-efsa: Self-aware in-context learning correction for entity-level financial sentiment analysis," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 4980–4992.

[17] K. Du, F. Xing, and E. Cambria, "Incorporating multiple knowledge sources for targeted aspect-based financial sentiment analysis," *ACM Transactions on Management Information Systems*, vol. 14, no. 3, pp. 1–24, 2023.

[18] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 782–796, 2014.

[19] K. Cortis, A. Freitas, T. Daudert, M. Huerlimann, M. Zarrouk, S. Handschuh, and B. Davis, "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)*, 2017, pp. 519–535.

[20] A. Sinha, S. Kedas, R. Kumar, and P. Malo, "Sentfin 1.0: Entity-aware sentiment analysis for financial news," *Journal of the Association for Information Science and Technology*, vol. 73, no. 9, pp. 1314–1335, 2022.

[21] Y. Tang, Y. Yang, A. Huang, A. Tam, and J. Tang, "Finentity: Entity-level sentiment classification for financial texts," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 15 465–15 471.

[22] A. Petrescu, C.-O. Truică, E.-S. Apostol, and A. Paschke, "Edsa-ensemble: an event detection sentiment analysis ensemble architecture," *IEEE Transactions on Affective Computing*, vol. 16, no. 2, pp. 555–572, 2024.

[23] T. Chen, Y. Zhang, G. Yu, D. Zhang, L. Zeng, Q. He, and X. Ao, "Efsa: Towards event-level financial sentiment analysis," *arXiv preprint arXiv:2404.08681*, 2024.

[24] A. Attri, A. Attri, P. Bhattacharyya, S. Banerjee, A. Patil, M. Chelliah, and N. Garera, "Why we feel what we feel: Joint detection of emotions and their opinion triggers in e-commerce," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 5515–5532.

[25] S. Yang, D. Feng, L. Qiao, Z. Kan, and D. Li, "Exploring pre-trained language models for event extraction and generation," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 5284–5294.

[26] M. Dredze, P. Kambadur, G. Kazantsev, G. Mann, and M. Osborne, "How twitter is changing the nature of financial news discovery," in *proceedings of the second international workshop on data science for macro-modeling*, 2016, pp. 1–5.

[27] Fleiss and L. Joseph, "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[28] R. A. PILIANG, E. A. Firdaus *et al.*, "Operationalizing model context protocol for multi-platform academic search: A comparative study of llm grounding," *NUANSA INFORMATIKA*, vol. 20, no. 1, pp. 59–71, 2026.

[29] Z. Lin, S. Piao, and A. Wang, "Comparative evaluation of chatgpt and chatglm performance in response to common queries on pediatric atopic dermatitis," *Pediatric Dermatology*, 2025.

[30] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[31] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[32] M. Li, Y. Zhao, B. Yu, F. Song, H. Li, H. Yu, Z. Li, F. Huang, and Y. Li, "Api-bank: A comprehensive benchmark for tool-augmented llms," *arXiv preprint arXiv:2304.08244*, 2023.

[33] W. Chen, Z. Li, and M. Ma, "Octopus: On-device language model for function calling of software apis," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, 2025, pp. 329–339.

[34] Z. Mai, J. Zhang, Z. Xu, and Z. Xiao, "Financial sentiment analysis meets llama 3: A comprehensive analysis," in *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)*, 2024, pp. 171–175.

[35] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao *et al.*, "Chatglm: A family of large language models from glm-130b to glm-4 all tools," *arXiv preprint arXiv:2406.12793*, 2024.

[36] A. Yang, B. Xiao, B. Wang, B. Zhang, C. Bian, C. Yin, C. Lv, D. Pan, D. Wang, D. Yan *et al.*, "Baichuan 2: Open large-scale language models," *arXiv preprint arXiv:2309.10305*, 2023.

[37] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu *et al.*, "Qwen2. 5-coder technical report," *arXiv preprint arXiv:2409.12186*, 2024.

[38] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.