No Language is an Island: Unifying Chinese and English in Financial Large Language Models, Instruction Data, and Benchmarks

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) have significantly advanced financial analysis, their scope has predominantly been limited to monolingual applications, with bilingual Chinese-English capabilities largely unexplored. To bridge this gap, we present ICE-PIXIU, integrating the ICE-INTENT model and ICE-FLARE benchmark for bilingual financial analysis. ICE-PIXIU uniquely incorporates a range of Chinese instruction tasks, alongside translated and original English datasets, extending the reach and depth of bilingual financial modeling. Our extensive analysis reveals that integrating these bilingual datasets, especially translation tasks and original English data, not only enhances the model's linguistic adaptability but also deepens its analytical acumen in financial contexts. ICE-INTENT, in particular, demonstrates a marked improvement over general-domain LLMs in bilingual settings, showcasing the substantial impact of rich bilingual data on the precision and effectiveness of financial NLP.

1 Introduction

Inspired by the success of general-domain large language models (LLMs) (Chang et al., 2023), the exploration of LLMs in the financial sector is gaining momentum. Starting with the broad capabilities of models like GPT-3.5 (Kalyan, 2023) and GPT-4 (OpenAI, 2023), the focus has shifted towards more specialized models such as BloombergGPT (Wu et al., 2023), which tackles the complex nuances of financial language and concepts. Subsequently, open-sourced models like PIXIU (Xie et al., 2023), FinGPT (Yang et al., 2023b) and InvestLM (Yang et al., 2023c) emerged, concentrating on various financial tasks. Notably, PIXIU marks an important advancement in overcoming the hurdles associated with open-source comprehensive instruction tuning data and evaluation benchmarks in this field.

Beyond the efforts in English-centric models, Chinese financial LLMs (FinLLMs) have also made notable advances. XuanYuan2.0 (Zhang, 2023) exemplifies this trend, focusing on the intricacies of Chinese financial language. This has been followed by models such as DISC-FinLLM (Chen et al., 2023) and CFGPT (Li et al., 2023a), which further the capabilities in Chinese financial analysis and reasoning. Recently, PanGu-π (Wang et al., 2023) continue pretrained a financial LLMs which improves the performance on Chinese financial examinations.

However, there is a significant challenge in their application and assessment in bilingual contexts, especially regarding model development, instruction dataset diversity, and evaluation methodologies in English and Chinese. As Table 1 illustrates, most FinLLMs are tailored exclusively for either English or Chinese, which underscores a significant shortfall in models proficient in both languages.

This issue is further compounded in the domain of instruction datasets. While there are tasks encompassing classification, extraction, and prediction, these datasets, particularly for Chinese, show a notable lack of diversity in each task category. Furthermore, the original English datasets and their subsequent translations into Chinese, are often overlooked in the development of comprehensive bilingual instruction datasets.

Correspondingly, the evaluation methodologies for these models tend to mirror these linguistic limitations, focusing mainly on monolingual assessments and failing to comprehensively evaluate bilingual proficiency. To address these challenges in the Financial LLM sector, there is a critical need for more diverse Chinese instruction datasets in both classification, extraction, and prediction, alongside a greater focus on the utilization and translation of original English datasets into Chinese.

To address this issue, we introduce ICE-PIXIU,

Model	Size	Language	Evaluation zh	en	oft	Open Source Model	Data	Release Date
BloombegGPT (Wu et al., 2023)	50B	en	0	5	0	✔	✗	03/30/23
InvestLM (Yang et al., 2023c)	65B	en	0	9	0	✔	✗	09/14/23
FinGPT (Yang et al., 2023b)	7/13B	en	0	6	0	✔	✗	11/10/23
PIXIU (Xie et al., 2023)	7/13B	en	0	12	3	✔	✔	06/01/23
XuanYuan2.0 (Zhang, 2023)	176B	zh	0	1	0	✗	✗	05/19/23
CFGPT (Li et al., 2023a)	7B	zh	6	0	0	✔	✗	09/01/23
DISC-FinLLM (Chen et al., 2023)	13B	zh	7	0	0	✔	✗	10/24/23
PanGu-π (Wang et al., 2023)	1/7B	zh	7	0	0	✗	✗	12/27/23
ICE-PIXIU	7B	zh,en	25	12	3	✔	✔	12/10/23

Table 1: Overview of various FinLLMs. We outline their model size, language proficiency, evaluation counts in Chinese [zh], English [en] and out-of-field [en-oft] tasks, open source, and release date.

a comprehensive framework that features ICE-INTENT and ICE-FLARE, the first cross-lingual bilingual financial model and evaluation benchmark, respectively. This framework is characterized by several key attributes:

- **Bilingual Proficiency**: ICE-INTENT, a component of ICE-PIXIU, demonstrates outstanding bilingual proficiency in English and Chinese, essential for global financial data processing.
- **Diverse Chinese Datasets**: ICE-PIXIU addresses gaps in Chinese financial NLP by incorporating a variety of Chinese classification, extraction, and prediction tasks, enhancing training and performance.
- **Incorporation of Translation and English Data**: The framework expands its capabilities by including translation tasks and English datasets, bolstering its bilingual training and application.
- **Cross-Lingual Evaluation with ICE-FLARE**: ICE-PIXIU introduces ICE-FLARE, a rigorous cross-lingual evaluation tool, ensuring consistent model performance in diverse linguistic contexts.
- **Open-Source Contribution**: With its open-access approach, ICE-PIXIU offers its resources to the research community, fostering collaborative advancement in financial NLP.

In the creation of ICE-INTENT, we meticulously gathered 40 datasets, which consisted of 1,185,076 raw data, 563,151 instruction data, and 95,091 evaluation data, strategically covering a wide range of financial tasks. This ensemble, as detailed in Table 1, comprises 9 datasets each in Chinese classification (zh-CLS) and English (en), and 8 each in Chinese information retrieval (zh-IR) and datasets translated from English (zh-TRAN). This diverse array not only underscores our method's commitment to linguistic breadth but also to task-specific depth, ensuring ICE-INTENT's proficiency across various financial scenarios. Complementing this, ICE-FLARE, our cross-lingual evaluation benchmark, incorporates 40 evaluation tasks—25 in Chinese and 15 in English. This robust framework highlights our approach's unique strength in providing comprehensive cross-lingual consistency assessments, a crucial aspect for bilingual financial language models.

In our evaluation conducted using ICE-FLARE, we observed distinct performance patterns of our financial language model in comparison with advanced general-domain LLMs. The results firstly emphasized the model's enhanced capabilities in cross-lingual tasks, particularly in handling complex and lengthy text, underscoring the critical role of bilingual data integration in financial language modeling. Subsequently, the model exhibited proficiency in translation tasks, highlighting its adeptness at managing linguistic variations crucial for global financial contexts. Finally, our model outperformed general LLMs in tasks specifically fine-tuned for the financial domain but faced challenges in less familiar tasks within ICE-FLARE. This delineates the specialized nature and unique challenges of the financial sector, reinforcing the need for tailored financial language models.

2 Method

2.1 Raw Data Construction

In this section, we introduce our bilingual Chinese-English dataset, designed for multi-task bilingual

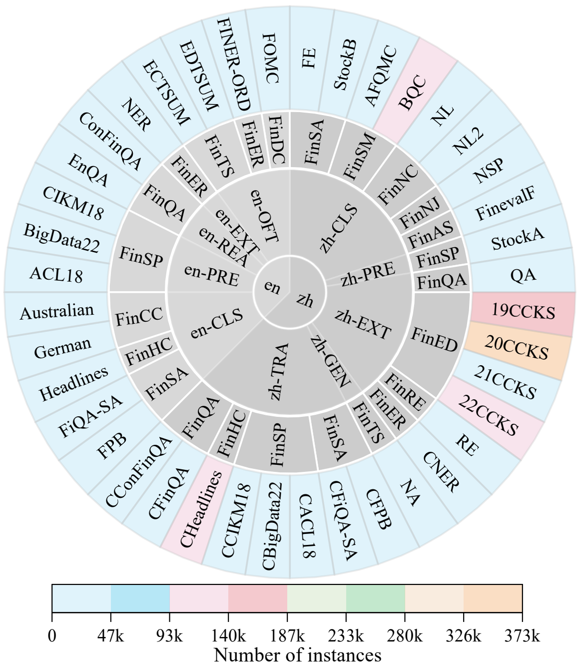

Figure 1: Sunburst chart of datasets with Chinese-English bilingual multiple financial specific tasks.

instruction tuning and evaluation benchmarks, featuring 34 data types across 13 financial tasks for applications like text analysis, generation, and prediction, derived from real-world financial contexts. This dataset stands out by utilizing expert-curated, high-quality sources, offering cost-efficiency over ChatGPT-based methods without usage restrictions, and providing varied data formats and multilingual support for flexible task adaptation. Details of our data are in Appendix 4, Table 3.

2.1.1 Chinese Financial Datasets

We present a comprehensive dataset catering to two pivotal aspects of Chinese financial text analysis including text classification (zh-CLS), information extraction (zh-EXT), text generation (zh-GEN), prediction (zh-PRE).

zh-CLS. For zh-CLS, our dataset covers essential tasks such as sentiment analysis (FinSA), semantic matching (FinSM), news classification (FinNC), negative judgment (FinNJ), and answer selection (FinAS). ***1) Sentiment Analysis.*** It is the process of analyzing and determining the sentiment expressed in financial texts (Sohangir et al., 2018; Araci, 2019). We've compiled well-known datasets for financial sentiment analysis including the FE (Lu et al., 2023) and the StockB [1] dataset. ***2) Negative Judgment.*** Financial news negative judgment focuses on identifying finance and economics-related negativity, unlike broader financial sentiment analysis. We use the FinNSP dataset (Lu et al., 2023), with social media as its source like the FE dataset, labeling data as "是" (indicating a match) or "否" (indicating a mismatch). ***3) Semantic matching.*** Financial semantic matching determines if financial texts or sentences are semantically similar, using two datasets: the Bank Question Corpus (BQC) (Chen et al., 2018) from Chinese bank customer logs and the Ant Financial Question Matching Corpus (AFQMC) (Xu et al., 2020) from the Alipay competition, both utilizing binary classification for match identification. ***4) News Classification.*** Financial news classification involves assigning financial articles to categories based on their content, using the FinNL (Lu et al., 2023) dataset from sources like Sina and Tencent Finance. We processed this dataset into binary and multi-class labeled data, referred to as NL and NL2, respectively. ***5) Answer Selection.*** Financial answer selection, a task requiring the choice of the correct option from "A", "B", "C", "D" based on a financial context, assesses financial reasoning and decision-making skills. The FinEval dataset (Zhang et al., 2023a), comprises 4,661 questions across finance, economy, accounting, and certification categories, from which we've curated data for ten key finance subjects as FinEvalF.

For **zh-EXT**, our dataset includes question answering (FinQA), named entity recognition (FinER), relation extraction (FinRE) and event detection (FinED). ***1) Question Answering.*** This task aims to extract accurate answers from the given text to answer questions related to an event mentioned in the text. We have collected a Chinese dataset named QA (Han et al., 2022). ***2) Relationship Extraction.*** The Financial RE task identifies connections between entities using the FinRE (Lu et al., 2023) dataset, which includes financial news and entity pairs across 44 categories (e.g. "合作(cooperation)", "持股(shareholding)", etc.). ***3) Name Entity Recognition(NER).*** The NER task identifies financial entities using the comprehensive CNER Chinese dataset (Jia et al., 2020). ***4) Event Detection.*** Financial event detection involves recognizing and comprehending specific incidents beyond industry categorization. We collected four datasets [2] from China Conference on Knowledge

[1] https://huggingface.co/datasets/kuroneko5943/stock11

[2] https://www.biendata.xyz/competition/{ccks_

Graph and Semantic Computing （CCKS). Unlike similar work (Lei et al., 2023) that focuses on detecting 254 different event types in four categories, our task further includes the task of identifying causal relationships between events (原因事件类型/Cause event type→结果事件类型/Result event type).

For **zh-PRE**, we focus on the ***Stock Prediction*** (FinSP) task. It involves categorizing stocks into three categories ("表现不佳(underperforming)","跑赢大盘(outperforming the market)","中性(neutral)") by combining stock data and financial news context. We collected the StockA dataset (Zou et al., 2022), features stock-specific news and factors for China's A-shares market, including minute-level price history, setting it apart from similar datasets (Xu and Cohen, 2018; Zhou et al., 2021) by offering comprehensive data on individual stocks. For **zh-GEN**, we include the ***Text Summarization*** (FinTS) task. It aims to condense complex financial information into concise summaries. We compiled the FinNA (Lu et al., 2023) Chinese dataset, referred to as NA, sourcing content from research reports and focusing on their conclusions and abstracts as targets.

2.1.2 English Financial Datasets

For our English dataset, we utilize datasets from PIXIU (Xie et al., 2023), a comprehensive source for English financial data tailored for LLMs. The dataset includes five parts: text classification (en-CLS), information extraction (en-EXT), prediction (en-PRE), reasoning (en-REA) and out-of-field task (en-OFT). **en-CLS** covers the sentiment analysis task using the Financial Phrase Bank (FPB) and FiQA-SA dataset, the ***Headline Classification*** (FinHC) task using the Headlines dataset (Sinha and Khandait, 2021), and the ***Credit Classification*** (FinCC) task. For credit classification that focusing on predicting whether a real-world user will default or not, we employ two datasets, German (Hofmann, 1994) and Australian (Quinlan). **en-EXT** includes the NER task using the CoNLL-2003 English dataset (Alvarado et al., 2015) named as NER, annotating financial texts for LOCATION, ORGANISATION, PERSON, and MISCELLANEOUS entities. **en-PRE** includes the stock prediction (FinSP) task using three commonly-used English datasets: ACL18 (Xu and Cohen, 2018), CIKM18 (Wu et al., 2018) and BigData22 (Soun et al., 2022). **en-REA** tackles the QA task using FinQA (Chen et al., 2021) as EnQA, and ConvFinQA (Chen et al., 2022) dataset, requiring logical and mathematical reasoning. Additionally, we introduce **en-OFT** including NER, text summarization and the ***Hawkish-dovish Classification*** (FinDC) task. In NER task, we utilize the FINER-ORD (Alvarado et al., 2015) dataset, which provides labels for entities such as PERSON, LOCATION, and ORGANIZATION. For text summarization, we use two datasets: ECTSUM (Mukherjee et al., 2022) and EDTSUM (Zhou et al., 2021). ECTSUM involves extracting key sentences from texts, while EDTSUM focuses on generating fitting news headlines through abstractive summarization. FinDC aims to classify sentences from monetary policy texts into a "hawkish" or "dovish" stance, unlike standard sentiment analysis, using the FOMC (Shah et al., 2023) dataset.

2.1.3 Translation Datasets

To bolster the model's proficiency in bilingual and cross-lingual tasks, we used ChatGPT to translate eight English datasets (zh-TRA) across four tasks: stock prediction, question answering, sentiment analysis, and headline classification—into Chinese, detailed in Table 3. We fully translated datasets for text-focused tasks like sentiment analysis and headline classification. For the more complex QA and stock prediction tasks, only validation sets were translated. The sentiment analysis datasets FPB and FiQA-SA became CFPB and CFiQA-SA in Chinese, labeled with sentiments: "消极" ("Negative"), "中性" ("Neutral"), and "积极" ("Positive") of sentiment polarity. The Headlines dataset was translated to CHeadlines, with labels "是" ("Yes"), and "否" ("No"). For the QA datasets including EnQA and ConvFinQA, we write specific prompt to retain the information such as tables, stock names, time series, and other relevant data, creating new Chinese datasets labeled CEnQA and CConvFinQA. For stock prediction tasks like ACL18, CIKM18, and BigData22, we employ regular expressions to extract the stock descriptions. After that, we use similar prompt to translate and retain information such as stock names and time. Finally, we combine them with historical prices to form the Chinese datasets named as CACL18, CCIKM18, and CBigData22, respectively. We discarded error-prone or untranslatable data to maintain data quality. An example of our translation method is in the Appendix.

2019_4,ccks_2020_4_1,ccks_2021_task6_2,ccks2022_eventext}

2.2 ICE-FIND: Bilingual Financial Instruction Data

Base on the different task of raw datasets, we further construct our b**I**lingual **C**hinese-**E**nglish **F**inancial **I**nstructio**N** **D**ataset (ICE-FIND), covering 13 tasks and 36 datasets with K data samples in total, as detailed in Appendix 4, Table 3. This dataset compilation combined raw data with expertly devised prompts, generating 20-30 unique prompts for each category. We assessed these prompts through human evaluations focusing on accuracy, naturalness, and informativeness using a 1-3 scale (See Appendix 4 for details), to identify the most effective ones. These prompts were tested on various LLMs using the ChatALL platform [3], showing satisfactory performance in most cases. For English tasks, we paired each dataset with all applicable prompts, excluding EnQA and ConvFinQA tasks. For Chinese tasks, due to the extensive amount of raw data, we selected one prompt per dataset at random. The instruction examples are presented in Appendix 4). The final step involved transforming these datasets into instruction-tuning samples, comprising human-crafted instructions, input texts, and responses, formatted in the following JSON format:

```
{"id": "{data_id}",
 "conversations": [
{"from": "human", "value": "{prompt} {input}"},
{"from": "agent", "value": "{answer}"}]}
```

{data_id} represents the ID of each data sample, {prompt} is the task-specific prompt, {input} refers to input, {answer} denotes the expected answer.

2.3 ICE-INTERN: Bilingual Financial Large Language Model

Leveraging InternLM-7B, a top-performing LLM for English-Chinese bilingual tasks, we developed ICE-INTERN-7B using the ICE-FIND dataset. We fine-tuned variants of ICE-INTERN to measure the effect of different data types, creating "ICE-INTERN-CEP-7B" with classification and prediction data, "ICE-INTERN-GE-7B" with extraction, generation, and reasoning data, "ICE-INTERN-TRA-7B" with English-to-Chinese translation data, and "ICE-INTERN-full-7B" with the complete dataset. We employed QLoRA (Hu et al., 2021), a parameter-efficient tuning technique, with uniform 2048-token sequence lengths. Optimization was done using AdamW with a 5e-5 initial learning rate and 1e-5 weight decay, plus a 1% total step warmup. All models were fine-tuned for one epoch in batches of 24 on eight A100 40GB GPUs, using consistent hyperparameters across all models.

2.4 ICE-FLARE: Bilingual Financial Evaluation Benchmark

Based on BiFinID, we refer to the evaluation metrics used in FLARE for similar tasks to design our bilingual multi-task evaluation benchmark, ICE-FLARE. The tasks, metrics and compared benchmarks ([1]FLARE (Xie et al., 2023), [2]CFBenchmark (Lei et al., 2023), [3]FinCUGE (Chen et al., 2023), [4]FLUE (Sanh et al., 2022), [5]Fineval (Zhang et al., 2023a), [6]CGCE (Zhang et al., 2023b)) for ICE-FLARE are presented in Table 2. (Xie et al., 2023) has already demonstrated the superiority of FLARE compared to FLUE in English benchmark. According to Table 2, it is evident that ICE-FLARE surpasses the existing benchmarks by a significant margin in terms of both task diversity and data scalability. Even though FinCUGE has a relatively higher number of financial evaluation tasks, the difference in quantity on the evaluation dataset compared to ICE-FLARE is still close to 8 times. These public benchmarks provide valuable references and inspiration for our test construction. However, they still face challenges in evaluating multilingual FinLLMs and there are limitations in applying them to real-world financial decision-making scenarios, such as credit classification. In conclusion, we provide ICE-FLARE, which offers bilingual evaluation benchmark and a more comprehensive assessment of diversity tasks, is necessary for improving bilingual FinLLMs.

3 Experiments

We select six representative and outstanding LLMs for evaluating performance.

1) **ChatGPT**(OpenAI [4]): ChatGPT-3.5-Turbo and ChatGPT-4, developed by OpenAI, are widely recognized LLMs with versatile functionalities. 2) **LLaMa** (Touvron et al., 2023): LLaMa and advanced Llama2 are powerful LLMs developed by Meta AI, with parameter sizes ranging from 7 billion to 650 billion. 3) **Baichuan** (Yang et al., 2023a): Baichuan is a LLM developed by Baichuan-inc for Chinese and English NLP tasks, available in Baichuan-7B and Baichuan-7B-Chat variants. 4) **ChatGLM** (Du et al., 2021):ChatGLM is an advanced bilingual LLM developed

[3] https://github.com/sunner/ChatALL

[4] Https://www.openai.com/chatgpt

Specific Task		Metric	Lang	Data	Test	Cover [1]	[2]	[3]	[4]	[5]	[6]
Sentiment Analysis	FinSA	F1 Accuracy	zh	FE	2,020	✔	✔	✔	✔	✗	✗
				StockB	1,962						
				CFPB	970						
				CFiQA-SA	233						
			en	FPB	970						
				FiQA-SA	235						
Semantic Matching	FinSM	F1 Accuracy	zh	Corpus	10,000	✗	✗	✗	✗	✗	✗
				AFQMC	4,316						
News Classification	FinNC	F1 Accuracy	zh	NL	884	✗	✔	✔	✗	✗	✗
				NL2	884						
Negative Judgment	FinNJ	F1 Accuracy	zh	NSP	500	✗	✗	✔	✗	✗	✗
Answer Selection	FinAS	F1 Accuracy	zh	FinevalF	222	✗	✗	✔	✗	✔	✗
Stock Prediction	FinSP	F1 Accuracy MCC	zh	StcokA	1,477	✔	✗	✗	✗	✗	✗
				CACL18	511						
				CBigData18	159						
				CCIKM18	86						
			en	ACL18	3,720						
				BigData18	1,472						
				CIKM18	1,143						
Relationship Extraction	FinRE	F1 Accuracy	zh	RE	1,489	✗	✗	✔	✗	✗	✗
Headline Classification	FinHC	Avg F1	zh	CHeadlines	2,051	✔	✗	✗	✔	✗	✗
			en	Headlines	20,547						
Credit Classification	FinCC	F1 Accuracy MCC	en	German	200	✗	✗	✗	✗	✗	✗
				Australian	139						
Hawkish-dovish Classification	FinDC	F1 Accuracy	en	FOMC	496	✗	✗	✗	✗	✗	✗
Question Answering	FinQA	EM Accuracy	zh	QA	2,469	✔	✔	✔	✔	✗	✔
				CEnQA	133						
				CConFinQA	237						
			en	EnQA	1,147						
				ConFinQA	1,490						
Entity Recognition	FinER	F1 Entity F1	zh	CNER	337	✔	✔	✔	✔	✗	✗
			en	NER	98						
			en	FINER-ORD	1,075						
Event Detection	FinED	F1 Recall Precision	zh	19CCKS	2,936	✗	✔	✗	✗	✗	✗
				20CCKS	9,159						
				21CCKS	1,400						
				22CCKS	11,829						
Text Summarization	FinTS	Rouge	zh	NA	3,600	✗	✔	✔	✗	✗	✗
		BERTScore	en	ECTSUM	495						
		BARTScore	en	EDTSUM	2000						
Open Source			✔			✔	✔	✔	✔	✔	✗
Task Tota			15			5	6	9	4	1	1
Data Tota			95k			43k	4k	11k	4k	5k	0.2k

Table 2: The detailed of our bilingual multi-task evaluation datasets and metrics in ICE-FLARE. "Cover" indicates whether the public benchmark ([1]FLARE, [2]CFBenchmark, [3]FinCUGE, [4]FLUE, [5]Fineval, [6]CGCE) cover the specific task of ICE-FLARE.

by Tsinghua University and ZhiPu AI, with versions like ChatGLM2-6B and ChatGLM3-6B. 5) **BLOOM** (Muennighoff et al., 2022): BLOOM, developed by the BigScience team, is a versatile LLM with parameter sizes ranging from 560M to 176B, including Bloomz-7b1. 6) **InternLM** (Team, 2023): InternLM is a language model developed by SenseTime, featuring different parameter configurations such as InternLM-7B. 7) **Qwen** (Bai et al., 2023): Qwen is a series of advanced LLMs introduced by Alibaba Cloud, including Qwen-7B for conversational capabilities in Chinese.

Following (Wu et al., 2023; Li et al., 2023b; Xie et al., 2023), we employed two common used zero-shot and few-shot evaluation methods to assess the model's adaptability and performance.

3.1 Results

We present evaluation results of various LLMs and ICE-INTERN variants on diverse tasks in Figure 2. The detailed results are shown in Appendix 5.

Model Overall Performance. Overall, ICE-INTERN-full-7B, a model fine-tuned on the entire dataset, emerged as the standout performer, achieving the best results in a significant majority of tasks. This model demonstrated exceptional superiority, securing the top spot in 21 out of the total tasks when including ChatGPT and GPT-4 results, and an even more impressive 24 tasks when excluding these two models. This underscores the effectiveness of domain-specific fine-tuning and the critical role of task data scale in enhancing LLM performance within specific domains. This fully demonstrates the exceptional superiority of ICE-INTERN and highlights the importance of domain-specific instruction fine-tuning and task data scale in improving the performance of LLMs in specific domains.

GPT-4 showcased robust performance, especially in the English dataset tasks, as indicated by the results cited from previous studies. While demonstrating strong performance on tasks within the English dataset, it evidenced less effectiveness in Chinese tasks, a trend consistent with other English financial fine-tuned and backbone LLMs. This highlights a significant language disparity, where models excel in English language tasks but face challenges in Chinese financial tasks. Comparatively, InternLM consistently surpasses the LLaMA2 model across all tasks, affirming our decision to utilize InternLM as the foundational model over LLaMA, as employed in related research (Yang et al., 2023c; Wu et al., 2023; Xie et al., 2023).

Ablation Study. In our ablation study, we observed distinct patterns in the performance of Large Language Models (LLMs) when applied to tasks

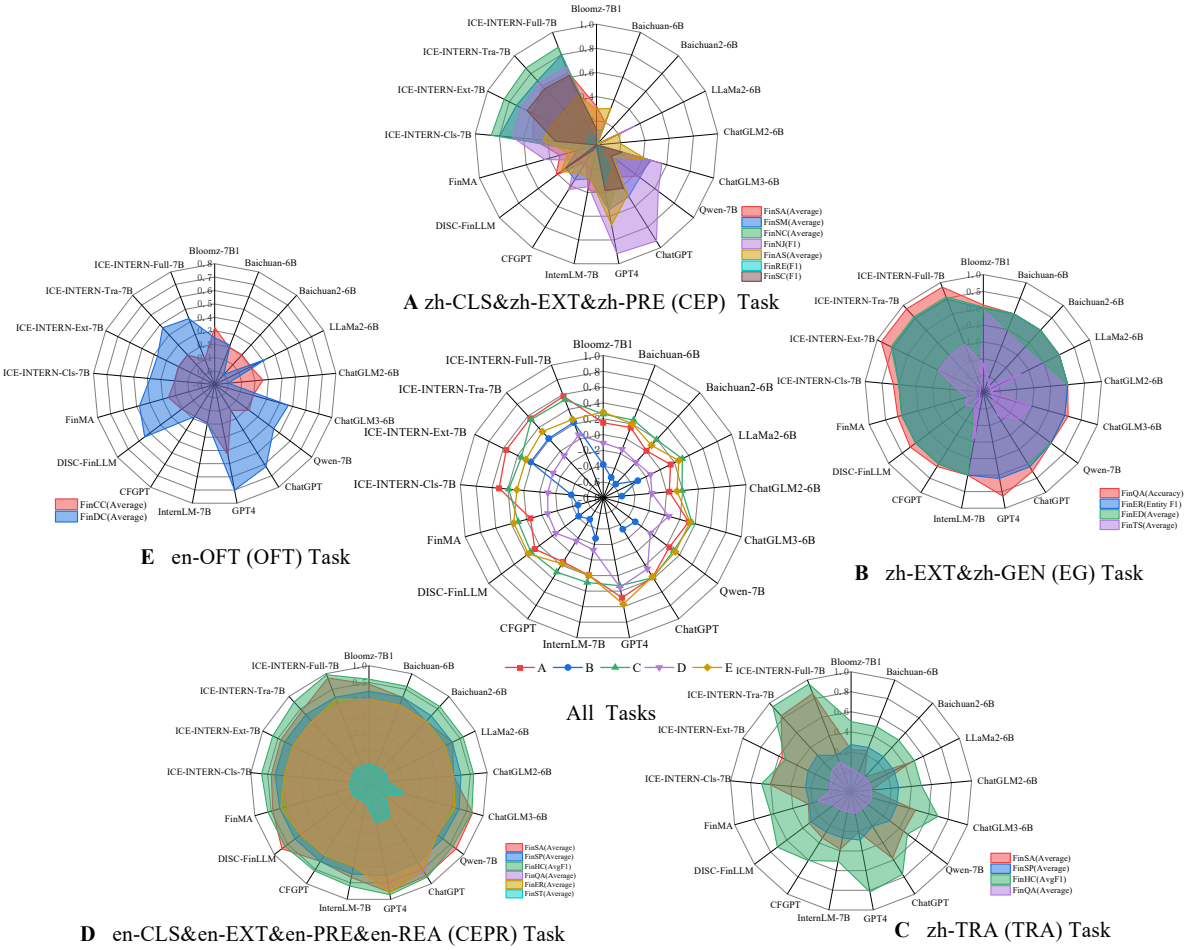

Figure 2: Multiple radar charts illustrating the performance of various LLMs and ICE-INTERN variants on diverse tasks.

involving limited resource languages compared to English, shedding light on critical considerations for the development of LLMs in the finance domain. Specifically, classification and prediction tasks had minimal impact on the performance of the fine-tuned ICE-INTERN-CEP-7B model. In contrast, the inclusion of extraction tasks significantly enhanced the capabilities of the ICE-INTERN-GE-7B model, underscoring the importance of comprehension and reasoning in financial applications.

Interestingly, the adaptation of translated instructions for fine-tuning in the ICE-INTERN-TRA-7B model did not foster improvements, and in some cases, it led to a decline in performance. This observation aligns with findings from previous studies (Chen et al., 2023; Xie et al., 2023), suggesting that while the InternLM backbone model demonstrates proficiency in classification tasks, extraction tasks—which demand a deeper level of understanding—are pivotal for augmenting model performance. Additionally, these results highlight the detrimental effect of low-quality translation data on model efficacy, suggesting that enhancements in translation quality are essential for optimizing the performance of LLMs in multilingual contexts.

Bilingual Case Analysis. In our investigation into cross-lingual tasks within the financial domain, specifically FinHC, FinSA, and FinSP, the ICE-INTERN-full-7B model demonstrated enhancements in performance, with gains ranging from 1% to 10% upon integrating corresponding translation data during the fine-tuning process. This pattern indicates a nuanced impact of translation data on fine-tuning outcomes, particularly evident in tasks involving short-text classifications such as FinSA and FinHC, where the benefits appear more modest. Conversely, in tasks requiring the classification of longer texts, such as FinSP which deals with time-series data, the inclusion of translation data markedly improves performance.

This differential impact can be attributed to the inherent characteristics of text length and complex-

ity. Short-text classification tasks, which typically involve simpler syntactic structures, allow our ICE-INTERN models to effectively leverage available Chinese data for reasoning. However, for longer texts, the addition of English data significantly augments the models' reasoning capabilities by providing richer contextual information. These findings underscore the critical role of text length and complexity in assessing the value of incorporating translation data for fine-tuning LLMs in cross-lingual settings. They suggest a tailored approach, where the decision to integrate translation data is informed by the specific nature of the text involved in each task, enhancing the effectiveness of LLMs in handling diverse financial linguistic tasks across languages.

4 Conclusion

In conclusion, we presented the ICE-PIXIU framework, which consists of the pioneering cross-lingual bilingual financial model, ICE-INTERN, and the evaluation benchmark, ICE-FLARE. The bilingual capability of ICE-INTERN empowers global finance by breaking language barriers, while ICE-FLARE enables comprehensive cross-lingual assessments for various financial tasks. The open access nature of ICE-PIXIU promotes collaboration and research in financial NLP. The framework provides a unified solution for diverse financial applications and ensures reliable cross-lingual consistency through rigorous evaluation. Overall, ICE-PIXIU contributes to the advancement of research and development in financial NLP, fostering innovation in the field.

Limitations

Despite the positive contributions of this study, we recognize the following limitations: 1)**Limitations of larger parameters**: Limited computational resources make it difficult to train models with more than 7B parameters, which hampers their potential performance with tuning large scale complete dataset. 2）**Adaptability of backbone LLMs**: the performance of fine-tuned models is influenced by the varying adaptability of backbone models to specific language tasks. Backbone model selection should be specific task-dependent. 3）**Inconsistent quality of prompts**: Even prompts with domain-expert annotation, instruction prompts may exhibit varying performance across different models. Quality differences can impact the assessment of poor model training. 4）**Potential negative impact**: Open-source models primarily aim to promote research, while commercial misuse can lead to financial risks. Proper regulation is needed to ensure responsible usage and mitigate potential harm.

Ethics Statement

Our financial LLMs and datasets adheres to ethical principles. We prioritize fairness, inclusivity, and non-discrimination in our language data. While we can't review all content, we use it as a text corpus without endorsing any views. Transparency, fairness, and accountability guide our AI practices. We comply with laws, protecting privacy and intellectual property. Responsible AI development is our focus, addressing biases and striving for unbiased outputs. We monitor and evaluate the model to meet ethical concerns. Our aim is a bilingual financial model with high ethical standards. Transparency, user well-being, and reliable information are our priorities. The responsible use of our financial LLMs and datasets ensures its positive impact in finance domain.

References

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

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang. 2018. The bq corpus: A large-scale domain-specific chinese corpus for sentence semantic equivalence identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4946–4951.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Lu, et al. 2023. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. *arXiv preprint arXiv:2310.15205*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

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

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, et al. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. 2022. Duee-fin: A large-scale dataset for document-level event extraction. In *International Conference on Natural Language Processing and Chinese Computing*, pages 172–183.

Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

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*.

Chen Jia, Yuefeng Shi, Qinrong Yang, and Yue Zhang. 2020. Entity enhanced bert pre-training for chinese ner. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, pages 6384–6396.

Katikapalli Subramanyam Kalyan. 2023. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, page 100048.

Yang Lei, Jiangtong Li, Ming Jiang, Junjie Hu, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023. Cfbenchmark: Chinese financial assistant benchmark for large language model. *arXiv preprint arXiv:2311.05812*.

Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023a. Cfgpt: Chinese financial assistant with large language model. *arXiv preprint arXiv:2309.10654*.

Xianzhi Li, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023b. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? an examination on several typical tasks. *arXiv preprint arXiv:2305.05862*.

Dakuan Lu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, et al. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Roberts, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, et al. 2022. Ectsum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. *arXiv preprint arXiv:2210.12467*.

OpenAI. 2023. Gpt-4 technical report.

Ross Quinlan. Statlog (Australian Credit Approval). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C59012.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Sutawika, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The 10th International Conference on Learning Representations*.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion dollar words: A new financial dataset, task & market analysis. *arXiv preprint arXiv:2305.07972*.

Ankur Sinha and Tanmay Khandait. 2021. Impact of news on the commodity market: Dataset and results. In *Proceedings of the 2021 Future of Information and Communication Conference (FICC)*, pages 589–601.

Sahar Sohangir, Dingding Wang, Anna Pomeranets, and Taghi M Khoshgoftaar. 2018. Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5(1):1–25.

Yejun Soun, Jaemin Yoo, Minyong Cho, Jihyeong Jeon, and U Kang. 2022. Accurate stock movement prediction with self-supervised learning from sparse noisy tweets. In *2022 IEEE International Conference on Big Data*, pages 1691–1700.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Lachaux, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yunhe Wang, Hanting Chen, Yehui Tang, Tianyu Guo, Kai Han, Ying Nie, Xutao Wang, Hailin Hu, Zheyuan Bai, Yun Wang, et al. 2023. Pangu-π : Enhancing language model architectures via nonlinearity compensation. *arXiv preprint arXiv:2312.17276*.

Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. 2018. Hybrid deep sequential modeling for social text-driven stock prediction. In *Proceedings of the International Conference on Information and Knowledge Management(CIKM)*, pages 1627–1630.

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

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. In *37th International Conference on Neural Information Processing Systems*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1970–1979.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

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

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

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Liao, et al. 2023a. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*.

Xuanyu Zhang, Bingbing Li, and Qing Yang. 2023b. Cgce: A chinese generative chat evaluation benchmark for general and financial domains. *arXiv preprint arXiv:2305.14471*.

Yang Qing Xu Dongliang Zhang, Xuanyu. 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.

Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the event: Corporate events detection for news-based event-driven trading. *arXiv preprint arXiv:2105.12825*.

Jinan Zou, Haiyao Cao, Lingqiao Liu, Yuhao Lin, Ehsan Abbasnejad, and Javen Qinfeng Shi. 2022. Astock: A new dataset and automated stock trading based on stock-specific news analyzing model. *arXiv preprint arXiv:2206.06606*.

Appendix

A Raw Data Statistics

Table 3 presents the specifics of the raw data used to construct the Chinese-English bilingual dataset for multi-task financial instruction tuning.

B Data Translation Example

Figure 3 presents the examples of translating English BigData22 into Chinese CBigData22 using written specific prompts.

C Instruction Prompt Examples

Table 4 presents the example of the Chinese prompt and corresponding English translation for each specific financial task.

D Experimental Results

Table 5 presents the performance of different 6B-7B LLMs, baseline FinLLMs and variants of ICE-INTERN on the ICE-FLARE bechmark.

E Financial Task Examples

Langugae	Type	Specific Task	Data	Raw	Instruction	Evaluation	Data Types	License
zh	zh-CLS	FinSA	FE	18,177	18,177	2,020	social texts	Public
			StockB	9,812	9,812	1,962	social texts	Apache-2.0
		FinSM	AFQMC	38,650	38,650	4,316	online chat service	Apache-2.0
			Corpus	120,000	110,000	10,000	bank service logs	Public
		FinNC	NL	7,955	7,955	884	news articles	Public
			NL2	7,955	7,955	884	news articles	Public
		FinNJ	NSP	4,499	4,499	500	social texts	Public
		FinAS	FinevalF	1,115	1,115	222	financial exam	Apache-2.0
	zh-EXT	FinQA	QA	22,375	22,375	2,469	QA pairs of news	Public
		FinED	19CCKS	156,834	14,674	2,936	social texts	CC BY-SA 4.0
			20CCKS	372,810	45,796	9,159	news, reports	CC BY-SA 4.0
			21CCKS	8,000	7,000	1,400	news, reports	CC BY-SA 4.0
			22CCKS	109,555	59,143	11,829	news, reports	CC BY-SA 4.0
		FinRE	RE	14,973	14,973	1,489	news, entity pairs	Public
		FinER	CNER	1,685	1,685	337	financial reports	Public
	zh-PRE	FinSP	StockA	14,769	14,769	1,477	news,historical prices	Public
	zh-GEN	FinTS	NA	32,400	32,400	3,600	news, announcements	Public
en	en-CLS	FinSA	FPB	4,845	4,845	970	economic news	CC BY-SA 3.0
			FiQA-SA	1,173	1,173	235	news headlines,tweets	Public
		FinHC	Headlines	11,412	102,708	20,547	news headlines	CC BY-SA 3.0
		FinCC	German	1,000	1,000	200	credit records	CC BY-SA 4.0
			Australian	690	690	139	credit records	CC BY-SA 4.0
	en-EXT	FinER	NER	609	609	98	financial agreements	CC BY-SA 3.0
	en-PRE	FinSP	ACL18	27,053	27,053	3,720	tweets, historical prices	MIT License
			BigData22	7,164	7,164	1,472	tweets, historical prices	Public
			CIKM18	4,967	4,967	1,143	tweets, historical prices	Public
	en-REA	FinQA	EnQA	8,281	8,281	1,147	earnings reports	MIT License
			ConFinQA	3,458	12,594	1,490	earnings reports	MIT License
zh	zh-TRA	FinSA	CFPB	4,845	4,838	970	economic news	MIT license
			CFiQA-SA	1,173	1,143	233	ews headlines,tweets	MIT license
		FinSP	CACL18	27,056	2,555	511	tweets, historical prices	MIT license
			CBigData22	7,167	798	159	tweets, historical prices	MIT license
			CCIKM18	4,970	431	86	tweets, historical prices	MIT license
		FinHC	CHeadlines	102,708	10,256	2,051	news headlines	MIT license
		FinQA	CEnQA	8,281	668	133	earnings reports	MIT license
			CConFinQA	12,594	1,189	237	earnings reports	MIT license
en	en-OFT	FinER	FINER-ORD	1,075	-	1,075	news articles	CC BY-SA 4.0
		FinTS	ECTSUM	495	-	495	earning call transcipts	Public
			EDTSUM	2,000	-	2,000	news articles	Public
		FinDC	FOMC	496	-	496	FOMC transcripts	CC BY-SA 4.0

Table 3: Details of raw data for the Chinese-English bilingual multi-task financial instruction and evaluation.

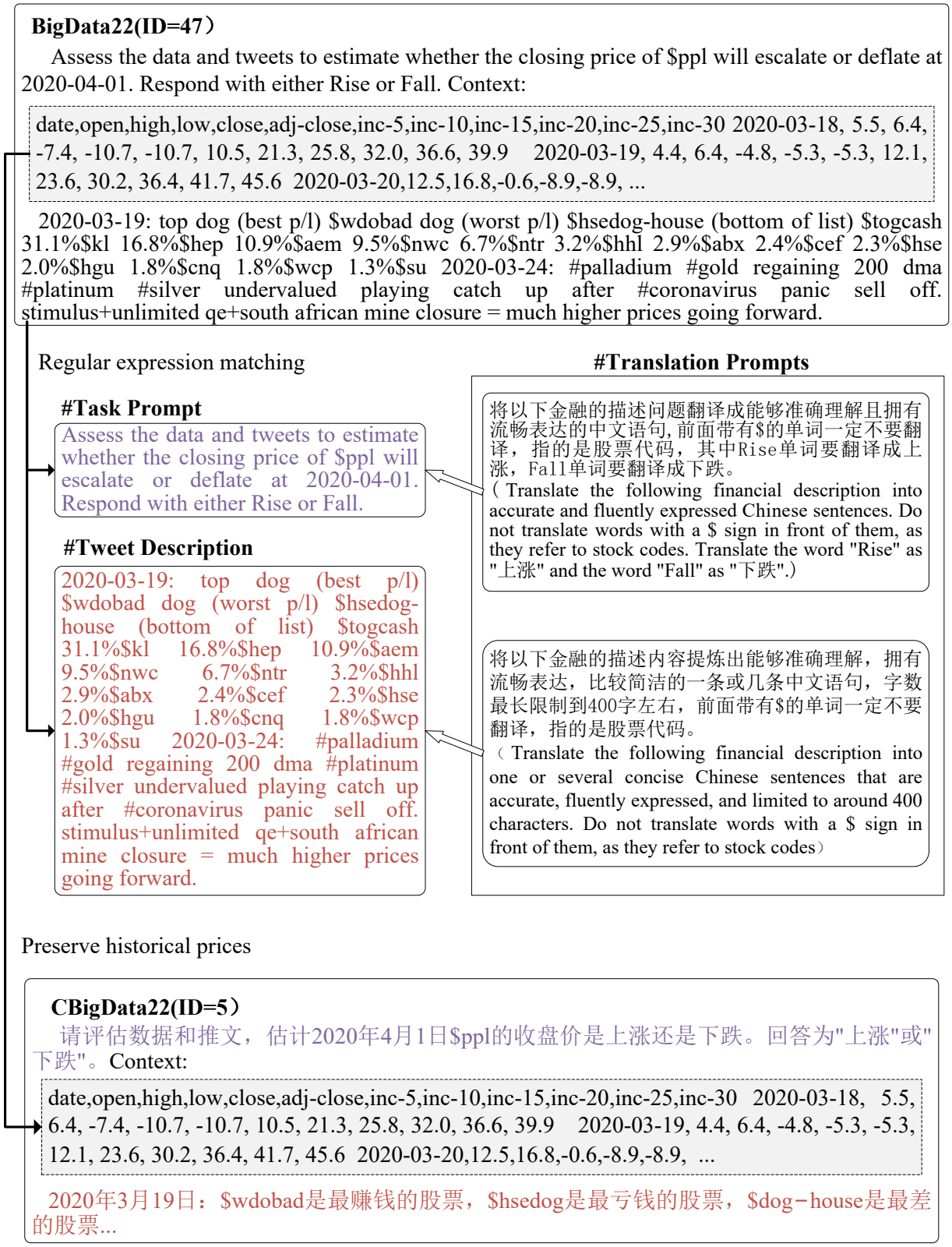

Figure 3: Example of translating English BigData22 into Chinese CBigData22 using written specific prompts.

Specific Task	Chinese Prompts	English Translations
Sentiment Analysis	确定所提供的金融新闻文章中的句子的情绪，识别其中的情绪是积极，消极，还是中性的。只需回答'积极'、'中性'或'消极'。	Determine the sentiment of the sentences in the provided financial news article, identifying whether the sentiment is positive, negative, or neutral. Simply answer 'positive', 'neutral', or 'negative'.
Sentiment Matching	判断给出的两个金融文本表达的意思是否相似，你只需要回答是或否。现在判断下面两个句子是否相似：	Determine whether the two provided financial texts convey similar meanings; you only need to answer 'yes' or 'no'. Now determine if the following two sentences are similar.
News Classification	根据金融报道的内容，判断其属于'中国'、'国际'还是'外国'，只需给出其中一个类别。现在分析下面的报道，回答它所属的类别：	According to the content of the financial report, determine whether it belongs to 'China', 'International', or 'Foreign', only one category needs to be given. Now analyze the following report and answer the category it belongs to:
Negative Judgment	根据所给的金融新闻及实体，你需要判断所给实体是否含有负面的消息，你只需简单回答'有'或者'无'	Based on the given financial news and entities, you need to determine whether the given entities contain negative messages. You only need to simply answer 'yes' or 'no'.
Stock Prediction	请仔细分析数据和推文，预测2017-10-13时$trv的收盘价格是上涨还是下跌。请确认是上涨还是下跌，只回答为'上涨'或者'下跌'。	Please carefully analyze the data and tweets, and predict whether the closing price of $TRV on 2017-10-13 will go up or down. Please confirm whether it is going up or down. You can answer 'Rise' or 'Fall'.
Stock Classification	综合考虑所给5天的市场数据和公司相关公告，请根据新闻对股票数据的影响判断该公司的股票运动走势是'跑赢大盘'、'中性'还是'表现不佳'。	Taking into account the market data over the given 5 days and the company's relevant announcements, please judge the movement trend of the company's stock based on the impact of the news on the stock data as 'outperforming the market', 'neutral', or 'underperforming'.
Answer Selection	根据所提出的金融问题，从以下四个选项中选择最合适的一个。你的输出应该是：'A'、'B'、'C' 或'D'。	Based on the financial issue presented, choose the most appropriate one from the following four options. Your output should be: 'A', 'B', 'C', or 'D'.
Headline Classification	请考虑标题是否讨论了与黄金相关的过去事件。新闻标题是否暗示黄金的过去新闻？您的回答应该为'是'或'否'。	Please consider whether the headline discusses past events related to gold. Does the news headline imply past news about gold? Your answer should be 'Yes' or 'No'.
Question Answering	你需要分析金融文本，根据内容回答相关问题。如果你对答案感到迷茫，可以回答'无相应参数'。	You need to analyze financial texts and answer related questions based on the content. If you feel confused about the answer, you can reply with 'No relevant parameters'.
Event Detection	阅读以下金融领域的公告，判断所有的事件类型及其对应主体，省略数字。请以:'事件类型, 事件主体'的格式回复 。其中事件类型应该在这里面：['信批违规',...'涉嫌欺诈']	The task is to read a financial announcement, determine all event types and their corresponding entities, and reply in the format 'event type, event subject'. The event types should be among those listed in the brackets ['credit approval violation',... 'suspected fraud'].
Entity Recognition	在分析中国证券交易委员会备案文件中的句子时，识别指明个人('PER')、组织('ORG')或地点('LOC')的特定命名实体。答案应遵循格式'实体名称, 实体类型' 。	When analyzing sentences in the China Securities Regulatory Commission filing documents, identify specific named entities that indicate individuals ('PER'), organizations ('ORG'), or locations ('LOC'). The answer should follow the format 'entity Name, entity Type'.
Relationship Extraction	请仔细分析所给金融报道和实体对，然后在['合并','竞争',...]中选择能准确描述该实体对关系的选项。请直接给出答案，如有疑虑可回答unknown。	Please carefully analyze the given financial report and entity pair, then choose the option from ['merge', 'compete', ...] that accurately describes the relationship of the entity pair. Please provide the answer directly; if in doubt, you may respond with 'unknown'
Text Summarization	通过阅读金融公告，你的任务是对所给的文本进行简短的总结，重点突出主要论点，长度保持在一到两句之间。	Your task is to provide a brief summary of the given text by reading financial announcements, emphasizing the main arguments, and keeping the length between one to two sentences

Table 4: Example of the Chinese prompt and corresponding English translation for each specific financial task.

Type	Specific Task	Metrics	Dataset	Bloomz-7B1	Baichuan-6B	Baichuan2-6B	LLaMa2-6B	ChatGLM2-6B	ChatGLM3-6B	Qwen-7B	ChatGPT	GPT4	InternLM-7B	CFGPT	DISC-FinLLM	FinMA	ICE-INTERN-CEP-7B	ICE-INTERN-GE-7B	ICE-INTERN-TRA-7B	ICE-INTERN-Full-7B
zh-CLS zh-EXT zh-PRE	FinSA	Accuracy	FE	0.587	0.442	0.002	0.423	0.000	0.572	0.366	0.329	0.349	0.553	0.306	0.624	0.534	0.760	**0.778**	0.743	0.760
			StockB	0.196	0.058	0.000	0.047	0.000	0.389	0.431	0.310	0.441	0.297	0.070	0.282	0.202	0.422	0.593	**0.636**	0.593
		F1	StockB	0.177	0.098	0.000	0.078	0.000	0.443	0.482	0.313	0.462	0.279	0.098	0.344	0.201	0.395	0.571	0.575	**0.578**
	FinSM	F1	Corpus	0.008	0.154	0.002	0.292	0.002	0.444	0.334	0.682	0.704	0.278	0.218	0.255	0.230	0.854	**0.859**	0.846	0.854
			AFQMC	0.004	0.296	0.014	0.266	0.012	0.470	0.560	0.323	0.391	0.292	0.460	0.373	0.095	0.743	0.609	0.594	**0.743**
	FinNC	F1	NL	0.245	0.219	0.002	0.280	0.000	0.453	0.348	0.645	0.816	0.375	0.164	0.267	0.261	0.950	0.922	**0.954**	0.950
			NL2	0.000	0.049	0.000	0.066	0.000	0.049	0.074	0.034	0.273	0.039	0.004	0.002	0.060	0.786	0.774	0.768	**0.786**
	FinNJ	F1	NSP	0.041	0.217	0.000	0.353	0.004	0.558	0.636	**0.935**	0.916	0.351	0.435	0.208	0.442	0.692	0.692	0.695	0.692
	FinAS	Accuracy	FinevalF	0.302	0.329	0.014	0.266	0.185	0.423	0.243	0.491	**0.662**	0.320	0.284	0.369	0.261	0.441	0.410	0.387	0.441
		F1	FinevalF	0.295	0.319	0.024	0.171	0.190	0.418	0.096	0.481	**0.686**	0.247	0.264	0.362	0.220	0.442	0.387	0.376	0.442
	FinRE	F1	RE	0.012	0.039	0.000	0.011	0.000	0.076	0.026	0.211	**0.340**	0.015	0.014	0.108	0.009	0.092	0.088	0.103	0.103
	FinSP	F1	StockA	0.131	0.000	0.000	0.000	0.000	0.220	0.154	0.422	0.383	0.003	0.050	0.315	0.000	0.340	**0.637**	0.630	0.617
zh-EXT zh-GEN	FinQA	Accuracy	QA	0.098	0.001	0.000	0.001	0.000	0.099	0.000	0.256	0.064	0.001	0.092	0.201	0.123	0.167	**0.864**	0.860	0.855
	FinER	Entity F1	CNER	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.101	0.103	0.000	0.000	0.000	0.000	0.000	**0.483**	0.482	0.468
	FinED	Precision	19CCKS	0.080	0.013	0.003	0.001	0.012	0.001	0.001	0.010	0.011	0.003	0.003	0.004	0.004	0.005	**0.831**	0.825	0.824
			20CCks	0.052	0.006	0.002	0.001	0.009	-	0.001	0.001	0.008	0.002	0.002	0.002	0.003	0.006	0.740	0.741	**0.743**
			21CCks	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	**0.004**	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.002
			22CCks	0.000	0.001	0.001	0.000	0.002	0.001	0.000	0.006	-	0.001	0.003	0.000	0.021	0.007	0.647	0.642	**0.660**
		F1	19CCKS	0.081	0.013	0.003	0.001	0.012	0.001	0.001	0.010	0.012	0.003	0.006	0.005	0.007	0.005	**0.831**	0.825	0.824
			20CCks	0.068	0.011	0.002	0.003	0.015	-	0.002	0.001	0.010	0.003	0.004	0.004	0.005	0.004	**0.705**	0.703	0.702
			21CCks	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.003	0.000	0.000	0.000	0.000	0.000	0.001	0.001	**0.001**
			22CCks	0.001	0.001	0.001	0.001	0.004	0.001	0.001	0.008	-	0.001	0.004	0.000	0.002	0.007	0.588	0.581	**0.595**
	FinTS	Rouge-1	NA	0.078	0.028	0.006	0.104	0.001	0.256	0.142	0.093	-	0.145	0.006	0.014	0.010	0.050	0.284	**0.289**	0.287
		BertScore	NA	0.676	0.527	0.482	0.649	0.496	0.748	0.603	0.710	-	0.651	0.603	0.607	0.603	0.707	0.824	0.823	**0.824**
		BartScore	NA	-5.704	-6.845	-7.274	-4.539	-7.293	-4.097	-4.301	-5.882	-	-4.153	-6.661	-6.194	-6.640	-6.044	-4.180	-4.183	**-4.027**
zh-TRA	FinSA	Accuracy	CFPB	0.005	0.203	0.000	0.098	0.000	0.424	0.239	0.699	0.196	0.273	0.256	0.085	0.038	0.545	0.498	0.846	**0.849**
		F1	CFPB	0.010	0.129	0.000	0.092	0.000	0.348	0.340	0.642	0.306	0.209	0.346	0.091	0.067	0.529	0.485	0.845	**0.846**
			CFiQA-SA	0.652	0.401	0.000	0.350	0.000	0.617	0.773	0.400	0.120	0.668	0.420	0.784	0.616	0.752	0.709	0.796	**0.868**
	FinSP	Accuracy	CACL18	0.550	0.585	**0.591**	0.577	0.589	0.485	0.556	0.544	0.568	0.566	0.581	0.564	0.585	0.438	0.442	0.581	0.417
			CBigData18	0.667	0.686	0.686	0.686	0.686	0.440	0.673	0.629	0.642	**0.686**	0.535	0.585	0.660	0.365	0.403	0.660	0.321
			CCIKM18	0.384	0.372	0.372	0.372	0.372	0.605	0.384	0.326	0.419	0.372	0.407	0.477	0.361	0.605	**0.628**	0.395	0.605
		MCC	CACL18	-0.034	0.008	0.000	0.021	-0.002	0.004	0.012	0.002	0.010	-0.017	**0.099**	0.076	0.022	0.044	0.042	0.039	-0.034
			CBigData18	-0.013	0.000	0.000	0.000	0.000	-0.042	-0.023	-0.024	0.066	0.000	0.067	0.037	-0.109	0.028	-0.056	**0.114**	-0.033
			CCIKM18	0.084	0.000	0.000	0.000	0.000	0.133	0.084	-0.250	0.032	0.000	-0.066	0.040	-0.141	0.037	**0.204**	-0.021	-0.119
	FinHC	AvgF1	CHeadlines	0.503	0.500	0.500	0.500	0.500	0.695	0.500	0.769	0.818	0.504	0.607	0.716	0.626	0.690	0.526	0.952	**0.957**
	FinQA	EM Accuracy	CEnQA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.008**
			ConFinQA	0.068	0.004	0.000	0.008	0.000	0.021	0.000	0.000	0.021	0.000	0.008	0.030	**0.276**	0.013	0.046	0.169	0.228
en-CLS en-EXT en-PRE en-REA	FinSA	Accuracy	FPB	0.388	0.231	0.000	0.286	0.001	0.741	0.696	0.780*	0.760*	0.319	0.278	0.808	0.435	0.353	0.384	0.398	**0.867**
		F1	FPB	0.231	0.171	0.000	0.127	0.002	0.744	0.702	0.780*	0.780*	0.353	0.292	0.809	0.436	0.349	0.331	0.415	**0.866**
			FiQA-SA	0.742	0.327	0.000	0.370	0.000	0.551	0.754	0.600*	0.800*	0.349	0.096	0.532	0.477	0.477	0.503	0.480	**0.848**
	FinSP	Accuracy	ACL18	0.503	0.500	0.487	0.508	0.488	0.511	0.490	0.500*	0.520*	0.510	0.483	0.492	0.493	0.515	0.519	**0.525**	0.502
			BigData18	0.550	0.525	0.448	0.484	0.448	0.4769	0.508	0.530*	0.540*	0.469	0.452	0.472	0.496	**0.555**	0.520	0.483	0.450
			CIKM18	0.460	0.478	0.419	0.496	0.418	0.499	0.470	0.550*	**0.570***	0.469	0.421	0.430	0.461	0.563	0.556	0.534	0.493
		MCC	ACL18	-0.034	-0.010	0.000	0.010	0.023	0.035	-0.038	0.005*	0.020*	0.019	-0.026	0.001	-0.014	0.017	0.029	**0.045**	0.023
			BigData18	0.002	-0.013	0.000	-0.012	0.000	0.008	-0.017	-0.025*	0.030*	-0.014	0.030	**0.063**	-0.002	0.042	0.030	0.017	-0.012
			CIKM18	-0.053	0.009	0.000	-0.028	-0.026	**0.043**	-0.056	0.010*	0.020*	-0.034	-0.016	-0.044	-0.038	-0.001	-0.007	0.005	0.042
	FinHC	AvgF1	Headlines	0.600	0.600	0.600	0.600	0.600	0.655	0.600	0.770*	0.860*	0.600	0.551	0.600	0.600	0.682	0.668	0.780	**0.965**
	FinCC	F1	German	**0.649**	0.525	0.525	0.525	0.525	0.525	0.476	0.200*	0.550*	0.525	0.525	0.548	0.192	0.525	0.525	0.525	0.318
			Australian	0.412	0.260	0.260	0.260	0.358	0.273	0.502	0.410*	**0.740***	0.260	0.288	0.260	0.599	0.260	0.260	0.260	0.421
		MCC	German	**0.231**	0.000	0.000	0.000	0.000	0.000	-0.166	-0.100*	-0.020*	0.000	0.000	0.012	0.002	0.000	0.000	0.000	-0.306
			Australian	0.000	0.000	0.000	0.000	0.153	-0.017	0.065	0.000*	**0.470***	0.000	-0.097	0.000	0.232	0.000	0.000	0.000	-0.030
	FinDC	Accuracy	FOMC	0.288	0.256	0.000	0.319	0.000	0.464	0.454	0.600*	**0.690***	0.252	0.208	0.557	0.476	0.385	0.373	0.488	0.429
		F1	FOMC	0.207	0.172	0.000	0.323	0.000	0.470	0.441	0.640*	**0.710***	0.131	0.155	0.541	0.480	0.375	0.372	0.455	0.424
	FinQA	EM Accuracy	EnQA	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.580*	**0.630***	0.000	0.000	0.004	0.001	0.000	0.000	0.001	0.001
			ConFinQA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.600*	**0.760***	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	FinER	Entity F1	NER	0.000	0.000	0.000	0.012	0.000	0.248	0.012	0.770*	**0.830***	0.000	0.000	0.039	0.392	0.000	0.006	0.000	0.362
			FINER-ORD	0.000	0.000	0.000	0.000	0.000	0.021	0.000	0.280*	**0.770***	0.000	0.000	0.008	0.000	0.000	0.000	0.000	0.000
	FinST	Rouge-1	ECTSUM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**
			EDTSUM	0.013	0.015	0.007	0.027	0.009	0.175	0.036	0.170*	**0.200***	0.032	0.008	0.051	0.033	0.009	0.018	0.017	0.015
		BertScore	ECTSUM	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**
			EDTSUM	0.522	0.447	0.445	0.483	0.445	0.599	0.536	0.660*	**0.670***	0.478	0.508	0.552	0.467	0.476	0.536	0.505	0.500
		BartScore	ECTSUM	-5.178	-5.178	-5.178	-5.178	-5.178	-5.178	-5.178	-5.178*	-5.178*	-5.178	-5.178	-5.178	-5.178	-5.178	-5.178	-5.178	**-5.178**
			EDTSUM	-6.996	-6.923	-7.036	-6.902	-6.957	-4.048	-6.845	-3.640*	**-3.620***	-6.973	-7.083	-6.951	-6.964	-7.107	-6.945	-6.925	-6.927
Best Result Count			Including ChatGPT	2	0	1	0	0	1	0	1	16	1	1	1	1	1	10	6	21
			excluding ChatGPT	2	0	1	0	0	5	0	-	-	1	2	5	4	2	10	7	24

Table 5: The performance of different 6B-7B LLMs, baseline FinLLMs and variants of ICE-INTERN on the ICE-FLARE bechmark. All results via our evaluations are the average of three runs. Results with "*" of ChatGPT-3.5(ChatGPT) and ChatGPT-4(GPT4) on English dataset are cited from (Xie et al., 2023). Results in bold indicate the best results across all models. Results with "_" indicate the best results among models excluding ChatGPT.

FE	Context: 国检集团和华测检测哪个比较漂亮？业务都差不多，都是检测，并且国检持有碳交易股份，两个都买了。 Answer: 金融的报道偏向积极评价
StockB	Context: 中烟集团金融网站,融通租赁注册100亿！揭牌仪式有望。 Answer: 金融的报道偏向中性评价
CFPB	Context: 维萨拉在2007年第三季度的净利润从2006年同期的680万欧元（980万美元）降至300万欧元（430万美元）。 Answer: 金融的报道偏向消极评价
CFiQA-SA	Context: $CERN-在50和200MA上方整合,这里是很好的长期进入点，止损位于10MA以下-目标区域$70。 Answer: 金融的报道偏向积极评价
FPB	Context: Hearst will be able to consolidate about 20% of all Russian market for advertising in press after the purchase. Answer: Sentiment analysis of financial news is positive.
FiQA-SA	Context: $SLV-4.44% at 18 now AWFUL, down from 42.50 Answer: Sentiment analysis of financial news is negative

Table 6: Examples with context description and annotated answer for financial semantic matching task.

BQC	Context: 1:两个小时还没有等到确认电话怎么办？明天会继续联系嘛？ 2:下次借款是否不需要电话确认 Answer: 两个金融表达语义不是相似的
AFQMC	Context: 1:借呗还款日当天不能再借款吗 2:蚂蚁借呗要一次性还清才能再借吗 Answer: 两个金融表达语义是相似的

Table 7: Examples with context description and annotated answer for news classification task.

NL	Context: [中航泰达：拟购买包钢节能34%股权]中航泰达公告，公司拟以2.09亿元现金认购包钢节能新增注册资本1.28亿元，同时以2.59亿元受让北方稀土持有标的公司的1.59亿元注册资本。本次交易前，北方稀土直接持有包钢节能100%股权。本次交易完成后，公司将直接持有包钢节能34%的股权。 Answer: 报道涉及中国区域
NL2	Context: 欧股集体高开，德国DAX30指数涨0.44%，英国富时100指数涨1.29%，法国CAC40指数涨0.7%，欧洲斯托克50指数涨0.63%。 Answer: 报道涉及国际大盘

Table 8: Examples with context description and annotated answer for financial sentiment analysis task.

CHeadlines	Context: 在亚洲早盘，金价略有下跌，市场正在关注美国的数据。 Answer: 标题中是提到金价会下降的意见。
Headlines	Context: Gold holds near 3-1/2 week low as investors opt for riskier assets. Answer: Yes, the headline suggests a downward direction for gold.

Table 9: Examples with context description and annotated answer for financial headline classification task.

StockA	Context: 新北洋公告，公司控股子公司荣鑫科技董事会审议通过了《关于拟申请公司股票在全国中小企业股份转让系统终止挂牌的议案》，具体详见荣鑫科技（证券代码：839288）披露在全国中小企业股份转让系统（以下简称"新三板"）的相关公告。荣鑫科技基于其自身经营发展及战略规划的需要，拟申请在新三板终止挂牌。日期 开盘价 闭盘价 2021-01-04 34.63 34.9763 2021-01-05 34.63 35.5122 2021-01-06 35.1659 34.0116 2021-01-07 33.8879 32.3048 2021-01-08 32.0987 32.3213. Answer: 股票的表现不佳
CACL18	Context: $codi近期的走向趋势如下，open, high, low, close, adj-close, inc-5, inc-10, inc-15, inc-20, inc-25, inc-30；2015-09-16, -0.8, 0.2, -1.3, 1.0, 1.0, -0.9, -0.1, -0.1, -0.3, 0.0, 0.5; 2015-09-17,-1.4, 0.9, -1.7, 1.6, 1.6, -2.0, -1.7, -1.5, -1.8, -1.6, -1.2; 2015-09-18, -0.9, 0.2, -1.5, 0.1, 0.1, -1.6, -1.8, -1.5, -1.8, -1.7, -1.3; 2015-09-21, -0.5, 0.3, -1.8, 0.7, 0.7, -1.5, -2.3, -2.1, -2.4, -2.3, -2.1; 2015 -09-22, 1.3, 3.4, -0.4, -3.1,-3.1,1.7,0.8,0.9,0.9,0.7,1.0;2015-09-23,-0.8,0.7,-0.8,0.7,0.7,1.0, 0.1,0.3,0.2,-0.0,0.2; 2015-09-24,-0.1,1.5,-0.7,0.1,0.1,0.7,0.1,0.1,0.2,-0.1,0.1; 2015-09-25,1.5, 1.9,0.0,-0.9, -0.9,1.0, 1.0, 0.8, 1.0, 0.8, 0.9; 2015-09-28, 1.2, 1.2, -0.8, -0.2, -0.2, 0.5, 1.2, 1.0, 1.2, 0.9, 1.0; 2015-09-29, 1.3, 2.1, -0.3, -1.8, -1.8, 2.0, 3.0, 2.6, 2.8, 2.9, 2.8. $codi评级为买入, 18.6%华尔街分析师根据平均评级/目标得出的上涨。 Answer: $codi的收盘价在2015-9-30会上涨
CCIKM18	Context: $chk近期的走向趋势如下，date, open, high, low, close, adj-close, inc-5, inc-10, inc-15, inc-20, inc-25, inc-30; 2017-10-05，-0.7，0.7，-0.9，0.7，0.7，0.0，0.6，-1.3，-3.8，-5.6, -7.0; 2017-10-06, 1.4，1.7，-1.9, -2.3, -2.3, 2.0, 2.9, 1.4, -0.8，-2.8，-4.5; 2017-10-09，0.2，0.7, -1.4,-0.2, -0.2, 1.4, 2.7, 1.9，0.1，-2.2，-3.9; 2017-10-10,7.9, 8.2, 0.0, -6.7, -6.7, 6.8, 8.9, 9.0, 7.5, 5.0，3.2; 2017-10-11, -0.8, 0.8, -4.8, 1.0, 1.0, 4.1, 6.6, 7.4, 6.4, 4.2, 2.4; 2017-10-12,1.3,2.1,-1.8,-2.5,-2.5，4.4，8.1，9.7，8.8，6.8，5.3; 2017-10-13, 0.5,1.3,-0.8, 0.8, 0.8，2.0，6.2，8.2，7.8，6.3，4.6。 Answer: $chk的收盘价在2027-10-19会上涨
CBigData22	Context: $msft近期的走向趋势如下，date, open, high, low, close, adj-close, inc-5, inc-10, inc-15, inc-20, inc-25, inc-30; 2020-10-09, -2.1, 0.0, -2.1, 2.5, 2.5, -2.5, -2.8, -3.6, -4.0, -4.0,-2.5 ; 2020-10-12, -1.2, 1.1, -2.1, 2.6, 2.6, -3.9, -4.7, -5.5, -6.1, -6.3, -5.1; 2020-10-13,-0.1,1.1,-1.1,0.7,0.7,-3.0, -4.6, -5.6, -6.4, -6.5, -5.8; 2020-10-14, 1.0, 1.5, -0.8, -0.9, -0.9, -1.2, -3.3,- 4.2, -5.2, -5.5, -5.0; 2020-10-15,-1.2, 0.3, -1.7, -0.5, -0.5,0.2,-2.4,-3.2,-4.3,-4.7,-4.7; 2020-10-16, 0.2, 1.2, -0.2, 0.0, 0.0, 0.6, -1.8, -2.8, -3.8, -4.5, -4.6; 2020-10-19, 2.9, 3.8, -0.2, -2.5, -2.5, 2.4, 0.9, -0.2, -1.1, -1.9, -2.2; 2020-10-20, 0.5, 1.3, -0.7, 0.2, 0.2, 1.5, 1.1, -0.2, -1.2, -2.0, -2.2; 2020-10-21,-0.8, 1.0, -0.8, 0.1, 0.1, 0.8, 1.2, -0.1, -0.9, -1.8, -2.2; 2020-10-22, -0.4, 0.5, -1.5, 0.0, 0.0, 0.4, 1.4, -0.1, -0.7, -1.7, -2.1。 Answer: $msft的收盘价在2020-10-23会上涨
ACL18	Context: The recent trend of $codi is as follows. date, open, high, low, close, adj-close, inc-5, inc-10, inc-15, inc-20, inc-25, inc-30; 2015-03-30, 0.0, 0.4, -0.9, 0.0, 0.0, -0.0, -0.8, -1.6, -2.1, -2.4, -2.6, 2015-03-31, 0.1, 0.3, -0.8, -0.3, -0.3, 0.2, -0.4, -1.0, -1.7, -2.0, -2.2; 2015-04-01, 1.1, 1.1, -0.7, -1.2, -1.2, 1.2, 0.9, 0.5, -0.4, -0.7, -0.9; 2015-04-02, -0.2, 0.4, -0.4, -0.2, -0.2, 1.0, 1.2, 0.7, -0.1, -0.4, -0.6; 2015-04-06, -0.6, 0.5, -0.9, 0.6, 0.6, 0.2, 0.6, 0.2, -0.6, -0.8, -1.1;2015-04-07, -0.4, 1.4, -0.4, 0.2, 0.2, -0.2, 0.4, 0.0, -0.5, -1.0, -1.2; 2015-04-08, 0.0, 0.2, -0.4, 0.1, 0.1, -0.4, 0.2, 0.0, -0.4,- 1.0, -1.2; 2015-04-09, -0.5, 0.2, -1.1, 0.6, 0.6, -0.7, -0.4, -0.5, -0.7, -1.4, -1.7; 2015-04-10, 0.1, 0.4, -0.5, 0.2, 0.2, -0.5, -0.6, -0.5,-0.8, -1.5, -1.7;2015-04-13, 1.4, 1.9, 0.0, -0.7, -0.7, 0.2, 0.1, 0.3, -0.0, -0.6, -0.9. $codi - current report filing (8-k) 2015-04-13: william blair starts compass diversified $codi at outperform. Answer: The closing price of $codi will rise at 2015-04-14.
CIKM18	Context: The recent trend of $aal is as follows. date, open, high, low, close, adj-close, inc-5, inc-10, inc-15, inc-20, inc-25, inc-30; 2017-01-17, 1.7, 2.1, -0.2, -1.9, -1.9, 2.5, 0.9, 1.3, 2.0, 2.3, 2.0; 2017-01-18, -0.5, 0.3, -2.1, 1.9, 1.9, 0.2, -0.7, -0.7, 0.1, 0.2, 0.2; 2017-01-19, 0.8, 1.8, -0.8, -0.8, -0.8, 0.5, 0.2, -0.1, 0.8, 1.0, 1.1; 2017-01-20, -1.0, 0.3, -1.6, 1.6, 1.6, -1.1, -0.9, -1.6, -0.9, -0.6, -0.3; 2017-01-23, 2.0, 2.4, -0.4, -2.2, -2.2,0.8,1.5,0.6,1.1,1.6,1.8 2017-01-24, -1.0, 0.5, -1.6, 1.3, 1.3, -0.1, 0.3, -0.5, -0.3, 0.2, 0.5; 2017-01-25,-0.0, 0.6, -0.6, 0.8, 0.8, -0.8, -0.6, -1.1, -1.1, -0.5, -0.4; 2017-01-26, -2.3, 0.2, -2.5, 3.5, 3.5, -3.2, -3.7, -4.1, -4.4, -3.8, -3.7 2017-01-27,6.5,6.5,-0.5,-5.3,-5.3, 1.8, 1.4, 1.5, 0.9, 1.4, 1.7; 2017-01-30, 1.6, 2.3, -2.3, -4.4, -4.4, 5.5, 5.5, 5.9, 5.3, 5.6, 6.1. Answer: The closing price of $aal will fall at 2020-9-28.
BigData22	Context: The recent trend of $intc is as follows. date, open, high, low, close, adj-close, inc-5, inc-10, inc-15, inc-20, inc-25, inc-30; 2020-09-14, -1.1, 0.6, -1.1, 0.3, 0.3, -0.4, 1.3, 1.0, 0.5, 0.1, -0.2; 2020-09-15, -0.4, 1.2, -0.5, 1.2, 1.2, -1.1, -0.1, -0.1, -0.6, -1.0, -1.3; 2020-09-16, 0.3, 1.3, -0.4, 0.7, 0.7, -1.5, -0.9, -0.7, -1.2, -1.5, -1.9; 2020-09-17, -1.9, 0.3, -2.0, -0.1, -0.1, -0.9, -1.2, -0.5, -0.9, -1.3, -1.7; 2020-09-18, 0.9, 1.2, -1.7, -0.9, -0.9, 0.2, -0.4, 0.4, 0.1,-0.4,-0.7; 2020-09-21, -0.7, 0.0, -1.8, -0.3, -0.3, 0.7, -0.1, 0.7, 0.5, 0.0, -0.3; 2020-09-22, -0.1, 0.5, -1.0, 0.5, 0.5, 0.2, -0.4, 0.1, 0.1, -0.4, -0.7; 2020-09-23, 2.1, 2.7, -0.3, -2.3, -2.3, 1.9, 1.7, 2.1, 2.3, 2.0, 1.6;2020-09-24, -1.3, 1.0, -1.5, 0.7, 0.7, 0.7, 1.1, 1.0, 1.6, 1.3, 0.9; 2020-09-25, -2.0, 0.7, -2.4, 1.6, 1.6, -0.8, -0.4, -0.6, 0.0, -0.2, -0.5. Answer: The closing price of $intc will rise at 2020-9-28.

Table 10: Examples with context description and annotated answer for financial stock prediction task.

QA	Context: 11月9日上午，佛山市人民政府与徐工集团工程机械有限公司签约，将在南海区建设总投资20亿元的广东生产基地项目。 Answer: 建设投资方是徐工集团。
CEnQA	Context: 下表比较了花旗五年普通股的累积总回报，该股在纽约证券交易所上市，代码为201cc201d。日期，花旗，标普500，标普金融指数；2012-12-31，100.0，100.0，100.0; 2013-12-31，131.8，132.4，135.6；2014-12-31，137.0，150.5，156.2；2015-12-31，131.4，152.6，153.9；2016-12-31，152.3，170.8，188.9；2017-12-31，193.5，208.1，230.9。 Answer: 五年回报百分比0.935。
CConFinQA	Context: 以下是比较2007年和2008年净收入变动的分析，金额（以百万计）。<table class='wikitable'><tr><td>1</td><td></td><td>金额（以百万计）</td> </tr> <tr> <td>2</td ><td>2007年净收入</td> <td>$991.1</td> </tr> <tr><td>3</td> <td>零售电价</td> <td>-17.1（17.1）</td> </tr> <tr> <td>4</td> <td>购买的电力容量</td> <td>-12.0（12.0）</td> </tr> <tr> <td>5</td> <td>净批发收入</td> <td>-7.4（7.4）</td> </tr> <tr> <td>6</td> <td>其他</td> <td>4.6</td> </tr> <tr> <td>7</td><td>2008年净收入</td> <td>$959.2</td> </tr> </table>。 Answer: 2007年的净收入是$991.1。
EnQA	Context: The following table provides a comparison of the accumulated total return of citi common stock over a period of five years, which is listed on the nyse under the ticker symbol 201cc201d. date, citi, s&p 500, s&p financials. 31-dec-2011, 100.0, 100.0, 100.0; 31-dec-2012, 150.6, 116.0, 128.8; 31-dec-2013, 198.5, 153.6, 174.7; 31-dec-2014，206.3, 174.6, 201.3; 31-dec-2015, 197.8, 177.0, 198.2; 31-dec-2016, 229.3, 198.2, 243.4. Answer: The percent of the growth for s&p financials cumulative total return from 2013 to 2014 is 26.6.
ConFinQA	Context: The following is an analysis comparing the changes in net income between 2007 and 2008. <table class='wikitable'><tr> <td>1</td> <td> </td> <td> amount (in millions)</td> </tr> <tr><td>2</td> <td>2007 net revenue</td> <td>$991.1</td> </tr> <tr> <td>3</td><td>retail electric price</td> <td>-17.1 (17.1)</td> </tr> <tr> <td>4</td> <td>purchased power capacity</td> <td>-12.0 (12.0)</td> </tr> <tr> <td>5</td> <td>net wholesale revenue</td><td>-7.4 (7.4)</td> </tr> <tr> <td>6</td> <td>other</td> <td>4.6</td></tr> <tr> <td>7</td> <td>2008 net revenue</td> <td>$959.2</td> </tr> </table>。 Answer: The net revenue in 2008 is $959.2.

Table 11: Examples with context description and annotated answer for financial question answering task.

NER	Context: The proceeds of the equipment advances will be used solely to reimburse Borrower for the purchase of eligible equipment. Answer: Entity name is Borrower, and the corresponding entity type is person (PER).
CNER	Context: 数、数年数月数日，银保监会核准王小林先生本行董事任职资格。 Answer: 实体名称是银保监会, 类型是ORG; 实体名称是王小林, 类型是PER。

Table 12: Examples with context description and annotated answer for financial entity recognition task.

19CCKS	Context: 恒立实业(000622)遭证监会立案调查 涉嫌信披违规上海家化(600315)复星退出家化集团股权竞购。 Answer: 事件类型是信披违规，事件主体为恒立实业
20CCKS	Context: 2018年10月，中弘股份于发布公告称，截至2018年9月28日，公司逾期债务本息合计金额超过55亿元，全部为各类借款。 Answer: 事件类型是债务违约，事件主体是中弘股份
21CCKS	Context: 反观需求,由于生猪存栏量持续下降,而水产养殖启动缓慢,豆粕需求难以有效放大。供需失衡,导致国内外豆粕价格在豆类板块中呈现弱势。 Answer: 原因事件类型是供给减少，原因实体是生猪，结果事件类型是市场价格下降，结果实体是豆粕
22CCKS	Context: 易投配资为什么无法出金；晨曦航空股东高文舍拟减持不超3%股份；招行逾期第三个月的情况。 Answer: 事件类型是股东减持，事件主体是晨曦航空

Table 13: Examples with context description and annotated answer for financial event detection task.

NA	Context: 美今日凌晨，苹果发布iOS15.4首个测试版，更新后 Face ID将支持戴口罩解锁。按照官方描述，在升级到戴口罩使用面容ID之后，开机就会提醒是否要使用此功能，用户选择使用后，会要求再录入一次面容ID，这次主要是记录使用者的眼睛周围特征，用于戴口罩时候识别。 Answer: 文本的摘要为苹果iOS15.4测试版支持戴口罩的面容解锁。

Table 14: Examples with context description and annotated answer for financial text summarization task.