Golden Touchstone: A Comprehensive Bilingual Benchmark for Evaluating Financial Large Language Models

Anonymous ACL submission

Abstract

As large language models (LLMs) increasingly permeate the financial sector, there is a pressing need for a standardized method to comprehensively assess their performance. Existing financial benchmarks often suffer from limited language and task coverage, low-quality datasets, and inadequate adaptability for LLM evaluation. To address these limitations, we introduce Golden Touchstone, the first comprehensive bilingual benchmark for financial LLMs, encompassing eight core financial NLP tasks in both Chinese and English. Developed from extensive open-source data collection and industry-specific demands, this benchmark thoroughly assesses models' language understanding and generation capabilities. Through comparative analysis of major models such as GPT-4o, Llama3, FinGPT, and FinMA, we reveal their strengths and limitations in processing complex financial information. Additionally, we open-source Touchstone-GPT, a financial LLM trained through continual pre-training and instruction tuning, which demonstrates strong performance on the bilingual benchmark but still has limitations in specific tasks. This research provides a practical evaluation tool for financial LLMs and guides future development and optimization.

1 Introduction

The rapid development of both proprietary (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023; Anthropic, 2024; Team et al., 2023) and open-source Large Language Models (LLMs) (Touvron et al., 2023a,b; AI@Meta, 2024; Bai et al., 2023; Yang et al., 2024a; DeepSeek-AI, 2024; Young et al., 2024; Zeng et al., 2023; Baichuan, 2023; Gan et al., 2023; Zhang et al., 2022) has led to their increasing application in various fields, including finance (Wu et al., 2023; Lopez-Lira and Tang, 2023), healthcare (Thirunavukarasu et al., 2023; Tian et al., 2023), and law (Cui et al., 2023; Xiao et al., 2021). Among these, the financial sector shown in Figure.1 stands out as a critical area for LLM application due to its rich textual information and high practical value.

In recent years, a variety of advanced financial large language models (FinLLMs) have emerged, capable of specialized tasks such as financial sentiment analysis, content summarization, stock movement prediction, and question answering (Yang et al., 2023; Xie et al., 2023; Li et al., 2023; Chen et al., 2023; Zhang and Yang, 2023). These models leverage unique frameworks and tuning methods to enhance their performance on domain-specific benchmarks, offering robust solutions for real-world financial applications. However, existing financial benchmarks often suffer from limited language and task coverage, low-quality datasets, and inadequate adaptability for LLM evaluation, leading to poor evaluation results (Shah et al., 2022; Lu et al., 2023; Xie et al., 2023, 2024; Yang et al., 2023; Lei et al., 2023; Zhang et al., 2023).

To address these challenges, we propose Golden Touchstone, the first comprehensive bilingual benchmark for financial LLMs, encompassing eight core financial NLP tasks in both Chinese and English. Golden Touchstone provides high-quality datasets, task-aligned metrics, and instructional templates to guide LLMs in generating task-appropriate responses. We evaluated several state-of-the-art models, including GPT-4o, Qwen-2, Llama-3, FinGPT, and FinMA, on this benchmark. Results indicate that while these models perform well on tasks such as sentiment analysis and entity extraction, there is significant room for improvement in areas like stock movement prediction and classification tasks. Additionally, we open-source Touchstone-GPT, a financial LLM trained through domain-specific continual pre-training and instruction tuning, which serves as a new baseline for future research.

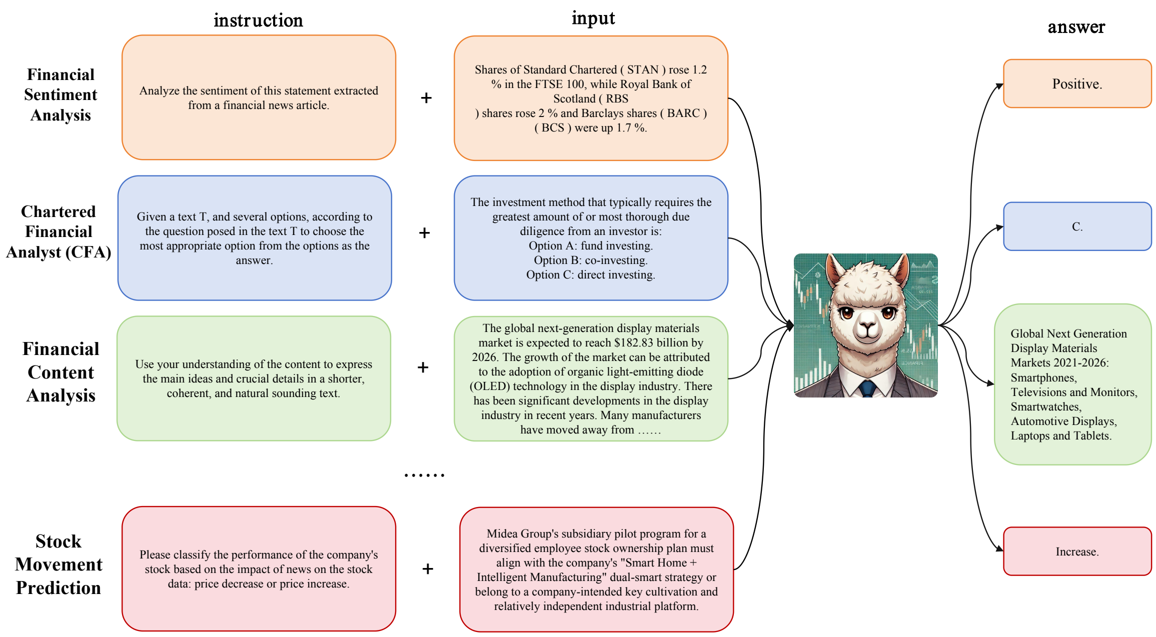

Figure 1: Financial large language models are designed to perform specialized tasks such as financial sentiment analysis, content analysis, stock movement prediction, and financial analyst level question answering by interpreting and processing structured instructions and various input data to generate precise outputs.

Our contributions are as follows:

- Introduction of Golden Touchstone, the first comprehensive bilingual benchmark for financial LLMs, encompassing 22 datasets across eight tasks in both Chinese and English.

- Evaluation of state-of-the-art LLMs and Fin-LLMs on Golden Touchstone, highlighting their strengths and limitations across various tasks.

- Open-sourcing of Touchstone-GPT, a financial LLM trained through domain-specific continual pre-training and instruction tuning, fostering further advancements in financial AI.

2 Benchmark Design

2.1 Current Benchmark Status

Existing open-source financial benchmarks have made significant strides in evaluating financial natural language processing (NLP) tasks. FLUE (Shah et al., 2022) pioneered English financial NLP evaluation, covering sentiment analysis, news classification, and other critical tasks. Subsequently, PIXIU (Xie et al., 2023) and FinBen (Xie et al., 2024) expanded task coverage, while in the Chinese domain, BBT-Benchmark (Lu et al., 2023) introduced the first comprehensive Chinese financial evaluation framework.

However, these benchmarks suffer from critical limitations:

- Inconsistent data quality across different tasks shown in Table.1.

- Challenges in numerical understanding by large language models (Shen et al., 2023; Akhtar et al., 2023; Schwartz et al., 2024)

- Lack of bilingual assessment capabilities

To address these limitations, we propose a unified benchmark that integrates high-quality financial datasets from both English and Chinese domains. Our approach aims to provide a more comprehensive and linguistically diverse evaluation of financial large language models (LLMs).

2.2 Golden Touchstone Benchmark Design

Addressing these critical limitations, we introduce the Golden Touchstone benchmark, a comprehensive bilingual evaluation framework designed to holistically assess financial language models. Conceptualized around two primary dimensions—task types and language coverage—our approach represents a significant departure from existing evaluation methodologies. The overview framework can be seen in Figure.2. Firstly, our Golden Touchstone

Table 1: Diversity of Financial Analysis Tasks Across Different Benchmarks

Benchmarks	Sent. Anal.	Classif.	Ent. Extr.	Rel. Extr.	Multi. Choice	Summ.	Quest. Ans.	Stock Pred.
FinGPT-Bench (Wang et al., 2023)	✓	✓	✓	✓				
FinBen (Xie et al., 2024)	✓	✓	✓	✓		✓	✓	✓
BBT-Fin (Lu et al., 2023)	✓	✓	✓	✓		✓	✓	
Fin-Eval (Zhang et al., 2023)					✓			
CFBenchmark (Lei et al., 2023)	✓	✓	✓			✓	✓	
Golden-Touchstone	✓	✓	✓	✓	✓	✓	✓	✓

benchmark categorizes financial NLP tasks across two dimensions:

1. Task type: Natural Language Understanding (NLU) and Natural Language Generation (NLG)

2. Language: English and Chinese

The benchmark strategically integrates English and Chinese datasets across eight sophisticated subtasks, spanning Natural Language Understanding (NLU) and Natural Language Generation (NLG). By carefully curating high-quality datasets from existing benchmarks and datasets. The detailed data statistics are shown in Appendix.A. The benchmark encompasses eight critical sub-tasks:

- **Sentiment Analysis**: Utilizing datasets like FPB (Malo et al., 2014) and FiQA-SA (Maia et al., 2018) and FinFE-CN (Lu et al., 2023)

- **Classification**: Integrating Headlines (Sinha and Khandait, 2021), FOMC (Shah et al., 2023), and LendingClub (Feng et al., 2023) and FinNL-CN (Lu et al., 2023) datasets

- **Entity Recognition**: Using NER (Alvarado et al., 2015) and FinESE-CN (Lu et al., 2023) datasets

- **Relation Extraction**: Incorporating FinRED (Sharma et al., 2022) and FinRE-CN (Lu et al., 2023) datasets

- **Multiple Choice**: Drawing from CFA (Yang et al., 2024b) and FinEval (Zhang et al., 2023) and CPA (Yang et al., 2024b) datasets

- **Summarization**: Employing EDTSUM (Zhou et al., 2021) and FinNA-CN (Lu et al., 2023) datasets

- **Question Answering**: Utilizing FinQA (Chen et al., 2021) and FinQA-CN and FinCQA-CN (Lu et al., 2023) datasets

- **Stock Movement Prediction**: Introducing news-based prediction using CMIN-US and CMIN-CN (Luo et al., 2023) datasets

By addressing previous benchmarks' limitations, our approach provides a more robust, comprehensive, and linguistically diverse framework for evaluating financial large language models. The benchmark not only expands task coverage but also addresses critical challenges in current financial NLP evaluation methodologies. Key methodological innovations include:

1. Replacing time-series tabular data with news-based stock prediction

2. Ensuring bilingual task alignment

3. Selecting datasets with consistent and high-quality labeling

4. Developing a unified evaluation approach across different financial NLP tasks

3 Experiments

3.1 Experimental Setup

Baselines. We conducted an extensive experimental evaluation against the Golden Touchstone Benchmark, incorporating a comprehensive array of models. For all models and inference tasks, we set the PyTorch and CUDA random seeds and configured the model with a greedy decoding strategy. This ensures reproducibility of experimental results and eliminates the influence of sampling decoding strategies on the final generated outputs. This included cutting-edge commercial models such as GPT-4o (OpenAI, 2023), alongside prominent open-source alternatives like Meta Llama-3 (AI@Meta, 2024) and Alibaba Qwen-2 (Yang et al., 2024a). Additionally, we integrated the latest and most influential financial language models (FInLLMs), namely FinGPT (Yang et al., 2023), FinMA (Xie et al., 2024), CFGPT (Li et al., 2023), and DISC-FinLLM (Chen et al., 2023).

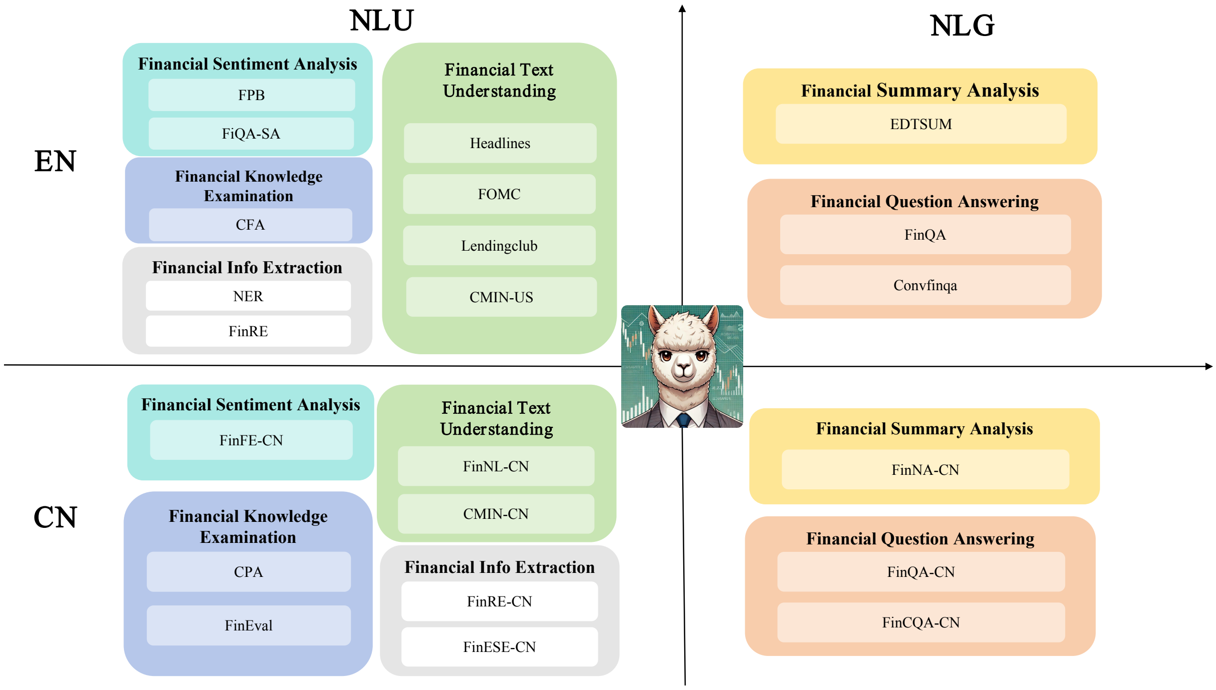

Figure 2: Financial NLP tasks are categorized along two dimensions: task types, divided into financial NLU (Natural Language Understanding) and financial NLG (Natural Language Generation), and language, categorized as English and Chinese. We organized the collected high-quality datasets along these axes.

These models were meticulously selected to represent a diverse spectrum of capabilities, ranging from general-purpose language understanding to specialized financial domain expertise. Our experiments aimed to rigorously assess the performance, robustness, and adaptability of each model within the context of financial data processing and analysis. The results provide valuable insights into the strengths and limitations of current state-of-the-art models, offering a foundation for future advancements in financial language modeling.

Touchstone-GPT Training. To further contribute to the research and development of FIn-LLMs and Financial benchmarks for LLMs, we have meticulously trained and open-sourced a Touchstone-GPT model. This initiative aims to serve as a valuable resource for advancing the field, providing a robust and versatile model that can be utilized for a wide range of financial language tasks. We adopted a two-stage training strategy comprising continuous pre-training and post-training, based on the Qwen-2 (Yang et al., 2024a) foundational model. During the continuous pre-training phase, we initially conducted pre-training on a high-quality financial corpus containing 100 billions tokens, which included textbooks, encyclopedias, research reports, news articles, and real-time analysis, all meticulously cleaned. In the post-training phase, we employed a standard instruction fine-tuning strategy, collecting, cleaning, and formatting a high-quality dataset of 300,000 instruction-response pairs shown in Tabel.4 and Table.5. To avoid catastrophic forgetting in general tasks, we also incorporated general-domain pre-training corpora (Gan et al., 2023) and instruction-tuning corpora (Peng et al., 2023) into continuous pre-training and post-training. This culminated in the final Touchstone-GPT model. We utilized Megatron(Shoeybi et al., 2019) for continuous pre-training and LlamaFactory(Zheng et al., 2024) for instruction post-training as our training frameworks, respectively. In this study, we employ an advanced model training setup using the AdamW optimizer (Kingma and Ba, 2014) with a learning rate of 1.0e-5, cosine annealing scheduler (Szegedy et al., 2016), and a 10% warmup ratio (He et al., 2016; Goyal et al., 2017) to enhance training stability and convergence. We enable gradient accumulation (Shoeybi et al., 2019) and checkpointing (Chen et al., 2016) to simulate larger batch sizes and reduce memory footprint. Training is conducted in mixed bfloat16 precision (Micikevicius et al., 2017; Wang and Kanwar, 2019) with DeepSpeed's ZeRO-1 optimization (Rajbhandari et al., 2020), reducing memory consumption and allowing for larger model training. This comprehensive setup optimizes efficiency and performance, providing an effective solution for large-scale deep learning model training.

Our training was conducted on 4 NVIDIA DGX servers, each equipped with 8 A100 GPUs, and spanned a period of 4 weeks. Inference was performed on a single NVIDIA DGX server with eight A100 GPUs, utilizing parallel batch inference. During the pre-training phase, we employed a data packing strategy (Krell et al., 2021) and batch dynamic right padding strategy in the instruction tuning phase (Wolf et al., 2020), while the inference phase incorporated a batch left padding strategy (Wolf et al., 2020).

3.2 Evaluation Results

In this section, we provide a detailed analysis of the evaluation and results for both the English and Chinese benchmarks. We discuss task-specific performances and identify key areas of strengths and weaknesses for each model. The following sections present insights for the English and Chinese benchmarks, each highlighting the differences in model capabilities across a variety of NLP tasks.

From the perspective of individual models in Figure.3, **GPT-4o** shows strong performance in sentiment analysis and structured tasks like multiple choice, indicating robust general language understanding capabilities. However, its weakness lies in relation extraction and detailed entity extraction, which require detailed understanding of complex financial relations. **FinMA-7B** stands out in sentiment tasks but lacks versatility, especially in question answering and summarization, likely due to the absence of targeted training for diverse NLP challenges. **Qwen-2-7B-Instruct** has a balanced yet modest performance, doing well in sentiment analysis but struggling significantly in question answering and summarization, which suggests a need for more specialized post-training. **Llama-3-8B-Instruct** excels in english NLU tasks, but shows limitations in tasks requiring chinese tasks, such as entity and relation analysis. The metrics of **FinGPT-8B-lora** indicating that the current level of domain-specific tuning is insufficient for complex financial tasks. Finally, **DISC-FinLLM-Full** and **CFGPT1-7B-Full** demonstrate moderate strengths in entity extraction tasks but lack the robustness needed for broader NLP capabilities, revealing significant gaps in financial language comprehension.

From a task perspective in Figure.4, we observe that **Sentiment Analysis** generally yields high scores across most models, particularly for the English benchmark, indicating that sentiment understanding, even in financial contexts, is relatively well addressed by these models. In contrast, **Relation Extraction** and **Question Answer** in financial domain exhibit notably lower performance, especially for the Chinese benchmark. These results suggest that capturing financial relationships and classifying detailed financial statements pose greater challenges, requiring more sophisticated training datasets or better model architectures. The LendingClub dataset in **Classification** is a specialized dataset in the field of risk control, requiring more targeted fine-tuning to achieve good results. **Stock Movement Prediction** also shows low performance across most models, with only a few models such as **GPT-4o** demonstrating relatively moderate performance, but it is still practically unusable, highlighting the inherent difficulty of this task. Market prediction relying solely on news information is likely insufficient; volume-price data and factor analysis can provide more comprehensive information. However, current large language models are unable to process these inputs, which is a significant area of future research. **Summarization** also stands out as a weak area for most models, with consistently low BLEU and Rouge scores, reflecting the challenges in generating concise, coherent summaries of complex financial text.

Overall, the insights suggest that while models like **GPT-4o**, **FinMA-7B**, and **Touchstone-GPT** have particular strengths in sentiment analysis and some structured tasks, the overall capability to handle comprehensive financial NLP tasks remains limited. Most models require targeted improvements, especially for relation extraction, summarization, question answering and stock movement prediction in both English and Chinese contexts. This calls for more domain-specific training and the development of specialized datasets that focus on capturing the detailed and often complex financial language, which is crucial for advancing the performance of financial large language models. Furthermore, while **Touchstone-GPT** demonstrates competitive performance across various tasks due to its robust pre-training and instruction tuning, ongoing refinements and specialized tuning efforts are needed to address specific deficiencies observed in tasks such as summarization, relation extraction, question an-

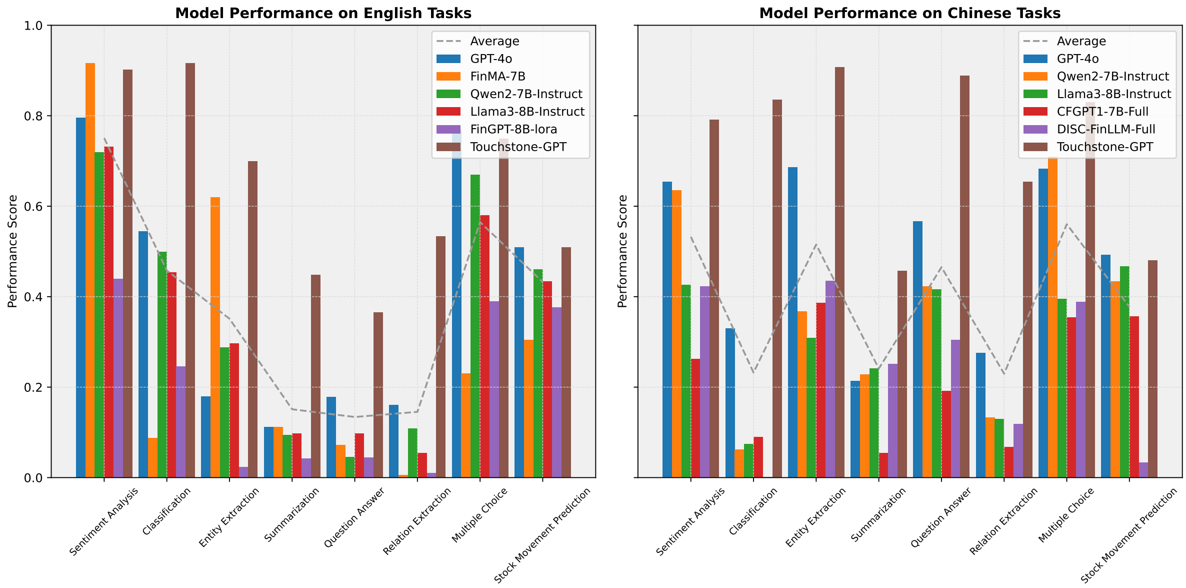

Figure 3: Comparison of Model Performance Across Tasks. Each subplot represents the performance of a models on both English and Chinese tasks. The bars indicate the model's performance on each task, while the dashed red line represents the average performance across all models for that task.

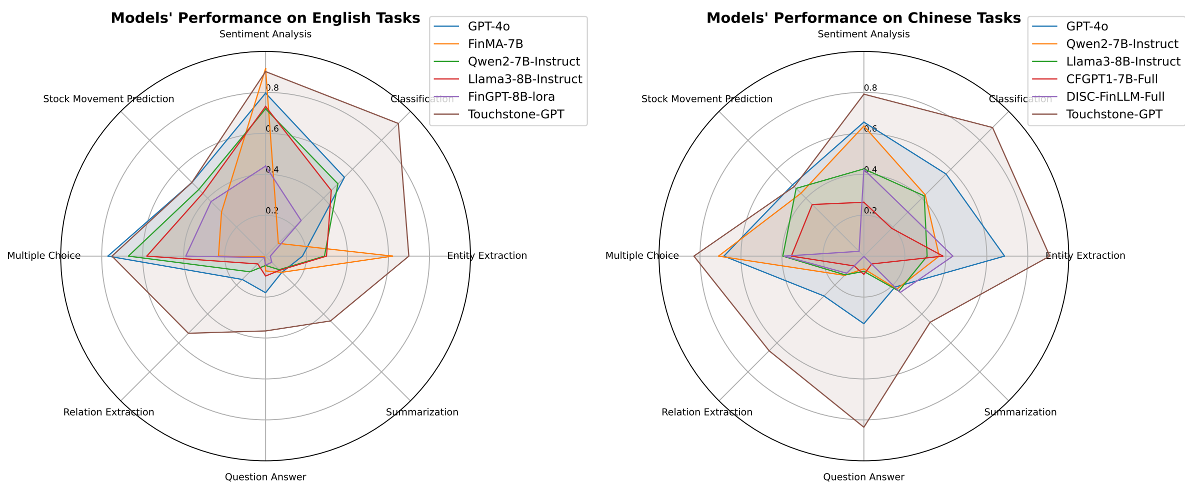

Figure 4: Comparison of different models' performance across tasks in the Golden Touchstone benchmark, illustrating average performance for English and Chinese tasks respectively.

swering and stock movement prediction.

4 Related Works

4.1 Financial Large Language Models

In recent years, large language models (LLMs) tailored for the financial domain have gained significant attention. BloombertGPT (Wu et al., 2023) marked the beginning of the FinLLM era. FinGPT (Yang et al., 2023) introduced an open-source framework emphasizing a data-centric approach with lightweight low-rank adaptation techniques. PIXIU (Xie et al., 2023) provided a comprehensive framework, presenting the first financial LLM fine-tuned on LLaMA with a 136K instruction dataset and evaluation benchmark. CFGPT (Li et al., 2023) developed a Chinese Financial Generative Pre-trained Transformer framework, encompassing dataset, model, and deployment capabilities. DISC-FinLLM (Chen et al., 2023) enhanced general LLMs through a multiple experts fine-tuning framework, expanding domain-specific capabilities.

4.2 Benchmarks for FinLLMs

The landscape of financial LLM benchmarks has evolved across English and Chinese domains. FLUE (Shah et al., 2022) introduced the first open-source benchmark for financial language understanding, covering five critical financial tasks. FinGPT (Yang et al., 2023) expanded the evaluation by introducing financial relation extraction and prompt-based instruction tuning. PIXIU (Xie et al., 2023) and FinBen (Xie et al., 2024) provided comprehensive financial task datasets. In the Chinese domain, BBT-Benchmark (Lu et al., 2023), FinEval (Zhang et al., 2023), and CFBenchmark (Lei et al., 2023) developed evaluation frameworks for financial NLP tasks.

Existing benchmarks still face significant challenges, including inconsistent data quality and task biases. This work aims to address these limitations by integrating high-quality bilingual datasets to create a more comprehensive FinLLM evaluation benchmark.

5 Conclusion

In this study, we introduce the Golden Touchstone benchmark, the inaugural structured and comprehensive bilingual benchmark specifically designed for English-Chinese financial NLP. This benchmark encompasses a wide array of financial NLP tasks, including Natural Language Understanding (NLU) and Natural Language Generation (NLG) across eight categories: Sentiment Analysis, Classification, Entity Extraction, Summarization, Stock Market Prediction, Question Answering, Relation Extraction, and Multiple Choice. By leveraging existing high-quality open-source financial datasets, we curated representative datasets and selected appropriate evaluation metrics for each task category. Utilizing these resources, we conducted extensive evaluations of current models such as GPT-4o and prominent open-source financial LLMs, including FinGPT and FinMA, thereby establishing performance benchmarks for financial LLMs within bilingual contexts. Moreover, we contributed to the community by open-sourcing Touchstone-GPT, a robust financial LLM that employs a two-stage training approach and has demonstrated superior input-based inference capabilities on the Golden Touchstone benchmark compared to GPT-4o. Our open-source initiative provides a bilingual English-Chinese evaluation framework aimed at fostering the sustainable development of LLMs in a multilingual financial environment.

6 Limitations

Despite these advancements, the benchmark currently exhibits certain limitations, including a limited range of NLG tasks and a focus solely on single-modality. Future enhancements will include the integration of additional NLG tasks, such as extended text generation for financial report analysis and more sophisticated sentiment assessments. Furthermore, we plan to expand the benchmark to cover other financial sectors such as insurance, cryptocurrency, and futures trading, thus broadening the scope and applicability of financial LLM assessments across diverse scenarios. Also, the performance of Touchstone-GPT on specific tasks within the Golden Touchstone benchmark, particularly in stock market prediction, requires further improvement. Our subsequent research will explore the incorporation of agent-based and retrieval-augmented generation (RAG) methods to augment the model's capabilities in numerical computation and real-time news analysis. Additionally, we aim to venture into multimodal modeling, integrating visual data and time-series data for tasks such as financial time-series forecasting, financial chart analysis, and content generation.

Table 2: Performance metrics of financial large language models across english tasks like Sentiment Analysis, Classification, and Summarization. Models include GPT-4o, Llama-3-8B, Qwen-2-7B, FinMA-7B, FinGPT-8B, and Touchstone-GPT. The best results of each dataset are marked in **bold**.

Task	Dataset	Metrics	GPT-4o	FinMA-7B full	Qwen-2-7B Instruct	Llama-3-8B Instruct	FinGPT-8B lora	Touchstone GPT
Sentiment Analysis	FPB	Weighted-F1	0.8084	**0.9400**	0.7965	0.7631	0.2727	0.8576
		ACC	0.8093	**0.9402**	0.8000	0.7660	0.3072	0.8557
	Fiqa-SA	Weighted-F1	0.8106	0.8370	0.6726	0.7515	0.5885	**0.8591**
		ACC	0.7702	0.8340	0.5957	0.7064	0.5872	**0.8638**
Classification	Headlines	Weighted-F1	0.7857	0.9739	0.7278	0.7006	0.4516	**0.9866**
		ACC	0.7931	0.9739	0.7252	0.7004	0.4331	**0.9866**
	FOMC	Weighted-F1	0.6603	0.3988	0.6112	0.4904	0.2758	**0.8788**
		ACC	0.6794	0.4274	0.6210	0.5625	0.2702	**0.8790**
	lendingclub	Weighted-F1	0.6730	0.1477	0.5938	0.5943	0.5480	**0.9783**
		MCC	0.1642	-0.6218	0.1714	0.1670	-0.1120	**0.9297**
Entity Extraction	NER	Entity-F1	0.1800	0.6200	0.2875	0.2973	0.0231	**0.6993**
Relation Extraction	FinRE	Relation-F1	0.1613	0.0054	0.1083	0.0540	0.0100	**0.5331**
Multiple Choice	CFA	Weighted-F1	**0.7700**	0.2200	0.6697	0.5800	0.3993	0.7497
		ACC	**0.7700**	0.2400	0.6700	0.5800	0.3800	0.7500
Summarization	EDTSUM	Rouge-1	0.1675	0.1566	0.1466	0.1467	0.0622	**0.5254**
		Rouge-2	0.0556	0.0491	0.0433	0.0429	0.0085	**0.3446**
		Rouge-L	0.1069	0.1060	0.0857	0.0930	0.0412	**0.4705**
		BLEU	0.1192	0.1361	0.0999	0.1085	0.0592	**0.4512**
Question Answering	Finqa	RMACC	0.1037	0.0497	0.0270	0.0470	0.0110	**0.2258**
	Convfinqa	RMACC	0.2540	0.0953	0.0644	0.1477	0.0772	**0.5053**
Stock Movement Prediction	CMIN-US	Weighted-F1	0.5025	0.2639	0.4112	0.3722	0.3379	**0.5036**
		ACC	**0.5149**	0.3446	0.5104	0.4955	0.4154	0.5144

Table 3: Performance metrics of financial large language models across chinese tasks like Sentiment Analysis, Classification, and Summarization. Models include GPT-4o, Llama-3-8B, Qwen-2-7B, CFGPT-7B, DISC-FinLLM, and Touchstone-GPT. The best results of each dataset are marked in **bold**.

Task	Dataset	Metrics	GPT-4o	Qwen-2-7B Instruct	Llama-3-8B Instruct	CFGPT1-7B Full	DISC-FinLLM Full	Touchstone GPT
Sentiment Analysis	FinFe-CN	Weighted-F1	0.6593	0.6274	0.3633	0.2528	0.4177	**0.7888**
		ACC	0.6500	0.6436	0.4891	0.2732	0.4292	**0.7936**
Classification	FinNL-CN	ORMACC	0.3303	0.0622	0.0747	0.0894	0.0011	**0.8360**
Entity Extraction	FinESE-CN	ORMACC	0.6867	0.3678	0.3088	0.3863	0.4346	**0.9074**
Relation Extraction	FinRE-CN	RMACC	0.2754	0.1330	0.1296	0.0678	0.1182	**0.6541**
Multiple Choice	FinEval	Weighted-F1	**0.7364**	0.7230	0.4432	0.3543	0.4288	0.7361
		ACC	**0.7353**	0.7235	0.4471	0.3529	0.4294	**0.7353**
	CPA	Weighted-F1	0.6312	0.6957	0.3421	0.3543	0.3451	**0.9238**
		ACC	0.6309	0.6960	0.3504	0.3553	0.3518	**0.9238**
Summarization	FinNA-CN	Rouge-1	0.3197	0.3326	0.3477	0.1018	0.3486	**0.5526**
		Rouge-2	0.1434	0.1597	0.1702	0.0263	0.1678	**0.3603**
		Rouge-L	0.2511	0.2644	0.2802	0.0650	0.2997	**0.5214**
		BLEU	0.1423	0.1541	0.1672	0.0238	0.1885	**0.3944**
Question Answering	FinQa-CN	RMACC	0.6578	0.5043	0.4540	0.1126	0.3949	**0.9214**
	FinCQa-CN	RMACC	0.4765	0.3422	0.3787	0.2714	0.2134	**0.8552**
Stock Movement Prediction	CMIN-CN	Weighted-F1	**0.4858**	0.3963	0.4497	0.3549	0.0329	0.4735
		ACC	0.4988	0.4723	0.4858	0.3584	0.0332	**0.4878**

References

AI@Meta. 2024. Llama 3 model card.

Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15391–15405.

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.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

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

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

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

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, et al. 2023. Discfinllm: 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, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Duanyu Feng, Yongfu Dai, Jimin Huang, Yifang Zhang, Qianqian Xie, Weiguang Han, Zhengyu Chen, Alejandro Lopez-Lira, and Hao Wang. 2023. Empowering many, biasing a few: Generalist credit scoring through large language models. *arXiv preprint arXiv:2310.00566*.

Ruyi Gan, Ziwei Wu, Renliang Sun, Junyu Lu, Xiaojun Wu, Dixiang Zhang, Kunhao Pan, Ping Yang, Qi Yang, Jiaxing Zhang, et al. 2023. Ziya2: Data-centric learning is all llms need. *arXiv preprint arXiv:2311.03301*.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv e-prints*, pages arXiv–1412.

Mario Michael Krell, Matej Kosec, Sergio P Perez, and Andrew Fitzgibbon. 2021. Efficient sequence packing without cross-contamination: Accelerating large language models without impacting performance. *arXiv preprint arXiv:2107.02027*.

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. 2023. Cfgpt: Chinese financial assistant with large language model. *arXiv preprint arXiv:2309.10654*.

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*.

Di Luo, Weiheng Liao, Shuqi Li, Xin Cheng, and Rui Yan. 2023. Causality-guided multi-memory interaction network for multivariate stock price movement prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12164–12176.

Macedo Maia, Siegfried Handschuh, Andre Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. pages 1941–1942.

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

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Eli Schwartz, Leshem Choshen, Joseph Shtok, Sivan Doveh, Leonid Karlinsky, and Assaf Arbelle. 2024. Numerologic: Number encoding for enhanced llms' numerical reasoning. *arXiv preprint arXiv:2404.00459*.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. 2022. Finred: A dataset for relation extraction in financial domain. In *Companion Proceedings of the Web Conference 2022*, pages 595–597.

Ruoqi Shen, Sébastien Bubeck, Ronen Eldan, Yin Tat Lee, Yuanzhi Li, and Yi Zhang. 2023. Positional description matters for transformers arithmetic. *arXiv preprint arXiv:2311.14737*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Ankur Sinha and Tanmay Khandait. 2021. Impact of news on the commodity market: Dataset and results. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*, pages 589–601. Springer.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2023. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. *arXiv preprint arXiv:2311.06025*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793*.

Shibo Wang and Pankaj Kanwar. 2019. Bfloat16: The secret to high performance on cloud tpus. *Google Cloud Blog*, 4(1).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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*.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Cehao Yang, Chengjin Xu, and Yiyan Qi. 2024b. Financial knowledge large language model. *arXiv preprint arXiv:2407.00365*.

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

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, et al. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *arXiv preprint arXiv:2209.02970*.

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

Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 4435–4439.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the event: Corporate events detection for news-based event-driven trading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2114–2124.

A Overview of Finance Evaluation Datasets by Task Type, Sample Sizes (Training, Validation, Test), and Evaluation Metrics

In the Table.4 and 5, we show the overview of finance evaluation datasets by task type, sample sizes (training, validation, test), and evaluation metrics

B Inference Template of Large Language Models

The Table.6 showcases how inference templates vary across different models. It is crucial to select the appropriate template for constructing correct inputs when inferring on the test sets of datasets. An incorrect template can significantly impair the performance of a model. We have observed that the underperformance of some large financial language models in some benchmarks is precisely due to not selecting the appropriate templates for evaluation. For more details of training and inference template, please refer to our open-source code repository on Github.

C Typical Case Study Analysis of Typical Financial NLP Tasks

This appendix provides a detailed case study analysis for some typical financial NLP tasks: financial sentiment analysis, text classification, entity extraction, and stock movement prediction. Each analysis is presented in a separate table, categorizing data sets, instructions, inputs, labels, and predictions from multiple models.

Financial Sentiment Analysis As demonstrated in Table 7, financial sentiment classification is one of the simpler tasks for benchmarks in financial NLP, resulting in high performance across all models tested. General-purpose models (GPT-4o, Qwen-2, Llama-3) provide not only the answer but also a detailed analysis, despite not being specifically fine-tuned on the FiQA-SA dataset. In contrast, specialized models (FinGPT, FinMA, Touchstone-GPT) that have undergone instruction tuning deliver straightforward, direct responses, illustrating their efficiency and focus in domain-specific applications.

Credit Rating Analysis In the classification task using the LendingClub dataset, which poses a challenging credit rating task, the models face a complex array of professional financial information evident in the input fields. Consequently, most models do not perform optimally. Among general models, GPT-4o exhibits the best performance, demonstrating the capabilities of large-scale models. In the realm of specialized financial language models, Touchstone-GPT, with its high-quality instruction tuning, significantly outperforms FinMA and FinGPT, which are only minimally tuned with Lora.

Financial NER In this information extraction task, most models demonstrated an understanding of the task intent and adhered to the instructions, signifying that even in the era of large language models, models like Qwen-2 and Llama-3 actually outperformed GPT-4o. In particular, specialized models such as FinMA and Touchstone-GPT, with more comprehensive instruction tuning, responded accurately and succinctly, highlighting their enhanced capability and focus on domain-specific tasks.

Stock Movement Prediction The Stock Movement Prediction task is one of the most challenging tasks, as it requires models to predict the daily fluctuations of the CMIN-US based solely on the 5-day news items. From the results in Table.4, it is evident that GPT-4o performed the best, yet it still falls short of practical utility. Even Touchstone-GPT, despite specialized instruction tuning, performed poorly. Our analysis suggests that the sentiment of news items may not reliably predict stock movements and that incorporating quantitative data is essential for achieving practical model performance. Similar conclusions were drawn from experiments with traditional machine learning methods like XGBoost. Nevertheless, aside from simpler tasks like financial sentiment analysis, we also pose challenging tasks such as stock prediction, which are closer to real-world applications, leaving more room for benchmark challenges and exploration. Multimodal fusion of news and quantitative data represents a promising future direction, and we look forward to seeing models excel in these tasks.

Due to the similar performance of models across corresponding task types on the Chinese benchmark, we will not reiterate these comparisons and analysis here.

Table 4: Overview of English Finance Evaluation Datasets by Task Type, Sample Sizes (Training, Validation, Test), and Evaluation Metrics

Task	Dataset	Train	Valid	Test	Metrics
Sentiment Analysis	**FPB**	3100	776	970	Weighted-F1 ACC
	FiQA-SA	750	188	235	Weighted-F1 ACC
Classification	**Headlines**	71900	10300	20500	Weighted-F1 ACC
	FOMC	1984	-	496	Weighted-F1 ACC
	lendingclub	9417	1345	2691	Weighted-F1 MCC
Entity Recognition	**NER**	408	103	98	Entity-F1
Relation Extraction	**FinRE**	27558	-	5112	Relation-F1
Multiple Choice	**CFA**	1884	100	20	Weighted-F1 ACC
Summarization	**EDTSUM**	8000	-	2000	ROUGE BLEU
Question Answering	**FinQa**	6251	883	1147	RMACC
	ConvfinQa	8890	2210	1490	RMACC
Stock Movement Prediction	**CMIN-US**	88297	9010	8480	Weighted-F1 ACC

Table 5: Overview of Chinese Finance Evaluation Datasets by Task Type, Sample Sizes (Training, Validation, Test), and Evaluation Metrics

Task	Dataset	Train	Valid	Test	Metrics
Sentiment Analysis	**FinFE-CN**	16157	2020	2020	Weighted-F1 ACC
Classification	**FinNL-CN**	7071	884	884	ORMACC
Entity Extraction	**FinESE-CN**	14252	1781	1782	ORMACC
Relation Extraction	**FinRE-CN**	13486	1489	3727	RMACC
Multiple Choice	**FinEval**	1071	170	3340	Weighted-F1 ACC
	CPA	6268	1444	6	Weighted-F1 ACC
Summarization	**FinNA-CN**	28800	3600	3600	ROUGE BLEU
Question Answering	**FinQa-CN**	19906	2469	2480	RMACC
	FincQa-CN	21965	2741	2745	RMACC
Stock Movement Prediction	**CMIN-CN**	214873	23904	23571	Weighted-F1 ACC

Table 6: Comparison of Inference Templates Across Different Models for Dataset Evaluation

<table>
<tr><th>Model</th><th>Template</th></tr>
<tr><td>GPT-4o</td><td>"<|im_start|>system{{system_prompt}}<|im_end|>\n"
"<|im_start|>user{{instruction}}{{input}}<|im_end|>\n"
"<|im_start|>assistant\n"</td></tr>
<tr><td>Qwen-2</td><td>"<|im_start|>system{{system_prompt}}<|im_end|>\n"
"<|im_start|>user{{instruction}}{{input}}<|im_end|>\n"
"<|im_start|>assistant\n"</td></tr>
<tr><td>Llama-3</td><td>"<|start_header_id|>system<|end_header_id|>"
"{{system_prompt}}<|eot_id|>\n"
"<|start_header_id|>user<|end_header_id|>"
"{{instruction}}{{input}}<|eot_id|>\n"
"<|start_header_id|>assistant<|end_header_id|>\n"</td></tr>
<tr><td>FinGPT</td><td>"Instruction:{{instruction}}"
"Input{{input}}\nAnswer:"</td></tr>
<tr><td>FinMA</td><td>"Human:{{instruction}}{{input}}\n"
"Assistant:\n"</td></tr>
<tr><td>CFGPT</td><td>"{{instruction}}{{input}}\n"</td></tr>
<tr><td>DISC-FinLLM</td><td>"<reserved_102>
{{instruction}}{{input}}<reserved_103>"</td></tr>
<tr><td>Touchstone</td><td>"<|im_start|>system{{system_prompt}}<|im_end|>\n"
"<|im_start|>user{{instruction}}{{input}}<|im_end|>\n"
"<|im_start|>assistant\n"</td></tr>
</table>

Table 7: Detailed Case Study Analysis of Financial Sentiment Analysis on the FiQA-SA dataset, Financial Text Classification on the LendingClub dataset, Financial Entity Extraction on NER dataset, Stock Movement Prediction on DJIA dataset.

Dataset	Instruction	Input	Label	Model	Prediction
FiQA-SA	What is the sentiment of the following financial post: Positive, Negative, or Neutral?	@Joerazorback This time I think $AAPL will miss again. Generally speaking market is near top when everyone thinks that market goes only up.	Negative	GPT-4o	The sentiment of this financial post is **Negative**. The user expresses doubt about Apple's performance, suggesting that they think the stock will miss expectations. Additionally, the phrase about the market being "near top" implies a bearish outlook, indicating that the market may decline after its current uptrend.
				Qwen-2	The sentiment of the given financial post is Negative. The user is predicting that Apple's stock price ($AAPL) will miss or fail to meet expectations, and they also suggest that the market might be nearing its peak when everyone is optimistic about continuous upward movement.
				Llama-3	The sentiment of this post is Negative. The author is predicting that Apple's stock (AAPL) will miss expectations, which is a negative view. Additionally, the author is suggesting that the market is near a top, implying that a downturn is likely, which is also a negative sentiment.
				FinGPT	Negative
				FinMA	Negative
				Touchstone GPT	Negative
LendingClub	Assess the client's loan status based on the following loan records from Lending Club. Respond with only 'good' or 'bad'.	The client has attributes as follows: ... The state of Employment Length is 2 years. The state of Home Ownership is RENT. The state of Verification Status is Verified. The state of Annual Income is 83000.0.	Good	GPT-4o	Good
				Qwen-2	Good
				Llama-3	Good
				FinGPT	Bad
				FinMA	Bad
				Touchstone GPT	Good
NER	In the sentences extracted from financial agreements in U.S. SEC filings, identify the named entities.	WHEREAS: (A) The Borrower intends to develop a plant in France for the production of solar grade silicon.	Borrower, PER; France, LOC	GPT-4o	France, LOC
				Qwen-2	Borrower, PER; France, LOC; solar grade silicon, ORG
				Llama-3	Borrower, PER; France, LOC
				FinGPT	The Borrower intends to develop a plant in France for the production of solar grade silicon
				FinMA	Borrower, PER; France, LOC
				Touchstone GPT	Borrower, PER; France, LOC
CMIN-US	Based on the news about the stock in the past five trading days......	NEWS1: Biden Urges CEOs to Improve U.S. Cybersecurity After Attacks...	0	GPT-4o	Rise
				Qwen2	Decrease
				Llama-3	Decrease
				FinGPT	""
				FinMA	Rise
				Touchstone GPT	Rise