FinNLP-AgentScen-2024 Shared Task: Financial Challenges in Large Language Models - FinLLMs

Qianqian Xie\textsuperscript{1} and Jimin Huang\textsuperscript{1} and Dong Li\textsuperscript{9} and Zhengyu Chen\textsuperscript{9} and Ruoyu Xiang\textsuperscript{1} and Mengxi Xiao\textsuperscript{9} and Yangyang Yu\textsuperscript{7} and VijayaSai Somasundaram\textsuperscript{8} and Kailai Yang\textsuperscript{2} and Chenhan Yuan\textsuperscript{2} and Zheheng Luo\textsuperscript{2} and Zhiwei Liu\textsuperscript{2} and Yueru He\textsuperscript{11} and Yuechen Jiang\textsuperscript{7} and Haohang Li\textsuperscript{7} and Duanyu Feng\textsuperscript{5} and Xiao-Yang Liu\textsuperscript{5,11} and Benyou Wang\textsuperscript{4} and Hao Wang\textsuperscript{5} and Yanzhao Lai\textsuperscript{6} and Jordan Suchow\textsuperscript{7} and Alejandro Lopez-Lira\textsuperscript{8} and Min Peng\textsuperscript{9} and Sophia Ananiadou\textsuperscript{2,10}.

\textsuperscript{1}The Fin AI, Singapore; \textsuperscript{2}University of Manchester, UK; \textsuperscript{3}Open Finance, USA; \textsuperscript{4}Chinese University of Hong Kong, Shenzhen, China; \textsuperscript{5}Sichuan University, China; \textsuperscript{6}Southwest Jiaotong University, China; \textsuperscript{7}Stevens Institute of Technology, USA; \textsuperscript{8}University of Florida, USA; \textsuperscript{9}Wuhan University, China; \textsuperscript{10}Archimedes RC, Greece; \textsuperscript{11}Columbia University, USA.

Abstract

Despite the promise of large language models (LLMs) in finance, their capabilities for comprehensive analysis and decision-making remain largely unexplored, particularly in areas such as financial text analysis, generation, and decision-making. To evaluate the capabilities of LLMs in finance, we introduce an LLMs-based financial shared task featured at IJCAI FinNLP-AgentScen-2024, FinLLMs Challenge. This challenge includes three subtasks: financial classification, financial text summarization, and single stock trading. In this paper, we provide an overview of these tasks and datasets, summarize participants’ methods, and present their experimental evaluations, highlighting the effectiveness of LLMs in addressing diverse financial challenges. To the best of our knowledge, the FinLLMs Challenge is one of the first challenges for assessing LLMs in the financial area. In consequence, we provide detailed observations and take away conclusions for future development in this area.

1 Introduction

FinNLP workshop is a platform committed to promoting international cooperation and the exchange of knowledge in applying Natural Language Processing (NLP) within the ever-evolving realm of FinTech. In recent years, the FinNLP series has delved into the intersection of FinTech and NLP, uncovering significant challenges and guiding future research directions, along with proposing a series of diverse share task in financial domain, involving Sentence boundary detection (Azzi et al., 2019; Au et al., 2020), learning semantic representations (Maarouf et al., 2020) and semantic similarities (Kang et al., 2021; Kang and El Maarouf, 2022; Chen et al., 2023).

Recent studies (Xie et al., 2024b, 2023; Lopez-Lira and Tang, 2023; Liu et al.; Xie et al., 2024a) have highlighted the significant potential of advanced large language models (LLMs) in finance, particularly for tasks involving financial text analysis and prediction. These models can transform traditional methodologies by boosting efficiency and enhancing the accuracy of predictive models. Although several approaches have achieved remarkable performance with LLMs, their capabilities of comprehensive analysis and decision-making for finance remain largely unexplored.

To explore the ability of LLMs from these facets, we propose a LLMs-based financial shared task, FinLLMs Challenge. This challenge includes three published datasets designed to address a range
of financial challenges effectively and comprehensively. These tasks include financial classification, financial text summarization, and single stock trading. For financial classification tasks, we utilize the FinArg AUC dataset (Chen et al.), which provides financial texts paired with two opinions. Using this data, we provide a prompt template to classify the text as either a claim or a premise. For financial text summarization tasks, we introduce the EDTSum dataset (Zhou et al., 2021), which is used to summarize given financial news articles, along with a recommended prompt template. For decision-making tasks, we provide the fintrade dataset (Xie et al., 2024a), which can be leveraged by FinMem (Yu et al., 2023) agent framework, allowing LLMs to generate one of three trading decisions from “buy”, “sell” or “hold.”

This paper overviews three subtasks and datasets in the FinLLMs Challenge, summarizes participant methods, and evaluates their experiments to explore LLM’s capabilities in financial analysis and prediction. Our comprehensive evaluation highlights the strengths and limitations of current methodologies, showcasing the effectiveness of LLMs across various financial tasks and the potential of domain-specific instruction tuning in the financial sector.

2 Tasks and Datasets

We provide three tasks for assessing the performance of LLMs in finance, as shown in Table 1.

Task 1: Financial Classification. This task, derived from FinBen (Xie et al., 2024a), concentrates on argument unit classification to identify and categorize individual units or segments of arguments within the discourse found in earnings conference call data. The objective of this task is to evaluate the capability of LLMs to distinguish and classify texts as premises or claims. The dataset (Chen et al.) includes 7.75k training examples and 969 testing examples for sentence categorization into claims or premises. We use two metrics to evaluate classification capability, including Macro F1 and Accuracy. Macro F1 score is used as the final ranking metric.

Task 2: Financial Text Summarization. Derived from FinBen (Xie et al., 2024a), this task aims to evaluate the ability of LLMs in producing coherent summaries. The dataset (Zhou et al., 2021) includes 8,000 training instances and 2,000 test instances for summarizing financial news articles succinctly. We utilize two metrics including ROUGE (1, 2, and L) (Lin, 2004) and BERTScore (Zhang et al., 2020), to evaluate generated summaries in terms of relevance. ROUGE-1 score is used as the final ranking metric.

Task 3: Single Stock Trading. Building on the Trading task in FinBen (Xie et al., 2024a), this evaluation aims to rigorously assess the ability of LLMs to execute complex trading decisions, addressing the critical challenge of human limitations in processing large volumes of data rapidly. We construct and provide the first public dataset of 291 distinct data points, which allows to test the models’ decision-making capabilities in stock trading based on the agent framework. Participants are required to analyze the dataset, adapt or develop LLM frameworks for financial data interpretation, and implement algorithms to generate sophisticated trading strategies based on the FinMem agent framework (Yu et al., 2023).

We employ the following prompt for model inputs:

Instruction: [task prompt] Context: [input context] Response: [output].

[input text] represents the financial investment information provided in the prompt. The [output] must adhere strictly to the following JSON format, without any additional content:

```json
{
    "investment_decision": string,
    "summary_reason": string,
    "short_memory_index": number,
    "middle_memory_index": number,
    "long_memory_index": number,
    "reflection_memory_index": number
}
```

We offer a comprehensive assessment of profitability, risk management, and decision-making prowess by a series of metrics, including Sharpe Ratio (SR) (Sharpe, 1994), Cumulative Return (CR), Daily (DV) and Annualized volatility (AV), and Maximum Drawdown (MD). Sharpe Ratio (SR) score is used as the final ranking metric, which is calculated by dividing the portfolio’s average excess return ($R_p$) over the risk-free rate ($R_f$) by its volatility ($\delta_p$).

$$\text{SharpeRatio} = \frac{R_p - R_f}{\delta_p}$$ (1)

Where $R_p$ represents the portfolio’s average excess return, $R_f$ is the risk-free rate, $\delta_p$ is the port-
3 Model Cheating Detection

To assess the risk of model cheating, where models improperly access test data during training (Zhou et al., 2023), we introduce a new metric called the Data Leakage Test (DLT). This metric builds on previous research (Wei et al., 2023; Xu et al., 2024) and aims to quantify the likelihood that a model is exposed to the test set during its training process.

The DLT measures the risk by comparing how well the LLM performs on the training data versus the test data. We feed the training and test sets separately into the model and measure its perplexity on each. The DLT score is then calculated by subtracting the perplexity on the training set from the perplexity on the test set:

\[
DLT = PPL(D_{test}) - PPL(D_{train}) \tag{2}
\]

where PPL is the perplexity given the dataset inputs.

A larger DLT score suggests the LLM is less likely to have been exposed to the test data during training. Conversely, a smaller DLT score implies the LLM is more likely to have seen the test data during training, suggesting a higher likelihood of cheating.

4 Participants and Automatic Evaluation

35 teams have registered for the FinLLMs Challenge, out of which 8 teams have submitted their LLMs solution papers. In this section, we provide a detail overview of the LLMs based solutions for each paper. For task 1 and 2, we employ two baseline models from (Xie et al., 2024a): GPT-4 (OpenAI et al., 2024) and LlaMA3-8B\(^1\). GPT-4, developed by OpenAI, is the state-of-the-art commercialized large language model, whereas LlaMA3-8B, created by MetaAI, is an open-source large language model built with more training data than its predecessor, LlaMA2.

\(^1\text{https://LlaMA.meta.com/LlaMA3/}

4.1 Task 1: Financial Classification

Table 2 presents the experimental results of task 1. BAI-Arg LLM (Srivastava, 2024) leverages LlaMA3-8B which is fine-tuned via QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023). L3iTC (Pontes et al., 2024), utilizes Mistral-7BInst-v0.3 to be finetuned with 4-bit quantization and LoRA (Hu et al., 2021) to reduce the memory usage of LLMs. Wealth Guide (Das et al., 2024) fine-tuned DistilBERT for financial text classification. CatMemo (Cao et al., 2024) finetuned Mistral-7B with fused datasets of both task 1 and task 2 via LoRA. Upaya (Jindal et al., 2024) utilizes distillation-based fine-tuning of the LlaMA3-8B method to learn the rationale generated by LlaMA3-3 (70B parameters) and labels.

4.2 Task 2: Financial Text Summarization

Table 3 presents the experimental results of task 2. University of Glasgow (Guo et al., 2024) investigated three common strategies: few-shot learning, fine-tuning, and reinforcement learning, to adapt LLMs to abstract news into concise summaries, with the fine-tuned model ranked first on the leaderboard. Upaya (Jindal et al., 2024) also utilized distillation-based fine-tuning of the LlaMA3-8B method, which leveraged the augmented datasets with a maximum of 5 relevant sentences from the original news text that are relevant to the given summary via LlaMA3-70B. Finance Wizard (Lee and Lay-Ki, 2024) introduced a pipeline approach. Based on LlaMA3-8B foundation, they first continual pretrained the model with the financial corpus, then they tailored it to the finance domain with multi-task instruction data, and finally fine-tune it for specific tasks. Revelata (Kawamura et al., 2024) first designed a set of prompts by systematically changing parts of the prompts and then finetuning Meta-LlaMA3-8B-Instruct on each of these prompts separately. L3iTC (Pontes et al., 2024) introduced Mistral-7B-Inst-v0.3 model, a finetuning model combining 4-bit quantization and LoRA to optimize the finetuning process.
Table 2: Evaluation Results of Task 1 - Financial Classification.

<table>
<thead>
<tr>
<th>Team</th>
<th>Method</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
<th>BERTScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAI-Arg LLM</td>
<td>LLaMA3-8B + QLoRA + Finetuning</td>
<td>0.7612</td>
<td>0.7626</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albatross</td>
<td></td>
<td>0.7575</td>
<td>0.7575</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3iTC</td>
<td>Mistral-7B + 4 Bit + Lora + Finetuning</td>
<td>0.7543</td>
<td>0.7544</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wealth Guide</td>
<td>DistilBERT + Finetuning</td>
<td>0.7509</td>
<td>0.7513</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finance Wizard</td>
<td></td>
<td>0.7262</td>
<td>0.7286</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CatMemo</td>
<td>Mistral-7B + Task 1 + Task 2 + QLoRA + Finetuning</td>
<td>0.7086</td>
<td>0.7110</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upaya</td>
<td>LLaMA3-8B + Distillation + Finetuning</td>
<td>0.7083</td>
<td>0.7090</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vidra</td>
<td></td>
<td>0.7070</td>
<td>0.7079</td>
<td></td>
<td></td>
</tr>
<tr>
<td>jt</td>
<td></td>
<td>0.4630</td>
<td>0.4933</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (Xie et al., 2024a) GPT-4</td>
<td></td>
<td>0.6000</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (Xie et al., 2024a) LLaMA3-8B</td>
<td></td>
<td>0.5100</td>
<td>–</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Evaluation results of Task 2 - Financial Text Summarization.

<table>
<thead>
<tr>
<th>Team</th>
<th>Method</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
<th>BERTScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Glasgow</td>
<td>LLaMA3-8B + 4 bit + Instruction tuning</td>
<td>0.5346</td>
<td>0.3581</td>
<td>0.4922</td>
<td>0.9117</td>
</tr>
<tr>
<td>Upaya</td>
<td>LLaMA3-8B</td>
<td>0.5295</td>
<td>0.3582</td>
<td>0.4860</td>
<td>0.9106</td>
</tr>
<tr>
<td>Finance Wizard</td>
<td>LLaMA3-8B</td>
<td>0.5210</td>
<td>0.3406</td>
<td>0.4735</td>
<td>0.9084</td>
</tr>
<tr>
<td>Reveleta</td>
<td>LLaMA3-8B-Instruction + Finetuning + Lead-in phrase</td>
<td>0.5004</td>
<td>0.3330</td>
<td>0.4644</td>
<td>0.9070</td>
</tr>
<tr>
<td>Albatross</td>
<td>Mistral-7B-Inst-v0.3 + Lora + Finetuning</td>
<td>0.3691</td>
<td>0.2011</td>
<td>0.3227</td>
<td>0.8720</td>
</tr>
<tr>
<td>L3iTC</td>
<td></td>
<td>0.3661</td>
<td>0.1872</td>
<td>0.3046</td>
<td>0.8750</td>
</tr>
<tr>
<td>Wealth Guide</td>
<td></td>
<td>0.3089</td>
<td>0.1795</td>
<td>0.2819</td>
<td>0.8596</td>
</tr>
<tr>
<td>Vidra</td>
<td></td>
<td>0.2850</td>
<td>0.1348</td>
<td>0.2286</td>
<td>0.8587</td>
</tr>
<tr>
<td>Baseline</td>
<td>GPT-4</td>
<td>0.2000</td>
<td>–</td>
<td>–</td>
<td>0.6700</td>
</tr>
<tr>
<td>Baseline</td>
<td>LLaMA3-8B</td>
<td>0.1400</td>
<td>–</td>
<td>–</td>
<td>0.6000</td>
</tr>
</tbody>
</table>

4.3 Task 3: single stock trading

Table 4 presents the experimental results of task 3. Wealth Guide (Das et al., 2024) utilizes the LLaMA2-13B model with zero-shot and few-shot fine-tuning, integrating sentiment scores and stock prices for trading predictions. CatMemo (Cao et al., 2024) also utilizes the Mistral-7B model fine-tuned using PEFT and LoRA techniques, integrating datasets from Task 1 and Task 2.

5 Discussion

5.1 Task 1: financial classification

As shown in Table 2, the experimental results highlight the remarkable performance of various teams in the financial text classification task, all of which employed fine-tuning with task-specific training data. Notably, BAI-Arg LLM, utilizing the LLaMA3-8B model with fine-tuning, carefully designed prompts, and semantically similar examples, achieved the best performance with an F1 score of 0.7612 and an accuracy of 0.7626. This performance surpasses both GPT-4 and the backbone model LLaMA3-8B, fully demonstrating the benefits of fine-tuning with task-specific data in financial classification tasks based on LLMs.

Compared to L3iTC and other teams, BAI-Arg LLM’s performance underscores the importance of both prompt templates and semantically similar examples for fine-tuning LLMs on financial classification tasks. This indicates the necessity for LLMs to be adapted to financial classification tasks through prompt engineering and few-shot learning. Moreover, their performance surpasses that of DistilBERT, proving the potential of LLMs compared to traditional BERT-based methods.

5.2 Task 2: financial text summarization

As shown in Table 3, the experimental results highlight the potential of LLMs in financial text summarization. Leveraging LLMs facilitates the generation of high-quality summaries, thereby enhancing both efficiency and quality. Similar to financial classification tasks, performance improves significantly with task-specific fine-tuning.

Notably, methods employing LLMs generally achieve high scores across various metrics. For instance, the University of Glasgow team achieved a ROUGE-1 score of 0.5346 using the instruc-
tion tuning method, while the Upaya team scored 0.5295 with a distillation-based fine-tuning approach. These results indicate that LLMs, when fine-tuned with appropriate methods, can effectively capture and condense the main information from financial texts into clear and concise summaries. The Finance Wizard team employed continual pretraining, multi-task fine-tuning, and specific task fine-tuning with LLaMA3-8B, demonstrating substantial benefits in overall performance. These approaches outperform GPT-4 and the backbone model LLaMA3-8B, underscoring that fine-tuning and continual pretraining can lead to significant improvements in financial text summarization tasks.

5.3 Task 3: single stock trading

Table 4 presents the performance of various teams using different LLMs in single stock trading tasks. The experimental results indicate that our challenge and provided resources have indeed contributed to advancements in financial investment decision-making. Participants utilized these resources to develop effective strategies and models, thereby improving their performance in this domain. The results reveal the potential of LLMs in financial investment decision-making, especially when integrated within an agent framework.

Notably, methods employing LLMs have achieved remarkable performance in key metrics. For instance, the Wealth Guide team achieved the highest Sharpe Ratio score of 0.9264 using a sentiment-score-based trading prediction model, indicating the effectiveness of LLMs in predicting market trends. In terms of Cumulative Return, the Wealth Guide team’s model also showed significant promise. These findings underscore the potential of LLMs to enhance trading strategies and improve investment outcomes when fine-tuned and applied within an agent framework. However, the CatMemo team’s use of the Mistral-7B method recorded lower performance, highlighting the variability in effectiveness depending on the specific model and approach used. Despite this, the overall results suggest that with proper tuning and integration, LLMs can be powerful tools in financial stock trading based on the agent framework.

5.4 Model Cheating Detection

We further conducted a Model Cheating Detection analysis using our Data Leakage Test (DLT) on teams that disclosed their training procedures in Task 1 and Task 2. The results, summarized in Table 5, reveal no evidence of model cheating among these teams.

<table>
<thead>
<tr>
<th>Team</th>
<th>Task</th>
<th>Rank</th>
<th>DLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAI-Arg LLM</td>
<td>Task1</td>
<td>1</td>
<td>38.90</td>
</tr>
<tr>
<td>L3iTC</td>
<td>Task1</td>
<td>3</td>
<td>2.26</td>
</tr>
<tr>
<td>Upaya</td>
<td>Task2</td>
<td>2</td>
<td>0.83</td>
</tr>
<tr>
<td>Finance Wizard</td>
<td>Task2</td>
<td>3</td>
<td>1.74</td>
</tr>
</tbody>
</table>

Table 5: Evaluation—Model Cheating Detection

For instance, “BAI-Arg LLM”, the top-performing team in Task 1, exhibited a DLT score of 38.90, significantly above zero, effectively ruling out any data leakage concerns. Similarly, teams like “L3iTC” and “Finance Wizard” consistently displayed DLT scores exceeding 1.5, indicating a negligible risk of data leakage.

These findings suggest that the majority of the participating teams adhered to the competition’s ethical guidelines. Furthermore, even with this strict adherence, the impressive performance improvements these teams achieved, exceeding the original benchmarks, underscore the immense potential of LLMs within the financial realm.

6 Conclusion

In this paper, the FinLLMs Challenge has demonstrated the efficacy and potential of LLMs in the domain of financial investment decision-making. Our challenge, along with the resources provided, has significantly contributed to advancing this field. Participants utilized these resources to develop effective strategies and models, which led to improved performance across various tasks. The experimental results from tasks such as financial classification, text summarization, and single stock trading highlight the considerable value of LLMs.
based approaches. The overall trend indicates that performance improves with increasing model size and advancements in fine-tuning and prompt engineering. These findings offer valuable insights for future research in financial tasks using LLMs. The success of this challenge underscores the importance and impact of collaborative efforts in pushing the boundaries of AI applications in finance.

Acknowledgments

We would like to thank all the anonymous reviewers and area chairs for their comments. This work is supported by the project JPNP20006 from New Energy and Industrial Technology Development Organization (NEDO). This work has also been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. This is also supported by National Science and Technology Major Project (No.2021ZZD0113304), National Natural Science Foundation of China (U23A20316), Key R&D Project of Hubei Province (2021BAA029), General Program of National Science Foundation of China (NSFC) (Grant No.62072346), and founded by Joint&Laboratory on Credit Technology.

References


Yupeng Cao, Zhiyuan Yao, Zhi Chen, and Zhiyang Deng. 2024. Catmemo@ijcai 2024 finllm challenge: Fine-tuning large language models using data fusion in financial applications. In proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP-AgentScen @ IJCAI 2024. International Joint Conference on Artificial Intelligence.


Sarmistha Das, R E Zera Marveen Lyngkhoi, Sriparna Saha, and Alka Maurya. 2024. Wealth guide: A sophisticated language model solution for financial trading decisions. In proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP-AgentScen @ IJCAI 2024. International Joint Conference on Artificial Intelligence.


Lubingzhi Guo, Javier Sanz-Cruzado, and Richard McCreadie. 2024. University of glasgow at the finllm challenge task: Adapting llama for financial news abstractive summarization. In proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP-AgentScen @ IJCAI 2024. International Joint Conference on Artificial Intelligence.


Ashvini Kumar Jindal, Pawan Kumar Rajpoot, and Ankur Parikh. 2024. Upaya at the finllm challenge task 1 and 2: Distfin: Distillation based fine-tuning for financial tasks. In proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP-AgentScen @ IJCAI 2024. International Joint Conference on Artificial Intelligence.


Ken Kawamura, Zeqian Li, Chit-Kwan Lin, and Bradley McDanel. 2024. Revelata at the finllm challenge task: Improving financial text summarization by restricted prompt engineering and fine-tuning. In proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP-AgentScen @ IJCAI 2024. International Joint Conference on Artificial Intelligence.

Meisin Lee and Soon Lay-Ki. 2024. ‘finance wizard’ at the finllm challenge task: Financial text summarization. In proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP-AgentScen @ IJCAI 2024. International Joint Conference on Artificial Intelligence.


Elvys Linhares Pontes, Carlos-Emiliano González-Gallardo, Mohamed Benjannet, Caryn Qu, and Antoine Doucet. 2024. L3ite at the finllm challenge task: Quantization for financial text classification summarization. In proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP-AgentScen @ IJCAI 2024. International Joint Conference on Artificial Intelligence.


Varad Srivastava. 2024. Bai-arg llm at the finllm challenge task: Earn while you argue - financial argument identification. In proceedings of Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen): FinNLP-AgentScen @ IJCAI 2024. International Joint Conference on Artificial Intelligence.


Qianqian Xie, Wei guang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yu eru He, Meng xiao Xiao, Dong Li, Yongfu Dai, Duan yu Feng, Yijing Xu, Hao qiang Kang, Zi yan Kuang, Chen han Yuan, Kai lai Yang, Zhe heng Luo, Tian lin Zhang, Zhi wei Liu, Guo jun Xiong, Zhi yang Deng, Yue chen Jiang, Zhi yuan Yao, Hao hang Li, Yang yang Yu, Gang Hu, Jia jia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Ben you Wang, Yan zhao Lai, Hao Wang, Min Peng, Sophia An a nadou, and Jimin Huang. 2024a. The finben: An holis tic financial benchmark for large language models. Preprint, arXiv:2402.12659.
