Exploring Large Language Models for Financial Applications: Techniques, Performance, and Challenges with FinMA

P. Djagba^{11,2} and A. Younoussi Saley²⁴

¹Lyman Briggs College, Michigan State University ²Department of Finance, Michigan State University ⁴African Institute for Mathematical Sciences, Rwanda

¹djagbapr@msu.edu

²saley.younoussi@aims.ac.rw

Abstract

This research explores the strengths and weaknesses of domain-adapted Large Language Models (LLMs) in the context of financial natural language processing (NLP). The analysis centers on FinMA, a model created within the PIXIU framework, which is evaluated for its performance in specialized financial tasks. Recognizing the critical demands of accuracy, reliability, and domain adaptation in financial applications, this study examines FinMA's model architecture, its instruction tuning process utilizing the Financial Instruction Tuning (FIT) dataset, and its evaluation under the FLARE benchmark. Findings indicate that FinMA performs well in sentiment analysis and classification, but faces notable challenges in tasks involving numerical reasoning, entity recognition, and summarization. This work aims to advance the understanding of how financial LLMs can be effectively designed and evaluated to assist in finance-related decision-making processes.

Keywords: Large Language Models (LLMs); Financial NLP; FinLLMs; FinMA; FLARE Benchmark; FIT Dataset; Sentiment Analysis; Financial Question Answering; Stock Movement Prediction; Named Entity Recognition; Financial Text Summarization; Instruction Tuning; Financial Reasoning; Domain Adaptation.

Contents

	0.1		3
	0.2	Context	
	0.3	Problem Statement	
	0.4	Objectives	3
	0.5	Methodology	4
	0.6	Thesis Structure	4
1	\mathbf{Lite}	rature Review	5
	1.1	Evolution of Financial Language Models	5
	1.2	Limitations of General-Domain LLMs in Finance	6
	1.3	Current Financial Language Models	7
	1.4	Adaptation Techniques for FinLLMs	8
		1.4.1 Pre-training Strategies	٤
		1.4.2 Fine-tuning and Adaptation Methods	٤
	1.5	Financial NLP Tasks and Datasets	
	1.6	Performance Benchmarks of FinLLMs	1
		1.6.1 Benchmarks for Basic Financial NLP Tasks	
		1.6.2 Benchmarks for complex Financial NLP Tasks	
		1.6.3 FinMA in the Literature	
		1.0.0 I minit in the Electwork	_
2	Met	chodology 1	3
	2.1	FinMA Architecture	
	2.2	Financial Instruction Tuning (FIT) Dataset	
		2.2.1 FIT construction	
		2.2.2 Dataset Format	
		2.2.3 Prompt Engineering and Domain Expertise	
	2.3	FLARE: Financial Evaluation Benchmark	
	2.0	2.3.1 Evaluation Protocol	
	2.4	Fine-Tuning Procedure	
	$\frac{2.4}{2.5}$	Inference and Evaluation Framework	
	2.0	2.5.1 Environment Setup	
		2.5.2 Downloading and Preparing Tools	
		0 1 0	
		1	
		2.5.4 Integration of Evaluation Tools	
		2.5.5 Running Financial NLP and Prediction Tasks	
		2.5.6 Special Handling for ConvFinQA Evaluation	
		2.5.7 Metrics and Evaluation Protocol	
	2.0	2.5.8 Experimental Environment	
	2.6	Reproducibility and Open Science	U
	D	ults and Analysis 2	1
•		· ·	
	3.1	Presentation of Results	
	0.0	3.1.1 Baseline Definition and Re-evaluation	
	3.2	Analysis and Interpretation	_
	3.3	Discussion	
		3.3.1 Practical Implications	
		3.3.2 Challenges 2	Δ

	3.3.3	Potential Future Directions	 	 	 	 	 	 25
4	Conclusion	1						27

0.1. Introduction

0.2. Context

Large Language Models (LLMs) have significantly influenced the advancement of Natural Language Processing (NLP), exhibiting strong performance across diverse linguistic tasks Lee et al. (2024a); Touvron et al. (2023). The evolution of LLMs was notably accelerated since the introduction of the Transformer architecture by Vaswani et al. (2017), which replaced recurrent structures with self-attention mechanisms. This innovation enhanced model scalability, parallel processing capabilities, and training efficiency, enabling the development of powerful models such as ChatGPT (Openai2022) until GPT-4 OpenAI et al. (2024) now.

These breakthroughs have opened new research frontiers across various disciplines, including mathematics, science, healthcare, and finance Chen et al. (2024). Within the financial sector, the inherent complexity, rapid fluctuations, and information asymmetries present substantial analytical challenges. LLMs have emerged as valuable tools due to their capacity to comprehend contextual nuances, process extensive unstructured data, and produce coherent, human like outputs Nie et al. (2024).

Given the domain-specific nature of financial language, which involves highly technical and context-dependent terminology, several LLMs have been developed to address these unique characteristics. Early efforts include FinBERT Araci (2019), a model adapted from BERTDevlin et al. (2019) for sentiment analysis in financial texts. Subsequent developments such as BloombergGPT Wu et al. (2023), FinGPT Lee et al. (2024b), and FinMA Xie et al. (2023) reflect a progression toward more sophisticated architectures tailored for financial applications, incorporating structured knowledge and large-scale web data.

These specialized models support a variety of financial NLP tasks, including sentiment analysis, named entity recognition, text classification, question answering, or stock trend forecasting. Their applications are instrumental in automating information extraction and enabling advanced analytics such as trading strategy formulation and risk evaluation Lee et al. (2024a); Nie et al. (2024).

The progress in this area underscores the necessity of designing LLMs that are finely tuned to financial domains and tasks, particularly where timing, accuracy, and interpretability are paramount.

0.3. Problem Statement

While general-domain LLMs have been extensively studied, Financial LLMs (FinLLMs) remain an emerging area with limited comprehensive surveys Lee et al. (2024a), and several challenges persist. The domain faces significant data accessibility constraints, where high-quality, specialized financial datasets remain largely proprietary or possess insufficient coverage for comprehensive model training Nie et al. (2024). Financial natural language processing introduces heightened complexity through tasks requiring advanced numerical reasoning and causal inference capabilities that extend beyond conventional text generation paradigms Lee et al. (2024a). Current evaluation frameworks present limitations in adequately assessing FinLLM performance, lacking specialized benchmarks that accurately reflect practical financial application requirements rather than general linguistic competency Lee et al. (2024a); Chen et al. (2024). Additionally, the financial sector imposes stringent reliability standards, necessitating exceptional accuracy, interpretability, and dependability given the substantial economic consequences of erroneous predictions or misinterpretations Lee et al. (2024a). These multifaceted challenges underscore the critical need for comprehensive analysis of existing FinLLM architectures, training approaches, evaluation methodologies, and application domains to inform future research directions in this rapidly advancing field.

0.4. Objectives

The primary aim of this study is to examine the strengths and limitations of domain-specific Large Language Models (FinLLMs) in executing financial natural language processing (NLP) tasks, with particular emphasis on the FinMA model Xie et al. (2023).

This research is guided by the following specific objectives:

- To investigate the architectural and training distinctions between general-purpose LLMs and those designed for financial applications.
- To assess the performance of FinMA-7B-full on benchmark financial tasks and datasets, using established models such as BloombergGPT and GPT-4 as comparative baselines.
- To identify the key challenges and opportunities associated with the development and deployment of FinLLMs in real-world financial settings.

To support these objectives, this research addresses the following questions:

- What architectural adaptations and training methodologies enable LLMs to effectively process and interpret financial information?
- How does FinMA compare with general-purpose models in financial reasoning, prediction tasks, and specialized applications?
- What evaluation frameworks most effectively measure LLM performance in financial contexts?

0.5. Methodology

This thesis addresses the identified gaps by conducting an analysis of FinMA, a prominent financial large language model from the PIXIU framework Xie et al. (2023). The model is selected due to its open-source availability and relevance to financial NLP tasks. Datasets and evaluation tools are sourced from the Hugging Face platform, ensuring accessibility and reproducibility.

The research employs a mixed-methods approach that integrates both quantitative and qualitative techniques to address the research questions. It begins with a systematic literature review that traces the development of financial large language models (FinLLMs), examining their architectural evolution, training methodologies, and evaluation benchmarks as documented in existing studies. This is followed by an empirical analysis that assesses the performance of FinMA in comparison to general-purpose models such as GPT-4 across a range of financial natural language processing (NLP) tasks, using established benchmark datasets. Finally, a qualitative discussion is conducted to interpret the evaluation outcomes, highlighting key challenges such as hallucination and data privacy, as well as emerging opportunities, including the potential for multimodal integration. A detailed account of the methodology, encompassing the FinMA architecture, the datasets utilized, and the evaluation framework, is presented in Chapter 2.

0.6. Thesis Structure

The rest of this thesis is organized as follows:

- Chapter 2: Literature Review This chapter reviews past research on LLMs, financial NLP applications, and specialized language models. It builds the literature background needed for this study.
- Chapter 3: Methodology This chapter explains the research design, evaluation methods, and analysis techniques used. It describes FinMA's architecture, training methods, and evaluation framework.
- Chapter 4: Results and Analysis This chapter presents the experimental setup, compares FinMA's performance with baselines like BloombergGPT and GPT-4, and discusses challenges and solutions.
- Chapter 5: Conclusion and Future Directions This chapter summarizes the main findings and suggests ideas for future research.

Chapter 1

Literature Review

The emergence of Financial Large Language Models (FinLLMs) corresponds with the broader evolution of transformer-based language models, which began with the introduction of the Transformer architecture by Vaswani et al. (2017). While general-purpose models like BERT Devlin et al. (2019) and GPT Radford and Narasimhan (2018) laid the foundation for language understanding, their application in financial contexts revealed performance limitations due to a lack of domain-specific adaptation. Financial texts often contain terminology and syntactic structures that differ significantly from general-language corpora, necessitating the development of specialized FinLLMs Lee et al. (2024a).

Early contributions to the field included FinBERT Araci (2019), which fine-tuned BERT on financial texts to improve sentiment analysis. Subsequent models, such as FLANG Shah et al. (2022) and FinGPT Lee et al. (2024b), expanded these ideas by incorporating larger datasets and instruction tuning. The release of BloombergGPT Wu et al. (2023) demonstrated the potential of large-scale financial pretraining, although its proprietary nature limits reproducibility. FinMA Xie et al. (2023), on the other hand, provides an open-source alternative designed to perform across a wide array of financial NLP tasks.

This open-access nature of some FinLLMs contrasts with closed-source models such as BloombergGPT, which restrict access to weights, training data, or model architecture. Open-source models offer transparency, reproducibility, and community collaboration, while closed-source models are typically proprietary and optimized for commercial use.

Recent advancements also emphasize multimodal learning and instruction tuning. Open-FinLLMs Huang et al. (2025) introduced multimodal variants capable of handling textual, numerical, and tabular data simultaneously. These models signify a shift toward generalist financial agents that can operate across varied data modalities.

This literature review examines the evolution of FinLLMs, their limitations compared to general LLMs, and the current state of the field. It explores the architectures, training methods, and performance benchmarks of notable FinLLMs.

1.1. Evolution of Financial Language Models

The development of financial language models has progressed through several distinct phases since the late 2010s. Initially, researchers focused on adapting existing transformer architectures for financial text processing, leading to the creation of specialized models such as FinBERT Araci (2019). This foundational work was subsequently extended through iterations including FinBERT-20 Yang et al. (2020) and FinBERT-21 (Liu2021finbert21), each building upon previous architectures while incorporating domain-specific improvements. The field saw methodological innovation with FLANG Shah et al. (2022), which departed from traditional approaches by implementing ELECTRA-based architectures Clark et al. (2020) combined with financial-specific token masking strategies. This period marked a shift toward more sophisticated pre-training techniques tailored to financial vocabularies and contexts. The landscape evolved significantly by 2023, characterized by the introduction of large-scale, comprehensive financial language models. Notable developments included BloombergGPT Wu et al. (2023), FinGPT Lee et al. (2024b), and FinMA Xie et al. (2023), each representing advances in model scale and capability. More recently, the field has expanded beyond text-only processing, with frameworks like Open-FinLLMs Huang et al. (2025) incorporating multimodal functionalities to handle diverse financial data types.

This evolution is shown in Figure 1.1, which illustrates the timeline from general LLMs to specialized FinLLMs. Figure 1.2 summurize this key milestones:

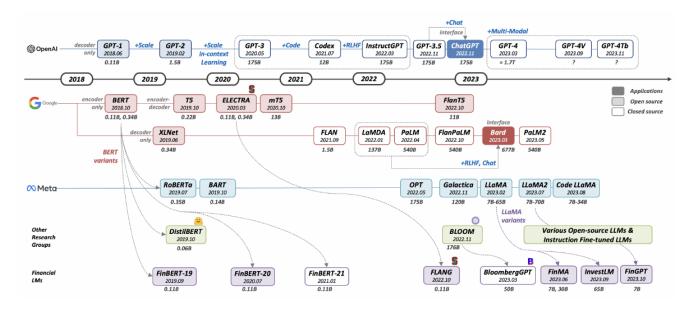


Figure 1.1: Chronological progression of notable pre-trained language models and large language models from general-purpose applications to finance-specific implementations. [Source: Lee et al. (2024a)]

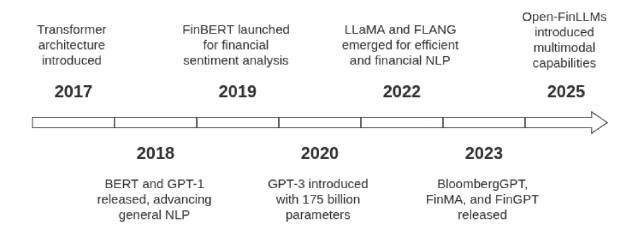


Figure 1.2: Timeline of key developments in financial natural language processing from 2017 to 2025.

1.2. Limitations of General-Domain LLMs in Finance

General-domain LLMs, such as BERT, GPT-3, or GPT-4, are trained on diverse datasets covering a wide range of topics, making them versatile for general language tasks. However, their application in finance is limited by several factors Nie et al. (2024); Chen et al. (2024):

- 1. Lack of Domain-Specific Knowledge: Financial texts contain specialized terminology and concepts that general LLMs may not fully understand. For instance, terms like "yield" or "derivative" have specific financial meanings that differ from their general usage, leading to potential misinterpretations Lee et al. (2024a).
- 2. **Data Efficiency:** General purpose language models require extensive fine-tuning on financial data to achieve competitive performance, which demands significant computational resources. In contrast, finance-specific

- language models (FinLLMs), which are pretrained or fine-tuned on financial corpora, tend to perform well with less training data, making them more data-efficient for financial tasks Nie et al. (2024).
- 3. Cost and Accessibility: Proprietary general LLMs like BloombergGPT GPT-4 are costly and less accessible, while many FinLLMs, such as FinGPT and FinMA, are open-source, making them more practical for researchers and smaller institutions Lee et al. (2024b).

While FinLLMs often outperform general purpose models like GPT-4 in specialized financial tasks such as sentiment analysis, numerical reasoning, and regulatory document analysis due to their targeted training on domain specific corpora Nie et al. (2024); Chen et al. (2024), GPT-4 generally achieves superior performance in broader reasoning tasks thanks to its scale, diverse pretraining data, and advanced zero/few-shot capabilities. However, GPT-4 remains a closed source model with limited transparency and deployment flexibility, which poses challenges for financial institutions concerning data privacy, compliance, and customizability. In contrast, several FinLLMs, particularly open source ones such as FinGPT or FinMA, offer greater control, adaptability, and potential for on premise deployment making them better suited for use cases where interpretability and data confidentiality are paramount Lee et al. (2024a).

1.3. Current Financial Language Models

The landscape of FinLLMs includes several models tailored for financial NLP, each with distinct features:

- **FinBERT:** Proposed by Araci, FinBERT is a domain-adapted variant of BERT, pre-trained on financial corpora. It has gained wide usage for sentiment analysis and text classification in finance-related applications Araci (2019).
- **FLANG:** Presented by Shah et al. (2022), FLANG is a transformer-based language model optimized for financial text. It achieves strong results in sentiment detection and classification tasks specific to the financial sector Shah et al. (2022).
- **BloombergGPT:** With 50 billion parameters, BloombergGPT is trained on a hybrid corpus combining general and domain-specific financial texts. It is built to handle a broad spectrum of financial NLP tasks Wu et al. (2023).
- **FinGPT:** Introduced by Lee et al. (2024b), FinGPT is an open-source initiative focused on enhancing accessibility in financial NLP. It utilizes large-scale financial web data to support real-time applications such as sentiment analysis and algorithmic trading Lee et al. (2024b).
- FinMA: As part of the PIXIU project, FinMA is fine-tuned on FIT datasets (see section 2.2). It integrates structured financial knowledge into language modeling to better capture domain-specific meanings Xie et al. (2023).
- InvestLM: Developed from the LLaMA-65B base model, InvestLM is specifically adapted for investment-focused tasks, showing effectiveness in various applications including content summarization, risk evaluation processes, and financial trend analysis Yang et al. (2023).
- FinTral: FinTral refers to a collection of advanced multimodal large language models constructed on the Mistral-7B architecture and fine-tuned for financial analysis. By incorporating diverse data types including text, numerical indicators, tabular information, and visual content, it is equipped to manage complex financial reasoning tasks. Empirical evaluations show that FinTral consistently outperforms ChatGPT-3.5 across all benchmarks and even exceeds GPT-4 in several cases, representing a significant leap forward in financial AI Bhatia (2024).
- Open-FinLLMs: Open-FinLLMs, released by The Fin AI initiative Huang et al. (2025), constitute a publicly available suite of financial LLMs developed to promote openness and innovation in the field. These models are based on Meta's LLaMA 3 framework and trained on a financial corpus exceeding 52 billion tokens. The suite encompasses multiple variants, each tailored for specialized tasks:
 - FinLLaMA: A foundational model pretrained on diverse financial data, including SEC filings, earnings calls, and market indicators. It demonstrates strong zero-shot performance in real-world financial scenarios Huang et al. (2025).

- **FinLLaMA-Instruct:** An instruction-tuned version of FinLLaMA, trained on 573,000 financial instruction examples to enhance reasoning capabilities, particularly in sentiment analysis, risk assessment, and numerical reasoning Huang et al. (2025).
- FinLLaVA: The first open-source multimodal financial LLM, capable of interpreting charts, tables, and text simultaneously, making it effective for financial decision-making and quantitative analysis Huang et al. (2025).
- Fin-R1: Fin-R1 is a purpose-built large language model optimized for financial reasoning and decision-making tasks. It adopts a two-phase training framework that combines supervised fine-tuning with reinforcement learning. Despite its relatively compact architecture of 7 billion parameters, Fin-R1 achieves competitive, and often superior, performance compared to larger models on financial benchmarks such as FinQA and ConvFinQA Liu (2025).

These models represent the state-of-the-art in the domain of FinLLMs, each contributing unique advancements to enhance NLP applications in the financial sector. Further details on FinMA's architecture and training are provided in Section 2.1 of Chapter 2.

1.4. Adaptation Techniques for FinLLMs

The evolution of Financial Pre-trained Language Models (FinPLMs) and Financial Large Language Models (FinLLMs)¹ reflects ongoing efforts to adapt general-purpose language models to the specific demands of financial natural language processing. This adaptation is achieved through targeted pre-training and fine-tuning methodologies Lee et al. (2024a). A comparative overview of these methods, as applied to four representative FinPLMs and four FinLLMs, is summarized in the work of Lee et al. (2024a).

1.4.1. Pre-training Strategies

Pre-training strategies aim to endow language models with financial domain knowledge by leveraging financial or hybrid textual corpora during the initial training phase. These strategies vary in approach, depending on whether they build upon existing models or initiate training from scratch.

Continual Pre-training. This approach involves extending the training of a pre-existing general-domain language model by further pre-training it on financial texts, thereby incrementally adapting it to domain-specific applications. For example, FinBERT-19 Araci (2019) begins with BERT, originally trained on 3.3 billion general-domain tokens, and is subsequently adapted through additional training on a 29-million-word financial corpus, followed by task-specific fine-tuning for sentiment analysis. This technique enables efficient domain adaptation of general-purpose models without full re-training, making it suitable for focused financial tasks.

Domain-Specific Pre-training from Scratch. In this strategy, models are trained entirely on financial-domain corpora from the ground up, using standard architectures and objectives tailored to the financial context. A representative example is FinBERT-20, which is trained from scratch using a 4.9-billion-token corpus of financial communication and incorporates a dedicated vocabulary known as FinVocab Lee et al. (2024a). This method produces highly specialized models that capture domain-specific linguistic patterns and terminology effectively.

Mixed-Domain Pre-training. This method combines general-domain and financial-domain texts during pre-training to balance broad language understanding with domain relevance. For instance, FinBERT-21 utilizes both general (3.3B tokens) and financial (12B tokens) corpora and applies multi-task learning across six self-supervised objectives. Similarly, FLANG Shah et al. (2022), built on ELECTRA, adopts this approach by training on 12 billion general and 69 billion financial documents. This strategy offers a compromise between generalization and specialization, supporting a wide array of financial NLP applications.

1.4.2. Fine-tuning and Adaptation Methods

Fine-tuning and adaptation techniques are employed to tailor pre-trained language models to specific financial tasks or to increase their flexibility through prompt-based mechanisms.

Mixed-Domain LLM with Prompt Engineering. This method involves training large language models on a blend of general and financial corpora. Once trained, the models are used in a frozen state during inference, with prompts designed to specify tasks often expressed in natural language and optionally accompanied by examples.

¹FinPLMs refer to smaller-scale models (typically around 110 million parameters) that are pre-trained or fine-tuned on financial data to perform domain-specific tasks. In contrast, FinLLMs are large-scale models (typically exceeding 7 billion parameters) that leverage advanced training strategies tailored for financial applications.

BloombergGPT Wu et al. (2023) is an example of this approach, utilizing a BLOOM-based model pre-trained on 343 billion general tokens and 363 billion financial tokens. It was evaluated across 47 tasks, including five financial NLP tasks and 42 general NLP tasks, using prompt engineering. This method facilitates flexible deployment of models without the need for additional weight updates, thereby reducing computational overhead across varied applications.

Instruction Fine-tuned LLM with Prompt Engineering. This approach fine-tunes language models using instruction-based datasets that contain explicit prompts and corresponding responses. Prompt engineering is then used at inference to guide task execution. Several recent FinLLMs exemplify the diverse strategies adopted to enhance model performance in financial contexts. For instance, FinMA (PIXIU) Xie et al. (2023) fine-tunes LLaMA models with 7B and 30B parameters on the Financial Instruction Tuning (FIT) dataset (See 2.2), which contains 136,000 samples and targets tasks such as sentiment classification and stock movement prediction. Similarly, InvestLM Yang et al. (2023) leverages the LLaMA-65B architecture and undergoes specialized training on a curated dataset comprising CFA examination content and Securities and Exchange Commission (SEC) regulatory documents. In another approach, FinGPT Lee et al. (2024b) applies Low-Rank Adaptation (LoRA) techniques Hu et al. (2021) to fine-tune six open-source LLMs Wang et al. (2023) using instruction datasets tailored to financial domains. Collectively, these models demonstrate how targeted fine-tuning strategies can significantly improve model adaptability and task-specific precision, particularly for complex reasoning tasks in financial natural language processing. This method improves model adaptability and task-specific precision, making it especially effective for complex reasoning tasks in financial NLP.

Table 1.1 provides a comparative overview of prominent financial language models. It summarizes key characteristics such as model type, parameter size, applied training techniques, data composition, and open-source accessibility adapted from Lee et al. (2024a).

Model	Type	Size	Technique	Training Data	Open Source
FinBERT-19	Disc	110M	Post-PT, FT	G: 3.3B words, F: 29M words	Yes
FinBERT-20	Disc	110M	PT, FT	F: 4.9B tokens	Yes
FinBERT-21	Disc	110M	PT, FT	G: 3.3B, F: 12B words	No
FLANG	Disc	110M	PT, FT	G: 3.3B, F: 696k documents	Yes
BloombergGPT	Gen	50B	PT, PE	G: 345B, F: 363B tokens	No
FinMA	Gen	7B, 30B	IFT, PE	G: 1T tokens	Yes
InvestLM	Gen	65B	PT, IFT, PE	G: 1.4T tokens	Noa
FinGPT	Gen	7B	IFT, PE, PEFT	G: 2T tokens	Yes

^a Although the code and model weights are publicly available on GitHub, InvestLM adheres to LLaMA's license Touvron et al. (2023), which permits research-only, non-commercial use.

Table 1.1: Summary of financial language models. Abbreviations: Disc = Discriminative, Gen = Generative, PT = Pretraining, FT = Fine-tuning, PST-PT = Post-pretraining, PST-PT = Post-

1.5. Financial NLP Tasks and Datasets

Financial Large Language Models (FinLLMs) are increasingly utilized to tackle a wide range of natural language processing (NLP) challenges within the financial domain. Their applications span from sentiment analysis of market-related text to the development of automated financial advisory systems Xie et al. (2023); Chen et al. (2024); Lee et al. (2024a).

An overview of representative financial NLP tasks is depicted in Figure 1.3, adapted from Chen et al. (2024). Table 1.2 complements this figure by presenting a structured summary of task categories and their associated datasets, as classified by Lee et al. (2024a) and aligned with widely used financial NLP benchmarks.

Task	Description
Sentiment Analysis (SA)	Assessing sentiment of financial texts
Text Classification (TC)	Categorizing texts into predefined labels
Named Entity Recognition (NER)	Identifying financial entities
Question Answering (QA)	Providing accurate answers to financial questions
Stock Movement Prediction (SMP)	Predicting stock price movements
Text Summarization (Summ)	Generating concise summaries of financial reports

Table 1.2: The most common Financial NLP tasks used for evaluating FinLLMs, as mentioned by Lee et al. (2024a)

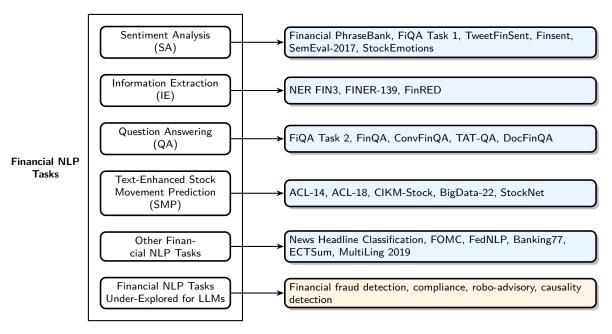


Figure 1.3: Overview of financial NLP tasks and representative datasets for FinLLM evaluation, adapted from Chen et al. (2024). Under-explored tasks are highlighted in yellow.

Evaluation process and scoring

Each model is evaluated using task-specific metrics tailored to individual datasets. In sentiment analysis tasks such as FPB, models classify sentences into positive, negative, or neutral categories, with performance measured through accuracy and F1 scores against ground-truth labels.

On Question-answering tasks like FinQA require precise numerical or textual responses to questions based on financial tables, evaluated using exact match scoring.

For Stock movement prediction tasks, exemplified by BigData22, involve forecasting directional price changes using news sentiment and historical data, assessed through accuracy and Matthews Correlation Coefficient (MCC) to measure prediction alignment with actual market outcomes. Table 1.3 summarizes these evaluation approaches.

Task	Metric	Illustration		
Classification	Accuracy	Ratio of correct predictions over total predictions:		
		(TP + TN) / Total.		
Classification	F1 Score	Harmonic mean of precision and recall. Xie et al.		
		(2023) report both weighted and macro versions.		
Classification	MCC	Correlation coefficient between true and predicted		
		classes. Ranges from -1 (inverse) to 1 (perfect).		
Sequential Labeling	F1 Score	Computed using seqeval, requiring exact match on		
		entity span and type. Xie et al. (2023) use this met-		
		ric.		
Sequential Labeling	Label F1 Score	Label-only correctness, ignores entity spans.		
Relation Extraction	F1 Score	Harmonic mean of precision and recall on predicted		
		relations.		
Summarization	ROUGE-N	Measures N-gram overlap (e.g., ROUGE-1, ROUGE-		
		2) with reference summary.		
Summarization	ROUGE-L	Based on longest common subsequence between gen-		
		erated and reference summaries.		
Question Answering	EMACC	Exact match between prediction and reference an-		
		swer.		

Table 1.3: Predefined task metrics (adapted from Xie et al. (2023))

1.6. Performance Benchmarks of FinLLMs

Financial Large Language Models (FinLLMs) are evaluated through comprehensive benchmarks like PIXIU's FLARE Xie et al. (2023) or with the recent one FinBen Xie et al. (2024), to assess their capabilities in financial NLP and prediction tasks, critical for applications in market analysis and financial advisory. As part of the literature review, this section presents the performance benchmarks from reported in literature for the six primary tasks introduced in Section 1.5: Sentiment Analysis (SA), Text Classification (TC), Named Entity Recognition (NER), Question Answering (QA), Stock Movement Prediction (SMP), and Text Summarization (Summ). The benchmarks, is summarized in Tables 1.4 and 1.5. The scores presented in these tables are drawn from the literature Lee et al. (2024a); Xie et al. (2023); Yang et al. (2023); Wang et al. (2023), as well as from the official repositories² ³.

1.6.1. Benchmarks for Basic Financial NLP Tasks

Lee et al. (2024a) identifies SA, TC, and NER as basic tasks, characterized by their reliance on classification and entity extraction. Table 1.4 presents the benchmarked performance on these tasks, as reported in the literature.

Table 1.4: Reported benchmarks for basic financial NLP tasks (SA = Sentiment Analysis, TC = Text Classification, NER = Named Entity Recognition).

Model	Category	SA	A (F1)	TC (F1)	NER (F1)	
1,10401		FiQA-SA	FPB	Headlines	FIN3	
BERT-base FinBERT-20 FLANG	General Domain Domain		86% 87% 92%	97% 97% 98%	79% 80% 82%	
BloombergGPT	Domain	75%	51%	82%	61%	
FinMA-7B-Full FinMA-30B FinGPT	Domain Domain Domain	79% 87% 87%	87% 88% 88.2%	97% 98% 94.2%	69% 62% 67.3%	
ChatGPT GPT-4	General General	78% 88%	75% 86%	91% 93%	77% 83 %	

²https://github.com/adlnlp/FinLLMs?tab=readme-ov-file

³https://github.com/AI4Finance-Foundation/FinGPT

For **SA**, FLANG achieving the highest Micro-F1 score of 92% on both FiQA-SA and Financial PhraseBank (FPB), followed by FinGPT (88.2%) and FinMA-30B (88%). General LLMs like ChatGPT score 75%, while BERT-base and GPT-4 reach 86%.

In **TC**, FLANG and FinMA-30B attain 98% F1 on the Headlines dataset, with BERT-base, FinBERT-20 and FinMA-7B-Full at 97%, and GPT-4 at 93%. For **NER**, GPT-4 leads with 83% F1 on the FIN3 dataset, followed by FLANG (82%), while FinMA-30B and BloombergGPT score 62% and 61%.

1.6.2. Benchmarks for complex Financial NLP Tasks

Same authors Lee et al. (2024a) classifies QA, SMP, and Summ as complex tasks, requiring numerical reasoning and generative capabilities. Table 1.5 summarizes the reported benchmarks for these tasks.

Model	Category	QA (EM)		SMP (Acc)	Summ (R1)
		FinQA	ConvFinQA	BigData22/ACL18/CIKM18	ECTSum
BERT-base	General	-	-	-	-
FinBERT-20	Domain	_	_	_	_
FLANG	Domain	-	-	_	_
BloombergGPT	Domain	-	43%	_	-
FinMA-7B-Full	Domain	4%	20%	53% / 56% / 53%	8%
FinMA-30B	Domain	11%	40%	47% / 49% / 43%	_
FinGPT	Domain	-	-		_
ChatGPT	General	49%	60%	53% / 50% / 55%	21%
GPT-4	General	76%	76%	54% / 52% / 57%	30%
SOTA (Task-specific)	General			55% / 61% / 59%	47%
Human Expert	Human	91%	89%	_	_
Gene	Human	51%	47%	_	_

Table 1.5: Reported benchmarks for advanced financial NLP tasks (QA = Question Answering, SMP = Stock Movement Prediction, Summ = Text Summarization).

In QA, GPT-4 achieving 76% EM on FinQA and ConvFinQA, with task-specific SOTA models at 89%. FinMA-30B scores 40%, BloombergGPT 43%, and ChatGPT 60%. For SMP, GPT-4 attains 54% accuracy on BigData22, FinMA-7B 52%, and SOTA models 58%, with FinMA-30B at 46%. In Summ, SOTA models lead with 47% ROUGE-1 on ECTSum, followed by GPT-4 (30%) and ChatGPT (21%), while FinMA-7B scores 8%.

1.6.3. FinMA in the Literature

FinMA has also been widely studied in the literature Lee et al. (2024a); Xie et al. (2023). Benchmarked using the FLARE framework, it demonstrates competitive performance in sentiment analysis (SA) and text classification (TC), often achieving results comparable to GPT-3.5 on stock sentiment tasks Xie et al. (2023). Further details regarding its training and evaluation procedures will be detailed in 2. Moreover, FinMA has been explored in advanced tasks such as relation extraction (e.g., the FinRED dataset) and multimodal understanding (e.g., the MAEC dataset), which, while less frequently benchmarked, represent promising directions for future research Lee et al. (2024a); Chen et al. (2024).

Chapter 2

Methodology

This chapter presents the methodology used to develop and evaluate **FinMA**, a financial large language model built within the PIXIU framework Xie et al. (2023). It covers the model architecture, the Financial Instruction Tuning (FIT) dataset, the fine-tuning process, and the Financial Language Understanding and Prediction Evaluation Benchmark (FLARE). The evaluation strategy, analysis techniques, and reproducibility considerations are also described to ensure transparency and replicability.

2.1. FinMA Architecture

FinMA is built upon Meta's LLaMA architecture, a transformer-based, decoder-only large language model optimized for efficiency and scalability Touvron et al. (2023). Two backbone models are used: LLaMA-7B and LLaMA-30B, which differ in parameter count and depth.

- Transformer Layers: 32 layers for LLaMA-7B and 60 layers for LLaMA-30B, each composed of multi-head self-attention and feed-forward networks using SwiGLU activation.
- Embedding Size: 4,096 for LLaMA-7B and 6,656 for LLaMA-30B, supporting rich contextual representation.
- Attention Mechanism: Multi-head self-attention with Root Mean Square Layer Normalization (RM-SNorm), no dropout, and enhanced numerical stability.
- Positional Encoding: Rotary Positional Embeddings (RoPE), enabling context lengths up to 2,048 tokens.

FinMA includes three model variants Xie et al. (2023), all fine-tuned on the Financial Instruction Tuning (FIT)(See 2.2) dataset:

- FinMA-7B: Fine-tuned from LLaMA-7B using instruction tuning on financial NLP tasks, such as in sentiment analysis, headline classification, named entity recognition (NER), and question answering.
- **FinMA-30B:** Fine-tuned from LLaMA-30B using instruction tuning on the same financial NLP tasks, offering enhanced capacity and performance.
- **FinMA-7B-full:** Fine-tuned from LLaMA-7B using full instruction tuning data, incorporating financial natural language understanding (NLP) and prediction applications (e.g., anticipating stock market fluctuations).

All FinMA variants were optimized using the AdamW algorithm, configured with an initial learning rate of $8e^{-6}$, a weight decay of $1e^{-5}$, and a warm-up schedule covering 5% of the total training steps. The maximum sequence length was capped at 2,048 tokens to accommodate long-form financial documents Xie et al. (2023).

The FinMA-7B model was trained over 15 epochs using 8 NVIDIA A100 GPUs with 40GB of memory. A variant version, FinMA-7B-full, underwent a shorter fine-tuning regime of three epochs under identical hardware conditions. In contrast, FinMA-30B required a more extensive training setup completing 20 epochs with a reduced batch size of 24 utilizing distributed training across 128 A100 40GB GPUs Xie et al. (2023). Table 2.1 provides an overview of these configurations across the variants of FinMA.

By leveraging the LLaMA architecture and applying comprehensive instruction tuning, FinMA effectively adapts large language models to the financial domain, demonstrating robust performance across diverse financial natural language processing and prediction tasks.

Table 2.1: Architectural and training configurations of FinMA variants Xie et al. (2023); Touvron et al. (2023).

Feature	FinMA-7B	FinMA-7B-full	FinMA-30B
Base Model	LLaMA-7B	LLaMA-7B	LLaMA-30B
Parameters	\sim 7B	~7B	~30B
Transformer Layers	32	32	60
Attention Heads	32	32	52
Hidden Size	4096	4096	6656
Adaptation	Full fine-tuning	Full fine-tuning	Full fine-tuning
Epochs	15	3	20
Batch Size	32	32	24
GPUs Used	$8 \times A100 40GB$	$8 \times A100 40GB$	$128 \times A100 40GB$
Input Length	2048 tokens	2048 tokens	2048 tokens
Tasks	NLP	NLP + Prediction	NLP

2.2. Financial Instruction Tuning (FIT) Dataset

The Financial Instruction Tuning (FIT) dataset was specifically designed by the creators of FinMA Xie et al. (2023) to address the scarcity of high-quality, instruction-tuning data tailored to the financial domain. It is among the first datasets to target this need, thereby enabling the training of models capable of responding to financial queries using diverse instruction formats.

The dataset leverages publicly accessible financial data and integrates multiple task types and modalities, aiming to support a broad spectrum of real-world applications. FIT contains 136,609 instruction samples that were drawn from nine publicly available sources. These samples span five task categories: sentiment analysis (SA), news headline classification (NC), named entity recognition (NER), question answering (QA), and stock movement prediction (SMP). The data sources include news headlines, tweets, earnings reports, and regulatory filings such as those from the SEC. Moreover, FIT incorporates a multimodal structure, with instances comprising textual content, tables, and time series data nabling the model to develop more nuanced reasoning abilities across different financial contexts.

Table 2.2 summarizes the datasets, task types, number of samples, and associated modalities.

2.2.1. FIT construction

Table 2.2: Datasets in the Financial Instruction Tuning (FIT) corpus.

Dataset	Task	Raw	Instruction	Modalities
Financial Phrase Bank	Sentiment Analysis	4,845	48,450	Text
FiQA-SA	Sentiment Analysis	1,173	11,730	Text
Gold News Headlines	News Headline Classification	11,412	11,412	Text
FIN Agreements	Named Entity Recognition	1,366	13,660	Text
FinQA	Question Answering	8,281	8,281	Text, Tables
ConvFinQA	Question Answering	3,892	3,892	Text, Tables
BigData22	Stock Movement Prediction	7,164	7,164	Text, Time-Series
ACL18	Stock Movement Prediction	27,053	27,053	Text, Time-Series
CIKM18	Stock Movement Prediction	4,967	4,967	Text, Time-Series

2.2.2. Dataset Format

Each FIT instance follows a structured JSON format suitable for instruction tuning:

- "id": Unique identifier
- "conversations": Human-agent instruction turns
- "text": Input data to be analyzed
- "label": Expected classification or response

2.2.3. Prompt Engineering and Domain Expertise

To enhance instruction-following capabilities, FIT reformulates raw data into instruction-based triplets: (instruction, input, response). Each task includes a diverse set of carefully crafted prompts, designed by *financial domain* experts, ensuring realism and generalization Xie et al. (2023).

Raw Data Example (from FPB):

"The company reported a 15% increase in quarterly profits." \rightarrow Positive

This raw example is converted into an instruction-based format for training, as shown in Listing 2.1 below.

```
1 {
    "instruction": "Determine the sentiment expressed in the following financial news
    excerpt:",
    "text": "The company reported a 15% increase in quarterly profits.",
    "response": "Positive"
}
```

Listing 2.1: Sample instruction-tuning entry from FIT (FPB)

This design helps the model learn how to follow instructions, which is important for real-world usage. For example, instead of only seeing direct labels like "positive", the model is trained to understand different ways people may ask for sentiment analysis, such as:

```
"Classify the sentiment of this statement:",
"What is the sentiment expressed in the following sentence?",
"Is the tone positive, negative, or neutral?",
etc.
```

Listing 2.2: Examples of sentiment analysis prompts

As can be referred to in Table 2.3 for the rest of the task.

Table 2.3: Illustrative prompts corresponding to selected financial NLP datasets. In FiQA-SA, {category} is a placeholder for content such as news headlines or tweets. In BigData22, {tid} and {point} represent stock tickers and time points, respectively Xie et al. (2023).

Dataset	Example Prompt
FPB	Determine the sentiment conveyed in the financial news statement: negative, positive, or neutral. For instance, "Stocks plummeted after the scandal" should be labeled as negative.
FiQA-SA	Assess the sentiment of the financial {category} (e.g., headline or tweet) and classify it as Positive, Negative, or Neutral.
Headline	Does the given headline refer to gold price movements? Respond with "Yes" or "No".
NER	Extract named entities from U.S. SEC filings and label them as Person (PER), Organization (ORG), or Location (LOC). Use the format: "entity name, entity type". Example: "Elon Musk, PER; SpaceX, ORG; Cape Canaveral, LOC".
FinQA	Provide a concise answer to the financial question using data and reasoning from available sources.
ConvFinQA	Use the given pretext, table, and document to answer the final financial question, applying necessary calculations and contextual understanding.
BigData22	Based on market signals and social media activity, predict whether the stock {tid} will increase or decrease at time {point}. Answer with: Rise or Fall.

2.3. FLARE: Financial Evaluation Benchmark

The **FLARE** benchmark (Financial Language Understanding and Prediction Evaluation) is also a benchmark specifically designed by the authors of FinMA Xie et al. (2023) to evaluate the performance of large language models (LLMs) on both financial natural language understanding and financial prediction tasks. Unlike earlier benchmarks such as FLUE (shah2022flue), which focus exclusively on NLP tasks, FLARE integrates financial prediction tasks such as stock movement forecasting, providing a more comprehensive assessment of LLM capabilities in real-world financial applications.

FLARE is constructed from the FIT dataset (see Section 2.2). The authors Xie et al. (2023) follow a principled approach by randomly selecting validation and test subsets from FIT, in line with established evaluation protocols Guo et al. (2023). This sampling strategy not only ensures statistical soundness but also facilitates meaningful comparison with proprietary models. In particular, the number and distribution of test samples in FLARE are aligned with those used in the evaluation of BloombergGPT Wu et al. (2023), whose original test data is not publicly released. This alignment allows researchers to benchmark open-source models under comparable settings.

The benchmark covers five task categories as FIT[2.2]: sentiment analysis, news headline classification, named entity recognition (NER), question answering (QA), and stock movement prediction (SMP). Each task is associated with one or more datasets and evaluated using appropriate metrics[1.3]. Sentiment analysis tasks (FPB, FiQA-SA) are evaluated using accuracy and F1 score; QA tasks (FinQA, ConvFinQA) use exact match (EM) accuracy; stock prediction tasks (BigData22, ACL18, CIKM18) are evaluated using both accuracy and Matthews correlation coefficient (MCC). Table 2.4 summarizes the tasks, datasets, and evaluation metrics.

Table 2.4: FLARE benchmark tasks, datasets, and evaluation metrics.

Task	Dataset(s)	${\bf Evaluation\ Metric(s)}$
Sentiment Analysis	FPB, FiQA-SA	F1 score, Accuracy
News Classification	Gold News Headlines	Average F1 score
Named Entity Recognition	FIN Agreements	Entity-level F1 score
Question Answering	FinQA, ConvFinQA	Exact Match (EM)
Stock Movement Prediction	BigData22, ACL18, CIKM18	Accuracy, MCC

2.3.1. Evaluation Protocol

The evaluation follows established practices in financial NLP, Guo et al. (2023); Wu et al. (2023):

Zero-shot vs Few-shot Settings:

- **Zero-shot:** Model receives only the instruction and input, with no examples shown before prediction. All results for FinMA are reported in the zero-shot setting Xie et al. (2023).
- Few-shot: Model is given a few examples of the task before making its prediction.
 - BloombergGPT: **20-shot** on FIN dataset, **5-shot** on FPB and FiQA-SA
 - Other baselines: **5-shot** on News dataset

Some baseline models that were not instruction-tuned had trouble generating answers in the expected format (e.g., they produce long text when only a label like "positive" is expected). In such cases, results were evaluated by humans instead of automatically Xie et al. (2023). This ensures a fair comparison between FinMA and BloombergGPT on the FLARE benchmark following Xie et al. (2023).

2.4. Fine-Tuning Procedure

FinMA models are fine-tuned using instruction-based learning on the FIT dataset (see Section 2.2). The goal is to align the model with financial domain tasks by training it to follow natural language instructions across multiple modalities and task types.

As seen in 2.2, each training sample follows a standard format:

(instruction, input, response)

This structure teaches the model to produce context-aware outputs in response to instructions, preparing it for zero-shot generalization. And this format enables FinMA to align language understanding with financial task objectives in a natural and structured way.

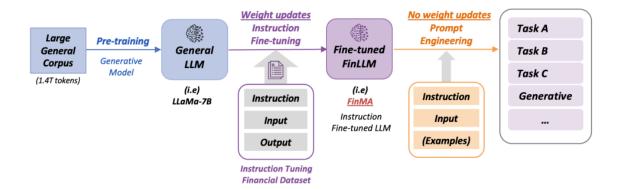


Figure 2.1: Instruction tuning pipeline for financial datasets Lee et al. (2024a).

Training Infrastructure

FinMA uses a causal language modeling objective, where the model predicts tokens autoregressively. Training is performed using the AdamW optimizer, with **Learning rate:** 8×10^{-6} and for **Weight decay:** 1×10^{-5} as mentioned in 2.1 and training was conducted using 8 NVIDIA A100 GPUs (40GB each), with implementation based on PyTorch and Hugging Face Transformers.

Model Variant	Hardware	Training Duration	Environment
FinMA-7B FinMA-7B-full	8 × A100 40GB 8 × A100 40GB	15 epochs 3 epochs	Google Colab Pro Google Colab Pro
FinMA-30B	$128 \times A100 \ 40 \text{GB}$	20 epochs	Distributed Tra

Table 2.5: Training infrastructure specifications for FinMA variants.

All variants are fully fine-tuned (no adapters or LoRAHu et al. (2021) techniques are used) ensuring deep adaptation to financial reasoning.

Prompt diversity and instruction design

As described in Section 2.2, domain experts created multiple instruction templates per task to improve generalization. For smaller datasets, up to 10 instructions were applied to each raw data point; for larger datasets, one instruction was randomly sampled to maintain scalability.

This diversity ensures that the model learns to handle different phrasings of the same task. For example, in sentiment analysis, prompts may vary from "What is the sentiment?" to "Classify the tone of this statement." This strategy contributes directly to FinMA's strong zero-shot performance (see Section 2.3).

2.5. Inference and Evaluation Framework

In this section, we describe the evaluation process for inference and our setup for inferring the model across tasks to reproduce the results and highlight our contributions. This also serves as a basis for the discussion of results presented in Chapter 3.

2.5.1. Environment Setup

To ensure reproducibility and avoid package conflicts, a dedicated Python virtual environment was created and activated. All dependencies were installed within this environment.

```
| !pip install virtualenv | !virtualenv | finma_venv
```

2.5.2. Downloading and Preparing Tools

Two main repositories were cloned:

- BARTScore¹: for advanced text generation evaluation.
- PIXIU: contains the financial language models and evaluation scripts.

```
| git clone https://github.com/neulab/BARTScore.git --recursive | git clone https://github.com/The-FinAI/PIXIU.git --recursive
```

2.5.3. Model and Metric Preparation

The FinMA-7B-full model² was downloaded from Hugging Face, and BARTScore was set up with its pretrained checkpoint.

¹https://github.com/neulab/BARTScore

 $^{^2} https://hugging face.co/Chance Focus/finma-7 b-full \\$

2.5.4. Integration of Evaluation Tools

To ensure BARTScore works with the evaluation scripts, the eval.py file was edited to include the correct path for the BARTScore module if not already present.

```
file_path = "/kaggle/working/PIXIU/src/eval.py"
path_to_add = "/kaggle/working/PIXIU/src/metrics/BARTScore"

with open(file_path, "r") as f:
    content = f.read()
if f"sys.path.append('{path_to_add}')" not in content:
    new_content = f"import sys\nsys.path.append('{path_to_add}')\n" + content
    with open(file_path, "w") as f:
    f.write(new_content)
```

2.5.5. Running Financial NLP and Prediction Tasks

The FinMA-7B-full model was evaluated across nine datasets covering six financial NLP tasks using the PIXIU benchmark. Inference was performed using the following command:

```
!python eval.py \
--model "hf-causal-llama" \
--model_args "use_accelerate=True,pretrained=TheFinAI/finma-7b-full,tokenizer=TheFinAI
/finma-7b-full,use_fast=False" \
--tasks "flare_fiqasa,flare_fpb,flare_ner,flare_headline,flare_finqa,flare_convfinqa,
flare_bigdata22,flare_skm,flare_cikm,flare_ectsumm"
```

2.5.6. Special Handling for ConvFinQA Evaluation

For the **ConvFinQA** dataset, we implemented a dedicated inference pipeline. The default evaluation script (eval.py) provided in the PIXIU repository could not be executed as intended due to technical issues previously reported by other users³. As such, a custom evaluation procedure was implemented for assessing the zero-shot performance of the FinMA-7B-full model on this benchmark. The detailed process is outlined below.

- a. Model Initialization. The FinMA-7B-full model was loaded from Hugging Face using the Transformers library. To support efficient inference, the model was cast to float16 precision and deployed with automatic device mapping. A generation pipeline was configured, and pad_token_id was explicitly defined to prevent runtime errors during generation.
- b. Dataset Loading. We loaded the flare-convfinqa test split using the datasets library. Each instance included a natural language financial question (query), a numerical answer (answer), the dialogue turn number (turn), and a unique identifier for the multi-turn dialogue (dialogue_id).
 - c. Prompt Design. Each question was embedded in a structured prompt as follows:

```
Question: {query}
Context: {context}
Answer with only a number:
Answer:
```

In most cases, the context field was left empty, as the questions are self-contained.

- **d. Answer Generation**. The prompts were tokenized and passed to the FinMA-7B-full model. Responses were generated with a maximum of 32 new tokens. The output was then parsed to extract only the predicted value following the final **Answer**: marker.
- **e.** Evaluation Metrics. To quantitatively assess performance, we computed two metrics widely used in question answering: Exact Match (EM) and F1 Score as summurize in Table 1.3. All answers were normalized by converting to lowercase, removing punctuation, and collapsing multiple whitespaces.

2.5.7. Metrics and Evaluation Protocol

The evaluation relied on predefined metrics per task, as summarized in 1.3 (Section 2.2). These include F1 score, Accuracy, Exact Match (EM) and ROUGE1. Readers are referred to that section for definitions of those metrics.

³https://github.com/chancefocus/PIXIU/issues

2.5.8. Experimental Environment

Table 2.6: Kaggle-based experimental environment specifications.

Component	Specification	Purpose		
Hardware	T4 GPUs (16GB), up to 4 CPU cores	Model inference in Kaggle kernels		
Software	PyTorch, HuggingFace Transformers, Python 3.11	Deep learning and model management		
Platform	Kaggle Notebooks (free tier)	Computational environment with GPU/TPU support		
Storage	20GB temporary disk, 5GB persistent output	Dataset and model storage		

2.6. Reproducibility and Open Science

The PIXIU repository⁴ provides comprehensive reproducibility support Xie et al. (2023). Additional resources, including custom scripts and documentation, are available in our GitHub repository⁵.

⁴https://github.com/chancefocus/PIXIU

⁵https://github.com/AbdelkaderYS

Chapter 3

Results and Analysis

This chapter evaluates the performance of FinMA-7B-full on the FLARE benchmark using the PIXIU framework Xie et al. (2023), as outlined in Chapter 2 (Section 2.5). The evaluation covers sentiment analysis (SA), news headline classification (TC), named entity recognition (NER), question answering (QA), stock movement prediction (SMP), and text summarization (Summ) in zero-shot, 5-shot, and 20-shot settings, following the evaluation methodology of BloombergGPT Wu et al. (2023). Results are compared against baselines from Xie et al. (2023) and other models, including BloombergGPT, GPT-4, and ChatGPT, as reviewed in Chapter 1 (Tables 1.4, 1.5).

3.1. Presentation of Results

3.1.1. Baseline Definition and Re-evaluation

In this study, the baseline refers to the zero-shot performance of FinMA-7B-full as reported in Xie et al. (2023) within the FLARE benchmark. This baseline measures the model's inherent capabilities without task-specific examples, a standard approach for assessing the generalizability of large language models (LLMs) in financial NLP Brown et al. (2020); Lee et al. (2024a). The re-evaluation of FinMA-7B-full in zero-shot settings verifies the reproducibility of these results, ensures alignment with the PIXIU framework, and extends the analysis to 5-shot and 20-shot settings for comprehensive benchmarking Guo et al. (2023). This methodology confirms the model's robustness, aligns with industry standards such as BloombergGPT Wu et al. (2023), and provides new insights into FinMA's performance across diverse financial tasks.

Performance results are summarized in Tables 3.1 (SA, TC, NER) and 3.2 (QA, SMP, Summ), presenting zero-shot, 5-shot, and 20-shot performance metrics compared to the zero-shot baseline from Xie et al. (2023).

Table 3.1: Financial NLP Benchmarking Results: Sentiment Analysis, Classification, and NER

Task & Dataset		Best vs Baseline			
	$\begin{array}{c} \hline \text{Baseline} \\ \text{(0-Shot)} \end{array}$	Zero-Shot	5-Shot	20-Shot	_ Dest vs Dasenne
SENTIMENT A	NALYSIS				
FiQA-SA					
o F1-Score	79.0	83.5	82.6	81.2	+4.5
FPB					
\circ F1-Score	87.0	93.9	93.4	93.4	+6.9
NEWS HEADLI	NE CLASSIF	ICATION			
Headlines					
• Avg F1-Score	97.0	97.5	93.5	_	+0.5
NAMED ENTIT	Y RECOGNI	TION			
Financial NER					

Continued on next page

Table 3.1 (continued)

Task & Dataset		Best vs Baseline			
	Baseline (0-Shot)	0-Shot	5-Shot	20-Shot	_ Best vs Baseline
o Entity F1-Score	69.0	53.9	64.1	65.8	-4.9

Financial NLP Benchmarking Results on SA, NER, and TC Tasks Across Their Datasets

Performance Comparison

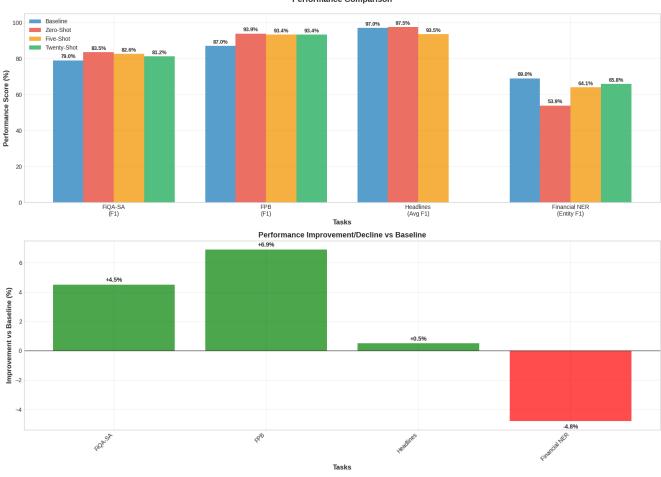


Figure 3.1: Zero-shot and Few-shot F1 Scores for Sentiment Analysis (FiQA-SA, FPB), News Headline Classification (Headlines), and Named Entity Recognition (Financial NER).

Table 3.2: Financial NLP Benchmarking Results: Question Answering, Stock Prediction, and Summarization

Task & Dataset	Performance Scores (%)				Best vs Baseline
	Baseline (0-Shot)	Zero-Shot	5-Shot	20-Shot	Dest vs Dasenne
QUESTION ANSV	VERING				
FinQA					
• Exact Match Acc.	4.0	7.4	7.0	_	+3.4
ConvFinQA					
• Exact Match Acc.	20.0	36.4	_	_	+16.4

Continued on next page

Table 3.2 (continued)

Task & Dataset		Best vs Baseline			
	$\begin{array}{c} \hline \text{Baseline} \\ \text{(0-Shot)} \end{array}$	0-Shot	5-Shot	20-Shot	- Best vs Baseline
STOCK MOVEM	ENT PRED	ICTION			
BigData22					
o Accuracy	53.0	50.5	51.5	_	-1.5
ACL18					
• Accuracy	56.0	51.2	51.1	_	-4.8
CIKM18					
• Accuracy	53.0	49.3	48.7	_	-4.3
TEXT SUMMAR	IZATION				
EctSumm Dataset					
\circ ROUGE Score	8.0	2.8	_	_	-5.2

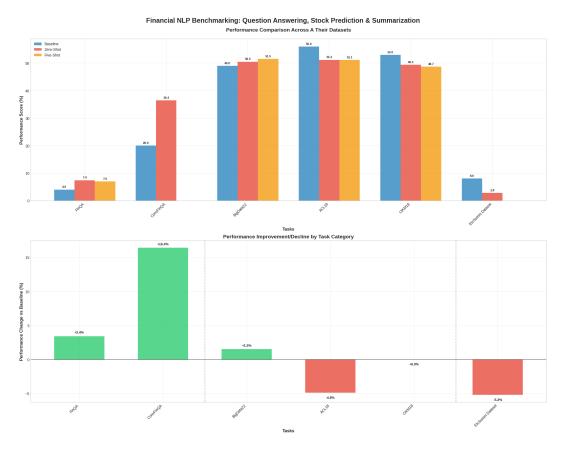


Figure 3.2: Zero-shot and Few-shot Performance for Question Answering (FinQA, ConvFinQA), Stock Movement Prediction (BigData22, ACL18, CIKM18), and Text Summarization (EctSumm).

3.2. Analysis and Interpretation

Sentiment Analysis, Headline Classification, and Named Entity Recognition Tasks

The FinMA-7B-full model demonstrates strong zero-shot performance on sentiment analysis tasks, achieving an F1 score of 93.9% on FPB and 83.5% on FiQA-SA, both exceeding the reported baselines of 87.0% and 79.0%, respectively (see Table 3.1). These scores also surpass those of FinMA-30B (88% on FPB) and GPT-4 (93% on FPB, reported in 5-shot settings) Xie et al. (2023). For news headline classification, FinMA-7B-full achieves an average F1 score of 97.5% on the Headlines dataset, marginally outperforming its baseline (97.0%) and significantly

surpassing GPT-4 (93%). However, a slight decline in 5-shot performance (e.g., 82.6% on FiQA-SA and 93.5% on Headlines) is observed, which may stem from the model's sensitivity to example quality, an issue previously identified in LLaMA-based architectures Touvron et al. (2023). These findings are consistent with the instruction tuning methodology adopted from the FIT dataset(2.2), and support downstream applications such as sentiment-driven trading or automated news monitoring.

By contrast, results in named entity recognition (NER) are less impressive. In a 5-shot setting, FinMA-7B-full reaches an F1 score of 64.1%, falling below its own zero-shot baseline of 69.0%. While it still outperforms BloombergGPT (61% in 20-shot), it remains far behind GPT-4 (83%). This underperformance likely stems from LLaMA's general-purpose architecture, which lacks optimized mechanisms for span detection, as well as the relatively limited representation of financial NER tasks in the FIT training set.

Question Answering, Stock Movement Prediction, and Text Summarization Tasks

In question answering (QA), FinMA-7B-full obtains zero-shot Exact Match scores of 7.4% on FinQA and 36.4% on ConvFinQA, improving on the baselines of 4.0% and 20.0%, respectively. Although these gains are significant, they remain below GPT-4's 69–76% and FinMA-30B's range of 11–40%. The relatively better performance on ConvFinQA can be linked to conversational formatting in the FIT dataset, which helps the model understand multi-turn questions. However, LLaMA's known limitations in numerical reasoning, demonstrated by its low accuracy on mathematical benchmarks Touvron et al. (2023), likely reduce its effectiveness on finance-specific QA tasks. The absence of 5-shot results on ConvFinQA further restricts the scope of comparative analysis.

For stock movement prediction (SMP), the model shows mixed results. It achieves 50.5% accuracy on BigData22, 51.2% on ACL18, and 49.3% on CIKM18, all below the respective baselines by 1.5 to 4.8 percentage points. Even in 5-shot settings, improvements are marginal. These results illustrate the difficulty the model faces in handling multimodal inputs, such as tweets combined with time series data, which are central to financial forecasting tasks. This limitation affects the model's applicability in real-world investment strategies.

Finally, the model's performance in text summarization is considerably lower. On the ECTSum dataset, FinMA-7B-full achieves a zero-shot ROUGE-1 score of just 2.8%, significantly below the 8.0% baseline and far from GPT-4 (30%) or specialized summarization models (around 47%). This gap likely results from the limited presence of summarization instructions in the FIT dataset, which hinders the model's ability to generate structured and informative financial reports. Moreover, the lack of 5-shot or 20-shot evaluations prevents further insights into whether few-shot prompting could mitigate this shortfall.

3.3. Discussion

This section examines FinMA-7B-full's performance on the FLARE benchmark, highlighting practical implications, challenges, and future directions, drawing on the PIXIU framework Xie et al. (2023), based on LLaMA's architecture Touvron et al. (2023).

3.3.1. Practical Implications

FinMA-7B-full excels in zero-shot SA (93.9% F1 on FPB, 83.5% on FiQA-SA) and TC (97.5% average F1 on Headlines), surpassing FinMA-30B (88% F1), GPT-4 (93% F1), and BloombergGPT (51–82% F1). Leveraging LLaMA's efficient transformer architecture with RMSNorm, SwiGLU activation, and RoPE embeddings Touvron et al. (2023), combined with FIT dataset tuning, enables cost-effective applications in sentiment-driven trading and automated news analysis. The model's QA performance (36.4% EM Acc on ConvFinQA) supports financial chatbots and analyst tools, though limitations in numerical reasoning restrict complex reporting tasks Lee et al. (2024a). Weaker performance in NER (64.1% F1), SMP (50.5–51.2% Acc), and summarization (2.8% ROUGE) limits applications requiring precise entity detection, multi-modal data integration, or concise report generation.

Ethical considerations include the risk of hallucinations, with LLaMA's 57% TruthfulQA score indicating potential misinformation in financial predictions. Biases in LLaMA's 67% CommonCrawl training data necessitate expert validation or Retrieval-Augmented Generation (RAG) to ensure reliability. Privacy concerns, particularly with sensitive financial data, require anonymization and adherence to regulatory standards Touvron et al. (2023); Lee et al. (2024a).

3.3.2. Challenges

FinMA-7B-full faces several challenges that limit its performance on the FLARE benchmark, reflecting common issues in financial LLMs. These challenges are supported by evidence from the FIT dataset's composition (Section 2.2, Table 2.2) and model characteristics.

- 1. Limited Financial Training: FinMA-7B-full relies on LLaMA's general-purpose 7B-parameter model, which struggles with financial-specific tasks. For named entity recognition (NER), the 5-shot Entity F1 score of 64.1% (-4.9% below baseline) is lower than GPT-4 (83%). This is due to the small size of the FIN Agreements dataset in FIT, with only 13,660 instruction samples (Table 2.2), limiting the model's exposure to financial entities like organizations or locations.
- 2. Poor Numerical Skills: FinMA's 7.4% Exact Match accuracy on FinQA (Table 3.2) reflects LLaMA's weak mathematical reasoning (10.6% on MATH) Touvron et al. (2023). The FIT dataset's QA components, FinQA (8,281 samples) and ConvFinQA (3,892 samples), focus more on textual reasoning than numerical calculations (Table 2.2), reducing FinMA's ability to handle quantitative financial tasks Lee et al. (2024a).
- 3. Multi-Modal Data Issues: Stock movement prediction (SMP) accuracies range from 50.5% to 51.2% (-1.5% to -4.8% below baseline) (Table 3.2), below state-of-the-art models (58%). The FIT dataset's SMP datasets (BigData22: 7,164 samples; ACL18: 27,053; CIKM18: 4,967) include text and time-series data (Table 2.2), but LLaMA's text-only architecture struggles to integrate these modalities, a key challenge in financial prediction Lee et al. (2024a).
- 4. Weak Summarization Training: The 2.8% ROUGE score on EctSumm is far below top models (47%). The FIT dataset lacks dedicated summarization tasks (Table 2.2), with no datasets explicitly designed for financial report summarization, limiting FinMA's ability to generate concise outputs Xie et al. (2023).
- 5. Few-Shot Instability: 5-shot performance drops, such as TC (93.5%) and FiQA-SA (82.6%) (Table 3.1), show LLaMA's sensitivity to example quality, consistent with variance in social intelligence tasks (SIQA) Touvron et al. (2023). The FIT dataset's prompt diversity (Section 2.2) helps zero-shot performance but does not fully stabilize few-shot learning.
- 6. Limited Computing Power: Restricted resources prevented full FinMA-30B fine-tuning (Section 2.4), as the FIT dataset's 136,609 samples required significant computational power (Table 2.2), impacting scalability Xie et al. (2023).
- 7. Bias and Ethical Risks: Biases in LLaMA's CommonCrawl training data (67% of training corpus) may skew financial outputs Touvron et al. (2023). The FIT dataset's reliance on public sources like tweets and news (Table 2.2) introduces potential biases, and LLaMA's 57% TruthfulQA score highlights risks of incorrect predictions, requiring ethical safeguards Lee et al. (2024a).
- 8. **Environment Constraints**: Also for this work, re-evaluation of FinMA-7B-full used Kaggle's environment with T4 GPUs (16GB, up to 4 CPU cores) and 20GB temporary disk (Table 2.6), which falls short of the 8× A100 40GB GPUs used for training (Section 2.4). This limited memory and processing power slowed inference and restricted experiments, such as larger batch sizes or FinMA-30B evaluation, impacting scalability.

Attempts to improve accuracy using prompt engineering techniques, such as Chain-of-Thought or Meta Prompting, did not yield significant gains, particularly for QA (7.4% FinQA) and SMP (50.5–51.2%, Table 3.2). The complexity of FIT's financial tasks, like numerical reasoning in FinQA or multi-modal analysis in BigData22 (Table 2.2), requires domain-specific prompts, which were limited by a lack of financial expertise in prompt design Lee et al. (2024a).

3.3.3. Potential Future Directions

To address the challenges outlined in Section 3.3.2 above and enhance FinMA-7B-full's performance within the PIXIU framework Xie et al. (2023), several research directions are proposed, drawing on experimental results (Tables 3.1, 3.2) and the FIT dataset (Table 2.2).

Fine-tuning on specialized financial datasets like FINER-139 for named entity recognition (NER) and EarningsCall for summarization can improve performance, as the FIT dataset's limited samples (e.g., 13,660 for FIN Agreements) and lack of summarization tasks hinder current results (Table 2.2).

Incorporating Retrieval-Augmented Generation (RAG) can further enhance NER accuracy. Training on numerical datasets like GSM8K can strengthen question answering (QA) capabilities, addressing FinMA's weak 7.4% Exact Match accuracy on FinQA (Table 3.2) Touvron et al. (2023); Lee et al. (2024a).

For stock movement prediction (SMP), developing hybrid models to integrate text and time-series data from FIT's datasets (e.g., BigData22) can boost accuracies (50.5–51.2%, Table 3.2) Lee et al. (2024a).

Active learning to select high-quality examples can stabilize few-shot performance in tasks like news headline classification (TC: 93.5%, Table 3.1) Touvron et al. (2023). The lack of accuracy gains from prompt engineering

techniques, such as Chain-of-Thought, suggests a need for financial domain expertise to design tailored prompts for FIT's complex tasks (Table 2.2) Lee et al. (2024a); Chen et al. (2024). Transitioning from Kaggle's T4 GPUs (Table 2.6) to cloud platforms like AWS can enable faster inference and FinMA-30B experiments.

Chapter 4

Conclusion

This work evaluated the performance of **FinMA-7B-full** model on the FLARE benchmark from the PIXIU framework Xie et al. (2023), as detailed in Chapter 3. The results show significant progress in sentiment analysis (SA)(93.9% F1 on FPB, +6.9% compared to the baseline) and news headline classification (TC) (97.5% Avg F1 on Headlines, +0.5% over the baseline), outperforming a little bit FinMA-30B and GPT-4 in zero-shot settings. These achievements, supported by the efficient LLaMA architecture Touvron et al. (2023) and fine-tuning on the FIT dataset, confirm the potential of FinMA-7B-full for financial applications such as sentiment-based trading and news analysis Lee et al. (2024a); Nie et al. (2024).

However, limitations remain in named entity recognition (NER) task (64.1% F1, -4.8% compared to the baseline), stock movement prediction (SMP) (50.5–53.0% accuracy), and summarization generation (Summ) (2.8% ROUGE). These issues are mainly due to the general-purpose nature of the LLaMA backbone and limited financial training Touvron et al. (2023); Xie et al. (2023). Biases in the LLaMA CommonCrawl data and hallucination risks (57% TruthfulQA) highlight the need for expert validation and ethical measures such as differential privacy Nie et al. (2024); Lee et al. (2024a).

The main contributions of this work include a rigorous evaluation of FinMA-7B-full, highlighting its strengths in financial language understanding and its weaknesses in numerical reasoning and multimodal tasks. These findings enrich the literature on FinLLMs, building on works like PIXIU Xie et al. (2023) and BloombergGPT, and offer perspectives for practical applications in finance.

Future Work: Future research will focus on the inference and evaluation of advanced conversational AI models applied to financial tasks. Although GPT-4.5 was initially considered due to its enhanced reasoning capabilities and reduced hallucination rates, its high inference cost and planned deprecation¹ present notable limitations for long-term use. Consequently, future research will consider more practical and cost-effective alternatives such as GPT4.1 and GPT4-turbo (o4), which represent the most capable models publicly accessible at the time of writing, along with open-source language models fine-tuned for the financial domain, including FinMA-7B-full. Model performance will be evaluated using selected tasks from the FLARE benchmark Xie et al. (2023), with a particular focus on sentiment analysis, question answering, and multimodal prediction.

For fine-tuning FinMA-7B-full, the first priority after securing access to suitable GPU resources will be to explore Low-Rank Adaptation (LoRA) techniques Hu et al. (2021). LoRA enables efficient and cost-effective fine-tuning by updating only a small subset of model parameters, significantly reducing GPU memory requirements and training time compared to full fine tuning. This approach is particularly suitable for large models like FinMA-7B and facilitates experimentation even on limited hardware. We will also investigate the potential merging or knowledge transfer between FinMA and FinGPT models to leverage complementary strengths. Additionally, the implementation of retrieval-augmented generation (RAG) methods will be considered to reduce hallucinations and enhance factual consistency in financial AI applications.

 $^{^{1}} https://medium.com/artificial-synapse-media/openai-deprecates-gpt-4-5-api-in-july-2025-forcing-developers-to-migrate-to-gpt-4-1-amid-backlash-417a4a31eb0d$

Bibliography

- Dogu Tan Araci. Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063, 2019. URL https://arxiv.org/abs/1908.10063.
- Gagan et al. Bhatia. Fintral: A family of gpt-4 level multimodal financial large language models. arXiv preprint arXiv:2402.10986, 2024.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.
- Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. A survey on large language models for critical societal domains: Finance, healthcare, and law. arXiv preprint arXiv:2405.01769, 2024. URL https://arxiv.org/abs/2405.01769.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020. URL https://arxiv.org/abs/2003.10555.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.
- Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. Evaluating large language models: A comprehensive survey. arXiv preprint arXiv:2306.05685, 2023.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.
- Jimin Huang, Mengxi Xiao, Dong Li, Zihao Jiang, Yuzhe Yang, Yifei Zhang, Lingfei Qian, Yan Wang, Xueqing Peng, Yang Ren, Ruoyu Xiang, Zhengyu Chen, Xiao Zhang, Yueru He, Weiguang Han, Shunian Chen, Lihang Shen, Daniel Kim, Yangyang Yu, Yupeng Cao, Zhiyang Deng, Haohang Li, Duanyu Feng, Yongfu Dai, VijayaSai Somasundaram, Peng Lu, Guojun Xiong, Zhiwei Liu, Zheheng Luo, Zhiyuan Yao, Ruey-Ling Weng, Meikang Qiu, Kaleb E Smith, Honghai Yu, Yanzhao Lai, Min Peng, Jian-Yun Nie, Jordan W Suchow, Xiao-Yang Liu, Benyou Wang, Alejandro Lopez-Lira, Qianqian Xie, Sophia Ananiadou, and Junichi Tsujii. Open-finllms: Open multimodal large language models for financial applications. arXiv preprint, 2025. URL https://arxiv.org/abs/2408.11878.
- Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. A survey of large language models in finance (finllms). arXiv preprint arXiv:2402.02315, 2024a. URL https://arxiv.org/abs/2402.02315.
- Jungwoo Lee, Renqian Wang, Bill Yuchen Lin, Xinyu Liu, Zhijing Gao, Di Jin, and Xiang Ren. Fingpt: Democratizing financial research with open-source large language models. arXiv preprint arXiv:2306.06031, 2024b. URL https://arxiv.org/abs/2306.06031.
- Zhaowei et al. Liu. Fin-r1: A large language model for financial reasoning through reinforcement learning. arXiv preprint arXiv:2503.16252, 2025.

- Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. A survey of large language models for financial applications: Progress, prospects and challenges. arXiv preprint arXiv:2406.11903, 2024. URL https://arxiv.org/abs/2406.11903.
- OpenAI, Josh Achiam, and et al. Steven Adler. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.
- Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.
- Nitish Shah, Zixuan Qin, Haonan Zhang, Zhen Yang, Peng Liu, Dong Yu, and Chin-Yew Lin. Flang: A pretrained language model for financial tasks. arXiv preprint arXiv:2210.05075, 2022. URL https://arxiv.org/abs/2006.08097.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023. URL https://arxiv.org/abs/2302.13971.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.
- Neng Wang, Hongyang Yang, and Christina Dan Wang. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets, 2023. URL https://arxiv.org/abs/2310.04793.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564, 2023. URL https://arxiv.org/abs/2303.17564.
- Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. arXiv preprint, 2023. URL https://arxiv.org/abs/2306.05443.
- Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. Finben: A holistic financial benchmark for large language models, 2024. URL https://arxiv.org/abs/2402.12659.
- Yi Yang, Mark Christopher Siy UY, and Allen Huang. Finbert: A pretrained language model for financial communications, 2020. URL https://arxiv.org/abs/2006.08097.
- Yi Yang, Yixuan Tang, and Kar Yan Tam. Investlm: A large language model for investment using financial domain instruction tuning. arXiv preprint arXiv:2309.13064, 2023. URL https://arxiv.org/abs/2309.13064.