

# Assessing the Capabilities and Limitations of FinGPT Model in Financial NLP Applications

Prudence Djabga\*

Chimezie A. Odinakachukwu<sup>†</sup>

## Abstract

This work evaluates FinGPT, a financial domain-specific language model, across six key natural language processing (NLP) tasks: Sentiment Analysis, Text Classification, Named Entity Recognition, Financial Question Answering, Text Summarization, and Stock Movement Prediction. The evaluation uses finance-specific datasets to assess FinGPT’s capabilities and limitations in real-world financial applications. The results show that FinGPT performs strongly in classification tasks such as sentiment analysis and headline categorization, often achieving results comparable to GPT-4. However, its performance is significantly lower in tasks that involve reasoning and generation, such as financial question answering and summarization. Comparisons with GPT-4 and human benchmarks highlight notable performance gaps, particularly in numerical accuracy and complex reasoning. Overall, the findings indicate that while FinGPT is effective for certain structured financial tasks, it is not yet a comprehensive solution. This research provides a useful benchmark for future research and underscores the need for architectural improvements and domain-specific optimization in financial language models.

**Keywords:** FinGPT, Financial NLP, Large Language Models, Sentiment Analysis, Named Entity Recognition, Stock Movement Prediction, Domain-Specific LLMs, Financial Question Answering, Text Summarization

## 1 Introduction

The financial industry has long been a pioneer in adopting cutting-edge technologies to enhance operational efficiency, accuracy, and strategic decision-making [2]. With the exponential growth of structured and unstructured data, particularly from news feeds, earnings reports, disclosures, and social media, there is an increasing demand for intelligent systems capable of processing human language at scale [11]. Initially, the industry relied on rule-based approaches and traditional statistical techniques such as bag-of-words and TF-IDF [28], which offered limited semantic understanding. As noted by Abubakar et al.[1], these limitations triggered a shift toward machine learning and deep learning models that, while better at capturing patterns, still required substantial domain-specific feature engineering.

This landscape was significantly transformed with the introduction of transformer-based architectures, most notably the Generative Pre-trained Transformer (GPT) family [5]. These models demonstrated the power of large-scale pretraining followed by task-specific fine-tuning, enabling generalization across diverse NLP tasks. Models such as GPT-3, GPT-4, BERT, and T5 have delivered state-of-the-art results in sentiment analysis, summarization, question answering, and named entity recognition [13]. Beyond LLMs, the broader field of Generative AI (GAI)—including GANs, VAEs, and diffusion models—has found increasing relevance in finance, facilitating applications such as synthetic data generation, automated reporting, and scenario simulation [32, 31].

LLMs have emerged as essential tools in processing unstructured financial text, especially models fine-tuned on finance-specific corpora like FinBERT, BloombergGPT, and FinGPT [4, 39]. Their capabilities in capturing complex linguistic nuances and domain-specific terminology make them highly effective in financial NLP tasks. However, challenges persist due to the specialized jargon, numerical reasoning, and the high-stakes context inherent in financial text. As highlighted by Qian et al.[29], understanding and improving the capabilities of these models through fine-tuning and rigorous benchmarking remains crucial to their reliable integration in finance. This intersection of GAI and finance thus provides a rich domain for innovation and intelligent decision-support systems.

### 1.1 Background

Recent advances in large language models (LLMs), such as GPT-3, GPT-4, and their instruction-tuned variants, have revolutionized general-purpose natural language understanding. However, applying these models effectively to domain-specific tasks, particularly in finance, remains a significant challenge. Financial texts often contain

\*Michigan State University djagbapr@msu.edu

<sup>†</sup>AIMS Senegal. chimezie.a.odinakachukwu@aims-senegal.org

specialized vocabulary, abbreviations, and implicit knowledge, which general models are not always optimized to handle.

While domain-specific models like **FinGPT** and **FinMA 7B** have emerged to address these gaps, their performance across a diverse set of financial NLP tasks remains uneven and often under-evaluated. Guo et al. [19] noted that many evaluations are narrow in scope, and Li et al. [25] emphasized the lack of systematic benchmarks and evaluation strategies across key financial tasks such as sentiment analysis, named entity recognition, and financial question answering.

Furthermore, there has been limited comparative analysis between general-purpose models like GPT-4 and domain-tuned models like FinGPT. This lack of comparison leaves unanswered questions about performance trade-offs in terms of accuracy, interpretability, and resource efficiency—critical considerations when deploying models in real-world financial systems where retraining is expensive.

To address this gap, this research evaluates the performance of FinGPT on six core financial NLP tasks and compares it with GPT-4 and FinMA 7B:

Table 1: Overview of the Six NLP Tasks and Their Corresponding Datasets

NLP Task	Dataset	Citation
Sentiment Analysis (SA)	FLARE-FPB, FinQA-SA	[35], [6]
Text Classification (TC)	FinGPT-Headline	[16]
Named Entity Recognition (NER)	FinGPT-NER	[17]
Financial Question Answering (QA)	ConvFinQA, FLARE-FinQA	[15], [9]
Stock Movement Prediction (SMP)	CIKM18, StockNet, BigData22	[7], [10], [34]
Text Summarization (Summ)	ECTSum	[8]

This research contributes a structured benchmark comparison and a practical methodology for evaluating large language models in financial NLP pipelines—balancing accuracy, efficiency, and domain alignment.

## 1.2 Research Questions

To guide this investigation, the following research questions were investigated:

- **RQ1:** How well does FinGPT perform on six core financial NLP tasks using real-world finance-specific datasets?
- **RQ2:** How does FinGPT’s performance compare with that of general-purpose models like GPT-4 and domain-tuned models like FinMA 7B on the same tasks?
- **RQ3:** What performance limitations emerge across tasks, and what insights can be drawn to guide future development of financial LLMs?

## 2 Evolution of Generative AI and Financial LLMs

The history of artificial intelligence (AI) dates back to the Dartmouth Summer Research Project (1956) [27], which envisioned that machine intelligence could replicate human learning. Since then, AI has undergone several milestones, particularly with the rise of generative models. Unlike discriminative models focused on classification, generative models aim to learn and recreate the data distribution, making them well-suited for text generation and structured reasoning.

Early innovations include Variational Autoencoders (VAEs) [22] and Generative Adversarial Networks (GANs) [18], which laid the foundation for modern generative systems. The introduction of Transformer architectures by Vaswani et al. in 2017 [38] marked a pivotal shift in deep learning, enabling scalable and context-aware sequence modeling. This breakthrough gave rise to the Generative Pre-trained Transformer (GPT) series—GPT-2, GPT-3, and GPT-4—which demonstrated exceptional performance across natural language processing (NLP) tasks including summarization, question answering, and translation [5, 43].

In finance, the shift from bag-of-words to transformer-based LLMs enabled models to capture domain-specific context more effectively. Pre-trained models like BERT [13] led to specialized variants such as FinBERT [4] and FinBERT-21 [20], improving sentiment analysis and entity recognition in financial texts. Proprietary models like BloombergGPT [39] trained on hybrid corpora expanded task coverage but remained inaccessible to most users due to licensing and compute barriers.

Open-source LLMs such as the LLaMA series [37] and FinGPT [26] offer scalable alternatives, often leveraging parameter-efficient fine-tuning (PEFT) and real-time training. Similarly, FinMA [40] and InvestLM [42] are optimized for financial tasks like market analysis and regulatory summarization. Figure 1 and Figure 2 illustrate the evolution from general to domain-specific LLMs in finance.

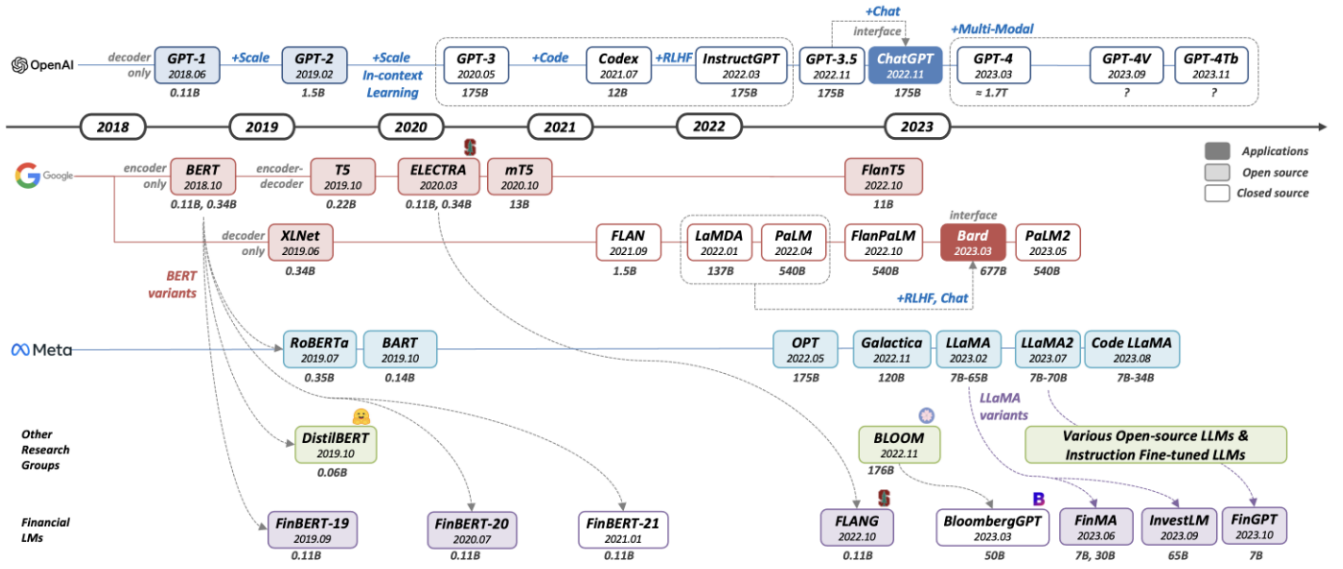


Figure 1: Timeline showing the evolution of selected PLM/LLM releases from the general domain to the financial domain [23].

## 2.1 Challenges in Financial AI Deployment

Deploying LLMs in finance introduces unique challenges. Model transparency, data privacy, and regulatory compliance are critical concerns. Black-box behaviors and hallucinations hinder trust, while limited labeled data restricts training quality [30, 24].

Retrieval-Augmented Generation (RAG) offers a promising solution by allowing models to access private databases during inference without retraining [24]. Meanwhile, evaluation metrics like accuracy and F1 often fall short in assessing financial task performance, prompting a shift toward expert-in-the-loop assessment.

Infrastructure constraints also limit adoption, especially due to high computational costs and memory demands. Open-access models such as FinGPT offer a cost-effective alternative but still lag behind GPT-4 in complex tasks. Models like FinGPT-13B remain underutilized due to resource limits.

Category	Model	Backbone	Paras.	Techniques	PT	Evaluation		Open Source			
					PT Data Size	Task	Dataset	Model	PT	IFT	Venue
FinPLM (Disc.)	FinBERT-19 [Araci, 2019]	BERT	0.11B	Post-PT, FT	(G) 3.3B words (F) 29M words	[SA]	FPB, FiQA-SA	Y	N	N	ArXiv Aug 2019
	FinBERT-20 [Yang <i>et al.</i> , 2020]	BERT	0.11B	PT, FT	(F) 4.9B tokens	[SA]	FPB, FiQA-SA, AnalystTone	Y	Y	N	ArXiv Jul 2020
	FinBERT-21 [Liu <i>et al.</i> , 2021]	BERT	0.11B	PT, FT	(G) 3.3B words (F) 12B words	[SA], [QA] [SBD]	FPB, FiQA-SA, FiQA-QA FinSBD19	N	N	N	IJCAI (S) Jan 2021
	FLANG [Shah <i>et al.</i> , 2022]	ELECTRA	0.11B	PT, FT	(G) 3.3B words (F) 696k docs	[SA], [TC] [NER], [QA], [SBD]	FPB, FiQA-SA, Headline FIN, FiQA-QA, FinSBD21	Y	Y	N	EMNLP Oct 2022
FinLLM (Gen.)	BloombergGPT [Wu <i>et al.</i> , 2023]	BLOOM	50B	PT, PE	(G) 345B tokens (F) 363B tokens	[SA], [TC] [NER], [QA]	FPB, FiQA-SA, Headline FIN, ConvFinQA	N	N	N	ArXiv Mar 2023
	FinMA [Xie <i>et al.</i> , 2023]	LLaMA	7B, 30B	IFT, PE	(G) 1T tokens	[SA], [TC], [NER], [QA] [SMP]	FPB, FiQA-SA, Headline FIN, FinQA, ConvFinQA, StockNet, CIKM18, BigData22	Y	Y	Y	NIPS (D) Jun 2023
	InvestLM [Yang <i>et al.</i> , 2023c]	LLaMA	65B	IFT, PE PEFT	(G) 1.4T tokens	[SA], [TC] [QA], [Summ]	FPB, FiQA-SA, FOMC FinQA, ECTSum	Y	N	N	ArXiv Sep 2023
	FinGPT [Wang <i>et al.</i> , 2023]	6 open-source LLMs	7B	IFT, PE PEFT	(G) 2T tokens (e.g. LLaMA2)	[SA], [TC] [NER], [RE]	FPB, FiQA-SA, Headline FIN, FinRED	Y	Y	Y	NIPS (W) Oct 2023

Figure 2: Summary of FinPLMs and FinLLMs. Abbreviations include Paras. = Parameter Size, PT = Pretraining, FT = Fine-Tuning, SA = Sentiment Analysis, QA = Question Answering, SMP = Stock Movement Prediction, Summ = Summarization, and others [26].

## 2.2 Enabling Responsible AI in Finance

To foster responsible AI use, federated learning (FL) [14] enables collaborative model training across institutions while preserving data privacy. Explainable AI (XAI), particularly SHAP values, improves model interpretability for sensitive decisions like fraud detection and credit scoring.

Lastly, upskilling professionals and bridging the AI-literacy gap are essential. Industry-academic partnerships and targeted curriculum reforms are crucial to ensure long-term adaptability and trust [12, 3, 21].

This shift from general-purpose to specialized financial LLMs highlights the growing importance of transparency, adaptability, and scalability in real-world financial NLP applications.

## 3 Methods

### 3.1 Overview of FinGPT Framework

FinGPT is an open-source framework designed to enable the application of Large Language Models (LLMs) in finance. It addresses challenges such as market volatility, heterogeneous data sources, and real-time text processing. As illustrated in Figure 3, the FinGPT architecture consists of four modular layers:

- **Data Source:** Collects structured and unstructured data from financial APIs, social media, news, and filings.
- **Data Engineering:** Handles cleaning, labeling, and streaming of data.
- **LLMs:** Utilizes transformer-based architectures (e.g., LLaMA) fine-tuned for finance.
- **Applications:** Deploys models to various tasks such as sentiment analysis, question answering, and stock movement prediction.

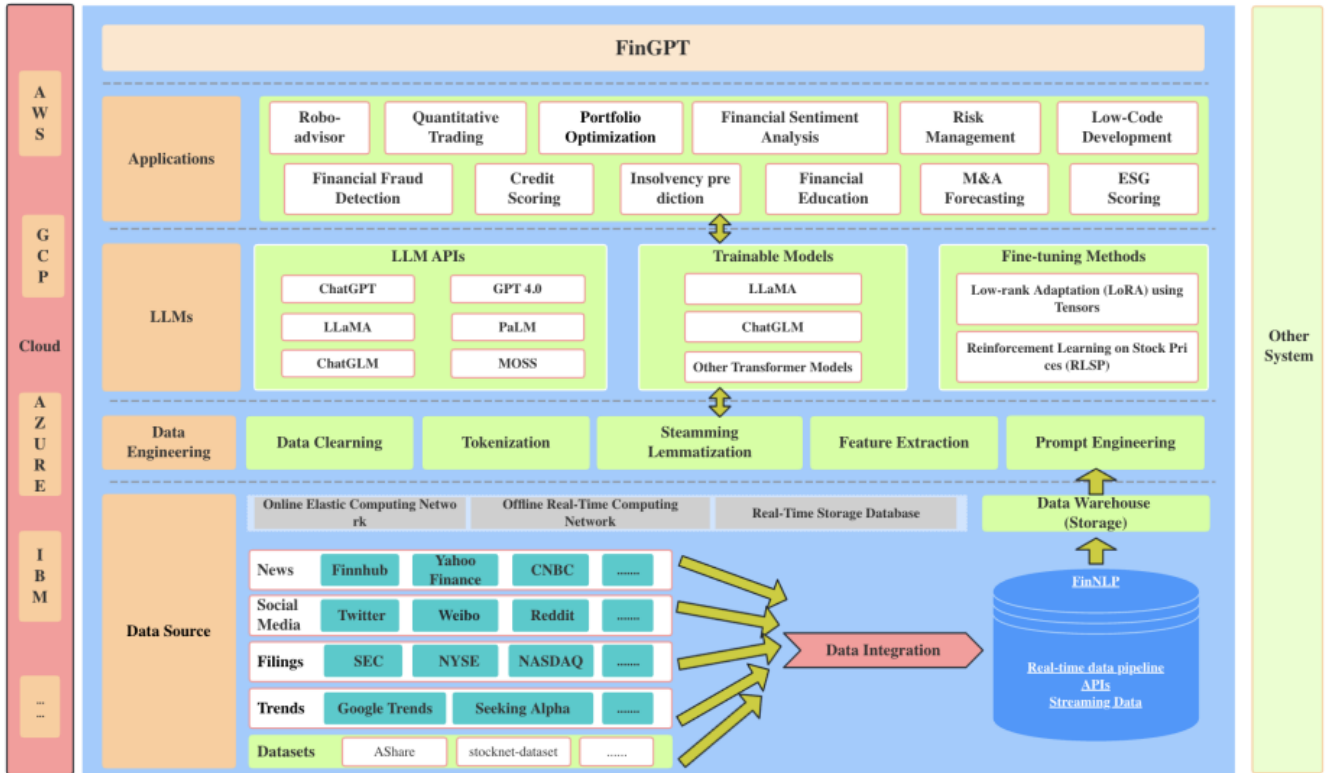


Figure 3: FinGPT system architecture highlighting the end-to-end data and modeling pipeline [41].

### 3.2 LLaMA2: Model Foundation for FinGPT

FinGPT is based on the LLaMA2 (Large Language Model Meta AI) architecture, a decoder-only transformer introduced by Touvron et al. [37]. The LLaMA2 model inherits the Transformer decoder architecture originally proposed by Vaswani et al. [38], optimized for autoregressive language modeling.

### 3.2.1 Transformer Decoder Architecture

The model comprises  $N$  stacked decoder blocks, each consisting of:

1. Multi-Head Self-Attention (MHSA)
2. Feed-Forward Networks (FFN)

For an input sequence  $X = (x_1, x_2, \dots, x_T)$  with embeddings  $x_i \in \mathbb{R}^d$ , the model computes token-wise representations through successive MHSA and FFN layers, incorporating residual connections and normalization.

### 3.2.2 Multi-Head Self-Attention

Each attention head is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where  $Q = XW^Q$ ,  $K = XW^K$ ,  $V = XW^V$  are projections with learnable matrices  $W^Q$ ,  $W^K$ , and  $W^V \in \mathbb{R}^{d \times d_k}$ .

The multi-head extension is given by:

$$\text{MHSA}(X) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

### 3.2.3 Rotary Positional Embedding (RoPE)

LLaMA replaces absolute positional encodings with Rotary Positional Embeddings (RoPE), introducing position-dependent rotations to maintain relative token information [33]. This enhances generalization across varying sequence lengths.

### 3.2.4 Feed-Forward Networks

Each token’s representation is passed through a two-layer FFN:

$$\text{FFN}(x) = W_2 \cdot \text{GELU}(W_1x + b_1) + b_2$$

### 3.2.5 Normalization and Residual Connections

Layer outputs are stabilized via pre-layer normalization and residual connections:

$$x' = x + \text{MHSA}(\text{LayerNorm}(x)), \quad x'' = x' + \text{FFN}(\text{LayerNorm}(x'))$$

### 3.2.6 Training Objective

The model is trained using next-token prediction by minimizing the negative log-likelihood:

$$\mathcal{L} = - \sum_{t=1}^T \log P(x_t \mid x_{<t}; \theta)$$

## 3.3 Adaptation in FinGPT

FinGPT applies domain-specific adaptation to LLaMA through fine-tuning on financial corpora. While the underlying transformer structure remains intact, adaptations include:

- Task-aligned instruction tuning (e.g., classification, QA)
- Financial dataset pretraining and finetuning
- Lightweight model variants for lower resource environments

These adaptations allow FinGPT to address complex financial NLP tasks with improved accuracy and relevance.

## 4 Evaluation

### 4.1 Task Overview

To evaluate FinGPT’s applicability across financial natural language processing (NLP) tasks, we designed a standardized evaluation pipeline involving six key tasks: sentiment analysis, text classification, named entity recognition, question answering, stock movement prediction, and text summarization. The methodology includes dataset selection, preprocessing, model configuration, inference, and metric-based evaluation.

### 4.2 Sentiment Analysis

#### 4.2.1 Datasets

Two open-source datasets were selected:

- **FLARE-FPB** [36]: Contains labeled financial texts (positive, neutral, negative); 970 samples from the test split were used.
- **FLARE-FIQASA** [6]: Comprises 235 financial QA-style texts; labels were extracted from the `answer` field.

#### 4.2.2 Preprocessing

The preprocessing stage involved formatting each sample as an instruction-style prompt, where the model was explicitly guided to identify sentiment from financial text. All inputs were lowercased and normalized to ensure consistency across label representations. Tokenization was performed using the `AutoTokenizer` from the Hugging Face Transformers library, with padding aligned to the end-of-sequence (EOS) token to ensure compatibility with the decoder-only architecture. Notably, no aggressive text cleaning was applied in order to preserve domain-specific entities such as stock tickers (e.g., `$AAPL`) and financial indicators.

#### 4.2.3 Model Configuration

The experiments employed the `NousResearch/Llama-2-13b-hf` model as the base architecture. This model was extended using a Low-Rank Adaptation (LoRA) module, specifically the adapter `oliverwang15/FinGPT_v33_...`, tailored for sentiment classification in finance. To optimize memory usage and enable efficient inference, the model was loaded in 8-bit precision with weights stored in `float16`. The model was set to evaluation mode using `model.eval()` to deactivate training layers and ensure deterministic behavior during generation.

#### 4.2.4 Inference

Inference was performed in batches of four for the FLARE-FIQASA dataset, while a single-batch mode was used for FLARE-FPB due to differing sequence lengths. Text generation was executed using greedy decoding (`do_sample=False`) with a maximum output length of 32 tokens for FLARE-FPB and 512 tokens for FLARE-FIQASA to accommodate longer responses. The generated outputs were subsequently normalized and mapped to the expected sentiment categories to facilitate evaluation.

### 4.3 Text Classification

#### 4.3.1 Dataset

The `FinGPT Headline Classification` dataset [16] was used, containing financial headlines labeled with binary sentiment (`yes/no`).

#### 4.3.2 Preprocessing

Each input headline was embedded into an instruction-style template to guide the model’s classification task. The prompt format followed the structure:

```
[INST] Classify the sentiment of the following financial headline: <HEADLINE> [/INST]
```

Tokenization was handled using the `AutoTokenizer` associated with the LLaMA-2 model. Left-padding was applied, and special attention was given to align special tokens with the EOS marker. To ensure consistent label evaluation, synonymous expressions such as “absolutely” or “nope” were normalized and mapped to binary sentiment labels. Ambiguous cases labeled as `maybe`, or outputs that could not be confidently categorized, were excluded from final evaluation to avoid metric distortion.

### 4.3.3 Model Setup

The classification model used `meta-llama/Llama-2-7b-hf` as the base architecture. To adapt it for financial sentiment classification without full fine-tuning, a Low-Rank Adaptation (LoRA) adapter from the FinGPT project `FinGPT/fingpt-mt_llama2-7b_lora`—was applied using the `peft` (Parameter-Efficient Fine-Tuning) library.

### 4.3.4 Inference and Prediction

During inference, the model generated sentiment predictions using the `generate()` function, constrained to a maximum of 10 tokens to enforce concise outputs. The generated outputs were decoded and passed through a normalization function that mapped the text to standardized sentiment categories (`yes`, `no`, or `unknown`) based on keyword matching:

```
def extract_label(output_text):
    output_text = output_text.lower()
    if "yes" in output_text:
        return "yes"
    elif "no" in output_text:
        return "no"
    else:
        return "unknown"
```

### 4.3.5 Evaluation Metrics

Model performance was assessed using standard classification metrics, including precision, recall, and F1-score, calculated over the `yes` and `no` classes. Any outputs categorized as `unknown`—due to lack of recognizable sentiment indicators—were excluded from the score aggregation to maintain the reliability of the evaluation.

## 4.4 Named Entity Recognition (NER)

### 4.4.1 Dataset Description

The Named Entity Recognition (NER) task was evaluated using the publicly available `FinGPT/fingpt-ner` dataset from Hugging Face (<https://huggingface.co/datasets/FinGPT/fingpt-ner>). This dataset contains financial-domain sentences annotated with entities from three classes: `person`, `organization`, and `location`. Each sample comprises a sentence and an expected output string in the format “Entity is a [type]”. The test set, consisting of 98 examples, was used as-is for zero-shot or few-shot evaluation without additional annotation or manual filtering.

### 4.4.2 Preprocessing

Each sentence was tokenized using the LLaMA-2 tokenizer (`meta-llama/Llama-2-7b-hf`) with left-padding and a maximum token length of 512 to suit decoder-style generation. A consistent instruction-style prompt was applied to all inputs:

Instruction: Please extract entities and their types from the input sentence ,  
entity types should be chosen from {person/organization/location}.

Input: <sentence>

Answer:

The generated outputs were post-processed to align with BIO tagging standards. Entities were extracted and mapped to token spans using the B-{TYPE}, I-{TYPE}, and O schema. Samples with token-output mismatches were excluded, although none were discarded during this experiment.

### 4.4.3 Model Configuration

The experiment used the `meta-llama/Llama-2-7b-hf` model as the base, with a parameter-efficient LoRA adapter from the FinGPT project (`FinGPT/fingpt-mt_llama2-7b_lora`) for financial-domain specialization. The model and adapter were integrated using the `transformers` and `peft` libraries. Evaluation was conducted using `float16` precision and automatic GPU allocation via `device_map="auto"`.

#### 4.4.4 Inference Procedure

The model was prompted using the instruction-based format and decoded using greedy decoding (`do_sample=False`). Extensive parameter tuning was required to ensure performance:

- **Maximum Length:** Set to 500 to balance output coherence and memory usage.
- **Max New Tokens:** Reduced from 64 to 34, which improved macro F1 from 38% to approximately 69%, minimizing hallucination.
- **Batch Size:** Set to 1 to accommodate hardware constraints and prevent instability during decoding.

These adjustments underscore the importance of careful generation parameter calibration when using autoregressive models like LLaMA for structured tasks such as NER. Despite architectural limitations, the model achieved robust entity extraction when appropriately tuned.

## 4.5 Financial Question Answering

### 4.5.1 Dataset Description

To assess FinGPT’s capability in financial question answering, we evaluated its performance on two benchmark datasets: **ConvFinQA** and **FLARE-FinQA**.

**ConvFinQA** is designed for multi-step numerical reasoning within a conversational context. It was accessed from Hugging Face at <https://huggingface.co/datasets/FinGPT/fingpt-convfinqa>, and the first 200 examples from the `test` split were selected. Each sample comprises a question and a numeric answer, requiring arithmetic reasoning and context comprehension.

**FLARE-FinQA** was obtained from <https://huggingface.co/datasets/ChanceFocus/flare-finqa>, and the first 50 test samples were used. It contains natural language queries with corresponding numeric or binary (“yes”/“no”) answers, testing both factual recall and logical reasoning in financial contexts.

### 4.5.2 Preprocessing

Preprocessing included prompt formatting, tokenization, answer extraction, and filtering. Both datasets were converted into a consistent prompt format to elicit concise numerical responses:

```
Please answer the given financial question based on the context.
Question: {question}
Answer:
```

The true answers were extracted from each dataset’s ground truth field. Tokenization was applied using the LLaMA-2 tokenizer with padding and truncation, using a maximum sequence length of 768 for FLARE-FinQA and 1012 for ConvFinQA. Batch sizes were set to 4 and 8 respectively.

Model outputs were cleaned with regular expressions to isolate numeric values. Only examples where both prediction and ground truth were valid numbers were retained for evaluation, ensuring fairness and consistency in scoring.

### 4.5.3 Model Configuration

The experiments utilized a parameter-efficient fine-tuning (PEFT) setup based on the `meta-llama/Llama-2-7b-hf` model. A domain-specific LoRA adapter from the FinGPT project (`FinGPT/fingpt-mt_llama2-7b_lora`) was integrated using the `peft` library.

The model was loaded with `float16` precision and assigned automatically to available GPUs via `device_map="auto"` to optimize runtime performance.

### 4.5.4 Inference Strategy

Inference was conducted using beam search with 2 beams and a temperature of 0.7. The maximum generation length was capped at 34 tokens to constrain output length and reduce hallucination. The model was executed within a `torch.no_grad()` context to minimize memory overhead during evaluation.

After generation, the predicted outputs were parsed to extract numerical values for comparison with ground truth. For FLARE-FinQA, if multiple numbers were present, the one with the smallest absolute difference from the true value was used as the final prediction.

This setup allowed us to benchmark FinGPT’s ability to perform structured, quantitative financial reasoning under realistic evaluation conditions.



## 4.6 Stock Movement Prediction

### 4.6.1 Dataset Description

To evaluate FinGPT’s performance on the Stock Movement Prediction (SMP) task, we employed three publicly available datasets accessed via Hugging Face. These datasets contain aligned financial news content and corresponding stock movement labels suitable for classification tasks:

- **CIKM18 (flare-ECTSum)** [7]  
*Source:* <https://huggingface.co/datasets/ChanceFocus/flare-ectsum>  
A benchmark dataset from the CIKM 2018 Financial Opinion Mining Challenge. It includes news summaries annotated with subsequent stock movement labels (up or down), providing insight into event-driven market response.
- **StockNet (flare-SM-ACL)** [10]  
*Source:* <https://huggingface.co/datasets/ChanceFocus/flare-sm-acl>  
Comprises sentiment-enriched financial news texts and corresponding price direction labels. It is widely adopted for evaluating sentiment-based stock forecasting models.
- **BigData22 (flare-SM-BigData)** [34]  
*Source:* <https://huggingface.co/datasets/TheFinAI/flare-sm-bigdata>  
A large-scale dataset integrating diverse financial sources such as news articles and social media with stock price annotations. It supports broader-scale evaluation under high-volume scenarios.

All datasets were loaded using the `datasets` library with the `split="test"` configuration. Labels were provided in binary or ternary format (`up`, `down`, `stable`), enabling a realistic multi-class prediction scenario.

### 4.6.2 Preprocessing

Standard preprocessing was performed to ensure consistency and compatibility across all datasets:

- **Text Cleaning:** Raw texts were cleaned by removing URLs, redundant whitespace, and special symbols. All inputs were lowercased for uniformity.
- **Date Alignment:** Stock movement labels were matched to news reports based on timestamps. Misaligned or incomplete entries were excluded.
- **Label Encoding:** Movement directions were mapped to numeric values: `up`  $\rightarrow$  1, `down`  $\rightarrow$  0, and `stable` handled conditionally based on task formulation.
- **Tokenization:** Cleaned text was tokenized using the LLaMA-2 tokenizer with padding and truncation applied to respect maximum sequence lengths.
- **Batching:** Preprocessed inputs were grouped using PyTorch’s `DataLoader` to enable efficient GPU-accelerated inference.

### 4.6.3 Model Configuration

The model used for this task was **FinGPT-Forecaster**, a parameter-efficient fine-tuned variant of **LLaMA-2-7B-chat**:

- **Base Model:** `meta-llama/Llama-2-7b-chat-hf`, a decoder-only language model.
- **LoRA Adapter:** `FinGPT/fingpt-forecaster_dow30_llama2-7b_lora`, a fine-tuned checkpoint tailored to Dow Jones-related stock prediction.
- **Tokenizer:** The matching LLaMA-2 tokenizer was used to prepare inputs.
- **Hardware and Precision:** The model was loaded using `float16` precision and automatically deployed to GPU using `device_map="auto"` for optimized execution.
- **Libraries:** Hugging Face’s `transformers` and `peft` libraries were used for model initialization and adapter integration.

#### 4.6.4 Inference Strategy

The inference procedure followed a structured and reproducible pipeline:

- **Prompt Construction:** Each news summary or article was embedded into a structured prompt suitable for the LLaMA-2-chat model.
- **Generation:** The model generated textual outputs (e.g., “up” or “down”) using greedy decoding or controlled generation parameters.
- **Prediction Extraction:** Output strings were parsed with regular expressions to extract the predicted class.
- **Batch Execution:** Predictions were performed in mini-batches using PyTorch, leveraging GPU acceleration.
- **Postprocessing:** Textual predictions were mapped back to numeric labels to facilitate quantitative evaluation.

This pipeline ensured compatibility across datasets, efficiency in execution, and consistency in prediction output, forming a solid foundation for analyzing FinGPT’s forecasting capabilities on financial time series events.

### 4.7 Text Summarization

While FinGPT demonstrated strong performance across several financial NLP tasks, its ability to perform abstractive text summarization was notably limited. This was assessed using the **flare-ECTSum** dataset, which consists of financial event summaries derived from economic news sources. Despite several attempts, FinGPT failed to generate coherent or informative summaries, often producing generic or fragmented text unrelated to the input.

This shortcoming prompted a technical investigation into the architectural limitations of FinGPT’s underlying model—LLaMA—particularly in the context of tasks requiring long-range dependency modeling and bidirectional context understanding.

#### 4.7.1 Limitations of Decoder-Only Architectures for Summarization

FinGPT is built on the LLaMA architecture, a decoder-only causal language model (CLM) trained using the standard autoregressive objective:

$$P(x_1, x_2, \dots, x_n) = \prod_{t=1}^n P(x_t \mid x_1, x_2, \dots, x_{t-1})$$

This unidirectional generation paradigm is highly effective for tasks such as next-token prediction, general text generation, and constrained question answering. However, it introduces inherent limitations when applied to abstractive summarization tasks, which require a global understanding of the entire input before generating a condensed version.

In contrast, encoder-decoder models like T5 and BART are designed specifically for conditional sequence generation. These models learn the probability of generating a summary sequence  $y = (y_1, \dots, y_m)$  given an input sequence  $x = (x_1, \dots, x_n)$  as follows:

$$P(y_1, y_2, \dots, y_m \mid x_1, x_2, \dots, x_n)$$

In this formulation:

- The encoder processes the full input  $x$  to create a context-rich representation.
- The decoder generates the output  $y$  conditioned on this representation, allowing it to consider the entire document structure and semantics.

This bidirectional encoding mechanism enables encoder-decoder models to excel at summarization, particularly in financial texts where salient information may be spread across multiple clauses or sentences.

#### 4.7.2 Implication for FinGPT

Given its decoder-only design, FinGPT lacks the full-context representation necessary for accurate and concise summarization. Additionally, financial text often contains complex structures, nested events, and high information density—all of which are poorly captured when the model can only attend to past tokens. This architectural constraint explains the model’s poor performance on the **flare-ECTSum** dataset.

Future work could involve incorporating retrieval-augmented generation (RAG) or transitioning to hybrid or encoder-decoder models to improve summarization performance in domain-specific LLMs such as FinGPT.

## 4.8 Summary Table

Table 2 provides a comprehensive summary of FinGPT’s performance across six key financial NLP tasks, benchmarked against FinMA 7B, GPT-4, human performance (where available), and traditional baselines.

FinGPT demonstrates strong performance in sentiment analysis and text classification, with F1-scores rivaling or surpassing GPT-4 in certain cases. In named entity recognition, the model achieves moderate accuracy, though it trails GPT-4, indicating room for improvement in structured output tasks.

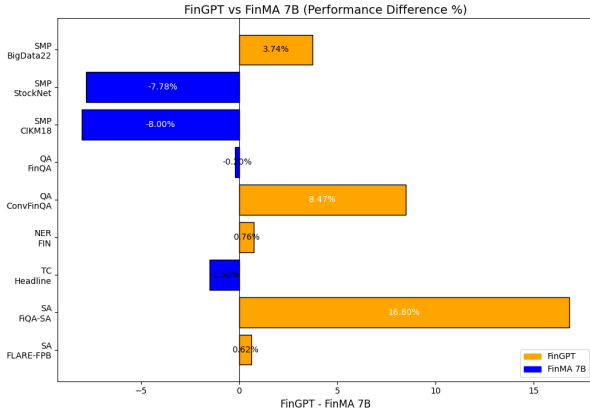
In contrast, FinGPT shows significant limitations in tasks that require numerical reasoning and deep context understanding—most notably in financial question answering (QA), where performance lags well behind GPT-4 and human baselines. Summarization remains the most challenging task, as expected for decoder-only architectures like LLaMA.

For stock movement prediction, FinGPT performs moderately across three financial datasets. While not state-of-the-art, these results highlight its generalization ability in high-volatility financial forecasting scenarios.

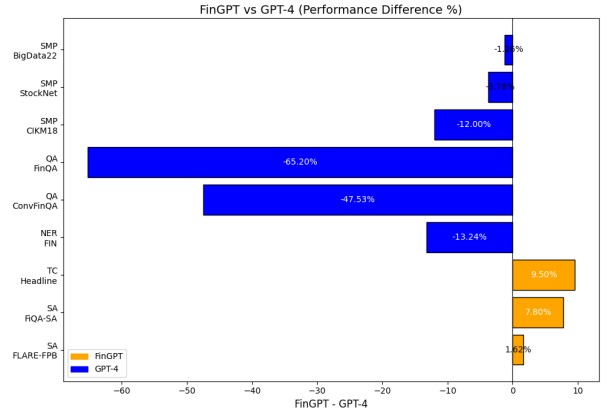
These findings underscore the importance of domain-specific tuning, model architecture, and task complexity in evaluating large language models for finance.

Table 2: Comparative Performance (%) Across Tasks for FinGPT, FinMA 7B, GPT-4, Human, and Baseline Benchmarks

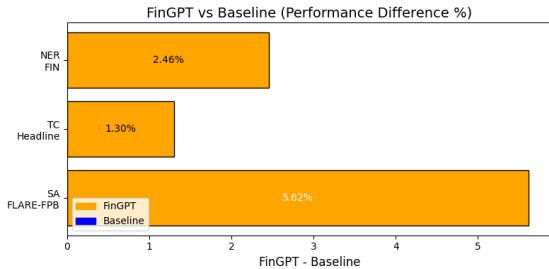
Task	Dataset	FinGPT	FinMA 7B	GPT-4	Human	Baseline
SA	flare-FPB	87.62 (F1)	87.00 (F1)	86.00 (F1)	—	82.00 (F1)
	FiQA-SA	95.80 (F1)	79.00 (F1)	88.00 (F1)	—	—
TC	Headline	95.50 (F1)	97.00 (F1)	86.00 (Avg. F1)	—	94.20 (F1)
NER	FIN	69.76 (Entity.F1)	69.00 (Entity.F1)	83.00 (Entity.F1)	—	67.30 (F1)
QA	ConvFinQA	28.47 (EM)	20.00 (EM)	76.00 (EM)	89.00 (EM)	—
	FinQA	3.80 (EM)	4.00 (EM)	69.00 (EM)	91.00 (EM)	—
SMP	CIKM18	45.00 (F1)	53.00 (F1)	57.00 (Acc)	—	—
	StockNet	48.22 (F1)	56.00 (F1)	52.00 (Acc)	—	—
	BigData22	52.74 (F1)	49.00 (F1)	54.00 (Acc)	—	—
TS	ECTSum	—	8.00 (ROUGE-1)	30.00 (ROUGE-1)	—	—



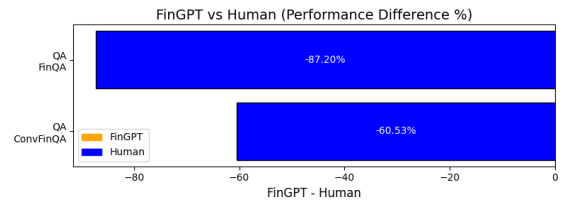
(a) FinGPT vs FinMA 7B



(b) FinGPT vs GPT-4



(a) FinGPT vs Baseline Acc



(b) FinGPT vs Human Acc

## 5 Results and discussions

This chapter presents the empirical results of the evaluation of FinGPT in six key financial NLP tasks. The analysis highlights performance strengths and weaknesses compared to domain-specific models (e.g. FinMA 7B), general-purpose models (e.g. GPT-4), and baseline or human references. Each section includes quantitative metrics and corresponding interpretations.

## 6 Sentiment Analysis

Table 3: Accuracy Comparison on Financial Sentiment Classification

Dataset	FinGPT (Acc)	FinMA 7B (F1)	GPT-4 (F1)	Human Acc	Baseline (F1)
FLARE-FPB	87.62%	87.00%	86.00%	–	82.00%
FLARE-FIQA-SA	95.74%	79.00%	88.00%	–	–

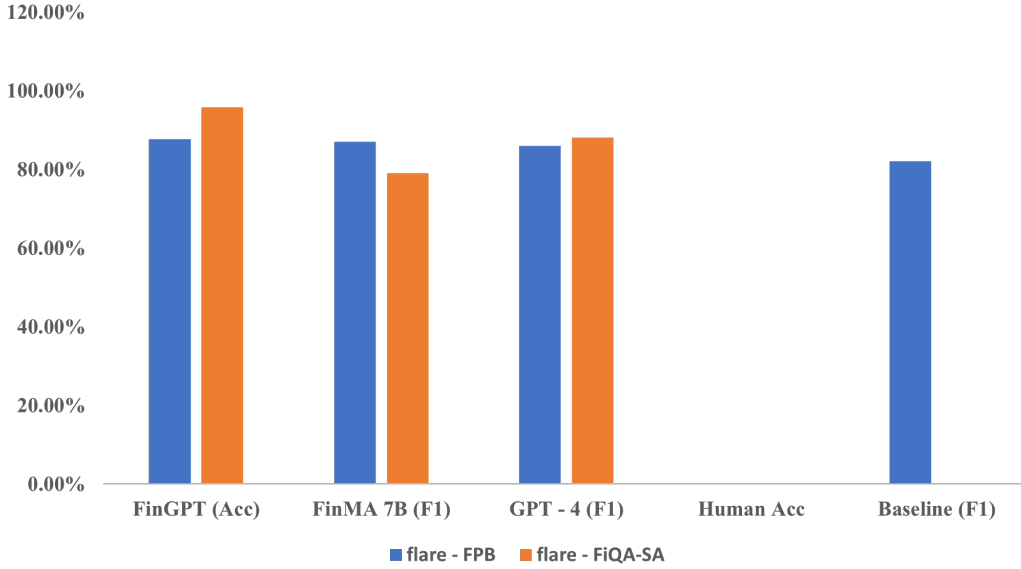


Figure 6: Performance Comparison on Financial Sentiment Datasets

**Interpretation:** FinGPT demonstrated strong performance on both sentiment analysis datasets. On FLARE-FPB, it marginally outperformed GPT-4 and FinMA 7B, and achieved the highest accuracy on FLARE-FIQA-SA. These results suggest that FinGPT is highly effective at interpreting sentiment in financial texts, outperforming the baseline and benefiting from domain-specific alignment.

## 7 Text Classification

Table 4: Average F1 Score Comparison for Financial Headline Classification

Model	FinGPT (Avg. F1)	FinMA 7B (F1)	GPT-4 (Avg. F1)	Human	Baseline (F1)
Headline Dataset	95.50%	97.00%	86.00%	–	94.20%

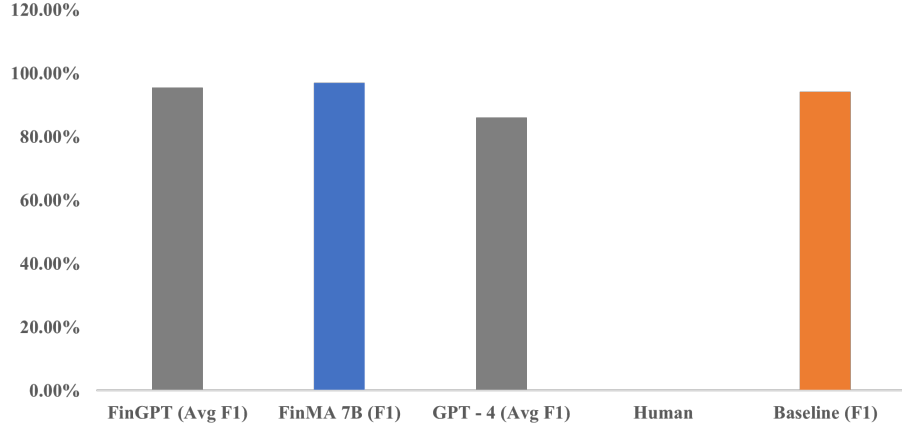


Figure 7: Performance on Headline-Based Text Classification

**Interpretation:** On the headline classification task, FinGPT performed competitively, achieving an average F1 score of 95.5%, slightly above the baseline and GPT-4. FinMA 7B attained the highest score, highlighting the potential advantage of more extensive fine-tuning. Nonetheless, FinGPT proves effective for financial text classification, reinforcing the utility of specialized language models in finance.

## 8 Named Entity Recognition (NER)

Table 5: Entity-Level F1 Comparison for Named Entity Recognition (NER)

Model	FinGPT (Entity F1)	FinMA 7B (F1)	GPT-4 (Entity F1)	Baseline (F1)
FIN Dataset	69.76%	69.00%	83.00%	67.30%

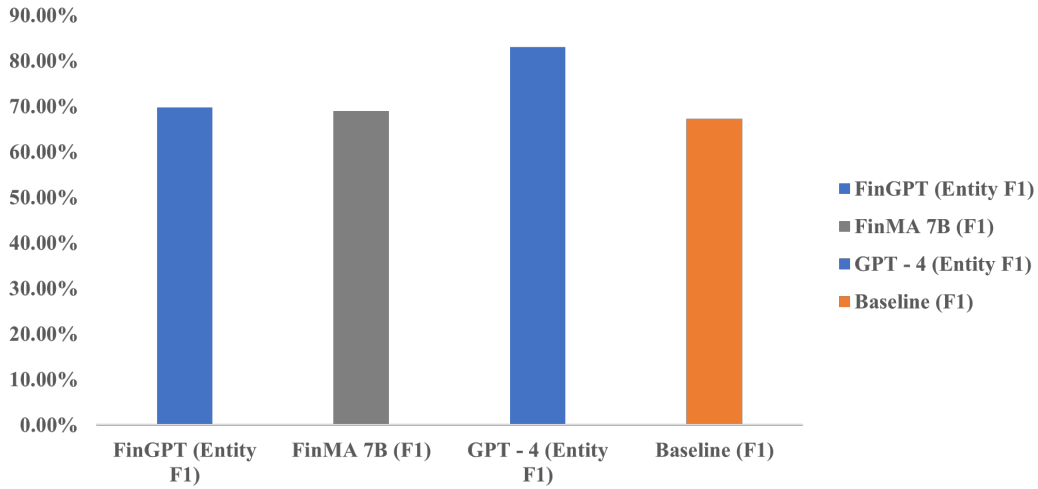


Figure 8: Comparison for Named Entity Recognition (NER)

**Interpretation:** FinGPT and FinMA 7B show comparable results, slightly outperforming the baseline. However, GPT-4 significantly surpasses both models, indicating a superior ability to extract and categorize financial entities. This underscores the challenge of complex entity recognition in finance and reveals areas where FinGPT can benefit from further refinement.

## 9 Financial Question Answering

Table 6: Exact Match (EM) Accuracy Comparison for Financial QA

Dataset	FinGPT (EM)	FinMA 7B (EM)	GPT-4 (EM)	Human (EM)
ConvFinQA	28.4%	20.0%	76.0%	89.0%
FLARE-FinQA	3.8%	4.0%	69.0%	91.0%

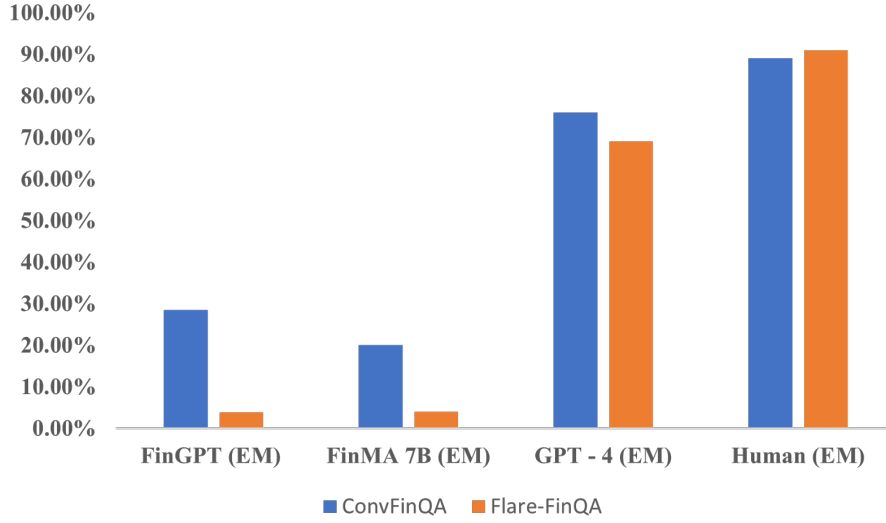


Figure 9: Comparison of Model Performance on Financial QA Datasets

**Interpretation:** FinGPT and FinMA 7B struggle with financial question answering, particularly on FLARE-FinQA, which demands complex reasoning and precise numerical answers. The gap between these models and GPT-4 or human performance is substantial, indicating that existing domain-specific models require deeper reasoning capabilities and enhanced numerical understanding.

## 10 Stock Movement Prediction (SMP)

Table 7: Accuracy Comparison for Stock Movement Prediction

Dataset	FinGPT (Acc)	FinMA 7B (Acc)	GPT-4 (Acc)	Human Acc
StockNet	48.47%	56.00%	52.00%	—
CIKM18	47.03%	53.00%	57.00%	—
BigData22	52.83%	49.00%	54.00%	—

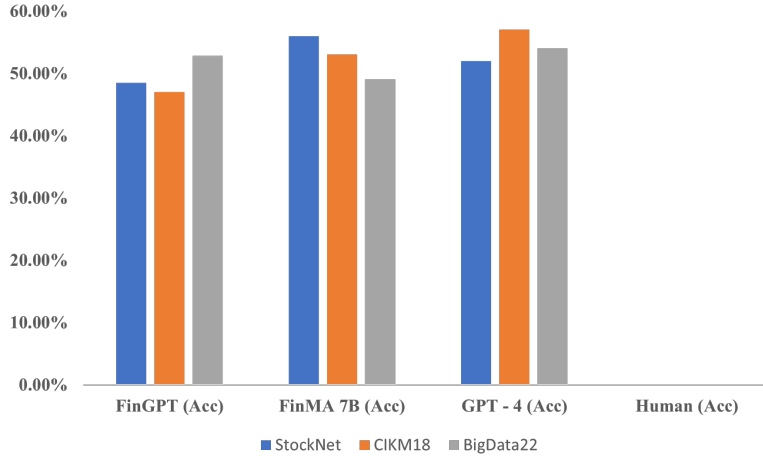


Figure 10: Stock Movement Prediction Accuracy Across Datasets

**Interpretation:** FinGPT demonstrates moderate accuracy across all stock movement datasets. While it generally trails behind GPT-4 and FinMA 7B, the results reflect the inherent difficulty of forecasting market trends using text data alone. These findings reinforce the challenge of financial prediction tasks, especially in the face of noise and volatility.

## 10.1 Directional Sensitivity Analysis in Stock Movement Prediction

To further assess FinGPT’s real-world applicability, we conducted a directional sensitivity analysis to evaluate whether it performs better in bullish or bearish market phases.

### 10.1.1 CIKM18 Dataset

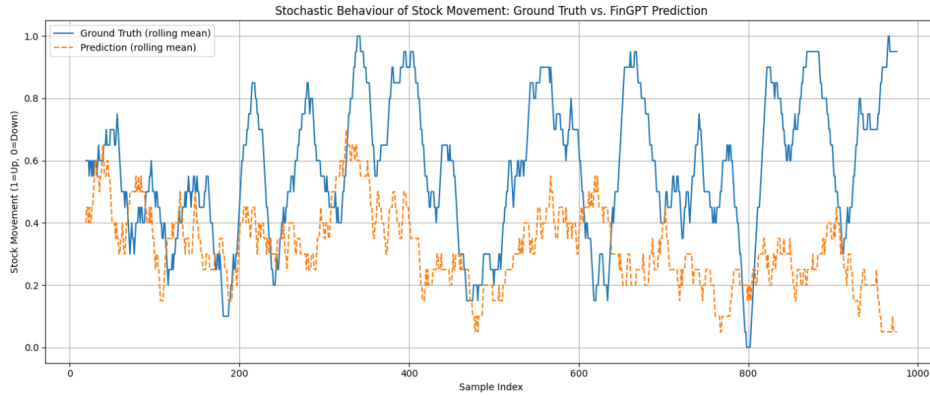


Figure 11: Rolling Mean of FinGPT vs Ground Truth on CIKM18

As shown in Figure 11, FinGPT aligns closely with upward trends but lags or flattens during market downturns. This suggests a directional bias toward bullish sentiment.

### 10.1.2 Directional Bias Observation

This bullish bias is consistent across datasets. Figures 12 and 13 display simulated portfolio performance, showing stronger returns when the model predicts buy signals versus sell signals.

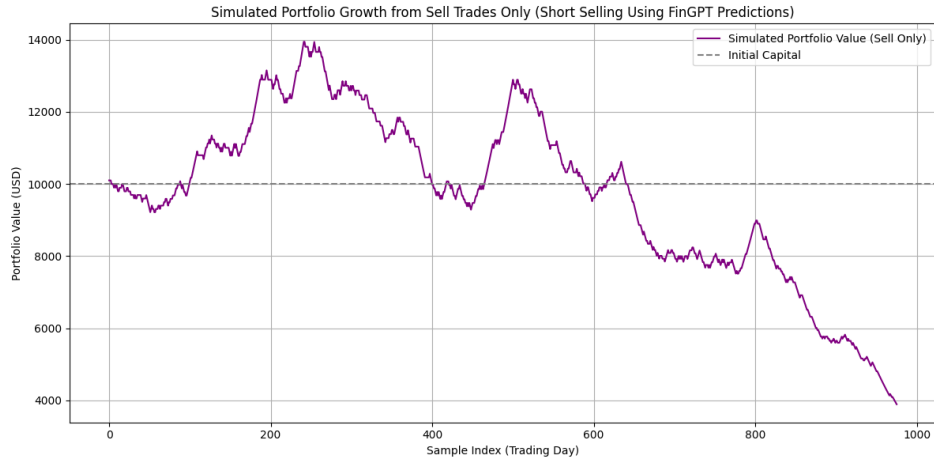


Figure 12: Portfolio Value (Bearish Trades) — CIKM18

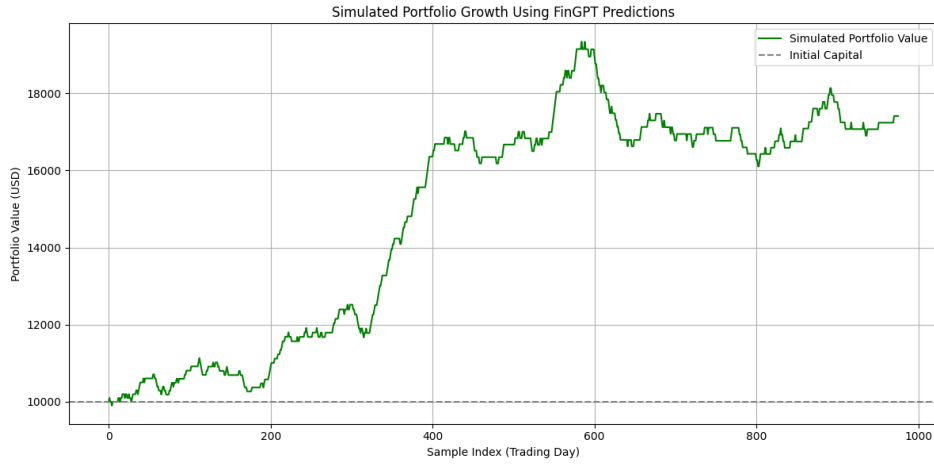


Figure 13: Portfolio Value (Bullish Trades) — CIKM18

### 10.1.3 BigData22 Dataset

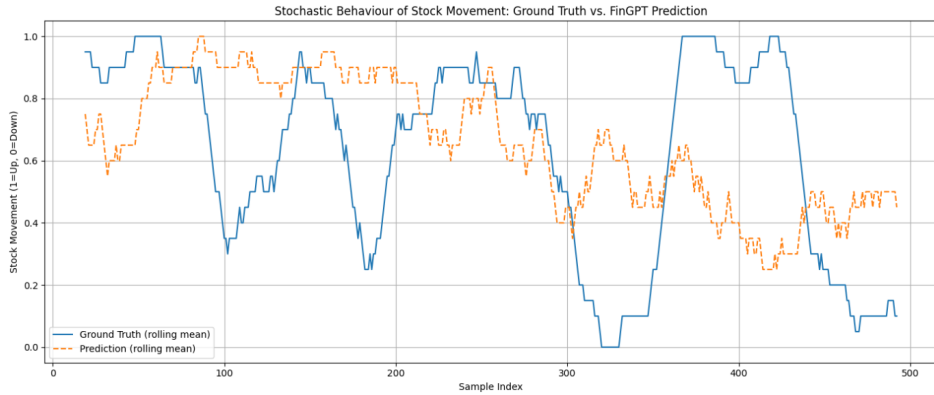


Figure 14: FinGPT vs Ground Truth (BigData22)

Figure 14 reaffirms FinGPT’s preference for bullish signal tracking. Its performance dips in response to sharp downturns but aligns more accurately during uptrends.



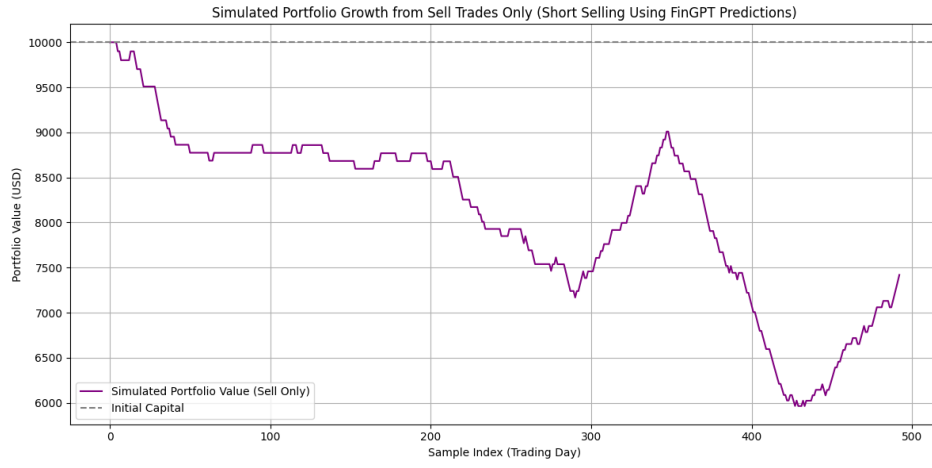


Figure 15: Portfolio Value from Bearish Trades (BigData22)

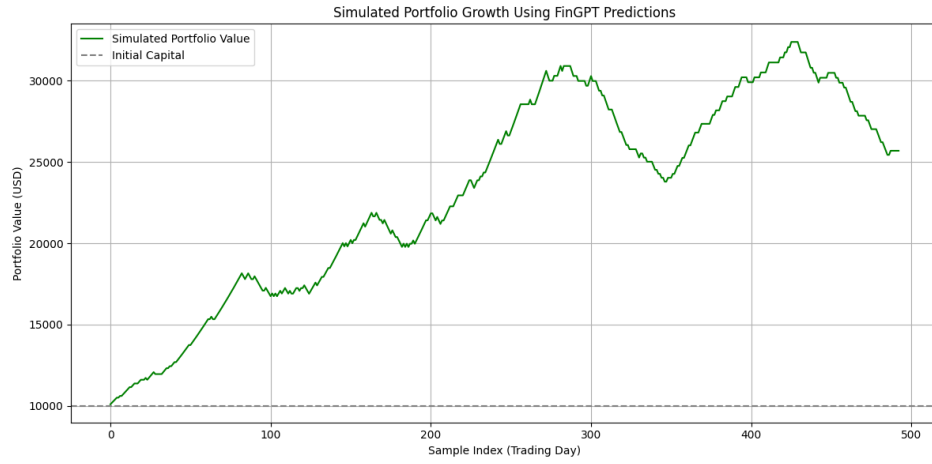


Figure 16: Portfolio Value from Bullish Trades (BigData22)

#### 10.1.4 StockNet Dataset

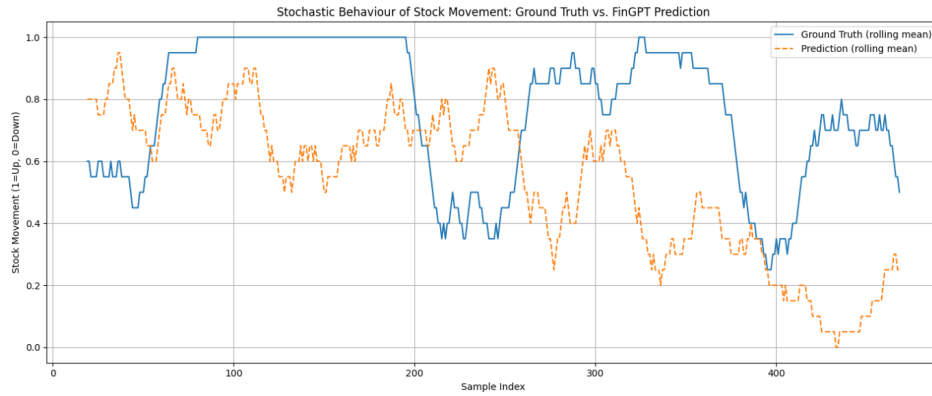


Figure 17: Smoothed Stock Movement on StockNet Dataset

**Trading Simulation:** Two strategies were tested:

- **Short-only:** Sell when FinGPT predicts market decline.
- **Long-only:** Buy when FinGPT predicts a rise.

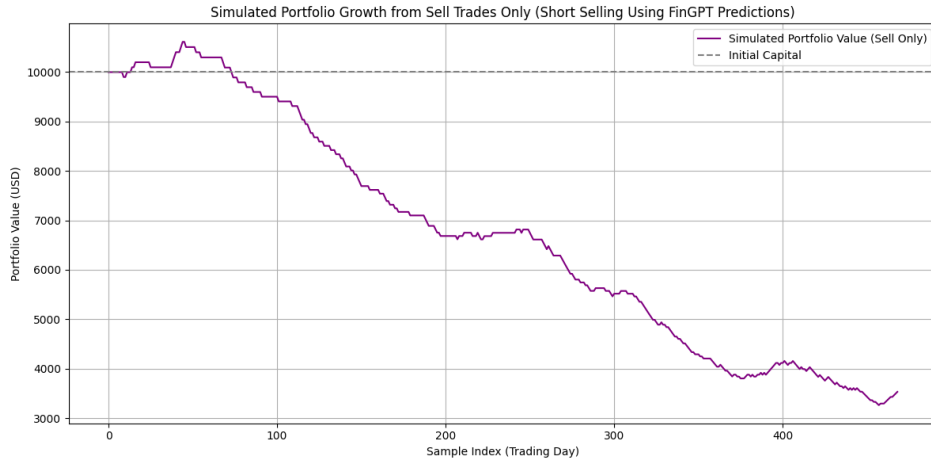


Figure 18: Short-Only Strategy Portfolio — StockNet

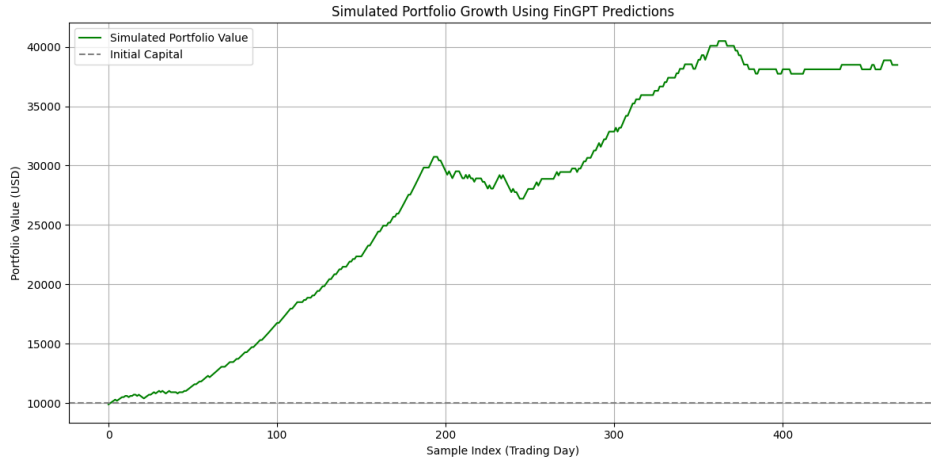


Figure 19: Long-Only Strategy Portfolio — StockNet

**Insights:** FinGPT’s predictive alignment with bullish markets leads to positive trading returns in upward trends, but its underperformance in bearish contexts highlights a key limitation. This directional imbalance is important for practical deployment in trading systems and suggests a need for improved calibration or training on diverse market conditions.

## 11 Conclusion

This research evaluated FinGPT performance on six key financial NLP tasks. The model showed strong results in classification tasks like sentiment analysis and headline classification, often matching or exceeding general models. However, it struggled with reasoning-heavy tasks like question answering and summarization. Benchmarking against GPT-4, FinMA 7B, and human performance provided valuable context. While FinGPT is promising for finance-specific NLP, it is not yet a full substitute for more advanced or general-purpose models. The findings underscore the importance of model architecture and computational resources in achieving reliable performance. This work lays a solid foundation for future improvements in domain-specific language models and highlights FinGPT’s potential in targeted financial applications.

## 12 Key Findings

### 12.1 Financial Question Answering (QA)

FinGPT demonstrated significant limitations in answering multi-step, numerically grounded financial questions, particularly when evaluated on datasets such as ConvFinQA and FLARE-FinQA. The Exact Match (EM) scores (3.8%–28.4%) fell well below human-level (89–91%) and GPT-4 (69–76%) benchmarks. These results indicate that

FinGPT, while domain-aligned, lacks the deep reasoning and arithmetic capabilities required to handle quantitative financial queries. This suggests a key limitation in FinGPT’s autoregressive architecture and its training setup. Unlike general-purpose models augmented with retrieval or external tools, FinGPT operates with limited reasoning depth, revealing a gap between task-specific expectations and model capacity. Future directions may involve augmenting FinGPT with symbolic reasoning, external calculators, or chain-of-thought prompting strategies.

## 12.2 Stock Movement Prediction (SMP)

The SMP task remains inherently challenging due to the stochastic nature of financial markets. In our experiments, FinGPT achieved modest accuracy scores across the CIKM18, StockNet, and BigData22 datasets (ranging from 47.0% to 52.8%). These results place it slightly behind FinMA 7B and GPT-4, but still ahead of traditional baselines in some configurations. What sets this task apart is the complexity of associating unstructured text with structured stock price movements—a task highly sensitive to both news content and market context. Despite lacking prior strong baselines, FinGPT offers a credible first benchmark, particularly when extended through our new directional analysis framework.

To expand the interpretability of FinGPT’s predictions beyond standard classification accuracy, we introduced a novel directional sensitivity analysis. This bi-directional performance evaluation measures how well the model detects and reacts to bullish versus bearish signals over time.

Across all three datasets, FinGPT exhibited a consistent bias toward bullish (upward) trends. While this bias enabled strong predictive alignment in rising markets, it also caused performance degradation during market downturns. Simulated trading portfolios built on FinGPT’s buy/sell signals revealed clear asymmetries: long-only strategies (buying on bullish predictions) consistently outperformed short-only strategies. This analysis offers two major insights:

1. **Model behavior is asymmetrical:** FinGPT is more attuned to upward sentiment, possibly due to training data imbalances or architectural bias in autoregressive decoding.
2. **Implications for real-world deployment:** In practice, FinGPT may be more reliable during growth phases than recessions. This insight is critical for designing trading strategies or decision-support systems that integrate model predictions.

Together, this bidirectional analysis introduces a valuable evaluation perspective for financial forecasting models, moving beyond flat metrics to behavioral insights. It also sets a precedent for future FinLLMs to incorporate risk-aware performance evaluations.

## References

- [1] Haisal Dauda Abubakar, Mahmood Umar, and Muhammad Abdullahi Bakale. Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec. *SLU Journal of Science and Technology*, 4(1):27–33, 2022.
- [2] Moses Alabi. Ai in financial services: Fraud detection, algorithmic trading, and risk assessment. 2022.
- [3] Ceray Aldemir and Tuğba Uçma Uysal. Ai competencies for internal auditors in the public sector. *Edpacs*, 69(1):3–21, 2024.
- [4] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. arxiv 2019. *arXiv preprint arXiv:1908.10063*, 2019.
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [6] ChanceFocus. Flare-fiqa-sa: Financial sentiment dataset. <https://huggingface.co/datasets/ChanceFocus/flare-fiqa-sa>, 2023. Accessed: 2025-06-10.
- [7] ChanceFocus. Cikm18 stock dataset. <https://huggingface.co/datasets/ChanceFocus/flare-sm-acl>, 2024. Accessed: 2025-06-10.
- [8] ChanceFocus. Flare-ectsum: Financial summarization dataset. <https://huggingface.co/datasets/ChanceFocus/flare-ectsum>, 2024. Accessed: 2025-06-10.
- [9] ChanceFocus. Flare-finqa: Financial qa dataset. <https://huggingface.co/datasets/ChanceFocus/flare-finqa>, 2024. Accessed: 2025-06-10.
- [10] ChanceFocus. Stocknet dataset. <https://huggingface.co/datasets/ChanceFocus/flare-sm-stocknet>, 2024. Accessed: 2025-06-10.
- [11] Dr Kostis Chlouverakis. How artificial intelligence is reshaping the financial services industry, 2024.
- [12] Tamanna Abdul Rahman Dalwai, Araby Madbouly, and Syeeda Shafiya Mohammadi. An investigation of artificial intelligence application in auditing. In *Artificial intelligence and COVID effect on accounting*, pages 101–114. Springer, 2022.
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
- [14] Vineet Dhanawat, Varun Shinde, Vishal Karande, and Kartik Singhal. Enhancing financial risk management with federated ai. In *2024 8th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI)*, pages 1–6. IEEE, 2024.
- [15] FinGPT. Convfinqa dataset. <https://huggingface.co/datasets/FinGPT/fingpt-convfinqa>, 2024. Accessed: 2025-06-10.
- [16] FinGPT. Fingpt-headline dataset. <https://huggingface.co/datasets/FinGPT/fingpt-headline>, 2024. Accessed: 2025-06-10.
- [17] FinGPT. Fingpt-ner dataset. <https://huggingface.co/datasets/FinGPT/fingpt-ner>, 2024. Accessed: 2025-06-10.
- [18] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
- [19] Yue Guo, Zian Xu, and Yi Yang. Is chatgpt a financial expert? evaluating language models on financial natural language processing. *arXiv preprint arXiv:2310.12664*, 2023.
- [20] Allen H Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023.

- [21] U Iwuanyanwu, AJ Apeh, OR Adaramodu, EC Okeleke, and OG Fakeyede. Analyzing the role of artificial intelligence in it audit: current practices and future prospects. *Computer Science & IT Research Journal*, 4(2):54–68, 2023.
- [22] David Kuo Chuen Lee, Chong Guan, Yinghui Yu, and Qinxu Ding. A comprehensive review of generative ai in finance. *FinTech*, 3(3):460–478, 2024.
- [23] Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. A survey of large language models in finance (finllms). *arXiv preprint arXiv:2402.02315*, 2024.
- [24] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- [25] Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? a study on several typical tasks. *arXiv preprint arXiv:2305.05862*, 2023.
- [26] Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. Fingpt: Democratizing internet-scale data for financial large language models. *arXiv preprint arXiv:2307.10485*, 2023.
- [27] John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. A proposal for the dartmouth summer research project on artificial intelligence. *WIRED*, 2012. Accessed: 2025-04-16.
- [28] Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. Deep learning for financial applications: A survey. *Applied soft computing*, 93:106384, 2020.
- [29] Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Han Yi, Jimin Huang, Qianqian Xie, and Jianyun Nie. Fino1: On the transferability of reasoning enhanced llms to finance. *arXiv preprint arXiv:2502.08127*, 2025.
- [30] Marko Ranković, Elena Gurgu, Oliva Martins, and Milan Vukasović. Artificial intelligence and the evolution of finance: opportunities, challenges and ethical considerations. *EdTech Journal*, 3(1):20–23, 2023.
- [31] Matteo Rizzato, Julien Wallart, Christophe Geissler, Nicolas Morizet, and Noureddine Boumlaik. Generative adversarial networks applied to synthetic financial scenarios generation. *Physica A: Statistical Mechanics and its Applications*, 623:128899, 2023.
- [32] Mr Ghiath Shabsigh and El Bachir Boukherouaa. *Generative artificial intelligence in finance: Risk considerations*. International Monetary Fund, 2023.
- [33] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- [34] TheFinAI. Bigdata22 dataset. <https://huggingface.co/datasets/TheFinAI/flare-sm-bigdata>, 2023. Accessed: 2025-06-10.
- [35] TheFinAI. Flare-fpb: Financial phrase bank dataset. <https://huggingface.co/datasets/TheFinAI/flare-fpb>, 2023. Accessed: 2025-06-10.
- [36] TheFinAI. Flare-fpb: Financial language analysis for real-world events - financial phrasebank. <https://huggingface.co/datasets/TheFinAI/flare-fpb>, 2024. Accessed: 2025-05-18.
- [37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [39] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kam-badur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.
- [40] 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 arXiv:2306.05443*, 2023.

- [41] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023.
- [42] 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.
- [43] Gokul Yenduri, M Ramalingam, G Chemmalar Selvi, Y Supriya, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, G Deepti Raj, Rutvij H Jhaveri, B Prabadevi, Weizheng Wang, et al. Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*, 2024.