A Survey of Large Language Models in Finance (FinLLMs)

Jean Lee[1] , **Nicholas Stevens**[2] , **Soyeon Caren Han**[1,3] , **Minseok Song**[4]

[1]The University of Sydney, Australia, [2]NRS Technology, Australia
[3]The University of Melbourne, Australia
[4]Pohang University of Science and Technology (POSTECH), South Korea

{jean.lee, caren.han}@sydney.edu.au, nick@nrs.technology, mssong@postech.ac.kr

Abstract

Large Language Models (LLMs) have shown remarkable capabilities across a wide variety of Natural Language Processing (NLP) tasks and have attracted attention from multiple domains, including financial services. Despite the extensive research into general-domain LLMs, and their immense potential in finance, Financial LLM (FinLLM) research remains limited. This survey provides a comprehensive overview of FinLLMs, including their history, techniques, performance, and opportunities and challenges. Firstly, we present a chronological overview of general-domain Pretrained Language Models (PLMs) through to current FinLLMs, including the GPT-series, selected open-source LLMs, and financial LMs. Secondly, we compare five techniques used across financial PLMs and FinLLMs, including training methods, training data, and fine-tuning methods. Thirdly, we summarize the performance evaluations of six benchmark tasks and datasets. In addition, we provide eight advanced financial NLP tasks and datasets for developing more sophisticated FinLLMs. Finally, we discuss the opportunities and the challenges facing FinLLMs, such as hallucination, privacy, and efficiency. To support AI research in finance, we compile a collection of accessible datasets and evaluation benchmarks on GitHub. [1]

1 Introduction

Research on Large Language Models (LLMs) has grown rapidly in both academia and industry, with notable attention to LLM applications such as ChatGPT. Inspired by Pre-trained Language Models (PLMs) [Devlin *et al.*, 2018; Radford *et al.*, 2018], LLMs are empowered by transfer learning and built upon the Transformer architecture [Vaswani *et al.*, 2017] using large-scale textual corpora. Researchers have discovered that scaling models [Radford *et al.*, 2019] to sufficient sizes not only enhances model capacity but also enables surprising emergent properties, such as in-context learning

[Brown *et al.*, 2020], that do not appear in small-scale language models. Language Models (LMs) can be categorized based on parameter size, and the research community has created the term "Large Language Models (LLM)" for PLMs of substantial size, typically exceeding 7 billion parameters [Zhao *et al.*, 2023]. The technical evolution of LLMs has resulted in a remarkable level of homogenization [Bommasani *et al.*, 2021], meaning that a single model could yield strong performance across a wide range of tasks. The capability of LLMs has facilitated the adaptation of various forms of multimodal data (e.g., text, image, audio, video, and tabular data) and multimodal models across AI and interdisciplinary research communities.

In the financial domain, there has been growing interest in applying NLP across various financial tasks, including sentiment analysis, question answering, and stock market prediction. The rapid advancement of general-domain LLMs has spurred an investigation into Financial LLMs (FinLLMs), employing methods such as mixed-domain LLMs with prompt engineering and instruction fine-tuned LLMs with prompt engineering. While general LLMs are extensively researched and reviewed [Zhao *et al.*, 2023; Yang *et al.*, 2023b; Zhang *et al.*, 2023], the field of financial LLMs [Li *et al.*, 2023b] is at an early stage. Considering the immense potential of LLMs in finance, this survey provides a holistic overview of FinLLMs and discusses future directions that can stimulate interdisciplinary studies. We acknowledge that this research focuses on LMs in English. The key contributions of this survey paper are summarized below.

- To the best of our knowledge, this is the first comprehensive survey of FinLLMs that explores the evolution from general-domain LMs to financial-domain LMs.

- We compare five techniques used across four financial PLMs and four financial LLMs, including training methods and data, and instruction fine-tuning methods.

- We summarize the performance evaluation of six benchmark tasks and datasets between different models, and provide eight advanced financial NLP tasks and datasets for the development of advanced FinLLMs.

- We discuss the opportunities and the challenges of FinLLMs, with regard to datasets, techniques, evaluation, implementation, and real-world applications.
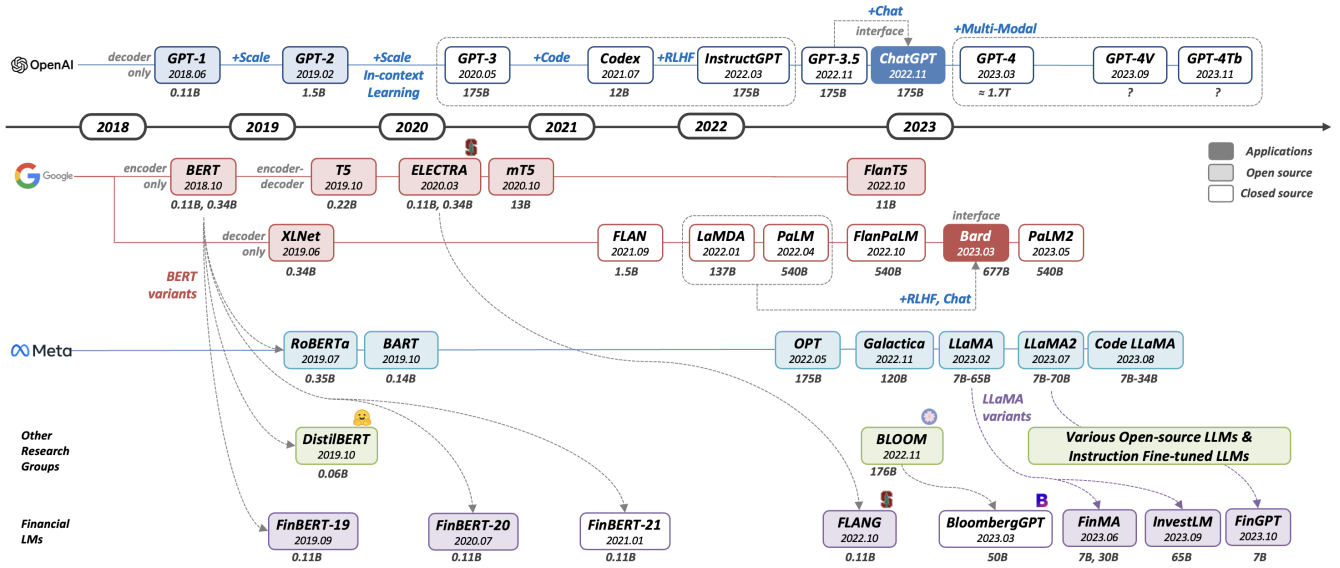
---

[1]https://github.com/adlnlp/FinLLMs

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

2 Evolution Trends: from General to Finance

2.1 General-domain LMs

Since the introduction of the Transformer [Vaswani *et al.*, 2017] architecture by Google in 2017, Language Models (LMs) are generally pre-trained with either **discriminative** or **generative** objectives. Discriminative pre-training uses a masked language model to predict the next sentence and features an encoder-only or an encoder-decoder architecture. Generative pre-training uses autoregressive language modeling to predict the next token and features a decoder-only architecture. Figure 1 illustrates the evolutionary timeline from general-domain LMs to financial-domain LMs.

GPT-Series

The Generative Pre-trained Transformer (GPT) series of models started with **GPT-1** (110M) [Radford *et al.*, 2018]. Since then, the OpenAI team focused on scaling the model, and **GPT-2** (1.5B) [Radford *et al.*, 2019] was released in 2019. GPT-2 identified the power of scaling and a probabilistic approach for multi-task problem-solving. In 2020, **GPT-3** with 175B parameters was released [Brown *et al.*, 2020]. This was a significant milestone for LLMs, as it introduced an emergent capability of LLMs; in-context learning. **In-context learning** refers to the model acquiring capabilities that were not explicitly trained, allowing language models to understand human language and produce outcomes beyond their original pre-training objectives.

Ongoing efforts to improve LLMs have resulted in the introduction of **ChatGPT**, in November 2022. This application combines GPT-3 (In-context learning), Codex (LLMs for code), and InstructGPT (Reinforcement Learning with Human Feedback, RLHF). The success of ChatGPT has led to further development of significantly larger models, including **GPT-4** (estimated 1.7T parameters). GPT-4 demonstrates

human-level performance, capable of passing law and medical exams, and handling multimodal data.

OpenAI continues to build extremely large language models, aiming to enhance the model's capabilities in handling multimodal data, as well as providing APIs for the development of real-world applications. Despite the mainstream popularity and adoption, real-world applications in finance utilizing their APIs have not yet been fully explored.

Open-source LLMs

Prior to the era of LLMs, the research community often released open-source PLMs such as Bidirectional Encoder Representations from Transformers (**BERT**, base-110M parameters) [Devlin *et al.*, 2018]. BERT is the foundational model for many early PLMs, including FinBERT. Since OpenAI shifted from open-source to closed-source LLMs, the trend across LLM research is a reduction in the release of open-source models. However, in February 2023, Meta AI released the open-source LLM, **LLaMA** (7B, 13B, 33B, 65B parameters) [Touvron *et al.*, 2023], and this encouraged the development of diverse LLMs using LLaMA. Similar to BERT variants, LLaMA variants quickly proliferated by adopting various techniques such as Instruction Fine-Tuning (IFT) [Zhang *et al.*, 2023] and Chain-of-Thought (CoT) Prompting [Wei *et al.*, 2022].

There have also been significant efforts by the research community to generate open-source LLMs to reduce the reliance on corporate research and proprietary models. **BLOOM** (176B) [Scao *et al.*, 2022] was built by a collaboration of hundreds of researchers from the BigScience Workshop. This open-source LLM was trained on 46 natural languages and 13 programming languages.

2.2 Financial-domain LMs

Domain-specific LMs, such as financial-domain LMs, are commonly built using general-domain LMs. In finance, there
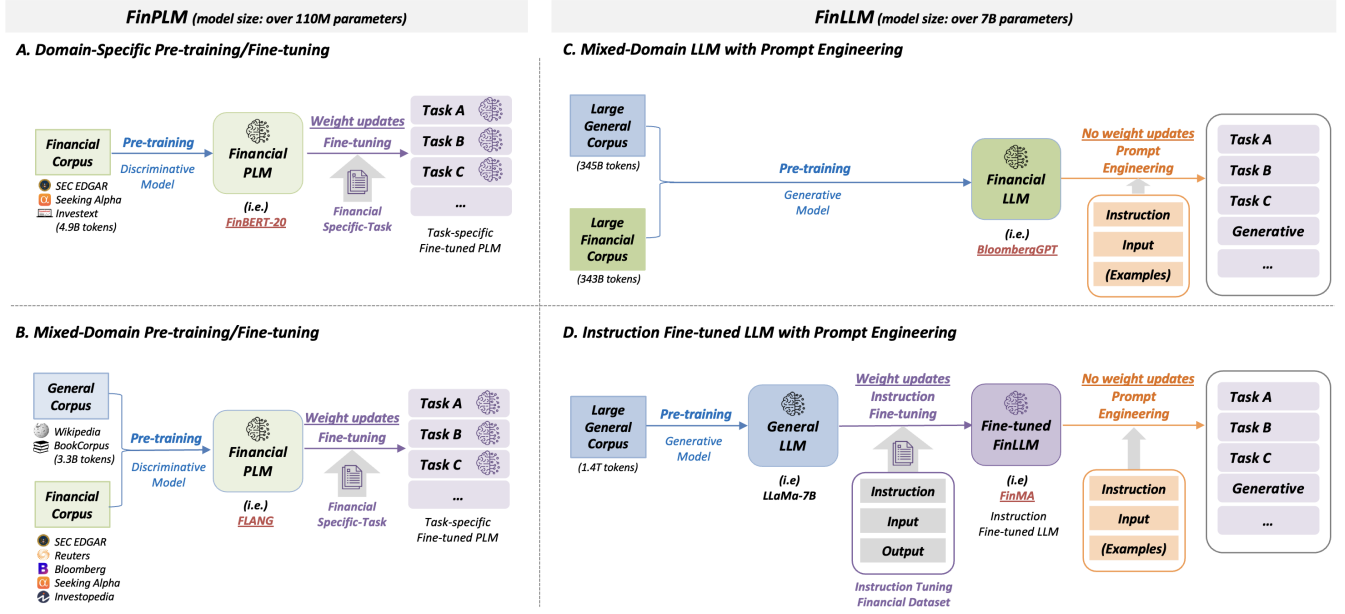
Figure 2: Comparison of techniques used in financial LMs: from FinPLMs to FinLLMs.

are primarily four financial PLMs (FinPLMs) and four financial LLMs (FinLLMs). Within the four FinPLMs, FinBERT-19 [Araci, 2019], FinBERT-20 [Yang *et al.*, 2020], and FinBERT-21 [Liu *et al.*, 2021] are all based on BERT, while FLANG [Shah *et al.*, 2022] is based on ELECTRA [Clark *et al.*, 2020]. Within the four FinLLMs, FinMA [Xie *et al.*, 2023], InvestLM [Yang *et al.*, 2023c], and FinGPT [Wang *et al.*, 2023] are based on LLaMA or other open-source-based models, while BloombergGPT [Wu *et al.*, 2023] is a BLOOM-style closed-source model.

3 Techniques: from FinPLMs to FinLLMs

While our survey focuses on FinLLMs, it is important to acknowledge that previous studies on FinPLMs as they formed the groundwork for FinLLM development. We reviewed three techniques used by the four FinPLMs and two techniques used by the four FinLLMs. Figure 2 illustrates technical comparisons of building financial LMs [2], and Table 1 shows a summary of FinPLMs/FinLLMs including pre-training techniques, fine-tuning, and evaluation.

3.1 Continual Pre-training

Continual pre-training of LMs aims to train an existing general LM with new domain-specific data on an incremental sequence of tasks [Ke *et al.*, 2022].

FinBERT-19 [Araci, 2019] [3] is the first FinBERT model released for financial sentiment analysis and implements three steps: 1) the initialization of the general-domain BERT PLM (3.3B tokens), 2) continual pre-training on a financial-domain corpus, and 3) fine-tuning on financial domain-specific NLP tasks. The fine-tuned financial LM is released

on HuggingFace, and this FinBERT-19 is a task-dependent model for the financial sentiment analysis task.

3.2 Domain-Specific Pre-training from Scratch

The domain-specific pre-training approach involves training a model exclusively on an unlabeled domain-specific corpus while following the original architecture and its training objective.

FinBERT-20 [Yang *et al.*, 2020] [4] is a finance domain-specific BERT model, pre-trained on a financial communication corpus (4.9B tokens). The author released not only the FinBERT model but also FinVocab uncased/cased, which has a similar token size to the original BERT model. FinBERT-20 also conducted a sentiment analysis task for fine-tuning experiments on the same dataset of FinBERT-19.

3.3 Mixed-Domain Pre-training

The mixed-domain pre-training approach involves training a model using both a general-domain corpus and a domain-specific corpus. The assumption is that general-domain text remains relevant, while the financial domain data provides knowledge and adaptation during the pre-training process.

FinBERT-21 [Liu *et al.*, 2021] [5] is another BERT-based PLM designed for financial text mining, trained simultaneously on a general corpus and a financial domain corpus. FinBERT-21 employs multi-task learning across six self-supervised pre-training tasks, enabling it to efficiently capture language knowledge and semantic information. FinBERT-21 conducted experiments on Sentiment Analysis as well as provided experiment results for two additional tasks; Sentence Boundary Detection and Question Answering.

---

[2]The continual pre-training diagram can be found on our GitHub.
[3]https://huggingface.co/ProsusAI/finbert

[4]https://github.com/yya518/FinBERT
[5]As of Dec 2023, the link mentioned in the paper does not exist.

Category	Model	Backbone	Paras.	Techniques	PT PT Data Size	Evaluation Task	Dataset	Open Source 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 *et al.*, 2020]	BERT	0.11B	PT, FT	(F) 4.9B tokens	[SA]	FPB, FiQA-SA, AnalystTone	Y	Y	N	ArXiv Jul 2020
	FinBERT-21 [Liu *et al.*, 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 *et al.*, 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 *et al.*, 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 *et al.*, 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 *et al.*, 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 *et al.*, 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

Table 1: A Summary of FinPLMs and FinLLMs. The abbreviations correspond to Paras.= Model Parameter Size (Billions); Disc. = Discriminative, Gen. = Generative; Post-PT = Post-Pre-training, PT = Pre-training, FT = Fine-Tuning, PE = Prompt Engineering, IFT = Instruction Fine-Tuning, PEFT = Parameter Efficient Fine-Tuning; (G) = General domain, (F) = Financial domain; (in Evaluation) [SA] Sentiment Analysis, [TC] Text Classification, [SBD] Structure Boundary Detection, [NER] Named Entity Recognition, [QA] Question Answering, [SMP] Stock Movement Prediction, [Summ] Text Summarization, [RE] Relation Extraction; (in Venue) (S) = Special Track, (D) = Datasets and Benchmarks Track, (W) = Workshop. In open source, it is marked as Y if it is publicly accessible as of Dec 2023.

FLANG [Shah *et al.*, 2022] [6] is a domain-specific model using financial keywords and phrases for masking, and follows the training strategy of ELECTRA [Clark *et al.*, 2020]. This research first introduces Financial Language Understanding Evaluation (**FLUE**), a collection of five financial NLP benchmark tasks. The tasks include Sentiment Analysis, Headline Text Classification, Named Entity Recognition, Structure Boundary Detection, and Question Answering.

3.4 Mixed-Domain LLM with Prompt Engineering

Mixed-domain LLMs are trained on both a large general corpus and a large domain-specific corpus. Then, users describe the task and optionally provide a set of examples in human language. This technique is called Prompt Engineering and uses the same frozen LLM for several downstream tasks with no weight updates. This survey does not explore prompt engineering but instead references recent surveys [Liu *et al.*, 2023].

BloombergGPT [Wu *et al.*, 2023] [7] is the first FinLLM that utilizes the BLOOM model [Scao *et al.*, 2022]. It is trained on a large general corpus (345B tokens), and a large financial corpus (363B tokens). The financial corpus, FinPile, contains data collected from the web, news, filings, press, and Bloomberg's proprietary data. The authors conducted financial NLP tasks (5 benchmark tasks and 12 internal tasks) as well as 42 general-purpose NLP tasks.

3.5 Instruction Fine-tuned LLM with Prompt Engineering

Instruction tuning is the additional training of LLMs using explicit text instructions to enhance the capabilities and controllability of LLMs. Research on instruction tuning can be classified into two main areas [Zhang *et al.*, 2023]: 1) the construction of instruction datasets, and 2) the generation of fine-tuned LLMs using these instruction datasets. In finance, researchers have started transforming existing financial datasets into instruction datasets and subsequently using these datasets for fine-tuning LLMs.

FinMA (or PIXIU) [Xie *et al.*, 2023] [8] consists of two fine-tuned LLaMA models (7B and 30B) [Touvron *et al.*, 2023] that use financial instruction datasets for financial tasks. It is constructed from a large-scale multi-task instruction dataset called Financial Instruction Tuning (FIT, 136k samples) by collecting nine publicly released financial datasets used across five different tasks. In addition to the five FLUE benchmark tasks, it includes the Stock Movement Prediction task.

InvestLM [Yang *et al.*, 2023c] [9] is a fine-tuned LLaMA-65B model using a manually curated financial domain instruction dataset. The dataset includes Chartered Financial Analyst (CFA) exam questions, SEC filings, Stackexchange quantitative finance discussions, and Financial NLP tasks. The downstream tasks are similar to FinMA but also include a financial text Summarization task.

FinGPT [Yang *et al.*, 2023a] [10] is an open-sourced and data-centric framework, which provides a suite of APIs for financial data sources, an instruction dataset for financial tasks, and several fine-tuned financial LLMs. The FinGPT team has released several similar papers that describe the framework and an experiment paper [Wang *et al.*, 2023] on the instruction fine-tuned FinLLMs using six open-source LLMs with the Low-Rank Adaptation (LoRA) [Hu *et al.*, 2021] method.

4 Evaluation: Benchmark Tasks and Datasets

As LLMs gain significant attention, evaluating them becomes increasingly critical. We summarize six financial NLP benchmark tasks and datasets, and review the evaluation results of models including FinPLMs, FinLLMs, ChatGPT, GPT-4, and task-specific State-of-the-Art (SOTA) models. The results [11] are referenced from original research or analysis research [Li *et al.*, 2023a], and SOTA results from task-specific models.

4.1 Sentiment Analysis (SA)

The Sentiment Analysis (SA) task aims to analyze sentiment information from input text, including financial news and microblog posts. Most FinPLMs and FinLLMs report the evaluation results of this task using the Financial PhraseBank (FPB) and the FiQA SA dataset. The **FPB** [Malo *et al.*, 2014] dataset consists of 4,845 English financial news articles. Domain experts annotated each sentence with one of three sentiment labels: Positive, Negative, or Neutral. The **FiQA-SA** [Maia *et al.*, 2018] dataset consists of 1,173 posts from both headlines and microblogs. The sentiment scores are on a scale of [-1, 1], and recent studies have converted this score into a classification task. Overall, FLANG-ELECTRA achieved the best results (92% on F1) while FinMA-30B and GPT-4 achieved similar results (87% on F1) with a 5-shot prompting. It suggests a practical approach for less complex tasks in terms of efficiency and costs.

For further evaluation of SA, we include two open-released datasets: SemEval-2017 (Task 5) and StockEmotions. The **SemEval-2017** [Cortis *et al.*, 2017] dataset comprises 4,157 sentences collected from both headlines and microblogs. Similar to FiQA SA, the sentiment scores are on a scale of [-1, 1]. The **StockEmotions** [Lee *et al.*, 2023] dataset consists of 10,000 sentences collected microblogs that annotate binary sentiment and 12 fine-grained emotion classes that span the multi-dimensional range of investor emotions.

4.2 Text Classification (TC)

Text Classification (TC) is the task of classifying a given text or document into predefined labels based on its content. In financial text, there are often multiple dimensions of information beyond sentiment such as price directions or interest rate directions. FLUE includes the gold news **Headline** [Sinha and Khandait, 2021] dataset for text classification. This dataset comprises 11,412 news headlines, labeled with a binary classification across nine labels such as "price up", or "price down". Similar to the SA task, FLANG-ELECTRA and FinMA-30B with a 5-shot prompting achieved the best

results (98% on Avg. F1) and the performance of BERT and FinBERT-20 was also noteworthy (97% on Avg. F1).

As TC is a broad task depending on the dataset and its predefined labels, we include three open-released financial TC datasets for further research: FedNLP, FOMC, and Banking77. The **FedNLP** [Lee *et al.*, 2021] dataset comprises documents sourced from various Federal Open Market Committee (FOMC) materials. The dataset is annotated with labels as Up, Maintain, or Down based on the Federal Reserve's Federal Funds Rate decision for the subsequent period. Similarly, the **FOMC** [Shah *et al.*, 2023] dataset is a collection of FOMC documents with the labels as Dovish, Hawkish, or Neutral, reflecting the prevailing sentiment conveyed within the FOMC materials. The **Banking77** [Casanueva *et al.*, 2020] dataset comprises 13,083 samples covering 77 intents related to banking customer service queries, such as "card loss" or "linking to an existing card". This dataset is designed for intent detection and developing conversation systems.

4.3 Named Entity Recognition (NER)

The Named Entity Recognition (NER) task is the extraction of information from unstructured text and categorizing it into predefined named entities such as locations (LOC), organizations (ORG), and persons (PER). For the financial NER task, the **FIN** dataset [Alvarado *et al.*, 2015] is included in FLUE benchmarks. The FIN dataset comprises eight financial loan agreements sourced from the US Security and Exchange Commission (SEC) for credit risk assessment. GPT-4 with a 5-shot prompting (83% on Entity F1) and FLANG-ELECTRA demonstrate notable performance (82% on Entity F1), while other FinLLMs exhibit suboptimal results (61%-69% on Entity F1).

For further research, we include a financial NER dataset, **FiNER-139** [Loukas *et al.*, 2022], consisting of 1.1M sentences annotated with 139 eXtensive Business Reporting Language (XBRL) word-level tags, sourced from the SEC. This dataset is designed for Entity Extraction and Numerical Reasoning tasks, predicting the XBRL tags (e.g., cash and cash equivalents) based on numeric input data within sentences (e.g., "24.8" million).

4.4 Question Answering (QA)

Question Answering (QA) is a task to retrieve or generate answers to questions from an unstructured collection of documents. Financial QA is more challenging than general QA as it requires numerical reasoning across multiple formats. **FiQA-QA** [Maia *et al.*, 2018] is for opinion-based QA, representing an early Financial QA dataset.

Over time, Financial QA has evolved to include complex numerical reasoning in multi-turn conversations. This evolution involves the introduction of hybrid QA, which is to create paths to connect hybrid contexts including both tabular and textual content. **FinQA** [Chen *et al.*, 2021] is a single-turn hybrid QA dataset having 8,281 QA pairs and annotated by experts from the annual reports of S&P 500 companies. **ConvFinQA** [Chen *et al.*, 2022], an extension of FinQA, is a multi-turn conversational hybrid QA dataset, consisting of 3,892 conversations with 14,115 questions. Instead of using the FiQA-QA dataset, all FinLLMs conducted experi-

---

[10]https://github.com/AI4Finance-Foundation/FinGPT

[11]Figure and tables of each task can be found on our Github.

ments on the FinQA and/or ConvFinQA datasets to assess their numerical reasoning capabilities. GPT-4 with a zero-shot prompting outperforms all other models (69%-76% on EM Accuracy), approaching the performance of human experts (Avg. 90% on EM Accuracy). BloombergGPT's result (43% on EM Accuracy) was slightly below the general crowd (47% on EM Accuracy).

4.5 Stock Movement Prediction (SMP)

The Stock Movement Prediction (SMP) task aims to predict the next day's price movement (e.g., up or down) based on historical prices and associated text data. As it requires the integration of time series problems with temporal dependencies from text information, it presents a complex task, where text data can act both as noise and signal. FinMA includes the SMP tasks for the first time, conducting experiments on three datasets; StockNet, CIKM18, and BigData22.

StockNet [Xu and Cohen, 2018] collected historical price data and Twitter data between 2014 and 2016 for 88 stocks listed in the S&P, and is widely used for SMP tasks. The task is framed as a binary classification with a threshold: a price movement higher than 0.55% is labeled as a rise (denoted as 1), while a movement less than -0.5% is labeled as a fall (denoted as 0). Similarly, **CIKM18** [Wu *et al.*, 2018] utilizes historical price and Twitter data in 2017 for 47 stocks in the S&P 500. **BigData22** [Soun *et al.*, 2022] compiled data between 2019 and 2020 for 50 stocks in the US stock markets. Like StockNet, it adopts a binary classification formulation with a threshold. On average across these three datasets, GPT-4 with a zero-shot prompting achieves higher performance (54% on Accuracy) than FinMA (52% on Accuracy) and slightly lower results than the SOTA model (58% on Accuracy). Although NLP metrics such as Accuracy are commonly used, these are insufficient for the SMP evaluation. It is important to consider financial evaluation metrics, such as the Sharpe ratio, as well as backtesting simulation results.

4.6 Text Summarization (Summ)

Summarization (Summ) is the generation of a concise summary from documents while conveying its key information via either an extractive or an abstractive approach. In finance, it has been relatively underexplored due to the lack of benchmark datasets, challenges with domain experts' evaluations, and the need for disclaimers when presenting financial advice. InvestLM includes summarization tasks for the first time, conducting experiments on the ECTSum dataset. **ECTSum** [Mukherjee *et al.*, 2022] consists of 2,425 document-summary pairs, containing Earnings Call Transcripts (ECTs) and bullet-point summarizations from Reuters. It reports evaluation results on various metrics, including ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore. Similar to other complex financial tasks, the task-specific SOTA model (47% on ROUGE-1) outperforms all LLMs. According to the authors of InvestLM, while GPT-4 with a zero-shot prompting (30% on ROUGE-1) shows superior performance compared to FinLLMs, the commercial models generate decisive answers.

The summarization task offers significant development opportunities, exploring whether FinLLMs can outperform task-specific SOTA models. For ongoing research, we include the financial summarization dataset, **MultiLing 2019** [El-Haj, 2019], containing 3,863 document-summary pairs extracted from UK annual reports listed on the London Stock Exchange (LSE). It provides at least two gold-standard summaries for each annual report.

4.7 Discussion

Within the six benchmarks, the performance of mixed-domain FinPLMs is noteworthy for the SA, TC, and NER tasks, suggesting that using a PLM with fine-tuning for a specific task can be a practical approach depending on the task complexity. For QA, SMP, and Summ tasks, the task-specific SOTA models outperform all LLMs, indicating areas for improvement in FinLLMs. Notably, GPT-4 shows impressive performance across all benchmarks except the Summ task, indicating that scaling models alone may not be adequate for optimal performance in finance. As most instruction-finetuned FinLLMs used the same datasets for their evaluation, we include additional datasets for future research.

5 Advanced Financial NLP Tasks and Datasets

Properly designed benchmark tasks and datasets are a crucial resource to assess the capability of LLMs, however, the current 6 benchmark tasks have yet to address more complex financial NLP tasks. In this section, we present 8 advanced benchmark tasks and compile associated datasets for each.

The **Relation Extraction** (RE) task aims to identify and classify relationships between entities implied in text. Similar to NER, this task is part of Information Extraction. The **FinRED** [Sharma *et al.*, 2022] dataset is released for RE and is curated from financial news and earning call transcripts, containing 29 relation tags (e.g. owned by) specific to the finance domain.

Event Detection (ED) in finance involves identifying the impact of how investors perceive and assess related companies. The **Event-Driven Trading** (EDT) [Zhou *et al.*, 2021] dataset is released for ED and includes 11 types of corporate event detection. EDT comprises 9,721 news articles with token-level event labels, and an additional 303,893 news articles with minute-level timestamps and stock price labels.

Causality Detection (CD) in finance seeks to identify cause-and-effect relationships within factual text, aiming to develop an ability to generate meaningful financial narrative summaries. The Workshop on Financial Narrative Processing (FNP) addresses this task every year and contributes datasets. One of the open-released dataset from FNP, **FinCausal20** [Mariko *et al.*, 2020] shares two tasks: detecting a causal scheme in a given text and identifying cause-and-effect sentences.

Numerical Reasoning (NR) in finance aims to identify numbers and mathematical operators in either digit or word form, in order to perform calculations and comprehend financial context (e.g. cash and cash equivalent). Some datasets introduced for NER and QA tasks are also designed for numerical reasoning, including: **FiNER-139** [Loukas *et al.*, 2022], **FinQA** [Chen *et al.*, 2021], **ConvFinQA** [Chen *et al.*, 2022].

Structure Recognition (SR) is a task focused on Structure Boundary Detection within a document (e.g. text, tables, or figures) and recognising the logical relationships between tables and surrounding content, or between cells within a table. IBM Research has released the **FinTabNet** [Zheng *et al.*, 2021] dataset, collected from earnings reports of S&P 500 companies. This dataset comprises unstructured PDF documents with detailed annotations of table structures. The FinQA and ConvFinQA datasets, included in QA tasks, have been further developed from FinTabNet.

Multimodal (MM) understanding is a challenging task across many domains. Recently, several multimodal financial datasets have been introduced. **MAEC** [Li *et al.*, 2020] compiles the multimodal data (text, time series, and audio) from earnings call transcripts on a larger scale, with 3,443 instances and 394,277 sentences. Additionally, **MONOPOLY** [Mathur *et al.*, 2022] introduces video data from monetary policy call transcripts across six central banks, sharing 24,180 samples from 340 videos with text scripts and time series.

Machine Translation (MT) in finance aims to not only translate sentences from a source language to a target language, but to also comprehend the financial contextual meaning in different languages. **MINDS-14** [Gerz *et al.*, 2021] consists of 8,168 samples of banking voice assistant data in text and audio formats across 14 different languages. **MultiFin** [Jørgensen *et al.*, 2023] includes 10,048 samples covering financial topics with 6 high-level labels (e.g., Finance) and 23 low-level labels (e.g., M&A & Valuations), sourced from public financial articles in 15 different languages.

Market Forecasting (MF) is an essential task in financial markets, involving the prediction of market price, volatility, and risk. This task extends beyond Stock Movement Prediction (SMP), which formulates problems as a classification task. The datasets introduced in Sentiment Analysis, Event Detection, and Multimodal tasks are also designed for Market Forecasting. Here, we include a list of datasets relevant to MF: **StockEmotions** (SA) [Lee *et al.*, 2023], **EDT** (ED) [Zhou *et al.*, 2021], **MAEC** (MM-audio) [Li *et al.*, 2020], and **MONOPOLY** (MM-video) [Mathur *et al.*, 2022].

6 Opportunities and Challenges

In this section, we highlight various aspects guiding the future directions of FinLLMs, covering datasets, techniques, evaluation, implementation, and real-world applications.

Datasets : High-quality data and multimodal data are significantly important for developing sophisticated FinLLMs. As most FinLLMs train general-domain LLMs on financial-specific data, the challenge lies in collecting high-quality financial data in diverse formats. Building **instruction fine-tuned financial datasets** by converting existing datasets for specific financial NLP tasks will facilitate the development of advanced FinLLMs. Also, the research on **financial multimodal datasets** will become increasingly important, enhancing the performance of FinLLMs on complex tasks.

Techniques : Major challenges in finance include utilizing internal data without privacy breaches, causing security issues, while also enhancing trust in the responses generated by FinLLMs. To address these challenges, some actively re-searched techniques on LLMs, such as **Retrieval Augmented Generation** (RAG) [Lewis *et al.*, 2020], can be implemented in the financial domain. The RAG system is similar to an open-book approach, which retrieves non-pre-trained external knowledge resources (e.g., queried private data) to enhance the pre-trained model's raw representation of information. RAG provides the model with access to factual information, enabling the generation of cross-referenced answers, therefore improving **reliability**, and minimizing **hallucination** issues. Moreover, RAG enables the use of internal non-trainable data without retraining the entire model, ensuring privacy is not breached.

Evaluation : The primary challenge in evaluation is incorporating **domain knowledge from financial experts** to validate the model's performance based on financial NLP tasks. The current evaluation results were presented using commonly used NLP metrics such as F1-score or Accuracy. However, the knowledge-driven tasks require human evaluation by financial experts, appropriate financial evaluation metrics over NLP metrics, and expert feedback for model alignment. Furthermore, **advanced financial NLP tasks**, including the eight further benchmarks we presented, would discover the hidden capabilities of FinLLMs. These complex tasks will assess whether FinLLMs can serve as general financial problem-solver models [Guo *et al.*, 2023], considering both cost and performance for specific tasks.

Implementation : The challenge in selecting suitable FinLLMs and techniques lies in the **trade-off between cost and performance**. Depending on the task complexity and inference cost, selecting general-domain LLMs with prompting or task-specific models might be a more practical choice than building FinLLMs. This requires **LLMOps engineering** skills, including soft prompt techniques such as Parameter-Efficient Fine-Tuning (PEFT) and monitoring operation systems with a Continuous Integration (CI) and Continuous Delivery (CD) pipeline.

Applications : The challenge of developing real-world financial applications relates to non-technical issues, including business needs, industry barriers, data privacy, accountability, ethics, and the understanding gap between financial experts and AI experts. To overcome these challenges, sharing FinLLM **use-cases** will be beneficial across various financial fields including robo-advisor, quantitative trading, and low-code development [Yang *et al.*, 2023a]. Furthermore, we encourage future directions towards **generative applications** including report generation and document understanding.

7 Conclusion

Our survey provides a concise yet comprehensive investigation of FinLLMs, by exploring their evolution from general-domain LMs, comparing techniques of FinPLMs/FinLLMs, and presenting six conventional benchmarks as well as eight advanced benchmarks and datasets. For future research, our big-picture view of FinLLMs, a relevant and extensive collection of datasets for more advanced evaluation, and opportunities and challenges for new directions for advanced FinLLMs will be beneficial to both the Computer Science and Finance research communities.

References

[Alvarado *et al.*, 2015] Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of ALTA Workshop*, pages 84–90, 2015.

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

[Bommasani *et al.*, 2021] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

[Casanueva *et al.*, 2020] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of NLP4ConvAI Workshop*, pages 38–45, 2020.

[Chen *et al.*, 2021] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, et al. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of EMNLP*, pages 3697–3711, 2021.

[Chen *et al.*, 2022] Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of EMNLP*, pages 6279–6292, 2022.

[Clark *et al.*, 2020] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[Cortis *et al.*, 2017] Keith Cortis, André Freitas, Tobias Daudert, et al. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of SemEval*, pages 519–535, 2017.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2018.

[El-Haj, 2019] Mahmoud El-Haj. Multiling 2019: Financial narrative summarisation. In *Proceedings of the Workshop MultiLing 2019*, pages 6–10, 2019.

[Gerz *et al.*, 2021] Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, et al. Multilingual and cross-lingual intent detection from spoken data. In *Proceedings of EMNLP*, pages 7468–7475, 2021.

[Guo *et al.*, 2023] Yue Guo, Zian Xu, and Yi Yang. Is chatgpt a financial expert? evaluating language models on financial natural language processing. In *Findings of EMNLP*, 2023.

[Hu *et al.*, 2021] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021.

[Jørgensen *et al.*, 2023] Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. Multifin: A dataset for multilingual financial nlp. In *Findings of EACL*, pages 864–879, 2023.

[Ke *et al.*, 2022] Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. In *ICLR*, 2022.

[Lee *et al.*, 2021] Jean Lee, Hoyoul Luis Youn, Nicholas Stevens, Josiah Poon, and Soyeon Caren Han. Fednlp: an interpretable nlp system to decode federal reserve communications. In *Proceedings of ACM SIGIR*, pages 2560–2564, 2021.

[Lee *et al.*, 2023] Jean Lee, Hoyoul Luis Youn, Josiah Poon, and Soyeon Caren Han. Stockemotions: Discover investor emotions for financial sentiment analysis and multivariate time series. *AAAI-24 Bridge*, 2023.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 33:9459–9474, 2020.

[Li *et al.*, 2020] Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of ACM CIKM*, pages 3063–3070, 2020.

[Li *et al.*, 2023a] Xianzhi Li, Samuel Chan, Xiaodan Zhu, et al. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of EMNLP: Industry*, pages 408–422, 2023.

[Li *et al.*, 2023b] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of ACM ICAIF*, pages 374–382, 2023.

[Liu *et al.*, 2021] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of IJCAI*, pages 4513–4519, 2021.

[Liu *et al.*, 2023] Pengfei Liu, Weizhe Yuan, Jinlan Fu, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[Loukas *et al.*, 2022] Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, et al. Finer: Financial numeric entity recognition for xbrl tagging. In *Proceedings of ACL*, pages 4419–4431, 2022.

[Maia *et al.*, 2018] Macedo Maia, Siegfried Handschuh, André Freitas, et al. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of WWW*, pages 1941–1942, 2018.

[Malo *et al.*, 2014] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *JASIST*, 65(4):782–796, 2014.

[Mariko *et al.*, 2020] Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, et al. The financial document causality detection shared task (fincausal 2020). In *Proceedings of the Workshop on FNP-FNS*, pages 23–32, 2020.

[Mathur *et al.*, 2022] Puneet Mathur, Atula Neerkaje, Malika Chhibber, Ramit Sawhney, et al. Monopoly: Financial prediction from monetary policy conference videos using multimodal cues. In *Proceedings of ACM MM*, pages 2276–2285, 2022.

[Mukherjee *et al.*, 2022] Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, et al. Ectsum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of EMNLP*, pages 10893–10906, 2022.

[Radford *et al.*, 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Scao *et al.*, 2022] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[Shah *et al.*, 2022] Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of EMNLP*, pages 2322–2335, 2022.

[Shah *et al.*, 2023] Agam Shah, Suvan Paturi, and Sudheer Chava. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of ACL*, 2023.

[Sharma *et al.*, 2022] Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, et al. Finred: A dataset for relation extraction in financial domain. In *Companion Proceedings of WWW*, pages 595–597, 2022.

[Sinha and Khandait, 2021] Ankur Sinha and Tanmay Khandait. Impact of news on the commodity market: Dataset and results. In *Proceedings of FICC*, pages 589–601. Springer, 2021.

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

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

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[Wang *et al.*, 2023] Neng Wang, Hongyang Yang, and Christina Dan Wang. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. In *NeurIPS Workshop on Instruction Tuning and Instruction Following*, 2023.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022.

[Wu *et al.*, 2018] Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. Hybrid deep sequential modeling for social text-driven stock prediction. In *Proceedings of ACM CIKM*, pages 1627–1630, 2018.

[Wu *et al.*, 2023] 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.

[Xie *et al.*, 2023] 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. In *Proceedings of NeurIPS Datasets and Benchmarks*, 2023.

[Xu and Cohen, 2018] Yumo Xu and Shay B Cohen. Stock movement prediction from tweets and historical prices. In *Proceedings of ACL*, pages 1970–1979, 2018.

[Yang *et al.*, 2020] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*, 2020.

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

[Yang *et al.*, 2023b] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*, 2023.

[Yang *et al.*, 2023c] 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.

[Zhang *et al.*, 2023] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, , Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[Zheng *et al.*, 2021] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF WACV*, pages 697–706, 2021.

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