Large Language Model Agents in Finance: A Survey Bridging Research, Practice, and Real-World Deployment

Yifei Dong[1,2,*], **Fengyi Wu**[1,*], **Kunlin Zhang**[3,*], **Heng Li**[1,4], **Jingdong Sun**[4], **Zhi-Qi Cheng**[1,*†]

[1]University of Washington, [2]Columbia University, [3]University of Alberta, [4]Carnegie Mellon University

yd2616@columbia.com, wufengyi98@gmail.com, kunlin2@ualberta.ca, {hengl2,jingdons}@andrew.cmu.edu, zhiqics@uw.edu

Abstract

Large language models (LLMs) have shown promise in finance, yet they often fall short of real-world demands due to static benchmarks, interpretability gaps, and limited integration within complex multi-agent financial workflows. This survey bridges these shortcomings through a dual-perspective framework. From a *practitioner-centric* standpoint, we provide a taxonomy linking five key financial functions—*Data Analysis*, *Investment Research*, *Trading*, *Investment Management*, and *Risk Management*—to essential tasks, datasets, and constraints such as regulatory mandates and interdepartmental coordination. From a *research-focused* view, we review state-of-the-art LLM techniques, encompassing *retrieval-augmented models*, *instruction-tuned architectures*, and *multi-agent systems*, highlighting unresolved issues in accuracy, transparency, and adaptability. To address these gaps, we review domain-specific fine-tuning, modular agent designs, and dynamic, market-aware benchmarks that incorporate real-time data. We emphasize deeper researcher–practitioner collaboration and transparent model architectures as critical pathways to safer and more scalable AI adoption in finance (see Project Website[1]).

1 Introduction

"In investing, what is comfortable is rarely profitable."
— Robert Arnott

The financial sector operates in a fast-paced, multifaceted environment, where decisions rely on vast, often unstructured datasets and must conform to stringent regulations. Practitioners need rapid, accurate insights for tasks ranging from investment forecasting and risk assessment to portfolio optimization. Yet, even skilled analysts struggle to extract actionable intelligence from disparate data sources under volatile conditions. Recent advances in *Large Language Models* (LLMs) offer a promising avenue for automating processes such as parsing regulatory filings, gauging market sentiment, and supporting trading strategies [Nie *et al.*, 2024; Chen *et al.*, 2024; Lee *et al.*, 2024]. By leveraging large-scale textual and numerical data, LLMs stand poised to streamline financial workflows and enhance decision quality.

Deploying LLMs requires more than high accuracy on benchmark tasks. Modern financial institutions comprise multiple departments—*Data Analysis*, *Investment Research*, *Trading*, *Investment Management*, and *Risk Management*—each fulfilling interdependent roles, as illustrated in Figure 1. Data analysts convert raw feeds into structured content, investment researchers generate insights for strategic and tactical decisions, traders execute market orders, portfolio managers optimize risk and returns, and risk managers ensure regulatory compliance and capital allocation. If LLM-based solutions fail to integrate smoothly into this multi-agent architecture—due to limited interpretability, inadequate real-time capabilities, or inability to handle specialized subtasks—they risk disrupting institutional workflows.

Although LLMs have demonstrated strong performance on subtasks such as *Text Summarization*, *Named Entity Recognition*, *Time Series Forecasting*, and *Fraud Detection* (see Table 1), they still face notable obstacles when operating in high-stakes financial contexts. Static, outdated benchmarks rarely capture the real-time demands of trading or the adaptive nature of regulatory compliance. Issues like hallucinations, incomplete logic chains, and insufficient coordination between specialized agents also undermine reliability in critical tasks such as portfolio optimization or default risk analysis [Nie *et al.*, 2024; Chen *et al.*, 2024].

1.1 A Dual-Perspective Framework

We address the gap between cutting-edge LLM research and concrete financial needs by proposing a dual-perspective framework: a *practitioner-centric perspective* and a *research-focused perspective* (see Table 1).

1. **Practitioner-Centric Perspective:** We present a taxonomy (cf. Section 2) mapping core financial roles—*Data Analysis*, *Investment Research*, *Trading*, *Investment Management*, and *Risk Management*—to primary subtasks, datasets, and evaluation metrics. This approach reveals pressing challenges such as regulatory adherence, heterogeneous data integration, and multifaceted

[1]https://f1y1113.github.io/fin_survey/

[*]Equal contribution.

[†]Corresponding author.

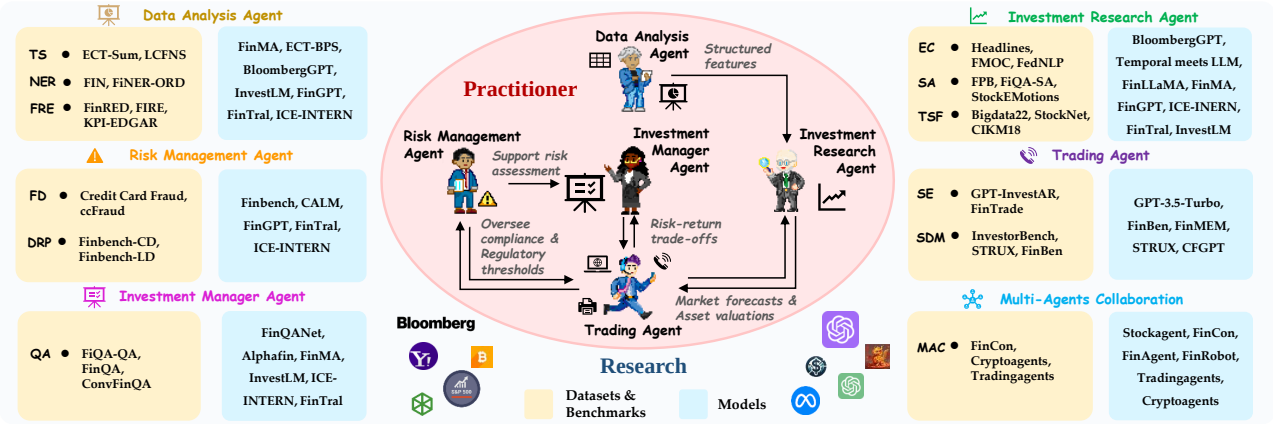

Figure 1: **Overview of LLM-based financial agents and their collaborative workflows.** Modern financial institutions rely on multiple departments—*Data Analysis*, *Investment Research*, *Trading*, *Investment Management*, and *Risk Management*—each handling specialized but interdependent roles. Key sub-tasks include *TS* (Text Summarization), *NER* (Named Entity Recognition), *FRE* (Financial Relation Extraction), *EC* (Event Classification), *SA* (Sentiment Analysis), *TSF* (Time Series Forecasting), *SE* (Strategy Execution), *QA* (Question Answering), *FD* (Fraud Detection), *DRP* (Default Risk Prediction), and *MAC* (Multi-Agent Collaboration). [Best viewed in color].

interdepartmental workflows, enabling a more grounded application of LLMs in real-world finance.

2. **Research-Focused Perspective:** We also survey state-of-the-art LLM methods—ranging from *retrieval-augmented architectures* and *instruction-tuned models* to *multi-agent frameworks*—and chart open research questions in interpretability, domain adaptation, and large-scale experimentation. As documented in Table 1, these methods underscore the interplay between financial decision-making and emerging LLM paradigms, illuminating key technical gaps.

1.2 Key Challenges and Research Gaps

Although Table 1 outlines state-of-the-art models, datasets, and metrics, it also highlights three overarching barriers that hinder LLM deployment in real-world financial contexts:

- **Static Benchmarks:** Most financial datasets in Table 1 remain fixed snapshots, ignoring market volatility, policy shifts, and real-time data [Xie *et al.*, 2024b; Yu *et al.*, 2024b]. This rigidity impedes robust stress testing and fails to reflect the dynamic conditions in which institutions operate.
- **LLM Limitations:** The "Representative Models" and "Limitations" columns in Table 1, along with prior studies [Xie *et al.*, 2023; Bhatia *et al.*, 2024], reveal recurring flaws—such as hallucinations, weak numeric reasoning, and narrow domain generalization—that compromise tasks like fraud detection and risk prediction. Accuracy alone proves inadequate in contexts demanding transparency and adaptiveness.
- **Coordination Bottlenecks:** Figure 1 depicts multi-agent collaboration across finance, yet Table 1 and related work [Xiao *et al.*, 2024; Luo *et al.*, 2025] show many frameworks neglect cross-departmental data exchange. LLMs lacking robust orchestration often yield inconsistent outputs when facing incomplete or conflicting information, undermining rebalancing, compliance, and overall risk management.

1.3 Actionable Solutions and Future Directions

To address these limitations and strengthen LLM-based systems in finance, we review four key strategies:

- **Dynamic, Market-Aware Benchmarks:** Benchmarks should integrate real-time updates and assess models across diverse market scenarios, as Table 1 and [Yu *et al.*, 2024b] propose. This strategy improves real-world applicability and tests resilience under volatility.
- **Domain-Specific Fine-Tuning:** Specialized training on financial corpora (see Table 1 and [Xie *et al.*, 2023; Bhatia *et al.*, 2024]) can boost interpretability and precision. Mechanisms for continuous learning allow models to adjust to shifting regulations and novel data patterns, mitigating the risk of obsolescence.
- **Modular Multi-Agent Architectures:** Figure 1 and Table 1 suggest that hierarchical or modular frameworks [Xiao *et al.*, 2024; Luo *et al.*, 2025] can enhance inter-agent communication and conflict resolution. Such designs reduce bottlenecks arising from fragmented data or misaligned decision-making processes.
- **Deep Researcher–Practitioner Collaboration:** Table 1 and [Chen *et al.*, 2024] highlight that shared repositories and real-world testbeds enable rapid feedback between academia and industry, refining metrics and compliance protocols essential for success.

Unlike prior surveys that focus on discrete tasks or narrowly defined benchmarks, our work embraces a holistic, practitioner-oriented viewpoint (see Project Website[†]). By aligning LLM development with the complex realities of institutional finance, we aim to enhance decision-making, fortify risk governance, and streamline operational workflows throughout the financial sector.

[†]https://f1y1113.github.io/fin_survey/

Table 1: **Overview of LLM-based financial tasks and datasets.** Organized by agent and subtask—Data Analysis, Investment Research, Trading, Investment Management, Risk Management, and Multi-Agent Collaboration—this table lists key datasets (size, period, source, license), along with diverse data modalities (text, tables, time series, structured reports), evaluation metrics, and representative LLM models. It also highlights primary challenges, emerging trends, and future research directions critical for real-world applications. [Best to zoom in].

Agent & Subtask	Datasets & Benchmarks	Modalities (Data Types)	Key Metrics	Representative Models	Limitations
Data Analysis Agent (data processing and extraction)					
Text Summarization (TS)	ECT-Sum [Mukherjee *et al.*, 2022], LCFNS [Li *et al.*, 2023a]	**Text** (earnings-call transcripts, expert bullet-point summaries, financial reports, news articles)	Recall-Oriented Understudy for Gisting Evaluation (ROUGE), BERTScore, Numerical Precision (Num-Prec.), Summarization Consistency (SummaC)	FinMA [Xie *et al.*, 2023], ECT-BPS [Mukherjee *et al.*, 2022], FinTral [Bhatia *et al.*, 2024], InvestLM [Yang *et al.*, 2023b], FinGPT [Yang *et al.*, 2023a], ICE-INTERN [Hu *et al.*, 2024]	**Datasets & Benchmarks:** (1) Lack of integrating both structured & unstructured data, (2) Limited annotated entity/relationship types, (3) Lack of dynamic data. **Models:** (1) High computational overhead (energy consumption), (2) Limited numeric reasoning & lack of online update.
Name-Entity Recognition (NER)	FIN [Alvarado *et al.*, 2015], FiNER-ORD [Shah *et al.*, 2023b]	**Text** (US Financial contracts, Exchange Commission (SEC) filings, financial news articles)	Precision, Recall, F1-score	FinMA [Xie *et al.*, 2023], BloombergGPT [Wu *et al.*, 2023], InvestLM [Yang *et al.*, 2023b], ICE-INTERN [Hu *et al.*, 2024]	**Datasets & Benchmarks:** (1) Small-scale coverage, (2) Limited annotated entity types, (3) Lack of dynamic data. **Models:** (1) Weak entity linking across documents, (2) Lack of domain-specific pretraining, (3) Limited numeric reasoning.
Financial Relation Extraction (FRE)	FinRED [Sharma *et al.*, 2022], FIRE [Hamad *et al.*, 2024], KPI-EDGAR [Deußer *et al.*, 2022]	**Text** (EDGAR filings, earnings-call transcripts, SEC fillings, KPI mentions)	Precision, Recall, F1, adjusted F1-score	FinTral [Bhatia *et al.*, 2024], ICE-INTERN [Hu *et al.*, 2024]	**Datasets & Benchmarks:** (1) Limited annotated entity/relationship types, (2) Lack of temporal data linking, (3) Inconsistent domain-specific labeling. **Models:** (1) Difficulty detecting event-based relationships, (2) Limited domain-specific pretraining, (3) Lack of online update.
Investment Research Agent (asset evaluation and market prediction)					
Event Classification (EC)	FOMC [Shah *et al.*, 2023a], FedNLP [Lee *et al.*, 2021], Headlines [Sinha and Khandait, 2021]	**Text** (policy statements, news headlines, earnings-call transcripts)	Accuracy, Precision, Recall, F1-score	BloombergGPT, FinLLaMA [Iacovides *et al.*, 2024], Temporal meets LLM [Yu *et al.*, 2023], FinMA [Xie *et al.*, 2023], FinGPT [Yang *et al.*, 2023a], ICE-INTERN [Hu *et al.*, 2024], FinTral [Bhatia *et al.*, 2024]	**Datasets & Benchmarks:** (1) No real-time market data, (2) Limited domain-specific event understanding, (3) Overlook multi-asset forecasting. **Models:** (1) Insufficient domain-specific pretraining, (2) Static fine-tuning hinders real-time adaptability.
Sentiment Analysis (SA)	FPB [Malo *et al.*, 2014], FiQA-SA [Maia *et al.*, 2018], StockEmotions [Lee *et al.*, 2023]	**Text** (news articles, microblogs, comments from StockTwits)	Accuracy, Precision, Recall, F1-score, Mean Squared Error (MSE)	FinGPT [Yang *et al.*, 2023a], FinMA [Xie *et al.*, 2023], BloombergGPT [Wu *et al.*, 2023], ICE-INTERN [Hu *et al.*, 2024], FinTral [Bhatia *et al.*, 2024], InvestLM [Yang *et al.*, 2023b]	**Datasets & Benchmarks:** (1) Reliance on short texts, no long-term context, (2) Lack of fundamental financial indicators, (3) Limited set of sentiment labels. **Models:** (1) Over-simplified sentiment or polarity classification, (2) Insufficient domain-specific pretraining, (3) Static fine-tuning hinders real-time adaptability.
Time Series Forecasting (TSF)	StockNet [Xu and Cohen, 2018], Bigdata22 [Soun *et al.*, 2022], CIKM18 [Wu *et al.*, 2018]	**Text** (tweets, microblogs) **Time Series** (stock prices)	Accuracy, Matthews Correlation Coefficient (MCC)	Temporal meets LLM [Yu *et al.*, 2023], FinLLaMA [Iacovides *et al.*, 2024], FinGPT [Yang *et al.*, 2023a], FinMA [Xie *et al.*, 2023]	**Datasets & Benchmarks:** (1) Lack of multi-asset coverage, (2) No real-time data, (3) Overlook fundamental indicators. **Models:** (1) Weak asset-specific feature integration, (2) Insufficient domain-specific pretraining, (3) Static fine-tuning hinders real-time adaptability.
Trading Agent (strategy execution and decision-making)					
Strategy Execution (SE)	GPT-InvestAR [Gupta, 2023], FinTrade [Xie *et al.*, 2024a]	**Text** (earnings reports, sentiment); **Tables** (historical prices)	Profitability, Sharpe Ratio (SR)	GPT-3.5-Turbo [Gupta, 2023], FinBen [Xie *et al.*, 2024a]	**Datasets & Benchmarks:** (1) Narrow market coverage, (2) Overlook high-frequency trading, (3) Lack of real-time data, (4) Ignore portfolio diversification. **Models:** (1) Conservative decision-making bias, (2) Dependency on closed-source backbone hinders domain adaptation.
Support Decision-Making (SDM)	InvestorBench [Li *et al.*, 2024a], STRUX [Lu *et al.*, 2024], FinBen [Xie *et al.*, 2024a]	**Text** (financial reports); **Tables** (crypto market data); **Time Series** (stock prices)	Cumulative Return (CR), Sharpe Ratio (SR), Annualized Volatility (AV), Maximum Drawdown (MDD)	FinMEM [Yu *et al.*, 2024a], STRUX [Lu *et al.*, 2024], CFGPT [Li *et al.*, 2023b]	**Datasets & Benchmarks:** (1) Narrow real-world asset coverage, (2) Limited multi-asset data integration, (3) Ignore risk-parity or correlation structures. **Models:** (1) Over-reliance on simplistic reward signals, (2) Lack of online adaptation, (3) Inconsistent performance under changing markets.
Investment Manager Agent (portfolio optimization and allocation)					
Question-Answering (QA)	FiQA-QA [Maia *et al.*, 2018], FinQA [Chen *et al.*, 2021], ConvFinQA [Chen *et al.*, 2022]	**Text** (financial news, social media posts, earnings statements); **Tables** (S&P 500 market tables)	Normalized Discounted Cumulative Gain (nDCG), Mean Reciprocal Rank (MRR), Execution Accuracy, Program Accuracy	FinQANet [Chen *et al.*, 2022], Alphafin [Li *et al.*, 2024b], FinMA [Xie *et al.*, 2023], InvestLM [Yang *et al.*, 2023b], ICE-INTERN [Hu *et al.*, 2024], FinTral [Bhatia *et al.*, 2024]	**Datasets & Benchmarks:** (1) Reliance on static & synthetic datasets, (2) Limited multi-modal support, (3) Oversimplification via synthetic data. **Models:** (1) Struggle with long & multi-hop reasoning, (2) Inability to adapt to dynamic financial data & incremental contexts.
Risk Management Agent (fraud detection and compliance)					
Fraud Detection (FD)	Credit Card Fraud [Balasubramanian *et al.*, 2022], ccFraud [Kamaruddin and Ravi, 2016]	**Text** (credit card transactions); **Tables** (financial logs)	Accuracy, Precision, Recall, F1-score, Area Under the Receiver Operating Characteristic Curve (AUC-ROC)	Finbench [Yin *et al.*, 2023], FinGPT [Yang *et al.*, 2023a], CALM [Feng *et al.*, 2023], FinTral [Bhatia *et al.*, 2024], ICE-INTERN [Hu *et al.*, 2024]	**Datasets & Benchmarks:** (1) Class imbalance with fewer fraudulent transactions, (2) Limited feature diversity, (3) Lack of long-term tracking of borrower behaviors. **Models:** (1) Poor scalability to real-time applications, (2) Struggle to adapt to evolving fraud patterns, (3) Inability to handle large data volumes effectively.
Default Risk Prediction (DRP)	Finbench-CD [Yin *et al.*, 2023], Finbench-LD [Yin *et al.*, 2023]	**Text** (home equity loans, vehicle loans); **Tables** (credit card client records)	Accuracy, Precision, Recall, F1-score	Finbench [Yin *et al.*, 2023], FinGPT [Yang *et al.*, 2023a], CALM [Feng *et al.*, 2023]	**Datasets & Benchmarks:** (1) Highly imbalanced data distribution, (2) Limited feature diversity, (3) Lack of real-time dynamic risk modeling. **Models:** (1) Struggle with ephemeral borrower behaviors, (2) Poor interpretability for credit decisions, (3) Difficult scaling for large corporate portfolios.
Multi-Agent Collaboration (MAC)	FinCon [Yu *et al.*, 2024b], Tradingagents [Xiao *et al.*, 2024], Cryptoagents [Luo *et al.*, 2025]	**Text** (financial news, company filling reports); **Tables** (cryptocurrency market data); **Audio** (ECC audio recordings)	Chain-of-Thought Accuracy (CoT Acc.), Profitability, Portfolio Performance, Cumulative Return (CR), Sharpe Ratio (SR), Max Drawdown (MDD)	Stockagent [Zhang *et al.*, 2024a], FinCon [Yu *et al.*, 2024b], Tradingagents [Xiao *et al.*, 2024], Cryptoagents [Luo *et al.*, 2025], FinAgent [Zhang *et al.*, 2024b], FinRobot [Yang *et al.*, 2024]	**Datasets & Benchmarks:** (1) Lack support for real-time/high-frequency trading, (2) Overlook multi-asset data sources, (3) Fail to capture order execution dynamics. **Models:** (1) Sensitive to prompt engineering, (2) Lack of online adaptation, (3) Inherent biases hamper collaborative synergy.

2 Taxonomy of LLM-based Agents in Finance

This section presents a taxonomy of Large Language Model (LLM) integration in financial workflows, categorizing five key agents: *Data Analysis*, *Investment Research*, *Trading*, *Investment Management*, and *Risk Management*. As shown in Figure 1, each agent specializes in tasks from unstructured data processing to market forecasting and portfolio optimization. Table 1 summarizes datasets, benchmarks, evaluation metrics, and state-of-the-art models, while Table 2 details dataset sizes, collection periods, sources, and licensing terms. We conclude with an analysis of LLM adoption's benefits, limitations, and future directions in financial workflows.

2.1 Data Analysis Agent

Definition and Scope. Data Analysis Agents form the foundation of modern financial workflows by aggregating, cleaning, and reconciling heterogeneous sources such as SEC filings, news feeds, and corporate disclosures. They integrate unstructured texts (e.g., annual reports, earnings-call transcripts) with structured data (e.g., prices, trading volumes) to produce a coherent market view. These refined outputs support downstream tasks in investment research, trading, and risk management, while also enabling real-time compliance.

2.1.1 Tasks & Benchmarks. Data Analysis Agents typically address three core tasks—*text summarization* (TS), *named*

entity recognition (NER), and *financial relation extraction* (FRE)—each critical for organizing unstructured data and extracting actionable insights. As summarized in Table 1 and detailed in Table 2, these tasks rely on domain-specific corpora and metrics tailored to financial contexts:

Text Summarization (TS). Financial text summarization task requires both numerical precision and robust contextual understanding. Benchmarks like ECT-Sum [Mukherjee *et al.*, 2022], with 2,425 document–summary pairs from earnings-call transcripts and Reuters, and LCFNS [Li *et al.*, 2023a], comprising over 430K news–headline pairs, typically apply ROUGE, BERTScore, Numerical Precision (Num-Prec.), and Summarization Consistency (SummaC) to assess accuracy. However, most corpora focus on single-document abstractive summaries and rarely incorporate structured data [Xie *et al.*, 2024b]. This gap restricts real-world applicability where robust, multi-document integrations are often essential.

Named Entity Recognition (NER). NER task identifies crucial entities such as companies, individuals, and financial terms. Datasets like FIN [Alvarado *et al.*, 2015] focus on SEC filings and legal documents, while FiNER-ORD [Shah *et al.*, 2023b] annotates 4,739 sentences within 201 financial news articles. Standard metrics include Precision, Recall, and F1. As shown in Table 1, NER datasets often suffer from narrow coverage and limited entity classes, omitting key domain-specific labels (e.g., *LoanType*, *DefaultIndicator*). Future directions include larger-scale annotations and continuous entity tracking across quarterly or annual disclosures.

Financial Relation Extraction (FRE). FRE task determines inter-entity relationships vital for tasks like M&A analysis, ownership tracking, and supply-chain risk assessment. FinRED [Sharma *et al.*, 2022], FIRE [Hamad *et al.*, 2024], and KPI-EDGAR [Deußer *et al.*, 2022] each provide thousands of annotated sentences covering various relation types. Precision, Recall, and F1 are standard metrics, but these benchmarks mainly feature static document snapshots. Incorporating temporal aspects (e.g., evolving subsidiary structures) and numeric ratios remains a challenge. Addressing these limitations would deepen FRE's applicability to real-time compliance and risk analytics.

2.1.2 LLM-Based Model Agents. Large language models (LLMs) have significantly advanced Data Analysis tasks in finance. FinMA [Xie *et al.*, 2023] fine-tunes LLaMA on 136K multi-task instructions, excelling at NER and summarization but remaining limited by quantitative reasoning and static updates [Bhatia *et al.*, 2024]. ECT-BPS [Mukherjee *et al.*, 2022] combines extractive (FinBERT [Liu *et al.*, 2021]) and abstractive (T5 [Raffel *et al.*, 2020]) methods for summarizing earnings-call transcripts, though pipeline architectures still risk factual inconsistencies. Additional strategies, including multi-granularity lattice frameworks [Li *et al.*, 2019] and chain-of-thought prompting in GPT-4 Turbo [Kim *et al.*, 2024], further refine domain-specific adaptation, improving interpretability and robustness in financial applications.

Challenges and Future Directions. Despite these advances, data analysis remains constrained by gaps in numerical reasoning, limited multi-document modeling, and a lack of real-time updates. As Table 1 and recent work [Xie *et al.*, 2023; Bhatia *et al.*, 2024] suggest, LLM-based methods must further integrate structured and unstructured data sources to maintain factual consistency under rapidly shifting market conditions. Future research should prioritize robust architectures that seamlessly blend text with numeric data, preserve domain-specific accuracy, and dynamically adjust to evolving corporate and regulatory disclosures.

2.2 Investment Research Agent

Definition and Scope. The Investment Research Agent conducts in-depth analyses of macroeconomic conditions, sector trends, and individual asset fundamentals to guide both strategic portfolio decisions and tactical trading. By synthesizing data from policy announcements, financial news, and social media, the agent merges qualitative market narratives with quantitative metrics. As outlined in Table 1, its core responsibilities span three tasks: *event classification* (EC), *sentiment analysis* (SA), and *time series forecasting* (TSF).

2.2.1 Tasks & Benchmarks. The Investment Research Agent transforms textual and numerical inputs into actionable insights, leveraging diverse datasets, advanced modeling techniques, and evaluation protocols (see Table 2).

Event Classification (EC). A primary goal of EC task is to identify significant market-moving events related to monetary policy or investor sentiment shifts. For instance, the FOMC dataset [Shah *et al.*, 2023a] includes meeting minutes, speeches, and press conferences (1996–2022), enabling classifications such as "hawkish" or "dovish." FedNLP [Lee *et al.*, 2021] adds more than 1,000 speeches and 100 press conferences (2015–2020), while the Headlines dataset [Sinha and Khandait, 2021] provides 11,412 annotated news headlines (2000–2019). Accuracy, Precision, Recall, and F1-score are standard metrics; however, real-time integration of yield curves or multi-asset information is often missing.

Sentiment Analysis (SA). This task gauges market sentiment by extracting opinions from textual data. FPB [Malo *et al.*, 2014] contains 4,840 annotated sentences, FiQA-SA [Maia *et al.*, 2018] covers financial microblogs, and StockEmotions [Lee *et al.*, 2023] compiles 10,000 StockTwits posts. Accuracy and F1 are common metrics, yet short-text constraints and limited label categories overlook multi-turn analyst calls and nuanced sentiment.

Time Series Forecasting (TSF). The TSF task fuses historical price data with textual signals to forecast future market behavior and trends. StockNet [Xu and Cohen, 2018] offers two years of S&P 500 prices for 88 stocks aligned with StockTwits commentary; Bigdata22 [Soun *et al.*, 2022] and CIKM18 [Wu *et al.*, 2018] integrate social media with price data. Accuracy and MCC are typical metrics, but many benchmarks lack multi-asset coverage and fundamental factors (e.g., P/E ratios), limiting practical utility.

2.2.2 LLM-Based Model Agents. Recent large language models (LLMs) have demonstrated significant promise in bolstering Investment Research. BloombergGPT [Wu *et al.*, 2023] (50B parameters) excels at sentiment analysis across

financial news and social media, though ambiguity in contextual interpretation remains a challenge. Temporal meets LLM [Yu *et al.*, 2023] harnesses GPT-4 for event classification and forecasting by merging company profiles, time series, and news sources within structured prompts. FinLLaMA [Iacovides *et al.*, 2024], a LoRA-based fine-tuning of Llama-3-7B [Touvron *et al.*, 2023], effectively classifies sentiment intensity and achieves competitive Sharpe ratios in portfolio simulations, yet static fine-tuning and limited domain-specific pretraining hinder adaptability in fast-evolving markets.

Challenges and Future Directions. Investment Research Agents must integrate diverse data streams, adapt in real time, and broaden beyond single-asset coverage to enhance decision-making [Wu *et al.*, 2023; Yu *et al.*, 2023]. High-frequency metrics from social media or policy announcements are vital for near-instant updates, while cross-asset modeling should encompass bonds, commodities, and cryptocurrencies to capture systemic trends more effectively. Advanced fine-tuning strategies that support continuous learning [Iacovides *et al.*, 2024] would further ensure resilience under volatile market shifts. Deeper data integration and refined model adaptation can strengthen both the reliability and granularity of insights in rapidly evolving financial environments (see Table 1).

2.3 Trading Agent

Definition and Scope. A Trading Agent executes buy and sell orders in real time, adapts strategies to evolving market conditions, and ensures compliance with internal and external regulations. By continuously monitoring price fluctuations, managing dynamic portfolio allocations, and fusing market-driven signals, it serves as a critical revenue driver for financial institutions. Typically, its functions include *Strategy Execution* and *Support Decision-Making*.

2.3.1 Tasks & Benchmarks. Trading Agents typically focus on two key tasks: *Strategy Execution* and *Support Decision-Making*. Table 1 and Table 2 provide an overview of relevant datasets and benchmarks, detailing their size, modalities, and evaluation metrics.

Strategy Execution. This task requires near-real-time processing of both textual disclosures (e.g., 10-K filings, earnings reports) and structured price data (open/high/low/close, volume) to guide precise and timely buy/sell orders. Representative datasets include GPT-InvestAR [Gupta, 2023], which connects 24,200 annual reports from 1,500 U.S. companies (2002–2023) with historical stock prices, and FinTrade [Xie *et al.*, 2024a], which integrates a year of daily price data for ten equities with corporate filings and market-moving news. While these benchmarks combine text and tabular data, they often omit high-frequency updates, macroeconomic variables (e.g., interest rates), and cross-asset correlations, restricting their utility in broader market modeling and long-horizon strategy testing. Typical evaluation metrics include profit, accuracy, Sharpe ratio, or maximum drawdown, though most remain at a daily or lower frequency, further limiting responsiveness to intraday market fluctuations and dynamic trading conditions.

Support Decision-Making. The second task leverages multimodal data—spanning textual insights, financial tables, and time-series signals—to optimize asset allocation and manage risk. InvestorBench [Li *et al.*, 2024a] offers 10,000 curated trading scenarios across asset classes (cryptocurrencies, equities, ETFs), assessing performance through metrics such as cumulative return, Sharpe ratio, annualized volatility, and maximum drawdown. STRUX [Lu *et al.*, 2024] provides 4,258 annotated earnings-call transcripts to classify the impact of favorable or adverse corporate factors. Although these datasets showcase diverse modalities and evaluation approaches, many remain constrained to single-asset scenarios, rely on delayed market data, and rarely incorporate real-world execution constraints like transaction costs, liquidity thresholds, or regulatory mandates. More comprehensive, multi-asset, and real-time benchmarks are needed to capture the complexities of institutional trading environments.

2.3.2 LLM-Based Model Agents. Recent advances in large language models (LLMs) show promise for Trading Agents. FinMEM [Yu *et al.*, 2024a] uses a memory-enhanced GPT-4-Turbo [Achiam *et al.*, 2023] architecture to adapt risk preferences to market volatility, though scalability and interpretability challenges persist. The STRUX approach [Lu *et al.*, 2024] converts earnings-call transcripts into concise tables and applies self-reflection to classify key facts as favorable or adverse, but depends heavily on transcript data, risking oversimplification when macro signals are missing.

Challenges and Future Directions. Real-world Trading Agents face numerous obstacles that restrict broader adoption. Many benchmarks and datasets concentrate on single-asset settings, limiting cross-asset hedging or multi-currency portfolios [Gupta, 2023; Xie *et al.*, 2024a]. Scarce high-frequency data impedes intraday or algorithmic trading, and few studies incorporate real-world constraints such as regulatory mandates or capital requirements. Future work should emphasize more diverse, multi-asset data streams and real-time pipelines, coupled with refined fine-tuning strategies that adapt to volatile market regimes. By integrating textual, numerical, and risk parameters more effectively (see Table 1), next-generation Trading Agents can better align trade execution with firm-wide objectives.

2.4 Investment Manager Agent

Definition and Scope. The Investment Manager Agent oversees portfolio decisions to balance risk and return under regulatory mandates. By analyzing market conditions, corporate fundamentals, and macroeconomic indicators, it designs long-term strategies to mitigate systemic and idiosyncratic risks. Although its remit includes scenario analysis, stress testing, and portfolio optimization, we focus on *Question-Answering (QA)* as a representative task requiring both textual and numerical reasoning to guide investment decisions.

2.4.1 Tasks & Benchmarks. In the QA sub-task, institutional investors query large-scale financial datasets to evaluate metrics such as valuation, regulatory disclosures, and risk-return profiles. FiQA-QA [Maia *et al.*, 2018] provides 5,676 question–answer pairs drawn from financial news and

Table 2: Comprehensive Overview of Representative Financial Datasets. The table summarizes key characteristics—including raw data size, collection period, data sources, and license types—of datasets used by various LLM-based agents in finance. [Best to zoom in].

Agent & Subtask		Dataset	Raw Data Size	Collection Period	Source	License
Data Analysis Agent	TS	ECT-Sum	2,425 document-summary pairs	Jan 2019 - Apr 2022	Earnings call transcripts, Reuters articles	GPL-3.0 license
		LCFNS	430,820 news-summary pairs	Jan 2013 - Jun 2020	Major financial portals	Public
	NER	FIN	54,256 words (8 annotated agreements)	-	U.S. SEC filings, CoNLL-2003 [Sang and De Meulder, 2003]	None Public
		FiNER-ORD	201 financial news articles, 4,739 sentences	Jul 2015 - Oct 2015	Webz.io	CC BY-NC 4.0
	RE	FinRED	7,775 sentences, 29 relation types	Jul 2015 - Oct 2015, Jun 2019 - Sep 2019	Financial news articles, earnings calls	Public
		FIRE	3,025 instances, 18 relation types	1993 - 2021	Financial news articles, SEC filings	CC BY 4.0
		KPI-EDGAR	1,355 sentences	-	EDGAR database annual reports	MIT license
Investment Research Agent	EC	FOMC	214 minutes, 1,026 speeches, 63 transcripts	1996 - 2022	Federal Open Market Committee communications	CC BY-NC 4.0
		FedNLP	1000+ speeches, 100+ press conferences	Jan 2015 - Jul 2020	Federal Reserve communications	Public
		Headlines	11,412 annotated news headlines	2000 - 2019	Gold commodity market	CC BY-NC-ND 4.0
	SA	FPB	4,840 sentences	-	Financial news articles	CC BY-SA 3.0
		FiQA-SA	529 annotated headlines and 774 financial microblogs	-	Financial news and social media	CC-BY-3.0
		StockEmotions	10,000 investor comments, 12 emotions	Jan 2020 - Dec 2020	StockTwits	Public
	TSF	StockNet	26614 price movement data of 88 stocks	2014 - 2016	S&P 500 stocks, StockTwits	MIT license
		Bigdata22	7,164 tweets	2014 - 2015	S&P 500 stocks	Public
		CIKM18	47 stocks from S&P 500	Jan 2017 - Nov 2017	Yahoo Finance, Twitter	Public
Trading Agent	SE	GPT-InvestAR	10-K filings with 24,200 documents	2002 - 2023	Annual SEC report filings	MIT license
		FinTrade	16137 news, 65 10-K/10-Q files, 4970 price data from 10 stocks	One year period	Stock prices, SEC filings, news	MIT license
	SDM	InvestorBench	5000 stock prices, 2000 earnings reports, 50000 cryptocurrency articles	2019 - 2023	Yahoo Finance, CoinMarketCap, CryptoPotato, CoinTelegraph	MIT license
		STRUX	11,950 quarterly earnings call transcripts	2017 - 2024	Motley Fool website, NASDAQ 500 and S&P 500 stocks	Public
Investment Management Agent	QA	FiQA-QA	17,072 QA pairs	-	Financial microblogs, reports, and news articles	CC-BY-3.0
		FinQA	8,281 QA pairs	-	Earnings reports (S&P 500)	MIT License
		ConvFinQA	3,892 conversations, 14,115 questions	-	Earnings reports (S&P 500)	MIT License
Risk Management Agent	FD	Credit Card Fraud	11,392 transactions	2013	European cardholders	DbCL v1.0
		ccFraud	10,485 transactions	2013	European cardholders	Public
	DRP	Finbench-CD	30k credit records	Apr - Sep 2005	Credit card clients in Taiwan	CC BY-NC 4.0
		Finbench-LD	10k credit records, 200k vehicle loan records	-	Loan records	CC BY-NC 4.0
Multi-Agent Collaboration	MAC	FinCon	Data size not specified	August 2020 - August 2023	Yahoo Finance, Form 10-Q, Form 10-K, Zacks Rank,Earning conference calls	CC BY-NC 4.0
		Tradingagents	Data size not specified	Jan - Mar 2024	S&P 500 stocks, Bloomberg, Yahoo, Reddit, Twitter	None Public
		Cryptoagents	Top 30 cryptocurrency data	Jun 2023 - Sep 2024	Blockchain.info, Coin Metrics, Cointelegraph	None Public

microblogs, with relevance assessed using metrics like Normalized Discounted Cumulative Gain (nDCG) and Mean Reciprocal Rank (MRR). FinQA [Chen *et al.*, 2021] comprises 8,281 expert-annotated QA pairs derived from S&P 500 earnings reports, emphasizing numerical reasoning. ConvFinQA [Chen *et al.*, 2022] extends QA to multi-turn dialogues, testing compositional reasoning across textual and tabular data in 3,892 dialogues (14,115 questions). Although these benchmarks capture essential aspects of financial QA, they often rely on static, archived reports rather than real-time market feeds, limiting their applicability in dynamic asset management where continuous data and frequent rebalancing are critical. They also provide limited coverage of constraints such as liquidity, leverage, or compliance thresholds, which are crucial for institutional settings.

2.4.2 LLM-Based Model Agents. Recent large language models (LLMs) enhance QA and decision support in portfolio management by combining textual reasoning with numerical analysis. ConvFinQA [Chen *et al.*, 2022] leverages GPT-3-based prompting for multi-turn queries, but encounters challenges with multi-hop dependencies, domain-specific numeric operations, and changing market conditions. AlphaFin [Li *et al.*, 2024b] employs a Retrieval-Augmented Generation pipeline to fetch real-time market data, mitigating hallucinations and improving decision accuracy. However, issues such as infrastructure overhead, latency in high-frequency scenarios, and the need for adaptive domain-specific training remain significant obstacles. Current QA metrics (e.g., execution accuracy, program accuracy) do not fully reflect portfolio performance over extended horizons or under stress-test scenarios.

Challenges and Future Directions. In practical settings, Investment Manager Agents must integrate a range of data streams—from real-time market feeds and corporate events to internal risk models—while coping with stress tests, capital constraints, and shifting macroeconomic conditions [Li *et al.*, 2024b; Chen *et al.*, 2022]. Addressing this complexity requires continuous pipelines that merge static QA benchmarks with dynamic scenarios, more advanced multi-step numeric reasoning, and finer modeling of risk factors. It also demands new evaluation frameworks that track long-term portfolio outcomes, stress resilience, and regulatory compliance under variable economic backdrops (see Table 1).

2.5 Risk Management Agent

Definition and Scope. The Risk Management Agent underpins a financial institution's stability by identifying, assessing, and mitigating diverse risks, including market, credit, and operational threats, while ensuring regulatory compliance. It continuously monitors transactions, counterparties, and external factors that may compromise institutional integrity. Although practical risk management extends to capital adequacy, liquidity stress testing, and scenario analysis, this survey highlights two representative tasks: *Fraud Detection* and *Default Risk Prediction*.

2.5.1 Tasks & Benchmarks. Table 1 and Table 2 outline key datasets and metrics for risk management tasks, focusing on fraud detection and default risk prediction. These datasets typically involve highly imbalanced classes and intricate feature spaces, reflecting real-world constraints.

Fraud Detection. This task must distinguish legitimate from malicious transactions under severe class imbalance and evolving attack patterns. The *Credit Card Fraud* dataset [Balasubramanian *et al.*, 2022] and *ccFraud* [Kamaruddin and Ravi, 2016] each contain around 10,000–11,000 records, with only a small fraction deemed fraudulent. Data modalities often include anonymized textual logs and tabular transaction attributes. Evaluation metrics such as Accuracy, Precision, Recall, F1-score, and AUC-ROC measure how effectively models cope with heavily skewed distributions. However, PCA-based transformations and privacy constraints limit contextual details (e.g., merchant profiles or user behavior), making real-world adaptation and generalization across different financial systems challenging.

Default Risk Prediction. Assessing the likelihood a borrower will fail to repay is another critical risk management

task with significant financial implications. *Finbench-CD* and *Finbench-LD* [Yin *et al.*, 2023] comprise credit card and loan datasets collected over defined periods (e.g., April–September 2005 in Taiwan), integrating textual descriptors and tabular indicators (annual income, credit history length). Accuracy, Precision, Recall, and F1-score are standard, although these datasets rarely incorporate macro-level shifts such as interest rate changes, inflation, or unemployment trends. Limited longitudinal tracking and a lack of cross-lender data further reduce applicability for evolving borrower behavior analysis and long-term financial risk modeling.

2.5.2 LLM-Based Model Agents. Recent work employs large language models (LLMs) to enhance risk management via natural-language representations of structured data. Finbench [Yin *et al.*, 2023] uses a *Profile Tuning* approach with GPT-2 [Radford *et al.*, 2019], outperforming traditional machine learning baselines through cost-sensitive learning. CALM [Feng *et al.*, 2023] leverages instruction-tuned models like Llama2-chat (with LoRA) on nine fraud and default datasets, attaining performance comparable to GPT-4 [Achiam *et al.*, 2023]. Nevertheless, the reliance on static, labeled corpora and high computational demands hamper adaptation to shifting fraud schemes or credit conditions, while real-time scalability remains a significant hurdle.

Challenges and Future Directions. Risk Management Agents must remain interpretable and compliant with evolving regulations, necessitating transparent audit trails for automated decisions [Yin *et al.*, 2023; Feng *et al.*, 2023]. Adversarial robustness is likewise crucial, given persistent efforts to evade fraud checks or manipulate credit evaluations [Achiam *et al.*, 2023]. Further, continuous adaptation to emerging threats demands real-time or near-real-time learning strategies that integrate external data (e.g., merchant details, credit bureau updates) alongside high-frequency transaction logs. Future work should prioritize bridging structured and unstructured inputs for richer context, refining online learning to ensure rapid updates, and developing evaluation protocols that measure not only predictive accuracy but also operational constraints (see Table 1).

2.6 Multi-Agent Collaboration

Definition and Scope. Multi-Agent Collaboration involves coordinated interaction among specialized agents, including Data Analysis, Investment Research, Trading, Investment Management, and Risk Management. Each agent contributes unique insights—ranging from extracting textual intelligence and performing quantitative analyses to executing trades and assessing risk. Their synchronized outputs drive informed decisions that meet shared objectives like regulatory compliance, operational efficiency, and profit maximization. As indicated in Table 1, this holistic approach addresses the complex challenges of modern finance.

2.6.1 Benchmarks. Multiple benchmarks assess how well agents collaborate in real-world scenarios. FinCon [Yu *et al.*, 2024b] compiles stock prices, daily news, regulatory filings, and earnings-call audio (2020–2023) for tasks such as stock trading and portfolio management. It leverages diverse data modalities, including long-term annual reports, medium-term quarterly updates, and daily news. Evaluations often measure cumulative returns, Sharpe ratios, and maximum drawdowns. Cryptoagents [Luo *et al.*, 2025] examines top-30 digital assets with real-time feeds and social sentiment, while Tradingagents [Xiao *et al.*, 2024] collects fundamentals, sentiment, and macroeconomic indicators for early 2024. Although these datasets highlight different asset classes and data modalities, most rely on daily or historical feeds, focus on single-asset scenarios, and omit market microstructure factors such as bid-ask spreads and execution latencies. These omissions hinder realistic simulation and compliance validation.

2.6.2 LLM-Based Model Agents. Recent work uses large language models (LLMs) to facilitate multi-agent collaboration across varied tasks. Stockagent [Zhang *et al.*, 2024a] employs GPT-3.5-Turbo/Gemini-Pro within an event-driven framework, while FinAgent [Zhang *et al.*, 2024b] augments LLMs with reflection layers that incorporate market metrics, historical actions, and sentiment analysis. FinCon [Yu *et al.*, 2024b] applies a hierarchical manager–analyst structure with daily Conditional Value at Risk (CVaR) monitoring and multi-episode refinement, whereas Tradingagents [Xiao *et al.*, 2024] uses specialized analyst, researcher, and trader roles in an institutional context. Cryptoagents [Luo *et al.*, 2025] targets digital assets, integrating real-time data feeds and specialized roles for allocation and execution. Although these approaches show promise, they often suffer from prompt sensitivity, inherent LLM biases, and the lack of high-frequency trading support.

Challenges and Future Directions. Despite recent progress [Zhang *et al.*, 2024a; Zhang *et al.*, 2024b], several obstacles still limit multi-agent systems in finance. Many benchmarks rely on historical or daily data [Xiao *et al.*, 2024; Luo *et al.*, 2025], failing to capture the real-time, high-frequency flows vital for algorithmic trading and rapid risk assessments. The absence of multi-asset integration constrains portfolio diversification, while neglecting market microstructure details reduces simulation fidelity. Current LLM-based frameworks face computational overhead, model bias, and regulatory restrictions like best-execution rules or privacy mandates. Future research should emphasize large-scale simulations with real-time data pipelines, refined risk and transaction-cost modeling, and adaptive LLM architectures aligned with evolving institutional norms.

3 Conclusion

This survey explores how large language models (LLMs) reshape financial workflows across data analysis, investment research, trading, investment management, and risk management. Despite progress in summarization, event classification, and multi-agent collaboration, gaps persist in data recency, numeric reasoning, and real-time orchestration. Future research must unite domain-specific fine-tuning with scalable architectures that adapt to evolving markets and regulations. By integrating robust benchmarks, fine-grained evaluation metrics, and continuous learning paradigms, LLM-based systems can deliver more transparent, resilient, efficient results for high-stakes financial contexts, yet success hinges on bridging academic insights with validation pipelines.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[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 *ALTA*, pages 84–90, 2015.

[Balasubramanian *et al.*, 2022] Natarajan Balasubramanian, Yang Ye, and Mingtao Xu. Substituting human decision-making with machine learning: Implications for organizational learning. *Academy of Management Review*, 47(3):448–465, 2022.

[Bhatia *et al.*, 2024] Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. Fintral: A family of gpt-4 level multimodal financial large language models. *arXiv preprint arXiv:2402.10986*, 2024.

[Chen *et al.*, 2021] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 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. *arXiv preprint arXiv:2210.03849*, 2022.

[Chen *et al.*, 2024] Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*, 2024.

[Deußer *et al.*, 2022] Tobias Deußer, Syed Musharraf Ali, Lars Hillebrand, Desiana Nurchalifah, Basil Jacob, Christian Bauckhage, and Rafet Sifa. Kpi-edgar: A novel dataset and accompanying metric for relation extraction from financial documents. In *ICMLA*, pages 1654–1659. IEEE, 2022.

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

[Gupta, 2023] Udit Gupta. Gpt-investar: Enhancing stock investment strategies through annual report analysis with large language models. *arXiv preprint arXiv:2309.03079*, 2023.

[Hamad *et al.*, 2024] Hassan Hamad, Abhinav Kumar Thakur, Nijil Kolleri, Sujith Pulikodan, and Keith Chugg. Fire: A dataset for financial relation extraction. In *NAACL*, pages 3628–3642, 2024.

[Hu *et al.*, 2024] Gang Hu, Ke Qin, Chenhan Yuan, Min Peng, Alejandro Lopez-Lira, Benyou Wang, Sophia Ananiadou, Jimin Huang, and Qianqian Xie. No language is an island: Unifying chinese and english in financial large language models, instruction data, and benchmarks. *arXiv preprint arXiv:2403.06249*, 2024.

[Iacovides *et al.*, 2024] Giorgos Iacovides, Thanos Konstantinidis, Mingxue Xu, and Danilo Mandic. Finllama: Llm-based financial sentiment analysis for algorithmic trading. In *ICAIF*, pages 134–141, 2024.

[Kamaruddin and Ravi, 2016] SK Kamaruddin and Vadlamani Ravi. Credit card fraud detection using big data analytics: use of psoaann based one-class classification. In *ICIA*, pages 1–8, 2016.

[Kim *et al.*, 2024] Alex Kim, Maximilian Muhn, and Valeri Nikolaev. Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*, 2024.

[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 *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. *arXiv preprint arXiv:2301.09279*, 2023.

[Lee *et al.*, 2024] 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.

[Li *et al.*, 2019] Ziran Li, Ning Ding, Zhiyuan Liu, Haitao Zheng, and Ying Shen. Chinese relation extraction with multi-grained information and external linguistic knowledge. In *ACL*, pages 4377–4386, 2019.

[Li *et al.*, 2023a] Haozhou Li, Qinke Peng, Xu Mou, Ying Wang, Zeyuan Zeng, and Muhammad Fiaz Bashir. Abstractive financial news summarization via transformer-bilstm encoder and graph attention-based decoder. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[Li *et al.*, 2023b] Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. Cfgpt: Chinese financial assistant with large language model. *arXiv preprint arXiv:2309.10654*, 2023.

[Li *et al.*, 2024a] Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, Koduvayur Subbalakshmi, Guojun Xiong, et al. Investorbench: A benchmark for financial decision-making tasks with llm-based agent. *arXiv preprint arXiv:2412.18174*, 2024.

[Li *et al.*, 2024b] Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, Jun Huang, and Wei Lin. Alphafin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. *arXiv preprint arXiv:2403.12582*, 2024.

[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 *IJCAI*, pages 4513–4519, 2021.

[Lu *et al.*, 2024] Yiming Lu, Yebowen Hu, Hassan Foroosh, Wei Jin, and Fei Liu. Strux: An llm for decision-making with structured explanations. *arXiv preprint arXiv:2410.12583*, 2024.

[Luo *et al.*, 2025] Yichen Luo, Yebo Feng, Jiahua Xu, Paolo Tasca, and Yang Liu. Llm-powered multi-agent system for automated crypto portfolio management. *arXiv preprint arXiv:2501.00826*, 2025.

[Maia *et al.*, 2018] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: financial opinion mining and question answering. In *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. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

[Mukherjee *et al.*, 2022] Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, et al. Ectsum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *EMNLP*, pages 10893–10906, 2022.

[Nie *et al.*, 2024] Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*, 2024.

[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.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[Sang and De Meulder, 2003] Erik Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *CoNLL*, pages 142–147, 2003.

[Shah *et al.*, 2023a] Agam Shah, Suvan Paturi, and Sudheer Chava. Trillion dollar words: A new financial dataset, task & market analysis. *arXiv preprint arXiv:2305.07972*, 2023.

[Shah *et al.*, 2023b] Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and Sudheer Chava. Finer: Financial named entity recognition dataset and weak-supervision model. *arXiv preprint arXiv:2302.11157*, 2023.

[Sharma *et al.*, 2022] Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. Finred: A dataset for relation extraction in financial domain. In *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 *FICC*, volume 2, 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 *Big Data*, pages 1691–1700. IEEE, 2022.

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

[Wu *et al.*, 2018] Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. Hybrid deep sequential modeling for social text-driven stock prediction. In *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.

[Xiao *et al.*, 2024] Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents llm financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024.

[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 *NIPS*, pages 33469–33484, 2023.

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

[Xie *et al.*, 2024b] Qianqian Xie, Jimin Huang, Dong Li, Zhengyu Chen, Ruoyu Xiang, Mengxi Xiao, Yangyang Yu, Vijayasai Somasundaram, Kailai Yang, Chenhan Yuan, et al. Finnlp-agentscen-2024 shared task: Financial challenges in large language models-finllms. In *FinNLP*, pages 119–126, 2024.

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

[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] 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.

[Yang *et al.*, 2024] Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo, Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou, Mao Guan, Runjia Zhang, et al. Finrobot: An open-source ai agent platform for financial applications using large language models. *arXiv preprint arXiv:2405.14767*, 2024.

[Yin *et al.*, 2023] Yuwei Yin, Yazheng Yang, Jian Yang, and Qi Liu. Finpt: Financial risk prediction with profile tuning on pretrained foundation models. *arXiv preprint arXiv:2308.00065*, 2023.

[Yu *et al.*, 2023] Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. Temporal data meets llm–explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*, 2023.

[Yu *et al.*, 2024a] Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. Finmem: A performance-enhanced llm trading agent with layered memory and character design. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 595–597, 2024.

[Yu *et al.*, 2024b] Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W Suchow, Zhenyu Cui, Rong Liu, et al. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In *NIPS*, 2024.

[Zhang *et al.*, 2024a] Chong Zhang, Xinyi Liu, Mingyu Jin, Zhongmou Zhang, Lingyao Li, Zhengting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, et al. When ai meets finance (stockagent): Large language model-based stock trading in simulated real-world environments. *arXiv preprint arXiv:2407.18957*, 2024.

[Zhang *et al.*, 2024b] Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, et al. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *KDD*, pages 4314–4325, 2024.