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

Yifei Dong[1,*], **Fengyi Wu**[1,*], **Kunlin Zhang**[1], **Yilong Dai**[1], **Sanjian Zhang**[1], **Wanghao Ye**[2], **Sihan Chen**[3], **Zhi-Qi Cheng**[1,†]

[1]University of Washington, [2]University of Maryland, [3]Carnegie Mellon University

[*]Equal contribution. [†]Corresponding author.

Abstract

Large language models (LLMs) are increasingly applied to finance, yet challenges remain in aligning their capabilities with real-world institutional demands. In this survey, we provide a systematic, dual-perspective review bridging financial practice and LLM research. From a *practitioner-centric* standpoint, we introduce a functional taxonomy covering five core financial domains—*Data Analysis*, *Investment Research*, *Trading*, *Investment Management*, and *Risk Management*—mapping each to representative tasks, datasets, and institutional constraints. From a *research-focused* perspective, we analyze key modeling challenges, including numerical reasoning limitations, prompt sensitivity, and lack of real-time adaptability. We comprehensively catalog over 30 financial benchmarks and 20 representative models, and compare them across modalities, tasks, and deployment limitations. Finally, we identify open challenges and outline emerging directions such as continual adaptation, coordination-aware multi-agent systems, and privacy-compliant deployment. 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.

However, effective deployment of LLMs in financial workflows demands more than synthesizing large-scale data, given the complex and interdependent structure of modern financial institutions (Lo, 2019). They comprise multiple departments—*Data Analysis*, *Investment Research*, *Trading*, *Investment Management*, and *Risk Management* (Eccles and Crane, 1988; Lo, 2019)—each fulfilling interdependent roles and subtasks, 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.

Although LLMs have demonstrated strong performance on some subtasks such as *Text Summarization*, *Named Entity Recognition*, *Time Series Forecasting*, and *Fraud Detection*, they still face systemic obstacles: benchmarks remain static and unimodal, model architectures struggle with numerical reasoning and long-horizon logic, and multi-agent systems exhibit fragility under real-world stress. Furthermore, privacy and compliance remain underexplored—most pipelines rely on centralized data and lack built-in regulatory auditing mechanisms (Zhao et al., 2025; Yao et al., 2024; Nie et al., 2024; Chen et al., 2024).

To address the gap between cutting-edge LLM research and concrete financial practice needs, we propose a dual-perspective–*practitioner-centric* and *research-focused*–framework:

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

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17889–17907
November 4-9, 2025 ©2025 Association for Computational Linguistics

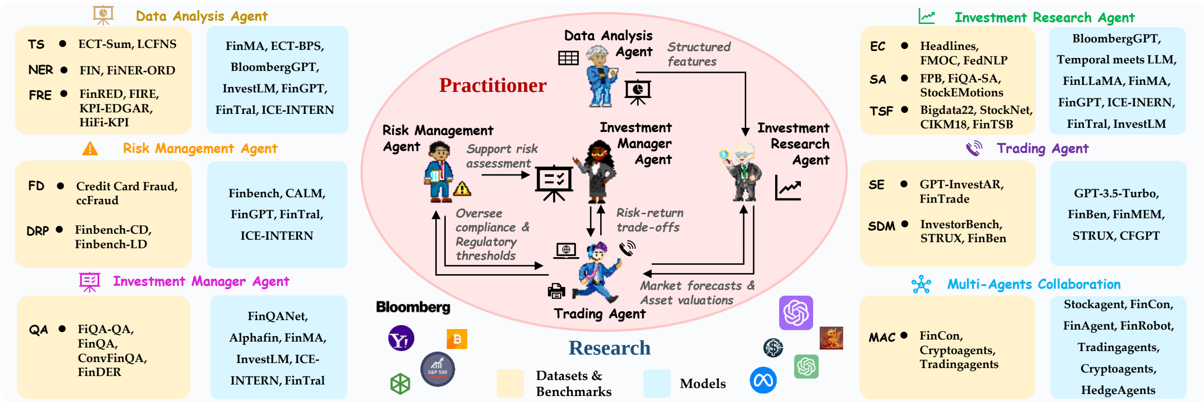

Figure 1: **Overview of LLM-based financial agents and their collaborative workflows.** Modern financial institutions rely on multiple departments—*Data Analysis*, *Investment Research*, *Trading*, *Investment Management*, and *Risk Management*—each handling specialized but interdependent roles, see pseudocode for each agent in Appx. B.5. 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].

• **Practitioner-Centric Perspective:** We present a taxonomy (Section 2) mapping core financial roles—*Data Analysis*, *Investment Research*, *Trading*, *Investment Management*, and *Risk Management*—to primary sub-tasks, datasets, and evaluation metrics. This approach reveals pressing challenges such as regulatory adherence, heterogeneous data integration, and multifaceted interdepartmental workflows, enabling a more grounded application of LLMs in real-world finance.

• **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 shown in Tables 1 and 2, these methods underscore the interplay between financial decision-making and emerging LLM paradigms, illuminating key technical gaps.

Unlike prior surveys (Lee et al., 2024; Nie et al., 2024; Chen et al., 2024) that focus on discrete tasks or narrowly defined benchmarks while mainly adopting a single perspective from LLMs, our work adopts a holistic, practitioner-oriented viewpoint (detailed related surveys comparison in Appx. A). This dual-perspective viewpoint allows us to synthesize over 30 benchmarks and 20 models across structured and unstructured modalities, and to contextualize technical progress within the real-world financial environment. We conclude our paper by discussing existing challenges and future research directions in this emerging and promising field.

2 Taxonomy of LLM-based Agents in Finance

Agent–Finance Taxonomy Alignment. To ensure the practical relevance of our agent taxonomy, we verify its consistency with established financial workflows (Appx. B). Financial institutions typically operate through five specialized divisions (Eccles and Crane, 1988; Lo, 2019): data analytics departments transform unstructured information into structured insights; research divisions generate investment theses and forecasts; trading operations execute market transactions; investment management teams make strategic allocation decisions; and risk management divisions ensure regulatory compliance and stability. This creates a consistent workflow where processed data becomes research insights, driving trades and portfolio strategies while undergoing continuous risk monitoring.

Our agent taxonomy mirrors this structure: *Data Analysis Agent* corresponds to financial data processing teams; *Investment Research Agent* to research departments; *Trading Agent* to trading desks; *Investment Manager Agent* to portfolio managers; and *Risk Management Agent* to risk divisions. As shown in Figure 1, each agent specializes in tasks from unstructured data processing to market forecasting and portfolio optimization (formalized in Alg. A1). Tables 1 and 2 summarize datasets, benchmarks, evaluation metrics, and state-of-the-art models, concluding with an analysis of their limitations, while Table 3 demonstrates model architectures and training details; Table 4 details dataset sizes, collection periods, and sources.

Table 1: **Overview** of Data Analysis, Investment Research, and Trading agents, showing datasets (size, period, source), data types (text, tables, time series, reports), metrics, and LLM models. Highlights key challenges for real-world applications for datasets, benchmarks, and corresponding models. [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, Summarization Consistency	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), HiFi-KPI (Aavang et al., 2025)	**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	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), FinTSB (Hu et al., 2025)	**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.

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 (Alg. A2). 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. Data Analysis Agents typically address three core tasks—*text summarization* (TS), *named entity recognition* (NER), and *financial relation extraction* (FRE).

2.1.1 Tasks & Benchmarks

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, and 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. 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*).

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. To further advance hierarchical KPI extraction, the HiFi-KPI dataset (Aavang et al., 2025) introduces annotated financial reports focusing on layered KPI entity recognition. However, these benchmarks mainly feature static document snapshots. Incorporating temporal aspects and numeric ratios remains a challenge.

2.1.2 LLM-Based Models for 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.

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 (Alg. A3). 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

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, FOMC dataset (Shah et al., 2023a) includes meeting minutes, speeches, and press conferences (1996–2022), enabling classifications like “hawkish” or “dovish.” FedNLP (Lee et al., 2021) adds more than 1,000 speeches and 100 press conferences (2015–2020), while Headlines dataset (Sinha and Khandait, 2021) provides 11,412 annotated news headlines (2000–2019). 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. FinTSB (Hu et al., 2025) unifies live-data ingestion, extreme-event simulation, and cost modeling. Many benchmarks lack multi-asset coverage and fundamental factors (e.g., P/E ratios), limiting practical utility.

2.2.2 LLM-Based Models for Agents

Recent 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.

Table 2: **Overview** of Investment Manager, Risk Management, and Multi-Agent Collaboration tasks, showing datasets (size, period, source), data types (text, tables, time series, reports), metrics, and LLM models. Highlights key challenges for real-world applications for datasets, benchmarks, and corresponding models. [Best to zoom in].

Agent & Subtask	Datasets & Benchmarks	Modalities (Data Types)	Key Metrics	Representative Models	Limitations
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), FinDER(Choi et al., 2025)	**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., 2024c), 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 multimodal 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, Sharpe Ratio, Max Drawdown	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), HedgeAgents (Li et al., 2025)	**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.

2.3 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 (Alg. A4). 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

Strategy Execution (SE). 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 and cross-asset correlations, restricting their utility in broader market modeling and long-horizon strategy testing.

Support Decision-Making (SDM). SDM 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, 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 or liquidity thresholds.

2.3.2 LLM-Based Models for Agents

Recent advances in LLMs show promise for Trading Agents. FinMEM (Yu et al., 2024a) uses a memory-enhanced GPT-4-Turbo (OpenAI et al., 2023) architecture to adapt risk preferences to market volatility, though scalability and interpretability challenges persist. STRUX (Lu et al., 2024) converts earnings-call transcripts into concise tables and applies self-reflection to classify key facts, but depends heavily on transcript data, risking oversimplification when macro signals are missing.

2.4 Investment Manager Agent

Definition and Scope. The Investment Manager Agent oversees portfolio decisions to balance risk and return under regulatory mandates (Alg. A5).

Table 3: **Overview of Representative LLM-Based Models for Financial Agents.** The table summarizes key characteristics, including related subtasks, model architecture (including backbone, parameters, and deployment cost), training details, involved dataset and benchmarks, and key observations of techniques in finance.

Model	Subtasks	Architectures	Training Details	Dataset & Benchmarks	Key Observations
ECT-BPS (Mukherjee et al., 2022)	TS	**Backbone**: FinBERT-based SummaRuNNer, T5, **Cost**: 1 P100 GPU	Two-stage separate training with Adam	ECTSum corpus	**Innovations**: Extract-then-paraphrase approach, new benchmark dataset ECTSum. **Performance**: ROUGE-1/2/L: 0.467/0.307/0.514, BERTScore: 0.764, Num-Prec.: 0.916.
BloombergGPT (Wu et al., 2023)	NER, SA	**Backbone**: BLOOM with Unigram tokenizer, **Parameters**: 50.6B, **Cost**: 512 A100 GPUs	Trained from scratch on 569B tokens	ConvFinQA, FiQA-SA, FPB, Headline	**Innovations**: Domain-specific yet general-purpose LLM. **Performance**: ConvFinQA (EM): 0.43, FiQA SA (F1): 0.75, FPB (F1): 0.51, Headline (F1): 0.82.
FinMA (Xie et al., 2023)	TS, NER, EC, SA, TSF, QA	**Backbone**: LLaMA, **Parameters**: 7 / 30B, **Cost**: 8 A100 / 128 A100 GPUs	Fine-tuned with multi-task and multi-modal instructions	FIT (combining FPB, Headline, FinQA, Bigdata22, etc)	**Innovations**: Fine-tuning LLaMA for finance. **Performance**: F1: 0.88 / 0.87 on FPB and FiQA-SA, Acc: 0.87 on FPB, MCC: 0.04 on BigData22.
FinGPT (Yang et al., 2023a)	TS, EC, SA, TSF, SE, FD, DRP	**Backbone**: ChatGLM, LLaMA, **Cost**: $300 per training	LoRA and RL on stock prices	Twitter, SEC Filings, Earnings Calls, Yahoo Finance	**Innovations**: Full-stack open-source FinLLM framework with RL using stock price feedback.
FinPT (Yin et al., 2023)	FD, DRP	**Backbone**: Flan-T5-Base, **Parameters**: 220M, **Cost**: 2 A40 GPU	Fine-tune pretrained foundation models with the profile	Finbench-CD, Finbench-LD	**Innovations**: Profile tuning for risk prediction. **Performance**: Average F1-score 49.17 across all FinBench datasets.
CALM (Feng et al., 2023)	FD, DRP	**Backbone**: LLaMA2-chat, **Parameters**: 7B, **Cost**: 4 A800 GPUs	LoRA instruction tuning on 75K samples	Credit scoring datasets	**Innovations**: Credit and Risk Assessment LLM. **Performance**: Credit Scoring (F1=0.545), Fraud Detection (Mcc=0.172), Financial Distress (Mcc=0.031).
InvestLM (Yang et al., 2023b)	TS, NER, SA, QA	**Backbone**: LLaMA, **Parameters**: 65B	LoRA finetuning and Linear Rope Scaling	FPB, FOMC, etc.	**Innovations**: Small diverse instruction dataset. **Performance**: Micro-F1 0.80 on ESG and 0.71 on FPB, accuracy 0.29 on FinQA.
CFGPT (Li et al., 2023b)	SDM	**Backbone**: InternLM, **Parameters**: 7B, **Cost**: 8 A800 GPUs	Two-stage training, continued pre-training	Self-build	**Innovations**: CFAPP framework.
Temporal meets LLM (Yu et al., 2023)	EC, TSF	**Backbone**: GPT-4, Open LLaMA, **Parameters**: 13B	Zero/few-shot prompting, instruction tuning	NASDAQ-100	**Innovations**: Explainable time series forecasting. **Performance**: Weekly Binary Precision: 64.7%, Bin Precision: 30.7%, MSE: 21.0.
FinLLaMA (Iacovides et al., 2024)	EC, TSF	**Backbone**: LLaMA-2-7B, **Parameters**: 7B, **Cost**: 1 A100 GPU	Finetuning with LoRA	S&P 500 (2015–2021)	**Innovations**: Sentiment intensity quantification. **Performance**: 308.2% cumulative return, 45.0% annualized return, 2.4 Sharpe ratio, 18.6% annualized volatility.
ICE-INTERN (Hu et al., 2024)	TS, NER, FRE, EC, SA, SE, SDM, QA, FD	**Backbone**: InternLM, **Parameters**: 7B, **Cost**: 8 A100 GPUs	Instruction finetuning with QLoRA	Self-build	**Innovations**: First open-source Chinese-English bilingual financial LLM framework. **Performance**: Bilingual.Avg: 0.117, CLS.Avg: 0.563, PRE.Avg: 0.434, EXT.Avg: 0.465.
FinTral (Bhatia et al., 2024)	TS, FRE, EC, SA, SDM, QA, FD	**Backbone**: Mistral, **Parameters**: 7B, **Cost**: 4 A100 GPUs	LoRA pretraining and QLoRA fine-tuning	FinanceBench, SA, NER, etc.	**Innovations**: Multimodal financial understanding. **Performance**: FinanceBench 90.67% correct, Hallu- cinations Index: 0.97, Stock Movement Prediction: 0.54.
FinMEM (Yu et al., 2024a)	SDM	**Backbone**: GPT-4	Prompt engineering, data retention in memory module	TSLA, NFLX, AMZN, MSFT, COIN	**Innovations**: Trading agent with layered memory. **Performance**: CR 61.78%, SR 2.68, DV 2.95%, AV 46.86%, MDD 11.00% on TSLA.
STRUX (Lu et al., 2024)	SDM	**Backbone**: LLaMA-3-Instruct, **Parameters**: 8B	Fine-tuning SFT, RL with GPT-4o-mini generated data	NASDAQ 500, S&P 500 (2017–2024)	**Innovations**: Structured explanation framework with reflection. **Performance**: Accuracy: 25.55%, F1: 19.80% in stock investment.
StockGPT (Li et al., 2024c)	QA	**Backbone**: ChatGLM2-6B, **Parameters**: 6B, **Cost**: 1 A800 GPU	LoRA on financial reports with chain of thought	Chinese stock market	**Innovations**: Stock-Chain retrieval QA. **Performance**: 30.8% maximum return.
FinCon (Yu et al., 2024b)	MAC	**Backbone**: GPT-4-Turbo, **Parameters**: API-based	Prompt optimization with Conceptual Verbal Reinforcement (CVRF)	FinCon dataset	**Innovations**: CVRF for multi-agent strategy updates. **Performance**: CR > 57%, SR: 0.825 .
TradingAgents (Xiao et al., 2024)	MAC	**Backbone**: o1-preview, GPT-4o, GPT-4o mini, **Parameters**: API-based	Zero/few-shot prompting; role-based agent assignment	Tradingagents custom dataset	**Innovations**: Simulates trading firm workflows via structured agent roles. **Performance**: 23.21% CR, 24.90%, 26% on $AAPL.
Cryptoagents (Luo et al., 2025)	MAC	**Backbone**: ChatGPT-4o, **Parameters**: API-based	Few-shot prompting and weekly rebalancing	Cryptoagents custom dataset	**Innovations**: Multi-agent prompt voting for crypto. **Performance**: Accuracy 0.52 (crypto), 0.58 (market).
FinAgent (Zhang et al., 2024b)	MAC	**Backbone**: GPT-4-preview/4V-preview, **Parameters**: API-based	Dual-level reflection	5 U.S. stocks and ETH prices	**Innovations**: RL agent with memory and reflection. **Performance**: 92.2% ARR on TSLA.
HedgeAgents (Li et al., 2025)	MAC	**Backbone**: GPT-4-preview, **Parameters**: API-based	Collaborative meetings of multiple agents	Bitcoin and the Dow Jones component stocks	**Innovations**: Hierarchical multi-agent hedging with memory and conferences. **Performance**: ARR: 72%, TR: 405%.

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 task, institutional investors query large-scale financial datasets. FiQA-QA (Maia et al., 2018) provides 5,676 question–answer pairs drawn from financial news and microblogs, with relevance assessed using metrics like nDCG and MRR. FinQA (Chen et al., 2021) comprises 8,281 expert-annotated QA pairs derived from S&P 500 earnings reports, emphasizing numerical reasoning. In addition, ConvFinQA (Chen et al., 2022) extends QA to multi-turn dialogues, testing compositional reasoning across diverse 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 or compliance thresholds.

2.4.2 LLM-Based Models for Agents

Recent 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., 2024c) 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 under stress-test scenarios.

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 (Alg. A6). 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

Fraud Detection (FD). 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 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), making generalization across different financial systems challenging.

Default Risk Prediction (DRP). Assessing the likelihood of a borrower failing 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., Apr–Sep 2005 in Taiwan), integrating textual descriptors and tabular indicators (annual income, credit history length). However, these datasets rarely incorporate macro-level shifts such as interest rate changes or unemployment trends. Limited longitudinal tracking and a lack of cross-lender data further reduce applicability for evolving borrower behavior analysis and long-term risk modeling.

2.5.2 LLM-Based Models for Agents

Recent work employs 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 (OpenAI et al., 2023). Nevertheless, the reliance on static, labeled corpora and high computational demands hamper adaptation to shifting fraud schemes, while real-time scalability remains a significant hurdle.

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 (Alg. A1, Alg. A7). 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. This holistic approach addresses the complex challenges of modern finance (Table 2).

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

Table 4: **Overview of Representative Financial Datasets.** The table summarizes key characteristics—including raw data size, collection period, data sources, and license with data links (if open-source)—of datasets used by various LLM-based agents in finance. [Best to zoom in].

Agent	Subtask	Dataset	Raw Data Size	Collection Period	Data Source	License and Link
Data Analysis Agent	**TS**	**ECT-Sum**	2,425 document-summary pairs	Jan 2019 - Apr 2022	Earnings call transcripts, Reuters articles	GPL-3.0
		LCFNS	430,820 news-summary pairs	Jan 2013 - Jun 2020	Major financial portals	None Public
	NER	**FIN**	54,256 words (8 annotated agreements)	-	U.S. SEC filings, CoNLL-2003	MIT license
		FiNER-ORD	201 financial news articles, 4,739 sentences	Jul 2015 - Oct 2015	Webz.io	CC BY-NC 4.0
	FRE	**FinRED**	7,775 sentences, 29 relation types	Jul 2015 - Oct 2015, Jun 2019 - Sep 2019	Financial news articles, earnings calls	CC BY-NC 4.0
		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
		HiFi-KPI	1.8M paragraphs, 5M entities	Jan 2017 – Jun 2024	SEC iXBRL Filings	Public
Investment Research Agent	**EC**	**FOMC**	214 minutes, 1,026 speeches, 63 transcripts	1996 - 2022	Federal Open Market Committee communications	Public
		FedNLP	122 FOMC docs, 1,300 speeches	Jan 2015 - Jul 2020	Federal Reserve communications	Public
		Headlines	11,412 annotated news headlines	2000 - 2019	Gold commodity market	Public
	SA	**FPB**	4,840 sentences	-	Financial news articles	CC BY-NC 3.0
		FiQA-SA	529 annotated headlines and 774 financial microblogs	-	Financial news and social media	Public
		StockEmotions	10,000 investor comments, 12 emotions	Jan 2020 - Dec 2020	StockTwits	Public
	TSF	**StockNet**	26614 price movement data of 88 stocks	Jan 2014 - Jan 2016	StockTwits , Yahoo Finance	MIT license
		Bigdata22	272,762 tweets of 50 stocks	Jul 2019 – Jun 2020	US high-trade-volume 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	3,384 samples (stock prices, 10-K/10-Q filings, news)	One year period	10 stocks (Yahoo Finance, SEC EDGAR, public 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	CC BY-NC-ND 4.0
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	1999 - 2019	Earnings reports (S&P 500)	MIT License
		ConvFinQA	3,892 conversations, 14,115 questions	1999 - 2019	Earnings reports (S&P 500)	MIT License
		FinDER	5,703 Triples	2023 - 2024	SEC EDGAR	None Public
Risk Management Agent	**FD**	**Credit Card Fraud**	11,392 transactions (train+test)	2013	European cardholders	DbCL v1.0
		ccFraud	10,485 transactions (train+test)	2013	credit card transactions	Public
	DRP	**Finbench-CD**	30k credit records	Apr 2005 - 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	Jan 2022 - Jun 2023	Yahoo Finance, Form 10-Q/10-K, Zacks Rank, Earnings conference calls	None Public
		Tradingagents	Data size not specified	Jan 2024 - Mar 2024	S&P 500 stocks, Bloomberg, Yahoo, Reddit, Twitter	None Public
		Cryptoagents	Top 30 cryptocurrency by market cap	Jun 2023 - Sep 2024	Blockchain.info, Coin Metrics, Cointelegraph	None Public

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 like bid-ask spreads and execution latencies.

2.6.2 LLM-Based Models for Agents.

Recent work uses LLMs to incorporate 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 historical actions and sentiment analysis. FinCon (Yu et al., 2024b) applies a hierarchical manager–analyst structure with daily Conditional Value at Risk monitoring and multi-episode refinement. Tradingagents (Xiao et al., 2024) and Cryptoagents (Luo et al., 2025) deploy specialized roles for institutional trading and digital assets, respectively. HedgeAgents (Li et al., 2025) coordinates fund management through conference mechanisms, while budget allocation research (Cardi et al., 2025) optimizes resource distribution. Despite their innovations, challenges still remain in prompt sensitivity, LLM biases, and high-frequency trading.

3 Challenges and Future Directions

3.1 Challenges

Benchmark Limitations. Despite the rise of benchmarks for financial LLM agents, several critical limitations persist: *(1). Lack of real-time adaptability.* Most benchmarks rely on historical archives that fail to capture real-time market dynamics, including volatility, policy changes, and shifting regulatory thresholds (Chen et al., 2021, 2022). *(2). Insufficient structured-unstructured integration.* Structured and unstructured modalities are treated independently, tasks such as *TS*, *NER*, and *FRE* are typically addressed in isolation, hindering holistic data interpretation (Mukherjee et al., 2022; Deußer et al., 2022). *(3). Limited coverage of scenarios. NER, FRE* datasets such as FIN and FinRED (Sharma et al., 2022) only support a narrow set of entity types (Section 2.1), while *SE, SDM* benchmarks remain constrained to single-asset scenarios (Section 2.3).

Model Design Challenges. Financial LLM systems still face core limitations: *(1). Weak numerical reasoning and multi-step logic.* Financial LLMs struggle with arithmetic chaining and compositional logic essential for *QA* and *TSF* tasks (Sections 2.2, 2.4). Output uncertainty and computational complexity compound over multi-turn interactions, weakening long-horizon planning (Cardi et al., 2025). *(2). Lack of adaptability to market shifts.* Most financial LLMs, such as (Yang et al., 2023a; Yu et al., 2024a), are fine-tuned offline and remain static. This undermines performance under market shifts (Sections 2.2–2.3). Real-world trading demands ultra-low latency and adaptability to market microstructure dynamics such as bid-ask spreads and liquidity constraints (Gupta, 2023; Xie et al., 2024a; Cheng et al., 2024b). *(3). Coordination issues in multi-agent systems.* Multi-agent frameworks suffer from prompt sensitivity and poor robustness under stress. Conflicting outputs with ambiguous cross-departmental data (Section 2.6) lead to degraded strategy alignment (Yu et al., 2024b; Luo et al., 2025) and introduce systemic risk, necessitating diversity-promoting coordination strategies (Nie et al., 2024; Zhang et al., 2024a; Yu et al., 2024b). *(4). Privacy and Compliance.* FinLLMs remain vulnerable to privacy breaches and regulatory gaps through centralized data handling practices (Nie et al., 2024).

3.2 Future Directions

Advancing Datasets & Benchmarks. To overcome current limitations in benchmark design—such as static data, modality gaps, and narrow coverage—future work should consider *(1).* Evaluating models under authentic market conditions across different states (normal, volatile, crisis events) (Nie et al., 2024), measuring performance variations and response speed. *(2).* Promoting multimodal benchmarks integrating seamlessly structured (e.g., financial indicators, tables) and unstructured data (e.g., filings, news) for complex tasks like *TS*, *NER*, and *FRE* (Lee et al., 2024; Xie et al., 2024a). *(3).* Developing temporal relationship modeling that extends *FinRED* and *FIRE*'s static approaches with timeline-aware annotations (Sharma et al., 2022; Hamad et al., 2024), scaling strategy execution frameworks from single-company limitations in GPT-InvestAR and FinTrade to comprehensive cross-asset coverage (Gupta, 2023; Xie et al., 2024a), and extending decision-making benchmarks to integrated multi-asset frameworks that capture correlation structures (Li et al., 2024a; Lu et al., 2024).

Improving Model Robustness and Adaptability. To address the former four challenges, future financial LLM agents could *(1).* Implement uncertainty-aware reasoning with error propagation tracking and excessive uncertainty verification modules (Blasco et al., 2024). Manage computational complexity through heuristic pruning (Cardi et al., 2025). *(2).* Apply diversity regularizers to agent behaviors to prevent synchronized actions and reduce systemic herd risk (Wang et al., 2023). Combine change-point detection to trigger rapid model adaptation when market regimes shift. *(3).* Equip agents with self-reflection (Bo et al., 2024), hierarchical messaging (shared memory, sequential communication), dynamic coalition formation during stress, and lightweight consensus protocols for high-risk decisions (Hooper et al., 2009). *(4).* Adopt privacy-preserving, compliant learning by deploying federated-learning frameworks alongside simulated-attack benchmarks (Zhao et al., 2025), and embedding executable regulatory rules via real-time compliance-auditor agents (Yao et al., 2024; Masoudifard et al., 2024).

4 Conclusion

We present this survey that systematically analyzes the deployment of large language model (LLM) agents across core financial functions, including Data Analysis, Investment Research, Trading, Investment Management, and Risk Management. For each functional division, we introduce representative subtasks, curated datasets, and state-of-the-art LLM-based solutions, along with their practical constraints in real-world finance. To support broader adoption, we also catalog benchmark datasets covering diverse modalities and detail their coverage, licensing, and evaluation metrics. Concluding the paper, we outline persistent challenges and emerging directions, including real-time adaptation, uncertainty-aware reasoning, and coordination among heterogeneous agents for future research in LLM-empowered financial AI.

Limitations

While this survey presents a comprehensive mapping of financial agents, tasks, datasets, and modeling approaches, it remains a descriptive and analytical study without conducting controlled empirical experiments. As such, our insights rely on reported results from existing literature. Moreover, although our agent framework is grounded in real-world institutional structures, we do not validate its effectiveness through deployment or benchmarking in operational environments, as our goal is to provide a conceptual and systematic overview rather than propose a specific implementable system. Given the survey nature and scope constraints, we leave empirical validations to future work.

Acknowledgements

This work was supported by the Graduate School Fund for Excellence & Innovation (GSFEI) Top Scholar Award from the University of Washington Graduate School, by the Vicky L. Carwein and William B. Andrews Endowment for Graduate Programs at UW Tacoma's School of Engineering & Technology, and by the faculty startup funding from the University of Washington.

References

Rasmus Aavang, Giovanni Rizzi, Rasmus Bøggild, Alexandre Iolov, Mike Zhang, and Johannes Bjerva. 2025. Hifi-kpi: A dataset for hierarchical kpi extraction from earnings filings. *Preprint*, arXiv:2502.15411.

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

Douglas W Arner, Janos Barberis, and Ross P Buckley. 2019. The evolution of fintech: A new post-crisis paradigm. *Georgetown Journal of International Law*, 47:1271–1319.

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

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

Txus Blasco, J. Salvador Sánchez, and Vicente García. 2024. A survey on uncertainty quantification in deep learning for financial time series prediction. *Neurocomputing*, 576:127339.

Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. 2024. Reflective multi-agent collaboration based on large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 138595–138631. Curran Associates, Inc.

Pierre Cardi, Laurent Gourvès, and Julien Lesca. 2025. On fair and efficient solutions for budget apportionment. *Autonomous Agents and Multi-Agent Systems*, 39(1):1–31.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting Hao Huang, Bryan Routledge, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv:2210.03849*.

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

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024a. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853.

Zhi-Qi Cheng, Yifei Dong, Aike Shi, Wei Liu, Yuzhi Hu, Jason O'Connor, Alexander G Hauptmann, and Kate S Whitefoot. 2024b. Shield: Llm-driven schema induction for predictive analytics in ev battery supply chain disruptions. *arXiv preprint arXiv:2408.05357*.

Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy yong Sohn, and Alejandro Lopez-Lira. 2025. Finder: Financial dataset for question answering and evaluating retrieval-augmented generation. *Preprint*, arXiv:2504.15800.

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

Yifei Dong, Fengyi Wu, Qi He, Heng Li, Minghan Li, Zebang Cheng, Yuxuan Zhou, Jingdong Sun, Qi Dai, Zhi-Qi Cheng, and 1 others. 2025. Ha-vln: A benchmark for human-aware navigation in discrete-continuous environments with dynamic multi-human interactions, real-world validation, and an open leaderboard. *arXiv preprint arXiv:2503.14229*.

Robert G. Eccles and Dwight B. Crane. 1988. *Doing Deals: Investment Banks at Work*. Harvard Business School Press, Boston, MA.

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

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

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

Daylond J. Hooper, Gilbert L. Peterson, and Brett J. Borghetti. 2009. Dynamic coalition formation under uncertainty. In *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IROS'09, page 4799–4804. IEEE Press.

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

Yifan Hu, Yuante Li, Peiyuan Liu, Yuxia Zhu, Naiqi Li, Tao Dai, Shu-tao Xia, Dawei Cheng, and Changjun Jiang. 2025. Fintsb: A comprehensive and practical benchmark for financial time series forecasting. *arXiv preprint arXiv:2502.18834*.

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

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

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

Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. 2024. A survey of large language models in finance (finllms). *arXiv preprint arXiv:2402.02315*.

Jean Lee, Hoyoul Luis Youn, Josiah Poon, and Soyeon Caren Han. 2023. Stockemotions: Discover investor emotions for financial sentiment analysis and multivariate time series. *arXiv preprint arXiv:2301.09279*.

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

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, Koduvayur Subbalakshmi, Guojun Xiong, Jimin Huang, Lingfei Qian, Xueqing Peng, Qianqian Xie, and Jordan W. Suchow. 2024a. Investorbench: A benchmark for financial decision-making tasks with llm-based agent. *arXiv preprint arXiv:2412.18174*.

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

Heng Li, Minghan Li, Zhi-Qi Cheng, Yifei Dong, Yuxuan Zhou, Jun-Yan He, Qi Dai, Teruko Mitamura, and Alexander G Hauptmann. 2024b. Human-aware vision-and-language navigation: Bridging simulation to reality with dynamic human interactions. *Advances in Neural Information Processing Systems*, 37:119411–119442.

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

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

Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2025. Hedgeagents: A balanced-aware multi-agent financial trading system. *arXiv preprint arXiv:2502.13165*.

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

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

Andrew W Lo. 2019. Adaptive markets: Financial evolution at the speed of thought. *Journal of Investment Management*, 17(1):1–44.

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

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

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *WWW*, pages 1941–1942.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Arsalan Masoudifard, Mohammad Mowlavi Sorond, Moein Madadi, Mohammad Sabokrou, and Elahe Habibi. 2024. Leveraging graph-rag and prompt engineering to enhance llm-based automated requirement traceability and compliance checks. *Preprint*, arXiv:2412.08593.

Niamh Moloney. 2019. *EU Securities and Financial Markets Regulation*, 3rd edition. Oxford University Press, Oxford.

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

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

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

Raksha Ramesh, Vishal Anand, Zifan Chen, Yifei Dong, Yun Chen, and Ching-Yung Lin. 2022. Leveraging text representation and face-head tracking for long-form multimodal semantic relation understanding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7215–7219.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023a. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679, Toronto, Canada. Association for Computational Linguistics.

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

Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. 2022. Finred: A dataset for relation extraction in financial domain. In *WWW*, pages 595–597.

Ankur Sinha and Tanmay Khandait. 2021. Impact of news on the commodity market: Dataset and results. In *FICC*, volume 2, pages 589–601. Springer.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jiangxing Wang, Deheng Ye, and Zongqing Lu. 2023. Mutual-information regularized multi-agent policy iteration. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Fengyi Wu, Yifei Dong, Zhi-Qi Cheng, Yilong Dai, Guangyu Chen, Hang Wang, Qi Dai, and Alexander G Hauptmann. 2025. Govig: Goal-conditioned visual navigation instruction generation. *arXiv preprint arXiv:2508.09547*.

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

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

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

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others. 2024a. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: a large language model, instruction data and evaluation benchmark for finance. In *NIPS*, pages 33469–33484.

Qianqian Xie, Xiao-yang Liu, Yangyang Yu, Dong Li, Benyou Wang, Alejandro Lopez-Lira, Yanzhao Lai, Min Peng, Sophia Ananiadou, Hao Wang, Jimin Huang, Zhengyu Chen, Ruoyu Xiang, VijayaSai Somasundaram, Kailai Yang, Chenhan Yuan, Zheheng Luo, Zhiwei Liu, Yueru He, and 3 others. 2024b. Finnlp-agentscen-2024 shared task: Financial challenges in large language models-finllms. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 119–126.

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

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

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

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023b. Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*.

Xu Yao, Xiaoxu Wu, Xi Li, Huan Xu, Chenlei Li, Ping Huang, Si Li, Xiaoning Ma, and Jiulong Shan. 2024. Smart audit system empowered by llm. *Preprint*, arXiv:2410.07677.

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

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

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. 2024a. 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.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Rong Liu, Zhenyu Cui, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024b. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In *NeurIPS*.

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

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

Joshua Zhao, Saurabh Bagchi, Salman Avestimehr, Kevin Chan, Somali Chaterji, Dimitris Dimitriadis, Jiacheng Li, Ninghui Li, Arash Nourian, and Holger Roth. 2025. The federation strikes back: A survey of federated learning privacy attacks, defenses, applications, and policy landscape. *ACM Computing Surveys*, 57(9):1–37.

Appendix

Appendix Contents

A Related Survey Comparison

As shown in Table A1, our survey makes several unique contributions while acknowledging certain inherent limitations in studying the rapidly evolving intersection of LLMs and finance. Unlike previous surveys that adopt a single perspective from LLM (Nie et al., 2024), our work uniquely bridges theory and practice through a dual-perspective framework, offering both practitioner-centric insights and research-focused analysis. This comprehensive approach enables us to thoroughly address finance orientation, datasets, benchmarks, applications, and challenges—areas where prior works like (Lee et al., 2024) and (Chen et al., 2024) showed only partial coverage. The practitioner-centric perspective provides concrete value by mapping financial roles to specific tasks, datasets, and metrics, making our findings directly applicable to real-world institutional finance.

B Detailed Financial Industry Practices and Agent Framework Alignment

This appendix provides additional details on financial industry practices and how they align with the LLM agent-based framework, expanding on the validation presented in Section 2.

B.1 Financial Institution Organization

Financial institutions have developed highly specialized departmental structures to manage complex information processing and decision-making requirements. These structures exhibit remarkable consistency across different types of institutions, from investment banks to asset managers:

Data and Analytics Departments form the foundation of financial institutions, processing vast quantities of structured and unstructured information from multiple sources. Bloomberg processes "millions of pieces of financial data a second" at market peaks (Wu et al., 2023), while J.P. Morgan has dedicated data teams that transform raw inputs into standardized formats for downstream consumption. These departments typically organize around three core functions that align with Data Analysis Agent: document processing (corresponding to text summarization task), entity identification (corresponding to named entity recognition), and relationship mapping (corresponding to financial relation extraction).

Research Departments generate insights that drive investment decisions. Goldman Sachs' Global Investment Research provides coverage across thousands of securities and dozens of economies (Shah et al., 2023a). Research departments typically classify market events (aligned with event classification task), assess sentiment from corporate communications (matching sentiment analysis task), and develop forecasts (corresponding to time series forecasting task). Lee et al. (2021) documents how financial research departments process Federal Reserve communications using methods that match Investment Research Agent's functions.

Trading Operations execute market transactions based on research insights and portfolio requirements. Xie et al. (2024a) demonstrate how trading desks incorporate both human judgment and algorithmic execution in processes that mirror Trading Agent's capabilities. Modern trading desks typically separate into two functional areas: execution mechanisms (corresponding to strategy execution task) and decision support systems (matching support decision-making task). Gupta (2023) documents how these functions operate in conjunction, with overlap with our surveyed framework.

Portfolio Management Teams make strategic asset allocation decisions within risk parameters. BlackRock, managing over $11.5 trillion in assets as of Q1 2025, organizes portfolio managers into specialized teams that develop investment theses and monitor performance. These teams consistently employ question-answering frameworks to evaluate investment opportunities, as in Chen et al. (2022) analysis of conversational financial QA systems. This validates Investment Manager Agent's QA functionality and demonstrates the centrality of this task in portfolio management processes.

Risk Management Divisions assess exposure across multiple dimensions to protect institutional stability. Yin et al. (2023) analyze how risk functions identify and mitigate various risks—functions encapsulated in Risk Management Agent. Financial institutions typically organize risk departments into specialized units focused on transaction monitoring (corresponding to fraud detection task) and credit assessment (matching default risk prediction task). Feng et al. (2023) documents how these functions operate in modern financial institutions, confirming alignment with LLM agent framework.

B.2 Detailed Agent-to-Function Mapping

Financial LLM agent framework maps to industry functions with a high degree of precision, as evidenced by detailed academic studies:

Table A1: Comparison between ours and related surveys. Half-correct indicates areas covered but lacking detail.

Survey Paper	Finance Oriented	Datasets & Benchmarks	Application	Challenges	Perspective
Lee *et al.* (Lee et al., 2024)	✓	✓	✓✗	✓✗	Single
Chen *et al.* (Chen et al., 2024)	✕	✓	✓	✓✗	Single
Nie *et al.* (Nie et al., 2024)	✓	✓	✓	✓✗	Single
Ours	✓	✓	✓	✓	Dual

Data Analysis Agent: Shah et al. (2023b) conducted a comprehensive analysis of financial data processing teams and Sharma et al. (2022) further documents how financial relation extraction is implemented in practice. Annual reports and earnings calls typically undergo processing that aligns precisely with LLM agent's workflow, beginning with summarization, proceeding through entity extraction, and culminating in relationship mapping (Deußer et al., 2022).

Investment Research Agent: Malo et al. (2014) analyzed financial sentiment analysis practices across institutional research departments. Their research demonstrated that financial analysts perform sentiment analysis on earnings calls. Sinha and Khandait (2021) similarly documented event classification practices in financial research, showing how analysts categorize market-moving events using approaches that align with the LLM agent framework. Time series forecasting methods in financial institutions Yu et al. (2023) exhibit striking similarities to the approach of LLM agents.

Trading Agent: A detailed study by Lu et al. (2024) examined trading desk operations across financial institutions, finding organizational structures that directly parallel Trading Agent design. Xie et al. (2024a) further documented how trading algorithms incorporate both execution mechanics and decision frameworks.

Investment Manager Agent: Chen et al. (2021) conducted extensive research on question-answering systems in portfolio management, analyzing how investment teams formulate and address complex financial questions. They demonstrate that the question-answering process in portfolio management is consistent with the LLM agent's design.

Risk Management Agent: Feng et al. (2023) surveyed risk management practices across financial institutions, documenting approaches to fraud detection and default risk prediction that align with Risk Management Agent. Kamaruddin and Ravi (2016) similarly documented how transaction monitoring and credit assessment operate in practice.

B.3 Multi-Agent Collaboration in Practice

The coordination mechanisms we survey in the multi-agent framework find direct parallels in financial institution practices:

Investment Committees: Xiao et al. (2024) analyzed how investment committees coordinate inputs from research, trading, portfolio management, and risk departments. Their research documented information flows with specialized units providing inputs that inform collective decision-making.

Morning Strategy Meetings: Zhang et al. (2024a) documented how daily strategy meetings coordinate activities across departments. Their research showed how insights flow from data analysis to research, from research to trading, and from trading to portfolio management—a pattern.

Risk Review Processes: Luo et al. (2025) analyzed how risk oversight functions interact with other departments. Their research demonstrated coordination patterns consistent with LLM agent framework, with risk considerations flowing back to inform portfolio decisions and trading actions.

B.4 Limitations in the Financial Industry

While LLM-based agents show promising potential in finance, several domain-specific (Cheng et al., 2024a; Wu et al., 2025) and sim-to-real (Li et al., 2024b; Dong et al., 2025) challenges require careful attention and targeted solutions. Financial institutions operate under strict regulatory frameworks (Basel III, MiFID II, Dodd-Frank) that demand transparent, auditable decision-making processes (Moloney, 2019; Arner et al., 2019), creating opportunities for developing explainable AI techniques tailored to regulatory compliance (Feng et al., 2023; Chen et al., 2024). The ultra-low latency requirements and complex market microstructure dynamics of financial markets—including bid-ask spreads, liquidity constraints, and execution costs—present technical challenges that could be addressed through optimized architectures and specialized training approaches (Gupta, 2023; Xie et al., 2024a; Wu et al., 2023). The interconnected nature of financial markets raises important ques-

Algorithm A1 Financial LLM Multi-Agent System

```
procedure FinSysTem(data, query, params)
    Initialize agents
    struct ← DataAgent(data)
    insight ← ResearchAgent(struct)
    strat ← TradeAgent(insight, params)
    port ← PortfolioAgent(strat, query)
    risk ← RiskAgent(port)
    if risk.level > params.threshold then
        Revise port based on risk
    end if
    return {port, risk}
end procedure
```

tions about systemic risks from correlated algorithmic behavior (Nie et al., 2024; Zhang et al., 2024a; Yu et al., 2024b), suggesting the need for coordination mechanisms and diversity requirements in deployment strategies. Current benchmarks and evaluation frameworks predominantly focus on single-asset scenarios with historical data (Li et al., 2024a; Chen et al., 2021), highlighting opportunities to develop more comprehensive multi-asset, real-time evaluation methodologies that better reflect institutional trading environments. Additionally, financial markets' structural regime changes and the inherent need for human judgment in client relationships and ethical considerations point toward promising research directions in adaptive learning systems and human-AI collaboration frameworks. While these challenges (Ramesh et al., 2022) are substantial, they represent important areas for future research that could unlock the full potential of LLMs in financial applications through domain-specific innovations and responsible deployment practices.

B.5 Pseudocode for Financial LLM Agents

Financial LLM Multi-Agent System (Alg. A1) orchestrates the entire workflow by coordinating specialized agents. It begins by processing raw data through the Data Analysis Agent, then passes structured information to the Research Agent for insight generation. These insights inform the Trading Agent's strategy development, which then feeds into Portfolio Agent's allocation decisions. Finally, a Risk Agent evaluates these decisions, prompting revisions if risk thresholds are exceeded.

Data Analysis Agent (Alg. A2) transforms unstructured financial data into structured insights

Algorithm A2 Data Analysis Agent

```
procedure DataAgent(raw)
    proc ← {}
    sum ← Summarize(raw.docs)
    proc.sum ← sum
    ent ← ExtractEntities(raw.docs)
    proc.ent ← ent
    rel ← ExtractRelations(raw.docs, ent)
    proc.rel ← rel
    final ← Integrate(proc, raw.struct)
    return final
end procedure
procedure Summarize(docs)
    Extract key info
    return summaries
end procedure
procedure ExtractEntities(docs)
    Identify financial entities
    return entity database
end procedure
procedure ExtractRelations(docs, ent)
    Find entity relationships
    return relationship graph
end procedure
```

through three core functions. The SUMMARIZE procedure distills key information from lengthy documents like earnings calls and financial reports. EXTRACTENTITIES identifies critical financial entities such as companies, regulators, and instruments. EXTRACTRELATIONS maps relationships between these entities, creating a graph structure. This agent's outputs form the foundation for downstream financial analysis, establishing standardized data representations from heterogeneous sources that other agents can effectively utilize.

Investment Research Agent (Alg. A3) analyzes structured data to generate actionable market insights. The CLASSIFYEVENTS procedure categorizes market-moving events like policy changes or earnings releases. ANALYZESENTIMENT evaluates opinions expressed in financial communications, extracting signal from noise. FORECAST integrates price patterns with text signals to predict market behavior. By merging these qualitative and quantitative analyses, this agent produces comprehensive market views that combine narrative context with numerical projections, directly supporting trading and portfolio management decisions.

Trading Agent (Alg. A4) translates research insights into executable trading strategies. EXECUTE procedure processes market data and generates spe-

Algorithm A3 Investment Research Agent

```
procedure ResearchAgent(data)
    insights ← {}
    events ← ClassifyEvents(data)
    insights.events ← events
    sentiment ← AnalyzeSentiment(data)
    insights.sentiment ← sentiment
    forecast ← Forecast(data)
    insights.forecast ← forecast
    merged ← Merge(insights)
    return merged
end procedure
procedure ClassifyEvents(d)
    Identify market events
    return classified events
end procedure
procedure AnalyzeSentiment(d)
    Extract opinion polarities
    return sentiment scores
end procedure
procedure Forecast(d)
    Combine price and text signals
    return predictions
end procedure
```

Algorithm A4 Trading Agent

```
procedure TradeAgent(insights, params)
    plan ← {}
    exec ← Execute(insights, params)
    plan.exec ← exec
    decide ← Support(insights, params)
    plan.decide ← decide
    optimal ← Optimize(plan, params)
    return optimal
end procedure
procedure Execute(i, p)
    Process market data
    Generate signals
    return execution plan
end procedure
procedure Support(i, p)
    Analyze assets
    Optimize allocation
    return framework
end procedure
```

cific buy/sell signals based on research insights and parameters like risk tolerance. SUPPORT analyzes assets and optimizes allocations, providing decision frameworks that adapt to changing market conditions. This agent balances algorithmic precision with adaptability, operating at junction between research insights and portfolio implementation, ensuring that strategies remain responsive to both systematic patterns and tactical opportunities.

Algorithm A5 Investment Manager Agent

```
procedure PortfolioAgent(strategy, query)
    p ← {}                                  ▷ Portfolio plan
    answers ← AnswerQuery(query, strategy)
    p.logic ← answers
    p.alloc ← Optimize(strategy, answers)
    p.metrics ← Measure(p.alloc)
    return p
end procedure
procedure AnswerQuery(q, s)
    Parse query components
    Apply numerical reasoning
    return answers with confidence
end procedure
procedure Optimize(s, a)
    Balance risk-return
    Apply portfolio constraints
    return optimized allocation
end procedure
```

Investment Manager Agent (Alg. A5) manages portfolio construction and optimization. The ANSWERQUERY procedure parses complex financial questions, applying numerical reasoning to address specific investment inquiries with confidence-scored responses. OPTIMIZE balances risk-return tradeoffs under portfolio constraints, converting strategic insights into concrete asset allocations. This agent encapsulates the core portfolio management function, combining quantitative optimization with explicable logic that maintains transparency across investment decisions while adhering to regulatory requirements and client mandates.

Risk Management Agent (Alg. A6) safeguards financial stability through risk assessment. DETECTFRAUD procedure analyzes transaction patterns to identify potential malfeasance. PREDICTDEFAULT evaluates creditworthiness across counterparties, incorporating both specific factors and broader macroeconomic indicators. CHECKCOMPLIANCE verifies adherence to regulatory frameworks and internal risk limits. This agent serves as final checkpoint before strategy implementation.

Multi-Agent Collaboration framework (Alg. A7) enables coordinated interaction among specialized financial agents. The procedure begins by decomposing complex tasks and assigning components to appropriate agents. The RESOLVE function han-

Algorithm A6 Risk Management Agent

procedure RISKAGENT($portfolio$)
    $risk \leftarrow \{\}$
    $fraud \leftarrow$ DETECTFRAUD($portfolio$)
    $risk.fraud \leftarrow fraud$
    $default \leftarrow$ PREDICTDEFAULT($portfolio$)
    $risk.default \leftarrow default$
    $risk.metrics \leftarrow$ RISKMETRICS($portfolio, fraud, default$)
    $risk.comply \leftarrow$ CHECKCOMPLIANCE($portfolio, risk$)
    return $risk$
end procedure
procedure DETECTFRAUD(p)
    Analyze transaction patterns
    return fraud score
end procedure
procedure PREDICTDEFAULT(p)
    Assess creditworthiness
    Include macro indicators
    return default risk
end procedure
procedure CHECKCOMPLIANCE(p, r)
    Verify regulations
    Check exposure limits
    return compliance status
end procedure

dles conflicts between agent outputs, weighting recommendations by domain expertise. SYNTHESIZE integrates cross-agent insights into a unified framework. This collaborative architecture mirrors institutional workflows, where cross-departmental coordination balances specialized expertise with integrated decision-making.

Algorithm A7 Multi-Agent Collaboration

procedure COLLABORATE($agents, task$)
    $subtasks \leftarrow$ DECOMPOSE($task$)
    $assigned \leftarrow$ ASSIGN($agents, subtasks$)
    $results \leftarrow \{\}$
    for each $\langle agent, task \rangle$ in $assigned$ **do**
        $results[task] \leftarrow$ RUN($agent, task$)
    end for
    $resolved \leftarrow$ RESOLVE($results$)
    $final \leftarrow$ SYNTHESIZE($resolved$)
    return $final$
end procedure
procedure RESOLVE($results$)
    Find conflicts between agents
    Weight by expertise
    return conflict-free results
end procedure
procedure SYNTHESIZE($resolved$)
    Integrate cross-agent insights
    Create unified framework
    return final output
end procedure