

FinanceLLM: A Survey of Large Language Models in Finance

Ojas Nimase¹ Zhengao Li² Xinyu Wei³ Jinglin Hu⁴ LuoJia Liu⁵
Kangyi Zhao⁶ Yue Zhao¹ Yushun Dong^{2*}

¹University of Southern California ²Florida State University ³Independent Researcher

⁴University of Florida ⁵UNC Chapel Hill ⁶University of Pittsburgh

{nimase, yzhao010}@usc.edu {zl23i, yd24f}@fsu.edu
sherryxyw25@gmail.com Jinglin.hu@warrington.ufl.edu
josie_liu@kenan-flagler.unc.edu KAZ78@pitt.edu

Abstract

Large Language Models (LLMs) have emerged as transformative, powerful tools for processing and acting on unstructured financial data. However, existing surveys suffer from methodological fragmentation, evaluation inconsistency, and application disconnect. This hinders the transition of research from academia to real-life applications in financial institutions. To address these gaps, we (1) introduce a unified taxonomy organizing approaches by computational complexity, (2) propose a three-stage evaluation framework, and (3) examine LLM deployment approaches suited to specific operational contexts. Furthermore, we identify key challenges including validation frameworks, data contamination, and operational compliance that need to be further investigated.

1 Introduction

Large language models (LLMs) have emerged as powerful tools in processing the vast amounts of unstructured text generated from financial markets. Real-world deployments exemplify this trend: Bloomberg integrated BloombergGPT into its Value-at-Risk stress-testing dashboard, reducing analysis times from hours to minutes (Wu et al., 2023). This underscores how LLMs are becoming embedded across the financial value chain, from data ingestion to trade execution and post-trade surveillance (Huang et al., 2023; Wu et al., 2023). Financial NLP is well-established at *ACL venues through dedicated workshops, shared tasks, and evaluation benchmarks (Xie et al., 2023, 2024; Malaysha et al., 2024). Organizing this rapidly growing body of work serves the ACL community by providing a unified framework for navigating computational trade-offs that generalize beyond finance to other domain-specific NLP applications.

While the rapid adoption of financial LLMs has generated significant academic interest, existing surveys fail to provide comprehensive guidance for practitioners. Most existing comprehensive surveys suffer from fundamental taxonomic confusion by mixing computational methods (machine learning, etc), application domains (trading, etc), and infrastructure (blockchain, etc) without clear separation (Vuković et al., 2025; Nie et al., 2024). Additionally, many surveys lack comprehensive evaluation frameworks to assess LLM performance across the full deployment lifecycle (Li et al., 2023; Lee et al., 2025; Tatarinov et al., 2025). While some surveys advocate for financial metrics over traditional ML measures, they fail to provide systematic evaluation frameworks connecting laboratory performance to real-world outcomes (Tatarinov et al., 2025; Lee et al., 2025). Other surveys lack unified validation frameworks and organize by application domains rather than computational architectures, preventing practical selection based on computational costs (Nie et al., 2024; Vuković et al., 2025). Since computational cost is a primary constraint for financial LLM deployment (Tavasoli et al., 2025), alternative dimensions such as supervision level or interpretability can serve as secondary selection criteria within each computational category.

As a result, we identify three critical gaps: methodological fragmentation, evaluation inconsistency, and application disconnect. *Methodological Fragmentation* is no unified framework for organizing the diverse approaches to financial LLMs. *Evaluation Inconsistency* is the lack of frameworks connecting laboratory performance to real-world deployment success. *Application Disconnect* represents the isolation of academic advances from the operational realities of financial institutions.

To address these gaps, we present **FinanceLLM**, a comprehensive survey that provides methodological clarity, evaluation coherence, and application guidance. To handle *Methodological Fragmenta-*

* Corresponding author.

Method	Representative Model	Metrics	Baseline	Training Dataset
Prompt-optimization Methods				
Example-Based Demonstration	Gemini, GPT (cluster) (Wei and Liu, 2025) GPT-4, BARD (Ahmed et al., 2024)	Acc. Human eval.	Random sel. Zero-shot	N/A N/A
Reasoning-Guided Instruction	Instruct-FinGPT-7B (Zhang et al., 2023a) LLaMA-7B + RAG (Zhang et al., 2023b)	Acc., F1 Acc., F1	FinBERT, LLaMA BloombergGPT	TFNS, FiQA TFNS, FiQA
Prompt-based Inference				
Zero-shot Deployment	ChatGPT (GPT-3.5/4) (Bond et al., 2023) GPT-4-32k (Papasotiriou et al., 2024)	Acc., corr. MAE	L&M dictionary Analyst ratings	N/A N/A
Chain-of-Thought	Qwen-7B-Chat (CoT) (Chen et al., 2024) PaLM (CoT) (Deng et al., 2023)	F1 Acc.	E2E-BERT FinBERT	N/A N/A
Representation Enhancement	FinBERT (quantum) (Yang et al., 2025)	F1	FT FinBERT	N/A
Agent-based Inference				
Sequential Processing	MarketSenseAI 2.0 (Fatouros et al., 2025)	Cum. return	SPY, OEF	N/A
Non-Sequential Collaboration	HAD (Xing, 2025) TradingAgents (Xiao et al., 2024)	Acc., F1 AR, SR	Naive prompt Buy-and-hold	N/A N/A
Parameter-efficient Methods				
Low-Rank Decomp. and Quantization	FinLlama (Iacovides et al., 2024) Llama-3-8B (QLoRA) (Ni et al., 2024) FinGPT (LoRA+RLSP) (Liu et al., 2023)	SR, AR Acc., F1, MCC F1, Acc.	FinBERT GPT-4, Mistral BloombergGPT	FPB, FiQA, TFNS, +GPT S&P 500 earnings News, social, filings
Reinforcement Learning	SAPPO (Kirtac and Germano, 2025b) I-RAG-RL (Zhao and Welsch, 2024)	SR, AR Acc., F1	PPO LLaMA-2-70B	Refinitiv news TFNS, FiQA-SA
Fully-tuning Methods				
Domain Transfer	BioFinBERT (Aparicio et al., 2024) Llama-2-7B (FT) (Luo and Gong, 2024)	Acc., F1 Acc.	FinBERT FinBERT, LSTM	Biopharma-Dive, FPB FPB
Contextual Adaptation	FinSentGPT (Ardekani et al., 2024) KR-FinBERT (Kim et al., 2023) DistilRoBERTa (Bhat and Jain, 2024)	Spearman, RMSE SR, AR Acc.	Linear reg. KOSPI 200 DistilRoBERTa	NYT, ECB policy Paxnet (Korean) 76 headlines
Synthetic Data Integration	FinBERT-LC (Atsiwo, 2024)	Acc., F1	FinBERT	FPB + synthetic

Table 1: Taxonomy of methods, ordered by ascending computational cost. Acc. = accuracy; AR = annualized return; SR = Sharpe ratio; L&M = Loughran–McDonald dictionary; FPB = Financial PhraseBank; TFNS = Twitter Financial News Sentiment; +GPT = GPT-labelled news. Dataset details are in Appendix C.

tion, we organize financial LLM approaches into three methodological pillars ordered by ascending computational cost: prompt-optimization methods, inference-based methods, and tuning-based methods. This enables practitioners to identify the most suitable approach given their computational budget. For *Evaluation Inconsistency*, we establish a three-stage validation framework: laboratory benchmarking, human evaluation, and production readiness. This connects academic metrics to performance requirements and operational constraints. For *Application Disconnect*, we examine LLM deployment across distinct financial functions (algorithmic trading, equity research, cross-asset domains) bridging

academic advances with practical implementation. We summarize the contribution of this paper as:

- **Unified Computational Taxonomy:** Methods organized by ascending computational cost: prompt optimization to full fine-tuning.
- **Multi-Layer Evaluation Framework:** Three progressive validation stages connecting laboratory benchmarking to production readiness.
- **Application-Method Mapping:** Computational approaches mapped to algorithmic trading, equity research, and cross-asset domains.

We present a comprehensive taxonomy in Table

1 that distinguishes three strategies ordered by ascending computational cost: optimizing prompts without parameter changes (prompt optimization), using models directly through API calls or multi-agent coordination (inference-based), and updating model parameters for domain adaptation (tuning-based). This reveals important trade-offs: prompt optimization methods offer maximum flexibility and interpretability at minimal cost, inference-based methods leverage pre-trained capabilities without modification, and tuning-based approaches provide superior domain adaptation at higher resource costs. We use financial sentiment analysis (FSA) as the primary illustrative domain because it offers the richest methodological diversity, though our taxonomy generalizes to tasks such as fraud detection, regulatory NER, and document extraction. This structured presentation helps practitioners navigate the cost-performance landscape: organizations with limited compute can begin with prompt optimization, scale to inference-based methods as needs grow, and invest in tuning only when domain adaptation justifies the resource cost. Our literature selection methodology is described in Appendix E.

2 Prompt Optimization Methods

At the lowest computational cost, prompt optimization adapts LLM behavior through strategic input design to leverage latent financial reasoning without modifying model parameters. This approach enables flexible deployment, compatibility with proprietary models, and greater interpretability via explicit reasoning chains. We explore prompt optimization through two input strategies reflecting the way we communicate tasks to LLMs. *Example-Based Demonstration* examines selecting optimal examples to help the model infer task requirements through demonstration. *Reasoning-Guided Instruction* goes further by crafting explicit instructions to guide models in accomplishing the task objective while also articulating their reasoning.

Example-Based Demonstration. The core intuition behind example-based approaches is that financial sentiment patterns can be conveyed through well-chosen demonstrations. In-context learning enables fast task adaptation and flexible deployment. A key challenge is selecting examples that effectively convey task-specific and domain-relevant patterns. Wei and Liu evaluate the efficacy of four demonstration selection strategies: random selection (picking K samples at random), distance-

based selection (choosing maximally dissimilar embeddings), difficulty-based selection (using an LLM to find the hardest examples), and clustering-based selection (selecting samples nearest to cluster centroids for diversity and representativeness) (Wei and Liu, 2025). Evaluation across ten LLMs on FiQA and Twitter datasets demonstrates that clustering-based approaches consistently outperform other methods by balancing diversity and representativeness (Wei and Liu, 2025). Ahmed et al. propose a structured prompt engineering approach that distinguishes between local (company-specific) and global (sector or macroeconomic) context in financial documents (Ahmed et al., 2024). In zero-shot settings, the framework is applied directly; in few-shot settings, it includes two AI-generated synthetic examples to illustrate the expected output (Ahmed et al., 2024). Evaluations on financial reports from Pakistani companies using four LLMs (ChatGPT, Bing Chat, BARD, Cohere) show that few-shot prompting improves accuracy, particularly in global sentiment detection, by offering clearer contextual guidance (Ahmed et al., 2024). While clustering-based selection provides the most robust demonstrations, it requires an embedding computation cost that instruction-based approaches avoid by shifting the burden to parameter updates.

Reasoning-Guided Instruction. Although instruction tuning modifies model parameters (as in Sections 4.1 and 4.2), it teaches prompt-following and reasoning articulation rather than encoding financial domain knowledge. This motivates its inclusion with prompt optimization strategies. Instruction tuning reframes FSA as a natural language generation task, training models to follow structured prompts and provide reasoning. How is this reformulation accomplished? Instruct-FinGPT involves three key steps: (1) converting existing financial sentiment datasets into instruction-following formats using 10 human-written templates, (2) performing supervised fine-tuning with sequence-to-sequence loss, and (3) mapping generated outputs back to sentiment labels during inference (Zhang et al., 2023a). The approach consistently outperforms both FinBERT and ChatGPT on Twitter validation datasets, while showing particular strength in contextual understanding scenarios (Zhang et al., 2023a). Their instruction-tuned model also exhibits strong zero-shot generalization capabilities, outperforming ChatGPT-3.5 and vanilla LLaMA-7B on the Financial PhraseBank dataset (Zhang et al., 2023a). Zhang et al. extend this approach by com-

binning instruction tuning with a RAG module that retrieves context from external financial sources, addressing the limited context in brief financial news (Zhang et al., 2023b). Their results show instruction-guided reasoning achieves optimal performance when paired with external knowledge retrieval (Zhang et al., 2023b). Compared to example-based demonstration, instruction tuning provides more consistent reasoning but blurs the boundary with tuning-based methods (Section 4), as both modify parameters. The shared finding across both strategies is that external knowledge retrieval consistently improves performance, suggesting FSA is often bottlenecked by context availability rather than reasoning capability.

3 Inference-based Methods

While prompt optimization focuses on input design, inference-based methods leverage the full capabilities of pre-trained LLMs by direct querying and multi-agent coordination at moderate computational cost. These methods require no parameter modification but incur higher inference-time costs than prompt optimization, particularly for multi-agent systems. We organize by interaction architecture: *Prompt-based Inference* uses direct querying while *Agent-based Inference* deploys specialized collaborating instances. This reflects the complexity shift from querying to multi-instance systems. Accordingly, techniques like RAG span multiple categories depending on their deployment cost.

3.1 Prompt-based Inference

Prompt-based inference directly queries a model. This approach prioritizes computational efficiency and deployment simplicity while leveraging the capabilities within a unified model. We explore three approaches that address different interaction patterns. *Zero-shot Deployment* directly queries the model. *Chain-of-Thought* structures complex analysis through multi-step reasoning chains. *Representation Enhancement* leverages learned representations to feed into the model via basic prompting. **Zero-shot Deployment.** This deployment approach uses pre-trained models through direct API calls without parameter modification. One study shows ChatGPT outperforms FinBERT and traditional sentiment classifiers (Loughran & McDonald, etc) in analyzing financial news summaries to predict S&P 500 returns (Bond et al., 2023). Papasotiriou et al. utilize GPT-4 to generate stock

ratings across S&P 500 companies using financial and news data, outperforming the consensus of 45,000 Wall Street analyst ratings (Papasotiriou et al., 2024). These findings establish prompting as a low-cost and efficient deployment strategy, though optimal input design (covered in Section 2) remains crucial for performance.

Chain-of-Thought. Chain-of-Thought (CoT) structures complex tasks into intermediate steps to improve reasoning. This structure can be useful for FSA, which involves determining emotional tone, market sentiment, and opinion polarity. Chen et al. apply a 4-step CoT framework decomposing FSA into sequential stages (company identification, sentiment classification, etc) and achieve strong performance on long, noisy texts and Chinese financial news (Chen et al., 2024). Meanwhile, Deng et al. employ multi-path CoT reasoning for cross-validation between reasoning chains, achieving 97.3% accuracy on FiQA News and 72.0% on Reddit sentiment analysis (Deng et al., 2023). In both cases, the reasoning structure comes entirely from the prompt while model parameters remain frozen. This is different from instruction-tuned reasoning (Section 2) where parameters are updated to internalize reasoning patterns. CoT thus trades increased latency and token usage for structured reasoning, a cost-robustness trade-off absent in zero-shot deployment, which is cheaper but fragile to prompt wording (Dmonte et al., 2024).

Representation Enhancement. This approach enhances the underlying data representations before feeding them to LLMs. Unlike prompt optimization strategies, the innovation lies in creating richer semantic embeddings. Yang et al. propose a quantum-inspired enhancement of FinBERT embeddings, encoding multiple interpretations of ambiguous financial terms (Yang et al., 2025). This approach achieves a zero-shot performance within 4.4 percentage points of fully fine-tuned FinBERT on Financial PhraseBank (Yang et al., 2025). Unlike the preceding approaches, representation enhancement improves input quality rather than inference strategy, but shares their core limitation: dependence on pre-trained knowledge that cannot adapt to novel financial terminology or market regimes.

3.2 Agent-based Inference

Agent-based inference deploys model instances with specialized roles collaborating to mimic financial institutions. We explore two approaches: *Sequential Processing* follows strict pipelines where

agents pass work through defined stages, while *Non-Sequential Collaboration* encompasses parallel processing and flexible coordination patterns.

Sequential Processing. MarketSenseAI developed a Chain-of-Agents system that assigns distinct LLM agents to sequential tasks (news summarization, fundamental analysis, etc) (Fatouros et al., 2025). Empirical evaluation shows the sequential approach significantly outperforming the benchmark indices (Fatouros et al., 2025). This demonstrates how sequential agent coordination can generate increasingly refined investment insights through structured information handoffs.

Non-Sequential Collaboration. Unlike rigid pipeline architectures, non-sequential coordination allows agents to interact flexibly. The Heterogeneous Agent Discussion (HAD) framework demonstrates this: seven specialized agents operate independently before a summative agent synthesizes their perspectives (Xing, 2025). TradingAgents combines parallel analysis with dynamic interactions. Analysts generate reports independently, bull-and-bear researchers engage in adaptive debates based on emerging conflicts, and risk management teams intervene opportunistically as conditions warrant (Xiao et al., 2024). Both approaches outperform their respective benchmark methods, suggesting adaptive coordination can be as effective while offering greater flexibility in agent interactions. However, neither study evaluates robustness to adversarial market conditions or measures latency overhead from multi-agent coordination, gaps that can limit deployment conclusions. Additionally, neither study quantifies coordination overhead relative to single-model baselines, leaving the cost-benefit ratio unclear for latency-sensitive applications.

4 Tuning-based Methods

While inference-based methods are effective, they are limited by their reliance on pre-trained knowledge. At the highest computational cost, tuning-based methods achieve in-domain adaptation by updating full or partial parameters. In FSA, this approach addresses a key limitation: pre-trained LLMs often lack the specialized knowledge needed for optimal performance. We organize tuning methods by the extent of parameter updates and computational cost. *Parameter-efficient Methods* update only small subsets of parameters, offering lower costs and faster deployment. *Full-tuning Methods*

adapt all model parameters for maximal domain integration, but require significant resources. This reflects the field’s move from exhaustive adaptation to scalable, efficient solutions.

4.1 Parameter-efficient Methods

While fully fine-tuning methods update every parameter, parameter-efficient fine-tuning only adjusts a small subset of parameters while keeping the rest frozen (Iacovides et al., 2024; Liu et al., 2023; Kirtac and Germano, 2025b; Zhao and Welsch, 2024). These methods (low-rank adaptation, quantization, reinforcement learning) have reduced computational requirements and faster training times while achieving domain specialization (Ni et al., 2024). We explore two complementary strategies that address distinct limitations of resource-constrained FSA deployment. *Low-Rank Decomposition and Quantization* reduces memory and computation costs while preserving performance. *Reinforcement Learning* shifts the training focus from laboratory-based performance metrics to optimizing for real-world financial outcomes. Together, these methods enable practical, cost-effective deployment aligned with market goals.

Low-Rank Decomposition and Quantization.

The foundational principle behind low-rank adaptation approaches is that task-specific learning can be effectively captured in low-dimensional parameter subspaces. This works by exploiting redundancy in model representations without modifying entire model architectures. FinLlama utilizes Low-Rank Adaptation (LoRA) during the fine-tuning of Llama 2 7B, resulting in FinLlama significantly outperforming FinBERT while having 4.2 million trainable parameters compared to FinBERT’s 110 million parameters (Iacovides et al., 2024). Quantization is another approach to reduce the memory footprint of a model by representing parameters with lower precision (e.g., 4-bit instead of 16-bit). Ni et al. apply a 4-bit QLoRA (quantized LoRA) quantization before fine-tuning LLaMA-3-8B and demonstrate superior predictive performance when integrating multiple types of financial information (company fundamentals, market signals, and analyst actions) for predictions (Ni et al., 2024). The FinGPT series demonstrates the full extent of low-rank decomposition, combining LoRA/QLoRA techniques with reinforcement learning to achieve state-of-the-art performance in just 4 hours on a single A100 GPU. This stands in stark contrast to pre-trained baselines: BloombergGPT (53 days)

and Llama2 (21 days) (Liu et al., 2023). This signifies that low-resource environments can still build cutting-edge, market-aligned models.

Reinforcement Learning. Reinforcement learning aligns model training with real-world financial outcomes rather than classification accuracy. Sentiment-Augmented Proximal Policy Optimization (SAPPO) extends the PPO framework (a reinforcement learning method that learns from market outcomes) by integrating real-time sentiment signals to guide portfolio decisions (Kirtac and Germano, 2025b). To generate consistent sentiment scores, Refinitiv financial news is fed into LLaMa 3.3, and the outputs are used as inputs for the sentiment component. This creates a stable information pipeline (Kirtac and Germano, 2025b). SAPPO outperforms the baseline PPO agent, demonstrating that integrating consistent real-time market sentiment can be valuable in applications of FSA (Kirtac and Germano, 2025b). Further work has advanced this paradigm by enabling models to dynamically select information sources based on current market conditions. Zhao and Welsch proposed a PPO-based reinforcement learning framework enhanced with an adaptive Retrieval-Augmented Generation (RAG) module that strategically retrieves context-relevant financial information from multiple sources (Zhao and Welsch, 2024). The base model, LLaMA2, operates within a dual framework: the RAG module ensures access to contextually optimal information, and the reinforcement learning component adapts trading decisions based on actual financial outcomes (Zhao and Welsch, 2024). Empirically, LLaMA I-RAG-RL achieved a 1–6% improvement in classification accuracy and F1 over GPT-4 Turbo, BloombergGPT, and RoBERTa on the Financial PhraseBank and TFNS benchmarks, with a 3.61% higher Sharpe ratio (Zhao and Welsch, 2024). Across parameter-efficient methods, LoRA achieves competitive performance with 4.2M parameters versus FinBERT’s 110M (Iacovides et al., 2024), questioning whether full fine-tuning (Section 4.2) is justified outside highly specialized domains.

4.2 Fully-tuning Methods

Full fine-tuning updates all the parameters of a model, offering strong domain adaptation but coming with the highest computational costs. Among tuning methods, it’s the most comprehensive and based on the assumption that FSA requires structural changes to model weights. We explore full-

tuning methods through three FSA challenges: knowledge acquisition, representation diversity, and data scarcity. *Domain Transfer* addresses the knowledge acquisition challenge by leveraging specialized knowledge (e.g., biomedical, general language) for finance. *Contextual Adaptation* explores the need to model linguistic and cultural diversity in global financial models (representation diversity challenge). *Synthetic Data Integration* mitigates the data scarcity challenge by examining the use of synthetic labeled data.

Domain Transfer. The intuition behind this approach is recognizing that effective FSA requires a strategic transfer of knowledge from source domains to financial contexts. BioFinBERT exemplifies cross-domain transfer by fine-tuning BioBERT on the Biopharma-Dive and Financial PhraseBank combined dataset, leveraging BioBERT’s biomedical grounding for FSA in biotechnology markets (Aparicio et al., 2024). However, BioFinBERT’s performance was comparable to FinBERT in biomedical markets. This suggests cross-domain transfer may not offer significant advantages even after fine-tuning on financial data (Aparicio et al., 2024). In contrast, Luo and Gong demonstrated highly effective general-to-financial domain adaptation. They examined the performance of the pre-trained model with few-shot prompting, unsupervised domain pre-training (mix of financial and general-domain text), and fine-tuning on the Financial PhraseBank dataset (Luo and Gong, 2024). This demonstrates the importance of strategic use of high-quality labeled data rather than relying on extensive domain pre-training.

Contextual Adaptation. This approach stems from the fact that financial markets operate across multiple languages, cultural frameworks, and emotional dimensions. These factors require FSA approaches to capture this context complexity. FinSentGPT fine-tuned OpenAI’s ADA GPT-3 on diverse datasets: 2,226 NYT articles on US inflation and European Central Bank’s (ECB) monthly policy statements in five European languages (Ardekani et al., 2024). FinSentGPT (ECB + NYT fine-tuning) outperformed FinSentGPT (ECB English-only fine-tuning), highlighting that linguistic and topical diversity help enhance model robustness (Ardekani et al., 2024). A Korean FSA study evaluated three LLMs: KoBERT (Korean-pretrained BERT), KR-FinBERT (FinBERT fine-tuned on a Korean financial corpora), and GPT-3.5 (off-the-shelf multilingual) (Kim et al., 2023). Both

GPT-3.5 and KR-FinBERT outperformed on the KOSPI 200 EWI benchmark model. Notably, KR-FinBERT achieved superior risk-adjusted returns, underscoring the value of domain-specialized fine-tuning for cultural grounding (Kim et al., 2023). Bhat and Jain show that fine-tuning DistilRoBERTa on Ekman’s six emotions outperforms binary sentiment for next-day stock movement prediction, suggesting FSA benefits from finer emotional granularity (Bhat and Jain, 2024). The need for linguistic diversity extends beyond European and East Asian languages. The AraFinNLP shared task introduced the first Arabic financial NLP evaluation covering intent detection and cross-dialectal financial text processing across Modern Standard Arabic, Tunisian, and Saudi dialects (Malaysha et al., 2024), while the ArBanking77 dataset provides Arabic banking intent classification across 77 intent categories (Jarrar et al., 2023). These efforts highlight that financial NLP requires not only language-specific fine-tuning but also dialect-aware adaptation for markets where informal financial communication diverges significantly from formal text.

Synthetic Data Integration. Integrating synthetic data generation can address the inherent limitations of the lack of available labeled financial data. BertNSP further pre-trains BERT on Bloomberg financial data to generate longer synthetic financial sentences (Atsiwo, 2024). This data, along with synthetic data from other LLMs and real Financial PhraseBank data, was used to train FinBERT-LC. The resulting model achieved a 3–4 point lift in both accuracy and macro F1 over models trained on human labels alone (Atsiwo, 2024). Across full-tuning methods, diminishing returns are evident: BioFinBERT performed comparably to FinBERT despite additional pre-training (Aparicio et al., 2024), while multilingual fine-tuning and synthetic data show more consistent gains. This suggests that data diversity matters more than model capacity for FSA.

5 Evaluations

Having explored the three methodological approaches to financial LLMs, we now examine how to systematically evaluate their effectiveness. We organize evaluation approaches into three progressive validation layers that mirror the development lifecycle. *Laboratory Benchmarking* establishes baseline capabilities on standardized datasets. *Human Evaluation* probes reasoning quality and in-

terpretability. *Production Readiness* validates performance under operational constraints. This progression reflects the field’s evolution from isolated performance metrics to comprehensive validation frameworks. We map representative works to their achieved validation stages in Appendix B.

Laboratory Benchmarking. Laboratory benchmarking is useful for understanding the strengths and weaknesses of different model architectures rather than simply measuring which achieves the highest accuracy scores. For example, Kirtac and Germano compare the performance of bidirectional models (FinBERT, RoBERTa) versus LLMs (GPT-4, OPT, LLaMA-3). They find the former excels at sentiment classification in formal financial documents (earnings reports, SEC filings), while the latter provides superior narrative understanding for informal financial news analysis (Kirtac and Germano, 2025a). Mahendran et al. confirm this pattern by testing on both formal financial datasets (PhraseBank) and informal social media (Reddit), demonstrating that models should be evaluated on data similar to their intended use (Mahendran et al., 2025). Standardized benchmarks further support this principle as the PIXIU framework provides a multi-task evaluation suite spanning sentiment analysis, question answering, and named entity recognition, enabling consistent cross-model comparison (Xie et al., 2023). Beyond PIXIU, several benchmarks expand evaluation scope. FLUE provides five English financial NLP tasks including sentiment, NER, and QA (Shah et al., 2022). FinBen extends this to 42 datasets spanning 24 tasks with bilingual English-Spanish evaluation (Xie et al., 2024). CFinBench offers 99,100 questions for Chinese financial contexts, where even GPT-4 achieves only 60.16% accuracy (Nie et al., 2025). Together, these benchmarks reveal that evaluation standardization remains fragmented across languages, with no single framework covering the full intersection of tasks, languages, and deployment scenarios. Chen and Kawashima find that incorporating LLaMA-3 sentiment analysis works best when combined with traditional forecasting methods rather than replacing them entirely (Chen and Kawashima, 2024). However, laboratory evaluation also reveals critical weaknesses. Dmonte et al. find that while GPT-4 can outperform specialized models like FinBERT, small changes in prompt wording cause dramatic accuracy drops (Dmonte et al., 2024). Fatemi et al. confirm this sensitivity problem persists even when models perform empir-

ically well (Fatemi and Hu, 2023). These findings establish three key evaluation principles: (1) test architectural fit for specific contexts (e.g., latency requirements, compute availability), (2) evaluate how models work with other financial processes rather than in isolation, and (3) include robustness testing beyond peak performance measurement.

Human Evaluation. While laboratory metrics capture statistical performance, human evaluation helps gauge the reasoning quality and reliability required by financial professionals. Du et al. evaluate financial reasoning by prompting models to consider semantics, numbers, time, comparisons, causality, and risk (Du et al., 2024). Without this guidance, GPT-3.5 and PaLM-2 show sharp drops on numerical and comparative tasks, exposing serious gaps in financial reasoning (Du et al., 2024). However, LLMs show strength in simpler tasks. Muhammad and Rospocher and Nakagawa et al. show that ChatGPT-4 matches specialized models in identifying company sentiment from financial headlines, suggesting it can naturally grasp company-sentiment links without task-specific training (Muhammad and Rospocher, 2025; Nakagawa et al., 2024). To better understand these reasoning patterns, Deng et al. have models verbalize their logic by breaking sentiment analysis into explicit steps (Deng et al., 2023). Human evaluation thus establishes that effective assessment must probe reasoning depth across multiple abstraction levels to evaluate explanatory power and predictive accuracy.

Production Readiness. Production evaluation tests whether lab results hold up in real-world deployment. It often reveals that infrastructure constraints matter more than laboratory performance. Trading studies confirmed this infrastructure reality: LLM-based strategies achieved strong returns (14.6% annually) and outperformed traditional methods, but higher trading costs and latency above 300ms significantly degraded performance (Liu et al., 2023). These resource constraints mean that while specialized models like FinBERT-C outperform general models, GPT-4 closes the gap through prompt engineering and serves as a practical, low-compute alternative (Lan et al., 2024). Further exploring these cost considerations, Roumeliotis et al. demonstrate that although ChatGPT/Claude match BERT performance with minimal examples, GPU costs double under dialogue streams. This forces cost-cutting strategies in real-world trading environments (Roumeliotis

et al., 2024). As a result, deployment requires testing real-world constraints and resource scenarios, since laboratory performance means little without operational viability.

6 Applications and Challenges

While evaluation frameworks guide model selection, practical deployment reveals performance across diverse financial functions. We examine these applications through three domains that reflect different operational needs. *Algorithmic Trading Systems* addresses the need for rapid, systematic execution in volatile markets. *Equity Research* focuses on synthesizing information for investment research and rating generation. *Cross-Asset Class Domains* examines LLM adaptation across different asset classes and regulatory environments. Finally, *Challenges and Future Directions* explores issues that hinder large-scale deployment. Our taxonomy generalizes beyond FSA to tasks such as fraud detection, regulatory NER, and document extraction; we illustrate this in Appendix A.

Algorithmic Trading Systems. Modern capital markets demand execution strategies with real-time processing and execution at institutional scale. Here, sentiment-augmented approaches consistently outperform traditional methods across asset types. For example, FinBERT integrated with econometric models achieves 304.65% gains in cryptocurrency trading (Liu et al., 2024). Additionally, fine-tuned LLMs outperform dictionary methods in U.S. equities over 13 years (Kirtac and Germano, 2024) and sentiment-enhanced reinforcement learning yields 52.6% profit improvements in portfolio management (Unnikrishnan, 2024). A common limitation across these studies is the absence of transaction cost modeling which can substantially erode reported gains in live trading. Furthermore, institutional trading requires collaborative expertise and explainable rationales. This prompted Xiao et al.’s TradingAgents to replicate institutional trading desks through specialized agents (fundamental analysts, sentiment analysts, technical analysts, risk management teams) that coordinate via structured protocols (Xiao et al., 2024). This yields returns of 26.62% (AAPL), 24.36% (GOOGL), and 23.21% (AMZN) versus baselines of 2.05%, 7.78%, and 17.1% (Xiao et al., 2024).

Equity Research. Beyond immediate trade execution, modern equity research involves synthesizing vast amounts of unstructured information into ac-

tionable investment insights and security ratings. This process suffers from information overload, filling inconsistencies, and cross-market complexity in global markets (linguistic, cultural, and regulatory understanding). Examining the efficacy of LLMs for these challenges, Papasotiriou et al. deploy GPT-4-32k for S&P 500 rating generation, outperforming human analysts with MAE of 1.417 versus 1.570 and validated through 45,000 ratings from 126 Wall Street firms (Papasotiriou et al., 2024). Complementing traditional rating methodologies, Chen and Kawashima demonstrate how sentiment-augmented approaches can enhance traditional equity research by providing systematic analysis across international stocks (Apple, Tencent, Toyota) (Chen and Kawashima, 2024). Their sentiment-guided approaches (Llama 3, FinBERT) outperform traditional approaches (LSTM, SVR, Random Forest) (Chen and Kawashima, 2024). These studies show how LLMs can enhance domestic ratings and global equity research, offering scalable solutions for broad market coverage.

Cross-Asset Class Domains. While general FSA provides a foundation, different asset classes exhibit distinctive market microstructures and operational constraints requiring specialized adaptations. Currency markets require interpretable explanations beyond sentiment classifications. Limonad et al. achieved 71.33% trend prediction accuracy through explainable AI frameworks across five major currency pairs while maintaining regulatory transparency (Limonad et al., 2024). On the other hand, cryptocurrency presents a unique challenge with its emerging nature. To address this, Roumeliotis et al. conduct systematic model evaluations across crypto news sources (CryptoNews, CryptoPotato, CoinTelegraph), fine-tuning multiple LLMs (GPT-4, BERT, FinBERT) using 31,037 articles to establish performance benchmarks (Roumeliotis et al., 2024). Results show that fine-tuned GPT-4 outperforms FinBERT, providing empirical evidence for model selection to guide crypto traders (Roumeliotis et al., 2024).

Challenges and Future Directions. One challenge LLM trading strategies face is validation when training data contains future information the model should not have (backtesting data). This creates a look-ahead bias as it artificially inflates performance that is misleading about real-world effectiveness (Glasserman and Lin, 2023). Glasserman and Lin demonstrate that removing company identifiers from headlines (entity anonymization)

used as training data for sentiment analysis trading strategies leads to an improvement in returns (5.89 basis points) (Glasserman and Lin, 2023). This suggests that an LLM's prior knowledge about specific companies can interfere with sentiment analysis—a “distraction effect” that limits LLM performance (Glasserman and Lin, 2023). Another challenge is data contamination from traditional noise (spam, misinformation) and LLM-generated content creating feedback loops detached from authentic market behavior (Nie et al., 2024). Nie et al. show that cloud-based LLMs amplify contamination risks by both spreading spurious content and reinforcing feedback loops of AI-generated financial information (Nie et al., 2024). These echo chambers disconnect models from real market dynamics, ultimately degrading performance in trading, risk management, and advisory services (Nie et al., 2024). Several concrete mitigation strategies have emerged: temporal train/test splits that enforce strict chronological separation to prevent look-ahead bias, n-gram decontamination filtering that identifies and removes memorized training sequences from evaluation data, and entity anonymization protocols (Glasserman and Lin, 2023). However, these approaches remain ad hoc and domain-general as no financial-LLM-specific decontamination protocol has been established, and reliable benchmarks that reflect real-world market conditions are still needed. Addressing these challenges is essential for safe LLM integration into financial systems.

7 Conclusion

This survey addresses shortcomings in financial LLM literature. One contribution is a unified computational taxonomy for informed method selection. Our three-stage evaluation framework connects laboratory benchmarking to human evaluation and production readiness, bridging the validation gap between academic metrics and real-world financial performance requirements. Application-method mapping demonstrates how different computational approaches suit specific operational contexts. Despite significant progress, challenges in validation, data contamination, and operational compliance require continued research for safe financial LLM deployment.

Limitations

While this survey provides a comprehensive framework for understanding LLM applications in finance, there are some limitations we want to acknowledge. As a survey paper, there are limitations in empirical testing and this paper could benefit from empirical testing to support the framework and application-method mapping. Furthermore, many financial LLM models are proprietary and undocumented in literature leading to better models and advances that aren't documented in this paper. Additionally the coverage focuses mostly on developed financial markets: America, Europe, China, Korea, etc. While recent benchmarks such as CFinBench (Nie et al., 2025) and AraFinNLP (Malaysha et al., 2024) have begun addressing Chinese and Arabic financial NLP respectively, comprehensive coverage of emerging markets (e.g., Southeast Asian, African, and Latin American financial systems) remains limited across the field. This potentially leads to a lack of geographic and cultural diversity found in financial markets.

References

- Rabbia Ahmed, Sadaf Abdul Rauf, and Seemab Latif. 2024. Leveraging large language models and prompt settings for context-aware financial sentiment analysis. In *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–9. IEEE.
- Valentina Aparicio, Daniel Gordon, Sebastian G Huayamares, and Yuhuai Luo. 2024. BioFinBERT: finetuning large language models (LLMs) to analyze sentiment of press releases and financial text around inflection points of biotech stocks. *arXiv preprint arXiv:2401.11011*.
- Aref Mahdavi Ardekani, Julie Bertz, Cormac Bryce, Michael Dowling, and Suwan Cheng Long. 2024. FinSentGPT: A universal financial sentiment engine? *International Review of Financial Analysis*, 94:103291.
- Abraham Atsiwo. 2024. Financial sentiment analysis: Leveraging actual and synthetic data for supervised fine-tuning. *arXiv preprint arXiv:2412.09859*.
- Rithesh Bhat and Bhanu Jain. 2024. Stock price trend prediction using emotion analysis of financial headlines with distilled LLM model. In *Proceedings of the 17th International Conference on Pervasive Technologies Related to Assistive Environments*, pages 67–73.
- Shaun A Bond, Hayden Klok, and Min Zhu. 2023. Large language models and financial market sentiment. *Available at SSRN 4584928*.
- Qizhao Chen and Hiroaki Kawashima. 2024. Stock price prediction using LLM-based sentiment analysis. In *2024 IEEE International Conference on Big Data (BigData)*, pages 4846–4853. IEEE.
- Tianyu Chen, Yiming Zhang, Guoxin Yu, Dapeng Zhang, Li Zeng, Qing He, and Xiang Ao. 2024. EFSA: Towards event-level financial sentiment analysis. *arXiv preprint arXiv:2404.08681*.
- Syed Affan Daimi and Asma Iqbal. 2024. A scalable data-driven framework for systematic analysis of SEC 10-k filings using large language models. *Preprint*, arXiv:2409.17581.
- Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2023. What do LLMs know about financial markets? A case study on Reddit market sentiment analysis. In *Companion Proceedings of the ACM Web Conference 2023*, pages 107–110.
- Alphaeus Dmonte, Eunmi Ko, and Marcos Zampieri. 2024. An evaluation of large language models in financial sentiment analysis. In *2024 IEEE International Conference on Big Data (BigData)*, pages 4869–4874. IEEE.
- Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. An evaluation of reasoning capabilities of large language models in financial sentiment analysis. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 189–194. IEEE.
- Sorouralsadat Fatemi and Yuheng Hu. 2023. A comparative analysis of fine-tuned LLMs and few-shot learning of LLMs for financial sentiment analysis. *arXiv preprint arXiv:2312.08725*.
- George Fatouros, Kostas Metaxas, John Soldatos, and Manos Karathanassis. 2025. MarketSenseAI 2.0: Enhancing stock analysis through LLM agents. *arXiv preprint arXiv:2502.00415*.
- Paul Glasserman and Caden Lin. 2023. Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. *arXiv preprint arXiv:2309.17322*.
- Allen H Huang, Hui Wang, and Yi Yang. 2023. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.
- Giorgos Iacovides, Thanos Konstantinidis, Mingxue Xu, and Danilo Mandic. 2024. FinLlama: LLM-based financial sentiment analysis for algorithmic trading. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 134–141.
- Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem. 2023. ArBanking77: Intent detection neural model and a new dataset in modern and dialectical Arabic. In *Proceedings of ArabicNLP 2023*, pages 276–287, Singapore (Hybrid). Association for Computational Linguistics.

- Seonmi Kim, Seyoung Kim, Yejin Kim, Junpyo Park, Seongjin Kim, Moolkyeol Kim, Chang Hwan Sung, Joohwan Hong, and Yongjae Lee. 2023. LLMs analyzing the analysts: Do BERT and GPT extract more value from financial analyst reports? In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 383–391.
- Kemal Kirtac and Guido Germano. 2024. Enhanced financial sentiment analysis and trading strategy development using large language models. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 1–10.
- Kemal Kirtac and Guido Germano. 2025a. Large language models in finance: what is financial sentiment? *arXiv preprint arXiv:2503.03612*.
- Kemal Kirtac and Guido Germano. 2025b. [Leveraging LLM-based sentiment analysis for portfolio allocation with proximal policy optimization](#). In *ICLR 2025 Workshop on Machine Learning Multiscale Processes*.
- Yinyu Lan, Yanru Wu, Wang Xu, Weiqiang Feng, and Youhao Zhang. 2024. [Chinese fine-grained financial sentiment analysis with large language models](#). *Neural Computing and Applications*.
- Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. 2025. [Large language models in finance \(FinLLMs\)](#). *Neural Computing and Applications*.
- Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. [Large language models in finance: A survey](#). In *Proceedings of the Fourth ACM International Conference on AI in Finance, ICAIF '23*, pages 374–382, New York, NY, USA. Association for Computing Machinery.
- Lior Limonad, Fabiana Fournier, Juan Manuel Vera Díaz, Inna Skarbovsky, Shlomit Gur, and Raquel Lazcano. 2024. Monetizing currency pair sentiments through LLM explainability. *arXiv preprint arXiv:2407.19922*.
- Chenghao Liu, Arunkumar Arulappan, Ranesh Naha, Aniket Mahanti, Joarder Kamruzzaman, and In-Ho Ra. 2024. Large language models and sentiment analysis in financial markets: A review, datasets and case study. *IEEE Access*.
- Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023. FinGPT: Democratizing internet-scale data for financial large language models. *arXiv preprint arXiv:2307.10485*. Presented at the Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.
- Yi-Te Lu and Yintong Huo. 2025. [Financial named entity recognition: How far can LLM go?](#) In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 164–168, Abu Dhabi, UAE. Association for Computational Linguistics.
- Wei Luo and Dihong Gong. 2024. Pre-trained large language models for financial sentiment analysis. *arXiv preprint arXiv:2401.05215*.
- Manish Barath Mahendran, Aswin Kumar Gokul, Poornima Lakshmi, and S Pavithra. 2025. Comparative advances in financial sentiment analysis: A review of BERT, FinBERT, and large language models. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pages 39–45. IEEE.
- Sanad Malaysha, Mo El-Haj, Saad Ezzini, Mohammed Khalilia, Mustafa Jarrar, Sultan Almujaivel, Ismail Berrada, and Houda Bouamor. 2024. [AraFinNLP 2024: The first Arabic financial NLP shared task](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 393–402, Bangkok, Thailand. Association for Computational Linguistics.
- Iftikhar Muhammad and Marco Rospocher. 2025. On assessing the performance of LLMs for target-level sentiment analysis in financial news headlines. *Algorithms*, 18(1):46.
- Kei Nakagawa, Masanori Hirano, and Yugo Fujimoto. 2024. Evaluating company-specific biases in financial sentiment analysis using large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pages 6614–6623. IEEE.
- Haowei Ni, Shuchen Meng, Xupeng Chen, Ziqing Zhao, Andi Chen, Panfeng Li, Shiyao Zhang, Qifu Yin, Yuanqing Wang, and Yuxi Chan. 2024. Harnessing earnings reports for stock predictions: A QLoRA-enhanced LLM approach. In *2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, pages 909–915. IEEE.
- Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, Yunhe Wang, and Dacheng Tao. 2025. [CFinBench: A comprehensive Chinese financial benchmark for large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 876–891, Albuquerque, New Mexico. Association for Computational Linguistics.
- 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](#). *Preprint*, arXiv:2406.11903.
- Kassiani Papasotiriou, Srijan Sood, Shayleen Reynolds, and Tucker Balch. 2024. AI in investment analysis: LLMs for equity stock ratings. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 419–427.

- Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. 2024. LLMs and NLP models in cryptocurrency sentiment analysis: A comparative classification study. *Big Data and Cognitive Computing*, 8(6):63.
- Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. [When FLUE meets FLANG: Benchmarks and large pre-trained language model for financial domain](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Xuwei Tan, Yao Ma, and Xueru Zhang. 2025. [Understanding structured financial data with LLMs: A case study on fraud detection](#). *Preprint*, arXiv:2512.13040. To appear in Proceedings of ACL 2026.
- Nikita Tatarinov, Siddhant Sukhani, Agam Shah, and Sudheer Chava. 2025. [Language modeling for the future of finance: A survey into metrics, tasks, and data opportunities](#). *Preprint*, arXiv:2504.07274.
- Ahmadreza Tavasoli, Maedeh Sharbaf, and Seyed Mohammad Madani. 2025. [Responsible innovation: A strategic framework for financial LLM integration](#). *Preprint*, arXiv:2504.02165.
- Ananya Unnikrishnan. 2024. Financial news-driven LLM reinforcement learning for portfolio management. *arXiv preprint arXiv:2411.11059*.
- Darko B. Vuković, Senanu Dekpo-Adza, and Stefana Matović. 2025. [AI integration in financial services: a systematic review of trends and regulatory challenges](#). *Humanities and Social Sciences Communications*, 12(562).
- Xinyu Wei and Luojia Liu. 2025. Are large language models good in-context learners for financial sentiment analysis? *arXiv preprint arXiv:2503.04873*.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabrovolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambar, David Rosenberg, and Gideon Mann. 2023. [BloombergGPT: A large language model for finance](#). *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, Ziyang Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others. 2024. [FinBen: A holistic financial benchmark for large language models](#). *Preprint*, arXiv:2402.12659.
- Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. [PIXIU: A comprehensive benchmark, instruction dataset and large language model for finance](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 33469–33484. Curran Associates, Inc.
- Frank Xing. 2025. [Designing heterogeneous llm agents for financial sentiment analysis](#). *ACM Trans. Manage. Inf. Syst.*, 16(1).
- Xi Yang, Jia Zhu, and Pasquale De Meo. 2025. [A quantum-like zero-shot approach for sentiment analysis in finance](#). *Journal of Intelligent Information Systems*, 63(3):705–721.
- Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023a. Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*. Presented at the International Symposium on Large Language Models for Financial Services (FinLLM) at IJCAI 2023.
- Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023b. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 349–356.
- Zijie Zhao and Roy E Welsch. 2024. Aligning LLMs with human instructions and stock market feedback in financial sentiment analysis. *arXiv preprint arXiv:2410.14926*.

A Taxonomy Generalization Beyond FSA

Our taxonomy organizes methods by computational cost and this structure applies to financial NLP tasks beyond FSA with each requiring practitioners to choose between the same deployment strategies (inference, tuning, and prompt-optimization) based on available resources. For example, financial fraud detection involves similar tradeoffs between zero-shot prompting LLMs to detect fraud and fine-tuning a model on institution-specific patterns (Tan et al., 2025). Named entity recognition (NER), which identifies organizational names and financial instruments, explores the same tradeoffs relative to performance (Lu and Huo, 2025). Similarly, document extraction from filings such as SEC 10-Ks involves the same choice between prompting pipelines and fine-tuned models for structured data parsing (Daimi and Iqbal, 2024). In each case, the core decision remains: how much computational investment is justified for the task? Our taxonomy provides a consistent framework regardless of domain.

B Evaluation Stage Mapping

Table 2 maps representative works to their achieved validation stages, illustrating gaps between laboratory evaluation and deployment readiness.

Work	Laboratory	Human Evaluation	Production Readiness
Prompt-optimization Methods			
Clustering Selection (Wei and Liu, 2025)	✓	–	–
Six-part prompt (Ahmed et al., 2024)	–	✓	–
Instruct-FinGPT (Zhang et al., 2023a)	✓	–	–
LLaMA RAG (Zhang et al., 2023b)	✓	–	–
Prompt-based Inference			
ChatGPT Zero-shot (Bond et al., 2023)	✓	–	–
GPT-4-32k Zero-shot (Papasotiriou et al., 2024)	✓	–	✓
Qwen-7B-Chat LoRA + CoT (Chen et al., 2024)	✓	–	–
PaLM CoT (Deng et al., 2023)	✓	✓	–
FinBERT (Yang et al., 2025)	✓	–	–
Agent-based Inference			
MarketSenseAI 2.0 (Fatouros et al., 2025)	✓	–	✓
Heterogeneous multi-Agent Discussion (Xing, 2025)	✓	–	–
TradingAgents (Xiao et al., 2024)	✓	–	✓
Parameter-efficient Methods			
FinLlama (Iacovides et al., 2024)	✓	–	✓
Llama-3-8b QLoRA (Ni et al., 2024)	✓	–	–
FinGPT (Liu et al., 2023)	✓	–	✓
SAPPO (Kirtac and Germano, 2025b)	✓	–	✓
I-RAG-RL (Zhao and Welsch, 2024)	✓	–	✓
Fully-tuning Methods			
BioFinBERT (Aparicio et al., 2024)	✓	–	–
Llama2-7B Fine-tuned (Luo and Gong, 2024)	✓	–	–
FinSentGPT (Ardekani et al., 2024)	✓	–	–
KR-FinBERT (Kim et al., 2023)	✓	–	✓
DistilRoBERTa Fine-tuned (Bhat and Jain, 2024)	✓	–	–
FinBERT-LC (Atsiwo, 2024)	✓	–	–

Table 2: Validation stages reached by representative works. Laboratory = benchmarking (accuracy, F1); Human Evaluation = expert reasoning evaluation; Production Readiness = deployment metrics (Sharpe ratio, latency, costs).

C Dataset Characterization

Table 3 provides size, domain coverage, and label information for key datasets referenced in this survey.

Dataset	Domain	Size	Labels
Financial PhraseBank (Aparicio et al., 2024; Luo and Gong, 2024; Atsiwo, 2024)	News	~4,840	3-class
FiQA Sentiment (Iacovides et al., 2024; Zhang et al., 2023a,b)	News, social	~1,170	Continuous
Twitter Fin. News (Iacovides et al., 2024; Zhang et al., 2023a,b)	Social media	~11,900	3-class
S&P 500 Earnings (Ni et al., 2024)	SEC filings	3,000+	Binary
Paxnet (Korean) (Kim et al., 2023)	Korean news	10,000+	3-class
Refinitiv News (Kirtac and Germano, 2025b)	Global news	100K+	Continuous
Biopharma-Dive (Aparicio et al., 2024)	Biotech news	2,500+	3-class
NYT + ECB (Ardekani et al., 2024)	News, policy	~2,200	Continuous
CryptoNews corpus (Roumeliotis et al., 2024)	Crypto news	~31,000	3-class
FiNER-ORD (Lu and Huo, 2025)	Fin. news	~200	NER
CFinBench (Nie et al., 2025)	Chinese finance	99,100	Multi-class
FinBen (Xie et al., 2024)	Multilingual (EN, ES)	42 datasets	Mixed
ArBanking77 (Jarrar et al., 2023)	Arabic banking	31,404	77 intents

Table 3: Key datasets used in surveyed works.

D Benchmark Desiderata

Existing financial LLM benchmarks each address different evaluation dimensions but none covers the full intersection of tasks, languages, and deployment scenarios. FLUE (Shah et al., 2022) targets English NLP tasks, PIXIU (Xie et al., 2023) adds multi-task coverage, FinBen (Xie et al., 2024) extends to bilingual evaluation and trading, and CFinBench (Nie et al., 2025) addresses Chinese financial contexts. Drawing on these efforts and our taxonomy, we outline desiderata for a unified financial LLM benchmark:

1. **Cost-tier coverage:** Evaluation tasks organized by computational cost (prompt optimization, inference-based, tuning-based) to enable method selection under resource constraints.
2. **Cross-domain tasks:** Tasks spanning algorithmic trading, equity research, and cross-asset domains rather than sentiment analysis alone.
3. **Validation stage alignment:** Metrics mapped to our three-stage framework: laboratory benchmarking (accuracy, F1), human evaluation (reasoning quality, explainability), and production readiness (latency, transaction costs, Sharpe ratio).
4. **Multilingual scope:** Coverage beyond English and Chinese to include Arabic, Spanish, and other underrepresented financial markets.
5. **Contamination controls:** Built-in temporal splits and decontamination protocols to prevent look-ahead bias.

E Literature Selection Methodology

We surveyed major NLP and AI venues (ACL, EMNLP, AAAI, NeurIPS, ICAIF, etc) from 2023–2025, supplemented by arXiv preprints matching queries for “financial LLM” and “financial sentiment analysis.” We prioritized works demonstrating methodological diversity across our three taxonomy categories (inference, tuning, prompt optimization) and excluded works focusing exclusively on time-series forecasting without language components, pure infrastructure surveys, and studies that do not evaluate on financial text data. Of the 23 works in Table 1, 12 were published in 2024, with six from 2023 and five from 2025. This concentration reflects the rapid growth of financial LLM research following the release of newer LLMs and open-source alternatives.