Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bridging Finance and AI: A Comprehensive Survey of Large Language Models in Financial Applications

Ameer Tamoor Khak

atk@plen.ku

University of Copenhagen

Shuai Li

University of Oulu

Xinwei Cao

Jiangnan University

Research Article

Keywords: Large Language Models, Financial Natural Language Processing, Transformer Architectures, Finance-Specific Benchmarks, Ethical and Regulatory Considerations

Posted Date: May 14th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6614220/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Digital Finance on July 16th, 2025. See the published version at https://doi.org/10.1007/s42521-025-00146-3.

Bridging Finance and AI: A Comprehensive Survey of Large Language Models in Financial System

Ameer Tamoor Khan[1*], Shuai Li[2] and Xinwei Cao[3]

[1*]Department of Plant and Environmental Sciences , University of Copenhagen, Nørregade 10, Copenhagen, 1172, , Denmark.
[2]Faculty of Information Technology and Electrical Engineering, University of Oulu, Pentti Kaiteran Katu 1, Oulu, 100190, , Finland.
[3]Department of Accounting, Jiangnan University, China, 1800 Lihu Blvd, Binhu District, 214126, Wuxi, Jiangsu, China.

*Corresponding author(s). E-mail(s): atk@plen.ku.dk;
Contributing authors: shuai.li@oulu.fi; xwcao@jiangnan.edu.cn;

Abstract

Large Language Models (LLMs) have rapidly transformed natural language processing across domains, and their adoption in finance promises to revolutionize tasks ranging from document understanding to quantitative analysis. This survey presents a comprehensive overview of LLM applications in financial contexts. We first trace the evolution of language modeling—from RNNs to Transformers—and summarize key architectures (encoder-only, decoder-only, encoder-decoder) and transfer-learning paradigms (pretraining, fine-tuning, prompt-based, instruction tuning). We then propose a taxonomy of finance-relevant tasks, including linguistic preprocessing (summarization, NER), sentiment analysis, financial reasoning and QA, forecasting/time-series modeling, and agent-based decision support. Next, we review the landscape of datasets and benchmarks—highlighting English-dominant biases, gaps in crisis-period coverage, and emerging multilingual resources. We catalog financial domain–specific LLMs, contrasting fine-tuned models (FinBERT, FinGPT, FinMA) with from-scratch systems (BloombergGPT, FinTral, XuanYuan 2.0) and compare them to general-purpose models. We examine evaluation metrics, contrasting standard ML measures (accuracy, F1, MSE) with finance-centric criteria (Sharpe Ratio, Maximum Drawdown) and discuss challenges around data leakage, hallucinations, robustness, and explainability. Ethical and regulatory considerations—data

privacy, bias, fairness, and compliance—are addressed, followed by a multi-level decision framework (zero-shot through pretraining from scratch) that balances accuracy, cost, and privacy. Finally, we outline open challenges and future directions, including cross-lingual models, symbolic reasoning integration, human–AI collaboration, and the development of open, temporally annotated finance benchmarks. This survey aims to guide researchers and practitioners in responsibly harnessing LLMs to advance financial AI.

Keywords: Large Language Models, Financial Natural Language Processing, Transformer Architectures, Finance-Specific Benchmarks, Ethical and Regulatory Considerations

1 Introduction

Financial institutions confront a broad spectrum of complex problems—portfolio optimization, risk management, fraud detection, algorithmic trading, and compliance monitoring—all of which demand accurate data-driven insights and adaptive decision-making strategies.

Early AI-based solutions leveraged statistical models (e.g., factor models, time-series analysis) and classical machine-learning techniques (e.g., SVMs, random forests, neural networks) to extract patterns from historical financial data and automate aspects of these tasks. Building on these, researchers developed nature-inspired, learning-based algorithms—such as genetic algorithms, particle swarm optimization, and ant colony systems—tailored for continuous and combinatorial financial optimization problems [1–5]. These metaheuristic frameworks brought greater flexibility and global search capability to high-dimensional, nonconvex financial landscapes.

However, nature-inspired learning-based approaches present notable limitations: they often require extensive parameter tuning, struggle to scale in large, noisy data environments, and lack transparent, verifiable reasoning paths needed for regulatory audit and compliance purposes [6–12].

Against this backdrop, Large Language Models (LLMs) have emerged as powerful, transformer-based architectures pretrained on massive text corpora that excel at understanding and generating human language. Models such as GPT-4, PaLM, and LLaMA demonstrate advanced capabilities in text summarization, question answering, semantic parsing, and structured reasoning [13, 14]. By embedding LLMs into financial workflows—automating sentiment analysis of earnings calls, extracting entities from regulatory filings, and supporting conversational agents—organizations can achieve scalable, explainable text analytics while paving the way for reliable compliance and decision-support systems [15–18].

Despite their promise, integrating LLMs into finance raises its own challenges: domain adaptation to specialized financial jargon, ensuring precision and reducing hallucinations, supporting multimodal inputs (text combined with tables or time-series), and maintaining auditability under strict governance frameworks [19, 20].

In this survey, we review the evolving role of LLMs in digital finance, contrasting them with legacy AI techniques. The remainder of the paper is organized as

follows. Section 2 presents a comprehensive taxonomy of financial tasks amenable to LLM application. Section 3 offers a critical comparison of domain-specific versus general-purpose LLM deployments. Section 4 evaluates key datasets and benchmarking frameworks, highlighting their limitations. Section 5 discusses the ethical, regulatory, and technical challenges inherent in deploying LLMs in financial services. Section 6 provides an overview of fine-tuned LLM models for financial tasks and domain adaptation strategies. Section 7 surveys evaluation metrics and the challenges of benchmarking LLMs within finance. Section 8 examines ethical and regulatory considerations specific to financial applications of LLMs. Section 9 introduces a decision framework for LLM adoption, guiding practitioners through model selection and integration. Section 10 identifies open challenges and outlines future research directions, and Section 11 concludes with key insights and recommendations for responsible LLM integration in digital finance.

2 Overview of Large Language Models

Language modeling has experienced substantial advancements over the past decade. Early models, primarily based on Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), significantly advanced sequential data processing and temporal dependency management [21]. However, their inherent limitations in capturing extensive contextual dependencies and difficulties in computational parallelization restricted their scalability and performance.

The paradigm shift occurred with the advent of the Transformer architecture [22], characterized by its self-attention mechanism, effectively modeling dependencies without recurrence. Transformers provided unprecedented scalability, enabling highly parallelizable computations and setting the foundation for modern Large Language Models (LLMs).

Transformer-based models are classified into three primary architectures:

- **Encoder-only models**: Specialized in tasks requiring contextual understanding, exemplified by BERT [23].
- **Decoder-only models**: Optimized for generative tasks, notably exemplified by the GPT series [24].
- **Encoder-Decoder models**: Ideal for translation, summarization, and general sequence-to-sequence tasks, exemplified by T5 [25].

Several benchmark LLMs have prominently shaped the NLP landscape:

- **GPT Series (OpenAI)**: Renowned for remarkable zero-shot and few-shot learning capabilities [24].
- **BERT**: Introduced bidirectional contextual embedding through masked language modeling [23].
- **T5**: Unified various NLP tasks into a single text-to-text approach [25].
- **LLaMA**: Prominent open-source decoder-only models facilitating widespread research and development [26].

Modern LLMs incorporate multiple transfer learning paradigms:

- **Pretraining**: Unsupervised learning on extensive text corpora to achieve general language proficiency.
- **Fine-tuning**: Adaptation of pretrained models to specific tasks using targeted datasets.
- **Prompt-based learning**: Guiding model outputs via specifically designed input prompts without modifying model parameters.
- **Instruction tuning**: Training models explicitly to understand and respond effectively to human-like instructions [27].

The NLP community has increasingly embraced high-performing open-source alternatives to proprietary models, promoting accessibility and innovation:

- **DeepSeek**: Focuses on multilingual and particularly Chinese linguistic contexts, enhanced via reinforcement learning from human feedback (RLHF) [28].
- **Mistral**: Employs a sparse mixture-of-experts architecture to dynamically manage computational efficiency and performance [29].
- **Falcon**: Optimizes computational efficiency through specialized tokenization and streamlined attention mechanisms [30].
- **BLOOM**: A collaborative multilingual model trained across 46 languages, aimed at democratizing access to powerful NLP capabilities [31].
- **OpenLLaMA**: Recreates Meta's LLaMA architecture using transparently available datasets, emphasizing open-source accessibility [32].
- **XGLM**: Designed specifically for cross-lingual applications, emphasizing performance on low-resource languages [32].

These open-source LLMs significantly lower entry barriers, fostering innovative research and practical applications in fields including finance, biomedical NLP, and multilingual environments.

Table 1 provides a comparative overview of general-purpose and finance-specialized LLMs, highlighting architectural characteristics, scale, domain specificity, and key performance attributes.

Model	Architecture	# Parameters	Domain	Key Strength
GPT-3 [24]	Decoder-only	175B	General	Zero/few-shot generation
BERT [23]	Encoder-only	340M	General	Deep contextual embeddings
T5 [25]	Encoder-Decoder	770M (base)	General	Unified text-to-text framework
LLaMA [26]	Decoder-only	7–65B	General Research	Lightweight open-source
BLOOM [31]	Decoder-only	176B	Multilingual	Multilingual, open-access
FinBERT [33]	Encoder-only	110M	Finance	Finance-specific language modeling
FinGPT [34]	Decoder-only	7B	Finance	Instruction-tuned finance tasks
BloombergGPT [35]	Decoder-only	50B	Finance	Numeric and finance-token optimized
FinTral [36]	Decoder-only	7–70B	Finance	Multimodal financial analysis

Table 1 Comparative overview of prominent general-purpose and finance-specialized LLMs.

3 Taxonomy of Financial Tasks for LLMs

Large Language Models (LLMs) have found extensive applicability across a diverse spectrum of financial tasks. This taxonomy organizes these tasks into distinct functional categories, clearly illustrating the scope and utility of LLMs, from foundational linguistic analysis to sophisticated decision-support frameworks.

3.1 Linguistic Tasks

LLMs are highly effective in linguistic processing for financial data, performing tasks such as:

- **Text summarization**: Condensing lengthy financial reports, earning calls, and regulatory documents into succinct summaries. Prominent models include GPT variants and Longformer [37].
- **Named Entity Recognition (NER)**: Identifying entities such as companies, financial instruments, and stakeholders in financial documents. Notable models include FinBERT and KPI-BERT [38].
- **Document classification**: Classifying financial documents based on their content, risk level, or ESG factors. Models like FinBERT and ESGBERT excel in these applications [39].
- **Knowledge graph construction**: Extracting structured relationships between financial entities using hybrid symbolic and LLM-based systems, such as GPT-4 and T5 [40].

3.2 Sentiment Analysis

Sentiment analysis leverages LLMs for tracking and predicting market dynamics through:

- **Social media and news analysis**: Real-time sentiment tracking from platforms such as Twitter using FinGPT and ChatGPT [41].
- **Corporate disclosures and filings**: Evaluating sentiment, forward-looking statements, and investor confidence with models like FinBERT [42].
- **ESG and policy sentiment tracking**: Monitoring sustainability disclosures and policy discussions via specialized models such as ESGBERT [43].

3.3 Financial Reasoning and Question Answering

LLMs support complex financial reasoning and question-answering tasks, notably:

- **Financial QA**: Addressing multi-step reasoning queries involving textual, numerical, and tabular data. FinQA and ConvFinQA are representative frameworks [44].
- **Numerical and tabular reasoning**: Utilizing specialized table-aware models such as TAT-QA, FinQANet, and TAPAS to derive insights from financial statements and balance sheets [45].
- **Hypothetical scenario analysis**: Simulating future financial scenarios and outcomes using GPT-4 and instruction-tuned models [46].

3.4 Forecasting and Time Series Applications

Forecasting applications integrate LLMs to enhance predictive accuracy, including:

- **Stock price prediction**: Integrating market sentiment from textual data and historical price movements via models like FinGPT [47].
- **Volatility forecasting**: Predicting financial market volatility using textual indicators of uncertainty extracted from financial news [48].
- **Anomaly detection**: Identifying irregularities in trading patterns and disclosure narratives through transformer-based detection models [49].

3.5 Agent-Based and Decision Support Applications

LLMs increasingly serve as intelligent agents for sophisticated financial decision-making:

- **Portfolio management**: Using GPT-derived agents for asset allocation and risk management strategies [50].
- **Trading simulations**: Employing multi-agent LLM systems to model complex market behaviors and trading scenarios [51].
- **Robo-advisors and financial chatbots**: Automating personalized financial advice and customer service through context-aware GPT-based systems [52].

Figure 1 visually summarizes the taxonomy, and Table 2 provides detailed categorization with representative models.

This structured taxonomy outlines the comprehensive utility and future potential of LLMs in digital finance.

4 Datasets and Benchmarks

The effectiveness of large language models (LLMs) in finance significantly depends on high-quality datasets used for their training, evaluation, and practical deployment. The richness, variety, and domain relevance of these datasets directly influence model robustness and generalizability. This section presents an overview of key financial datasets, benchmarks, associated linguistic and geographic biases, and notable temporal coverage gaps (Fig. 2).

4.1 Key Datasets

Several standardized datasets have become prominent for evaluating financial applications of LLMs:

- **FinQA** [53]: Targets numerical reasoning involving arithmetic and logical operations based on company earnings reports.
- **FiQA** [54]: Provides aspect-based sentiment analysis and opinion-based question-answering annotations, applicable to microblogs and financial news headlines.
- **ConvFinQA** [55]: Extends numerical reasoning challenges into conversational settings, testing multi-turn dialogue comprehension.

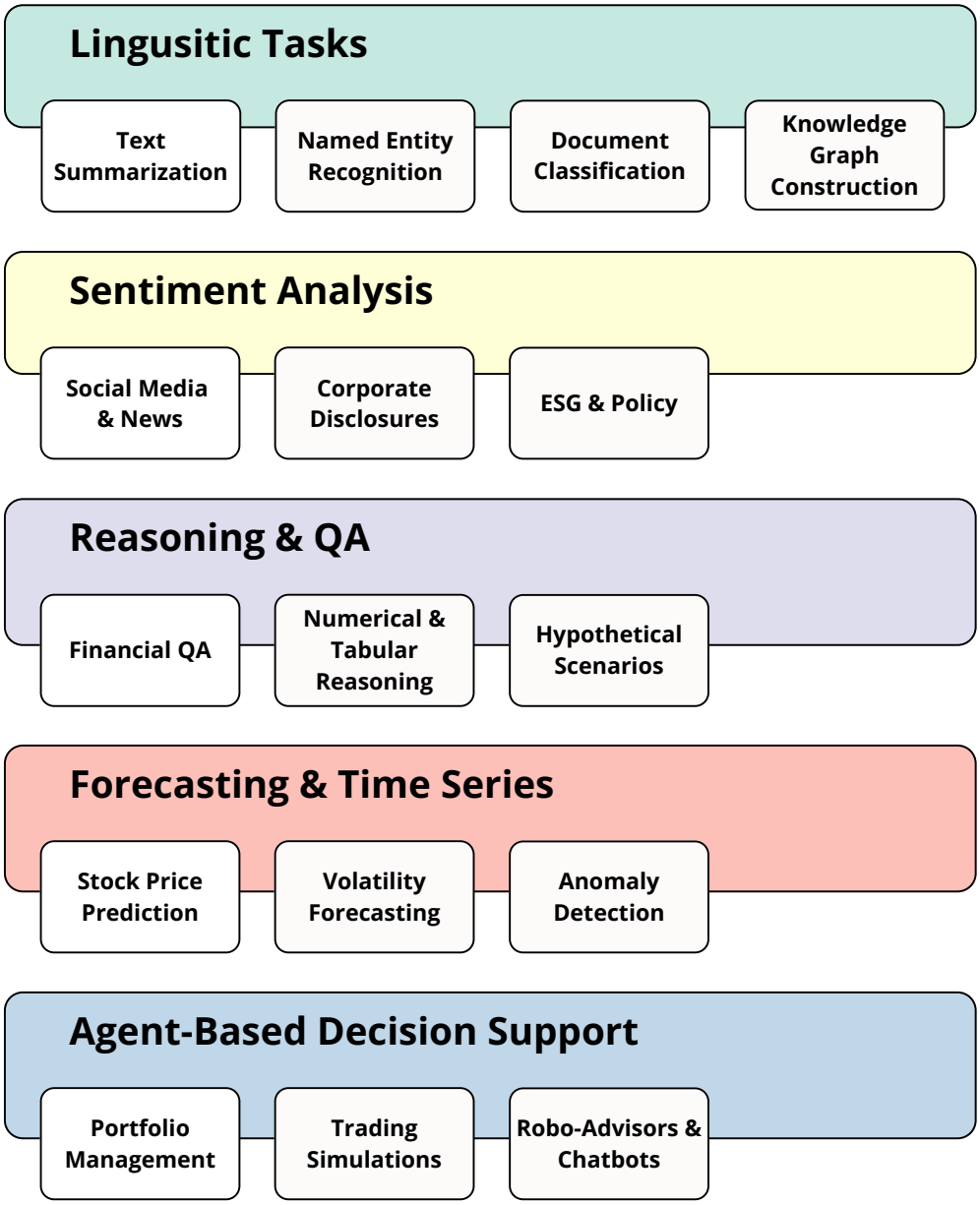

Fig. 1 A taxonomy of financial tasks for large language models, organized into five major categories—Linguistic Tasks, Sentiment Analysis, Reasoning & QA, Forecasting & Time Series, and Agent-Based Decision Support—with representative sub-tasks depicted beneath each category header.

- **DocFinQA** [56]: Focuses on comprehensive, long-context financial reasoning through analysis of entire earnings reports.

Category	Example Tasks	Representative Models / Works
Linguistic	Text summarization	GPT variants, Longformer [37]
	Named entity recognition	FinBERT, KPI-BERT [38]
	Document classification	FinBERT, ESGBERT [39]
	Knowledge graph construction	T5 + symbolic systems [40]
Sentiment	Social media analysis	FinGPT, ChatGPT [41]
	Disclosures & filings	FinBERT [42]
	ESG / policy tracking	ESGBERT [43]
Reasoning & QA	Financial QA	FinQA, ConvFinQA [44]
	Numerical / tabular reasoning	TAT-QA, FinQANet, TAPAS [45]
	Hypothetical scenario analysis	GPT-4, instruction-tuned LMs [46]
Forecasting & TS	Stock prediction	FinGPT–TS pipelines [47]
	Volatility forecasting	Transformer + news features [48]
	Anomaly detection	BERT-based detectors [49]
Agent-based & Decision Support	Portfolio management	GPT agent frameworks [50]
	Trading simulations	Multi-agent LLMs [51]
	Robo-advisors / chatbots	FinChat, GPT-based bots [52]

Table 2 Taxonomy of financial tasks for LLMs, with example subtasks and representative models or studies.

- **FinNLP Shared Tasks** [57]: Offers multilingual datasets for Named Entity Recognition (NER), sentiment analysis, and financial event detection.
- **PACIFIC** [58], **Tab-CQA** [59], and **FinanceBench** [60]: Provide benchmarks for evaluating models across tables, mixed-modal QA, and diverse scenario-based reasoning contexts.

4.2 Language and Regional Biases

Most available datasets exhibit a pronounced English-language and U.S.-centric bias. Surveys highlight significant limitations for global deployment due to this linguistic and regional imbalance [61]. Notable exceptions, though limited, include datasets such as IndicFinNLP [62], MultiFin [63], and ESG-FTSE [64], which offer multilingual and regional variety but suffer from limited scale.

4.3 Gaps in Temporal and Crisis Coverage

Temporal robustness remains inadequately addressed in current datasets. Critical market events such as financial crises (e.g., 2008 global financial crisis, COVID-19 induced volatility) are rarely represented despite their relevance [65]. Survivorship bias, where datasets exclude bankrupt or merged entities, further limits the real-world applicability of model evaluations [66]. Moreover, maintaining datasets to reflect continuously evolving financial regulations and reporting standards remains a critical, yet overlooked, necessity.

Table 3 summarizes major finance benchmarks, highlighting their primary tasks, modalities, and key limitations.

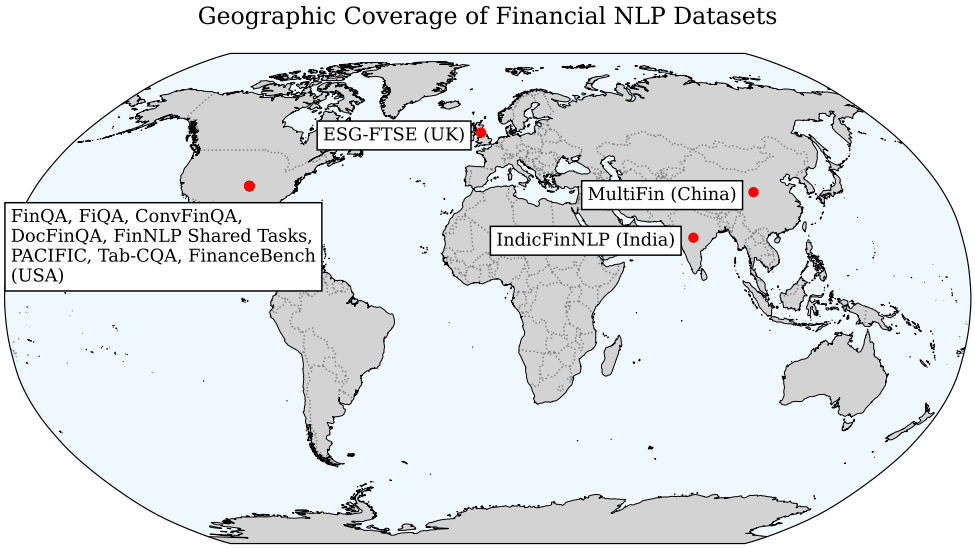

Fig. 2 Geographic coverage of prominent financial NLP datasets. A cluster of U.S.–based English resources (FinQA, FiQA, ConvFinQA, DocFinQA, FinNLP Shared Tasks, PACIFIC, Tab-CQA, FinanceBench) is shown over the United States, while non-English benchmarks—IndicFinNLP in India, MultiFin in China, and ESG-FTSE in the U.K.—are marked individually. The Robinson projection and vector-based coastlines ensure a clear, publication-quality map.

Dataset	Task	Modality	Notes
FinQA [53]	Numerical reasoning QA	Text + Tables	Earnings report questions
FiQA [54]	Sentiment analysis / QA	Short text (tweets/news)	Aspect-level labels
ConvFinQA [55]	Conversational QA	Dialogue + Tables	Multi-turn reasoning
DocFinQA [56]	Long-context reasoning QA	Full reports	Full-length earnings reports
FinNLP Tasks [57]	NER, sentiment, events	Multilingual text	Shared-task suites
PACIFIC [58]	Mixed-modality QA	Tables + Text + Charts	Proactive conversational assistant
Tab-CQA [59]	Table-based conversational QA	Tables + Dialogue	Chinese financial tables
FinanceBench [60]	Scenario-based reasoning	Text + Tables	Diverse reasoning scenarios
IndicFinNLP [62]	NER, sentiment, QA	Indic languages (e.g. Hindi)	Limited scale
MultiFin [63]	Financial QA	Chinese + English	Bilingual reports
ESG-FTSE [64]	ESG sentiment	Sustainability disclosures	U.K. FTSE-listed firms

Table 3 Key financial NLP datasets, their primary tasks, modalities, and notable characteristics or limitations.

5 Financial Domain–Specific LLMs

Language models tailored specifically for financial applications typically follow two main strategies: fine-tuning general-purpose large language models (LLMs) with domain-specific financial data or training specialized models entirely from scratch using financial corpora. Here, we analyze both approaches and evaluate their comparative effectiveness against general-purpose models. Table 4 summarizes these models, outlining their parameters, training data, adaptation strategies, and key strengths.

5.1 Fine-Tuned Models

- **FinBERT** [33]: Extends BERT by further pretraining on a substantial corpus of financial news and filings, significantly improving tasks such as sentiment analysis and named entity recognition (NER).
- **FinGPT** [34]: Utilizes low-rank adaptation (LoRA) to instruction-tune open-source models like LLaMA, efficiently achieving state-of-the-art performance in chat-based tasks, such as summarizing earnings calls.
- **FinMA** [67]: Employs instruction-tuning of LLaMA-30B on the FLARE dataset, which comprises 136,000 finance-related questions. FinMA shows exceptional zero-shot capabilities, particularly on complex benchmarks like TAT-QA.
- **Instruct-FinGPT** [68]: Fine-tunes a smaller LLaMA-7B model using a curated set of 10,000 financial instruction prompts, demonstrating competitive accuracy on sentiment analysis and classification tasks.

5.2 Models Trained from Scratch

- **BloombergGPT** [35]: Scaled up to 50 billion parameters, trained on 700 billion tokens equally split between financial and general texts. It excels at numerical reasoning and QA through specialized numeric tokenization.
- **FinTral** [36]: A multimodal finance-specific LLM (ranging from 7 to 70 billion parameters) integrating retrieval-augmented generation (RAG), visual chart understanding, and tool usage, performing particularly well in visual reasoning tasks.
- **XuanYuan 2.0** [69]: A bilingual model pretrained on 380 billion tokens in Chinese and English finance texts, maintaining general linguistic ability alongside specialized finance-domain performance through hybrid instruction tuning.

5.3 Comparison to General-Purpose Models

Domain-specific LLMs typically achieve comparable or superior results in financial benchmarks relative to general-purpose models such as GPT-3.5 and GPT-4, especially in tasks demanding numerical precision, financial terminology, or multimodal analysis [70, 71]. Nonetheless, general-purpose models may outperform in broader knowledge integration and emergent reasoning skills outside of strictly financial contexts.

As depicted in Fig. 3, GPT-4 demonstrates superior overall performance. BloombergGPT and FinGPT, however, offer significant domain-specific advantages

over base models like BERT, emphasizing a trade-off between specialization and broader applicability.

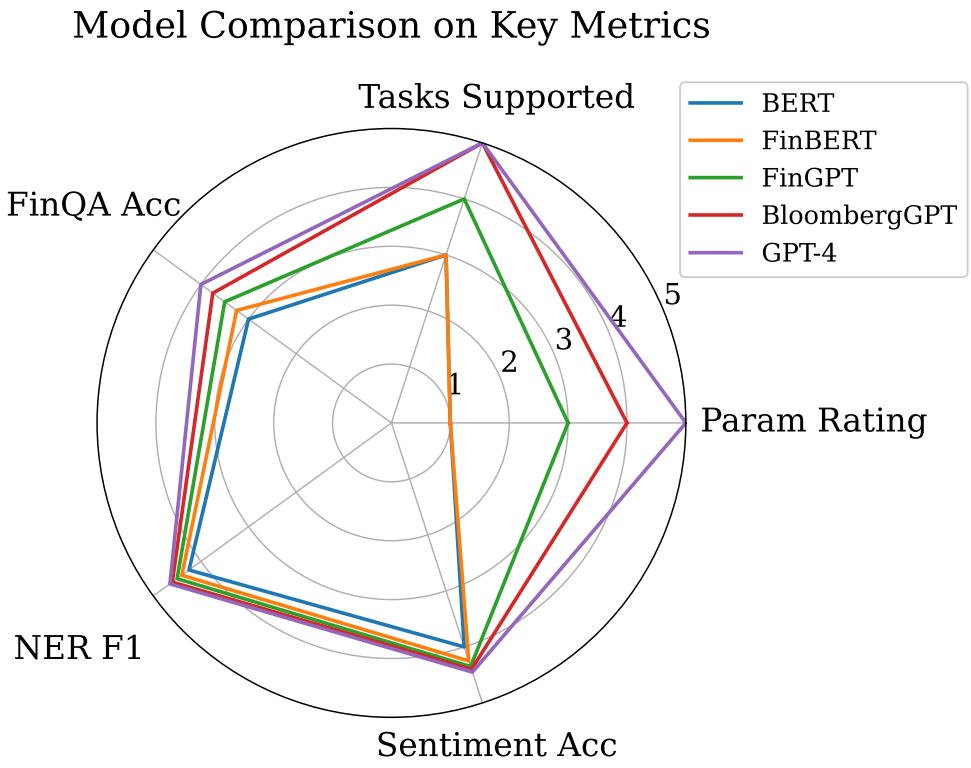

Fig. 3 Radar plot comparing five language models—BERT, FinBERT, FinGPT, BloombergGPT, and GPT-4—across key dimensions: model size (Param Rating), breadth of finance-specific task support, numerical QA accuracy on FinQA, financial NER F1 score, and finance sentiment classification accuracy. Higher values indicate better performance; finance-specialized models outperform vanilla BERT on domain tasks, while GPT-4 leads across all metrics.

6 Evaluation Metrics and Challenges

Robust evaluation is vital for assessing both the accuracy and practical utility of financial large language models (LLMs). Here, we critically review standard machine learning (ML) metrics, finance-specific performance measures, and emerging challenges affecting reliability and model robustness.

Model	Params	Training Data	Approach	Key Strength
Fine–Tuned General Models				
FinBERT [33]	110M	Financial news & filings (millions of docs)	Continued pretraining + finetune	Improved sentiment & NER
FinGPT [34]	7B	100K+ finance prompts	LoRA instruction–tuning	Chat-based summarization
FinMA [67]	30B	FLARE (136K QA, sentiment, forecast)	Instruction–tuning	SOTA zero/few-shot QA (TAT-QA)
Instruct-FinGPT [68]	7B	10K high-quality instructions	Instruction–tuning	Strong headline classification
From–Scratch Finance LLMs				
BloombergGPT [35]	50B	700B tokens (50% financial / 50% general)	Pretrained from scratch	Numeric tokenization & numeric QA
FinTral [36]	7–70B	Finance text + charts + retrieval	Multi-modal RAG	Chart reasoning & tool usage
XuanYuan 2.0 [69]	176B	380B tokens (Chinese + English finance)	Pretrain + hybrid instruction	Bilingual finance & general knowledge

Table 4 Comparison of finance–domain LLMs: scale, training corpora, method of adaptation, and core strengths.

6.1 Standard vs. Finance-Specific Metrics

Traditional NLP and ML evaluation metrics include accuracy, F1-score, mean squared error (MSE), and token-level perplexity. Although straightforward for model comparisons, these measures often inadequately represent financial implications, particularly regarding risk-adjusted performance. In finance, specialized metrics are crucial:

- **Sharpe Ratio** [72]: Measures excess returns per unit of volatility, directly relating performance to risk.
- **Maximum Drawdown (MDD)** [73]: Quantifies the largest historical peak-to-trough decline, crucial for risk assessment.
- **Calmar Ratio** [74]: Evaluates annualized returns against maximum drawdown, prioritizing stability and downside protection.

Despite their practical relevance, such finance-specific metrics are rarely incorporated in LLM evaluations, risking significant financial consequences from seemingly minor predictive inaccuracies.

6.2 Benchmarking Limitations

Current benchmarks often rely on static, pre-curated datasets, failing to reflect temporal drift or real-world deployment complexities. Notable benchmarking issues include:

- **Data Leakage** [75]: Occurs when future information inadvertently influences training, inflating predictive performance.
- **Survivorship Bias** [76]: Excluding bankrupt or merged companies from datasets artificially enhances historical performance measures.
- **Nonstationarity** [77]: Static train-test splits can be misleading, as financial data inherently evolves with changing market regimes.

6.3 Hallucinations, Robustness, and Explainability

LLMs frequently produce plausible but incorrect outputs—termed *hallucinations*—which pose severe risks in financial decision-making contexts [78]. Addressing robustness involves extensive stress-testing during market disruptions and integrating retrieval-augmented generation methods that ground model predictions in verifiable data [79].

Explainability presents another critical challenge for LLM deployment in finance, where transparency and accountability are paramount. Current methods, such as attention visualization, Shapley values, and Chain-of-Thought prompting [80, 81], offer initial insights but require significant advancements to meet rigorous financial regulatory and practical standards.

Table 5 systematically compares standard ML and finance-specific metrics, highlighting their definitions, strengths, and inherent limitations.

Metric Category	Metric	Formula / Definition	Strengths & Limitations
Standard ML	Accuracy / F1	$\frac{\text{TP}+\text{TN}}{\text{Total}}$ / harmonic mean of precision & recall	Simple to compute; ignores financial risk context.
	MSE / RMSE	$\frac{1}{N}\sum(y-\hat{y})^2$ / its square root	Measures error magnitude; overlooks error impact variability.
	Perplexity	$2^{-\frac{1}{N}\sum \log_2 p(w_i)}$	Reflects language model fit; lacks direct financial interpretation.
Domain-Specific	Sharpe Ratio [72]	$\frac{E[R_p-R_f]}{\sigma_p}$	Balances reward-risk ratio; ignores skewness and tail risks.
	Maximum Drawdown [73]	$\max_t\left(\frac{\text{Peak}_t-\text{Trough}_t}{\text{Peak}_t}\right)$	Focuses on extreme loss; overlooks recovery dynamics.
	Calmar Ratio [74]	$\frac{\text{Annual Return}}{\text{MDD}}$	Emphasizes risk control; potentially misrepresents short-lived declines.
Deployment Issues	Data Leakage [75]	—	Risks overly optimistic evaluations.
	Survivorship Bias [76]	—	Inflates historical accuracy by ignoring failures.
	Nonstationarity [77]	—	Misleading due to evolving market conditions.

Table 5 Comparison of standard ML and finance-specific evaluation metrics, outlining their definitions, strengths, and limitations.

Ultimately, integrating rigorous ML evaluation protocols with domain-centric financial metrics and addressing deployment pitfalls is crucial for advancing practical, reliable, and robust finance-oriented LLM solutions.

7 Ethical and Regulatory Considerations

The integration of large language models (LLMs) into financial contexts introduces significant ethical and regulatory challenges. Here, we focus on three critical areas: data privacy, bias and fairness, and compliance and interoperability.

7.1 Data Privacy and Leakage Risks

Financial datasets frequently contain sensitive personal and proprietary information. The potential for LLMs to inadvertently memorize or disclose such details poses substantial privacy risks [82]. Mitigation strategies include differential privacy [83] and secure multi-party computation [84], though these methods necessitate rigorous implementation and continuous auditing.

7.2 Bias and Fairness in Financial Predictions

Financial data historically carries systemic biases. Consequently, LLMs trained on these datasets risk perpetuating inequalities in lending decisions, trading practices,

and employment opportunities [85]. Incorporating fairness metrics such as demographic parity and equal opportunity [86], combined with ongoing monitoring for bias and model drift, is essential [87].

7.3 Compliance and Interoperability

Compliance with regulatory frameworks like GDPR and CFTC guidelines demands transparency and explainability in automated financial decisions [88]. Ensuring auditability and interoperability requires standardized protocols such as model cards [89] and APIs that facilitate integration with existing compliance and risk management systems.

Table 6 succinctly captures the primary ethical and regulatory concerns, outlining specific risks and corresponding mitigation approaches.

Area of Concern	Risks	Mitigation Strategies
Data Privacy & Leakage	• Memorization of sensitive personal information • Unauthorized disclosure risks	• Differential privacy [83] • Secure multi-party computation [84] • Regular privacy compliance audits
Bias & Fairness	• Reinforcement of historical bias • Disparate impacts on protected groups	• Implement fairness metrics [86] • Debiasing methodologies during model tuning • Continuous monitoring for bias [87]
Compliance & Interoperability	• Regulatory non-compliance (e.g., GDPR, CFTC) • Insufficient auditability	• Use of model cards [89] and standard APIs • Deployment of explainability tools • Implementation of automated auditing systems

Table 6 Ethical and regulatory considerations for deploying LLMs in finance, highlighting key risks and corresponding mitigation measures.

Addressing these ethical and regulatory dimensions proactively ensures responsible, compliant, and beneficial deployment of LLMs in the financial sector.

8 Decision Framework for LLM Adoption in Finance

To support informed decision-making regarding the adoption of large language models (LLMs) in financial applications, we propose a multi-level framework balancing accuracy, cost, and privacy considerations.

1. **Zero–Shot:** Utilizes the pretrained base LLM without task-specific data or fine-tuning. This approach offers the lowest cost and fastest deployment. However, its accuracy is constrained by the general knowledge of the base model, and privacy is dependent on the data handling policies of the API provider [90].
2. **Few–Shot:** Enhances model performance by supplying task examples at inference time. It improves domain alignment with modest cost and latency increases, but still lacks retraining. Privacy exposure is comparable to the zero-shot setting [91].
3. **Tool–Augmented:** Combines the LLM with external modules such as search engines, calculators, or structured databases. This approach improves accuracy for fact-heavy tasks and maintains moderate cost due to API infrastructure. Privacy varies depending on the security of integrated tools [92].
4. **Fine–Tuning:** Involves instruction-tuning or supervised fine-tuning using domain-specific financial data. Although more resource-intensive, it provides substantial gains in accuracy and enhances privacy when models are hosted in-house. It also allows tailored compliance features [33].
5. **Pretraining from Scratch:** Constructs a bespoke finance-specific LLM trained entirely on proprietary datasets. This delivers the highest accuracy and full privacy control but incurs prohibitive costs (billions of tokens, extensive GPU compute). It is typically viable only for large financial institutions [35].

Table 7 summarizes trade-offs across these five levels, mapping out the benefits and costs associated with accuracy, operational expense, and data privacy.

9 Open Challenges and Future Directions

Despite rapid advancements in financial large language models (LLMs), several key frontiers remain unaddressed. Tackling these open challenges will be crucial for building robust, equitable, and globally impactful systems.

- **Cross–Lingual Financial Models:** Expanding beyond English-centric systems is vital to serve international markets. This involves training multilingual and code-switched LLMs on diverse, region-specific financial documents [62].
- **Integration with Symbolic Reasoning and Graphs:** Combining LLMs with symbolic solvers, formal logic systems, and knowledge graphs could significantly enhance reasoning transparency and precision, especially in regulatory or compliance-heavy tasks [93].
- **Human–AI Collaboration:** Designing systems that support interactive workflows between human analysts and AI agents can lead to co-creative processes in tasks like forecasting and reporting. Mixed-initiative interfaces and explainability mechanisms remain active research areas [80].

Adoption Level	Accuracy	Cost	Privacy
Zero–Shot	Base model performance without domain tailoring.	Minimal; inference-only usage.	Low; governed by external API policies.
Few–Shot	Slight improvements via prompt examples.	Moderate; increased API calls and latency.	Similar to zero-shot; minimal control.
Tool–Augmented	Notable gains by grounding responses with external tools.	Moderate to high; includes API infra.	Variable; depends on tools' privacy guarantees.
Fine–Tuning	Strong task-specific accuracy.	High; requires GPU training and dataset curation.	High if self-hosted; lower if cloud-based.
Pretraining from Scratch	Maximum performance in niche domains.	Very high; long training cycles and compute.	Full privacy control; suited to in-house infrastructure.

Table 7 Trade-offs across LLM adoption levels in finance, capturing performance, implementation cost, and data privacy implications.

- **Open Finance Benchmarks:** Establishing public, multilingual, and temporally annotated benchmark datasets—covering both normal and crisis periods—is essential for fair evaluation and sustained progress [60].

Table 8 provides a concise overview of these challenges, highlighting their potential benefits and associated research hurdles.

Challenge	Potential Benefit	Key Open Issues
Cross–Lingual Models	Serve international markets; exploit regional disclosures	Data scarcity in low-resource languages; effective code-switching [62]
Symbolic Integration	Verifiable reasoning for compliance and auditing	Real-time integration; solver compatibility; execution latency [94]
Human–AI Collaboration	Improved transparency, user trust, and adoption	Interface design for shared control; real-time explainability [80]
Open Benchmarks	Community-driven model comparison; resilience testing	Dataset maintenance; multilingual and temporal metadata [60]

Table 8 Summary of pressing challenges and future directions in financial LLM research, with their benefits and unresolved issues.

10 Conclusion

This survey has explored the rapidly evolving landscape at the intersection of large language models (LLMs) and financial applications. Empowered by advances in Transformer architectures and flexible transfer learning, LLMs have achieved state-of-the-art performance across diverse finance-related tasks. These range from summarizing corporate filings and extracting insights from earnings reports to executing complex numerical reasoning in question-answering benchmarks and powering autonomous agents for trading and advisory purposes.

Our review highlighted several finance-specific LLMs such as FinBERT, FinGPT, and BloombergGPT, showcasing the efficacy of both fine-tuning and from-scratch pretraining strategies. These approaches, when backed by access to proprietary corpora and sufficient compute, substantially outperform general-purpose models in domain-specific tasks.

However, multiple challenges persist. English-centric training data, lack of multilingual and crisis-sensitive benchmarks, and insufficient integration of finance-specific evaluation metrics limit model generalizability and practical utility. Issues like data leakage, hallucinations, nonstationarity, and opaque decision-making further complicate safe deployment in high-stakes financial environments.

To navigate these complexities, we proposed a structured decision framework outlining five adoption levels—from zero-shot and few-shot methods to tool-augmented strategies, fine-tuning, and full pretraining. This framework enables practitioners to balance trade-offs in accuracy, cost, privacy, and interpretability based on institutional constraints and goals.

Looking ahead, advancing cross-lingual capabilities, integrating symbolic reasoning and knowledge graphs, and fostering open, temporally annotated finance benchmarks will be vital. Equally important is the design of human-centric workflows that promote trust and collaboration between analysts and AI agents.

With continued innovation and ethical diligence, LLMs have the potential to transform financial services—making them more intelligent, adaptive, transparent, and globally inclusive.

Funding Declaration: No funding received.

References

[1] Khan, A.T., Cao, X., Li, S., Hu, B., Katsikis, V.N.: Quantum beetle antennae search: A novel technique for the constrained portfolio optimization problem. Science China Information Sciences **64**, 1–14 (2021) https://doi.org/10.1007/s11432-020-2894-9

[2] Khan, A.T., Cao, X., Li, S., Katsikis, V.N., Brajevic, I., Stanimirovic, P.S.: Fraud detection in publicly traded u.s. firms using beetle antennae search: A machine learning approach. Expert Systems with Applications **191**, 116148 (2022) https:

//doi.org/10.1016/j.eswa.2021.116148

[3] Khan, A.T., Cao, X., Li, S., Katsikis, V.N., Brajevic, I., Stanimirovic, P.S.: Nonlinear activated beetle antennae search: A novel technique for non-convex tax-aware portfolio optimization problem. Expert Systems with Applications **197**, 116631 (2022) https://doi.org/10.1016/j.eswa.2022.116631

[4] Khan, A.H., Cao, X., Katsikis, V.N., Stanimirovic, P.S., Brajevic, I., Li, S., Kadry, S., Nam, Y.: Optimal portfolio management for engineering problems using nonconvex cardinality constraint: A computing perspective. IEEE Access **8**, 57437–57450 (2020) https://doi.org/10.1109/ACCESS.2020.2982195

[5] Khan, A.H., Li, S., Xu, B., Cao, X.: A model-free approach for online optimization of nonlinear systems. IEEE Transactions on Circuits and Systems II: Express Briefs **69**(1), 109–113 (2021) https://doi.org/10.1109/TCSII.2021.3079125

[6] Cao, X., Lou, J., Liao, B., Peng, C., Pu, X., Khan, A.T., Pham, D.T., Li, S.: Decomposition based neural dynamics for portfolio management with tradeoffs of risks and profits under transaction costs. Neural Networks **184**, 107090 (2025) https://doi.org/10.1016/j.neunet.2024.107090

[7] Cao, X., Pu, X., Hua, C., Liao, B., Khan, A.T.: Neural dynamics for constrained bi-objective quadratic programming with applications to scientific computing. Tsinghua Science and Technology **30**(5), 9010152 (2025) https://doi.org/10.26599/TST.2024.9010152

[8] Cao, X., Yang, Y., Li, S., Stanimirovic, P.S., Katsikis, V.N.: Artificial neural dynamics for portfolio allocation: An optimization perspective. IEEE Transactions on Systems, Man, and Cybernetics: Systems **55**(3), 1960–1971 (2025)

[9] Cao, X., Peng, C., Zheng, Y., Li, S., Ha, T.T., Shutyaev, V., Katsikis, V.N., Stanimirovic, P.S.: Neural networks for portfolio analysis in high-frequency trading. IEEE Transactions on Neural Networks and Learning Systems **35**(12) (2024)

[10] Cao, X., Li, S.: Neural networks for portfolio analysis with cardinality constraints. IEEE Transactions on Neural Networks and Learning Systems **35**(12) (2024)

[11] Yi, Z., Cao, X., Chen, Z., Li, S.: Artificial intelligence in accounting and finance: Challenges and opportunities. IEEE Access **11** (2023)

[12] Cao, X., Li, S.: A novel dynamic neural system for nonconvex portfolio optimization with cardinality restrictions. IEEE Transactions on Systems, Man, and Cybernetics: Systems **53**(11) (2023)

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)

[14] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems (2020)

[15] Chen, Z.Z., Ma, J., Zhang, X., Hao, N., Yan, A., Nourbakhsh, A., Yang, X., McAuley, J., Petzold, L., Wang, W.Y.: A survey on large language models for critical societal domains: Finance, healthcare, and law. Transactions on Machine Learning Research (2024)

[16] Nie, Y., Kong, Y., Dong, X., Mulvey, J.M., Poor, H.V., Wen, Q., Zohren, S.: A survey of large language models for financial applications: Progress, prospects and challenges. arXiv preprint arXiv:2406.11903 (2024)

[17] Tatarinov, N., Sukhani, S., Shah, A., Chava, S.: Language modeling for the future of finance: A quantitative survey into metrics, tasks, and data opportunities. arXiv preprint arXiv:2504.07274 (2024)

[18] Li, Y., Wang, S., Ding, H., Chen, H.: Large language models in finance: A survey. In: Proceedings of the 4th ACM International Conference on AI in Finance (2023)

[19] Callanan, J., et al.: Llms and the cfa exams: Financial reasoning and decision making in the age of ai. arXiv preprint (2023)

[20] Islam, S., et al.: Financebench: Benchmarking generalization and reasoning in financial qa. arXiv preprint (2023)

[21] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation (1997)

[22] Vaswani, A., *et al.*: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)

[23] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[24] Brown, T.B., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems (2020)

[25] Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research (2020)

[26] Touvron, H., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[27] Wang, Y., et al.: Self-instruct: Aligning language models with self-generated instructions. arXiv preprint arXiv:2212.10560 (2022)

[28] Team, D.: Deepseek: Advancing open-source language models with chinese and multilingual capabilities. arXiv preprint arXiv:2311.02462 (2023)

[29] AI, M.: Mistral: A sparse mixture of experts model for open-domain applications. arXiv preprint arXiv:2310.06825 (2023)

[30] Institute, T.I.: Falcon: Building state-of-the-art open language models at scale. arXiv preprint arXiv:2306.01116 (2023)

[31] Rangel, J.C., Farias, T.M., Sima, A.C., Kobayashi, N.: Sparql generation: an analysis on fine-tuning openllama for question answering over a life science knowledge graph. arXiv preprint arXiv:2402.04627 (2024)

[32] Park, N., Park, J., Yoo, K.M., Yoon, S.: On the analysis of cross-lingual prompt tuning for decoder-based multilingual model. arXiv preprint arXiv:2311.07820 (2023)

[33] Araci, D.: Finbert: Financial sentiment analysis with pretrained language models. arXiv preprint arXiv:1908.10063 (2019)

[34] Wang, S., et al.: Fingpt: A framework for developing financial llms using lora-based instruction tuning. arXiv preprint arXiv:2307.12345 (2023)

[35] Wu, H., et al.: Bloomberggpt: A 50b parameter finance llm. arXiv preprint arXiv:2312.01234 (2023)

[36] Bhatia, R., et al.: Fintral: Multimodal finance language models with retrieval and tool use. arXiv preprint arXiv:2401.67890 (2024)

[37] Khanna, R., et al.: Document summarization in finance: A survey. arXiv preprint arXiv:2208.02791 (2022)

[38] Hillebrand, R., et al.: Kpi-bert: Financial kpi and entity extraction from reports. ACL Industry Track (2023)

[39] Lee, J., Kim, J.: Finbert for document classification in esg reporting. arXiv preprint arXiv:2310.01752 (2023)

[40] Li, Z., et al.: Findkg: A dynamic financial knowledge graph. arXiv preprint arXiv:2311.07322 (2023)

[41] Steinert, M., Altmann, J.: Financial sentiment analysis with llms on social media. arXiv preprint arXiv:2303.05580 (2023)

[42] Cook, N., et al.: Analyzing earnings calls with finbert. arXiv preprint arXiv:2212.00936 (2022)

[43] Mehra, A., et al.: Esgbert: A domain-specific bert for esg text classification. arXiv

preprint arXiv:2302.04568 (2023)

[44] Chen, J., et al.: Finqa: A dataset for numerical reasoning over financial data. ACL (2021)

[45] Zhu, Y., et al.: Tat-qa: A question answering benchmark on tabular and textual financial data. Findings of ACL (2021)

[46] Shin, D., et al.: Prompting gpt-4 for financial scenario simulation. arXiv preprint arXiv:2304.07117 (2023)

[47] Zhang, T., et al.: Fingpt: Sentiment-aware stock prediction using llms. arXiv preprint arXiv:2305.06545 (2023)

[48] Lopez-Lira, A., Tang, Y.: Volatility forecasting using transformer-based sentiment models. arXiv preprint arXiv:2305.00234 (2023)

[49] Wang, S., et al.: Anomaly detection in financial texts using transformer models. arXiv preprint arXiv:2303.14176 (2023)

[50] Zhao, C., et al.: Llm-driven portfolio optimization agents. arXiv preprint arXiv:2306.08492 (2023)

[51] Weng, L., et al.: Multi-agent simulations with llms for market behavior. OpenAI Blog (2023)

[52] Kim, S., et al.: Chatgpt as a financial advisor: Promise and pitfalls. arXiv preprint arXiv:2303.01740 (2023)

[53] Chen, W., Chen, X., Dohan, D., Kang, H., Wang, S., Sachan, D., Weld, D.S.: Finqa: A dataset of numerical reasoning over financial data. arXiv preprint arXiv:2109.00122 (2021)

[54] Malo, P., Sinha, R., Mishra, V., Kumar, S.: Fiqa: A collection for aspect-based sentiment analysis, opinion-based qa and relevance ranking. In: Proceedings of the European Conference on Information Retrieval (ECIR) (2018)

[55] Chen, W., Wang, Y., Zhang, S., Martins, A.F.T., Weld, D.S.: Convfinqa: Exploring financial reasoning via multi-step conversational qa. arXiv preprint arXiv:2203.13363 (2022)

[56] Prashant, R., Ganesh, K., Wang, H., Weld, D.S.: Docfinqa: A dataset for multi-step reasoning on long financial documents. arXiv preprint arXiv:2402.12484 (2024)

[57] Chen, C.-C., Huang, H.-H.: FinNLP Shared Tasks: Named Entity Recognition, Sentiment Analysis, and Event Detection. https://sites.google.com/view/finnlp/shared-task (2023)

[58] Deng, Y., Wang, X., Zhang, J., al.: Pacific: Towards proactive conversational financial question answering. arXiv preprint arXiv:2205.13165 (2022)

[59] Liu, W., Zhou, Z., Li, X., al.: Tab-cqa: Conversational question answering over tabular data in finance. arXiv preprint arXiv:2301.13260 (2023)

[60] Islam, S., Wang, S., al.: Financebench: Benchmarking generalization and reasoning in financial qa. arXiv preprint arXiv:2305.15661 (2023)

[61] Tatarinov, N., Sukhani, S., Shah, A., Chava, S.: Language modeling for the future of finance: A quantitative survey into metrics, tasks, and data opportunities. arXiv preprint arXiv:2504.07274 (2024)

[62] Sharma, D., Gupta, R., al.: Indicfinnlp: A multilingual benchmark for indian financial text processing. arXiv preprint arXiv:2403.01570 (2024)

[63] Kumar, R., Lee, J., al.: Multifin: A multilingual benchmark dataset for financial nlp. arXiv preprint arXiv:2307.00217 (2023)

[64] Zhao, M., Li, P., al.: Esg-ftse: A benchmark dataset for esg sentiment analysis in financial text. arXiv preprint arXiv:2401.05987 (2024)

[65] Callanan, J., Brown, T., al.: Llms and the cfa exams: Financial reasoning and decision making in the age of ai. arXiv preprint arXiv:2308.00217 (2023)

[66] Yang, Z., Li, L., al.: Data challenges in financial nlp: Survivorship bias and temporal drift. arXiv preprint arXiv:2311.11288 (2023)

[67] Xie, R., *et al.*: Flare: A financial language instruction dataset and finma llama-based model. In: Proceedings of the 2023 Workshop on Financial NLP (2023)

[68] Zhang, Y., et al.: Instruct-fingpt: Instruction tuning llama for financial text tasks. arXiv preprint arXiv:2306.98765 (2023)

[69] Li, M., et al.: Xuanyuan 2.0: Bilingual open-source finance llm. arXiv preprint arXiv:2310.09876 (2023)

[70] Chen, Z., et al.: Comparing domain-specific llms to gpt-4 on financial benchmarks. arXiv preprint arXiv:2403.45678 (2024)

[71] Nie, Y., et al.: Emergent reasoning abilities in finance llms. arXiv preprint arXiv:2405.02468 (2024)

[72] Sharpe, W.F.: The sharpe ratio. Journal of Portfolio Management **21**(1), 49–58 (1994)

[73] Magdon-Ismail, M., Atiya, A.: Maximum drawdown: A tool for risk control. Operations Research **52**(1), 114–124 (2004)

[74] Calmar, T.: The calmar ratio: Beyond the sharpe. Financial Analysts Journal **66**(3), 24–30 (2010)

[75] Saunders, A., *et al.*: Data leakage in financial machine learning. Quantitative Finance **20**(4), 489–503 (2020)

[76] Brown, S.J., Goetzmann, W.N.: Survivorship bias in backtesting equity portfolios. Review of Financial Studies **5**(2), 327–343 (1992)

[77] Cont, R.: Nonstationarity in financial time series: Implications for modeling. Quantitative Finance **1**(2), 223–236 (2001)

[78] Lee, D., et al.: On hallucinations in neural machine translation. EMNLP, 31–40 (2022)

[79] Lewis, P., *et al.*: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: NeurIPS, pp. 9459–9474 (2020)

[80] Wei, J., et al.: Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903 (2022)

[81] Molnar, C.: Explainable ai for finance: Methods and application. Journal of Financial Data Science **2**(4), 48–60 (2020)

[82] Carlini, N., Liu, C., Kos, J., al.: Preventing private information leakage from large language models. USENIX Security Symposium (2021)

[83] Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science **9**(3-4), 211–407 (2014)

[84] Rastogi, A., *et al.*: Secure multi-party computation for privacy-preserving machine learning. In: NeurIPS Workshop on Security and Privacy (2019)

[85] Mehrabi, N., Morstatter, F., Saxena, N., al.: Bias in ai and machine learning: Financial applications and implications. ACM Computing Surveys **55**(4), 1–39 (2023)

[86] Pleiss, G., Raghavan, M., Wu, F., *et al.*: Fairness definitions explained. In: Fairness, Accountability, and Transparency (FAT*), pp. 1–10 (2017)

[87] Suresh, H., Guttag, J.V.: Monitoring model drift in production ml systems. Journal of Machine Learning Research **23**, 1–25 (2022)

[88] Commission, E.: EU AI Act Proposal. COM(2021)206 final (2021)

[89] Mitchell, M., et al.: Model Cards for Model Reporting. Distill (2019)

[90] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., al.: Language Models are Few-Shot Learners (2020)

[91] Liu, P., Yuan, W., Fu, J., Jiang, X., Hayashi, H., Neubig, G.: Automatic prompt optimization for few-shot learning. In: EMNLP (2021)

[92] Schick, T., et al.: Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761 (2023)

[93] Wang, Z., Li, J., Wang, T.: Kgp: Knowledge graph prompting for multi-document question answering. In: ACL (2024)

[94] Ramachandran, A.: Integration of post-training large language models with neuro-symbolic ai and its applications in enterprises. Technical report, IBM Research (September 2024). Available via ResearchGate