A Survey of Financial Large Language Models: From Domain Adaptation to Agentic Financial Intelligence

Yue Dai
PBOC Research Institute
dyue@pbc.mail.cn

Abstract

Financial large language models (FinLLMs) have rapidly evolved from domain-adapted encoders for sentiment analysis into instruction-tuned, retrieval-augmented, multimodal, and agentic systems for financial analysis. This survey synthesizes recent work on FinLLMs across data construction, model adaptation, retrieval-augmented generation, benchmarks, applications, and risk control. We organize the literature into six layers: financial corpora and data governance, domain-specific language modeling, financial reasoning and numerical analysis, retrieval and evidence grounding, multimodal and time-series integration, and agentic workflow automation. We also discuss adjacent financial research on analyst information environments, sustainability disclosure, satellite imagery, segment reporting, and volatility modeling, because these domains define realistic use cases and evaluation targets for FinLLM systems. Our central argument is that the next generation of FinLLMs should be evaluated less as isolated text generators and more as auditable decision-support systems operating under temporal, regulatory, and economic constraints.

1 Introduction

Financial NLP has long relied on domain dictionaries, supervised classifiers, and task-specific representations for filings, news, earnings calls, analyst reports, and risk disclosures [87, 64, 43, 37]. Large language models (LLMs) changed the technical baseline by enabling instruction following, few-shot learning, long-document summarization, retrieval-grounded question answering, and tool-using agents [25, 6, 50, 35]. Finance is an unusually demanding domain for these capabilities. Financial statements contain dense numeric and tabular information; market data are non-stationary; disclosures are governed by regulation; and outputs can affect investment, audit, and risk-management decisions.

The FinLLM literature now includes finance-specialized pretraining [107], open-source financial LLM frameworks [113, 59, 67], Chinese and multilingual financial models [11, 36, 112, 72, 45], data-centric construction methods [23, 104], multimodal systems [103, 5, 38, 116], retrieval-based financial question answering [40, 79, 10, 99, 21, 123], and agentic platforms [89, 33, 22, 17, 18]. Surveys and readiness studies have begun to map the space [53, 48, 122, 57, 31].

This paper contributes a structured synthesis of the current FinLLM landscape. It emphasizes three ideas. First, financial language modeling is data-centric: timestamps, licenses, source provenance, and market microstructure assumptions are part of the model design. Second, benchmark design must move beyond static answer accuracy toward evidence-grounded and temporally realistic workflows. Third, FinLLMs should be connected to substantive financial research problems, including disclosure quality, analyst monitoring, ESG greenwashing, alternative data, and volatility regimes [63, 24, 62, 15, 16].

2 Scope and Literature Collection

The survey focuses on papers that construct, adapt, evaluate, or deploy language models for financial tasks. We include verified arXiv papers, selected peer-reviewed finan-

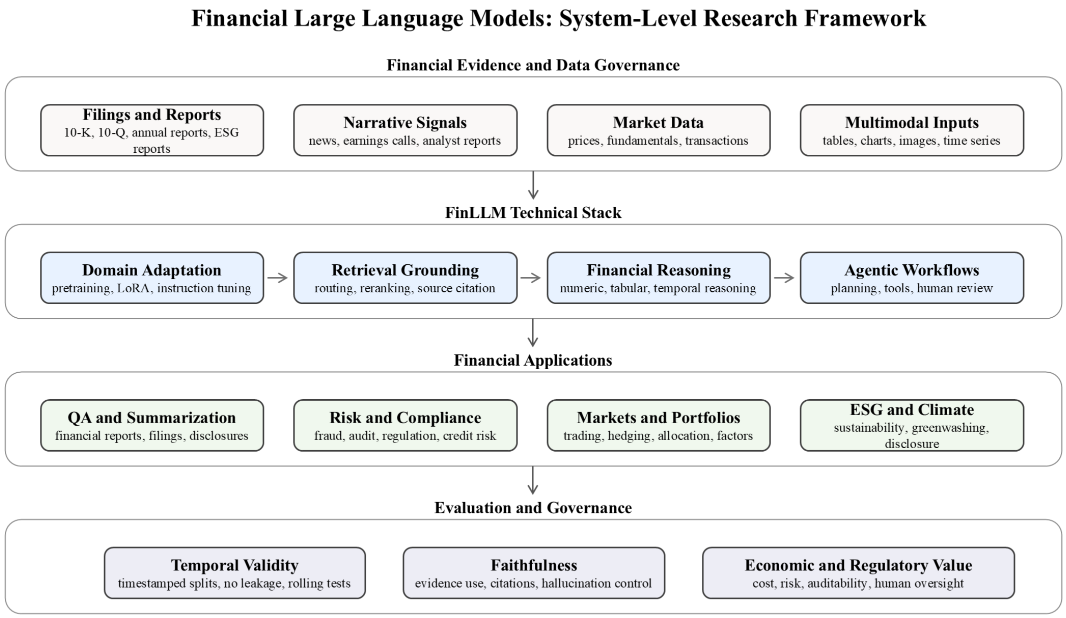

Figure 1: Overview of financial large language model research. The field can be read as a pipeline from governed financial evidence to domain-adapted models, retrieval-grounded and multimodal reasoning, agentic workflows, application-specific deployment, and evaluation under temporal, evidential, economic, and regulatory constraints.

cial NLP and accounting/finance papers, and user-specified papers with DOI, arXiv, SSRN, or conference metadata. The reference set covers more than one hundred papers, but the local PDF download policy is narrower: only core, high-relevance, open-access, or repeatedly cited technical papers are downloaded to the project folder. Less central papers are retained in BibTeX and cited in the narrative.

We treat FinLLM work as a layered system rather than a single model class. The layers are: (i) corpora, data pipelines, and governance; (ii) pretraining and adaptation; (iii) instruction tuning and preference alignment; (iv) retrieval and grounding; (v) multimodal and temporal modeling; (vi) agents and workflow orchestration; and (vii) evaluation, safety, and deployment.

3 From Financial NLP to FinLLMs

3.1 Pre-LLM Financial Text Analysis

Traditional financial NLP created task-specific representations for sentiment, risk, fraud, and disclosure analysis. Dictionary-based approaches, especially domain-specific financial dictionaries, demonstrated that general sentiment lexicons often fail in accounting and finance contexts [64]. Supervised learning and topic modeling were used to infer risk from 10-K filings, identify analyst information roles, and quantify media sentiment [43, 87, 37]. These studies remain important because they define financial validity: a model's language output matters only if it connects to disclosure, governance, risk, or market outcomes.

Table 1: A layered taxonomy of financial large language model research.

Layer	Main technical question	Representative topics	Example citations
Data and governance	What financial evidence can be used, updated, licensed, and audited?	filings, earnings calls, reports, news, synthetic instructions, timestamps, contamination checks	[23, 104, 118]
Model adaptation	How should general LLMs be specialized for finance?	pretraining, continual pretraining, instruction tuning, LoRA, quantization, multilingual adaptation	[107, 113, 11, 94]
Retrieval and grounding	How can answers be tied to authoritative documents?	RAG, reranking, document routing, evidence curation, financial QA	[50, 40, 79, 17]
Numerical and multimodal reasoning	How can models handle tables, charts, PDFs, time series, and imagery?	report tables, chart QA, visual citation, satellite data, time-series forecasting	[3, 103, 5, 24]
Agents and workflows	Can systems complete multi-step financial analysis tasks?	tool use, search agents, investment memos, portfolio agents, compliance workflows	[89, 33, 22]
Evaluation and risk	How should reliability be measured under financial constraints?	benchmarks, hallucination, temporal leakage, bias, auditability, economic validity	[57, 31, 41]

3.2 Encoder Models and Financial Sentiment

FinBERT and related encoder models made financial sentiment analysis a central benchmark for domain adaptation [2, 32, 66, 39]. Later work explored instruction-tuned sentiment models, Chinese financial sentiment, company-specific bias, bond-market sentiment, and in-context learning for financial sentiment analysis [119, 47, 30, 120, 28, 77, 9, 88, 81, 70, 4, 127, 69, 105, 61, 93]. This line shows a recurring pattern: generic LLMs can be competitive, but domain adaptation, label quality, and context design remain decisive.

4 Financial Data and Corpus Construction

Financial data are heterogeneous, high-volume, and temporally ordered. Filings, earnings calls, news, prices, analyst reports, macro releases, and social media each have different noise patterns and legal constraints. Data-centric FinLLM work therefore emphasizes collection, cleaning, deduplication, instruction synthesis, and governance [23, 104, 118]. FinGPT's open-source pipeline approach explicitly foregrounds dynamic financial data acquisition and update cycles [113, 59, 67, 96].

Data quality is also a central issue in disclosure-oriented work. LLMs have been proposed for financial report extraction, segment disclosure comparability, long-document information extraction, and regulatory interpretation [78, 62, 8, 18, 95]. These tasks require traceable document spans, numeric consistency, and consistent definitions across firms and periods. In sustainability and ESG contexts, FinLLMs may help parse sustainability reports, detect greenwashing-related language, and connect textual disclosures to governance mechanisms [63, 106, 68, 92]. The risk is that a fluent model may normalize weak or performative disclosure unless evaluation explicitly measures evidence and comparability.

5 Model Families and Adaptation Strategies

5.1 Finance-specialized Pretraining

BloombergGPT is a landmark finance-specialized model trained on a mixture of financial and general corpora [107]. It demonstrates that domain-intensive pretraining can improve financial NLP while preserving broad language ability. Related work explores high-performance pretraining, domain-specific Japanese and Dutch models, and multilingual or bilingual financial LLMs [60, 34, 72, 36, 45]. These models suggest that finance is not a single universal domain: language, regulation, disclosure rules, and market institutions differ across jurisdictions.

5.2 Open-source FinLLM Frameworks

Open-source FinLLM research aims to reduce dependence on proprietary financial models. FinGPT and its extensions emphasize democratized data pipelines, instruction tuning, and open benchmarks [113, 59, 96, 54, 27]. DISC-FinLLM and other Chinese systems highlight expert-guided tuning, localized data, and Chinese benchmark construction [11, 112, 124, 85]. KodeX, FinLoRA, and leaderboard-oriented fine-tuning studies explore practical adaptation under compute constraints [75, 94, 76].

5.3 Efficient and On-device Adaptation

Parameter-efficient tuning and small language models are important for financial deployment because data may be proprietary and latency-sensitive. LoRA-style adaptation is a general technical foundation [35]; finance-specific work has explored quantization, small-language-model readiness, efficient fine-tuning, and on-device or energy-aware RAG [74, 44, 19, 12]. These studies are relevant when financial institutions need local inference, auditability, and lower operational cost.

6 Retrieval-Augmented Financial Reasoning

6.1 Why Retrieval Matters

Finance is a natural setting for retrieval-augmented generation (RAG) because many answers must be grounded in current filings, tables, policies, or news. General RAG improves knowledge-intensive NLP by conditioning generation on external documents [50]. Financial RAG extends this idea to noisy PDFs, standardized documents, filings, regulatory material, and real-time news [79, 42, 99, 21, 20, 17, 18].

6.2 Financial Question Answering

Financial QA has become one of the most important FinLLM evaluation settings. Early work studied DeBERTa-based financial QA, zero-shot QA over financial documents, and FinanceBench [101, 73, 40]. More recent work introduces long-form financial QA, multi-document QA, end-to-end QA pipelines, FinBERT-QA, FinDER, FinSage, FinGEAR, and multi-reranker systems [10, 80, 83, 49, 117, 21, 99, 97]. The key challenge is not only answer correctness but also evidence sufficiency: a system should retrieve the right documents, cite the right spans, and avoid unsupported inference.

6.3 Reranking, Routing, and Evidence Curation

Recent RAG work increasingly treats retrieval as a multi-stage workflow. Financial reports contain multiple document types and sections, so retrieval may benefit from query routing, document-level routing, hybrid sparse-dense search, and local fine-tuning [79, 49, 123, 20, 17]. The user-provided RAG papers fit this line: one emphasizes hybrid document-routed retrieval for the robustness-precision trade-off [17]; the other studies financial report QA with reranking analysis [18]. A broader RAG robustness question is whether a system can detect and resist misleading retrieved context under knowledge conflict [13]. These papers should be positioned as financial RAG system papers rather than generic FinLLM model papers.

7 Numerical, Tabular, and Multimodal Finance

Financial reasoning requires numbers, tables, charts, and time series. Numerical reasoning over financial reports is a known bottleneck [3]; numeric-sensitive models and Text2SQL-style systems target this gap [85, 71]. Multimodal work extends LLMs to financial charts, long reports, and visual evidence [103, 7, 5, 38, 116, 123]. FinRAGBench-V is especially important because it asks for visual citation, moving evaluation closer to auditable financial analysis [123].

Alternative data also matters. Satellite imagery and geospatial information can reveal local economic activity and information asymmetry, as shown in work on Chinese B-share discounts [24]. Although such studies are not FinLLM papers by themselves, they motivate multimodal financial foundation models that combine text, imagery, geospatial signals, and market data.

8 Markets, Forecasting, and Trading

FinLLMs are frequently applied to stock movement prediction, factor discovery, volatility forecasting, portfolio management, and trading agents. Work in this area includes multimodal stock prediction, breakout trading, stock-chain RAG, interpretable stock movement prediction, time-series-specialized LLMs, sentiment-driven trading systems, multi-agent trading, and fund investment benchmarks [110, 121, 91, 52, 98, 125, 109, 114, 51, 100, 55]. Adjacent reinforcement-learning studies on option hedging and portfolio optimization provide useful baselines and task designs for LLM-assisted trading research [14, 84, 102]. Agent and ensemble approaches include institutional portfolio management, end-to-end trading systems, and multimodal trading agents [1, 33, 26, 125].

This literature must be evaluated carefully. Backtests can be contaminated by look-ahead bias, survivorship bias, execution assumptions, and prompt selection. Volatility-regime studies and cross-market forecasting papers are relevant because they remind FinLLM researchers that market dynamics are state-dependent [15, 16]. A credible LLM trading evaluation should report timestamps, data availability, transaction costs, baseline strategies, turnover, drawdown, and regime sensitivity.

9 Benchmarks and Evaluation

9.1 Knowledge, Reasoning, and Readiness Benchmarks

FinLLM benchmarks evaluate domain knowledge, reasoning, sentiment, summarization, QA, and readiness. FinEval, FinDABench, FinBen, FLAME, Open FinLLM Leaderboard, Golden Touchstone, and advanced financial reasoning benchmarks are representative [58, 31, 57, 108, 82]. Some benchmarks focus on Chinese finance or bilingual evaluation, while others evaluate institutional readiness or CFA-style reasoning. Benchmarking suites for agents and financial LLMs broaden the unit of evaluation from text completion to task performance [57].

9.2 Task-specific Benchmarks

Task-specific work covers NER, report summarization, report analysis, financial tagging, climate finance, regulatory challenges, and financial education QA [65, 115, 10, 100, 68, 95]. These datasets are useful because financial competence is multidimensional. A model that performs well on sentiment may fail at numeric table reasoning; a model that summarizes filings may still hallucinate unsupported trends.

9.3 Hallucination, Bias, and Robustness

Finance magnifies hallucination risk. Studies have examined hallucination in financial reports, empirical deficiencies of LLMs in finance, company-specific sentiment bias, retrieval vulnerability, and detection/editing of hallucinations [78, 41, 70, 123]. RAG helps but does not eliminate hallucination; retrieved evidence can be irrelevant, outdated, or misread. Robustness evaluation should therefore test adversarial headlines, missing documents, stale filings, conflicting evidence, and market regime shifts.

10 Agents and Workflow Automation

Financial tasks are multi-step: collect evidence, clean data, compute metrics, retrieve filings, compare peers, write an investment memo, and expose uncertainty. Agent systems

Table 2: Representative FinLLM systems and benchmarks covered in this survey. The table is illustrative rather than exhaustive.

Work	Type	Primary focus	Why it matters
BloombergGPT	Domain-pretrained model	financial corpora plus general language ability	Establishes large-scale finance-specific pretraining as a strong baseline [107].
FinGPT	Open-source framework	data pipelines and instruction tuning	Makes financial LLM adaptation reproducible and easier to update [113, 59].
DISC-FinLLM	Localized model	Chinese financial expert tuning	Shows the importance of jurisdiction, language, and expert data [11].
FinBench / FinanceBench	QA benchmark	financial document question answering	Pushes evaluation toward realistic document-grounded analysis [40].
FinBen / FLAME	Benchmark suite	broad financial capability assessment	Moves beyond single-task sentiment or QA evaluation [31].
FinTral / Open-FinLLMs	Multimodal model family	reports, tables, charts, and multimodal finance	Expands FinLLMs beyond text-only reasoning [5, 38].
FinSage / FinDER	RAG systems and datasets	financial filings QA and retrieval evaluation	Makes evidence quality and retrieval strategy central [99, 21].
FinRobot / FinAgentBench	Agent platform and benchmark	multi-step financial workflows	Evaluates tool use and agentic retrieval rather than isolated generation [22].

attempt to automate these workflows through decomposition, tool use, and reflection. Examples include FinRobot, FinGPT search agents, heterogeneous agents for sentiment, multi-agent reflection for QA, FinSphere, FinTeam, role-aware education QA, and agentic retrieval benchmarks [111, 89, 29, 33, 22].

Agentic systems require stricter evaluation than single-turn models. Intermediate actions should be logged; tool calls should be reproducible; retrieved evidence should be cited; and final answers should separate fact from inference. Secure and auditable RAG platforms are relevant because financial institutions need traceability and permission control [99, 17]. Short-text embedding and semantic retrieval studies, including work outside finance, can inform system design when retrieval must operate over noisy snippets rather than long filings [46].

11 Applications Beyond Core NLP

11.1 Disclosure, Audit, and Regulation

LLMs are increasingly used for disclosure completeness, segment reporting, regulatory interpretation, auditing, fraud detection, and compliance verification [62, 8, 95, 90, 86]. These applications differ from market prediction because they prioritize consistency, completeness, and audit trails. A disclosure assistant should not simply generate plausible text; it should identify missing segment information, compare peer disclosures, and cite authoritative rules.

11.2 Sustainability and Climate Finance

Sustainability reporting and climate finance are important FinLLM applications because they combine narrative disclosure, quantitative targets, and external verification. Work on analyst following and greenwashing shows that information intermediaries can affect firms' disclosure behavior [63]. SusGen-GPT and Climate Finance Bench illustrate how LLMs can support sustainability-report generation and climate-finance reasoning [106, 68]. These applications require careful treatment of greenwashing: the model should detect weak evidence rather than transform vague claims into polished language.

11.3 Credit Risk and Model Explainability

Credit risk management is another natural deployment setting for financial AI because model explanations must be stable enough for review, governance, and user trust. Work on SHAP stability in credit-card default modeling illustrates the need to evaluate not only predictive performance but also explanation robustness [56]. This is relevant for FinLLMs because generated rationales, retrieved evidence, and post-hoc explanations can all become part of the decision record.

11.4 Financial Education and Literacy

LLMs can also support financial literacy, education QA, and decision support for non-experts [44, 126]. This setting requires calibration and guardrails, because users may interpret fluent text as personalized financial advice. Evaluation should include refusal behavior, explanation quality, and user comprehension.

12 Open Problems

Temporal validity. Financial language and market relationships change over time. Benchmarks should use timestamped corpora, rolling-window tests, and post-release evaluation to avoid leakage.

Evidence-grounded generation. FinLLMs should cite source spans, preserve numeric values, and distinguish retrieved facts from model inference. RAG evaluation should score retrieval, reranking, evidence use, and final answer separately.

Data governance. Financial corpora may include licensed reports, proprietary data, personal information, and non-public material. Dataset cards, source licenses, deduplication logs, and contamination tests should become standard.

Economic validity. For trading and forecasting, NLP metrics are insufficient. Evaluation should include transaction costs, risk-adjusted returns, drawdowns, regime analysis, and economic baselines.

Multimodal reliability. Charts, tables, PDFs, satellite images, and time series require different error checks. Multimodal FinLLMs should be evaluated on source-grounded visual and numeric claims, not only final summaries.

Auditable agents. Agent systems need logs, deterministic replay, permission control, and human review points. This is especially important for compliance, investment research, and financial advisory settings.

13 Conclusion

FinLLMs are becoming an integrated research area connecting NLP, machine learning systems, accounting, finance, and responsible AI. The field has moved beyond applying generic LLMs to financial text: it now includes domain pretraining, open-source adaptation, RAG systems, multimodal models, agent platforms, and financial readiness benchmarks. The next step is methodological consolidation. FinLLM research should prioritize verified data provenance, temporally realistic evaluation, evidence-grounded outputs, and auditable workflows. These requirements are not peripheral; they define whether FinLLMs can support real financial analysis rather than merely produce fluent financial prose.

References

[1] Yoshia Abe, Shuhei Matsuo, Ryoma Kondo, and Ryohei Hisano. Leveraging large language models for institutional portfolio management: Persona-based ensembles, 2024.

[2] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019.

[3] Abhinav Arun, Ashish Dhiman, Mehul Soni, and Yibei Hu. Numerical reasoning for financial reports, 2023.

[4] Abraham Atsiwo. Financial sentiment analysis: Leveraging actual and synthetic data for supervised fine-tuning, 2024.

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

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[7] Tianyu Cao, Natraj Raman, Danial Dervovic, and Chenhao Tan. Characterizing multimodal long-form summarization: A case study on financial reports, 2024.

[8] Zhiyu Cao and Zachary Feinstein. Large language model in financial regulatory interpretation, 2024.

[9] Chung-Chi Chen, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. Enhancing financial sentiment analysis with expert-designed hint, 2024.

[10] Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. Fintextqa: A dataset for long-form financial question answering, 2024.

[11] Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, and Zhongyu Wei. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning, 2023.

[12] Wei Chen, Liangmin Wu, Yunhai Hu, Zhiyuan Li, Zhiyuan Cheng, Yicheng Qian, Lingyue Zhu, Zhipeng Hu, Luoyi Liang, Qiang Tang, Zhen Liu, and Han Yang. Autoneural: Co-designing vision-language models for npu inference, 2025.

[13] Yihang Chen, Pin Qian, Su Wang, Sipeng Zhang, Huan Xu, Shuhuai Lin, and Xinpeng Wei. Does rag know when retrieval is wrong? diagnosing context compliance under knowledge conflict, 2026.

[14] Ziheng Chen, Minxuan Hu, Jiayu Yi, and Wenxi Sun. Reinforcement learning for option hedging: Static implied-volatility fit versus shortfall-aware performance, 2026.

[15] Kai Cheng, Xiaoxi Qi, Zhiyuan Cheng, and Longying Lai. Volatility persistence and model choice in cross-market volatility forecasting. *Available at SSRN 6610278*, 2026.

[16] Kai Cheng, Xiaoxi Qi, Zhiyuan Cheng, Longying Lai, and Xuan Liu. Regime-dependent volatility dynamics: Evidence from time-series analysis. In *Proceedings of the 2026 3rd International Conference on Applied Economics, Management Science and Social Development*, pages 179–189, 2026.

[17] Zhiyuan Cheng, Longying Lai, and Yue Liu. Resolving the robustness-precision trade-off in financial rag through hybrid document-routed retrieval, 2026.

[18] Zhiyuan Cheng, Longying Lai, Yue Liu, Kai Cheng, and Xiaoxi Qi. Enhancing financial report question-answering: A retrieval-augmented generation system with reranking analysis, 2026.

[19] Zhiyuan Cheng, Longying Lai, Yue Liu, and Yu Sun. Toward sustainable on-device intelligence: A survey on energy-efficient rag systems with small language models. *Available at SSRN 6698538*, 2026.

[20] Jaeyoung Choe, Jihoon Kim, and Woohwan Jung. Hierarchical retrieval with evidence curation for open-domain financial question answering on standardized documents, 2025.

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

[22] Chanyeol Choi, Jihoon Kwon, Alejandro Lopez-Lira, Chaewoon Kim, Minjae Kim, Juneha Hwang,

Jaeseon Ha, Hojun Choi, Suyeol Yun, Yongjin Kim, and Yongjae Lee. Finagentbench: A benchmark dataset for agentic retrieval in financial question answering, 2025.

[23] Zhixuan Chu, Huaiyu Guo, Xinyuan Zhou, Yijia Wang, Fei Yu, Hong Chen, Wanqing Xu, Xin Lu, Qing Cui, Longfei Li, Jun Zhou, and Sheng Li. Data-centric financial large language models, 2023.

[24] Yue Dai, Ming Chen, and Zhi Zuo. Neighbors in space: Satellite imagery and chinese b-share discount. *China Economic Review*, 82:102063, 2023.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[26] Qianggang Ding, Haochen Shi, Jiadong Guo, and Bang Liu. Tradexpert: Revolutionizing trading with mixture of expert llms, 2024.

[27] Prudence Djagba and Chimezie A. Odinakachukwu. Assessing the capabilities and limitations of fingpt model in financial nlp applications, 2025.

[28] Sorouralsadat Fatemi and Yuheng Hu. A comparative analysis of fine-tuned llms and few-shot learning of llms for financial sentiment analysis, 2023.

[29] Sorouralsadat Fatemi and Yuheng Hu. Enhancing financial question answering with a multi-agent reflection framework, 2024.

[30] Georgios Fatouros, John Soldatos, Kalliopi Kouroumali, Georgios Makridis, and Dimosthenis Kyriazis. Transforming sentiment analysis in the financial domain with chatgpt, 2023.

[31] Jiayu Guo, Yu Guo, Martha Li, and Songtao Tan. Flame: Financial large-language model assessment and metrics evaluation, 2025.

[32] Asier Gutierrez-Fandino, Miquel Noguer i Alonso, Petter Kolm, and Jordi Armengol-Estape. Fineas: Financial embedding analysis of sentiment, 2021.

[33] Shijie Han, Jingshu Zhang, Yiqing Shen, Kaiyuan Yan, and Hongguang Li. Finsphere, a real-time stock analysis agent powered by instruction-tuned llms and domain tools, 2025.

[34] Masanori Hirano and Kentaro Imajo. Construction of domain-specified japanese large language model for finance through continual pre-training, 2024.

[35] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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

[37] Allen H. Huang, Reuven Lehavy, Amy Y. Zang, and Rong Zheng. Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, 64(6):2833–2855, 2018.

[38] Jimin Huang, Mengxi Xiao, Dong Li, Zihao Jiang, Yuzhe Yang, Yifei Zhang, Lingfei Qian, Yan Wang, Xueqing Peng, Yang Ren, Ruoyu Xiang, Zhengyu Chen, Xiao Zhang, Yueru He, Weiguang Han, Shunian Chen, Lihang Shen, Daniel Kim, Yangyang Yu, Yupeng Cao, Zhiyang Deng, Haohang Li, Duanyu Feng, Yongfu Dai, VijayaSai Somasundaram, Peng Lu, Guojun Xiong, Zhiwei Liu, Zheheng Luo, Zhiyuan Yao, Ruey-Ling Weng, Meikang Qiu, Kaleb E Smith, Honghai Yu, Yanzhao Lai, Min Peng, Jian-Yun Nie, Jordan W. Suchow, Xiao-Yang Liu, Benyou Wang, Alejandro Lopez-Lira, Qianqian Xie, Sophia Ananiadou, and Junichi Tsujii. Open-finllms: Open multimodal large language models for financial applications, 2024.

[39] Pau Rodriguez Inserte, Mariam Nakhle, Raheel Qader, Gaetan Caillaut, and Jingshu Liu. Large language model adaptation for financial sentiment analysis, 2024.

[40] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering, 2023.

[41] Haoqiang Kang and Xiao-Yang Liu. Deficiency of large language models in finance: An empirical examination of hallucination, 2023.

[42] Sejong Kim, Hyunseo Song, Hyunwoo Seo, and Hyunjun Kim. Optimizing retrieval strategies for financial question answering documents in retrieval-augmented generation systems, 2025.

[43] Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. Predicting risk from financial reports with regression. *Proceedings of NAACL-HLT*, pages 272–280, 2009.

[44] Tagore Rao Kosireddy, Jeffrey D. Wall, and Evan Lucas. Exploring the readiness of prominent small language models for the democratization of financial literacy, 2024.

[45] KBTG Labs, Atthakorn Petchsod, Pornchanan Balee, Danupat Khamnuansin, Anuruth Lertpiya, Chanatip Saetia, Tawunrat Chalothorn, Thadpong Pongthawornkamol, and Monchai Lertsutthiwong. Thai financial domain adaptation of thalle – technical report, 2024.

[46] Longying Lai, Zhiyuan Cheng, Kai Cheng, and Xiaoxi Qi. Do transformers always win? an empirical study of semantic embeddings for short-text e-commerce reviews. In *2026 9th International Symposium on Big Data and Applied Statistics (ISBDAS)*, pages 525–529, 2026.

[47] Yinyu Lan, Yanru Wu, Wang Xu, Weiqiang Feng, and Youhao Zhang. Chinese fine-grained financial sentiment analysis with large language models, 2023.

[48] Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. A survey of large language models in finance (finllms), 2024.

[49] Joohyun Lee and Minji Roh. Multi-reranker: Maximizing performance of retrieval-augmented generation in the financerag challenge, 2024.

[50] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

[51] Changlun Li, Yao Shi, Chen Wang, Qiqi Duan, Runke Ruan, Weijie Huang, Haonan Long, Lijun Huang, Nan Tang, and Yuyu Luo. Time travel is cheating: Going live with deepfund for real-time fund investment benchmarking, 2025.

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

[53] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey, 2023.

[54] Yixuan Liang, Yuncong Liu, Neng Wang, Hongyang Yang, Boyu Zhang, and Christina Dan Wang. Fingpt: Enhancing sentiment-based stock movement prediction with dissemination-aware and context-enriched llms, 2024.

[55] Luyun Lin, Lixing Lin, Zhen Zhang, Moxuan Zheng, and Yiqing Wang. A volume-price-adjusted macd trading strategy with sensitivity calibration for u.s. equity indices, 2026.

[56] Luyun Lin and Yiqing Wang. Shap stability in credit risk management: A case study in credit card default model. *Risks*, 13(12):238, 2025.

[57] Shengyuan Colin Lin, Felix Tian, Keyi Wang, Xingjian Zhao, Jimin Huang, Qianqian Xie, Luca Borella, Matt White, Christina Dan Wang, Kairong Xiao, Xiao-Yang Liu Yanglet, and Li Deng. Open finllm leaderboard: Towards financial ai readiness, 2025.

[58] Shu Liu, Shangqing Zhao, Chenghao Jia, Xinlin Zhuang, Zhaoguang Long, Jie Zhou, Aimin Zhou,

Man Lan, Qingquan Wu, and Chong Yang. Findabench: Benchmarking financial data analysis ability of large language models, 2024.

[59] Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. Fingpt: Democratizing internet-scale data for financial large language models, 2023.

[60] Xiao-Yang Liu, Jie Zhang, Guoxuan Wang, Weiqing Tong, and Anwar Walid. Fingpt-hpc: Efficient pretraining and finetuning large language models for financial applications with high-performance computing, 2024.

[61] Yiwei Liu, Junbo Wang, Lei Long, Xin Li, Ruiting Ma, Yuankai Wu, and Xuebin Chen. Mulfsa: Multi-level financial sentiment analysis framework for bond market, 2025.

[62] Yue Liu, Zhiyuan Cheng, and Longying Lai. Improving the completeness and comparability of segment disclosures: A large language model approach. *Available at SSRN 6720239*, 2026.

[63] Yuxiang Liu, Jiewei Zhang, and Yue Dai. Analyst following and greenwashing decision. *Finance Research Letters*, 58:104510, 2023.

[64] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[65] Yi-Te Lu and Yintong Huo. Financial named entity recognition: How far can llm go?, 2025.

[66] Wei Luo and Dihong Gong. Pre-trained large language models for financial sentiment analysis, 2024.

[67] Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Le Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. Fingpt: Large generative models for a small language, 2023.

[68] Rafik Mankour, Yassine Chafai, Hamada Saleh, Ghassen Ben Hassine, Thibaud Barreau, and Peter Tankov. Climate finance bench, 2025.

[69] Hamid Moradi-Kamali, Mohammad-Hossein Rajabi-Ghozlou, Mahdi Ghazavi, Ali Soltani, Amirreza Sattarzadeh, and Reza Entezari-Maleki. Market-derived financial sentiment analysis: Context-aware language models for crypto forecasting, 2025.

[70] Kei Nakagawa, Masanori Hirano, and Yugo Fujimoto. Evaluating company-specific biases in financial sentiment analysis using large language models, 2024.

[71] Quang Hung Nguyen, Phuong Anh Trinh, Phan Quoc Hung Mai, and Tuan Phong Trinh. Finstat2sql: A text2sql pipeline for financial statement analysis, 2025.

[72] Sander Noels, Jorne De Blaere, and Tijl De Bie. A dutch financial large language model, 2024.

[73] Karmvir Singh Phogat, Chetan Harsha, Sridhar Dasaratha, Shashishekar Ramakrishna, and Sai Akhil Puranam. Zero-shot question answering over financial documents using large language models, 2023.

[74] Elvys Linhares Pontes, Carlos-Emiliano Gonzalez-Gallardo, Mohamed Benjannet, Caryn Qu, and Antoine Doucet. L3itc at the finllm challenge task: Quantization for financial text classification summarization, 2024.

[75] Neel Rajani, Lilli Kiessling, Aleksandr Ogaltsov, and Claus Lang. Kodexv0.1: A family of state-of-the-art financial large language models, 2024.

[76] Varun Rao, Youran Sun, Mahendra Kumar, Tejas Mutneja, Agastya Mukherjee, and Haizhao Yang. Llms meet finance: Fine-tuning foundation models for the open finllm leaderboard, 2025.

[77] Tohida Rehman, Raghubir Bose, Samiran Chattopadhyay, and Debarshi Kumar Sanyal. Transfer learning and transformer architecture for financial sentiment analysis, 2024.

[78] Bhaskarjit Sarmah, Tianjie Zhu, Dhagash Mehta, and Stefano Pasquali. Towards reducing hallucination in extracting information from financial reports using large language models, 2023.

[79] Spurthi Setty, Harsh Thakkar, Alyssa Lee, Eden Chung, and Natan Vidra. Improving retrieval for rag based question answering models on financial documents, 2024.

[80] Shalin Shah, Srikanth Ryali, and Ramasubbu Venkatesh. Multi-document financial question answering using llms, 2024.

[81] Yanxin Shen and Pulin Kirin Zhang. Financial sentiment analysis on news and reports using large language models and finbert, 2024.

[82] Pranam Shetty, Abhisek Upadhayaya, Parth Mitesh Shah, Srikanth Jagabathula, Shilpi Nayak, and Anna Joo Fee. Advanced financial reasoning at scale: A comprehensive evaluation of large language models on cfa level iii, 2025.

[83] Kuldeep Singh, Simerjot Kaur, and Charese Smiley. Finqapt: Empowering financial decisions with end-to-end llm-driven question answering pipeline, 2024.

[84] Zitao Song, Yining Wang, Pin Qian, Sifan Song, Frans Coenen, Zhengyong Jiang, and Jionglong Su. From deterministic to stochastic: An interpretable stochastic model-free reinforcement learning framework for portfolio optimization. *Applied Intelligence*, 2022.

[85] Huan-Yi Su, Ke Wu, Yu-Hao Huang, and Wu-Jun Li. Numllm: Numeric-sensitive large language model for chinese finance, 2024.

[86] Wenxi Sun, Qiannan Shen, Yijun Gao, Qinkai Mao, Tongsong Qi, and Shuo Xu. Objective over architecture: Fraud detection under extreme imbalance in bank account opening. *Computation*, 13(12):290, 2025.

[87] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[88] Graison Jos Thomas. Enhancing tinybert for financial sentiment analysis using gpt-augmented finbert distillation, 2024.

[89] Felix Tian, Ajay Byadgi, Daniel Kim, Daochen Zha, Matt White, Kairong Xiao, and Xiao-Yang Liu Yanglet. Customized fingpt search agents using foundation models, 2024.

[90] Yong-En Tian, Yu-Chien Tang, Kuang-Da Wang, An-Zi Yen, and Wen-Chih Peng. Template-based financial report generation in agentic and decomposed information retrieval, 2025.

[91] Hanshuang Tong, Jun Li, Ning Wu, Ming Gong, Dongmei Zhang, and Qi Zhang. Ploutos: Towards interpretable stock movement prediction with financial large language model, 2024.

[92] Saeid Ario Vaghefi, Aymane Hachcham, Veronica Grasso, Jiska Manicus, Nakiete Msemo, Chiara Colesanti Senni, and Markus Leippold. Ai for climate finance: Agentic retrieval and multi-step reasoning for early warning system investments, 2025.

[93] Dimitris Vamvourellis and Dhagash Mehta. Reasoning or overthinking: Evaluating large language models on financial sentiment analysis, 2025.

[94] Dannong Wang, Daniel Kim, Bo Jin, Xingjian Zhao, Tianfan Fu, Steve Yang, and Xiao-Yang Liu. Finlora: Finetuning quantized financial large language models using low-rank adaptation, 2024.

[95] Keyi Wang, Jaisal Patel, Charlie Shen, Daniel Kim, Andy Zhu, Alex Lin, Luca Borella, Cailean Osborne, Matt White, Steve Yang, Kairong Xiao, and Xiao-Yang Liu Yanglet. A report on financial regulations challenge at coling 2025, 2024.

[96] Neng Wang, Hongyang Yang, and Christina Dan Wang. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets, 2023.

[97] Peng Wang, Xuesi Hu, Jiageng Wu, Yuntao Zou, Qiancheng Zhang, and Dagang Li. What factors affect llms and rllms in financial question answering?, 2025.

[98] Shengkun Wang, Taoran Ji, Linhan Wang, Yanshen Sun, Shang-Ching Liu, Amit Kumar, and Chang-Tien Lu. Stocktime: A time series specialized large language model architecture for stock price prediction, 2024.

[99] Xinyu Wang, Jijun Chi, Zhenghan Tai, Tung Sum Thomas Kwok, Muzhi Li, Zhuhong Li, Hailin He, Yuchen Hua, Peng Lu, Suyuchen Wang, Yihong Wu, Jerry Huang, Jingrui Tian, Fengran Mo, Yufei Cui, and Ling Zhou. Finsage: A multi-aspect rag system for financial filings question answering, 2025.

[100] Yan Wang, Lingfei Qian, Xueqing Peng, Yang Ren, Keyi Wang, Yi Han, Dongji Feng, Fengran Mo, Shengyuan Lin, Qinchuan Zhang, Kaiwen He, Chenri Luo, Jianxing Chen, Junwei Wu, Chen Xu, Ziyang Xu, Jimin Huang, Guojun Xiong, Xiao-Yang Liu, Qianqian Xie, and Jian-Yun Nie. Fintagging: Benchmarking llms for extracting and structuring financial information, 2025.

[101] Yanbo J. Wang, Yuming Li, Hui Qin, Yuhang Guan, and Sheng Chen. A novel deberta-based model for financial question answering task, 2022.

[102] Yining Wang, Zhixiang Lu, Pin Qian, Jionglong Su, Mian Zhou, Chong Li, and Zhengyong Jiang. Gems-llm: Integrating large language models with goal-aware exploration for rl-based portfolio optimization. In *Advanced Intelligent Computing Technology and Applications*, Communications in Computer and Information Science. 2025.

[103] Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. Finvis-gpt: A multimodal large language model for financial chart analysis, 2023.

[104] Ziao Wang, Jianning Wang, Junda Wu, and Xiaofeng Zhang. An effective data creation pipeline to generate high-quality financial instruction data for large language model, 2023.

[105] Xinyu Wei and Luojia Liu. Are large language models good in-context learners for financial sentiment analysis?, 2025.

[106] Qilong Wu, Xiaoneng Xiang, Hejia Huang, Xuan Wang, Yeo Wei Jie, Ranjan Satapathy, Ricardo Shirota Filho, and Bharadwaj Veeravalli. Susgen-gpt: A data-centric llm for financial nlp and sustainability report generation, 2024.

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

[108] Xiaojun Wu, Junxi Liu, Huanyi Su, Zhouchi Lin, Yiyan Qi, Chengjin Xu, Jiajun Su, Jiajie Zhong, Fuwei Wang, Saizhuo Wang, Fengrui Hua, Jia Li, and Jian Guo. Golden touchstone: A comprehensive bilingual benchmark for evaluating financial large language models, 2024.

[109] Mengxi Xiao, Zihao Jiang, Lingfei Qian, Zhengyu Chen, Yueru He, Yijing Xu, Yuecheng Jiang, Dong Li, Ruey-Ling Weng, Min Peng, Jimin Huang, Sophia Ananiadou, and Qianqian Xie. Retrieval-augmented large language models for financial time series forecasting, 2025.

[110] Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. The wall street neophyte: A zero-shot analysis of chatgpt over multimodal stock movement prediction challenges, 2023.

[111] Frank Xing. Designing heterogeneous llm agents for financial sentiment analysis, 2024.

[112] Liang Xu, Lei Zhu, Yaotong Wu, and Hang Xue. Superclue-fin: Graded fine-grained analysis of chinese llms on diverse financial tasks and applications, 2024.

[113] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models, 2023.

[114] Jie Yang, Yiqiu Tang, Yongjie Li, Lihua Zhang, and Haoran Zhang. Dynamic hedging strategies in derivatives markets with llm-driven sentiment and news analytics, 2025.

[115] Xinqi Yang, Scott Zang, Yong Ren, Dingjie Peng, and Zheng Wen. Evaluating large language models on financial report summarization: An empirical study, 2024.

[116] Xiao-Yang Liu Yanglet, Yupeng Cao, and Li Deng. Multimodal financial foundation models (mffms): Progress, prospects, and challenges, 2025.

[117] Bithiah Yuan. Finbert-qa: Financial question answering with pre-trained bert language models, 2025.

[118] Ziqiang Yuan, Kaiyuan Wang, Shoutai Zhu, Ye Yuan, Jingya Zhou, Yanlin Zhu, and Wenqi Wei. Finllms: A framework for financial reasoning dataset generation with large language models, 2024.

[119] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models, 2023.

[120] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models, 2023.

[121] Kang Zhang, Osamu Yoshie, Lichao Sun, and Weiran Huang. Finllm-b: When large language models meet financial breakout trading, 2024.

[122] Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Hanqi Jiang, Yi Pan, Junhao Chen, Yifan Zhou, Zeyu Zhang, Ruitong Sun, Gengchen Mai, Ninghao Liu, and Tianming Liu. Revolutionizing finance with llms: An overview of applications and insights, 2024.

[123] Suifeng Zhao, Zhuoran Jin, Sujian Li, and Jun Gao. Finragbench-v: A benchmark for multimodal rag with visual citation in the financial domain, 2025.

[124] Yuhang Zhou, Zeping Li, Siyu Tian, Yuchen Ni, Sen Liu, Guangnan Ye, and Hongfeng Chai. Silversight: A multi-task chinese financial large language model based on adaptive semantic space learning, 2024.

[125] Ziyao Zhou and Ronitt Mehra. An end-to-end llm enhanced trading system, 2025.

[126] Andy Zhu and Yingjun Du. A role-aware multi-agent framework for financial education question answering with llms, 2025.

[127] Senbin Zhu, Chenyuan He, Hongde Liu, Pengcheng Dong, Hanjie Zhao, Yuchen Yan, Yuxiang Jia, Hongying Zan, and Min Peng. Silc-efsa: Self-aware in-context learning correction for entity-level financial sentiment analysis, 2024.