

# The New Quant: A Survey of Large Language Models in Financial Prediction and Trading

Weilong Fu\*

## Abstract

Large language models are reshaping quantitative investing by turning unstructured financial information into evidence-grounded signals and executable decisions. This survey synthesizes research with a focus on equity return prediction and trading, consolidating insights from domain surveys and more than fifty primary studies. We propose a task-centered taxonomy that spans sentiment and event extraction, numerical and economic reasoning, multimodal understanding, retrieval-augmented generation, time series prompting, and agentic systems that coordinate tools for research, backtesting, and execution. We review empirical evidence for predictability, highlight design patterns that improve faithfulness such as retrieval first prompting and tool-verified numerics, and explain how signals feed portfolio construction under exposure, turnover, and capacity controls. We assess benchmarks and datasets for prediction and trading and outline desiderata-for time safe and economically meaningful evaluation that reports costs, latency, and capacity. We analyze challenges that matter in production, including temporal leakage, hallucination, data coverage and structure, deployment economics, interpretability, governance, and safety. The survey closes with recommendations for standardizing evaluation, building auditable pipelines, and advancing multilingual and cross-market research so that language-driven systems deliver robust and risk-controlled performance in practice.

**Keywords** large language models; financial prediction; return prediction; trading; portfolio construction

## 1 Introduction

Large Language Models enable a shift from feature-centric text mining to end-to-end decision systems in markets. We refer to this emerging paradigm as the new quant, by which we mean investment processes where language models read and reason over heterogeneous disclosures, generate auditable hypotheses, interact with external tools and data, and translate textual understanding into risk-controlled positions. This survey concentrates on the pipeline components that matter most for investment outcomes, namely financial prediction with an emphasis on equity return prediction and trading with portfolio construction. We systematize advances from 2023 to 2025 through this lens (Zhao et al., 2024; Lee et al., 2024; Nie et al., 2024; Liu, 2024; Kong et al., 2024; Xu, 2024).

Transformer pretraining and instruction tuning produced general purpose models with non-trivial reasoning and tool use (Devlin et al., 2019; Brown, Mann, et al., 2020; Achiam et al., 2023). Domain-specific financial language models and open weight ecosystems make finance-grade adaptation feasible under privacy, governance, and cost constraints (Wu, Irsøy, et al., 2023; Scao, Fan, et al., 2023; Touvron et al., 2023b; Touvron et al., 2023a; Meta AI, 2024). Efficient tuning through low-rank adaptation, quantization-sensitive fine-tuning, and one-bit optimization lowers the barrier to controlled deployment in trading environments (Hu, Shen, et al., 2021; Dettmers et al., 2024; Ma et al., 2024b). In parallel, task-level capabilities have matured across sentiment, information extraction and knowledge graphs, numerical question answering, long document understanding, multi-modal analysis, and agentic decision support. These capabilities feed, constrain, and explain predictive signals and trade decisions (Kong et al., 2024; Nie et al., 2024; Zhao et al., 2024).

Evidence already suggests that model derived views on news, filings, earnings calls, and policy communications can predict returns in certain settings, although evaluation practice often falls short of trading standards. Leakage control, stress testing, market microstructure realism, and cost or capacity reporting

---

\*Columbia University, USA. wf2232@columbia.edu

remain inconsistent. Governance and interpretability requirements, such as evidence-based rationales, audit logs, and a clear separation between signal generation and portfolio allocation, are likewise unevenly addressed.

Our contributions are fourfold. First, we frame the design space for the development of new quantitative and review models relevant to finance together with the efficiency techniques that make financial language models practical (Devlin et al., 2019; Brown, Mann, et al., 2020; Achiam et al., 2023; Wu, Irsoy, et al., 2023; Hu, Shen, et al., 2021; Dettmers et al., 2024; Ma et al., 2024b). Second, we offer a task taxonomy centered on prediction and trading that clarifies how the upstream natural language processing components feed tradable signals (Nie et al., 2024; Zhao et al., 2024; Kong et al., 2024). Third, we synthesize the literature on return prediction in Section 4 and on trading with portfolio construction in Section 5. We cover interpretable financial language models, retrieval augmented pipelines, time series aware prompting, and multi-agent trading systems. Fourth, we consolidate benchmarks and datasets in Section 6 and articulate challenges in Section 7 that involve temporal leakage, faithfulness, evaluation realism, cost and latency, and governance. The audience includes researchers who build financial language models for tradable use cases, quantitative practitioners who evaluate language model signals, and leaders who design audit-ready deployment strategies (Xu, 2024; Kong et al., 2024).

## 2 Foundations for Prediction and Trading with FinLLMs

### 2.1 From transformers to tool using language models

The transformer replaced recurrent networks with attention and enabled scalable pretraining on large corpora (Vaswani, Shazeer, et al., 2017). Decoder only GPT models demonstrated emergent in context and few shot abilities (Radford, Narasimhan, et al., 2018; Radford, Wu, et al., 2019; Brown, Mann, et al., 2020). Encoder only models such as BERT and RoBERTa delivered state of the art text understanding for classification and span extraction (Devlin et al., 2019; Liu, Ott, et al., 2019). The GPT 4 technical report and instruction tuning advances, as exemplified by FLAN, established language models as general purpose controllers with meaningful reasoning and tool use (Achiam et al., 2023; Wei et al., 2021).

### 2.2 Open models and efficient adaptation for finance

Open releases including BLOOM and the LLaMA families catalyzed a flourishing ecosystem that supports controlled adaptation and on premise deployment (Scao, Fan, et al., 2023; Touvron et al., 2023b; Touvron et al., 2023a; Meta AI, 2024). Additional families such as Qwen, Baichuan, and InternLM expand the menu of base models (Bai et al., 2023; Baichuan Inc., 2023; InternLM, 2024). Efficient training and post training alignment through low rank adaptation, quantization aware finetuning, and one bit optimization enable domain tuned models with modest compute budgets, which aligns with privacy and reproducibility constraints that are common in trading environments (Hu, Shen, et al., 2021; Dettmers et al., 2024; Ma et al., 2024b).

### 2.3 Financial PLMs and FinLLMs

Before instruction following language models, financial applications built on encoder style pretrained language models such as the FinBERT line and showed early domain transfer benefits (Araci, 2019; Yang, Uy, and Huang, 2020; Liu et al., 2021). Recent finance specific language models include BloombergGPT (Wu, Irsoy, et al., 2023), PIXIU (Xie et al., 2023), FinGPT (Yang, Liu, Wang, et al., 2023; Liu, Zhang, et al., 2024a), InvestLM (Yang, Tang, and Tam, 2023), Instruct FinGPT (Zhang, Yang, et al., 2023b), DISC FinLLM (Chen, Wang, et al., 2023), and CFGPT (Li et al., 2023). Language and market specific adaptations such as SilverSight and FinVisGPT tailor models to regional corpora and workflows (Zhou et al., 2024; Wang, Li, et al., 2023). FinTral reports a multimodal family with performance at the reported level of general purpose models (Bhatia et al., 2024). New compact bases such as Mistral 7B have also been adopted in finance settings (Jiang et al., 2023a).

## 2.4 Implications for prediction and trading

For return prediction, decoder models support rationale generation and in context composition over news, filings, and macro text, while encoder models remain strong for narrow sentiment or extraction tasks that feed signals. Hybrid systems that combine retrieval augmented language models, language driven graph or sequence models, and mixture of experts pipelines appear frequently, often coupled with faithfulness checks and backtesting aware evaluation (Kong et al., 2024; Nie et al., 2024). For trading, agentic frameworks with tool use, memory, and role specialization begin to structure research, critique, and execution under constraints, which motivates new benchmarks and simulation protocols that we discuss in later sections.

## 3 Task taxonomy for financial prediction and trading

This section codifies a taxonomy that maps language model capabilities to finance workflows and clarifies how prediction and trading depend on upstream natural language processing. We group tasks by the primary function they serve in production pipelines, while noting that deployed systems often combine several capabilities in a single workflow.

### 3.1 Sentiment and opinion as signal inputs

The objective is to infer polarity, stance, and intensity from heterogeneous sources such as news, social media, earnings calls, and analyst notes, and to transform these assessments into features for event studies, return prediction, or risk monitoring. Domain tuned encoders in the FinBERT line demonstrate strong transfer on finance texts (Araci, 2019; Yang, Uy, and Huang, 2020; Liu et al., 2021). Instruction tuned language models can score sentiment and produce justifications and they sometimes outperform classical lexicon baselines on complex material (Lopez-Lira and Tang, 2023; Steinert and Altmann, 2023; Luo and Gong, 2024). Work on FOMC minutes and ECB press conferences indicates that policy tone can be quantified and linked to market responses (Gössi et al., 2023; Kanelis and Siklos, 2024). Classic resources remain useful as baselines and diagnostic tools (Loughran and McDonald, 2011; Pennebaker, Francis, and Booth, 2001; Stone, Dunphy, and Smith, 1966; Mishev et al., 2020; Tan, Lee, and Lim, 2023; Bordoloi and Biswas, 2023). In a trading context sentiment features must be timestamped, free of look ahead leakage, and aligned to realistic rebalancing schedules.

### 3.2 Information extraction and knowledge graphs for point in time signals

Information extraction converts unstructured documents into structured entities, relations, and events that can feed screens, factor engines, and retrieval modules. Datasets such as FiNER, FinRED, and REFinD support supervised training and enable point in time knowledge curation (Hillebrand et al., 2022; Shah, Vithani, et al., 2023; Sharma et al., 2022; Kaur et al., 2023). Large language models can assist IE through prompting or lightweight finetuning for named entity recognition and relation extraction (Covas, 2023; Rajpoot and Parikh, 2023). Event detection and relation modeling in Chinese and English demonstrate cross market applicability (Tian, Zhao, and Ren, 2019; Wan et al., 2023). As institutions deploy knowledge graphs for research, LLMs act as controllers and generators that populate and query the graphs while surveys outline integration patterns and governance requirements (Xue et al., 2023; Jiang et al., 2023b; Pan et al., 2023; Pan et al., 2024; Li, 2023; Liang et al., 2024a; Zwam et al., 2020). For prediction and trading these components provide upstream signal generation and retrieval that improves evidence quality.

### 3.3 Numerical question answering and reasoning for thesis validation

These systems execute multi step reasoning over tables, text, and formulas in filings, earnings calls, and macro releases and they answer questions while computing key performance indicators. Benchmarks such as FinQA, FinanceBench, BizBench, DocMathEval, and EconLogicQA probe numerical correctness, long document understanding, and economics logic that underlies valuation and surprise based signals (Chen, Chen, et al., 2021; Islam et al., 2023; Koncel-Kedziorski et al., 2023; Zhao et al., 2023; Quan and Liu, 2024). Retrieval augmented generation with layout aware encoders and tool calls improves verifiability and accuracy

and verification or constrained decoding reduces hallucination risk (Phogat et al., 2023; Srivastava, Malik, and Ganu, 2024; Arun et al., 2023; Wang et al., 2024). In production pipelines these systems are most valuable when they produce intermediate calculations and citations that can be audited before the trade decision.

### 3.4 Summarization and document understanding for evidence condensation

Abstractive and extractive hybrids and instruction tuned models can compress long financial narratives such as ten K filings, management discussion and analysis, and earnings calls, which accelerates research and supports hypothesis generation (El-Haj, 2019; Quatra and Cagliero, 2020; Mukherjee et al., 2022; Abdaljalil and Bouamor, 2021; Zmandar et al., 2021). Retrieval aware chunking and long document architectures reduce information loss and yield more stable summaries (Yepes, You, et al., 2024; Beltagy, Peters, and Cohan, 2020). For trading use the outputs should be materiality aware and time stamped and they should reference evidence spans that analysts can verify.

### 3.5 Multimodal cues for predictive signals

Beyond text, systems fuse audio from calls, visuals such as charts, or structured time series to inform predictive modeling. Datasets and models for multimodal analysis of earnings calls and policy communication supply prosodic and visual cues (Li and Zhang, 2020; Mathur et al., 2022). FinVisGPT addresses chart reading and explanation, and language model informed graph or sequence models use textual context to guide stock movement prediction (Wang, Li, et al., 2023; Chen, Zheng, et al., 2023; Wimmer and Rekabsaz, 2023). RiskLabs illustrates multi source fusion for risk prediction (Cao, Chen, et al., 2024). For trading deployment multimodal models must meet latency constraints and must ensure that any audio or visual evidence is available at decision time.

### 3.6 Agentic workflows for trading and execution

Agentic frameworks operationalize language models as decision support agents with memory, tools, and role specialization. TradingGPT introduces layered memory and distinct analyst characters and later systems expand tool use, multi agent debate, and evaluation in leakage controlled simulations (Li, Yu, et al., 2023; Zhang, Zhao, et al., 2024; Yu et al., 2024; Wang, Yuan, et al., 2024; Wang, Yuan, et al., 2023; Yuan, Wang, and Guo, 2024). Surveys of LLM based agents, computational experiments, memory mechanisms, and trust or safety provide design guidance and highlight open problems such as planning reliability and tool misuse (Xi et al., 2023; Guo et al., 2024; Ma et al., 2024a; Zhang, Bo, et al., 2024; Hua et al., 2024). In practice these agents should separate research from order routing and they should log prompts, retrievals, and tool calls for audit.

### 3.7 Governance functions that constrain trading systems

Financial institutions explore LLMs for auditing support, contradiction detection, and regulatory interpretation (Berger et al., 2023; Deußer, Leonhard, et al., 2023; Cao and Feinstein, 2024; Choi and Kim, 2024). These capabilities do not execute trades and they do shape acceptable model behavior, guardrails, and evidence requirements for production systems. Trading oriented deployments benefit from explicit policies that bind language to timestamped evidence and that restrict action when verification fails.

## 4 LLMs for return prediction

This section surveys how language models produce equity return signals and how those signals can be translated into investable decisions. We organize methods by evidence channels and modeling patterns and then discuss evaluation practice and practical guidance. We focus on studies from 2023 to 2025 with direct implications for trading.

Table 1: Mapping of tasks to trading relevance

Task	Representative artifacts	Typical outputs	Contribution to trading
Sentiment and opinion	FinBERT line, instruction tuned LLM scorers	Polarity, stance, justifications	Event features, risk monitoring, regime filters
Information extraction and knowledge graphs	FiNER, FinRED, REFinD, FinDKG, WeaverBird	Entities, relations, events, KG triples	Point in time factors, high precision retrieval, constraint checks
Numerical QA and reasoning	FinQA, FinanceBench, DocMathEval, EconLogicalQA	Computed KPIs, verified answers, reasoning chains	Thesis validation, surprise based signals, audit trails
Summarization and document understanding	ECTSum, MultiLing FNS, Longformer, RAG chunking	Condensed briefs with citations	Faster research, explanation for signals and trades
Multimodal analysis	MAEC, MONOPOLY, FinVisGPT	Prosody features, chart readings, fused embeddings	Additional cues for selection or sizing under latency limits
Agentic workflows	TradingGPT, FinAgent, FinMem, QuantAgent	Tool calls, debate traces, memory states	Orchestration of research, verification, and execution
Governance and compliance	Audit and regulation tools	Contradiction flags, policy checks	Guardrails that shape allowable actions and documentation

## 4.1 Problem formulation and evidence channels

The objective is to map text and related modalities to expected returns at horizons that range from intraday to monthly. Common channels include news and social media, corporate disclosures and earnings calls, and policy communications and macro releases. Work using general purpose language models reports that zero or few shot prompts on event text can be predictive in some settings (Lopez-Lira and Tang, 2023; Steinert and Altmann, 2023). Domain specific corpora such as earnings call transcripts provide richer cues in the narrative and question and answer segments and several studies evaluate models on this setting (Cook et al., 2023). Central bank statements and press conferences also encode information that matters for assets that are sensitive to interest rates and risk appetite (Gössi et al., 2023; Kanelis and Siklos, 2024).

## 4.2 Modeling patterns for text to return signals

**Zero and few shot scoring with general models** General purpose models can be prompted to classify event direction or to assign a return score together with a rationale. Early finance studies report out of sample predictability from such scores on news and social media (Lopez-Lira and Tang, 2023; Steinert and Altmann, 2023). Practical systems often add calibration layers, confidence filtering, and symbol mapping before portfolio construction.

**Domain and instruction tuned FinLLMs** Instruction tuning on finance instructions and curated corpora improves robustness and reduces prompt brittleness. Representative models include InvestLM, InstructFinGPT, FinGPT and its high performance computing variants, and FinLlama (Yang, Tang, and Tam, 2023; Zhang, Yang, et al., 2023b; Yang, Liu, Wang, et al., 2023; Liu, Zhang, et al., 2024a; Konstantinidis et al., 2024). Value aligned or preference tuned variants have also been explored (Yu, Huber, and Tang, 2024). These models typically strengthen sentiment and event classification and they can generate explanations that are easier to audit.

**Retrieval augmented modeling and knowledge grounded signals** Retrieval augmented generation reduces hallucination risk and improves faithfulness by binding predictions to timestamped evidence. Financial pipelines add retrieval aware chunking and layout features for long documents (Yepes, You, et al., 2024; Wang et al., 2024). Knowledge graphs and retrieval over company specific graphs further structure the evidence and stabilize factor construction (Xue et al., 2023; Li, 2023). RAG enhanced sentiment has been shown to improve downstream accuracy on finance tasks (Zhang, Yang, et al., 2023a).

**LLM guided structured models** Language models can supply features or supervisory signals to graph and sequence models that are optimized for price movement prediction. Examples include graph neural networks whose edges or node priors are informed by language model judgments and vision language systems that read price charts (Chen, Zheng, et al., 2023; Wimmer and Rekabsaz, 2023). Risk oriented work fuses model derived signals with other sources to predict adverse events (Cao, Chen, et al., 2024).

**Time series aware prompting and forecasting** Several studies examine how to connect language models to time series forecasting. Approaches include reprogramming prompts for temporal inputs, using language models as zero shot forecasters, and using specialized long horizon transformer architectures alongside language models for reasoning and explanation (Jin et al., 2023; Gruver et al., 2024; Zhou et al., 2022; Nie et al., 2022; Wen et al., 2022; Liang et al., 2024b). Although these papers are not always finance specific, the techniques inform how to condition signals on regime and horizon.

**Interpretable designs and auditability** Production systems require transparent rationales and clear provenance for each predicted effect. Interpretable pipelines generate structured explanations and highlight evidence spans and they expose ablation or counterfactual checks. Recent work proposes interpretable stock movement modeling with finance specific rationale templates and self reflective explanations (Tong et al., 2024; Koa et al., 2024). These designs help risk teams to understand when and why a signal should be trusted.

### 4.3 Evaluation protocols that meet trading standards

Return prediction requires time safe evaluation. We recommend rolling walk forward splits with document availability enforced at decision time and with an embargo to prevent label leakage from post event commentary. Report signal quality with correlation and calibration and report economics with returns that include explicit commission and spread assumptions, Sharpe and drawdown, turnover and capacity, and sensitivity to universe and rebalancing frequency. Given growing evidence of look ahead issues in pretrained models, evaluation should check for time machine effects using dated corpora and explicit filters (Sarkar and Vafa, 2024; Drinkall et al., 2024). Baselines should include naive and factor models and trend benchmarks to avoid overstating the incremental value of language signals (Jiang, Kelly, and Xiu, 2023).

### 4.4 What works when and practical guidance

Language signals often add value around identifiable events and narrative changes and they can complement price based factors during regime shifts. News and social media sentiment tends to matter at shorter horizons when coverage is fast and dense. Earnings call analysis matters at announcement and in the following days when management tone and detail resolve uncertainty. Policy communication sentiment is most relevant for rate sensitive sectors and for broad risk appetite proxies. In all cases the portfolio should separate signal generation from allocation and risk and it should include materiality filters, confidence gating, and exposure controls.

Table 2: Representative papers for equity return prediction with one sentence summaries

Paper	Setting or channel	One sentence summary
Lopez-Lira and Tang (2023)	News and filings	Zero and few shot scores from a general model predict cross sectional returns in several universes with controls for headline leakage.
Steinert and Altmann (2023)	Social media	Microblog sentiment from a large model correlates with next day stock moves and improves on lexicon baselines.
Cook et al. (2023)	Earnings calls	Locally hosted language models score call tone and deliver signals that survive controls for known factors.
Gössi et al. (2023)	Policy minutes	FinBERT tuned for policy text extracts sentiment from FOMC minutes that aligns with market responses.
Kanelis and Siklos (2024)	Press conferences	A sentiment indicator from ECB statements explains euro area asset movements and complements macro variables.
Yang, Tang, and Tam (2023)	Finance tuned LLM	Instruction tuned InvestLM improves investment specific judgments and produces auditable rationales.
Zhang, Yang, et al. (2023b)	Finance tuned LLM	Instruct FinGPT strengthens finance sentiment and can act as a robust scoring component in pipelines.
Yang, Liu, Wang, et al. (2023) and Liu, Zhang, et al. (2024a)	Open finance LLM	FinGPT provides open models and recipes that enable cost aware domain adaptation for finance tasks.
Konstantinidis et al. (2024)	Sentiment for trading	FinLlama demonstrates instruction tuned scoring for trading oriented sentiment classification.
Yu, Huber, and Tang (2024)	Preference tuned LLM	GreedLlama studies value alignment for financial reasoning and highlights the effect on moral or risk trade offs.
Yepes, You, et al. (2024)	Retrieval and chunking	Retrieval aware chunking improves long document question answering for filings and earnings analysis.
Wang et al. (2024)	Layout aware modeling	A layout aware generator improves numerical reasoning over tables and reduces errors in KPI extraction.
Xue et al. (2023)	Knowledge grounded RAG	A system that couples language models with a knowledge base and search engine improves decision support quality.
Li (2023)	Dynamic knowledge graphs	A dynamic finance knowledge graph supports point in time retrieval for research and signal construction.

*Continued on next page*

*Table 2 continued*

Paper	Setting or channel	One sentence summary
Chen, Zheng, et al. (2023)	Text guided GNN	A graph neural network informed by language model judgments improves stock movement prediction.
Wimmer and Rekabsaz (2023)	Vision language	A vision language approach uses chart images to detect granular market changes that relate to returns.
Cao, Chen, et al. (2024)	Multi source risk	A multi source pipeline with a language model integrates diverse data to predict financial risk events.
Jin et al. (2023)	Time series prompting	A reprogramming approach adapts language models to time series forecasting and yields competitive accuracy.
Gruver et al. (2024)	Zero shot forecasting	Large language models used as zero shot forecasters provide reasonable baselines for several temporal datasets.
Zhou et al. (2022)	Long horizon TS	A frequency enhanced transformer delivers strong long horizon forecasting and can complement language signals.
Nie et al. (2022)	Tokenization for TS	A tokenization approach converts time series into compact sequences that are well suited to transformers.
Tong et al. (2024)	Interpretable stock movement	An interpretable finance specific model produces rationales that link text spans to predicted movement.
Koa et al. (2024)	Self reflective explanations	A method that uses self reflection yields explainable stock predictions with improved plausibility of rationales.
Jiang, Kelly, and Xiu (2023)	Baseline for trends	A comprehensive study of trend models supplies strong baselines that are useful when measuring incremental value.

## 5 LLM assisted trading systems and portfolio construction

This section analyzes how language models support trading decisions from idea generation to execution and how they interact with portfolio construction. We organize the discussion around the life cycle of a trade and we emphasize designs that produce auditable, time safe, and economically meaningful outcomes.

### 5.1 From assisted research to executable strategies

Agentic systems transform language models into research assistants that read disclosures, propose hypotheses, and coordinate tools such as retrieval, calculators, and backtesters. TradingGPT introduces layered memory and distinct analyst roles that debate and refine theses before handing off to tools (Li, Yu, et al., 2023). FinAgent expands the toolkit to include multimodal inputs and broker like actions under a tool governance layer (Zhang, Zhao, et al., 2024). FinMem focuses on memory design that stabilizes multi day workflows and preserves analyst intent during iteration (Yu et al., 2024). QuantAgent explores self improvement loops that

critique prompts and strategies and that then retest within a controlled simulator (Wang, Yuan, et al., 2024). Alpha GPT and its successor Alpha GPT 2.0 formalize analyst in the loop alpha discovery with critique, ranking, and evaluation gates to reduce overfitting (Wang, Yuan, et al., 2023; Yuan, Wang, and Guo, 2024). Together these systems show how assisted research can evolve into executable strategies while keeping human oversight in the loop.

## 5.2 Prompting and language to strategy

Several studies convert natural language descriptions into screen definitions, factor recipes, or backtest scripts. Work on code generation for trading strategies indicates that language models can scaffold usable code with human review and unit tests (Alonso and Dupouy, 2024). Conversational research tools support exploratory analysis and rapid what if checks for fundamental and event driven theses (Yue and Au, 2023). Effective practice includes canonicalizing prompts into machine readable templates, validating data access permissions, and compiling prompts and code into immutable artifacts that can be audited later.

## 5.3 Retrieval verified analysis loops

Hallucination and numerical brittleness motivate retrieval verified workflows. Retrieval aware chunking and layout aware modeling improve KPI extraction from filings and reduce reasoning errors in long documents (Yepes, You, et al., 2024; Wang et al., 2024). Systems that couple a language model with a curated knowledge base and a search engine demonstrate higher faithfulness for decision support (Xue et al., 2023). In trading contexts the loop proceeds as propose, retrieve, verify, and only then simulate or trade. Each step produces traces with timestamps and evidence spans to support review by risk and compliance.

## 5.4 From signals to orders and execution

Language models that score text still require a conversion to orders and an execution policy that respects market microstructure. A practical pattern separates signal generation from order placement and routing. Execution quality depends on latency, slippage, queue priority, and the balance between limit and market orders. Recent work on generative modeling for limit order book message flow offers realistic simulators for policy testing (Nagy et al., 2023). Decision systems should log order intents, parameter choices, and realized costs to enable attribution and continuous improvement.

## 5.5 Portfolio construction with language model support

Portfolio construction benefits from language models in two ways. First, LLM derived signals enter a classical optimizer or a rules based allocator with exposure and turnover controls. Second, language models can assist with constraint elicitation and documentation by translating investment beliefs and policy rules into machine readable constraints. Studies that evaluate the impact of conversational assistance on portfolio choices suggest that language models can improve portfolio hygiene when paired with clear prompts and risk constraints (Ko and Lee, 2024). In production settings the optimizer and the signal engine should remain distinct services with independent monitoring and fallback policies.

## 5.6 Evaluation protocols and guardrails for live trading

Trading evaluation must be time safe and economically grounded. Walk forward backtests should enforce document availability and embargo periods and they should report returns with explicit cost and impact assumptions, Sharpe and drawdown, turnover and capacity, and sensitivity to universe and rebalancing cadence. Work on lookahead bias in pretrained models and on time machine effects underscores the need for dated corpora and strict filters during both training and evaluation (Sarkar and Vafa, 2024; Drinkall et al., 2024). Cost and latency management are essential for live use and hybrid query routing can reduce spend while maintaining quality by steering easy queries to lightweight models and reserving high capacity models for hard cases (Ding et al., 2024). Safety and governance require agent constitutions and risk aware judges that flag unsafe tool uses or policy violations (Hua et al., 2024; Yuan et al., 2024). Systems should also detect contradictions in reports and maintain audit logs to support regulatory reviews (Deußer, Leonhard, et al.,

2023; Cao and Feinstein, 2024). Strong baselines such as trend models help contextualize the incremental value of language driven workflows (Jiang, Kelly, and Xiu, 2023).

## 5.7 Design patterns and practical guidance

A robust design separates research and execution and binds language to verifiable evidence. Retrieval first prompting, tool verified numerics, and debate or critique before simulation reduce false positives. Confidence gating, materiality thresholds, and exposure caps stabilize portfolios. Human review remains important for new strategies, high impact actions, and regime changes. Regular stress tests and post trade analysis complete the loop and help teams decide when to promote a research signal into a production strategy.

Table 3: Representative papers for LLM assisted trading and portfolio construction with one sentence summaries

Paper	Contribution or setting	One sentence summary
Li, Yu, et al. (2023)	Multi agent research to trade	A layered memory and role based framework proposes, critiques, and verifies trade ideas before execution in a controlled simulator.
Zhang, Zhao, et al. (2024)	Tool augmented multimodal agent	A generalist agent integrates text and visuals and coordinates broker like tools under governance to produce executable decisions.
Yu et al. (2024)	Memory design for trading agents	A layered memory with character design improves persistence of analyst intent and boosts performance across multi day workflows.
Wang, Yuan, et al. (2024)	Self improving agent loop	A system that critiques prompts and strategies and that retests within a simulator yields more stable trading policies.
Wang, Yuan, et al. (2023)	Human AI alpha mining	An interactive workflow uses critique and ranking to surface promising alphas with guardrails against overfitting.
Yuan, Wang, and Guo (2024)	Human in the loop alpha mining	The second version formalizes review gates and improves reliability when promoting ideas to production.
Liu, Zhang, et al. (2024b)	Trading in realistic environments	A benchmarked environment evaluates LLM based traders with market frictions and supports ablation studies.
Alonso and Dupouy (2024)	Code generation for strategies	An empirical study shows that language models can scaffold trading code that passes unit tests when supervised by practitioners.
Yepes, You, et al. (2024)	Retrieval aware analysis	A chunking method improves retrieval and long document analysis for filings and earnings research that feeds trading.
Wang et al. (2024)	Layout aware modeling	A layout aware generator improves numerical reasoning over tables which reduces errors in research that precedes trades.

*Continued on next page*

*Table 3 continued*

Paper	Contribution or setting	One sentence summary
Nagy et al. (2023)	Limit order book simulation	A token level generative model of message flow produces realistic microstructure that is useful for execution policy testing.
Ding et al. (2024)	Cost and latency control	A hybrid routing approach reduces inference cost while maintaining answer quality which benefits live trading systems.
Hua et al. (2024)	Safety for agent systems	A constitution guided method constrains tool use and reduces unsafe actions during autonomous or semi autonomous operation.
Yuan et al. (2024)	Risk aware judging for agents	A benchmark and judge detect unsafe patterns in agent traces which complements trading evaluation.
Deußer, Leonhard, et al. (2023)	Governance and auditing	A contradiction detection pipeline highlights inconsistencies in financial reports and contributes to audit readiness.
Cao and Feinstein (2024)	Regulatory interpretation	A study outlines how language models can support interpretation of financial regulation which aids deployment governance.
Jiang, Kelly, and Xiu (2023)	Baseline for execution value add	A comprehensive trend study provides strong baselines that help measure the incremental value of language driven trading.

## 6 Benchmarks and datasets for prediction and trading

The growth of financial language models has outpaced the availability of standardized and time safe benchmarks that connect textual understanding to tradable decisions. We organize the landscape into prediction oriented reasoning benchmarks that produce signals, trading and agent benchmarks that evaluate decision quality under constraints, and corpora and datasets that supply supervision or retrieval evidence. Across categories three design principles are foundational. First, temporal integrity ensures point in time documents and rolling and non overlapping out of sample evaluation with embargoed validation. Second, economically grounded metrics require profit and loss with costs, Sharpe, drawdown, turnover and capacity, and hit rate at realistic rebalancing frequencies. Third, reproducibility demands seeded data releases, fixed symbol universes with survivorship bias controls, and code to reconstruct splits.

### 6.1 Prediction oriented reasoning and understanding

A first class of resources evaluates whether models can extract and reason over financial information that plausibly feeds return prediction. FinQA targets numerical reasoning over text and tables and signals derived from correct KPI computation are often used upstream of event driven strategies (Chen, Chen, et al., 2021). FinanceBench and BizBench probe quantitative reasoning and business logic and they stress mathematical consistency that underlies valuation or surprise based signals (Islam et al., 2023; Koncel-Kedziorski et al., 2023). DocMathEval isolates long document numerical reasoning with tables which is a frequent failure point in earnings analysis (Zhao et al., 2023). EconLogicQA evaluates economics sequential reasoning that matters for macro sensitive trade selection (Quan and Liu, 2024). The FinBen proposes a holistic financial benchmark that covers multiple tasks in finance (Xie et al., 2024). AlphaFin frames analysis as a retrieval augmented

stock chain that aligns evaluation with multi step reasoning workflows (Li, Li, et al., 2024). These resources are not trading simulators and they measure signal fidelity since a failure on numerical reasoning or economic logic makes the trade premise unsound.

## 6.2 Trading and agent evaluations

Benchmarks that are tailored to trading decisions remain emergent. Agent frameworks report simulation results using internal market environments together with layered memory and tool use such as retrieval, backtesting, and data application programming interfaces (Li, Yu, et al., 2023; Zhang, Zhao, et al., 2024; Yu et al., 2024; Wang, Yuan, et al., 2024; Wang, Yuan, et al., 2023; Yuan, Wang, and Guo, 2024). These works advance methodology through role specialization, verifier checks, and reflection and two gaps persist. First, there is limited standardization of market microstructure such as latency, slippage, queue priority, and the limit or market order mix. Second, there is heterogeneous choice of universes and horizons that complicates cross paper comparisons. Safety risk awareness for agents is emerging through R Judge which can complement trading evaluations by detecting unsafe tool usage or risk insensitive actions (Yuan et al., 2024).

## 6.3 Domain corpora and supervision for predictive pipelines

Upstream datasets support sentiment, information extraction, event detection, and summarization that feed predictive engines. FiNER, FinRED, and REFinD supervise extraction of entities, relations, and events that populate knowledge graphs and enable cleaner point in time factors (Shah, Vithani, et al., 2023; Sharma et al., 2022; Kaur et al., 2023). ECTSUm and MultiLing FNS provide summarization targets for earnings calls and reports, while MAEC and MONOPOLY supply multimodal earnings and policy material (Mukherjee et al., 2022; El-Haj, 2019; Li and Zhang, 2020; Mathur et al., 2022). FinSBD focuses on structural boundary detection in unstructured filings and DocLLM demonstrates layout aware modeling that improves numerical question answering and KPI retrieval (Au, Ait-Azzi, and Kang, 2021; Wang et al., 2024). These resources help construct evidence grounded signals that can survive audit.

## 6.4 Multilingual and regional benchmarks

Financial markets are multilingual and regulatory regimes differ across regions. Several efforts broaden coverage to non English disclosures. CFBenchmark, FinEval, and CFLUE provide Chinese financial evaluation resources and SuperCLUE Fin offers a fine grained analysis of Chinese tasks (Lei et al., 2023; Zhang et al., 2023; Zhu et al., 2024; Xu et al., 2024). Hirano constructs a Japanese financial benchmark that expands regional testing (Hirano, 2024). A study on bilingual prowess examines English and Spanish which is valuable for cross listings and American depositary receipts (Zhang et al., 2024).

## 6.5 Evaluation desiderata and a practical proposal

A prediction and trading benchmark should enforce time safe document availability, include standardized universes, rebalancing schedules, and cost models, and report both signal metrics and portfolio metrics with ablations for retrieval, verifiers, and tool latency. It should publish agent traces with evidence links for auditability and include stress periods and regime slices together with multilingual tracks. A practical path is to couple AlphaFin or The FinBen style reasoning tasks with an open microstructure simulator and R Judge style safety checks.

Table 4: Datasets and benchmarks that are most relevant to prediction and trading

Resource	Modality	Primary task	Relevance to trading
FinQA (Chen, Chen, et al., 2021)	Text and tables	Numerical question answering	KPI correctness supports earnings surprise and event driven signals
FinanceBench (Islam et al., 2023)	Text and numbers	Financial question answering	Valuation and logic checks help thesis validation
BizBench (Koncel-Kedziorski et al., 2023)	Text and numbers	Quantitative reasoning	Business logic consistency matters for fundamental theses
DocMathEval (Zhao et al., 2023)	Long documents and tables	Numerical reasoning	Reduces miscalculation risk in filings driven research
EconLogicQA (Quan and Liu, 2024)	Text	Economics sequential reasoning	Supports macro sensitive selection and hedging decisions
The FinBen (Xie et al., 2024)	Multi task	Holistic finance evaluation	Broad coverage aligns with diverse production workflows
AlphaFin (Li, Li, et al., 2024)	RAG with stock chain	Financial analysis	Multi step reasoning for equity research with RAG
FiNER and FinRED and REFinD (Shah, Vithani, et al., 2023; Sharma et al., 2022; Kaur et al., 2023)	Text	IE and NER and relation extraction	Populates knowledge graphs and supports time safe factors and retrieval
ECTSum and MultiLing FNS (Mukherjee et al., 2022; El-Haj, 2019)	Text	Summarization	Generates research briefs that accelerate analysis before trading
MAEC and MONOPOLY (Li and Zhang, 2020; Mathur et al., 2022)	Audio and video and text	Multimodal earnings and policy	Supplies prosodic and policy cues for selection and sizing
FinSBD and DocLLM (Au, Ait-Azzi, and Kang, 2021; Wang et al., 2024)	Text and layout	Structure detection and layout aware modeling	Stabilizes retrieval and improves numerical accuracy in long documents
CFBenchmark and FinEval and CFLUE and SuperCLUE Fin and JP benchmark (Lei et al., 2023; Zhang et al., 2023; Zhu et al., 2024; Xu et al., 2024; Hirano, 2024)	Text	Regional evaluation	Enables non English disclosures and cross market strategies
R Judge (Yuan et al., 2024)	Agent traces	Safety risk awareness	Adds guardrails for tool using agents during evaluation
TradingGPT and FinAgent and FinMem and QuantAgent and Alpha GPT (Li, Yu, et al., 2023; Zhang, Zhao, et al., 2024; Yu et al., 2024; Wang, Yuan, et al., 2024; Wang, Yuan, et al., 2023; Yuan, Wang, and Guo, 2024)	Agent frameworks	Trading simulations	Provide methodology and protocols without standard data releases

## 7 Challenges and open problems in LLM based prediction and trading

### 7.1 Temporal leakage and time machine effects

Return prediction with general web pretraining risks look ahead leakage because models may memorize future facts and surface them during prompting. Recent critiques show that even without explicit future documents at inference time the latent knowledge can leak into answers (Sarkar and Vafa, 2024; Drinkall et al., 2024). Effective mitigation combines corpora with strict publication cutoffs per evaluation fold, training data that is filtered by crawl date and source type, embargo windows for validation, and rationales that cite evidence published before the decision timestamp.

### 7.2 Evaluation realism and economic significance

Many studies report accuracy or correlation without a trading grade evaluation. Credible claims for language model signals require rolling walk forward backtests, conservative cost and impact models, turnover and capacity analysis, stress tests across regimes, and risk controlled performance with Sharpe, Sortino, drawdown, and tail loss. Benchmarks should include materiality filters so that statistically significant and economically trivial effects are not over interpreted.

### 7.3 Faithfulness, hallucination, and numerical robustness

Language models can produce confident but wrong rationales and can show brittle numerical reasoning. Evidence bound generation with citations to retrieved passages and tables, constrained tool use such as calculators and parsers, post hoc verification methods, and dual model cross checking reduce risk (Krishna et al., 2024). For trading the system should never change risk based on unverifiable rationales.

### 7.4 Data coverage, point in time structure, and retrieval

Filings, press releases, calls, and macro statements have heterogeneous formats and chunking and indexing must be point in time and stable across refactors. Layout aware encoders improve KPI extraction, structure boundary detection stabilizes retrieval, and financial information extraction datasets support higher precision evidence graphs (Wang et al., 2024; Au, Ait-Azzi, and Kang, 2021; Shah, Vithani, et al., 2023; Sharma et al., 2022; Kaur et al., 2023). Coverage gaps persist for small capitalization firms and non English issuers and multilingual resources help reduce these gaps.

### 7.5 Cost, latency, and deployment economics

Real time trading requires bounded latency and cost. Hybrid query routing can steer easy queries to cheaper models and reserve high capacity models for hard cases and low rank and quantized adaptation can further lower the footprint (Ding et al., 2024; Hu, Shen, et al., 2021; Dettmers et al., 2024; Ma et al., 2024b). System level reporting should include wall clock latency per decision and amortized compute cost per basis point of excess return.

### 7.6 Interpretability, governance, and regulatory alignment

Trading decisions must be explainable to risk, audit, and regulators. Desirable properties include rationales that are grounded in timestamped evidence, decomposition of effect that links evidence to predicted return, clear separation between signal generation and portfolio allocation, and audit logs for prompts, retrieved passages, and tool calls. Studies on regulatory interpretation and auditing support illustrate patterns for compliance ready pipelines (Cao and Feinstein, 2024; Berger et al., 2023; Deußer, Leonhard, et al., 2023).

## 7.7 Security, privacy, and safety

Financial language models raise attack surfaces that include prompt injection and alignment breaking attacks and they create privacy concerns. Agent frameworks need constitutions and safety checks to prevent unauthorized orders, personal data leakage, or policy violations (Hua et al., 2024; Yao et al., 2024). Ethical codes and evolving artificial intelligence regulations should inform deployment gates and operational controls.

## 7.8 Robust generalization and regime shifts

Language model signals can overfit a disclosure style, sector, or macro regime. Techniques that help include domain adaptation with retrieval from diverse sources, regime aware training through explicit slicing or adversarial invariance, multilingual modeling for cross listed firms, and ensembling with classical factors to stabilize exposures. Reporting should include sector breakdowns and regime wise performance.

## 7.9 Data synthesis and augmentation

Language model based augmentation can improve label efficiency and synthetic data can introduce biases or leakage if generated with non time safe context. Synthetic examples should be marked, confined to training, and stress tested for bias. Evaluation sets should never contain synthetic items.

## 7.10 Minimum reporting standard

As a minimum reporting standard, studies should enforce time safe data and splits with document availability at the decision timestamp; present a full cost model with commissions, spreads, and market impact calibrated to the universe and size; report turnover, capacity, and the effect of transaction costs on net performance; analyze stress periods and regimes with sector level breakdowns; provide evidence grounded rationales and verified calculations for key examples; include ablations for retrieval, verifiers, and query routing together with wall clock latency and compute cost per decision; release seeds and code to reconstruct time splits and point in time indices or a protocol that supports replication; and compare against strong trend and factor baselines to quantify incremental value.

# 8 Conclusion

Large Language Models are redefining quantitative investing by turning unstructured financial information into auditable signals and coordinated actions. In the new quant, language models do not replace classical statistics or portfolio theory and they compose with them. Models read filings and calls, cite timestamped evidence, invoke calculators and parsers for numerics, and hand verified signals to risk aware allocation engines. Evidence accumulated in recent work suggests real potential for excess returns in selected regimes and universes, especially when models are domain adapted, retrieval grounded, and evaluated with trading grade procedures.

The field should adopt three practical principles. Separate concerns means keeping signal generation with retrieval and verification distinct from portfolio construction so that objectives and accountability remain clear. Bind language to evidence means requiring timestamped citations and tool verified calculations before any position changes and logging prompts, retrievals, and tool calls for auditability. Evaluate like a practitioner means enforcing time safe splits with document availability checks and realistic costs and slippage, reporting turnover and capacity, analyzing stress regimes, and disclosing latency and compute cost per decision rather than only reporting accuracy.

A focused research program follows from these principles. The community should design standardized prediction to trading benchmarks that couple reasoning tasks such as FinQA, FinanceBench, and AlphaFin with open and time safe market simulators and with safety audits that detect risky tool use (Chen, Chen, et al., 2021; Islam et al., 2023; Li, Li, et al., 2024; Yuan et al., 2024). Training and evaluation should emphasize temporal robustness through filtered corpora, explicit publication cutoffs, and diagnostics for look ahead effects (Sarkar and Vafa, 2024; Drinkall et al., 2024). Explainable financial language models should produce evidence anchored rationales that map to portfolio exposures to meet governance needs. Multilingual

and low resource finance should receive sustained attention to support global coverage and cross listing dynamics. Systems should be cost aware through hybrid query routing and efficient adaptation methods such as low rank tuning, quantization aware finetuning, and one bit optimization (Ding et al., 2024; Hu, Shen, et al., 2021; Dettmers et al., 2024; Ma et al., 2024b). Human and AI collaboration should be central and analyst in the loop critique and debate agents can increase faithfulness without sacrificing speed while agent constitutions and judges improve safety (Hua et al., 2024; Yuan et al., 2024).

With transparent benchmarks, temporal discipline, and audit ready system design, financial language models can progress from promising prototypes to reliable building blocks in modern investment processes. The promise of the new quant is to translate textual understanding into robust, risk controlled, and economically meaningful trades.

## Disclosure

Portions of this paper were drafted or paraphrased with the assistance of ChatGPT (OpenAI). The author reviewed, edited, and takes full responsibility for the intellectual content and conclusions presented in this work.

## References

- Abdaljalil, S. and H. Bouamor (2021). “An Exploration of Automatic Text Summarization of Financial Reports”. In: *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*.
- Achiam, J. et al. (2023). “GPT-4 Technical Report”. In: *arXiv preprint arXiv:2303.08774*. URL: <https://arxiv.org/abs/2303.08774>.
- Alonso, M. N. i and H. Dupouy (2024). *Evaluating LLMs in Financial Tasks: Code Generation in Trading Strategies*. SSRN Working Paper.
- Araci, D. C. (2019). “FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models”. In: *arXiv preprint arXiv:1908.10063*.
- Arun, A., A. Dhiman, M. Soni, and Y. Hu (2023). “Numerical Reasoning for Financial Reports”. In: *arXiv preprint arXiv:2312.14870*. URL: <https://arxiv.org/abs/2312.14870>.
- Au, W., A. Ait-Azzi, and J. Kang (2021). “FinSBD-2021: The 3rd Shared Task on Structure Boundary Detection in Unstructured Text in the Financial Domain”. In: *Companion Proceedings of the Web Conference 2021*.
- Bai, J. et al. (2023). “Qwen Technical Report”. In: *arXiv preprint arXiv:2309.16609*.
- Baichuan Inc. (2023). *Baichuan-13B*. <https://github.com/baichuan-inc/Baichuan-13B>.
- Beltagy, I., M. E. Peters, and A. Cohan (2020). “Longformer: The Long-Document Transformer”. In: *arXiv preprint arXiv:2004.05150*.
- Berger, A. et al. (2023). “Towards Automated Regulatory Compliance Verification in Financial Auditing with Large Language Models”. In: *2023 IEEE International Conference on Big Data (BigData)*, pp. 4626–4635. DOI: [10.1109/BigData59044.2023.10386420](https://doi.org/10.1109/BigData59044.2023.10386420).
- Bhatia, G., E. M. B. Nagoudi, H. Cavusoglu, and M. Abdul-Mageed (2024). “FinTral: A Family of GPT-4-Level Multimodal Financial Large Language Models”. In: *arXiv preprint arXiv:2402.10986*.
- Bordoloi, M. P. and S. K. Biswas (2023). “Sentiment Analysis: A Survey on Design Framework, Applications and Future Scopes”. In: *Artificial Intelligence Review*.
- Brown, T. B., B. Mann, et al. (2020). “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*.
- Cao, Y., Z. Chen, et al. (2024). “RiskLabs: Predicting Financial Risk Using Large Language Model Based on Multi-Sources Data”. In: *arXiv preprint arXiv:2404.07452*.
- Cao, Z. and Z. Feinstein (2024). “Large Language Model in Financial Regulatory Interpretation”. In: *arXiv preprint arXiv:2405.06808*.
- Chen, W., Q. Wang, et al. (2023). “DISC-FinLLM: A Chinese Financial Large Language Model Based on Multiple Experts Fine-Tuning”. In: *arXiv preprint arXiv:2310.15205*.
- Chen, Z., L. N. Zheng, et al. (2023). “ChatGPT Informed Graph Neural Network for Stock Movement Prediction”. In: *arXiv preprint arXiv:2306.03763*.

- Chen, Z., W. Chen, et al. (2021). "FinQA: A Dataset of Numerical Reasoning over Financial Data". In: *arXiv preprint arXiv:2109.00122*.
- Choi, G. and A. G. Kim (2024). *Firm-Level Tax Audits: A Generative AI-Based Measurement*. Chicago Booth Research Paper No. 23-23, SSRN. URL: <https://ssrn.com/abstract=4645865>.
- Cook, T. R., S. Kazinnik, A. L. Hansen, and P. McAdam (2023). "Evaluating Local Language Models: An Application to Financial Earnings Calls". In: *SSRN 4627143*.
- Covas, E. (2023). "Named Entity Recognition Using GPT for Identifying Comparable Companies". In: *arXiv preprint arXiv:2307.07420*.
- Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer (2024). "QLoRA: Efficient Finetuning of Quantized LLMs". In: *Advances in Neural Information Processing Systems*.
- Deußer, T., D. Leonhard, et al. (2023). "Uncovering Inconsistencies and Contradictions in Financial Reports Using Large Language Models". In: *2023 IEEE International Conference on Big Data*.
- Devlin, J., M. Chang, K. Lee, and K. Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805*. URL: <https://arxiv.org/abs/1810.04805>.
- Ding, D. et al. (2024). "Hybrid LLM: Cost-Efficient and Quality-Aware Query Routing". In: *arXiv preprint arXiv:2404.14618*.
- Drinkall, F., E. Rahimikia, J. B. Pierrehumbert, S. Roberts, and S. Zohren (2024). "Time Machine GPT". In: *arXiv preprint arXiv:2404.18543*.
- El-Haj, M. (2019). "MultiLing 2019: Financial Narrative Summarisation". In: *Proceedings of the MultiLing 2019 Workshop*.
- Gössi, S., Z. Chen, W. Kim, B. Bermeitinger, and S. Handschuh (2023). "FinBERT-FOMC: Fine-Tuned FinBERT with Sentiment Focus Method for Enhancing Sentiment Analysis of FOMC Minutes". In: *Proceedings of the ACM International Conference on AI in Finance*.
- Gruver, N. et al. (2024). "Large Language Models Are Zero-Shot Time Series Forecasters". In: *Advances in Neural Information Processing Systems*.
- Guo, T., X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang (2024). "Large Language Model Based Multi-Agents: A Survey of Progress and Challenges". In: *arXiv preprint arXiv:2402.01680*. URL: <https://arxiv.org/abs/2402.01680>.
- Hillebrand, L. et al. (2022). "KPI-BERT: A Joint NER and Relation Extraction Model for Financial Reports". In: *Proceedings of the 26th International Conference on Pattern Recognition*.
- Hirano, M. (2024). "Construction of a Japanese Financial Benchmark for Large Language Models". In: *arXiv preprint arXiv:2403.15062*.
- Hu, E. J., Y. Shen, et al. (2021). "LoRA: Low-Rank Adaptation of Large Language Models". In: *arXiv preprint arXiv:2106.09685*. URL: <https://arxiv.org/abs/2106.09685>.
- Hua, W. et al. (2024). "TrustAgent: Toward Safe and Trustworthy LLM-Based Agents Through Agent Constitution". In: *arXiv preprint arXiv:2402.01586*.
- InternLM (2024). *InternLM*. <https://github.com/InternLM>.
- Islam, P. et al. (2023). "FinanceBench: A New Benchmark for Financial Question Answering". In: *arXiv preprint arXiv:2311.11944*.
- Jiang, A. et al. (2023a). "Mistral 7B". In: *arXiv preprint arXiv:2310.06825*.
- Jiang, J., B. Kelly, and D. Xiu (2023). "(Re-)Imag(in)ing Price Trends". In: *The Journal of Finance* 78.6, pp. 3193–3249.
- Jiang, X. et al. (2023b). "On the Evolution of Knowledge Graphs: A Survey and Perspective". In: *arXiv preprint arXiv:2310.04835*.
- Jin, M. et al. (2023). "Time-LLM: Time Series Forecasting by Reprogramming Large Language Models". In: *arXiv preprint arXiv:2310.01728*.
- Kanelis, D. and P. L. Siklos (2024). "The ECB Press Conference Statement and a New Sentiment Indicator for the Euro Area". In: *International Journal of Finance & Economics*.
- Kaur, S. et al. (2023). "REFinD: Relation Extraction Financial Dataset". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3054–3063.
- Ko, H. and J. Lee (2024). "Can ChatGPT Improve Investment Decisions from a Portfolio Management Perspective". In: *Finance Research Letters*.

- Koa, K. J. et al. (2024). "Learning to Generate Explainable Stock Predictions Using Self-Reflective Large Language Models". In: *arXiv preprint arXiv:2402.03659*.
- Koncel-Kedziorski, R. et al. (2023). "BizBench: A Quantitative Reasoning Benchmark for Business and Finance". In: *arXiv preprint arXiv:2311.06602*.
- Kong, Y. et al. (2024). "Large Language Models for Financial and Investment Management: Applications and Benchmarks". In: *arXiv preprint arXiv:2402.09171*. URL: <https://arxiv.org/abs/2402.09171>.
- Konstantinidis, T., G. Iacovides, M. Xu, T. G. Constantinides, and D. Mandic (2024). "FinLlama: Financial Sentiment Classification for Algorithmic Trading Applications". In: *arXiv preprint arXiv:2403.12285*. URL: <https://arxiv.org/abs/2403.12285>.
- Krishna, K. et al. (2024). "GenAudit: Fixing Factual Errors in Language Model Outputs with Evidence". In: *arXiv preprint arXiv:2402.12566*.
- Lee, J., N. Stevens, S. C. Han, and M. Song (2024). "A Survey of Large Language Models in Finance (FinLLMs)". In: *arXiv preprint arXiv:2402.02315*. URL: <https://arxiv.org/abs/2402.02315>.
- Lei, Y. et al. (2023). "CFBenchmark: Chinese Financial Assistant Benchmark for Large Language Model". In: *arXiv preprint arXiv:2311.05812*.
- Li, J. et al. (2023). "CFGPT: Chinese Financial Assistant with Large Language Model". In: *arXiv preprint arXiv:2309.10654*.
- Li, Q. and Q. Zhang (2020). "A Unified Model for Financial Event Classification, Detection and Summarization". In: *Proceedings of IJCAI*.
- Li, X., Z. Li, C. Shi, Y. Xu, Q. Du, M. Tan, J. Huang, and W. Lin (2024). "AlphaFin: Benchmarking Financial Analysis with Retrieval-Augmented Stock-Chain Framework". In: *arXiv preprint arXiv:2403.12582*. URL: <https://arxiv.org/abs/2403.12582>.
- Li, X. V. (2023). "FinDKG: Dynamic Knowledge Graph with Large Language Models for Global Finance". In: *SSRN*. URL: <https://ssrn.com/abstract=4608445>.
- Li, Y., Y. Yu, et al. (2023). "TradingGPT: Multi-Agent System with Layered Memory and Distinct Characters for Enhanced Financial Trading Performance". In: *arXiv preprint arXiv:2309.03736*.
- Liang, Y. et al. (2024a). "Aligning Large Language Models to a Domain-Specific Graph Database". In: *arXiv preprint arXiv:2402.16567*.
- Liang, Y. et al. (2024b). "Foundation Models for Time Series Analysis: A Tutorial and Survey". In: *arXiv preprint arXiv:2403.14735*.
- Liu, J. (2024). "A Survey of Financial AI: Architectures, Advances and Open Challenges". In: *arXiv preprint arXiv:2403.06761*. URL: <https://arxiv.org/abs/2403.06761>.
- Liu, X., J. Zhang, et al. (2024a). "FinGPT-HPC: Efficient Pretraining and Finetuning Large Language Models for Financial Applications with High-Performance Computing". In: *arXiv preprint arXiv:2402.13533*.
- Liu, X., C. Zhang, et al. (2024b). *When AI Meets Finance (StockAgent): Large Language Model-Based Stock Trading in Simulated Real-World Environments*. Manuscript.
- Liu, Y., M. Ott, et al. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv:1907.11692*. URL: <https://arxiv.org/abs/1907.11692>.
- Liu, Z. et al. (2021). "FinBERT: A Pre-Trained Financial Language Representation Model for Financial Text Mining". In: *Proceedings of IJCAI*.
- Lopez-Lira, A. and Y. Tang (2023). "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models". In: *arXiv preprint arXiv:2304.07619*.
- Loughran, T. and B. McDonald (2011). "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks". In: *The Journal of Finance* 66.1, pp. 35–65.
- Luo, W. and D. Gong (2024). "Pre-trained Large Language Models for Financial Sentiment Analysis". In: *arXiv preprint arXiv:2401.05215*.
- Ma, Q. et al. (2024a). "Computational Experiments Meet Large Language Model Based Agents: A Survey and Perspective". In: *arXiv preprint arXiv:2402.00262*. URL: <https://arxiv.org/abs/2402.00262>.
- Ma, S. et al. (2024b). "The Era of 1-Bit LLMs: All Large Language Models Are in 1.58 Bits". In: *arXiv preprint arXiv:2402.17764*.
- Mathur, R. et al. (2022). "MONOPOLY: A Multimodal Dataset of Monetary Policy Press Conferences". In: *Proceedings of LREC*.
- Meta AI (2024). *Introducing Meta Llama 3*. <https://ai.meta.com/blog/meta-llama-3/>.

- Mishev, K. et al. (2020). "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers". In: *IEEE Access* 8, pp. 131662–131682.
- Mukherjee, R. et al. (2022). "ECTSum: A New Benchmark Dataset for Bullet Point Summarization of Long Earnings Call Transcripts". In: *arXiv preprint arXiv:2210.12467*.
- Nagy, P. et al. (2023). "Generative AI for End-to-End Limit Order Book Modelling". In: *Proceedings of the ACM International Conference on AI in Finance*.
- Nie, Y. et al. (2022). "A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers". In: *arXiv preprint arXiv:2211.14730*.
- (2024). "A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges". In: *arXiv preprint arXiv:2404.01328*. URL: <https://arxiv.org/abs/2404.01328>.
- Pan, J. Z. et al. (2023). "Large Language Models and Knowledge Graphs: Opportunities and Challenges". In: *arXiv preprint arXiv:2308.06374*.
- Pan, S. et al. (2024). "Unifying Large Language Models and Knowledge Graphs: A Roadmap". In: *IEEE Transactions on Knowledge and Data Engineering*.
- Pennebaker, J. W., M. E. Francis, and R. J. Booth (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Lawrence Erlbaum Associates.
- Phogat, K. S., C. Harsha, S. Dasaratha, S. Ramakrishna, and S. A. Puranam (2023). "Zero-shot Question Answering over Financial Documents Using Large Language Models". In: *arXiv preprint arXiv:2311.14722*. URL: <https://arxiv.org/abs/2311.14722>.
- Quan, Y. and Z. Liu (2024). "EconLogicQA: A Question Answering Benchmark for Evaluating Large Language Models in Economic Sequential Reasoning". In: *arXiv preprint arXiv:2405.07938*.
- Quatra, M. L. and L. Cagliero (2020). "End-to-End Training for Financial Report Summarization". In: *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*.
- Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever (2018). "Improving Language Understanding by Generative Pre-Training". In: *OpenAI Technical Report*.
- Radford, A., J. Wu, et al. (2019). "Language Models are Unsupervised Multitask Learners". In: *OpenAI Blog* 1.8.
- Rajpoot, P. K. and A. Parikh (2023). "GPT-FinRE: In-Context Learning for Financial Relation Extraction Using Large Language Models". In: *arXiv preprint arXiv:2306.17519*.
- Sarkar, S. and K. Vafa (2024). "Lookahead Bias in Pretrained Language Models". In: *SSRN*.
- Scao, T. L., A. Fan, et al. (2023). "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model". In: *arXiv preprint arXiv:2211.05100*. URL: <https://arxiv.org/abs/2211.05100>.
- Shah, A., R. Vithani, et al. (2023). "FiNER: Financial Named Entity Recognition Dataset and Weak-Supervision Model". In: *arXiv preprint arXiv:2302.11157*.
- Sharma, S. et al. (2022). "FinRED: A Dataset for Relation Extraction in the Financial Domain". In: *Companion Proceedings of The Web Conference*.
- Srivastava, P., M. Malik, and T. Ganu (2024). "Assessing LLMs' Mathematical Reasoning in Financial Document Question Answering". In: *arXiv preprint arXiv:2402.11194*. URL: <https://arxiv.org/abs/2402.11194>.
- Steinert, R. and S. Altmann (2023). "Linking Microblogging Sentiments to Stock Price Movement: An Application of GPT-4". In: *arXiv preprint arXiv:2308.16771*.
- Stone, P. J., D. C. Dunphy, and M. S. Smith (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.
- Tan, L., C.-P. Lee, and K. Lim (2023). "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research". In: *Applied Sciences* 13.7, p. 4550.
- Tian, C., Y. Zhao, and L. Ren (2019). "A Chinese Event Relation Extraction Model Based on BERT". In: *Proceedings of ICAIBD*.
- Tong, H. et al. (2024). "Ploutos: Towards Interpretable Stock Movement Prediction with Financial Large Language Model". In: *arXiv preprint arXiv:2403.00782*.
- Touvron, H. et al. (2023a). "LLaMA 2: Open Foundation and Fine-Tuned Chat Models". In: *arXiv preprint arXiv:2307.09288*. URL: <https://arxiv.org/abs/2307.09288>.
- (2023b). "LLaMA: Open and Efficient Foundation Language Models". In: *arXiv preprint arXiv:2302.13971*. URL: <https://arxiv.org/abs/2302.13971>.

- Vaswani, A., N. Shazeer, et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*.
- Wan, Q. et al. (2023). "CFERE: Multi-Type Chinese Financial Event Relation Extraction". In: *Information Sciences* 630, pp. 119–134.
- Wang, D. et al. (2024). "DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding". In: *arXiv preprint arXiv:2401.00908*.
- Wang, S., H. Yuan, et al. (2023). "Alpha-GPT: Human-AI Interactive Alpha Mining for Quantitative Investment". In: *arXiv preprint arXiv:2308.00016*.
- (2024). "QuantAgent: Seeking Holy Grail in Trading by Self-Improving Large Language Model". In: *arXiv preprint arXiv:2402.03755*.
- Wang, Z., Y. Li, et al. (2023). "FinVisGPT: A Multimodal Large Language Model for Financial Chart Analysis". In: *arXiv preprint arXiv:2308.01430*.
- Wei, J. et al. (2021). "Finetuned Language Models Are Zero-Shot Learners". In: *arXiv preprint arXiv:2109.01652*.
- Wen, Q. et al. (2022). "Transformers in Time Series: A Survey". In: *arXiv preprint arXiv:2202.07125*.
- Wimmer, C. and N. Rekabsaz (2023). "Leveraging Vision-Language Models for Granular Market Change Prediction". In: *arXiv preprint arXiv:2301.10166*.
- Wu, S., O. Irsoy, et al. (2023). "BloombergGPT: A Large Language Model for Finance". In: *arXiv preprint arXiv:2303.17564*. URL: <https://arxiv.org/abs/2303.17564>.
- Xi, Z. et al. (2023). "The Rise and Potential of Large Language Model Based Agents: A Survey". In: *arXiv preprint arXiv:2309.07864*. URL: <https://arxiv.org/abs/2309.07864>.
- Xie, Q. et al. (2023). "PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance". In: *arXiv preprint arXiv:2306.05443*.
- (2024). "The FinBen: An Holistic Financial Benchmark for Large Language Models". In: *arXiv preprint arXiv:2402.12659*.
- Xu, J. (2024). "GenAI and LLM for Financial Institutions: A Corporate Strategic Survey". In: *arXiv preprint arXiv:2403.03375*. URL: <https://arxiv.org/abs/2403.03375>.
- Xu, L. et al. (2024). "SuperCLUE-Fin: Graded Fine-Grained Analysis of Chinese LLMs on Diverse Financial Tasks and Applications". In: *arXiv preprint arXiv:2404.19063*.
- Xue, S. et al. (2023). "WeaverBird: Empowering Financial Decision-Making with Large Language Model, Knowledge Base, and Search Engine". In: *arXiv preprint arXiv:2308.05361*.
- Yang, H., X. Liu, C. Wang, et al. (2023). "FinGPT: Open-Source Financial Large Language Models". In: *arXiv preprint arXiv:2306.06031*.
- Yang, Y., M. C. S. Uy, and A. Huang (2020). "FinBERT: A Pretrained Language Model for Financial Communications". In: *arXiv preprint arXiv:2006.08097*.
- Yang, Y., Y. Tang, and K. Y. Tam (2023). "InvestLM: A Large Language Model for Investment Using Financial Domain Instruction Tuning". In: *arXiv preprint arXiv:2309.13064*.
- Yao, Y. et al. (2024). "A Survey on Large Language Model Security and Privacy". In: *High Confidence Computing*, p. 100211.
- Yepes, A. J., Y. You, et al. (2024). "Financial Report Chunking for Effective Retrieval Augmented Generation". In: *arXiv preprint arXiv:2402.05131*.
- Yu, J., M. Huber, and K. Tang (2024). "GreedLlama: Performance of Financial Value-Aligned Large Language Models in Moral Reasoning". In: *arXiv preprint arXiv:2404.02934*. URL: <https://arxiv.org/abs/2404.02934>.
- Yu, Y. et al. (2024). "FinMem: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design". In: *Proceedings of the AAAI Symposium Series*.
- Yuan, H., S. Wang, and J. Guo (2024). "Alpha-GPT 2.0: Human-in-the-Loop AI for Quantitative Investment". In: *arXiv preprint arXiv:2402.09746*.
- Yuan, T. et al. (2024). "R-Judge: Benchmarking Safety Risk Awareness for LLM Agents". In: *arXiv preprint arXiv:2401.10019*.
- Yue, T. and D. Au (2023). *GPTQuant's Conversational AI: Simplifying Investment Research for All*. SSRN 4380516.
- Zhang, B., H. Yang, et al. (2023a). "Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models". In: *Proceedings of the ACM International Conference on AI in Finance*.

- Zhang, B., H. Yang, et al. (2023b). "Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose LLMs". In: *arXiv preprint arXiv:2306.12659*.
- Zhang, L. et al. (2023). "FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models". In: *arXiv preprint arXiv:2308.09975*.
- Zhang, W., L. Zhao, et al. (2024). "FinAgent: A Multimodal Foundation Agent for Financial Trading". In: *arXiv preprint arXiv:2402.18485*.
- Zhang, X. et al. (2024). "Dólares or Dollars? Unraveling the Bilingual Prowess of Financial LLMs Between Spanish and English". In: *arXiv preprint arXiv:2402.07405*.
- Zhang, Z., X. Bo, C. Ma, R. Li, X. Chen, Q. Dai, J. Zhu, Z. Dong, and J.-R. Wen (2024). "A Survey on the Memory Mechanism of Large Language Model Based Agents". In: *arXiv preprint arXiv:2404.13501*. URL: <https://arxiv.org/abs/2404.13501>.
- Zhao, H. et al. (2024). "Revolutionizing Finance with LLMs: An Overview of Applications and Insights". In: *arXiv preprint arXiv:2401.11641*. URL: <https://arxiv.org/abs/2401.11641>.
- Zhao, Y. et al. (2023). "DocMathEval: Evaluating Numerical Reasoning Capabilities of LLMs in Understanding Long Documents with Tabular Data". In: *arXiv preprint arXiv:2311.09805*.
- Zhou, T. et al. (2022). "FEDformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting". In: *Proceedings of the International Conference on Machine Learning*.
- Zhou, Y. et al. (2024). "SilverSight: A Multi-Task Chinese Financial Large Language Model Based on Adaptive Semantic Space Learning". In: *arXiv preprint arXiv:2404.04949*.
- Zhu, J. et al. (2024). "Benchmarking Large Language Models on CFLUE: A Chinese Financial Language Understanding Evaluation Dataset". In: *arXiv preprint arXiv:2405.10542*.
- Zmandar, N. et al. (2021). "Joint Abstractive and Extractive Method for Long Financial Document Summarization". In: *Proceedings of the 3rd Financial Narrative Processing Workshop*.
- Zwam, M. van et al. (2020). *Knowledge Graphs for Financial Services*. <https://www2.deloitte.com/content/dam/Deloitte/de/Documents/operations/knowledge-graphs-pov.pdf>.