FinTradeBench: A Financial Reasoning Benchmark for LLMs

Yogesh Agrawal **Aniruddha Dutta** **Md Mahadi Hasan**
Santu Karmaker **Aritra Dutta**
University of Central Florida

Abstract

Real-world financial decision-making is a challenging problem that requires reasoning over heterogeneous signals, including company fundamentals derived from regulatory filings and trading signals computed from price dynamics. Recently, with the advancement of Large Language Models (LLMs), financial analysts have begun to use them for financial decision-making tasks. However, existing financial question answering benchmarks for testing these models primarily focus on company balance sheet data and rarely evaluate reasoning over how company stocks trade in the market or their interactions with fundamentals. To take advantage of the strengths of both approaches, we introduce **FinTradeBench**, a benchmark for evaluating financial reasoning that integrates company fundamentals and trading signals. FinTradeBench contains 1,400 questions grounded in NASDAQ-100 companies over a ten-year historical window. The benchmark is organized into three reasoning categories: fundamentals-focused, trading-signal-focused, and hybrid questions requiring cross-signal reasoning. To ensure reliability at scale, we adopt a calibration-then-scaling framework that combines expert seed questions, multi-model response generation, intra-model self-filtering, numerical auditing, and human–LLM judge alignment. We evaluate **14 LLMs** under zero-shot prompting and retrieval-augmented settings and witness a clear performance gap. Retrieval substantially improves reasoning over textual fundamentals, but provides limited benefit for trading-signal reasoning. These findings highlight fundamental challenges in the numerical and time-series reasoning for current LLMs and motivate future research in financial intelligence.

1 Introduction

Real-world financial analysis requires reasoning on two complementary information sources: company fundamentals and market dynamics. *Company fundamentals* are accounting-based metrics derived from company balance sheets or Securities and Exchange Commission (SEC) filings, such as profitability, leverage, and valuation ratios, that capture a company's underlying financial health (Fama and French, 1992; Harvey et al., 2016). In contrast, *trading signals,* computed from historical price and volume data, capture market dynamics and investor sentiment, including momentum, volatility, and trend reversals (Brock et al., 1992; Jegadeesh and Titman, 1993; Lo et al., 2000; Andersen et al., 2003; Park and Irwin, 2007; Choi, 2021). Effective financial analysis and valuation, therefore, require integrating these two perspectives rather than relying on either source in isolation. Additionally, synthesizing heterogeneous information sources, reasoning over numerical indicators, and interpreting market behavior under uncertainty make financial analysis an inherently challenging task, even for expert human analysts.

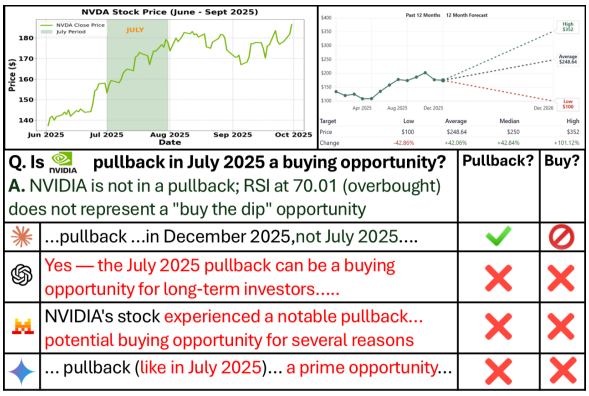

Figure 1: **Performance comparison of proprietary LLMs on a trading signal-focused question.** There was *no pullback* in Nvidia's stock in July 2025, and it was *not a lucrative buying opportunity*; only Claude correctly identified the pullback component. On the buying component, all LLMs failed.

In the advent of artificial intelligence (AI), LLMs are increasingly used to assist analysts with financial analysis tasks such as summarizing earnings calls, interpreting regulatory filings, and answer-

ing questions about company performance and risk (Lee et al., 2025; Yang et al., 2025, 2023; Djagba and Saley, 2025; Wu et al., 2023; Shah et al., 2022; Araci, 2019). Existing benchmarking datasets such as FinQA (Chen et al., 2022a), ConvFinQA (Chen et al., 2022b), and TAT-QA (Zhu et al., 2021) focus primarily on numerical reasoning over financial reports and tables. Recent benchmarks evaluate long-context reasoning and retrieval-based question answering over financial text (Li et al., 2024; Reddy et al., 2025; Islam et al., 2023; Choi et al., 2025). For a comprehensive overview of FinLLMs and their benchmarking, see Lee et al. (2025).

Nevertheless, these datasets rarely assess reasoning over trading signals derived from historical price dynamics and typically do not require models to integrate both sources of information. Consequently, it remains unclear whether current LLMs can answer financial questions that require joint reasoning of company fundamentals and market behavior. E.g., answering the trading question

> *Is NVIDIA's pullback in July 2025 a buying opportunity?*

as in Figure 1, requires reasoning over both company fundamentals and market dynamics. An analyst must consider metrics such as return on assets and cash-flow strength while interpreting trading signals reflected in price momentum and trading volume to determine if the company's stock is undervalued or trading at a premium. As Figure 1 shows, most LLMs fail to reason over the relevant trading signals and give incorrect answers.

A related class of ambiguities arises when fundamentals and market behavior conflict. *E.g., in April 2025, despite weak first-quarter earnings (Earnings per share (EPS) $0.27 vs. $0.42 expected and revenue $19.34B vs. $21B expected), and a generally cautious analyst consensus, Tesla's stock rallied by nearly 20% within days, driven by a forward-looking market narrative rather than contemporaneous fundamentals* (Investing.com, 2025; McDade, 2025). This is a copybook case when investor sentiment and market narratives can drive stock prices independently of current fundamentals (De Bondt and Thaler, 1985; Baker and Wurgler, 2006; Shiller, 2017; Bybee et al., 2023). If one needs to know whether to buy or sell Tesla stock in April 2025, they may not find a reliable answer using financial statements alone[1]. Evaluating such reasoning is challenging, since high-quality annotations require domain expertise, and LLMs often fail to capture numerical fidelity or alignment with expert judgment.

In this paper, we address these challenges by making the following contributions:

(*i*) **FinTradeBench: A Financial Reasoning Benchmark (§3).** We introduce FinTradeBench, a benchmark for evaluating financial reasoning over company fundamentals (from SEC filings) and trading signals (from historical price data). We curate a compact set of signals commonly used in financial analysis, including valuation ratios, leverage metrics, momentum indicators, and volatility measures, and integrate them to support structured reasoning across heterogeneous data sources; see §B Table 6 for the full set of signals. Questions are organized into *fundamentals-focused*, *trading-signal-focused*, and *hybrid reasoning categories* for granular model evaluation. Using a *calibration-then-scaling pipeline*, we combine 150 expert-authored seed questions (50 per category), each with golden key indicators, and scale them across firms and time periods to yield 1,400 total benchmark questions.

(*ii*) **Benchmarking & Evaluation (§4 & §5).** We benchmark 14 LLMs in zero-shot prompting and retrieval-augmented settings and witness a clear performance gap in financial reasoning. Retrieval substantially improves performance on fundamentals-focused questions (↑+37% higher accuracy), and hybrid reasoning questions (↑+55% higher accuracy), but offers limited or negative gains for trading-signal questions derived from time-series data; see Table 2. This suggests that while current LLMs can effectively leverage textual financial information, they struggle to interpret quantitative market dynamics.

2 Background and Related Work

Financial Question-Answering (QA) Benchmarks. The last decade witnessed a surge in question-answering and numerical reasoning datasets in finance. E.g., FinQA (Chen et al., 2022a) and TAT-QA (Zhu et al., 2021) numerical reasoning datasets based on financial reports, tables, and textual disclosures. While ConvFinQA (Chen et al., 2022b) extended these tasks to conversational settings, FinanceBench (Islam et al., 2023), FinDER (Choi et al., 2025), and DocFinQA (Reddy et al., 2025) expanded evaluation to long-context financial reasoning and retrieval tasks over financial documents. These benchmarks signifi-

[1]This is not an isolated event, see §A for analogous cases.

cantly advanced financial QA, with a primary focus on reasoning over textual financial disclosures and accounting-derived indicators. However, they rarely evaluate reasoning over trading signals derived from historical price dynamics or require models to integrate both sources of financial information; see (Lee et al., 2025).

Trading Signals and Quantitative Finance. Trading signals derived from price and volume data play a key role in understanding market behavior and risk dynamics (Lo et al., 2000). Indicators such as momentum, volatility, moving averages, and drawdowns have been widely studied in asset pricing and quantitative trading (Fama and French, 1992; Jegadeesh and Titman, 1993; Lo et al., 2000). Volatility measures are also used to capture perceived market risk and regime changes (Engle, 2004; Ang and Timmermann, 2012; Bollerslev et al., 2015, 2018). While machine learning approaches have recently been applied to forecasting volatility and detecting market regimes (Han et al., 2025; Mishra et al., 2024; Moreno-Pino and Zohren, 2024; Li, 2024), they typically frame the problem as prediction rather than question answering. As a result, existing NLP benchmarks rarely evaluate whether LLMs can reason about these signals for financial analysis.

LLM Evaluation and Benchmark Design. Carefully designed benchmarks for evaluating the reasoning capabilities of the LLMs are important. Prior studies emphasize controlled evaluation settings, high-quality reference answers, and scalable annotation strategies when constructing benchmarks for complex reasoning tasks (Chen et al., 2024a; Ye et al., 2025; Hossain et al., 2025). Financial QA presents additional challenges due to numerical fidelity, domain expertise, and the integration of heterogeneous data sources (Yang et al., 2023; Ran et al., 2019; Zhang et al., 2024).

Comparison with Existing Benchmarks. Table 1 summarizes key differences between existing financial benchmarks and ours. Prior financial QA benchmarks, (Islam et al., 2023; Chen et al., 2022b) emphasize reasoning over textual financial documents, and quantitative finance research (Oberlechner, 2001; Jegadeesh and Titman, 1993), focuses on predictive modeling of trading signals. None explicitly evaluates reasoning over trading signals or the joint interaction between fundamentals and market dynamics. FinTradeBench bridges these two areas by introducing a benchmark that evaluates financial reasoning across both company fundamentals and trading signals within a unified evaluation framework. By explicitly modeling these two signal types, we perform financial reasoning tasks that closely reflect real-world financial analysis.

Table 1: **Comparison of financial QA benchmarks.** The columns indicate different features such as retrieval support (RAG), time-series trading signals (TS), multi-hop reasoning (MH), cross-modal joint reasoning over fundamentals and trading signals (F+T), and LLM-oriented design. For datasets that support these features, we use ✓, × for not supported, and ○ for partially supported.

Dataset	RAG	TS	MH	F+T	LLM
FinQA (Chen et al., 2022a)	×	×	×	×	×
DocFinQA (Reddy et al., 2025)	○	×	×	×	×
ConvFinQA (Chen et al., 2022b)	×	×	×	×	×
FinDER (Choi et al., 2025)	✓	×	○	×	○
FinTextQA (Chen et al., 2024b)	✓	×	○	×	✓
AlphaFin (Li et al., 2024)	✓	×	×	×	✓
FinanceBench (Islam et al., 2023)	✓	×	○	×	✓
FinTradeBench (This paper)	✓	✓	✓	✓	✓

3 FinTradeBench: Benchmark Design

In this section, we construct *FinTradeBench* using the *calibration-then-scaling* paradigm that grounds expert financial intuition in automated large-scale evaluation (Srivastava et al., 2023; Liang et al., 2022; Thrush et al., 2022; Cobbe et al., 2021). This benchmark curation has three primary components, each with multiple sub-components; see the pipeline in Figure 2.

(*1*) **Scope and Data Sources.** FinTradeBench covers NASDAQ-100 companies over a ten-year window (2015–2025), ensuring reporting consistency and availability of both regulatory filings and trading data. For each company-quarter pair, we aggregate two primary sources: (*i*) *Regulatory Filings (10-K/10-Q):* SEC filings from which we extract company fundamentals such as profitability, leverage, valuation, and efficiency ratios. (*ii*) *Daily Trading Data:* OHLCV (Open, High, Low, Close, and Volume) data used to compute trading signals such as momentum, volatility, drawdown, and moving averages. All signals are aligned by ticker & financial quarter to ensure benchmark questions correspond to verifiable historical data; see Table 6 for the full signal list.

Signal selection follows three principles, *Interpretability, Empirical Relevance, and Liquidity*, which are consistent with established asset pricing/trading literature (Fama and French, 1992; Zheng et al., 2023; Jegadeesh and Titman, 1993; Harvey et al., 2016). Signals are organized into: (*i*) *Com-*

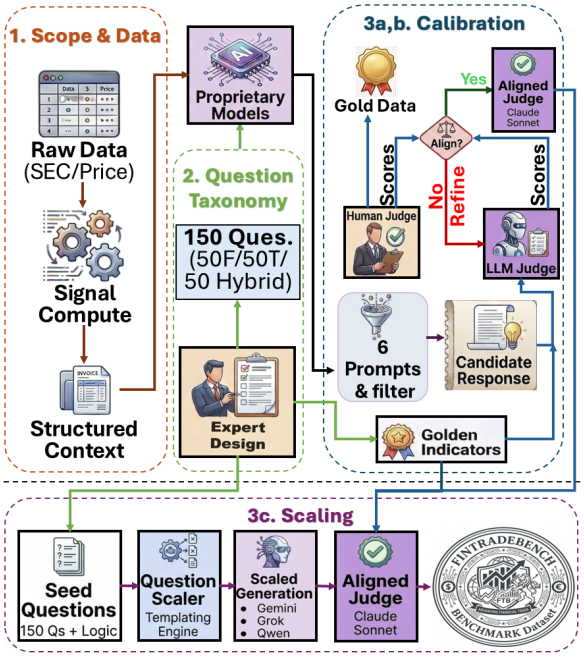

Figure 2: **FinTradeBench design pipeline.** We sketch the 3 primary components and their sub-components on the top block of the pipeline: *1. Data & Design (top left), 2. Question Taxonomy (top middle), and 3. Calibration (top right).* The lower block emphasizes the *scaling phase*, which is a sub-pipeline on its own.

pany Fundamentals, which are accounting-based indicators including return on assets (ROA), return on equity (ROE), earnings-to-price, book-to-price, debt-to-equity, and sales-to-assets. (*ii*) *Trading Signals* which are price time-series-derived indicators including moving averages, momentum, realized volatility, drawdowns, and volume measures.

(*2*) **Question Taxonomy.** We divide the questions into three reasoning categories: (*i*) *Fundamentals-Focused (F-type):* reasoning over accounting-based indicators like ROA, ROE; (*ii*) *Trading-Focused (T-type):* reasoning over market trading signals like price momentum, volatility, and market dynamics; and (*iii*) *Hybrid (FT-type):* joint reasoning across both signals. This taxonomy enables diverse performance analysis and tests whether models can integrate heterogeneous financial signals. See sample questions with gold responses in Table 9.

(*3*) **Calibration-Then-Scaling Framework.** To make our benchmark scalable, unbiased, and reproducible, we use a *calibration-then-scaling* framework; see Figure 2. The framework proceeds in three phases: first, expert-guided seed construction, followed by evaluation and calibration, and finally automated scaling using a calibrated LLM judge (Zheng et al., 2023; Gu et al., 2024).

(*3a*) **Phase 1: Multi-Model Candidate Generation and Self-Selection.** In this phase, we generate responses via LLMs as potential candidates for the gold answer in our benchmark by following three prompting techniques: (*i*) **Multi-model, multi-prompt sampling.** For each question Q, we generate $N = 6$ candidate responses per model using distinct prompt templates derived from the TELeR taxonomy (Santu and Feng, 2023), which defines a structured hierarchy for reasoning-level prompts; see Table 7. This promotes intra-model response diversity while maintaining cross-model comparability, paralleling best-of-N sampling (Chow et al., 2025) and self-reflective refinement (Shinn et al., 2023). (*ii*) **Intra-model self-filtering.** Each model independently selects its best response $a^\star$ by comparing its N candidates on factual accuracy, reasoning completeness, and relevance. This symmetric, bias-neutral design avoids cross-model preference leakage (Li et al., 2025), as each model evaluates only its own outputs. Because all candidates within a model share identical priors, the selection process remains symmetric and acts as a bias-neutral quality filter. This is consistent with prior self-evaluation work (Lee et al., 2024; Yuan et al., 2024; Wu et al., 2024). (*iii*) **Automated numerical audit.** Each self-selected response is audited against a structured financial knowledge base by an independent LLM auditor. Numerical claims are classified as `SUPPORTED`, `CONTRADICTED`, or `NOT_FOUND`, yielding a binary `is_accurate` indicator. To quantify filtering effectiveness, we compute mean numerical accuracy before and after self-filtering, as well as precision, recall, and F1 over the overlap between referenced financial metrics in the response generated (M_{gen}) and ground-truth reference metrics (M_{ref}); see §E.1.

(*3b*) **Phase 2: Evaluation and Calibration.** Following Liang et al. (2022), we ask a financial expert and an independent LLM to evaluate the self-filtered response and then align their evaluation as follows: (*i*) **Human Evaluation.** Domain experts evaluate self-filtered responses $\{a_m^\star\}$ across all models on a 5-point Likert scale for four criteria: factual and numerical accuracy, completeness, relevance, and clarity. Evaluation is performed double-blind to prevent rater bias (Zheng et al., 2023). These human judgments serve as the gold standard against which the automated judge is calibrated. (*ii*) **LLM-as-a-Judge Evaluation.** We evaluate each question Q, self-selected response, $a_m^\star$, and numerical audit summary by giving a structured

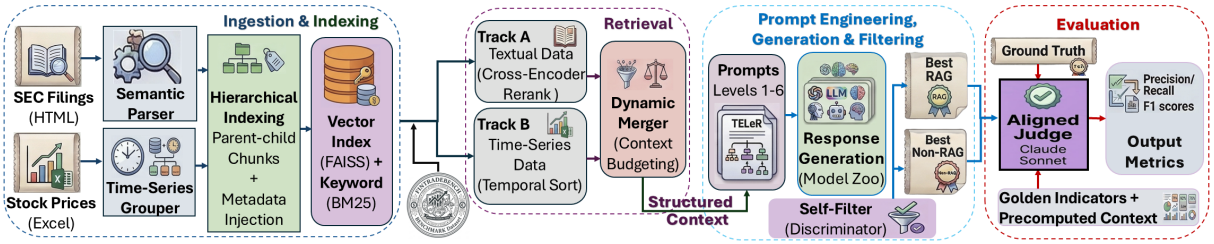

Figure 3: **Overview of the RAG architecture.** The pipeline features a dual-track retrieval engine that processes unstructured text and structured time-series separately. The generation phase utilizes TELeR taxonomy to produce multiple candidate responses across the model zoo, which are filtered via self-selection before final evaluation.

rubric mirroring human criteria to *Claude Sonnet 4.5* (Anthropic, 2025) which acts as an independent LLM judge that is distinct from all generator models. This separation of generator and evaluator mitigates known self-preference biases in LLM-judge systems (Chen et al., 2024a; Ye et al., 2025, 2024; Zheng et al., 2023). We measure Human-LLM-judge-alignment via mean absolute error (MAE) per model and aggregate across models to assess inter-rater consistency (Hossain et al., 2025); see §E.1. We achieve human-LLM alignment through prompt engineering (see §F) on our seed set of 150 questions. Each of them were annotated with *golden indicators*; see sample question in Table 9.

(*3c*) **Phase 3: Scaling.** In this phase, we generate scaled versions of seed questions using companies in NASDAQ 100 spanning 2015-2025 data. As in Phase 1 of §3, we use the same pipeline to generate and filter responses. Finally, using the human aligned LLM judge (MAE$<10\%$,see §G), we evaluate these responses, scaling the benchmark to 1,400 historically grounded financial reasoning questions (Kiela et al., 2021).

4 Experimental Setup

We evaluate LLMs on FinTradeBench under zero-shot (No-RAG) and realistic-RAG conditions to isolate the contribution of heterogeneous information sources: textual/tabular SEC filings vs. numerical time-series market data. This design follows recent recommendations for high-stakes RAG evaluation (Friel et al., 2025; Niu et al., 2024).

Models Evaluated. We evaluate 14 LLMs across FinTradeBench. We categorize the LLMs into large (proprietary and open-source LLMs with better reasoning and parameters $\gtrapprox$ 100B), mid, and small (open-source and distilled or instruction-tuned, parameters ranging from $1-70$B) categories by parameter scale and reasoning capability; see Table 2. Evaluating this diverse set under different signal combinations (Fundamentals (F), Trading (T), and Hybrid (FT)) allows us to isolate how model size and architecture affect performance across heterogeneous financial signals.

Domain-Aware Hybrid Retrieval via RAG. We design a RAG-based architecture for the dual nature of financial analysis by integrating text and tabular data with numerical time-series data. Our four-part RAG implements a Dual-Track Retrieval Engine followed by TELeR-guided generation and self-filtering before evaluation; see Figure 3.

(*i*) **Data Ingestion and Indexing.** Financial documents often contain high token density and complex tables that standard chunking corrupts. We address this via two strategies: (*a*) *Hierarchical Indexing:*A parent-child strategy segments documents by logical boundaries (e.g., SEC "Item 7") (Shaukat et al., 2026; Zhou et al., 2026; Lewis et al., 2021, 2020). Retrieval matches on smaller child chunks ($L = 300$ tokens) trigger the loading of the full parent context, preserving narrative coherence. (*b*) *Metadata Injection:* Structured metadata (ticker, fiscal year) is prepended to every chunk embedding to mitigate temporal hallucination.

(*ii*) **Dual-Track Retrieval Engine.** We use a dual-track retrieval architecture to handle the asymmetric structure of financial evidence; see Figure 3. *Track A* indexes SEC filings using parent–child chunking: smaller child chunks are embedded for retrieval, while larger parent sections are returned to preserve document-level context. Retrieval over this track combines dense embeddings (`BAAI/bge-large-en-v1.5`), BM25 lexical matching, and cross-encoder re-ranking (`ms-marco-MiniLM-L-6-v2`). *Track B* indexes market data as time period-aligned price chunks retrieved via an auxiliary temporal query mechanism; these chunks bypass cross-encoder re-ranking, as semantic relevance models tend to underweight structured time-series evidence. At query time, we dynamically merge the outputs from both tracks: after ticker detection, the sys-

Category	Model	Fundamental (F)			Trading (T)			Hybrid (FT)			Overall		
		No-RAG	RAG	Δ	No-RAG	RAG	Δ	No-RAG	RAG	Δ	No-RAG	RAG	Δ
Large LLMs	DeepSeek-R1	34	42	+23.6%	24.8	24.8	0%	33.2	46.4	+39.8%	30.7	37.7	+23%**
	Gemini 2.5 Flash	31	38.4	+23.8%	24.1	21.6	-10.3%	30.8	40.4	+31.2%	28.6	33.4	+16.7%**
	Gemini 2.5 Flash-Lite	34.4	34.4	0.0%	26.4	21.2	-19.7%	33.2	37.6	+13.1%	31.3	31	-1%
	GPT-5-mini	37.1	42.0	+13.1%	27.8	23.2	-16.4%	34.8	44.1	+26.7%	33.2	36.4	+9.4%*
Mid LLMs	R1-Distill-LLaMA (70B)	34.2	32.3	-5.5%	24.6	20.1	-18.4%	27.1	27.2	+0.3%	28.5	26.3	-7.7%
	R1-Distill-Qwen (32B)	31.7	43.5	+37%	21.6	22	+2.2%	24.1	37.4	+55.1%	25.7	33.9	+32%**
	LLaMA 3.3 70B	29.4	34.8	+18.4%	21.2	21.6	+1.8%	26.8	30.2	+12.7%	25.8	28.9	+11.8%**
	LLaMA 3.3 Instruct (70B)	40.1	36.9	-7.9%	25	20.5	-18.0%	29.6	28.5	-3.7%	31.4	28.4	-9.5%**
	Qwen 2.5 Instruct (32B)	42.3	47	+11.3%	24.8	22.7	-8.5%	33.4	40.2	+20.3%	33.2	36.2	+8.9%**
Small LLMs	LLaMA 3.1 Instruct (8B)	35.7	35	-2%	23.2	20.9	-9.8%	28.2	30	+6.3%	28.9	28.4	-1.7%
	Phi-4 (14B)	36.8	38.6	+4.9%	23.6	23.2	-1.6%	29.6	31.3	+5.8%	29.8	30.8	+3.3%*
	Mistral v0.2 (7B)	33.7	34.2	+1.3%	24.3	29.9	+23.2%	27.4	30	+9.7%	28.3	31.3	+10.6%**
	R1-Distill-Qwen (14B)	30.8	41.3	+33.9%	21	21.8	+3.7%	23.6	35.3	+49.5%	25	32.5	+29.6%**
	LFM 2.5 (1.2B)	24.8	23.5	-5.2%	20.1	20	-0.4%	21.3	21.5	+0.8%	22	21.6	-1.8%

Table 2: **Overall and category-specific accuracy (%) across LLMs.** Here Δ denotes the relative improvement of RAG over No-RAG ($\Delta = (\text{RAG} - \text{No-RAG})/\text{No-RAG} \times 100\%$). Significance assessed via paired t-test ($^{*}p < 0.05$, $^{**}p < 0.01$).

tem retrieves evidence independently per track, applies source-specific quotas, filters by temporal relevance, and removes duplicate parent contexts. We assemble the final prompt under a global token budget, balancing long-form financial texts containing company fundamentals, with short-horizon market evidence needed to calculate trading signals, while preserving signal-specific retrieval strengths.

(*iii*) **TELeR-guided response generation and self-filtration.** In the generation phase, we use the TELeR taxonomy (Santu and Feng, 2023) to generate six distinct prompts per question, ranging from simple directives (L1) to complex RAG-aware reasoning tasks (L6); see §C Table 8. This reduces reasoning errors associated with any single prompt structure. A self-selection module evaluates these candidates against the retrieved context to identify the best RAG and best No-RAG response per model, consistent with the self-filtering approach used during benchmark construction; see §3.

(*iv*) **Evaluation.** An LLM judge evaluates the selected best RAG and best No-RAG responses against ground-truth answers and a set of *golden indicators*, i.e., the key financial metrics required for a correct answer, ensuring evaluation captures both reasoning quality and factual precision.

Evaluation Metrics & Statistical Testing. We evaluate model performance along four dimensions: (*i*) *Absolute Accuracy* normalizes the judge's 1-5 correctness score to a percentage. (*ii*) *Relative Retrieval Delta (Δ)* measures the relative accuracy shift of RAG architectures compared to No-RAG, with statistical significance assessed via paired-samples t-test over question-level scores. (*iii*) *Golden Indicator F1* measures precision and recall over expert-defined financial metrics in model responses. (*iv*) *Integration Score* assesses how well models synthesize textual and tabular signals; see full metric definitions in Table 6.

5 Results and Discussion

Table 2 reports the performance comparison of RAG-based and No-RAG architectures of 14 evaluated LLMs on FinTradeBench. Paired t-tests on question-level correctness scores assess the statistical reliability of RAG-induced changes. Table 3 (and Figure 5 in §H.3) complement this with global generative quality metrics, revealing how RAG reshapes model reasoning behavior beyond raw accuracy. Our analysis surfaces the following findings:

(*1*) **RAG strongly benefits fundamental reasoning (F) and degrades trading signal (T) reasoning.** Fundamental (F) questions require extracting accounting metrics such as debt-to-equity ratios from SEC 10-K/10-Q filings. On them, RAG produced large, statistically significant gains for reasoning-capable LLMs. E.g., *R1-Distill-Qwen (32B)* improved by 37% relative to its No-RAG baseline on F-type questions. Among proprietary models, *Gemini 2.5 Flash* gained 23.8% and *DeepSeek-R1* gained 23.6% on fundamentals. This confirms that hybrid retrieval over text-heavy financial disclosures effectively anchors generation and mitigates hallucination when pre-training representations of fundamental concepts are strong.

Trading (T) questions require computing technical indicators (e.g., momentum, RSI) from raw OHLCV price series. LLMs with RAG perform systematically worse on them. E.g., the performance drops of *Gemini 2.5 Flash-Lite*, *GPT-5-mini*, and *LLaMA 3.3 Instruct (70B)* are in the range of $16.4 - 19.7\%$ relative to their No-RAG base-

lines. Even when the auxiliary temporal retrieval track correctly surfaced the relevant price chunks, models could not reliably parse unrolled numerical tables to derive trend indicators. This suggests that quantitative market data demands intermediate computational steps, such as code execution, rather than retrieval alone. Our study indicates that LLMs struggle to compute metrics on the fly, based only on retrieval. This observation aligns with Cobbe et al. (2021), where LLMs fail to perform robustly on multi-step mathematical reasoning and opens a broader research direction beyond finance that can systematically explore those failure cases.

Takeaway Message: ***The pre-training data corpora of LLMs play a significant role in modulating their performance.*** SEC filings are legally mandated public records, indexed by EDGAR and widely represented in financial datasets (Islam et al., 2023; Yang et al., 2025; Choi et al., 2025; Chen et al., 2022a). As a result, LLMs enter evaluation with strong latent representations of fundamental financial concepts, and RAG acts as a grounding anchor that activates this prior knowledge. In contrast, tick-level trading data and proprietary technical trading signals are commonly behind a paywall (e.g., Bloomberg, Refinitiv). Their scarcity in pre-training mixtures leaves models without the representational framework needed to reason over retrieved numerical price tables; injecting this unfamiliar structure into the context window causes distraction rather than augmentation.

(*2*) **Reasoning LLMs Dominate Hybrid Questions.** Hybrid (FT) questions impose the highest cognitive load in our benchmark, requiring a model to retrieve company fundamentals, trading signal context, and reason across both. Models equipped with latent chain-of-thought reasoning capabilities outperformed standard instruction-tuned models in this category. *DeepSeek-R1* achieved the highest Hybrid RAG accuracy at 46.4%, which is a $\uparrow 39.8\%$ relative gain over its No-RAG baseline. This capability transferred to distilled open-weight models: *R1-Distill-Qwen (32B)* gained $\uparrow 55.1\%$ and *R1-Distill-Qwen (14B)* gained $\uparrow 49.5\%$ on FT questions. By contrast, instruction-tuned models without explicit reasoning steps showed modest or even negative gains in this category. We attribute this to the fact that latent reasoning models allocate additional inference-time computation to reconcile conflicting signals types, precisely the capability that hybrid financial questions demand.

Metric	No-RAG	RAG	Δ (%)
Golden Indicator Precision (%)	0.44	0.2	-55.6%
Golden Indicator Recall (%)	0.22	0.1	-55.8%
Golden Indicator F1 (%)	0.27	0.12	-56.5%
Fundamental Integration (1–5)	1.60	1.81	+13.4%
Trading signal Integration (1–5)	1.54	1.47	-4.6%
Reasoning Depth (1–5)	2.74	2.44	-10.8%

Table 3: **Global Quality Metrics** average across all LLMs.

(*3*) **RAG actively harms certain model families** regardless of their parameter count. While Qwen models and their DeepSeek distillations showed significant improvements ($p < 0.01$), the LLaMA models exhibited systematic degradation. *LLaMA 3.3 Instruct (70B)* declined by $\downarrow 9.5\%$ overall ($p < 0.01$), with drops across all three question categories: $\downarrow 7.9\%$ on F, $\downarrow 18.0\%$ on T, and $\downarrow 3.7\%$ on FT. *R1-Distill-LLaMA (70B)* declined by $\downarrow 7.7\%$ overall, and *LLaMA 3.1 Instruct (8B)* dropped $\downarrow 1.7\%$. Notably, the 14B *R1-Distill-Qwen* model outperformed all three LLaMA models under RAG despite having fewer parameters. This pattern suggests that architecture and pre-training data mixture matter more than scale; LLaMA base weights appear highly susceptible to distraction when presented with dense, jargon-heavy SEC text and unstructured numerical tables, causing the model to abandon its internal reasoning in favor of surface-level context summarization.

(*4*) **LLMs are Distracted by RAG-Based Extra Information Sources.** Table 3 shows that, despite the LLMs' grounding their answers on the fundamental texts when aided with RAG architecture, the reasoning across the golden indicator decreases. This shows LLMs are prone to distraction when extra information is provided. Only *Fundamental Integration* scores improved by $\uparrow 13.4\%$ with RAG. But the precise extraction of expert-defined *Golden Indicators* collapsed; Golden Indicator F1 dropped by $\downarrow 56.5\%$, and overall *Reasoning Depth* dropped by $\downarrow 10.8\%$. This shows that dense financial context improves surface-level factual grounding but actively suppresses the abstract analytical reasoning that expert evaluation requires. Models absorb the retrieved documents and produce fluent, citation-heavy summaries, yet fail to isolate the specific financial signals an expert would prioritize. To resolve this, injecting rich evidence while preserving an LLM's analytical depth remains the central challenge for financial RAG architectures.

5.1 A Case Study on The Impact of Context Quality on Reasoning

To qualitatively illustrate the distraction effect (the final finding), we compare *Gemini 2.5 Flash-Lite* outputs on a Hybrid (FT) query about Apple Inc. (AAPL) under three retrieval conditions: No-RAG, with RAG, and finally, a new class of model, which we call *ideal RAG*. The ideal RAG model is retrieved from a precomputed context. Table 4 shows that the No-RAG model produces generic textbook definitions with no actual data. The standard RAG model grounds its answer in real figures but is overwhelmed by raw revenue totals and daily price points, failing to surface the required Golden Indicators. Only the ideal RAG condition, where precomputed context is provided, activates precise reasoning; the model correctly identifies the Book/Price ratio (0.02), the RSI (60.39), and synthesizes a concrete investment conclusion. This confirms that the bottleneck is not the model capability but the context structure. When numerical financial signals are pre-computed rather than raw, a mid-tier model can reason over them effectively. We provide unimodal ablations (either with precomputed company fundamentals or trading signals) in §H.1, §H.2 extends this analysis across market sectors and query types to confirm the pattern is systemic rather than instance-specific.

Case Study: AAPL Valuation and Momentum

Question: *As of August 2025, is Apple a good buy given its valuation and price trend?*
Golden Indicators: Earnings/Price, Book/Price, RSI, One Day Reversal.

1. Zero-Shot (No-RAG) Output:

"To assess Apple's valuation, we would examine several key metrics... We would need to know Apple's projected EPS... A P/E significantly above its historical average might suggest the stock is expensive. *[Model proceeds to list textbook definitions of P/E, P/S, P/B, and RSI without providing a single actual data point or conclusion for August 2025]...*"

2. Standard RAG Output:

"...Total net sales show a consistent upward trend from 2023 to 2025. Services segment has shown strong growth (14% in 2025)... August 2025 prices show a strong upward trend, starting around $202.38 and reaching $232.56. *[Model gets distracted by raw text/prices, completely failing to calculate or cite the Golden Indicators.]* Without forward-looking earnings estimates, a definitive valuation assessment is not possible..."

3. Ideal RAG (Precomputed Context):

"...Apple's fundamental data presents a mixed picture. A low Book/Price ratio (0.02) suggests the market values Apple significantly higher than its book value, indicating a premium valuation. However, it demonstrates an excellent Return on Equity (35.36%). The market data for August 2025 shows a strong positive price trend... The RSI is at 60.39, suggesting the stock is leaning towards bullish sentiment without being overbought. **Conclusion:** Given the strong positive price trend and excellent profitability, Apple presents a compelling case, though investors must be comfortable with its premium valuation."

Table 4: **Comparison of generative reasoning paths.** Only the Ideal RAG successfully isolates the Golden Indicators to form a concrete, data-backed conclusion.

6 Conclusion and Future Direction

We introduce FinTradeBench to evaluate financial reasoning across company fundamentals, trading signals, and hybrid queries. FinTradeBench is based on a decade of NASDAQ-100 data using a calibration-then-scaling framework that combines expert-authored seeds with scalable automated evaluation. We benchmark 14 LLMs under No-RAG and RAG settings and observe several key patterns. Quantitatively, we witness the highest performance gains across all models in hybrid reasoning-based questions (up to ↑39.8% for large, ↑55.1% for mid, and ↑49.5% for small LLMs). RAG yields statistically significant improvements for fundamental reasoning, where retrieved SEC filings activate relevant prior knowledge. But RAG degrades performance on trading-signal questions, where raw numerical time-series data confuses models rather than assists them. It also introduces information overload, improving surface-level grounding while reducing the precise indicator extraction and reasoning depth required for financial analysis. Additionally, we witness that model architecture plays a significant role. Latent reasoning models (e.g., DeepSeek-R1 and its distillations) substantially outperform instruction-tuned models on hybrid questions, suggesting that inference-time chain-of-thought computation is crucial for reasoning over heterogeneous financial signals. Model families also exhibit different sensitivities to context; Qwen models generally benefit from RAG, while LLaMA models degrade overall, even at 70B parameters. We expect FinTradeBench to support consistent benchmarking for financial reasoning and provide a scalable, industry-standard dataset for researchers and practitioners.

We will release FinTradeBench upon acceptance; presently, we provide a representative subset and evaluation scripts for review. Greater data augmentation and evaluation diversity are essential, and in the future, we plan to augment FinTradeBench, diversify the evaluation pipeline, and explore agentic RAG.

Acknowledgments

Aritra Dutta is partially supported by the Florida Department of Health Grant, AWD00007072, and the National Science Foundation Grant, 2321986.

References

Torben G Andersen, Tim Bollerslev, Francis X Diebold, and Paul Labys. 2003. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625.

Andrew Ang and Allan Timmermann. 2012. Regime changes and financial markets. *Annu. Rev. Financ. Econ.*, 4(1):313–337.

Anthropic. 2025. Claude Sonnet 4.5. https://www.anthropic.com/claude. Accessed: 2025-12-21.

Dogu Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Malcolm Baker and Jeffrey Wurgler. 2006. Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4):1645–1680.

Tim Bollerslev, Benjamin Hood, John Huss, and Lasse Heje Pedersen. 2018. Risk everywhere: Modeling and managing volatility. *The Review of Financial Studies*, 31(7):2729–2773.

Tim Bollerslev, Lai Xu, and Hao Zhou. 2015. Stock return and cash flow predictability: The role of volatility risk. *Journal of Econometrics*, 187(2):458–471.

William Brock, Josef Lakonishok, and Blake LeBaron. 1992. Simple technical trading rules and the stochastic properties of stock returns. *The Journal of finance*, 47(5):1731–1764.

Leland Bybee, Bryan Kelly, and Yinan Su. 2023. Narrative asset pricing: Interpretable systematic risk factors from news text. *The Review of Financial Studies*, 36(12):4759–4787.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.

Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024b. Fintextqa: A dataset for long-form financial question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 6025–6047. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2022a. Finqa: A dataset of numerical reasoning over financial data. *Preprint*, arXiv:2109.00122.

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

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

Jaehyung Choi. 2021. Maximum drawdown, recovery, and momentum. *Journal of Risk and Financial Management*, 14(11):542.

Yinlam Chow, Yu Li, Xuchao Han, Prateek Jain, Ruiqi Sun, Huazhe Xu, Jiawei Gao, Dong Zhou, and Bin Yu. 2025. Inference-aware fine-tuning for best-of-n sampling in large language models. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Werner FM De Bondt and Richard Thaler. 1985. Does the stock market overreact? *The Journal of finance*, 40(3):793–805.

Prudence Djagba and Abdelkader Y. Saley. 2025. Exploring large language models for financial applications: Techniques, performance, and challenges with finma. *Preprint*, arXiv:2510.05151.

Robert Engle. 2004. Risk and volatility: Econometric models and financial practice. *American economic review*, 94(3):405–420.

Eugene F. Fama and Kenneth R. French. 1992. The cross-section of expected stock returns. *The Journal of Finance*, 47(2):427–465.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2025. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *Preprint*, arXiv:2407.11005.

Robin Greenwood, Samuel G Hanson, and Gordon Y Liao. 2018. Asset price dynamics in partially segmented markets. *The Review of Financial Studies*, 31(9):3307–3343.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Beining Han, Anqi Liu, Jing Chen, and William Knottenbelt. 2025. Can machine learning models better volatility forecasting? a combined method. *The European Journal of Finance*, pages 1–22.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. 2016. ...and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

MD Kamrul Hossain, Runpeng Zhang, Haoran Hu, and Kelvin Lo. 2025. Llms as meta-reviewers' assistants: Benchmarking reliability, calibration, and bias in automated paper evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. To appear.

Investing.com. 2025. Earnings call transcript: Tesla's Q1 2025 results fall short, stock rises post-call. investing.com. Accessed: 2026-03-15.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *Preprint*, arXiv:2311.11944.

Narasimhan Jegadeesh and Sheridan Titman. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48:65–91.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, and 1 others. 2021. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pages 4110–4124.

Jean Lee, Nicholas Stevens, and Soyeon Caren Han. 2025. Large language models in finance (finllms). *Neural Computing and Applications*, 37(30):24853–24867.

Soomin Lee, Hyounghun Kim, Sungdong Kim, Joonho Kim, Soojin Park, and Juneyoung Park. 2024. Aligning large language models by on-policy self-judgment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. SELF-JUDGE: self-evaluation and on-policy alignment for LLMs.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Alex Li. 2024. Volatility forecasting in global financial markets using timemixer. *Preprint*, arXiv:2410.09062.

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

Zeming Li, Zhuowan Jiang, Bowen Cheng, Tianle Zhao, Tianyi Zhang, and Yizhe Wang. 2025. Preference leakage: A contamination problem in llm-as-a-judge. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Andrew W Lo, Harry Mamaysky, and Jiang Wang. 2000. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The journal of finance*, 55(4):1705–1765.

Aaron McDade. 2025. Tesla's Q1 earnings miss estimates. investopedia.com. Accessed: 2026-01-15.

Aswini Kumar Mishra, Jayashree Renganathan, and Aaryaman Gupta. 2024. Volatility forecasting and assessing risk of financial markets using multi-transformer neural network based architecture. *Engineering Applications of Artificial Intelligence*, 133:108223.

Fernando Moreno-Pino and Stefan Zohren. 2024. Deepvol: Volatility forecasting from high-frequency data with dilated causal convolutions. *Preprint*, arXiv:2210.04797.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *Preprint*, arXiv:2401.00396.

Thomas Oberlechner. 2001. Importance of technical and fundamental analysis in the european foreign exchange market. *International Journal of Finance & Economics*, 6(1):81–93.

Cheol-Ho Park and Scott H Irwin. 2007. What do we know about the profitability of technical analysis? *Journal of Economic surveys*, 21(4):786–826.

Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. 2019. Numnet: Machine reading comprehension with numerical reasoning. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2474–2484.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2025. Docfinqa: A long-context financial reasoning dataset. *Preprint*, arXiv:2401.06915.

Shubhra Kanti Karmaker Santu and Dongji Feng. 2023. Teler: A general taxonomy of llm prompts for benchmarking complex tasks. *Preprint*, arXiv:2305.11430.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *Preprint*, arXiv:2211.00083.

Muhammad Arslan Shaukat, Muntasir Adnan, and Carlos C. N. Kuhn. 2026. A systematic investigation of document chunking strategies and embedding sensitivity. *Preprint*, arXiv:2603.06976.

Robert C Shiller. 2000. Irrational exuberance. *Philosophy & Public Policy Quarterly*, 20(1):18–23.

Robert J. Shiller. 2017. Narrative economics. *American Economic Review*.

Noah Shinn, G. Labash, D. Gopinath, T. Khot, and A. Sabharwal. 2023. Reflexion: An autonomous agent with dynamic memory and self-refinement. *arXiv preprint arXiv:2303.11366*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Tristan Thrush, Kushal Tirumala, Anmol Gupta, Max Bartolo, Pedro Rodriguez, Tariq Kane, William Gaviria Rojas, Peter Mattson, Adina Williams, and Douwe Kiela. 2022. Dynatask: A framework for creating dynamic ai benchmark tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 174–181.

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

Yikai Wu, Xiao Liu, Yuchen Xu, Yichong Zhou, Zihao Liu, Jiahai Zhou, Geng Cui, and Maosong Sun. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2405.15000*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2025. Fingpt: Open-source financial large language models. *Preprint*, arXiv:2306.06031.

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

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, and 1 others. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Xiang Ye, Bowen Zhang, Yujie Zhu, Jiayi Hu, Xinyi Wang, Pengfei Liu, and Yue Zhang. 2025. Justice or prejudice? quantifying biases in llm-as-a-judge. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

Hongyi Yuan, Runxin Sun, Jie Zhang, Zekun Xu, Geng Cui, Zeyu Zhang, Yang Gao, Zhiyuan Liu, and Maosong Sun. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, and 1 others. 2024. A careful examination of large language model performance on grade school arithmetic. *Advances in Neural Information Processing Systems*, 37:46819–46836.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yongjie Zhou, Shuai Wang, Bevan Koopman, and Guido Zuccon. 2026. Beyond chunk-then-embed: A comprehensive taxonomy and evaluation of document chunking strategies for information retrieval. *Preprint*, arXiv:2602.16974.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *Preprint*, arXiv:2105.07624.

FinTradeBench: A Financial Reasoning Benchmark for LLMs

Supplementary Material

Organization. We organize the Supplementary Material as follows. §A lists additional historical examples that motivate the need for joint reasoning over company fundamentals and trading signals. §B lists all financial signals used in the benchmark construction. §C presents the TELeR prompt taxonomy used for generation. §D provides sample seed questions alongside their expert-authored golden answers. §E details the mathematical formulae and definitions for all evaluation metrics. §F outlines the automated LLM-as-a-Judge prompt and the corresponding human annotation rubric. §G presents the alignment statistics between human experts and the automated judge. §H provides qualitative case studies illustrating failures based on signal types and the RAG distraction effect. §I discusses the limitaions of the paper. Finally, we discuss the ethical considerations in §J.

A Extended Motivating Examples: Fundamentals–Market Divergence

The Tesla April 2025 case discussed in §1 is illustrative of a broader and well-documented phenomenon in financial markets: sustained divergences between company fundamentals and market price dynamics driven by investor sentiment, narrative momentum, and forward-looking expectations. We present three additional historical examples that further motivate the need for benchmarks capable of reasoning over both types of signals jointly.

(*i*) **Amazon (1999–2001).** During the dot-com era, Amazon sustained extremely high market valuations despite persistent operating losses and negative earnings. Rather than anchoring on contemporaneous accounting metrics, investors priced in long-run platform dominance and e-commerce adoption narratives. This example illustrates how growth narratives can sustain valuations that are entirely disconnected from current fundamental signals (Shiller, 2000). A system reasoning solely from financial statements would have consistently flagged Amazon as financially distressed during this period, while the market priced in the opposite trajectory.

(*ii*) **Nvidia (2016–2017).** NVIDIA's stock appreciated substantially in 2016-2017, well before its earnings reports fully reflected the revenue impact of GPU adoption in deep learning workloads. The rally was driven primarily by forward-looking narratives around artificial intelligence and autonomous vehicles, with price movements leading rather than following the fundamental confirmation (Greenwood et al., 2018; Bybee et al., 2023). This case illustrates the temporal asymmetry that motivates hybrid reasoning: trading signals captured the market's forward expectation long before quarterly filings reflected it, meaning neither signal alone would have been sufficient to correctly make the investment decision.

(*iii*) **Tesla (2020): Sentiment-Driven Rally Under Modest Profitability.** Tesla's stock rose dramatically in 2020 despite modest profitability at the time, as investors priced in narratives around electric vehicle adoption and autonomous driving at scale (Shiller, 2017; Baker and Wurgler, 2006). This example predates the main example we give in §1 (the April 2025 earnings miss) and hence establishes that the fundamentals–market divergence is a recurring feature of Tesla's price history rather than an isolated anomaly. It also shows how investor sentiment (Baker and Wurgler, 2006) and narrative economics (Shiller, 2017) can sustain multi-year divergences, not merely short-term reactions.

Takeaway. From these examples, we can see a consistent pattern emerging: price dynamics reflect investor expectations and narrative momentum that often precede or contradict the signal available in the company's financial statements. Table 5 summarises the key characteristics of each case.

These cases collectively motivate FinTradeBench's dual-signal design. A benchmark that evaluates reasoning over fundamentals alone would reward models that correctly identify Amazon, Nvidia, and Tesla as weak or fairly valued, missing the market trading signal entirely. A benchmark that evaluates trading signals alone would capture momentum but provide no mechanism for assessing whether that momentum is anchored in improving fundamentals or purely sentiment-driven. Only by evaluating both types of signals jointly, and by including hybrid questions that require reconciling conflicting signals, can

Table 5: Summary of fundamentals–market divergence episodes. "Dominant Signal" refers to which signal better characterised the market outcome at the time.

Company	Period	Fundamental Signal	Dominant Signal
Amazon	1999–2001	Persistent losses, negative EPS	Market (narrative)
Nvidia	2016–2017	Moderate earnings, GPU revenue lag	Market (forward expectation)
Tesla	2020	Modest profitability, high P/E	Market (sentiment)
Tesla	Apr. 2025	EPS miss ($0.27 vs. $0.42), revenue miss	Market (narrative rally)

we create a benchmark that reflects the reasoning demands of real-world financial analysis.

B Financial Signal Reference

Table 6 summarizes all company fundamentals and trading signals used in question design, including their formulae and economic interpretation. Trading signals are derived from historical OHLCV data; fundamental signals originate from standardised SEC 10-K and 10-Q filings. These variables are among the most widely studied features in the financial domain (Fama and French, 1992; Harvey et al., 2016).

C TELeR Prompt Taxonomy

Table 7 describes the six TELeR-inspired prompt variants used for multi-prompt candidate generation during benchmark construction §3. Prompts vary systematically in instruction richness and reasoning explicitness, from Level 0 (minimal context) to Level 6 (maximal justification)

Table 8 describes the TELeR taxonomy used during RAG evaluation §4. Prompts vary systematically in instruction richness and reasoning explicitness, from Level 1 (single sentence high-level directive) to Level 6 (maximalist, evidence-citing, self-justified). Levels 0–4 are used without RAG; Levels 5–6 have RAG context.

D Sample Questions and Golden Answers

Table 9 presents representative seed questions from each of the three FinTradeBench categories alongside their expert-authored golden answers. Each answer cites the specific golden indicators required for a complete response; these indicators serve as the reference set for Golden Indicator F1 scoring.

E Evaluation Metrics and Statistical Testing

This section provides complete definitions and formulae for metrics used across the benchmark construction pipeline §3 and the RAG evaluation §4.

E.1 Benchmark Construction Metrics

Below, we describe the metrics used during the benchmark construction phase §3.

(*i*) **Numerical Accuracy ($\mathbf{Acc_{num}}(M_m)$)** measures the fraction of numerically accurate statements produced by model M_m across its N candidate responses, computed before and after self-filtering to quantify the filtering benefit

$$\mathrm{Acc}_{\mathrm{num}}(M_m) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\texttt{is\_accurate}(a_i^m) = 1]\,, \tag{1}$$

where a_i^m is the i-th candidate response from model M_m, and $\texttt{is\_accurate}(\cdot) \in \{0, 1\}$ is the binary indicator returned by the automated numerical auditor ($\texttt{SUPPORTED} = 1$, otherwise 0).

(*ii*) **Metric Extraction Precision ($\mathbf{P}(a_m^\star)$), Recall ($\mathbf{R}(a_m^\star)$), and F1 ($\mathbf{F1}(a_m^\star)$)** quantifies the overlap between the set of financial metrics cited in a generated response and the expert-defined reference set. For a self-selected response $a_m^\star$, let $M_{\mathrm{gen}}(a_m^\star)$ denote the set of financial metrics mentioned in the response and M_{ref} the ground-truth reference set. Precision, recall, and F1 are then given by:

$$\mathrm{P}(a_m^\star) = \frac{|M_{\mathrm{gen}} \cap M_{\mathrm{ref}}|}{|M_{\mathrm{gen}}|}, \tag{2}$$

$$\mathrm{R}(a_m^\star) = \frac{|M_{\mathrm{gen}} \cap M_{\mathrm{ref}}|}{|M_{\mathrm{ref}}|}, \tag{3}$$

$$\text{and } \mathrm{F1}(a_m^\star) = \frac{2 \cdot \mathrm{P}(a_m^\star) \cdot \mathrm{R}(a_m^\star)}{\mathrm{P}(a_m^\star) + \mathrm{R}(a_m^\star)}, \tag{4}$$

respectively. The macro-averaged F1 across all K models is:

$$\overline{\mathrm{F1}} = \frac{1}{K}\sum_{m=1}^{K} \mathrm{F1}(a_m^\star), \tag{5}$$

where K is the total number of evaluated models and $a_m^\star$ is the self-selected best response from model M_m.

Table 6: Summary of trading signals and company fundamentals used in the question design.

Signal	Formula	Definition / Description
Trading Signals		
Notation: P_t = Price at time t, V_t = Volume, R_t = Return, N, k = Lookback periods, α = Smoothing factor		
MA (Moving Average)	$\frac{1}{N}\sum_{i=1}^{N} P_{t-i}$	Average price of a stock over a fixed lookback window. Smooths out short-term fluctuations to reveal longer-term trends.
EMA (Exp. Moving Average)	$\alpha P_t + (1-\alpha)\text{EMA}_{t-1}$	Weighted moving average emphasizing recent prices; reacts faster to new information (α is the smoothing factor).
MACD	$\text{EMA}_{\text{short}} - \text{EMA}_{\text{long}}$	Momentum indicator based on the difference between short- and long-term EMAs; captures trend strength and reversal signals.
RSI (Relative Strength Index)	$100 - \frac{100}{1+\text{RS}}$ ($\text{RS} = \frac{\text{Avg Gain}}{\text{Avg Loss}}$)	Scaled measure of recent gains vs. losses; high RSI suggests overbought conditions, low RSI indicates oversold conditions.
OBV (On-Balance Volume)	$\text{OBV}_{t-1} + V_t \cdot \text{sgn}(P_t - P_{t-1})$	Cumulative volume measure linking price movement and trading volume; rising OBV indicates accumulation, falling OBV suggests distribution.
One-Day Reversal	$\frac{P_t - P_{t-1}}{P_{t-1}}$	Daily return from previous close to current close; captures immediate short-term reversals or shocks.
Max Return (20-day)	$\max_{1\le i\le 20}\left(\frac{P_{t-i}-P_{t-i-1}}{P_{t-i-1}}\right)$	Maximum single-day return observed over the past 20 trading days; indicates short-term volatility extremes.
Medium-Term Momentum	$\prod_{i=1}^{k}(1+R_{t-i}) - 1$	Price persistence over several weeks or months; positive values indicate sustained trends.
Long-Term Mean Reversal	$-(P_t - \bar{P}_{long})$	Tendency of price to revert toward historical average, representing equilibrium-seeking behavior in markets.
Company Fundamentals		
Cash Flow / Assets	$\frac{\text{Operating Cash Flow}}{\text{Total Assets}}$	Operating cash flow divided by total assets; measures how efficiently assets generate cash.
Book / Price (Quarterly)	$\frac{\text{Book Value of Equity}}{\text{Market Capitalization}}$	Ratio of book value to market price; high values may indicate undervaluation.
Earnings / Price (Quarterly)	$\frac{\text{Earnings Per Share (EPS)}}{\text{Price Per Share}}$	Inverse of price-to-earnings ratio; higher values imply cheaper valuations relative to earnings.
Forecast Earnings / Price	$\frac{\text{Expected Future EPS}}{\text{Price Per Share}}$	Forward-looking E/P ratio using analyst forecasts; reflects expected profitability.
Sales / Assets (Quarterly)	$\frac{\text{Total Sales}}{\text{Total Assets}}$	Asset turnover ratio; measures how effectively a company uses assets to generate revenue.
Debt / Assets (Quarterly)	$\frac{\text{Total Debt}}{\text{Total Assets}}$	Leverage ratio showing the proportion of assets financed by debt.
Debt / Equity (Quarterly)	$\frac{\text{Total Debt}}{\text{Shareholders' Equity}}$	Ratio of total debt to shareholders' equity; higher values indicate greater financial leverage.
Dividend Yield (Quarterly)	$\frac{\text{Dividends Per Share}}{\text{Price Per Share}}$	Dividends per share divided by stock price; represents cash return to shareholders.
Return on Assets (Quarterly)	$\frac{\text{Net Income}}{\text{Total Assets}}$	Net income divided by total assets; gauges profitability relative to firm size.
Return on Equity (Quarterly)	$\frac{\text{Net Income}}{\text{Shareholders' Equity}}$	Net income divided by shareholders' equity; measures profitability relative to owners' capital.

Table 7: TELeR-inspired prompt variants used for multi-prompt generation during benchmark creation §3.

ID	Attributes	Description (Example Behaviour)
L0	Data-only, no task framing	Provides only *Trading Signals Context* and *Fundamental Data Context* with no explicit question or role; baseline for spontaneous reasoning.
L1	Single-turn, low detail, instruction-style, role-specified	Instructs the model in simple one-sentence instructions focusing on the high-level goal.
L2	Single-turn, moderate detail, instruction-style, role-specified	Paragraph-style instructions expressing the high-level goal and sub-tasks that need to be performed to achieve the goal
L3	Step-by-step reasoning, moderate detail, decomposed	adds a structured reasoning template (bulleted-list-style): clarify goal, decompose into sub-questions, answer sub-parts using both contexts, and synthesize a final answer.
L4	Step-by-step reasoning, moderate detail, decomposed	"Level 3 Prompt" + "It provides a guideline on how LLMs will be evaluated."
L5	Step-by-step reasoning, high detail, decomposed	"Level 4 Prompt" + "Provide additional relevant gathered via RAG"
L6	Step-by-step reasoning, Maximalist, high detail, evidence-citing, justified	"Level 5 Prompt" + "Provide explicit statement asking LLM to explain its own output."

Lvl	Type	Strategy	RAG
L1	Baseline	Single-sentence high-level directive.	No
L2	Focus	Role specification + Paragraph style breakdown.	No
L3	CoT	Bulleted Step-by-Step Chain-of-Thought.	No
L4	Auditor	Adds explicit evaluation criteria (e.g., coherence).	No
L5	Context	Injects retrieved evidence and citations.	**Yes**
L6	Explain	Adds self-justification for data usage.	**Yes**

Table 8: The TELeR Prompt Taxonomy used for Benchmarking and evaluation. The framework systematically increases prompt complexity, with Levels 5 and 6 specifically designed to leverage retrieved context.

(*iii*) **LLM-Judge Agreement ($\overline{\text{MAE}}$)** measures the calibration between human expert scores and automated LLM-judge scores. The primary alignment metric is mean absolute error (MAE(M_m)), which is invariant to score variance and directly measures practical disagreement magnitude (Hossain et al., 2025):

$$\text{MAE}(M_m) = \frac{1}{n_m} \sum_{i=1}^{n_m} \left| S^m_{\text{h},i} - S^m_{J,i} \right|, \qquad (6)$$

and its macro average across all models:

$$\overline{\text{MAE}} = \frac{1}{K} \sum_{m=1}^{K} \text{MAE}(M_m), \qquad (7)$$

where n_m is the number of annotated responses for model M_m, $S^m_{\text{h},i} \in [1, 5]$ is the human Likert score

Sample Golden Answers by Category

1. Company Fundamentals

Q: *Is Nvidia's profitability sustainable with its valuation in September 2025?*
A: Nvidia's profitability is highly sustainable and supported by strong operational efficiency. Its exceptional financial health is characterized by a ROE of 33.38% and a ROA of 21.31%, achieved with low leverage (Debt/Equity of 0.102). However, the valuation is extremely premium, evidenced by a Book/Price ratio of 0.023.

2. Trading Signals

Q: *Based on trading in H1 2025, which stocks show the strongest EMA support?*
A: INTC, MU, and LRCX show the strongest support, trading approx. 18.5%, 16.8%, and 14.4% above their 20-day EMAs. However, LRCX provides the most sustainable profile, as its technical trend is backed by strong fundamentals (18.7% ROE), whereas INTC's support coincides with negative profitability metrics.

3. Hybrid Reasoning

Q: *Is Microsoft overvalued based on Q3 2025 prices despite strong performance?*
A: Microsoft is trading at a premium valuation, but it is not necessarily overvalued. While the Book/Price ratio of 0.0929 indicates the stock is expensive relative to its book value, this premium is supported by high operational efficiency (Cash Flow/Assets of 0.0689) and strong profitability (ROE of 0.0890). Technical indicators show strong momentum (RSI 65.42, MACD > Signal) without yet reaching extreme "overbought" conditions (>70 RSI).

Table 9: Sample seed questions & corresponding golden answers; see details about key financial terms in §E Table 6.

Metric	Definition / Description
Benchmark Construction	
Numerical Accuracy	Fraction of numerically accurate statements per model before and after self-filtering; measures factual grounding (Eq. 1).
Metric Extraction F1	Precision, recall, and F1 over overlap between generated and reference financial metrics; quantifies topical completeness (Eq. 4).
LLM–Judge Agreement (MAE)	Mean Absolute Error (MAE) between human and LLM-judge scores per model (primary metric); Spearman ρ reported where score variance permits (Eq. 6).
Self-Critique Effectiveness	Fraction of models where self-selected response matches the numerically best candidate; evaluates internal self-consistency (Eq. 8).
Prompt Sensitivity	Intra-model F1 variance across TELeR prompt levels; lower variance implies robustness to prompt formulation (Eq. 9).
Overall Composite Score	Weighted aggregate of F1, numerical accuracy, inverse MAE, and prompt variance for cross-model ranking (Eq. 10).
RAG Evaluation	
Absolute Accuracy (%)	Judge's 1–5 correctness score normalised to a percentage; we report it overall and by question type (Eq. 11).
Retrieval Delta (Δ)	Relative accuracy shift of RAG over No-RAG; positive values indicate grounding benefit, negative values indicate distraction (Eq. 12).
Statistical Significance	Paired t-test on question-level score differences; $^{*}p < 0.05$, $^{**}p < 0.01$ (Eq. 13).
Golden Indicator F1	Precision and recall of expert-defined financial signals in model responses; drop under RAG signals distraction effect (Eq. 14).
Context Integration (1–5)	Separate judge scores for fundamental (10-K/10-Q) and trading signal integration; isolates signal-specific failures (Eq. 15).
Reasoning Depth (1–5)	Judge score for logical chain quality independent of factual correctness; decline under RAG signals information overload (Eq. 16).

Table 10: Summary of evaluation metrics used across the benchmark construction and RAG evaluation pipelines.

for response i, and $S^m_{J,i} \in [1,5]$ is the corresponding LLM-judge score. All scores are on the raw 1–5 Likert scale; an MAE of 0.27 corresponds to a 5.4% relative deviation.

(*iv*) **Self-Critique Effectiveness (SCR)** measures how often each model's self-selection identifies its most numerically correct response, assessing internal self-consistency (Lee et al., 2024; Yuan et al., 2024; Wu et al., 2024):

$$\text{SCR} = \frac{\left|\left\{M_m : a^\star_m = \arg\max_{a^m_i} \text{Acc}_{\text{num}}(a^m_i)\right\}\right|}{K}, \tag{8}$$

where $a^\star_m$ is the model's self-selected best response, $\text{Acc}_{\text{num}}(a^m_i)$ is the numerical accuracy of candidate i from model M_m, and K is the total number of models. A higher SCR indicates that self-filtering reliably surfaces the most factually accurate response.

(*v*) **Prompt Sensitivity and Robustness ($\mathbf{Var_{prompt}}(M_m)$)** measures the variance of intra-model F1 scores across TELeR prompt levels, capturing how sensitive a model's output quality is to prompt structure (Chow et al., 2025):

$$\text{Var}_{\text{prompt}}(M_m) = \frac{1}{N}\sum_{i=1}^{N}\left(\text{F1}(a^m_i) - \overline{\text{F1}}_m\right)^2, \tag{9}$$

where $\text{F1}(a^m_i)$ is the indicator F1 score of the i-th candidate response from model M_m, and $\overline{\text{F1}}_m = \frac{1}{N}\sum_{i=1}^{N}\text{F1}(a^m_i)$ is the mean F1 for that model across all N prompt variants. A lower variance indicates greater robustness to prompt formulation.

(*vi*) **Overall Composite Score ($\mathbf{S_{ov}}(M_m)$)** aggregates the four primary signals into a single cross-model ranking metric that balances factuality, alignment reliability, and prompt robustness:

$$\begin{aligned}\text{S}_{\text{ov}}(M_m) = {} & w_1\,\overline{\text{F1}}_m + w_2\,\text{A}_{\text{num}}(M_m) \\ & + w_3\,\overline{\text{MAE}}_m^{-1} \\ & + w_4\left(1 - \text{V}_{\text{p}}(M_m)\right),\end{aligned} \tag{10}$$

where $\overline{\text{F1}}_m$ is model M_m's mean indicator F1 score, $\text{A}_{\text{num}}(M_m)$ is its numerical accuracy, $\overline{\text{MAE}}_m^{-1}$ is the inverse MAE (so that lower judge disagreement yields a higher composite score), $\text{V}_{\text{p}}(M_m)$ is its prompt sensitivity, and weights $\{w_1, w_2, w_3, w_4\}$ balance the four components. All terms are normalised to $[0,1]$ before aggregation.

E.2 RAG Evaluation Metrics

(*i*) **Absolute Accuracy (A**(M_m) measures the factual correctness and overall alignment of the model's response with the expert-provided gold answer. The LLM judge assigns a correctness score on a 1–5 Likert scale, which is normalised to a percentage for comparability across models and question categories:

$$\mathrm{A}(M_m) = \frac{S_J^m}{5} \times 100\%, \tag{11}$$

where $S_J^m \in \{1, 2, 3, 4, 5\}$ is the LLM-judge correctness score for model M_m. We report accuracy globally (Overall) and stratified by question type (F, T, FT); see §2.

(*ii*) **Retrieval Delta ($\Delta(M_m)$)** measures the relative performance shift induced by retrieval augmentation compared to a zero-shot baseline. A positive Δ indicates successful grounding; a negative Δ indicates context distraction or information overload:

$$\Delta(M_m) = \frac{\mathrm{A}_{\text{RAG}}(M_m) - \mathrm{A}_{\text{No-RAG}}(M_m)}{\mathrm{A}_{\text{No-RAG}}(M_m)} \times 100\%, \tag{12}$$

where $\mathrm{A}_{\text{RAG}}(M_m)$ and $\mathrm{A}_{\text{No-RAG}}(M_m)$ are the normalised accuracy scores of model M_m under RAG and No-RAG conditions respectively.

(*iii*) **Statistical Significance (Paired t-test (t))** assesses whether RAG-induced accuracy changes are statistically reliable. We apply a paired samples t-test on question-level correctness scores:

$$t = \frac{\overline{d}}{s_d/\sqrt{N}}, \quad \overline{d} = \frac{1}{N}\sum_{i=1}^{N} d_i,$$
$$d_i = x_{i,\text{RAG}} - x_{i,\text{No-RAG}}, \tag{13}$$

where d_i is the per-question score difference, $\overline{d}$ is the mean difference, s_d is the standard deviation of differences, and N is the number of questions. We report $p < 0.05$ (denoted *) and $p < 0.01$ (denoted **); see §2. A significant result at the ** level indicates, with 99% confidence, that the RAG systematically alters model reasoning rather than producing localised fluctuations on a few queries.

(*iv*) **Golden Indicator F1 ($\mathbf{F1_{GI}}(a_m)$)** measures the precision and recall of expert-defined financial signals in model-generated responses. Using the same formulation as Eq. 4, let $M_{\text{gen}}(a_m)$ denote the financial metrics cited in a model response under evaluation and M_{ref} the expert-defined golden indicator set:

$$\mathrm{P}_{\text{GI}}(a_m) = \frac{|M_{\text{gen}} \cap M_{\text{ref}}|}{|M_{\text{gen}}|},$$
$$\mathrm{R}_{\text{GI}}(a_m) = \frac{|M_{\text{gen}} \cap M_{\text{ref}}|}{|M_{\text{ref}}|},$$
$$\mathrm{F1}_{\text{GI}}(a_m) = \frac{2 \cdot \mathrm{P}_{\text{GI}} \cdot \mathrm{R}_{\text{GI}}}{\mathrm{P}_{\text{GI}} + \mathrm{R}_{\text{GI}}}, \tag{14}$$

where a_m is the model response under evaluation (best RAG or best No-RAG). A drop in $\mathrm{F1}_{\text{GI}}$ under RAG signals the distraction effect: the model is absorbing retrieved content without isolating the specific indicators an expert would prioritize.

(*v*) **Context Integration Scores (FI**(M_m) & **TI**(M_m)**)** are the scores given by the LLM judge, separately scoring two signal-specific integration dimensions, each on a 1-5 scale:

$$\mathrm{FI}(M_m) \in [1, 5], \quad \mathrm{TI}(M_m) \in [1, 5], \tag{15}$$

where FI represents the use of 10-K/10-Q fundamental content, and TI represents processing of numerical time-series data. These scores isolate signal-specific failures: a model may score highly on FI while failing on TI, directly evidencing the gap between the signals. We report macro-averages across all models and separately by question type.

(*vi*) **Reasoning Depth (RD**(M_m)**)** evaluates the quality of the model's logical reasoning chain independently of factual correctness, specifically its ability to chain intermediate analytical steps (e.g., observing momentum → checking leverage → concluding the stock price rally is over-leveraged). This is scored by the LLM judge on a 1-5 scale:

$$\mathrm{RD}(M_m) \in [1, 5]. \tag{16}$$

A decline in ReasonDepth under RAG alongside stable or rising Accuracy indicates the model is summarising retrieved text rather than reasoning analytically over it, which is a clear indication of the information overload effect reported in §5.

A complete summary of all metrics is provided in Table 10.

F Evaluation Prompts and Human Annotation Rubric

This section presents the complete LLM-as-a-Judge prompt and the human annotation rubric used during the calibration phase (§3). Both instruments share the same four scoring dimensions to enable

Group	Bias	MAE
Overall & Dimensions		
Overall composite	−0.021	0.404
Accuracy	−0.059	0.163
Completeness	+0.401	0.545
Relevance	+0.314	0.399
Clarity	+0.317	0.698
By Generator Model		
Gemini 3 Pro	+0.159	0.277
Grok-4.1	+0.345	0.495
Qwen3-235B	+0.222	0.437
By Question Type		
Fundamental (F)	−0.083	0.417
Hybrid (FT)	−0.060	0.348
Trading (T)	+0.087	0.452

Bias = LLM score − Human score.

Table 11: Human–LLM judge alignment results. Scores are on the raw 1–5 Likert scale.

direct human–LLM alignment measurement. The LLM judge additionally performs metric extraction (computing M_{gen}, precision, recall, and F1 against M_{ref}), which is not required of human annotators. Figure 4 presents both instruments side by side.

G Human–LLM Judge Alignment

To validate the LLM-as-a-Judge pipeline described in §3, we measure alignment between automated judge scores and human expert annotations on set of annotated responses to our seed questions spanning all three question types (F, T, FT) and three generator models (Gemini 3 Pro, Grok-4.1, Qwen3-235B). For each response, both a human expert and the LLM judge independently assigned scores on the same four 1–5 Likert dimensions: factual and numerical accuracy, completeness, relevance, and clarity. Table 11 reports the resulting alignment statistics.

(*i*) **Overall Alignment.** We find an overall MAE of 0.40 across all 404 responses, corresponding to an 8.0% relative deviation on the 1–5 scale. At the dimension level, relevance achieves the tightest alignment (MAE = 0.40, 8.0%), followed by completeness (MAE = 0.54, 10.9%) and clarity (MAE = 0.70, 14.0%). The dimension-level accuracy MAE is notably low (MAE = 0.16, 3.3%), which we attribute to both raters awarding near-perfect accuracy scores for filtered responses generated by strong proprietary models. All MAE values fall below 1.0 on a 5-point scale, and the overall bias of −0.02 is close to zero, indicating the LLM

Ablation: AAPL Valuation and Momentum

Question: *As of August 2025, is Apple a good buy given its valuation and price trend?*

Ablation 1: Precomputed Fundamental Context ONLY

"Based on the provided fundamental data for AAPL, I cannot determine if it is a good buy as of August 2025. The fundamental data provided **does not include information about its price trend** or any forward-looking projections that would be necessary to assess its valuation in August 2025."

Ablation 2: Precomputed Trading signal Context ONLY

"Based on the provided volatility and price data for Apple (AAPL) in August 2025, the stock experienced a significant positive price change of 14.84%... The RSI of 60.3940 indicates bullish momentum. **However, I cannot assess Apple's valuation**. Valuation requires fundamental data such as earnings, book value, or cash flow, which are not included in the provided market context."

Table 12: Signal based ablation study for §5.1.

Extended Case 1: Hybrid Reasoning (TSLA)

Question: *As of mid-2025, is Tesla showing signs of fundamental deterioration beneath its technical strength?*
Golden Indicators: Return on Equity, Cash Flow/Assets, EMA, One Day Reversal

1. Zero-Shot (No-RAG) Output

"As a financial analyst, I will assess Tesla's financial health as of mid-2025 by examining key fundamental indicators and comparing them against its observed technical strength... *[Model hallucinates a generic methodology without answering the prompt]*"

2. Standard RAG Output: Textual Distraction

"Based on the provided market data and risk factor disclosures, it's challenging to definitively assess fundamental deterioration beneath its technical strength as of mid-2025... *[Model gets distracted by SEC risk factor boilerplate text and fails to extract any quantitative metrics to prove or disprove the premise]*"

3. Ideal RAG (Precomputed Context)

"...Tesla is not showing clear signs of fundamental deterioration; in fact, some metrics suggest a stable position.
Fundamental Analysis: For every dollar of assets, Tesla generates about 2 cents in Cash Flow (0.0198). Tesla has a very low level of Debt/Assets (0.0562), a strong indicator of financial health. Return on Equity is 1.64%, which is modest but expected for a high-investment growth company.
Technical Analysis: The stock experienced a price increase of 18.45%. The current price is above the 20-day EMA (415.60), suggesting an uptrend. RSI (56.48) suggests balanced market sentiment."

Table 13: Extended Case study for a hybrid focused question.

judge neither systematically inflates nor deflates scores relative to the human expert.

(*ii*) **Variation across models.** Breaking down by generator model, all three models fall within a narrow MAE band (0.28–0.49). This indicates that judge calibration is broadly consistent across generators, with the LLM judge reliably matching human evaluation regardless of which model produced the response.

Variation across question types. Across question types, FT (Hybrid) questions yield the lowest MAE (0.35, 7.0%), while T (Trading) questions yield the highest (0.45, 9.0%). The higher MAE for Trading questions is consistent with their numerical complexity: both human and automated raters show greater disagreement when evaluating responses that require precise computation over time-series data. Nonetheless, all three question types remain within an acceptable alignment range, and no systematic bias toward any single category is observed (bias values: -0.060 for FT, -0.083 for F, $+0.087$ for V).

H Qualitative Case Studies

This section expands on the qualitative analysis in §5.1 with two additional components: unimodal ablations confirming that hybrid reasoning requires both data tracks, and extended cross-sector examples demonstrating that the distraction effect is systemic rather than instance-specific.

H.1 Unimodal Ablations: What Happens When One Signal Is Withheld?

We test Gemini 2.5 Flash-Lite on the AAPL Hybrid (FT) query from §5.1 by providing each signal's precomputed context in isolation. As shown below, the model exhibits strong prompt adherence: when starved of one signal type, it explicitly declines to answer the missing component rather than hallucinating, confirming that hybrid financial reasoning strictly requires a multimodal retrieval pipeline.

H.2 Extended Case Studies: Systemic Distraction Across Sectors

The two examples in Table 13 and 14 sample from different market sectors and question types to confirm that the RAG distraction effect observed for AAPL is not an isolated instance. In each case we compare Zero-Shot, Standard RAG, and Ideal RAG outputs from Gemini 2.5 Flash-Lite.

Extended Case 2: Tabular Trading Signal Failure

Question: *Based on trading in the first half of 2025, which stocks are showing the strongest exponential moving average support?*
Golden Indicators: EMA (Exponential Moving Average)

1. Zero-Shot (No-RAG) Output
"As a financial analyst, I can provide a detailed response to your question about identifying stocks showing strong Exponential Moving Average (EMA) support... *[Model provides a textbook definition of EMA without naming a single stock]*"

2. Standard RAG Output: Tabular Blindness
"I cannot answer your question. The provided context does not contain information about exponential moving averages or any technical analysis indicators. The data provided includes market data (close prices), but... *[Standard semantic search completely failed to map the tabular numerical array to the technical concept of EMA.]*"

3. Ideal RAG (Precomputed Context)
"To determine the strongest EMA support, we look for stocks where the current price is significantly above their 20-day EMA. Based on the analysis, the stocks showing the strongest support are:
1. INTC: Last Close = 36.83, EMA_20 = 31.06. Difference = 18.56%
2. MU: Last Close = 187.83, EMA_20 = 160.82. Difference = 16.80%
3. LRCX: Last Close = 145.81, EMA_20 = 127.48. Difference = 14.37%
These stocks have their current prices trading at a significant premium to their 20-day EMA, suggesting it is acting as a strong support level."

Table 14: Extended Case study for trading focused question.

H.3 Global Quality Metric Figures

Figure 5 presents the grouped bar chart (No-RAG vs. RAG scores per metric) and the relative Δ horizontal chart discussed in Table 3.

I Limitations

While FinTradeBench provides a new benchmark for evaluating financial reasoning across company fundamentals and trading signals, several limitations should be noted.

(*i*) **Market Coverage.** The benchmark focuses on companies in the NASDAQ-100 index over a ten-year period (2015–2025). These firms represent only the technology sector; the benchmark may not fully capture financial reasoning challenges present in all other sectors, emerging markets, or other asset classes such as commodities or fixed-income instruments.

(*ii*) **Signal Coverage.** FinTradeBench includes a curated set of widely used financial indicators derived from SEC filings and historical price data. However, financial analysis in practice may involve additional signals such as macroeconomic variables, analyst forecasts, alternative data sources, or high-frequency market features. Future benchmarks could extend the signal set to incorporate these additional sources of information.

(*iii*) **Temporal Generalization.** The benchmark is based on historical price data from 2015 to 2025. Questions are designed to be answerable from publicly available filings and price data, but the benchmark does not cover forward-looking predictions, real-time market events, or macroeconomic shocks that post-date the evaluation window. Models evaluated on future data releases may exhibit different performance gaps as pre-training corpora evolve.

(*iv*) **Evaluation with LLM Judges.** Our evaluation pipeline relies on an LLM-as-a-Judge calibrated against expert annotations on a seed set of 150 questions. Despite strong measured human–LLM alignment, the judge may not fully replicate the nuanced judgments of professional financial analysts, particularly for subjective or context-dependent reasoning steps (Zheng et al., 2023; Ye et al., 2024). All judge scores should therefore be interpreted as approximations of expert assessment rather than ground truth.

(*v*) **Ideal RAG replicability.** The ideal RAG architecture in §5.1 represents a manually curated upper bound rather than a realistic RAG system. The observed performance ceiling under an ideal context should not be interpreted as achievable by current automated pipelines without further engineering.

(*vi*) **Benchmark Scope.** FinTradeBench focuses on question answering tasks that require reasoning over structured financial indicators and historical market data. The benchmark does not evaluate other important financial tasks such as portfolio optimization, trading strategy generation, or risk management decisions. Therefore, performance on FinTradeBench should be interpreted as measuring financial reasoning capabilities rather than overall financial decision-making ability.

J Ethical Considerations

(*i*) **Financial decision-making risk.** FinTradeBench evaluates language model reasoning over real financial data for named public companies. Scores on this benchmark should not be interpreted as endorsements of any model for live trading, investment advisory, or automated financial decision-making. Even the highest-performing models in our evaluation exhibit substantial error rates, and financial decisions based on LLM outputs carry real economic risk to end users. (*ii*) **Benchmark misuse.** While FinTradeBench is designed for research evaluation, we acknowledge that fine-tuning models specifically to maximise FinTradeBench scores without genuine improvement in financial reasoning could inflate reported performance. We encourage the community to treat benchmark results as one signal among many and to complement automated evaluation with human expert review before drawing strong conclusions about financial reasoning capability. (*iii*) **Annotator and expert involvement.** Human financial experts involved in seed question authoring were part of the research team. Evaluation was conducted double-blind to minimise rater bias. (*iv*) **Societal impact.** Improvements in LLM financial reasoning could benefit retail investors and analysts by democratising access to structured financial analysis. However, the same capabilities could be exploited to automate misleading financial narratives or market manipulation at scale. We call for responsible disclosure norms and human-in-the-loop oversight in any deployment of LLM-based financial analysis tools.

(A) LLM-as-a-Judge Prompt

System: *You are an expert financial analyst and a meticulous fact-checker.*

Input Fields: `[Question]`, `[Reference Metrics]` (M_{ref}), `[Automated Audit Report]`, `[Generated Answer]`

Evaluation Rubric (1–5 scale):

(*1*) **Factual & Numerical Accuracy** –Relies heavily on the Numerical Audit Report.
- **5:** All numerical claims are audit-supported.
- **3:** Minor errors that do not change the overall thesis.
- **1:** Severe hallucinations or math errors that invalidate the conclusion.

(*2*) **Completeness & Context [Critical Human Alignment Rule]** –Does not penalise omission of reference metrics if the response fully answers the question with a highly relevant subset.
- **5:** Fully addresses the prompt with strong explanatory power.
- **3:** Addresses main points but leaves minor sub-questions unanswered.
- **1:** Fails to address the core question or omits critical context.

(*3*) **Relevance & Utility** –Usefulness to an investor or financial decision-maker.
- **5:** Highly actionable, directly answers the prompt without digressing.
- **3:** Generally relevant but includes some tangential information.
- **1:** Misses the point or provides information of no practical value.

(*4*) **Clarity & Rationale [Critical Human Alignment Rule]** –Rewards structured, step-by-step breakdowns; penalises verbosity and repetitive formatting.
- **5:** Crisp, highly readable, actionable, gets straight to the point.
- **3:** Understandable but overly wordy or clunky in formatting.
- **1:** Confusing, disjointed, or buried in jargon.

Few-Shot Anchor Examples:
- *Completeness Anchor:* If a response perfectly answers the question using 2 metrics with strong reasoning, do **not** dock Completeness for omitting a 3rd or 4th reference metric - score it a 5.
- *Clarity Anchor:* If a response is accurate but opens with a long definition of basic concepts, or uses highly repetitive step headers that waste space, cap Clarity at 3.

Output: JSON object containing `qualitative_scores` (four scored dimensions with justifications) and `metric_analysis` (M_{gen}, $M_{\text{ref}} \cap M_{\text{gen}}$, precision, recall, F1).

(B) Human Annotation Rubric

Task: Score AI-generated answers as a financial expert. For each response, provide scores on the five criteria below, and optionally supply a golden answer or comments.

Annotation Criteria:

(*1*) **Audit Validation Agreement (0/1)** – Does your independent review agree with the automated numerical audit's `is_numerically_accurate` flag?
- **1:** Agree –the audit conclusion is correct.
- **0:** Disagree –the audit missed an error, or incorrectly flagged a correct claim.

(*2*) **Factual & Numerical Accuracy (1–5)** –Based on your own review (and the audit), what is the final accuracy score?
- **5:** 100% correct.
- **1:** Contains significant, misleading numerical errors.

(*3*) **Completeness & Context (1–5)** –Does the answer fully address the question and correctly use and contextualise the golden indicators?
- **5:** Excellent. Uses required metrics in a deep, integrated analysis.
- **1:** Superficial. Misses most required metrics or necessary context.

(*4*) **Relevance & Utility (1–5)** –Is every piece of information relevant? Does the response avoid fluff or potentially misleading tangents?
- **5:** High precision; no fluff, no harmful information.
- **1:** Cluttered with irrelevant or misleading content.

(*5*) **Clarity & Rationale (1–5)** –Is the answer clear, well-structured, and does it explain its reasoning?
- **5:** Exceptionally clear and well-reasoned.
- **1:** Confusing, poorly written, or reasoning is opaque.

Output columns to complete: `H_Audit_Agreement` (0/1), `H_Accuracy` (1–5), `H_Completeness` (1–5), `H_Relevance` (1–5), `H_Clarity` (1–5), `H_Golden_Answer` (optional), `H_Notes` (optional).

Figure 4: Evaluation instruments used for human–LLM calibration. **(A)** LLM-as-a-Judge prompt, which additionally extracts M_{gen} and computes Golden Indicator F1 against M_{ref}. **(B)** Human annotation rubric administered as a CSV task. Both instruments share the same four scored dimensions ((*1*)–(*4*) in A, (*2*)–(*5*) in B), enabling direct Spearman ρ and MAE alignment measurement.

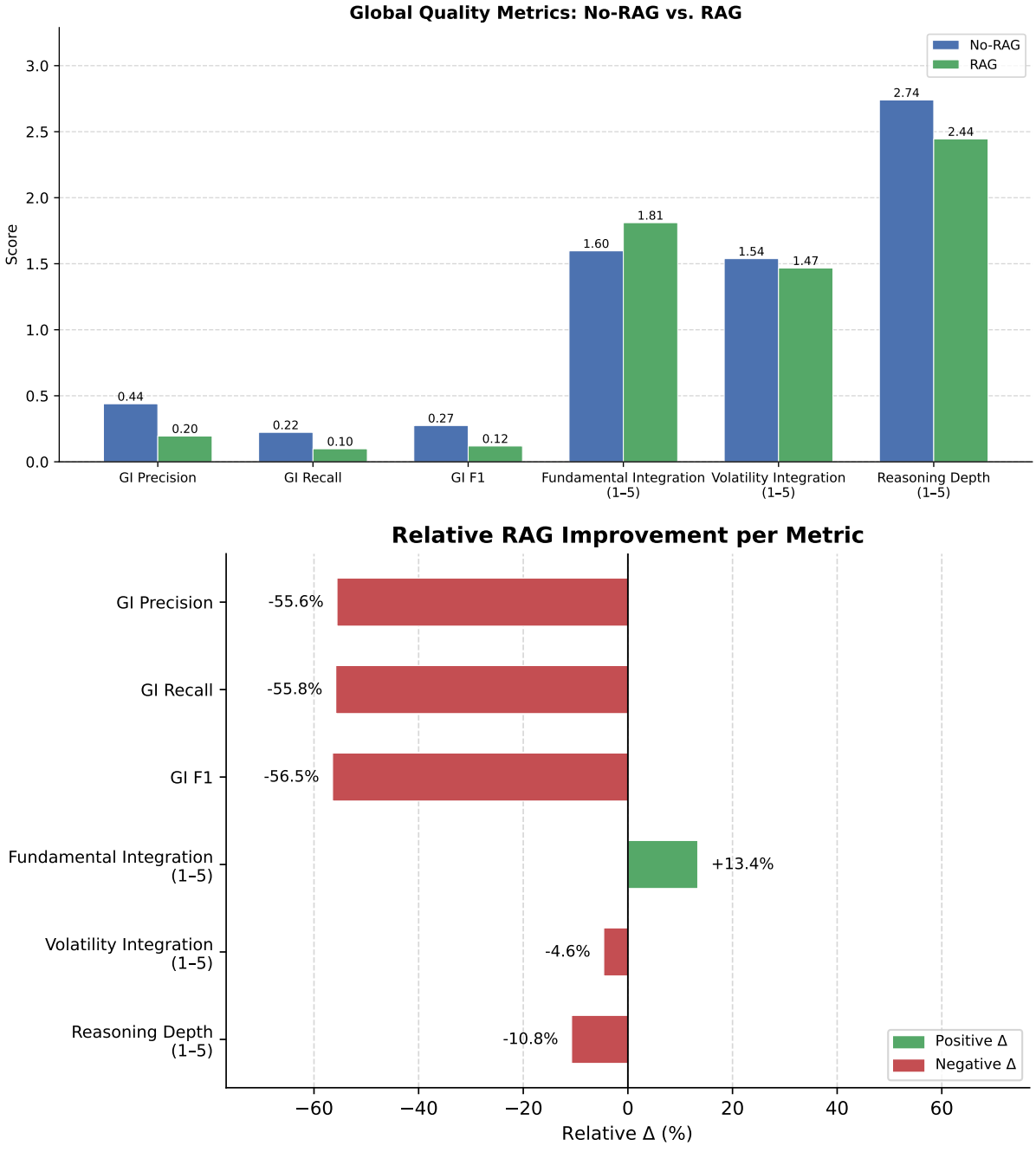

Figure 5: Global Quality metrics: RAG vs No-RAG and improvement