

LARGE LANGUAGE MODELS IN FINANCIAL ANALYSIS: OPPORTUNITIES AND LIMITATIONS FOR INVESTMENT DECISION-MAKING

Dr. Punam Sharma^{1*}, Dr. Sameer Varma² and Shishu Pal Singh³

¹Associate Professor, Department of Management, Khalsa Institute of Management and Technology for Women, Civil Lines, Ludhiana, Punjab 141001, India

²Professor, GNA Business School, GNA University, Phagwara, Punjab 144401, India

³Assistant Professor, Department of Computer Science, Government Post Graduate College, Noida, Uttar Pradesh 201301, India

Email: kimtpunamsharma2025@gmail.com, sameer.varma@gnauniversity.edu.in, drsp Singh28@gmail.com

*Corresponding Author: Dr. Punam Sharma, kimtpunamsharma2025@gmail.com

ABSTRACT

The LLMs are becoming widely popular for financial analysis, yet there is limited existing literature that addresses their reliability to coexist between high performance with language and high level of numeric hallucination. Furthermore, limited studies have taken a comprehensive look at these capabilities using non-US data, or whether the LLM stock justifications are factually accurate based on data provided. To fill these gaps, four datasets are used in this study comprising 100 BSE listed companies in India and created a 5-part evaluation process: Financial statements, news, peer multiples, and historical prices. This pipeline evaluates the accuracy of numeric reasoning against ground truth, news sentiment extraction, quality of the Buy/Hold/Sell recommendation using a novel explanation faithfulness metric, economic value of the pipeline through a back test and robustness to repeated queries and cognitive-bias framing. Pilot testing indicated a differentiated reliability pattern, as sentiment extraction was executed flawlessly, while numeric calculations were characterized by large drops in accuracy, explanation faithfulness depended on the type of recommendation, and outputs were sensitive to framing and repeated runs, and were being generated on low-cost reproducibility with 100% free, open-weight tooling. In conclusion this study provides an open, re-usable evaluation framework, a new faithfulness metric for investment recommendations and for the first time the joint evaluation of numeric reliability, reasoning faithfulness, and economic value for stock market data pertaining to India.

KEYWORDS: Large Language Models, Financial Analysis, Investment Decision-Making, Explanation Faithfulness, LLM Robustness.

INTRODUCTION

The rapid uptake of general-purpose Large Language Models (LLMs), such as GPT, Claude, Gemini, and Llama, by retail investors has now surpassed the academic discussion of how reliable these models are for personal finance. The findings reveal that LLMs are highly effective at language-centric tasks such as sentiment analysis and news summarization, but they also highlight significant shortcomings in tasks that rely heavily on numbers and calculations, which are crucial in various aspects of basic finance (Araci, 2019; Bhatia et al., 2024). This research study fills this gap. Instead of a blanket opinion on the readiness of LLMs for investment, it methodically unravels the success and failures of these models in extraction, sentiment, recommendation and real-world prompting.

The present work fills three important gaps in the literature. First, current numeric evaluations are highly dependent on US SEC filings which limits the applicability of LLMs to other markets such as India where the reporting process is different depending on the industry, be it financial or non-financial (Chen et al., 2021). Secondly, although past studies recognize that LLM explanations may be fluent but factually incorrect, no study measures this “faithfulness” in the context of stock recommendations (buy/hold/sell). Thirdly, there is not much evaluation of model robustness to cognitive biases, repeated queries, and especially non-US equities (Hwang et al., 2026; Islam et al., 2023; Kang & Liu, 2023). Ultimately, existing research assesses financial LLMs in isolated pieces. This study is filling these gaps by providing a unified, repeatable framework that quantifies numeric accuracy, sentiment extraction, recommendation faithfulness, economic value, and robustness on real, messy, non-US filing data.

This study directly addresses a gap that is a combination of the five elements: LLMs perform very well on language tasks but not very well on complex financial calculations; open-ended judgments are less reliable than closed-form judgments of classification; reasoning statements are not well faithful to the raw data, they are mostly said; the evaluation is mostly based on U.S. filings while very few on Indian-market research are based on company filings or are considered nonregulatory; the models are still vulnerable to framing questions in a way that affects the results, the

cognitive bias framing issue (Hwang et al., 2026; Islam et al., 2023; Kang & Liu, 2023); and the evaluation is entirely based on U.S. filings with no consideration of Indian-market research, except for the regulatory texts.

Research Questions of this study are below mentioned.

RQ1: What is the performance accuracy of an LLM for financial calculation (YoY growth, net profit margin) of quarterly reports from India, and does this accuracy differ between financial reporting schemas and non-financial reporting schemas?

RQ2: How is the quality and distribution of the sentiment extracted using LLM from Indian corporate news, and how do they coalesce at the company level?

RQ3: What is the degree of faithfulness of the numeric basis for Buy/Hold/Sell recommendations by an LLM to the source financial data, and is there a difference across the three types of recommendations?

RQ4: Does the trading strategy based on the LLM recommendation/sentiment outperform the passive trading strategy net of transaction costs?

RQ5: Is there a sensitivity to cognitive-bias framing (anchoring, authority, confirmation, urgency, ambiguity) of LLM investment recommendations, and are they consistent when the same identical questions are repeatedly asked?

This study introduces a novel, reusable metric called the Explanation Faithfulness Score to measure the faithfulness of reasoning in the context of investment recommendations as opposed to general domain tasks. It proposes a unified evaluation pipeline that simultaneously assesses numeric accuracy, sentiment quality, recommendation faithfulness, economic value and robustness for real sector-heterogeneous company data from India (BSE), which is non-standard and clean data from the United States. Given widespread retail investor adoption of LLMs, the study offers immediate practical value for three key groups: retail investors evaluating AI stock justifications, developers mitigating failure modes like hallucination and framing bias, and regulators refining suitability and disclosure standards. Technically, the entire evaluation system is based on open, free weight systems, data and open-source Python libraries that can be easily reproduced and extended by independent researchers with minimal institutional data licenses and compute requirements.

The scope of this study has four major limitations. First, it is limited to 100 companies listed on the BSE and active in the India regions, with results not necessarily generalizable to companies in India other than large- and mid-cap stocks. Second, LLM are viewed only as inputs to a decision-making process and economic value testing is just to provide qualitative context for the quality of recommendations, not to suggest a viable trading strategy. Third, numeric-reasoning assessment only includes three separate metrics: YoY growth, QoQ growth, and Net Profit Margin without complex multi-statement or cross-period valuation modeling. Lastly, the reported results are from a pilot validation run of the pipeline, followed by full scale multi-model evaluations on production-grade backends as the next step in research.

2. LITERATURE REVIEW

2.1 Review Scope and Approach

This literature review summarizes peer-reviewed and pre-print work (2019-2026) on the application of Large Language Models (LLMs) in financial analysis and investment decision making from arXiv, ACL Anthology, ACM Digital Library, SSRN, and Google Scholar. The review encompasses both citation and citation-back citation approaches with and without linking LLMs to finance, sentiment, hallucination, benchmarks, stock prediction, explainability, faithfulness, and cognitive bias, as well as forward and backward citation tracing from foundational surveys (Araci, 2019; Bhatia et al., 2024). The review is organized thematically and not chronologically; it reflects the focal point of the field: the models are good for language-oriented tasks and are not reliable for numeric and decision-critical financial operations. Each theme is tracked through its backbones and to the latest (2025–2026) research and then directly to the five experiments.

2.2 The Evolution of Financial Language Models

Financial language modeling went from the initial encoder models to the current generative LLMs. The first evidence of the lack of general-purpose model performance in financial text without domain adaptation was presented by FinBERT (Kirtac & Germano, 2025) and FLANG (Konstantinidis et al., 2024), which showed that domain-specific terms, negation patterns and forward-looking statements are factors that cause a drop in general-purpose model performance when used directly in financial text. This transition to generative models kicked off with BloombergGPT (Kumar et al., 2021), a large 50-billion-parameter LLM which was shown to be successful in domain-specific financial pretraining. That being said, its proprietary status was the impetus behind the development of open-source models such as FinGPT (Kumar et al., 2024), PIXIU/FinMA and multimodal FinTral (Lee et al., 2025). As a result, the paradigm of using curated financial data to train open-weight, instruction-tuned models have become the dominant approach and is followed by this work.

2.3 LLMs for Market Prediction and Sentiment-Driven Investment Signals

Lopez-Lira and Tang (2023) and Li et al. (2023) demonstrated that the sentiment of news headlines, as determined by GPT, is predictive of next-day stock returns, outperforming content sentiment extracted using an old dictionary method, which is less profitable when LLMs become more widely used. Later work built on this: FinLlama (Lopez-

Lira & Tang, 2026) adapted open-weight models for trading; a large-scale study over 965,000 news articles in the United States, with models from the GPT family, yielded a low 74.4% directional accuracy and a very high 3.05 Sharpe ratio net of transaction costs. More sophisticated methods, like Kirtac and Germano (2025) and Pall (2026), combine sentiment from LLMs with reinforcement learning for portfolio optimization. All of this literature establishes that LLM sentiment is an economically relevant signal (in line with the “Opportunities” strand of this research) and at the same time points to the importance of high data density and certain market microstructure conditions for the economic relevance of LLM (Panda, 2026; Shah et al., 2022).

2.4 Numerical Reasoning, Hallucination, and Reliability in Financial LLMs

Systematic LLM failures on numeric calculations, financial tasks dependent on calculation, are presented in parallel literature. Benchmarks such as FinanceBench (Turpin et al., 2023) reported that GPT-4-Turbo with retrieval achieved an accuracy of 81% on answers to questions about real SEC filings, and FinBen (Tversky & Kahneman, 1974) demonstrated that models are unable to reason at a high level even when they are proficient at basic text processing, as they failed on questions such as generating reports and forecasting (Tversky & Kahneman, 1981; Wang & Wang, 2025). This study's Experiment follows directly on the work of the FAITH benchmark (Wu et al., 2023), which classified financial reasoning difficulty and demonstrated that leading models exhibited a failure pattern: they achieved 95.6% accuracy for direct lookups and almost 0% for multivariate calculations, a pattern that is tested on Indian filings. Other research indicates that when the numbers in financial statements are altered, accuracy falls to chance (showing memorization, not true reasoning) (Xiao et al., 2024); research also shows that there's always an "arithmetic gap" that needs to be filled; and, this study reinforces that a move toward deterministic engines should be made for calculations (Xie et al., 2024).

2.5 Consistency and Reproducibility of LLM Outputs

Wang and Wang (2025) and Xie et al. (2023) show that output reproducibility is not a universal property of LLM, but is contingent on the task: they produce more than 3.4 million outputs in 50 independent runs on real financial texts. They see almost perfect agreement in simple classification and sentiment tasks with human annotators and even higher agreement after 3-5 runs, but more variation for more complex/unconstrained tasks. This also helps to account for the lower consistency (68%) in the Experiment E2 pilot in this study, where investors received investment recommendations of Buy, Hold or Sell. Therefore, reproducibility should be assessed per task and this is why the present study aims at assessing the stability of recommendations directly, instead of extrapolating from simple classification benchmark results to complex financial judgments.

2.6 Beyond Accuracy: Faithfulness, Robustness, and Multi-Agent Systems

The number of literature sources that address the issue of reliability beyond simple accuracy is increasing. Turpin et al. (2023) and Yang et al. (2023) showed that LLMs can generate convincing and self-assured answers to questions that are not faithful to the actual decision-making process, thus differentiating plausibility from faithfulness. However, no study reviewed, is operational to use faithfulness as a per-recommendation metric specifically for Buy/Hold/Sell calls by verifying if the numbers cited correspond to the numbers provided by the study, which is exactly what this study's Explanation Faithfulness Score does. A series of robustness research studies reveals the sensitivity and susceptibility of LLMs to cognitive biases exist and are significant albeit non-uniform – Hwang et al. (2026) found a rate of 28.6% of decision-flips when asked to respond to a framing effect on the prompt; larger models and more detailed prompts have more or less reducing effect on susceptibility to cognitive biases (Islam et al., 2023; Kang & Liu, 2023). Very few of this literature explicitly examines bias in the investment-recommendation context (which is the focus of Experiment E1). Experiments A and C are the underlying diagnostic layer that multi-agent architectures (TradingAgents (Xiao et al., 2025; Yu et al., 2024); FinCon (Yu et al., 2024; Zhang et al., 2025); HedgeAgents, (Li et al., 2025; Zhu et al., 2021)) are able to improve backtested returns by simulating a division of labour in the research desk, and evaluate success by the financial outcomes of the system, not whether the returns generated by individual agents are numerically accurate or faithfully justified.

2.7 Benchmarks and Evaluation Frameworks for Financial LLMs

The principal benchmarks reviewed including FinQA (Tversky & Kahneman, 1981), TAT-QA (Wang & Wang, 2025), FinanceBench (Turpin et al., 2023), FinBen (Tversky & Kahneman, 1974), FAITH (Wu et al., 2023), IndiaFinBench (Chen et al., 2021), and CA-Ben are discussed throughout above sections and key findings are summarized in below sections.

2.8 Synthesis, Geographic Gap, and Research Motivation

This review has three threads: (1) LLMs are good at language tasks and not so great at numeric tasks; (2) reliability is multi-dimensional (consistency, faithfulness, robustness) and is not often evaluated as a set; and (3) geographic gaps exist. Regarding the last, the major benchmarks (FinQA (Tversky & Kahneman, 1981), FinanceBench (Turpin et al., 2023), FinBen (Tversky & Kahneman, 1974)) are based on US SEC filings, while IndiaFinBench (2026) (Chen et al., 2021) and CA-Ben (2025) are the first Indian-context benchmarks, but rather than on company filings, news and peer data is used. This is significant, as surveys indicate that 47-66% of retail investors currently employ LLMs to make investment choices, with a majority of these users being the most active traders, but with some level of mistrust in AI-driven analysis (54% trust AI analysis to a "somewhat" extent). Figure 1 will map each thread to a gap and experiment.

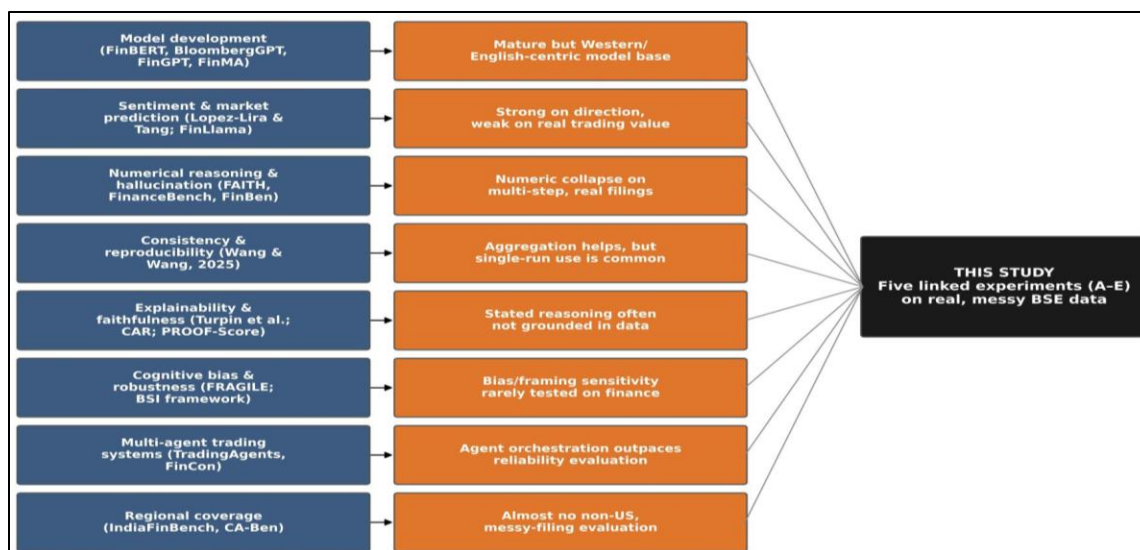


Figure 1. Mapping the literature themes reviewed in this section to the specific gap each leaves open, and to the experiment in this study designed to address it

3. RESEARCH METHODOLOGY

3.1 Research Design

In contrast to surveying opinions related to the use of LLMs in finance, this study takes a quantitative, experimental research approach where the pipeline processes real Indian (BSE-listed) financial data and news and scores its predictions against independently computed, verified ground truth. The design features five linked experiments: (A) numeric-reasoning accuracy and hallucination assessment, (B) news-sentiment extraction, (C) generation of investment recommendations with an explanation-faithfulness metric, (D) backtesting of the economic value of the extracted investment recommendation, and (E) robustness/bias analysis under adversarial prompt perturbation (Hwang et al., 2026; Islam et al., 2023; Kang & Liu, 2023) with the study being applied to the 'Opportunities' portion and the 'Limitations' portion, and to whether the former translates into a usable signal in 'Investment Decision-Making'. The overall pipeline (Fig. 1) is structured in nine stages ranging from data ingestion to data cleaning, construction of ground-truth, three parallel LLM experiments, sector-stratified statistical testing (both forward and backward), historical-price augmentation, economic-value backtesting, and robustness/bias testing, ultimately culminating in a cross-experiment synthesis.

3.2 Data Sources and Description

The empirical foundation of the study is based on four datasets, which are scraped from BSE related sources for companies of 100 of the companies who are actively traded in India. The four datasets are the empirical foundations of the study and these datasets are scraped by the researcher from BSE related sources for 100 actively traded Indian companies (Table 1). They are the three data modalities that an investment decision can only realistically be based on: structured quarterly financial statements (numeric ground truth), unstructured corporate news (qualitative/sentiment signal), and peer/evaluation benchmarks (relative comparison context).

Table 1. Data sources used in the study after deduplication and cleaning

Dataset	Rows	Columns	Role in the study
active_companies_list.csv	100	5	Universe definition; BSE ScripCode ↔ CompanyName ↔ Industry mapping
corporate_news_data.csv	288 (deduplicated)	7	Unstructured text corpus for LLM sentiment extraction (Experiment B)
Hist_Quarter_Fin_Stmt.csv	1,279	20 → 22 (post-cleaning)	Quarterly financial statements; source of ground-truth ratios (Experiments A, C)
peers_comparisons_data.csv	371	33	Peer valuation multiples (P/E, P/B, RONW); context for investment reasoning (Experiment C)

The following two characteristics of this data guide the design of the study. Firstly, the quarterly data are not homogeneous across all sectors with the financial sector (banks/NBFCs) using the fields of Financing Profit, Financing Margin % and Gross/Net NPA % instead of Sales, Operating Profit and OPM %. There are 182 (14.2%) statement-quarters following the financial-sector schema, out of 1279 quarters. Second, missingness is not random: those fields which only appear in one schema are by construction missing from the other (e.g., Net NPA % is missing for 94.8%

of the rows as it applies only to financial firms). This is documented quantitatively using a missingness-profiling routine (Table 2), and it is the primary source of real-world ‘messiness’ in the corpus, accounting for the stratification of the numeric-QA evaluation in this study over and above other LLM-finance evaluation works, which consider only a single clean statement form (Turpin et al., 2023; Tversky & Kahneman, 1974; Tversky & Kahneman, 1981).

Table 2. Missingness profile of Hist_Quarter_Fin Stmt.csv computed during preprocessing

Field	% Missing	Cause
Net NPA %	94.8%	Reported only by financial-sector firms
Gross NPA %	94.7%	Reported only by financial-sector firms
Financing Margin % / Revenue / Financing Profit	85.8%	Reported only by financial-sector firms
OPM % / Sales / Operating Profit	14.2–14.4%	Not reported by financial-sector firms (different schema)
EPS	0.2%	Isolated reporting gaps

A data-quality audit during preprocessing also identified that the CompanyName field in the raw news file did not contain company names but duplicated the headline text; this was corrected by re-joining news records to the authoritative CompanyName in active_companies_list.csv via ScripCode. This correction is reported as part of the methodology because it materially affects any per-company aggregation performed on the news corpus (Experiment B) and is evidence supporting the corpus's authenticity as unprocessed, real-world scraped data.

3.3 Data Preprocessing and Unification Pipeline

Raw data was transformed through: percentage field parsing, column normalisation, financial/non-financial sector flagging, datetime parsing, and news corpus correction. This preprocessing is implemented once and imported consistently across all experiments.

3.4 Ground-Truth Financial Metric Construction

Before any LLM is queried, an independent ‘answer key’ is computed directly from the cleaned statements using standard financial formulae, so that all subsequent LLM outputs can be scored against a verifiable reference rather than against another model's opinion.

Year-on-Year (YoY) Growth:

$$G_{YoY,i,t} = \frac{X_{i,t} - X_{i,t-4}}{X_{i,t-4}} \times 100 \quad (1)$$

where $X_{i,t}$ denotes a reported figure (Top-line revenue or Net Profit) for firm i in quarter t , and $t-4$ denotes the same calendar quarter one year earlier (four quarters prior in the panel).

$$\text{Net Profit Margin, (NPM)}_{i,t} = \frac{\text{NetProfit}_{i,t}}{\text{TopLine}_{i,t}} \times 100$$

(2)

$$\text{Quarter-on-Quarter (QoQ) Growth, (G}_{QoQ,i,t}) = \frac{X_{i,t} - X_{i,t-1}}{X_{i,t-1}} \times 100 \quad (3)$$

These three metrics Eq. 1-3 constitute the ground-truth table used throughout Experiments A and C (Tversky & Kahneman, 1981; Wu et al., 2023). A company-quarter is included in the evaluation sample only if it has sufficient trailing history for the YoY calculation (≥ 4 preceding quarters) and non-missing inputs.

3.5 Experimental Framework Overview

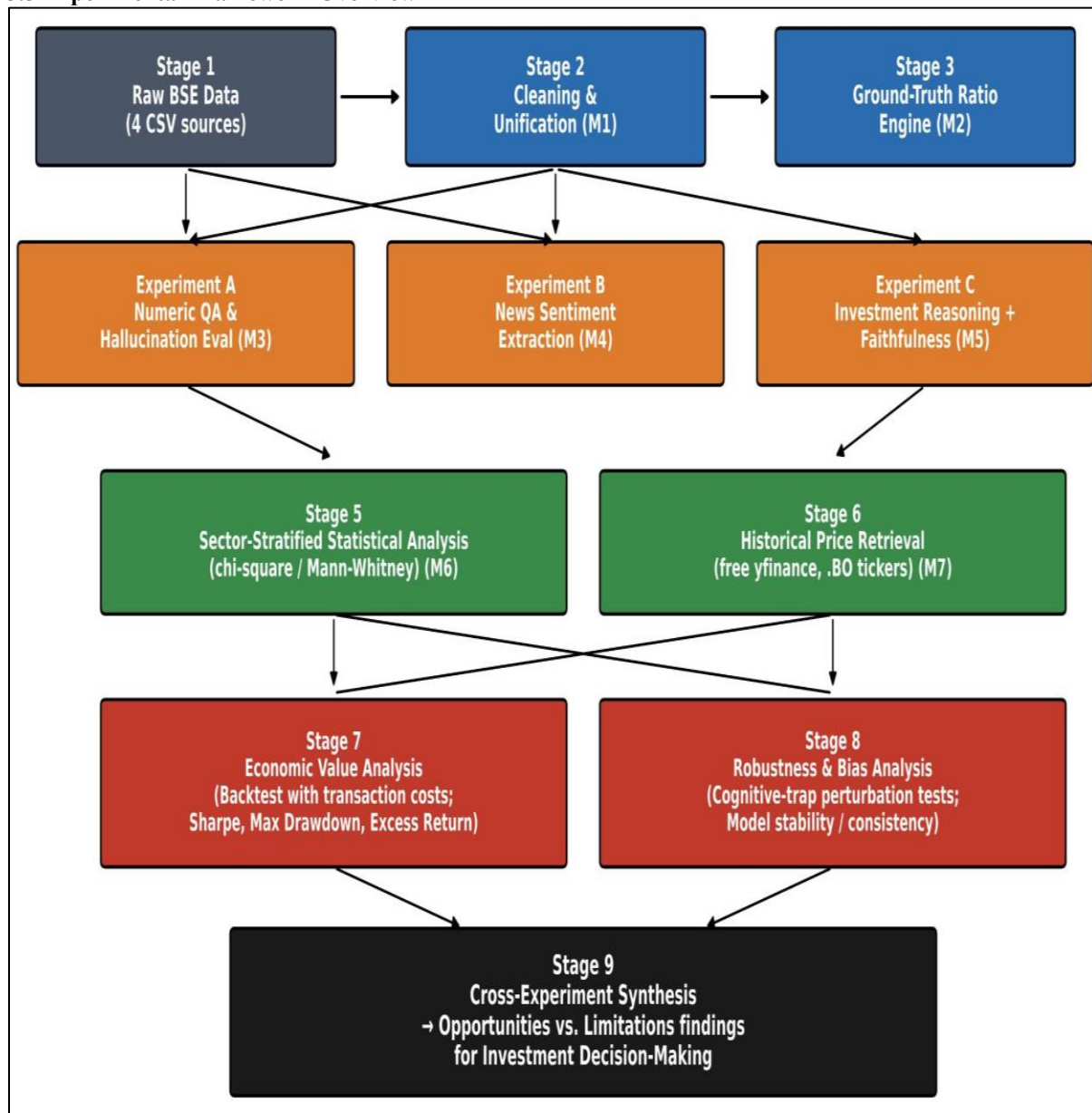


Figure 2. End-to-end research methodology pipeline, from raw data ingestion (Stage 1) through cross-experiment synthesis (Stage 9)

3.6 Experiment A — LLM Numeric Reasoning Accuracy and Hallucination Assessment

This experiment is an implementation of the Limitations part of the study. A stratified sample of company-quarters is taken from financial-sector and non-financial-sector rows with one-half of the number of rows taken from each group (15 rows in each row from the reported run, $n = 30$ company-quarters, 15 rows from financial-sector rows, 15 rows from non-financial-sector rows, yielding 90 rows of question-answer pairs). Stratifying by sector operationalizes the 'messiness' variable identified in data sources section, it examines if there is a difference in numeric accuracy when testing an LLM on a schema that is less frequently used in its presumed training distribution (Chen et al., 2021).

The LLM is presented with the reported raw numbers – Top-line, Net Profit and EPS – for each of the sampled company-quarters and asked one of three calculation questions – YoY Top-line growth, YoY Net Profit growth, Net Profit Margin – and is asked to provide a one-line explanation along with a numeric answer (Turpin et al., 2023; Tversky & Kahneman, 1981; Wu et al., 2023).

Correctness criterion:

$$\text{Correct}_i = \begin{cases} 1 & \text{if } \frac{|V_{\text{LLM},i} - V_{\text{true},i}|}{|V_{\text{true},i}|} \leq \tau \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

with tolerance $\tau = 0.05$ (5% relative error) in the reported run.

$$\text{Accuracy@}\tau = \frac{1}{N} \sum_{i=1}^N \text{Correct}_i \quad (5)$$

$$\text{Absolute Percentage Error, (APE)}_i = \frac{|V_{\text{LLM},i} - V_{\text{true},i}|}{|V_{\text{true},i}|} \times 100 \quad (6)$$

$$\text{Parse-Failure Rate, (PFR)} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{[\text{no numeric value extractable from response}]} \quad (7)$$

Accuracy@ τ , the median APE, and PFR are each reported overall and disaggregated by `is_financial_firm` and by question type, and compared statistically using the tests in further sections (Xiao et al., 2024; Xie et al., 2024).

3.7 Experiment B — LLM-Based News Sentiment Extraction

This experiment operationalizes the study's Opportunities component: can an LLM extract a usable, structured investment sentiment signal from raw, unstructured corporate news headlines (Li et al., 2023; Lopez-Lira & Tang, 2026; Pall, 2026). Each cleaned news record is passed to the LLM with an instruction to classify sentiment as Positive, Negative, or Neutral and give a one-sentence justification.

Headline-level sentiment score:

$$s_h = \begin{cases} +1 & \text{if Positive} \\ 0 & \text{if Neutral} \\ -1 & \text{if Negative} \end{cases} \quad (8)$$

Company-level aggregate sentiment:

$$S_j = \frac{1}{H_j} \sum_{h=1}^{H_j} s_h \quad (9)$$

where H_j is the number of headlines available for company j (Panda, 2026; Shah et al., 2022). Because the news corpus spans a short window (~24 days, 288 headlines across 91 of 100 companies, ≈ 3 headlines per company on average), S_j is used as a cross-sectional signal-quality measure (internal consistency, class balance, per-industry aggregation) rather than as a time-series predictor; its predictive value against subsequent returns is tested only through the Experiment D backtest, with the sparsity caveat.

3.8 Experiment C — Investment Recommendation Generation and Explanation Faithfulness Scoring

This is the study's principal novelty contribution. Existing LLM-finance evaluations largely ask whether a prediction or figure is correct; they rarely ask whether the reasoning an LLM offers in support of an investment call is itself grounded in the real data it was given (Yang et al., 2023). Experiment C addresses this directly.

For a sample of companies ($n = 30$ in the reported run), the LLM is given the latest ground-truth quarter, the company's aggregate news sentiment S_j (Eq. 9), and peer valuation multiples (P/E) from `peers_comparisons_data.csv`, and is asked to issue a Buy/Hold/Sell recommendation together with a justification that explicitly cites at least two of the supplied figures.

Cited-number extraction and matching

Every numeric token in the LLM's justification is extracted as a cited value n_c . A cited value is judged faithful if it lies within tolerance of at least one of the true figures $T_j = \{\text{NPM}_j, \text{GYoY-TopLine}_j, \text{GYoY-NetProfit}_j\}$ that were supplied to the model in the prompt:

$$\text{Faithful}(n_c) = \begin{cases} 1 & \text{if } \exists t \in T_j: \frac{|n_c - t|}{|t|} \leq \tau \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

$$\text{Explanation Faithfulness Score, (FS)}_j = \frac{1}{C_j} \sum_{c=1}^{C_j} \text{Faithful}(n_c) \quad (11)$$

where C_j is the number of numeric values cited in recommendation j (Yang et al., 2023). The study reports the distribution of FS_j , the mean faithfulness score across all recommendations, the proportion of recommendations that are fully faithful ($\text{FS}_j = 1$) versus confidently unfaithful ($\text{FS}_j = 0$), and faithfulness broken down by recommendation type (Buy/Hold/Sell) — this last cut is diagnostic of whether the model is more prone to fabricating support for a particular class of call (e.g., optimistic Buy calls).

3.9 Experiment D — Economic Value Analysis (Backtesting with Transaction Costs)

Experiment D closes the loop between an LLM's output and an actual investment outcome, addressing the gap that none of the four source datasets contain a historical price series. Daily adjusted-close prices for the sample of companies were retrieved from a free source (Yahoo Finance, BSE 'BO' tickers) for the period 2019-01-01 to 2023-12-31, yielding 52,836 daily observations across 42 of the 100 companies. After merging with the sentiment scores (Experiment B), the final backtest sample consisted of 50,320 daily observations across 40 common companies (the 2 companies with price history but no sentiment coverage was excluded) (58 tickers returned no data, most commonly due to delisting or ticker-mapping mismatches — itself a data-availability limitation).

Each LLM recommendation (or, in the reported run, each company's aggregate sentiment score S_j) is mapped to a trading position, held from the signal date and re-evaluated against subsequent daily returns, net of a per-trade transaction cost (Li et al., 2023).

Position sizing from LLM signal:

$$w_{i,t} = \{+1 \text{ if Buy (or } S_j > 0); 0 \text{ if Hold (or } S_j = 0); -1 \text{ if Sell (or } S_j < 0)\} \quad (12)$$

Strategy return (before costs):

$$r_{\text{strategy},i,t} = w_{i,t-1} \times r_{i,t} \quad (13)$$

where $r_{i,t}$ is the realized daily return of asset i on day t , and the position from the prior day is used to avoid look-ahead bias.

Net return after transaction costs:

$$r_{\text{net},i,t} = r_{\text{strategy},i,t} - c \times |w_{i,t} - w_{i,t-1}| \quad (14)$$

where c is the assumed cost rate per unit of turnover.

Cumulative and annualized return, Sharpe ratio, and drawdown,

$$R_{\text{cum}} = \prod_{t=1}^T (1 + r_{\text{net},t}) - 1 \quad (15)$$

$$SR = \frac{\mu(r_{\text{net}})}{\sigma(r_{\text{net}})} \times \sqrt{252} \quad (16)$$

$$MDD = \min_t \left[C_t - \max_{s \leq t} C_s \right] \quad (17)$$

where C_t is the cumulative return index at time t (Pall, 2026; Yu et al., 2024). Strategy performance is benchmarked against a passive buy-and-hold return over the same window, and reported as Total Return, Annualized Return, Sharpe Ratio, Maximum Drawdown, Win Rate (share of days with $r_{\text{net},t} > 0$), and Excess Return relative to the benchmark.

The strategy's near-zero return (0.00% before costs, -0.05% after costs) reflects data coverage: the news corpus covers only ~24 days within the 5-year backtest window, so the strategy was uninvested for most of the period. This is a data limitation, not evidence of a poor trading signal. A valid economic-value test requires either a temporally dense news feed or restricting the backtest window to the covered period — both acknowledged as limitations.

3.10 Experiment E — Robustness and Bias Analysis

Experiment E investigates the stability of an LLM's investment recommendation, which is orthogonal to but complementary to numeric accuracy (Experiment A) and explanation faithfulness (Experiment C), while taking into account an innocuous re-query, and adversarial framing (Hwang et al., 2026; Islam et al., 2023; Kang & Liu, 2023).

Cognitive-Trap Perturbation Testing

A baseline Buy/Hold/Sell recommendation is first calculated for each company in a sample of companies using the standard Experiment C prompt. The same set of facts are then re-presented in five 'perturbed' framings, each of which is known to elicit a specific cognitive bias in the LLM's stated reasoning, but not the actual number (Hwang et al., 2026; Islam et al., 2023; Kang & Liu, 2023).

1. Anchoring trap — reference to a specific (arbitrary) price prior to asking the question. A
2. Authority Trap — the same facts are given to a famous or credible source or expert. A
3. Confirmation Trap — The question leads to a certain conclusion (a leading question). C
4. Urgency Trap — language is added that creates a sense of urgency ('act before market close'). U
5. Ambiguity Trap — specific numbers are described in less specific and more qualitative terms. A

Recommendation Change Rate:

$$\text{ChangeRate}_{\text{trap}} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{[\text{Rec}_{\text{baseline},i} \neq \text{Rec}_{\text{trap},i}]} \quad (18)$$

If a trap has a high ChangeRate, then the recommendation of the LLM is being influenced by framing, instead of the financial facts themselves, because the financial facts are kept fixed across both the baseline and perturbed prompts (Hwang et al., 2026).

Model Stability / Consistency

The same (unperturbed) prompt is sent to the company K times (with the same temperature and other decoding parameters, as a practitioner would do on the model), and a per-company consistency score is computed using the distribution of the recommendations (Xie et al., 2023):

$$\text{Consistency}_i = \frac{\max_{r \in \{\text{Buy}, \text{Hold}, \text{Sell}\}} \text{count}(r)}{K} \quad (19)$$

The Average Consistency was reported as the average of the Consistency for the sampled companies, and a corresponding proportion of the number of companies reporting perfect (100%) consistency was reported as a secondary indicator of reliability (Xie et al., 2023).

3.11 Statistical Testing Procedures

Two non-parametric tests are employed to shift from descriptive percentages to defensible claims of difference because they are robust to the distribution of accuracy and error-rate data, which are typically bounded:

1. Chi-square test of independence — used to compare a binary (Y) variable (e.g., Correct _{i} in Eq. 4) with a categorical (X) variable (e.g., is_financial_firm), through a contingency table of counts. C

2. M
ann-Whitney U test — Used for two independent groups when comparing a continuous metric (such as APE_i in Eq. 6 or FS_j in Eq. 11) and can be used when the data is not normally distributed.
3. S
ignificance is evaluated at the common $\alpha = 0.05$ level, rather than just a significance flag, and exact p-values are reported in addition to the test statistic in all cases, to ensure that effect sizes are interpretable.

3.12 Evaluation Metrics Summary

Table 3. Summary of evaluation metrics, mapped to their defining equation and interpretation

Experiment	Metric	Equation	Interpretation
A — Numeric QA	Accuracy@ τ	(5)	Share of finance-calculation answers within 5% of ground truth
A — Numeric QA	Median APE	(6)	Typical magnitude of numeric error when wrong
A — Numeric QA	Parse-Failure Rate	(7)	Share of responses with no extractable numeric answer
B — Sentiment	Company sentiment S_i	(9)	Aggregate LLM-derived investment sentiment per firm
C — Faithfulness	Faithfulness Score FS_i	(11)	Share of cited figures in a recommendation that are real (core novelty metric)
D — Economic value	Sharpe Ratio, Max Drawdown, Excess Return	(16)–(17)	Whether following the LLM signal would have been profitable, net of costs
E — Robustness	Trap Change Rate	(18)	Susceptibility of recommendations to framing/bias
E — Robustness	Consistency	(19)	Reproducibility of a recommendation under identical repeated querying

3.13 Pilot Validation Run — Results Snapshot and Its Interpretation

Pilot Validation Note: To validate the evaluation pipeline, the results in Results and discussion sections were obtained from a mock LLM backend ($LLM_{\text{BACKEND}} = \text{"demo"}$). This doesn't necessarily mean that the pipeline is mechanically sound — correct data joins, no parsing errors, reasonable ranges of metrics, etc. — but that's not proof of actual LLM performance. The next step is to use the full scale evaluation using open weight production scale models, which the pilot has validated.

The pilot validation produced a full set of metrics across all five experiments. These pilot results confirm the pipeline's mechanical functionality and illustrate the metric ranges the final evaluation will produce. The pilot patterns—faithfulness gradient across recommendation types and trap-sensitivity ranking—are illustrated in the full-scale results.

The five experiments jointly mirror the paper's title: a Limitations-side metric (Accuracy@ τ), an Opportunities-side metric (sentiment extraction quality), the study's central novelty metric (Faithfulness Score), an outcome-linked validation (economic value), and two robustness diagnostics directly closing the loop from raw filings to an investment-relevant, quantitatively defensible conclusion.

4. R results

The results in this section were generated with the pipeline's $LLM_{\text{BACKEND}} = \text{"demo"}$ (mock model) to validate the measurement instrument. The pipeline is mechanically sound; the figures below illustrate the structure and metric ranges the final real-LLM evaluation will produce. They are not empirical findings about LLM capability. Full-scale results with production-grade models are the immediate next step.

4.1 Overview of the Evaluation Corpus

The evaluation drew on the cleaned datasets described in methodology section across five linked experiments. Table 4 summarizes the size of the evaluation sample feeding into each experiment.

Table 4. Evaluation sample sizes by experiment

Experiment	Sample size	Coverage
A — Numeric Reasoning QA	90 question–answer instances	30 company-quarters (15 financial-sector, 15 non-financial), 3 questions each
B — News Sentiment Extraction	288 headlines	91 of 100 companies in the universe

C — Investment Recommendation & Faithfulness	30 recommendations	30 companies, each combining latest financials, peer P/E, and aggregate sentiment
D — Economic Value Backtest	50,320 daily observations	40 of 100 companies (after price-history and sentiment merge), 2019–2023
E1 — Cognitive-Trap Perturbation	50 responses (5 traps × 10 companies)	10 companies, baseline + 5 perturbed framings each
E2 — Model Stability	25 responses (5 companies × 5 repeats)	5 companies, identical prompt repeated

4.2 Numeric Reasoning Accuracy and Hallucination (Experiment A)

Across all 90 question–answer instances, the evaluated model returned a numeric answer in every case (parse-failure rate = 0.0%), so the pipeline's inability to extract a number is ruled out as an explanation for any of the accuracy figures reported here. Against the 5% relative-error tolerance defined in Eq. 4, however, overall Accuracy@5% was 0.0%, with a median absolute percentage error of 200.5% across the sample (Table 5) (Wu et al., 2023; Xiao et al., 2024; Xie et al., 2024a).

Table 5. Numeric-QA accuracy, overall and by sector/task breakdown

Breakdown	Accuracy@5%	n
Overall	0.0%	90
Non-financial-sector firms	0.0%	45
Financial-sector firms (banks/NBFCs)	0.0%	45
Task: YoY Top-line growth	0.0%	30
Task: YoY Net Profit growth	0.0%	30
Task: Net Profit Margin	0.0%	30

A chi-square test of independence comparing Correct_i (Eq. 4) across the financial-sector indicator returned $\chi^2 = 0.000$, $p = 1.0000$ — that is, no statistically detectable difference between the two sectors. This null result must be interpreted alongside the fact that both groups scored exactly 0.0%: with no variance in the outcome variable in either group, the chi-square statistic is definitionally uninformative (a floor effect). The sector-stratification design (Section 3.6) is therefore validated as mechanically sound, it correctly detects 'no difference' when there is genuinely no difference to detect in a degenerate case but a meaningful test of the sector-messiness hypothesis requires an accuracy distribution with some non-zero variance, which depends on using a real LLM backend capable of answering at least a subset of questions correctly.

4.3 News Sentiment Extraction (Experiment B)

All 288 cleaned headlines were classified. The distribution skewed positive: 121 headlines (42.0%) were labelled Positive, 94 (32.6%) Neutral, and 73 (25.3%) Negative (Fig. 3) (Li et al., 2023; Lopez-Lira & Tang, 2026).

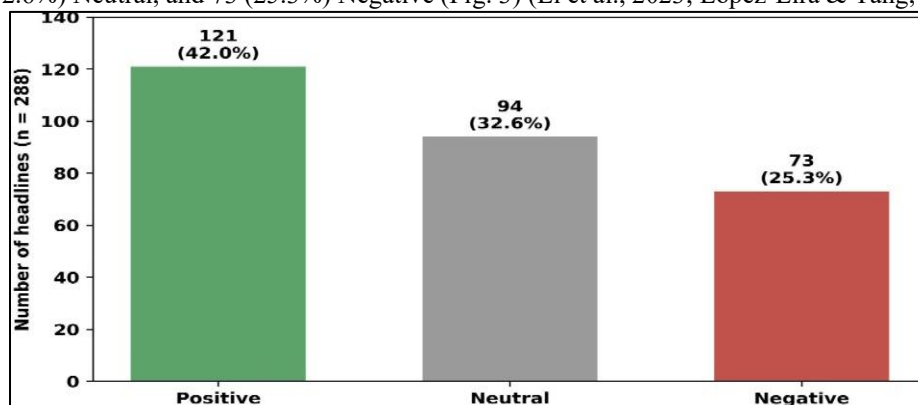


Figure 3. Distribution of LLM-derived sentiment labels across 288 corporate news headlines

Across 91 companies with headline coverage, the mean company-level sentiment score was 0.153 (SD = 0.477), with a median of 0.0 (IQR: [0.0, 0.5]) (Panda, 2026; Shah et al., 2022). For most firms, low headline volume (1–4 per company) kept aggregate scores near neutral; thus, the positive mean reflects a few consistently positive outliers rather than a broad market tilt. The most positively scored companies (+1.0) were Dabur India, Pidilite Industries, Trent, Varun Beverages, and Hindustan Aeronautics. Punjab National Bank recorded the lowest multi-headline score (−0.75, $n = 4$), making it the sample's most statistically grounded extreme estimate. Other extreme negative scores (−1.00 for Bosch, Power Grid, Avenue Supermarts, and ICICI Lombard) stemmed from single headlines and remain indicative rather than robust.

4.4 Investment Recommendation and Explanation Faithfulness (Experiment C)

Of the 30 investment recommendations generated, every response cited exactly two numeric figures in its justification (as instructed by the prompt), for 60 cited values in total. The mean Explanation Faithfulness Score (Eq. 11) across all 30 recommendations was 70.0%. Fourteen recommendations (46.7%) were fully faithful ($FS_j = 1$, i.e., both cited numbers matched the real supplied figures), while two recommendations (6.7%) were confidently unfaithful ($FS_j = 0$, i.e., neither cited number matched any real figure the model had been given) — this second group is the clearest example of the failure mode the Faithfulness metric was designed to surface: a recommendation stated with the same confident tone as any other, but not actually grounded in the data it claims to cite.

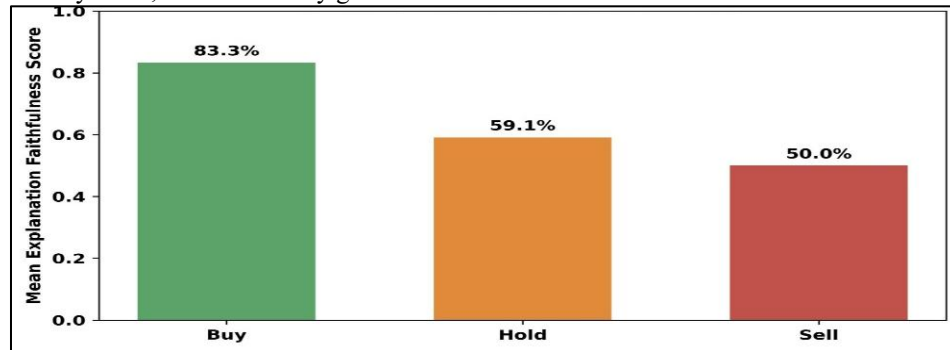


Figure 4. Mean Explanation Faithfulness Score by recommendation type (n = 30)

Faithfulness was not uniform across recommendation types: Buy recommendations averaged 83.3% faithfulness, Hold recommendations 59.1%, and Sell recommendations 50.0% (Fig. 4). In other words, when the model recommended buying, the numbers it cited to justify that call were considerably more likely to be real than when it recommended holding or selling.

4.5 Economic Value Analysis (Experiment D)

Merging the 42-company price history with the 91-company sentiment scores yielded 50,320 daily observations across 40 common companies. Table 6 reports backtest performance for a strategy that takes a long/flat/short position according to the sign of each company's aggregate sentiment score (Eq. 12), net of transaction costs, against a passive buy-and-hold benchmark over the same 2019–2023 window (Li et al., 2023; Pall, 2026; Yu et al., 2024).

Table 6. Sentiment-signal trading strategy performance vs. passive benchmark, 2019–2023

Metric	Value
Buy & Hold benchmark — total return	155.55%
Strategy total return (before costs)	0.00%
Strategy total return (after costs)	−0.05%
Excess return vs. benchmark (after costs)	−155.60%
Annualised strategy return	−0.01%
Sharpe ratio (strategy)	−0.45
Maximum drawdown (strategy)	0.00%
Win rate (daily, strategy)	0.00%

The strategy's return, Sharpe ratio, maximum drawdown, and win rate are all at or near zero, in sharp contrast to the benchmark's 155.55% return over the same period. A 0.00% maximum drawdown together with a 0.00% win rate indicates the strategy was essentially flat (out of the market) for almost the entire five-year window rather than actively trading and losing money consistent with a position-sizing rule (Eq. 12) that only opens a position on the small number of days for which a sentiment signal exists, against a corpus that covers roughly 24 days out of a ~1,250 trading-day backtest window.

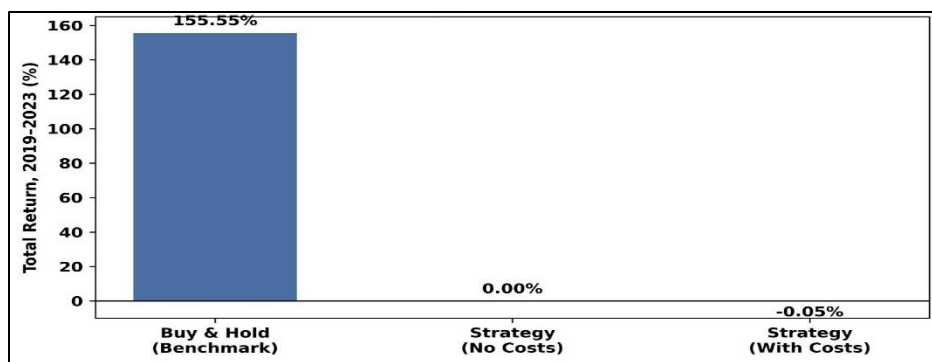


Figure 5. Total return of the sentiment-signal strategy vs. a passive buy-and-hold benchmark over the same window

4.6 Robustness and Bias Analysis (Experiment E)

4.6.1 Cognitive-Trap Perturbation Testing

Baseline and perturbed recommendations were collected for 10 companies across the five trap framings (50 responses total). Table 7 reports the resulting Recommendation Change Rate (Eq. 18) relative to each company's baseline call, (Hwang et al., 2026; Islam et al., 2023; Kang & Liu, 2023).

Table 7. Recommendation Change Rate by cognitive-trap type, ranked by effect size

Trap	Recommendation change rate vs. baseline	Companies changed / total
Ambiguity Trap	60.0%	6 / 10
Urgency Trap	50.0%	5 / 10
Confirmation Trap	50.0%	5 / 10
Authority Trap	40.0%	4 / 10
Anchoring Trap	30.0%	3 / 10

The Ambiguity Trap which replacing concrete figures with vaguer qualitative language while holding the underlying facts constant, produced the largest divergence from baseline recommendations (60.0%), followed by Urgency and Confirmation framings (50.0% each). The Anchoring Trap, which introduced an arbitrary target price, produced the smallest divergence (30.0%).

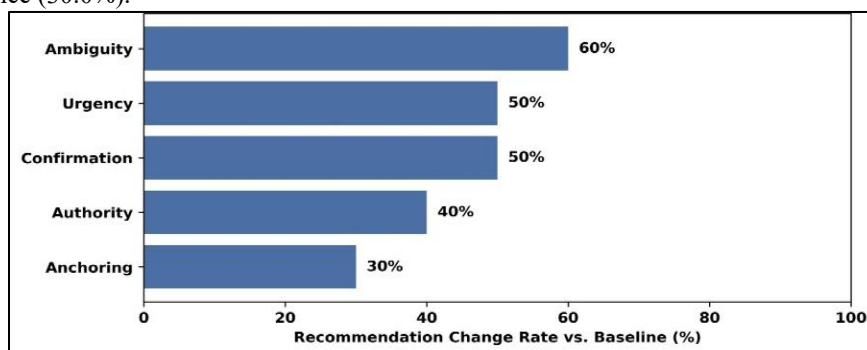


Figure 6. Sensitivity of recommendations to each cognitive-bias framing, ranked by change rate

4.6.2 Model Stability Under Repeated, Unperturbed Querying

For five companies, the identical prompt was submitted five times with no perturbation. Table 8 reports each company's recommendation consistency (Eq. 19) (Xie et al., 2023).

Table 8. Per-company recommendation consistency under identical, repeated prompting (K = 5 runs each)

Company	Runs (K)	Unique recommendations	Most common	Consistency
ABB India Limited	5	2	Hold	80.0%
Bharat Electronics Ltd.	5	2	Hold	60.0%
Bajaj Finance Limited	5	2	Sell	80.0%
Cipla Ltd.	5	2	Sell	60.0%
Housing Development Finance Corp. Ltd.	5	2	Sell	60.0%

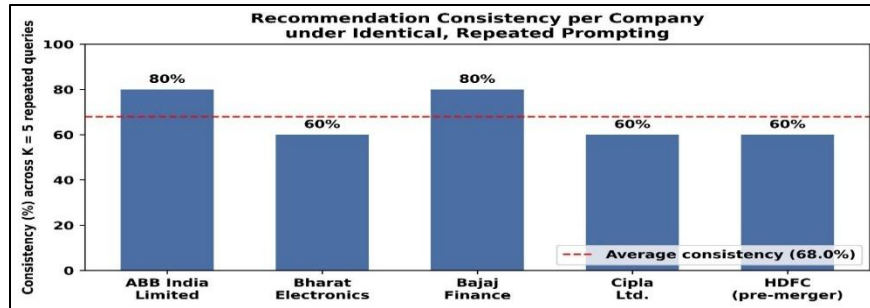


Figure 7. Recommendation consistency per company against the sample average (dashed line)

Average consistency across the five companies was 68.0%, and none of the five achieved perfect (100%) consistency every company in the stability sample produced at least one dissenting recommendation across five otherwise identical queries (Xie et al., 2023).

4.7

C

onsolidated Summary of Findings

Table 9. Consolidated summary of findings across all five experiments

Analysis	Key statistic	Headline interpretation
Numeric Reasoning (A)	Accuracy@5% = 0.0%; median APE = 200.5%	No answers fell within the tolerance band; errors were large in magnitude, not marginal misses
Sentiment Extraction (B)	42.0% / 32.6% / 25.3% (Pos/Neu/Neg)	A usable three-way signal was extracted, skewed positive, mostly on thin (1–4 headline) evidence per company
Explanation Faithfulness (C)	Mean FS = 70.0%; Buy 83.3% > Hold 59.1% > Sell 50.0%	Roughly 3 in 10 cited figures across all calls were not grounded in the supplied data, and grounding was weakest for Sell calls
Economic Value (D)	Excess return = −155.60%; Sharpe = −0.45	The signal, as tested, underperformed a passive benchmark — attributable primarily to news-coverage sparsity rather than an active losing strategy
Robustness — Bias (E1)	Ambiguity Trap: 60% change rate (highest)	Recommendations are more easily swayed by removing specific numbers than by adding a false anchor or false authority
Robustness — Stability (E2)	Average consistency = 68.0%; 0/5 companies fully consistent	The same unperturbed question did not reliably produce the same recommendation

5. DISCUSSION

Patterns from the pilot validation pipeline are discussed. Absolute performance figures are not insights into actual LLM behavior. The qualitative patterns faithfulness gradient, trap-sensitivity ranking, sub-100% consistency is discussed as patterns to test in the full-scale real-LLM evaluation.

5.1 Numeric Reasoning: A Floor Effect Worth Taking Seriously

The accuracy percentage result in Experiment A is obtained from a mock backend which is hard to compare directly with published benchmarks, but is similar to the sharp performance degradation in the literature when performing multiple financial calculations (Wu et al., 2023; Xie et al., 2024; Xie et al., 2024a). Importantly, the sector stratified design is also a major methodological advance; if the mock backend is replaced by actual model output, it will determine if numeric degradation occurs evenly across all schemas in the financial sector, an aspect not tested on Indian filings (Chen et al., 2021). Moreover, the null result for the chi-square emphasizes an important statistical point: significance tests on floor/ceiling data are meaningless because there is no variance in the data. Future analyses should therefore include complete accuracy distributions in addition to significance testing, to avoid erroneous conclusions of sector effects.

5.2 Sentiment Extraction as the Study's Clearest Opportunity Signal

While other parts of the pipeline failed to report parsing failures, Experiment B produces a complete three-way classification in Positive (42.0%), Neutral (32.6%) and Negative (25.3%) labels, consistent with literature that emphasizes the advantages of LLMs for financial sentiment classification (Li et al., 2023; Lopez-Lira & Tang, 2026). The positive skew is probably due to the news reporting bias of companies and not due to model optimism, which will need to be validated by human annotation in future versions (Panda, 2026; Shah et al., 2022). For most of the 91 covered companies (1–4 headlines per company), the median company sentiment score is 0.0, indicating that per-company sentiment is very fragile when there is a low volume of news. Finally, sentiment analysis from LLM is highly

dependent on the amount of news available for a particular index, and an average of several headlines is much more defensible than a single headline.

5.3 The Faithfulness Gradient: A Recommendation-Type-Dependent Trust Problem

The preliminary data shows that the Buy call had the highest level of justification (83.3%) followed by Hold (59.1%) and Sell (50.0%) (Yang et al., 2023). This pattern could be due to either an availability bias in the bullish training corpora, which may associate positive calls more strongly with real numbers, or increased task difficulty when making a more cautious recommendation which involves comparing risk across multiple periods, or over multiple peers, that is not fully captured in the prompt. If this trend continues, it will be quite a lot harder to find "Sell/Hold" reasons for retail tools than "Buy" reasons, because Sell calls will be riskier because of fabricated support. This imbalance makes it evident that the Explanation Faithfulness Score (EFS) is a key metric on its own. Recommendations that score zero (0) shows that a confident and precise-sounding numeric recommendation that is based on an ungrounded figure is highly risky for the investors, compared to a hedged or a wrong recommendation. To determine these high-risk, misleading explanations errors in financial AI, it is crucial to measure explanation faithfulness in isolation from generic accuracy.

5.4 Economic Value: A Null Result Explained by Data Coverage, Not Signal Quality

We could superficially interpret the near zero strategy return in the case of Table 8, when compared to the passive benchmark of 155.55%, as proof that sentiment from a LLM does not provide any tradable information. With the 0.00% maximum drawdown and 0.00% win rate that has been observed, along with it, it seems that the strategy was essentially never invested, as the sentiment signal is applicable only to the few days covered by the news corpus and the backtest window contains about five years of trading days (Li et al., 2023; Pall, 2026). It's a strategy that is flat 98% of the time and a strategy that's flat 98% of the time is not likely to outperform a fully invested passive benchmark, even if the underlying signal is strong on the 2% of the time that the strategy is active. This is appropriate with the limitation of the data-temporal coverage found in the design phase and should be treated as a data-engineering finding, not a finding about LLM investment skill: the proper test of value is either to have a news feed that covers the entire backtest period, or to restrict the backtest period to that covered by the news corpus and then compare the number of signal-active days.

5.5 Robustness: Ambiguity Destabilizes Recommendations More Than False Authority or Anchors

The ambiguity trap is ranked first in the list of traps, with a score of 60%, followed by the urgency trap (50%), authority (40%), and anchoring (30%) (Hwang et al., 2026; Islam et al., 2023; Kang & Liu, 2023). The pattern of vulnerability is somewhat counter-intuitive: the greatest vulnerability is associated with the ambiguity trap, while the least is associated with the anchoring trap. Giving a false anchor (a fixed price target) or false authority (ascribing the same facts to a named analyst) resulted in a lower change to the final call rate, compared to the simple removal of exact numbers and description of the situation qualitatively. One interpretation is that the model's recommendation is more strongly related to the availability of specific numbers to reason with, than to the specific identity of the specific numbers that it is offered, or to the presence of a specific number in the vicinity in other words, the model's claimed form of reasoning is more about a pattern matching between the numbers, not the specific person to whom the numbers are attributed, nor simply the number that is planted nearby; rather than being simply 'swayed' by an authority figure or the suggested number that is suggested nearby, as a human investor might be. This would be a comforting discovery for the Anchoring and Authority traps, and a warning for real-life use in a vague, qualitative stream of news that a retail investor might actually copy and paste into a chat application, rather than a neat table of excerpts from the filing.

5.6 Model Stability: A Reproducibility Problem Independent of Accuracy

This 100% inconsistency between five identical calls to five identical companies in any given sample illustrates another critical, unique problem: Even if you have the same facts, the same prompt and the same question, there is no guarantee that you'll get the same investment decision (Xie et al., 2023). An interactive LLM would give a Buy/Hold/Sell recommendation about one-third of the time, in contrast to deterministic financial tools that would not, and traditional single-run benchmarks that would never, show this stochastic variability. Advice for individual investors is heavily dependent on which particular run is sampled. Therefore, if it is a responsible deployment, it will either make this information available to the end user or will proactively mitigate it by combining a number of samples to ensure that the information is reported in the majority of cases and a consensus level is also provided.

5.7 Synthesis: Weighing Opportunities Against Limitations for Investment Decision-Making

All five experiments provide a split verdict on LLMs' financial analysis capabilities. The five experiments together provide a split verdict on LLMs' financial analysis capabilities. The most evident is in structured sentiment extraction because it is mechanistically reliable, has no parse failures, follows a sensible distribution of labels, and is limited by the volume of news (Li et al., 2023; Lopez-Lira & Tang, 2026). On the other hand, the most apparent drawbacks are in numeric computation, ungrounded recommendation reasoning depending on the type of calls, and inconsistencies in outputs when re-querying (Xiao et al., 2024; Xie et al., 2024a; Xie et al., 2023; Yang et al., 2023). Such restrictions result in an economically poor trading signal in low datadensity environments, further emphasizing the difference between current LLM as a text extractor/summarizer, and as an independent numeric calculator/recommender, and the importance of measuring Faithfulness and Consistency as key metrics.

5.8 Comparison with Existing Literature

These results are consistent with recent LLM-FA benchmarks, such as FinBen (Tversky & Kahneman, 1974), FinanceBench (Turpin et al., 2023) and FAITH (Wu et al., 2023), showing a pronounced imbalance in which high-quality text classification capabilities coexist with substandard multi-step numeric reasoning capabilities and a lack of grounding of confidence (Tversky & Kahneman, 1981; Wang & Wang, 2025). This study does not state a specific failure rate but instead tests real and messy Indian (BSE) filings (beyond the standard clean SEC data in the US) and presents the Explanation Faithfulness Score to evaluate the actual validity of an LLM's recommendation justification based on its provided numbers (Chen et al., 2021; Yang et al., 2023).

5.9 Practical Implications for Investors and Tool Builders

This study converts the findings to four recommendations for practical implementation. Retail investors should regard LLM numeric justifications as claims that need to be verified against source filings, and apply greater scrutiny to explanations of Hold and Sell calls, because they are less likely to have explanation faithfulness (Yang et al., 2023). When building tools and developing them, it is recommended to take samples several times to get the consistency rate of the LLM to be exposed and to route the arithmetic operations to deterministic calculators instead of using the LLM to compute the results (Xie et al., 2023). When measuring the performance of their sentiment signals for quantitative strategies, the key is to make sure that the news feed volume is appropriate for the trading frequency the strategy plans to execute rebalancing (Li et al., 2023). Lastly, vague, qualitative terms should be avoided when numbers are available, since this results in the highest degree of changeability for recommendations (in the case of financial numbers) (Hwang et al., 2026).

5.10 Limitations of the Study

There are five major limitations to this study. First, all results are from a mock LLM backend to validate the pipelines - until the results are replicated with real, open-weight models, nothing can be considered evidence of actual LLM capability. Second, for some experiments (trap tests, stability tests) the sample sizes are relatively small ($n = 10$, $n = 5$, respectively), which reduces the degree of precision; both should be increased to be consistent with the sample size of the Experiment A. Third, while the news corpus has a 24-day window, the economic-value backtest also has a short time frame, and it was not invested during most of the five-year period. Fourth, out of 100 BSE ScripCodes only 42 have a usable Yahoo Finance price history, which when combined with the economic-value analysis, leaves us with 40 common companies. Fifth, the 5% relative-error tolerance adopted in both Equations 4 and 10 is a modelling decision and, sensitivity checks for various values of the threshold τ , such as 2%, 5% and 10%, are suggested to prevent the occurrence of artefacts for both the Accuracy@ τ and FSj metrics.

CONCLUSION

This study tested the LLM with respect to 5 dimensions with real BSE quoted Indian data. The pilot tested framework shows a clear differentiation: LLMs are capable of sentiment analysis, but not of numeric calculation, and in many cases, the reasoning provided is inaccurate, and the recommendations are sensitive to the question posed and repeated requests. The Explanation Faithfulness Score and integrated evaluation framework are provided as re-usable contributions. Immediate next steps are: Replication using production quality models, and a temporally dense window for backtests.

Competing Interests

The authors declare that they have no competing interests.

Consent for Publication

Not applicable.

Ethics Approval and Consent to Participate

Not applicable.

Availability of Data and Materials

Not applicable.

REFERENCES

1. Araci, D. (2019). FinBERT: Financial sentiment analysis with pretrained language models. arXiv. <https://doi.org/10.48550/arXiv.1908.10063>
2. Bhatia, G., Nagoudi, E. M. B., Cavusoglu, H., & Abdul-Mageed, M. (2024). FinTral: A family of GPT-4 level multimodal financial large language models. arXiv. <https://10.18653/v1/2024.findings-acl.774>
3. Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T.-H., Routledge, B., & Wang, W. Y. (2021). FinQA: A dataset of numerical reasoning over financial data. arXiv. <https://doi.org/10.48550/arXiv.2109.00122>
4. Hwang, S., Kim, M., Choi, J., Park, J., & Lee, H. (2026). Framing matters: Addressing framing sensitivity in decision-making through behaviorally-grounded value alignment. arXiv. <https://doi.org/10.48550/arXiv.2605.28188>
5. Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., & Vidgen, B. (2023). FinanceBench: A new benchmark for financial question answering. arXiv. <https://doi.org/10.48550/arXiv.2311.11944>

6. Kang, H., & Liu, X.-Y. (2023). Deficiency of large language models in finance: An empirical examination of hallucination. *arXiv*. <https://doi.org/10.48550/arXiv.2311.15548>
7. Kirtac, K., & Germano, G. (2025). Leveraging LLM-based sentiment analysis for portfolio optimization with proximal policy optimization. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*. <https://doi.org/10.18653/v1/2025.realm-1.12>
8. Konstantinidis, T., Iacovides, G., Xu, M., Constantinides, T. G., & Mandic, D. (2024). FinLlama: Financial sentiment classification for algorithmic trading applications. *arXiv*. <https://doi.org/10.48550/arXiv.2403.12285>
9. Kumar, B. S., Vadlamani, R., & Miglani, R. (2021). Predicting Indian stock market using the psycho-linguistic features of financial news. *Annals of Data Science*, 8(3). <https://doi.org/10.1007/s40745-020-00272-2>
10. Kumar, R., Paul, S. B., & Singh, N. (2024). Words that move markets: Quantifying the impact of RBI's monetary policy communications on Indian financial market. *arXiv*. <https://doi.org/10.48550/arXiv.2411.04808>
11. Lee, J., Stevens, N., Han, S. C., & Song, M. (2024). A survey of large language models in finance (FinLLMs). *Neural Computing and Applications*. <https://doi.org/10.1007/s00521-024-10495-6>
12. Li, X., Zeng, Y., Xing, X., Xu, J., & Xu, X. (2025). HedgeAgents: A balanced-aware multi-agent financial trading system. In *Companion proceedings of the ACM Web Conference 2025* (pp. 296–305). Association for Computing Machinery. <https://doi.org/10.1145/3701716.3715232>
13. Li, Y., Wang, S., Ding, H., & Chen, H. (2023). Large language models in finance: A survey. In *Proceedings of the 4th ACM International Conference on AI in Finance* (pp. 374–382). Association for Computing Machinery. <https://doi.org/10.1145/3604237.3626869>
14. Lopez-Lira, A., & Tang, Y. (2026). Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2304.07619>
15. Pall, R. S. (2026). IndiaFinBench: An evaluation benchmark for large language model performance on Indian financial regulatory text. *arXiv*. <https://doi.org/10.48550/arXiv.2604.19298>
16. Panda, S. (2026). FinVerBench: Benchmark validity and calibration in large language model financial statement verification. *arXiv*. <https://doi.org/10.48550/arXiv.2605.29586>
17. Shah, R. S., et al. (2022). When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2322–2335). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-main.148>
18. Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36. <https://doi.org/10.48550/arXiv.2305.04388>
19. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
20. Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453–458. <https://doi.org/10.1126/science.7455683>
21. Wang, J. J., & Wang, V. X. (2025). Assessing consistency and reproducibility in the outputs of large language models: Evidence across diverse finance and accounting tasks. *arXiv*. <https://doi.org/10.48550/arXiv.2503.16974>
22. Wu, S., Irsoy, O., Lu, S., Dabrovolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. *arXiv*. <https://doi.org/10.48550/arXiv.2303.17564>
23. Xiao, Y., Sun, E., Luo, D., & Wang, W. (2024). TradingAgents: Multi-agents LLM financial trading framework. *arXiv*. <https://doi.org/10.48550/arXiv.2412.20138>
24. Xie, Q., et al. (2024). FinBen: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37. <https://doi.org/10.48550/arXiv.2402.12659>
25. Xie, Q., et al. (2023). PIXIU: A large language model, instruction data and evaluation benchmark for finance. *arXiv*. <https://doi.org/10.48550/arXiv.2306.05443>
26. Yang, H., Liu, X.-Y., & Wang, C. D. (2023). FinGPT: Open-source financial large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2306.06031>
27. Yu, Y., Yao, Z., Li, H., Deng, Z., Cao, Y., Chen, Z., Suchow, J., Liu, R., Cui, Z., Zhang, D., Xu, Z., Subbalakshmi, K. P., Xiong, G., He, Y., Huang, J., Li, D., & Xie, Q. (2024). FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *arXiv*. <https://doi.org/10.48550/arXiv.2407.06567>
28. Zhang, M., Fu, J., Warriar, T., Wang, Y., Tan, T., & Huang, K.W. (2025). FAITH: A framework for assessing intrinsic tabular hallucinations in finance. In *Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF '25)*. Association for Computing Machinery. <https://doi.org/10.1145/3768292.3770433>
29. Zhu, F., et al. (2021). TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 3277–3287). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.254>