# Strat-LLM: Stratified Strategy Alignment for LLM-based Stock Trading with Real-time Multi-Source Signals

1st Wenliang Huang*
*School of Economics*
*Zhejiang University of Technology*
Hangzhou, China
302023065053@zjut.edu.cn

2nd Zengyi Yu
*Faculty of Education*
*East China Normal University*
Shanghai, China
51284118014@stu.ecnu.edu.cn

***Abstract*—Large Language Models (LLMs) are evolving into autonomous trading agents, yet existing benchmarks often overlook the interplay between architectural reasoning and strategy consistency. We propose Strat-LLM, a framework grounded in Stratified Strategy Alignment. Operating in a *live-forward* setting throughout 2025, it integrates heterogeneous data including sequential prices, real-time news, and annual reports to eliminate look-ahead bias. Extensive stress tests on A-share and U.S. markets reveal: (1) reasoning-heavy models achieve peak utility in Free Mode via internal logic, whereas standard models require Strict Mode as a vital risk anchor; (2) alignment utility is regime-dependent, with Free and Guided modes capturing momentum in uptrending markets, while Strict Mode mitigates drawdowns in downtrends; (3) mid-scale models (35B) show optimal fidelity under strict constraints, whereas ultra-large models (122B) suffer an alignment tax under rigid rules but gain a performance premium in Guided Mode; (4) standard LLMs often fall into a high win-rate trap, optimizing for small gains at the expense of total returns, which can only be mitigated through deep reasoning or strict external guardrails. Project details are available at https://Strat-LLM.github.io.**

***Index Terms*—Stratified Strategy Alignment, Real-time Dynamic Backtesting, Multi-Source Heterogeneous Trading Agents**

## I. Introduction

In recent years, large language models (LLMs) have transcended the paradigm of pure text generation, evolving into **autonomous agents** endowed with Multi-Source perception and complex decision-making capabilities. The very recent emergence of local-resident agents like **OpenClaw** confirms that LLMs can now technically function as persistent, autonomous operators rather than mere chatbots. However, this operational capacity merely exposes a critical knowledge gap: we have yet to discern specifically where their decision-making capabilities excel or falter when subjected to varying model parameters, trading horizons, risk attitudes, and distinct strategic execution modes (ranging from strict adherence to full autonomy). Within this paradigm shift, **financial trading**, owing to its direct linkage to economic value and its stringent requirements on decision logic, has emerged as a *litmus test* for evaluating agent competence.

Existing evaluation frameworks generally fall into two categories. The first focuses on static knowledge, exemplified by benchmarks such as **FinBen** [1] and **Pixiu** [2], which assess fundamental financial literacy. The second comprises dynamic environments like **FinTSB** [3], **StockBench** [4], which simulate trading behaviors. Recently, Multi-Source datasets such as **FinMME** [5] and **Time-MMD** [6] have begun to address heterogeneous data integration.

Despite this progress, three critical limitations persist. First is the absence of **strategic executability**. Most benchmarks emphasize final profits while neglecting whether the decision process adheres to predefined risk rules. As Wallace et al. [7] argue, LLMs inherently struggle to prioritize privileged system instructions such as risk limits over conflicting lower-priority inputs like market noise, rendering them operationally dangerous. Second, existing evaluations overlook agents' **vulnerability to behavioral biases**. Standard frameworks fail to audit the structural soundness of an agent's trading behavior. Without rigid external constraints or intrinsic deep reasoning, LLMs frequently fall into a "High Win-Rate Trap", exhibiting a retail-like Disposition Effect where they optimize for frequent small gains but tolerate massive drawdowns during volatile regimes. Third, **data contamination** remains unresolved, as models may simply memorize historical patterns rather than reason dynamically through temporal shifts [8].

To address these challenges, we propose **Strat-LLM**, a **stratified strategy alignment framework** designed for the real financial market of 2025. We define three progressively constrained modes: **Free**, **Guided**, and **Strict** to explore the boundaries between intrinsic intuition and compliance. To audit strategic consistency, we require models to operate under explicitly predefined execution protocols that isolate the effect of rule adherence from autonomous reasoning. Furthermore, we employ a **T+1 rolling-window backtest** using real-time 2025 data to eliminate look-ahead bias. Our extensive evaluation reveals a fundamental **Reasoning Dichotomy** in how models respond to strategic constraints, and identifies an **Alignment Tax** where rigid rule enforcement incurs measurable performance costs that vary systematically with model scale

and market regime.

## II. Related Work

### A. Evolution of Financial Agents & Dynamic Evaluation

The evaluation paradigm for Financial Large Language Models (FinLLMs) has shifted from static text analysis to dynamic trading simulations. Early works [9] established the baseline for predicting stock movements from textual signals. This evolved into static knowledge benchmarks like **FinBen** [1] and **PIXIU** [2], which tested fundamental literacy but lacked temporal complexity.

To capture real-world market dynamics, recent frameworks adopted rolling-window backtesting. **FinTSB** [3] and **StockBench** [4] introduced environments where agents adapt to shifting market regimes. The push for automation is further exemplified by **KRX Bench** [10], which automates financial benchmark creation, and **InvestorBench** [11], which assesses financial decision-making tasks for agents. However, these dynamic evaluations prioritize outcome-oriented metrics (e.g., annualized return) over **strategic executability**, often failing to distinguish whether a model is reasoning financially or merely exploiting probability patterns [8].

### B. Multi-Source Decision-Making

Financial trading is a process of conflict resolution across heterogeneous, multi-source information streams. While recent advances in multi-modal proxy learning and subspace clustering [12], [13] have improved the foundational representation of diverse data structures, applying these paradigms to dynamic finance remains challenging. Datasets like **FinMME** [5] and **Time-MMD** [6] have advanced the integration of chart and text analysis but often treat Multi-Source reasoning as a static task. In contrast, real-world markets present conflicting signals (e.g., positive earnings reports amidst technical breakdowns). **FinTMMBench** [14] benchmarks temporal-aware Multi-Source retrieval-augmented generation (RAG), highlighting the difficulty agents face in synthesizing heterogeneous data streams without succumbing to information overload.

### C. Strategic Alignment and Behavioral Biases

Financial trading demands strict adherence to risk management protocols, a requirement that poses a significant challenge for current autonomous agents. Research shows that LLMs inherently struggle with instruction hierarchy; as Wallace et al. [7] demonstrate, models often fail to prioritize privileged system instructions (e.g., rigid risk limits or strategy rules) when confronted with conflicting lower-priority inputs (e.g., volatile market data and news). In dynamic financial environments, this alignment failure frequently manifests as emergent behavioral biases. Instead of optimizing for mathematical expectancy, unconstrained standard models often default to retail-like heuristics, such as the Disposition Effect prematurely realizing small gains to inflate win rates while failing to execute stop-losses effectively. Strat-LLM addresses this gap by shifting the evaluation paradigm from static profitability to dynamic behavioral auditability. By mandating stratified interaction modes (Free, Guided, Strict), our framework explicitly measures an agent's ability to overcome intrinsic cognitive biases, balance intuitive reasoning with rule adherence, and maintain strategic consistency across shifting market regimes.

## III. METHODOLOGY

We propose **Strat-LLM**, a comprehensive framework designed to audit the strategic alignment of Large Language Models (LLMs) in financial trading. The system simulates a realistic institutional trading environment and consists of three tightly coupled modules: (1) a *Multi-Source Data Integration Module* that processes heterogeneous market signals and static knowledge; (2) a *Stratified Strategy Scaffolding Engine* responsible for dynamically injecting varying levels of strategy constraints; and (3) a *Dynamic Execution & Evaluation Engine* that executes trades via a rolling-window mechanism and computes multi-dimensional risk metrics.

### A. Multi-Source Data Integration

To construct a *contamination-free* evaluation environment, we integrate heterogeneous data sources covering two major markets: A-Shares and U.S. Stocks.

*1) Data Construction & Time Horizons:* The system was deployed in a *live-forward* setting throughout the 2025 experimental window, utilizing robust knowledge fusion paradigms [15] to ensure interpretable state representation across the multi-source streams. Unlike retrospective backtesting, the agent received data streams sequentially in real-time, strictly preserving the chronological flow of information to eliminate look-ahead bias. All decisions were executed based on the information available at the specific timestamp of the 2025 trading sessions. To accommodate different trading calendars and ensure data integrity, we set staggered evaluation intervals:

- **U.S. Market:** Evaluation period from January 1, 2025, to June 30, 2025.
- **A-Share Market:** Evaluation period from June 1, 2025, to September 30, 2025.

Based on this, we designed a dual time-scale to comprehensively examine agent adaptability:

*a) Short-term Tactical Windows:* We selected 3 independent short-term windows within each market's evaluation period, each containing approximately 15 trading days. For U.S. stocks, these windows are distributed from early to mid-2025, covering high-volatility periods post-earnings releases. For A-shares, windows focus on the oscillation and trend transition periods from June to September. This design tests the model's rapid reaction capabilities and tactical flexibility in the face of sudden microstructural changes (e.g., policy announcements, breaking news).

*b) Long-term Strategic Window:* To evaluate capital management and strategic consistency over a long cycle, we set 1 continuous long-term window per market, covering approximately 90 trading days. The U.S. window covers January to May, experiencing a complete semi-annual market cycle. The A-share window covers June to late September. This requires the model to navigate a full market volatility cycle

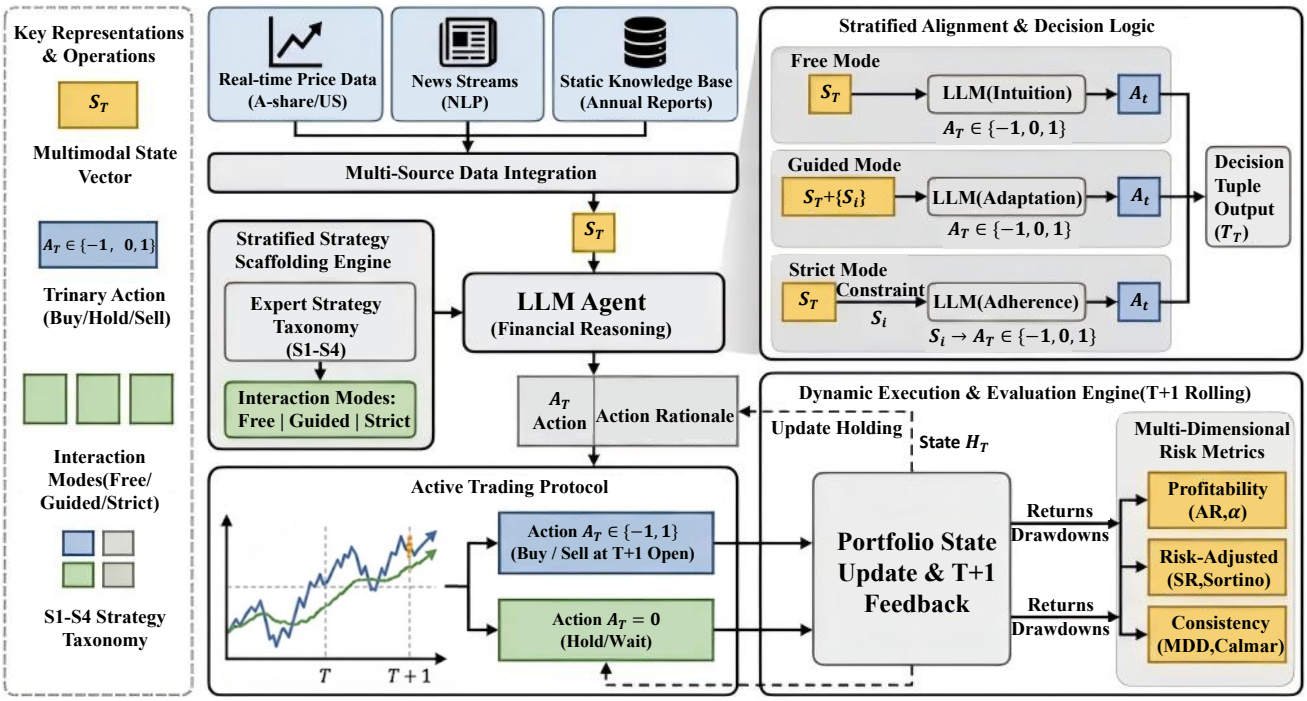


Fig. 1. **The overall architecture of Strat-LLM.** The framework operates in three sequential phases: (1) *Multi-Source Data Integration*, which fuses real-time price data, news streams, and static knowledge into a unified state vector $S_T$; (2) the *Stratified Strategy Scaffolding Engine*, which directs the LLM to generate trading actions $A_T$ under three varying levels of autonomy (Free, Guided, and Strict) based on the Expert Strategy Taxonomy (S1–S4); and (3) the *Dynamic Execution & Evaluation Engine*, which executes orders using a T+1 rolling-window mechanism and computes multi-dimensional risk metrics (e.g., Sharpe Ratio, Max Drawdown) to close the feedback loop.

(oscillation, rally, or correction), verifying its ability to avoid "Strategy Drift" during long-term holding.

*c) Data Hierarchy:* The system integrates three data streams: (1) Minute-level price data; (2) Real-time news streams (sentiment and key events extracted via NLP); (3) Static Expert Knowledge Base (structured annual reports processed via hierarchical summarization, containing only information public as of trading day $T$).

### B. Stratified Strategy Alignment Protocol

To decouple the model's "reasoning capability" from its "instruction following capability," Strat-LLM introduces a core stratified strategy scaffolding mechanism.

*1) Expert Strategy Taxonomy:* We formally define four classic quantitative trading strategies as evaluation baselines:

- **S1 (Short-Term Reversal):** Based on the overreaction hypothesis in behavioral finance, capturing mean reversion opportunities after short-term plunges.
- **S2 (Breakout Momentum):** A trend-following strategy that triggers a buy signal when the price breaks the 3-day high.
- **S3 (Volatility Compression):** A volatility regime-switching strategy identifying accumulation patterns in low-volatility zones.
- **S4 (Price-Volume Confirmation):** Based on price-volume theory, validating the effectiveness of price upward trends.

*2) Interaction Modes:* We implement three autonomy-progressive interaction modes via prompt engineering:

- **Free Mode:** Simulates a zero-shot reasoning scenario, where the model relies on its native financial intuition, but this can lead to unstable decision-making as it has no external strategy constraints to guide it.
- **Guided Mode:** Simulates a human-AI collaboration scenario where strategies S1-S4 are provided as reference, allowing the model to adjust them based on real-time news. However, the model may still face internal cognitive conflicts, leading to behavioral inconsistency when the externally provided strategies don't fully align with its pre-trained intuition.
- **Strict Mode:** Simulates a compliance risk control scenario, where the model must strictly adhere to predefined strategies (S1-S4). While this enhances stability and reduces overtrading, it can also induce "Cognitive Dissonance," causing the model's internal intuition to clash with rigid strategy rules. This cognitive dissonance often manifests as execution paralysis or suboptimal trade timing, allowing us to explicitly audit the "Alignment Tax."

*3) T+1 Rolling Decision & Feedback:* The system employs day-by-day simulation logic:

- **Input:** On trading day $T$, the agent receives the Multi-Source state vector $S_T$.
- **Decision:** The model outputs a structured instruction containing the action signal $Action \in \{-1, 0, 1\}$, rationale.

$Action = 1$ indicates Buy, $Action = -1$ indicates Sell, and $Action = 0$ indicates Hold. In "Strict Mode," the rationale must explicitly cite specific clauses from the strategy library.

- **Execution:**
  - If $Action = 1$ and cash is sufficient, buy at the opening price of $T + 1$ (deducting costs). If cash is insufficient, execute "Cash Truncation" (buy the maximum affordable shares).
  - If $Action = -1$, sell the model-designated volume of shares at the opening price of $T + 1$ (deducting costs).
  - If $Action = 0$, maintain the current position.
- **Update:** After market close, update unrealized PnL and feed the new holding state $H_T$ to the next day's prompt.

*4) Evaluation Metrics:* We employ a multi-dimensional set of metrics to comprehensively evaluate the agent's trading performance. These metrics are categorized into three primary dimensions:

- **Profitability:** Measured by Annualized Return (AR) and Alpha ($\alpha$), which capture the absolute and excess returns generated by the model's trading strategy.
- **Risk-Adjusted Returns:** Evaluated using the Sharpe Ratio and Sortino Ratio, providing insights into the returns earned per unit of total risk and downside risk, respectively.
- **Strategic Consistency:** Assessed via Maximum Drawdown (MDD) and the Calmar Ratio, which quantify the portfolio's resilience and risk-return profile during market downturns.

## IV. EXPERIMENTS

### A. Experimental Settings

*1) Models and Baselines:* To establish a comprehensive evaluation of large language models (LLMs) in financial decision-making, we select a diverse suite of state-of-the-art models, ranging from prominent open-source weights to leading proprietary architectures.

- **Open-Source Frontier Models:** We incorporate an array of highly capable open models, including Qwen3.5-Plus, Qwen3.5-122B-A10B, Qwen3.5-35B-A3B, Qwen3.5-9B, DeepSeek-V3.2, GLM-5, Kimi-k2.5, and Minimax-m2.5. This selection allows us to thoroughly assess the baseline trading acumen of accessible, top-tier general-purpose LLMs.
- **Proprietary State-of-the-Art Models:** To gauge the approximate upper bound of current artificial intelligence capabilities within our benchmark, we evaluate leading closed-source systems such as GPT-5.4, Claude-4.5-Sonnet, and Gemini-3.1-Pro.

To guarantee a fair and rigorous comparison, all evaluated models operate under a unified testing framework, utilizing standardized efficient sampling and generation configurations [16]. This entails identical broker configurations, standardized interaction interfaces, and consistent input modalities across all experimental conditions.

*2) Data and Evaluation Protocol:* All experiments are executed under an **Active Trading Protocol**, allowing models to dynamically issue both *buy* and *sell* orders based on changing market conditions. The primary experimental configurations are summarized as follows.

*a) Markets and Assets.:* Our evaluation spans the **A-share** and **U.S. equity markets**. To enhance generalizability and mitigate idiosyncratic risks, we select representative equities across diverse sectors, including technology, manufacturing, and consumer goods. Performance metrics are then aggregated and averaged across these assets.

*b) Initial Capital Allocation.:* To reflect realistic trading scenarios representative of typical investors, we initialize the simulated trading accounts with a fixed capital threshold. Specifically, the starting capital is set to 1,000,000 in the respective local currency (i.e., $1,000,000 for the U.S. market and ¥1,000,000 RMB for the A-share market). Unlike an infinite-capital assumption, this realistic constraint forces the models to make strategic portfolio allocation and liquidity management decisions when executing their buy and sell signals.

### B. Research Questions

**RQ1 (The Reasoning Dichotomy):** Does an LLM's intrinsic reasoning architecture dictate its reliance on external strategic guardrails versus autonomous intuition?

**RQ2 (Regime Dependence):** Is the utility of strict strategic alignment and risk anchoring fundamentally contingent upon prevailing market trends?

**RQ3 (The Scaling Paradox):** Does strategic execution fidelity scale monotonically with model size, or does an optimal mid-scale parameter region exist?

**RQ4 (Behavioral Biases):** Do standard LLMs inherently manifest retail-like biases, such as the Disposition Effect, and can strict alignment mitigate them?

### C. Main Results

In this section, we provide a comprehensive analysis of the experimental results. By comparing the performance of models across different architectures (Reasoning vs. Standard), strategic constraints, model scales, and temporal horizons, we identify a series of key behavioral patterns that define the efficacy of LLM-based financial agents.

*1) Market-Regime Dependent Utility of Strict Alignment:* Our results demonstrate that the utility of the *Strat-LLM* modes is highly sensitive to market regimes. In the uptrending A-share sample, the flexibility of *Free* and *Guided* modes allows models to capture momentum. However, in the downtrending US stock environment, the *Strict Mode* provides a critical safety net.

As evidenced in Panel A of Table I, the proprietary model **GPT-5.4** drastically reduces its Maximum Drawdown (MDD) from 20.83% (Guided) to 11.66% (Strict), effectively preserving capital in a declining market. This phenomenon is similarly reflected in open-source models, confirming that *Strict Alignment* is not merely a performance-enhancing tool, but a universal, regime-dependent insurance mechanism that mitigates extreme downside risks.

TABLE I

PERFORMANCE COMPARISON OF STRATEGY MODES: PROPRIETARY VS. OPEN-SOURCE MODELS IN A-SHARES AND US STOCKS. TO MAINTAIN FOCUS, THIS TABLE PRESENTS THE MOST REPRESENTATIVE VARIANTS OF EACH MODEL SERIES. THE BEST AND SECOND-BEST RESULTS WITHIN EACH METRIC ARE HIGHLIGHTED IN **BOLD** (WITH ORANGE SHADING) AND *italics* (WITH BLUE SHADING), RESPECTIVELY.

| Model | Strategy | A-shares | | | | | | US Stocks | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | TR(%) | SR | MDD(%) | Vol(%) | WR(%) | $\alpha$(%) | TR(%) | SR | MDD(%) | Vol(%) | WR(%) | $\alpha$(%) |
| **Panel A: Proprietary State-of-the-Art Models** | | | | | | | | | | | | | |
| GPT-5.4 | Strict | 12.31 | 2.19 | *3.17* | 15.95 | 54.55 | 17.88 | **−0.40** | −0.04 | 11.66 | 20.88 | 25.81 | 6.72 |
| GPT-5.4 | Guided | 6.36 | 1.15 | 4.76 | 15.87 | 57.58 | −1.87 | −10.50 | −1.31 | 20.83 | 23.22 | 21.62 | −21.97 |
| Claude-Sonnet-4.5 | Guided | 11.52 | 1.81 | 4.76 | 18.36 | 52.38 | 6.68 | −4.33 | −0.44 | 18.13 | 25.16 | 44.83 | −1.64 |
| Gemini-3.1-Pro | Strict | 0.60 | 0.12 | 7.86 | 25.10 | 46.15 | −32.53 | −8.60 | −0.51 | 29.03 | 38.68 | 33.33 | −6.66 |
| Gemini-3.1-Pro | Free | 6.38 | 0.80 | 5.88 | 25.43 | 41.18 | −17.64 | −14.73 | −0.56 | 29.22 | 37.38 | 36.84 | −8.14 |
| **Panel B: Open-Source Frontier Models** | | | | | | | | | | | | | |
| Qwen3.5-122B-A10B | Guided | **13.68** | 1.62 | 6.93 | 25.06 | 70.00 | 1.84 | −9.19 | −0.82 | 25.01 | 29.95 | 35.48 | −13.63 |
| Qwen3.5-35B-A3B | Strict | *12.33* | 1.36 | 5.84 | 27.61 | 56.25 | −5.74 | −13.67 | −0.48 | 29.24 | 38.17 | 20.00 | −4.86 |
| DeepSeek-Reasoner | Free | 7.62 | 1.57 | **1.98** | 21.14 | 33.33 | 1.77 | −5.35 | −0.35 | 22.65 | 33.63 | 34.29 | 0.28 |
| DeepSeek-Chat | Strict | 7.04 | 1.09 | 5.40 | 19.13 | 40.00 | −5.79 | −3.62 | −0.23 | 19.48 | 31.61 | 31.43 | 4.05 |
| GLM-5_think | Free | 7.82 | 1.23 | 5.33 | 18.64 | 61.76 | −3.27 | −12.01 | −0.35 | 20.76 | 31.77 | 20.83 | 0.52 |
| GLM-5_nothink | Strict | 6.32 | 1.24 | 4.73 | 14.42 | 44.12 | −2.89 | *−0.51* | −0.01 | 16.81 | 24.41 | 50.00 | 8.83 |
| Kimi-K2.5_think | Free | 11.68 | 1.40 | 6.79 | 25.13 | 75.00 | −3.86 | *−0.51* | −0.34 | **1.61** | 42.18 | 33.33 | 0.53 |
| Kimi-K2.5_nothink | Guided | 12.05 | 1.45 | 7.19 | 24.92 | 65.22 | −2.50 | −1.03 | 0.02 | 19.55 | 32.65 | 42.86 | 12.02 |
| Minimax-M2.5 | Guided | 10.55 | 1.70 | 4.67 | 17.54 | 66.67 | 5.23 | −4.61 | −0.22 | *10.28* | 28.74 | 30.77 | 4.36 |

*2) The Reasoning Dichotomy: Internal Logic vs. External Guardrails:* Beyond market regimes, our analysis reveals a fundamental dichotomy in how architectural paradigms interact with the *Strat-LLM* framework. As shown in the ablation study (Table II), the necessity of strategic constraints is inversely proportional to the model's intrinsic reasoning capacity (CoT).

*a) Reasoning Models (Think/Reasoner)::* Models with native reasoning chains (e.g., Kimi-K2.5_think, DeepSeek-Reasoner, GLM-5_think) consistently achieve peak performance in *Free Mode*. For instance, returning to Table I, Kimi-K2.5_think reaches an impressive TR (11.68%) and Win Rate (75.00%) when unconstrained. For these architectures, external strict rules often clash with their internal multi-step deduction, leading to suboptimal trade timing.

*b) Standard Models (Chat/Non-think)::* In contrast, standard conversational models exhibit high vulnerability to market noise when granted full autonomy. Our ablation data (Table II) shows that standard models frequently experience a collapse in performance without rules (e.g., DeepSeek-Chat achieving its optimal 7.04% TR under *Strict Mode*). This confirms that for architectures lacking deep sequential reasoning, *Strict Mode* acts as an essential "Risk Anchor," preventing logical divergence in volatile environments.

*3) The Scaling Paradox: Mid-Scale Sweet Spot in Financial Alignment:* A pivotal finding of our study is the non-linear relationship between model scale and alignment efficacy within the open-source array. While Table I presents the top-performing representatives, Figure 2 illustrates the full ablation across the Qwen series, revealing a distinct "Mid-Scale Sweet Spot" for strategic execution.

TABLE II

ABLATION STUDY: THE IMPACT OF REASONING (THINK) ON STRATEGIC PREFERENCE

| Model Architecture | Strategy | TR(%) | SR | MDD(%) |
|---|---|---|---|---|
| **Kimi-K2.5_think** | **Free** | **11.68** | 1.40 | 6.79 |
| | Guided | 7.71 | 1.00 | 9.53 |
| | Strict | 9.10 | 1.10 | 8.09 |
| **Kimi-K2.5_nothink** | Strict | 10.93 | 1.41 | 7.36 |
| | **Guided** | **12.05** | 1.45 | 7.19 |
| | Free | 9.15 | 1.29 | 5.75 |
| **DeepSeek-Reasoner** | **Free** | **7.62** | 1.57 | 1.98 |
| | Guided | 5.13 | 1.57 | 2.26 |
| | Strict | 3.52 | 0.55 | 6.06 |

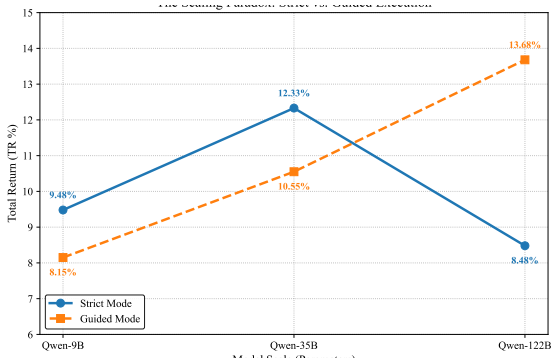


Fig. 2. The Scaling Paradox: Total Return comparison across Qwen variants (9B, 35B, 122B). The mid-scale 35B model exhibits optimal execution fidelity under strict constraints.

Specifically, the **Qwen3.5-35B-A3B** model achieves an exceptional Total Return (TR) of 12.33% under *Strict Mode*, maintaining robust performance in constrained settings. We hypothesize that mid-scale models possess sufficient knowledge to interpret financial signals while exhibiting high fidelity

to instruction-following, avoiding the "reasoning inertia" often found in ultra-large models. Conversely, the ultra-large **Qwen3.5-122B-A10B** excels under *Guided Mode* (13.68% TR), suggesting that larger models require a degree of strategic autonomy to leverage their complex internal heuristics. This indicates that the "Alignment Tax" is highest for ultra-large models when subjected to rigid rule-sets, but transforms into an "Alignment Premium" when shifted to guided frameworks.

*4) The High Win-Rate Trap: Emergent Disposition Effect:* In quantitative trading, a high Win Rate (WR) is frequently misconstrued as a definitive proxy for strategy robustness. However, our empirical analysis reveals a counterintuitive asymmetry between WR and Total Return (TR) / Alpha ($\alpha$), exposing a critical vulnerability in models that lack deep reasoning or strict alignment constraints.

As evidenced in Table I, higher win frequencies do not linearly translate to superior portfolio growth. In the A-share environment, **Minimax-M2.5** (Guided) achieves an inflated WR of 66.67% but yields a TR of 10.55% and an $\alpha$ of 5.23%. Conversely, the proprietary **GPT-5.4** (Strict) secures a dominant TR of 12.31% and an unparalleled $\alpha$ of 17.88%, despite a significantly lower WR of 54.55%. This stark divergence is even more pronounced in the US bear market: **GLM-5_nothink** (Strict) boasts a 50.00% WR yet remains in negative territory (TR $-0.51\%$), whereas **GPT-5.4** (Strict) dictates the best bear-market performance (TR $-0.40\%$, $\alpha$ 6.72%) with a mere 25.81% WR.

Viewed through behavioral finance, standard LLMs inherently exhibit the *Disposition Effect* by prematurely securing small gains while tolerating massive drawdowns. In contrast, advanced reasoning architectures act as cognitive debiasing mechanisms that prioritize mathematical expectancy and risk-reward ratios over inflated win rates. Consequently, overcoming these emergent retail-investor biases in LLM-driven agents necessitates either intrinsic Chain-of-Thought processing or extrinsic *Strict Mode* constraints.

*5) Temporal Horizon Sensitivity:* A consistent performance gap is observed across temporal scales. The **long-term strategic window** (90 trading days) significantly outperforms the *short-term tactical window* (15 trading days). Over short horizons, models struggle with signal ambiguity, resulting in indecisive trading behavior. In contrast, extended horizons provide sufficient latency for the models' entry logic to materialize, supporting the view that LLM-based reasoning is better suited for strategic positioning rather than high-frequency tactical adjustments.

## V. Conclusion

In this study, we introduced *Strat-LLM*, a framework that shifts financial agent evaluation from static profitability to **stratified strategic auditability**. Extensive stress tests reveal that the optimal level of strategic constraint depends heavily on model architecture and market conditions. Specifically, reasoning-heavy models thrive under full autonomy, whereas standard models require rigid alignment to mitigate downside risks. Furthermore, we observe a **mid-scale advantage** at approximately 35B parameters and find that strict rules penalize ultra-large models unless they are granted guided autonomy, challenging conventional scaling laws in financial tasks. We also expose an emergent disposition effect where models misleadingly optimize for high win rates over total returns, a behavioral bias that can only be mitigated through deep reasoning or strict external guardrails. Ultimately, *Strat-LLM* provides a rigorous baseline for trustworthy autonomous trading, bridging the gap between unregulated generation and professional compliance.

## References


[1] Q. Xie, W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, M. Xiao, D. Li, Y. Dai, D. Feng *et al.*, "Finben: A holistic financial benchmark for large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 95 716–95 743, 2024.

[2] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, and J. Huang, "Pixiu: A large language model, instruction data and evaluation benchmark for finance," *arXiv preprint arXiv:2306.05443*, 2023.

[3] Y. Hu, Y. Li, P. Liu, Y. Zhu, N. Li, T. Dai, S.-t. Xia, D. Cheng, and C. Jiang, "Fintsb: A comprehensive and practical benchmark for financial time series forecasting," *arXiv preprint arXiv:2502.18834*, 2025.

[4] Y. Chen, Z. Yao, Y. Liu, J. Ye, J. Yu, L. Hou, and J. Li, "Stockbench: Can llm agents trade stocks profitably in real-world markets?" *arXiv preprint arXiv:2510.02209*, 2025.

[5] J. Luo, Z. Kou, L. Yang, X. Luo, J. Huang, Z. Xiao, J. Peng, C. Liu, J. Ji, X. Liu *et al.*, "Finmme: Benchmark dataset for financial multi-modal reasoning evaluation," *arXiv preprint arXiv:2505.24714*, 2025.

[6] H. Liu, S. Xu, Z. Zhao, L. Kong, H. Prabhakar Kamarthi, A. Sasanur, M. Sharma, J. Cui, Q. Wen, C. Zhang *et al.*, "Time-mmd: Multi-domain multimodal dataset for time series analysis," *Advances in Neural Information Processing Systems*, vol. 37, pp. 77 888–77 933, 2024.

[7] E. Wallace, K. Xiao, R. Leike, L. Weng, J. Heidecke, and A. Beutel, "The instruction hierarchy: Training llms to prioritize privileged instructions," *arXiv preprint arXiv:2404.13208*, 2024.

[8] A. Lopez-Lira, Y. Tang, and M. Zhu, "The memorization problem: Can we trust llms' economic forecasts?" *arXiv preprint arXiv:2504.14765*, 2025.

[9] Y. Xu and S. B. Cohen, "Stock movement prediction from tweets and historical prices," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1970–1979.

[10] G. Son, H. Jeon, C. Hwang, and H. Jung, "Krx bench: Automating financial benchmark creation via large language models," in *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing@ LREC-COLING 2024*, 2024, pp. 10–20.

[11] H. Li, Y. Cao, Y. Yu, S. R. Javaji, Z. Deng, Y. He, Y. Jiang, Z. Zhu, K. Subbalakshmi, J. Huang *et al.*, "Investorbench: A benchmark for financial decision-making tasks with llm-based agent," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 2509–2525.

[12] J. Yao, Q. Qian, and J. Hu, "Multi-modal proxy learning towards personalized visual multiple clustering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 066–14 075.

[13] J. Yao, Q. Qian, and J. Hu, "Customized multiple clustering via multi-modal subspace proxy learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 82 705–82 725, 2024.

[14] F. Zhu, J. Li, L. Pan, W. Wang, F. Feng, C. Wang, H. Luan, and T.-S. Chua, "Fintmmbench: Benchmarking temporal-aware multi-modal rag in finance," *arXiv preprint arXiv:2503.05185*, 2025.

[15] T. Zhou, G. Li, Y. Li, R. Zhang, and S. Ju, "Dual-domain knowledge fusion for interpretable knowledge tracing," *Knowledge-Based Systems*, p. 115715, 2026.

[16] J. Yao, C. Li, and C. Xiao, "Swift sampler: Efficient learning of sampler by 10 parameters," *Advances in Neural Information Processing Systems*, vol. 37, pp. 59 030–59 053, 2024.