Refining LLM Trading by Slowing Decisions: Weekly Rebalancing Improves Risk-Adjusted Performance over Daily Control

Erik Ely and Idea-Explorer

Abstract

Large language model (LLM) trading agents are usually evaluated with frequent daily reallocation, but this setting can induce noise chasing and high turnover. Our main question is whether reducing decision cadence improves out-of-sample portfolio quality. We run a controlled experiment on U.S. equities with real GPT-4.1 API calls, using 15 tickers from 2010–2026 and a fixed test window from 2025-01-02 to 2026-01-30. We hold the model family, feature schema, and constraints constant, and vary only rebalance cadence (daily, weekly, monthly) plus a weekly position-awareness ablation. Weekly LLM control is the strongest LLM variant: Sortino 2.84 versus 2.17 for daily, lower drawdown (-0.1518 vs -0.1432, with monthly best at -0.1347), and much lower turnover (24.20 vs 59.49). Weekly outperforms daily on paired daily returns under Wilcoxon signed-rank testing ($p = 0.0272$), while monthly shows practical gains but no significant daily-return difference ($p = 0.4382$). Robustness checks across 0/5/10 bps transaction costs keep weekly and monthly above daily. A tuned non-LLM weekly momentum baseline remains best overall (Sortino 2.98), so we do not claim absolute dominance. The contribution is a cadence-controlled estimate showing that medium-horizon LLM allocation is more stable and more deployable than day-to-day LLM control in this setting.

1 Introduction

LLM agents can produce coherent portfolio rationales, but many reported gains come from dense daily decisions that are costly to execute and hard to trust out of sample [Chen et al., 2025, Li et al., 2025, Benhenda, 2026]. Our main question is simple: does an LLM trade better when we ask it to act less often?

Why this matters. In practice, turnover and drawdown matter as much as raw return. If lower-frequency LLM control improves risk-adjusted performance, then LLM agents become more realistic for deployment and benchmarking.

What is missing in prior work? Existing finance-agent papers cover memory, multi-agent workflows, and contamination-aware evaluation [Liu and Dang, 2025, Zhang et al., 2023, Xiao et al., 2024, Huang et al., 2024]. However, controlled studies that isolate decision cadence while fixing model, data, and prompt family are still limited. This leaves an attribution gap: are observed gains from better reasoning, or from a better action horizon?

Our approach. We run a cadence-controlled backtest with real GPT-4.1 calls on daily U.S. equities data (15 tickers, 2010–2026). We compare daily, weekly, and monthly LLM rebalancing under the same lagged technical features, long-only constraints, and 35% per-asset cap. We also test position-aware versus memoryless prompting at weekly cadence.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

What do we find? Weekly LLM control improves Sortino from 2.165 (daily) to 2.835 and reduces turnover from 59.49 to 24.20 at 5 bps cost. Monthly also beats daily on risk-adjusted metrics with even lower turnover (12.03). Weekly versus daily is significant under Wilcoxon testing ($p = 0.0272$). A weekly momentum rule still ranks first overall, which clarifies the current boundary of LLM value.

Our contributions are:

- We isolate decision cadence as the independent variable in an apples-to-apples LLM trading experiment.
- We show that weekly and monthly LLM control improve practical risk-adjusted outcomes and trading stability versus daily LLM control.
- We provide a position-awareness ablation showing similar returns but lower turnover and better drawdown than memoryless weekly prompting.
- We release a reproducible evaluation artifact with metrics, statistical tests, and cost-sensitivity outputs.

The rest of the paper is organized as follows. section 2 summarizes prior work, section 3 describes the setup, section 4 presents quantitative findings, and section 5 discusses implications and limits.

2 Related Work

LLM trading benchmarks and long-run realism. Recent benchmark papers show that short-window gains often weaken under realistic horizons. StockBench emphasizes contamination-aware, multi-month evaluation and reports that many LLM agents do not reliably beat passive alternatives [Chen et al., 2025]. FINSABER extends this concern over longer horizons and broader universes [Li et al., 2025]. Look-Ahead-Bench further highlights temporal leakage as a major threat to validity [Benhenda, 2026]. We follow this line by using lagged features and an out-of-sample recent window.

Position and memory in financial agents. FinPos argues that position-aware formulations improve stability and risk-adjusted behavior [Liu and Dang, 2025]. FinMem introduces memory-centric LLM finance agents [Zhang et al., 2023], and TradingAgents/FinAgent develop multi-role systems for richer decision pipelines [Xiao et al., 2024, Huang et al., 2024]. Our study is complementary: we do not introduce a new architecture, but instead isolate cadence and test whether position-awareness changes outcomes at fixed weekly cadence.

Baseline framing. Prior work consistently recommends comparing LLM agents against transparent quantitative baselines such as momentum and volatility-based allocators [Li et al., 2025, Chen et al., 2025]. We adopt equal-weight, weekly momentum top-k, and inverse-volatility parity baselines to contextualize LLM performance.

Our position. Unlike prior papers that vary many knobs at once, we vary one primary control variable: action frequency. This design directly addresses whether medium-horizon control is a better operating point for LLM portfolio allocation.

3 Methodology

Task definition. At each rebalance date t, the agent receives lagged per-ticker features and outputs long-only portfolio weights $\boldsymbol{w}_t \in \mathbb{R}^{15}$ with $\sum_i w_{t,i} = 1$ and $0 \leq w_{t,i} \leq 0.35$. The portfolio is held until the next rebalance. Daily strategy return is net of transaction costs applied on rebalance days.

Data. The primary dataset is YAHOO-OHLCV-15: Yahoo Finance daily OHLCV from 2010-01-04 to 2026-01-30 (60,660 rows, 15 tickers). We build lagged technical features per ticker: 1/5/21-day returns, 21-day volatility, 63-day momentum, and 63-day drawdown proxy. All features are shifted by one day to prevent look-ahead leakage. The fixed out-of-sample evaluation window is 2025-01-02 to 2026-01-30 (270 trading days). We also retain TWITTER-FINSENT (9,543 train / 2,388 validation) as auxiliary future context but do not feed it into the current allocator.

Compared methods. We compare five LLM policies and three non-LLM baselines:

- LLM daily position-aware (LLM-DAILY).

Method	CumReturn	AnnRet	Sharpe	Sortino	MaxDD	Turnover
MOMENTUM-WEEKLY	**0.4527**	**0.4169**	**2.2372**	**2.9801**	-0.1613	16.9333
LLM-WEEKLY-POSAWARE	0.4261	0.3928	2.0966	2.8352	-0.1518	24.2000
LLM-WEEKLY-MEMORYLESS	0.4274	0.3940	2.0820	2.7305	-0.1660	37.2333
LLM-MONTHLY	0.3804	0.3511	1.8274	2.4603	**-0.1347**	12.0333
LLM-DAILY	0.3091	0.2858	1.6095	2.1650	-0.1432	59.4933
EQUAL-WEIGHT	0.2841	0.2629	1.4577	1.8574	-0.1897	**0.0000**
INVVOL-WEEKLY	0.2260	0.2094	1.3125	1.7189	-0.1752	6.3687

Table 1: Main results on 2025-01-02 to 2026-01-30 with 5 bps transaction costs. Higher is better except MaxDD where values closer to zero are better. Best values are in bold.

- LLM weekly position-aware (LLM-WEEKLY-POSAWARE).
- LLM monthly position-aware (LLM-MONTHLY).
- LLM weekly memoryless (LLM-WEEKLY-MEMORYLESS).
- Equal-weight buy-and-hold (EQUAL-WEIGHT).
- Weekly momentum top-k (MOMENTUM-WEEKLY, $k = 5$, 63-day momentum).
- Weekly inverse-volatility parity (INVVOL-WEEKLY, 21-day volatility).

LLM configuration. We use real GPT-4.1 API calls (OpenAI SDK 2.21.0), temperature 0.0, JSON weight output, and caching for deterministic reruns. Main run usage is 752,721 tokens across about 391 rebalance decisions.

Backtest protocol. We apply a 5 bps proportional transaction cost as default and run sensitivity checks at 0 and 10 bps. Rebalance cadence is the independent variable; model family, feature schema, constraints, and test dates are fixed.

Metrics and statistics. We report cumulative return, annualized return, annualized volatility, Sharpe, Sortino, maximum drawdown, Calmar, and turnover. For inference, we use paired Wilcoxon signed-rank tests on aligned daily returns, Cliff's delta as effect size, and circular block bootstrap (1,000 samples, block size 5) for Sortino differences.

Implementation details. Experiments run in Python 3.12.2 with pandas 2.3.1, numpy 2.3.5, scipy 1.17.0, matplotlib 3.10.8, and seaborn 0.13.2. Missingness in engineered features comes only from rolling-window warmup (max 1.583% for 63-day momentum), with no raw OHLCV missing values and no duplicates.

4 Results

Main comparison. Table 1 reports the primary 5 bps results. Among LLM variants, LLM-WEEKLY-POSAWARE is best on risk-adjusted return (Sortino 2.8352, Sharpe 2.0966), while LLM-MONTHLY has the lowest drawdown (-0.1347) and lower turnover than weekly. Daily LLM is consistently weaker and has the highest turnover (59.4933).

Cadence effect within LLMs. Weekly vs daily improves Sortino by +0.6702 and cuts turnover by 35.2933. Monthly vs daily improves Sortino by +0.2953 and cuts turnover by 47.4600. These results support the claim that slower control reduces churn and improves practical risk-adjusted behavior.

Statistical tests. Weekly position-aware versus daily position-aware yields a significant Wilcoxon signed-rank result ($p = 0.0272$) on paired daily returns, with mean daily difference $+3.18 \times 10^{-4}$. The bootstrap 95% CI for Sortino difference is wide ([-0.355, 2.073]), indicating directional improvement with moderate uncertainty. Monthly versus daily is not significant ($p = 0.4382$). Weekly position-aware versus weekly memoryless is also not significant on daily returns ($p = 0.5787$), suggesting the main gain comes from cadence rather than per-day return distribution shifts.

Ablation on position-awareness. LLM-WEEKLY-POSAWARE and LLM-WEEKLY-MEMORYLESS reach similar cumulative return (0.4261 vs 0.4274), but position-awareness lowers drawdown (-0.1518 vs -0.1660) and turnover (24.20 vs 37.23). This supports position continuity as a stability mechanism.

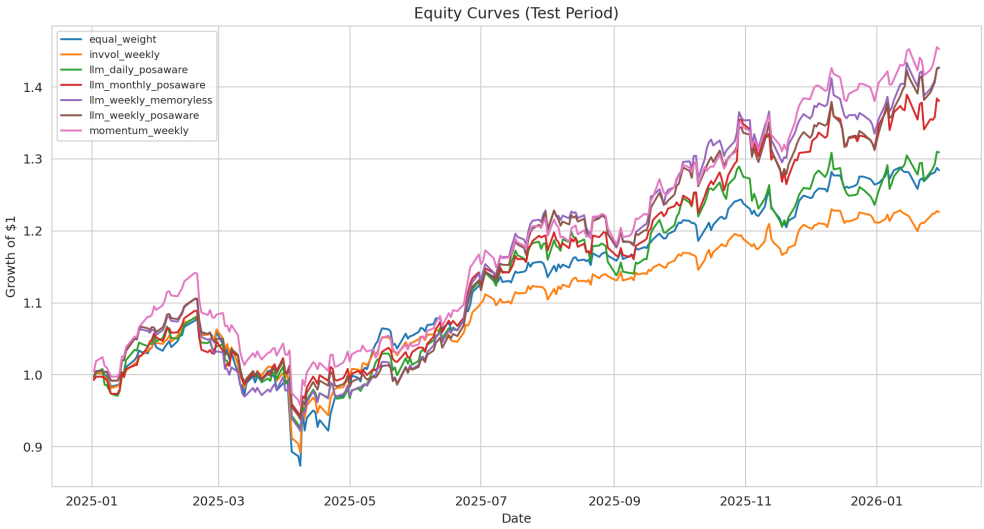

Figure 1: Equity curves for all methods on the 2025-01-02 to 2026-01-30 test period. Weekly and monthly LLM policies stay above daily LLM for most of the window, while weekly momentum remains highest overall.

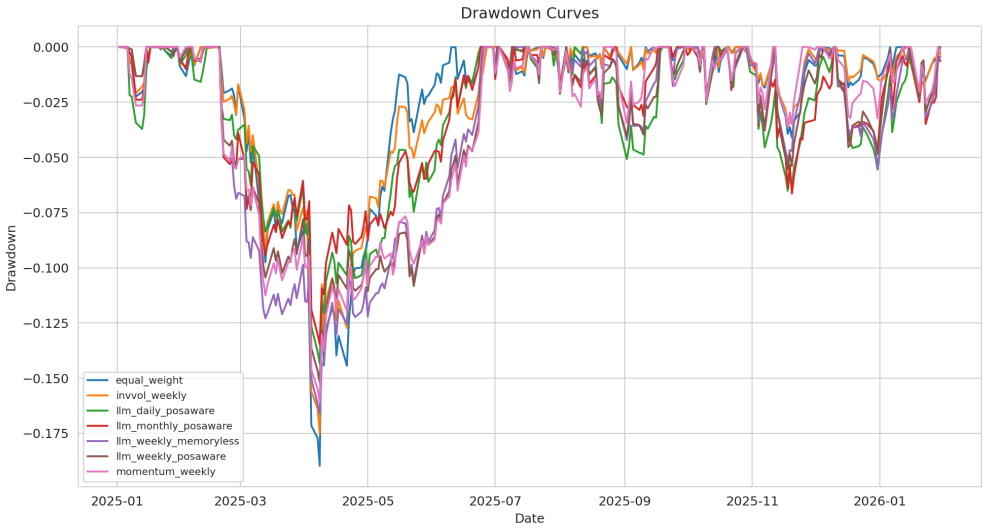

Figure 2: Drawdown trajectories. Monthly LLM has the shallowest trough among LLM variants, and weekly position-aware avoids some deeper drops seen in weekly memoryless.

Comparison to non-LLM baselines. MOMENTUM-WEEKLY remains best overall (Sortino 2.9801, Sharpe 2.2372), outperforming all LLM variants in this window. Still, weekly and monthly LLM beat EQUAL-WEIGHT and INVVOL-WEEKLY on major risk-adjusted metrics.

Transaction-cost robustness. Under 0, 5, and 10 bps settings, weekly and monthly LLM policies remain above daily LLM in key risk-adjusted metrics. This pattern suggests the cadence advantage is not an artifact of one cost assumption.

5 Discussion

Interpretation. The main signal is consistent with a noise-chasing account: daily control offers more chances to react, but in this setting it mostly increases churn. Weekly and monthly cadences

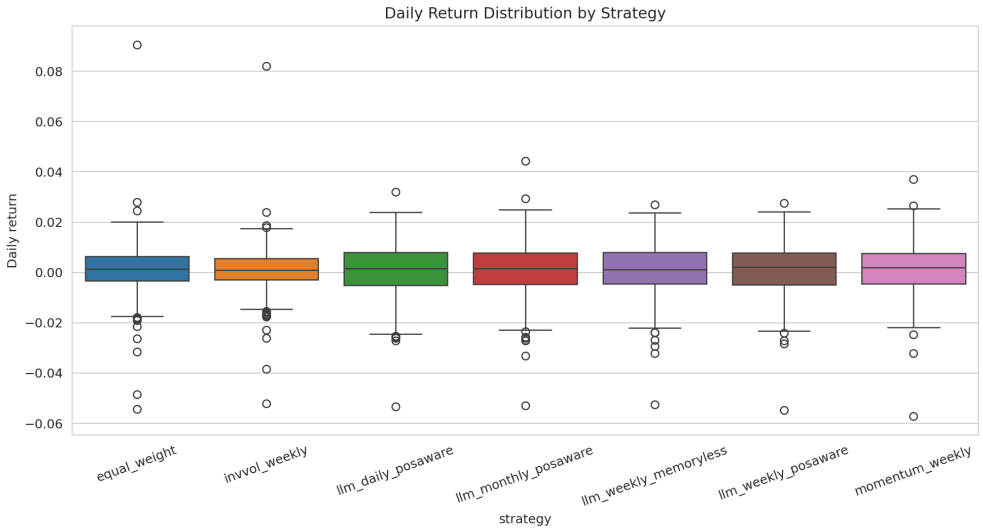

Figure 3: Distribution of daily returns. Distribution overlap explains small effect-size estimates even when cumulative and turnover outcomes diverge materially.

appear to regularize behavior by forcing the agent to commit longer, improving net performance after costs.

Why position-awareness helps. Weekly memoryless and position-aware variants have similar cumulative return, but memoryless trading turns over much more. This suggests that adding explicit portfolio-state context mainly improves execution stability rather than directional alpha.

Limits of current LLM trading. A transparent weekly momentum rule still outperforms all LLM variants in our test window. This matters: better prompting cadence improves LLM behavior, but does not yet replace tuned quantitative baselines for absolute performance.

Limitations. We study one model family (GPT-4.1), one equity universe (15 U.S. tickers), and one recent out-of-sample window. We model frictions with proportional costs only, without detailed slippage or market impact. Bootstrap intervals for Sortino differences are wide, so effect magnitude needs more data.

Broader implications. For benchmark design, cadence should be a standard ablation axis in financial LLM papers. For deployment, medium-horizon LLM control is more credible than day-to-day reallocation because it lowers turnover and operational burden.

6 Conclusion

We presented a controlled cadence study for LLM portfolio allocation using real GPT-4.1 calls on U.S. equity daily data. By fixing model, features, constraints, and evaluation dates, we isolated action frequency as the main variable.

Weekly and monthly LLM policies outperform daily LLM on risk-adjusted metrics and turnover, with weekly position-aware giving the strongest LLM result (Sortino 2.835 vs 2.165 daily). Weekly-vs-daily daily-return differences are significant under Wilcoxon testing ($p = 0.0272$), while monthly improvements are practical but not statistically significant in this window.

The central takeaway is that LLM agents in this setup work better as medium-horizon allocators than as daily controllers. Future work should test cross-model generalization (e.g., GPT-5 and Claude-class models), add point-in-time news/fundamental context, and expand to rolling multi-regime evaluations across longer histories and additional asset classes.

References

Mostapha Benhenda. Look-ahead-bench: Measuring temporal leakage and robustness in llm financial evaluation. *arXiv preprint arXiv:2601.13770*, 2026. URL https://arxiv.org/abs/2601.13770.

Yanxu Chen et al. Stockbench: A contamination-aware benchmark for realistic multi-month llm trading evaluation. *arXiv preprint arXiv:2510.02209*, 2025. URL https://arxiv.org/abs/2510.02209.

Xuan Huang et al. Finagent: A multimodal foundation agent for financial trading. *arXiv preprint arXiv:2402.18485*, 2024. URL https://arxiv.org/abs/2402.18485.

Weixian Waylon Li et al. Can llm-based financial investing strategies outperform the market in long run? *arXiv preprint arXiv:2505.07078*, 2025. URL https://arxiv.org/abs/2505.07078.

Bijia Liu and Ronghao Dang. Finpos: Position-aware financial trading with large language models. *arXiv preprint arXiv:2510.27251*, 2025. URL https://arxiv.org/abs/2510.27251.

Kai Xiao et al. Tradingagents: Multi-agents llm financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024. URL https://arxiv.org/abs/2412.20138.

Yufeng Zhang et al. Finmem: A performance-enhanced llm trading agent with layered memory and character design. *arXiv preprint arXiv:2311.13743*, 2023. URL https://arxiv.org/abs/2311.13743.