TIME SERIES FOUNDATION MODELS IMPROVE LLM DECISIONS: A CASE STUDY IN STOCK TRADING

Shifeng Xie*
Université Paris Cité, France

Ziwei Li*
King Abdullah University of Science and Technology, Saudi Arabia

Themis Palpanas
Université Paris Cité, France

Peilin Zhao
Shanghai Jiao Tong University, China

Chenghao Liu
Datadog

ABSTRACT

Time series foundation models provide strong numerical forecasting capabilities, while large language models are strong at high-level reasoning, suggesting a complementary synergy for decision-making tasks. We study this integration in a stock trading setting showing that TSFM signals statistically significant improvements in trading performance over LLM-only baselines.

Track: Industry & Applications

1 INTRODUCTION

Time series foundation models (TSFMs) have recently attracted growing attention due to their strong cross task generalization capabilities across forecasting, classification, and anomaly detection(Rasul et al., 2024; Garza et al., 2024; Feofanov et al., 2025; Lan et al., 2025). While substantial effort has been devoted to designing new TSFM architectures, their practical integration into downstream systems remains underexplored(Gnassounou et al., 2025; Liang et al., 2024), particularly in combination with large language models (LLMs)(Xie et al., 2025). Recent years have witnessed a surge of LLM-driven trading agents, with multiple representative systems(Zhang et al., 2024a; Yu et al., 2024; Chen et al., 2025). Despite their impressive reasoning abilities, LLMs are inherently suboptimal at processing numerical signals(Mirzadeh et al., 2025; Yang et al., 2025; Singh & Strouse, 2024). This limitation contrasts with TSFMs, which are explicitly optimized for numerical modeling and temporal prediction. The complementary strengths of the two foundation models suggest a synergy: TSFMs provide accurate numerical forecasts, while LLMs leverage these signals for high-level reasoning and decision making. Motivated by this observation, we hypothesize that integrating TSFMs can systematically improve LLM-based decisions across time series driven applications.

Stock trading combines two capabilities: forecasting time series signals and making sequential, high-level decisions. As such, it provides an ideal testbed for studying whether TSFMs can improve LLM decision making. We conduct a systematic case study on the U.S. "Magnificent Seven" stocks, using a strong open-source LLM(Team, 2025a), Qwen-30B(Team, 2025b), as the decision-making agent, and augmenting it with predictions from multiple state-of-the-art TSFMs, including Chronos-2(Ansari et al., 2024), TimesFM-2.5(Das et al., 2024), Moirai-2(Woo et al., 2024; Liu et al., 2026), and Toto(Cohen et al., 2024). The LLM consumes TSFM outputs and produces trading actions based on its reasoning process. Our empirical results reveal two key findings:

- Incorporating TSFM predictions leads to statistically significant improvements in the trading performance of LLM-based agents compared to using the LLM alone.
- The performance gains are robust to the specific format in which TSFM predictions are presented, suggesting that the benefit of TSFM integration is not sensitive to prompt-level formatting choices.

*Equal contribution.

2 EXPERIMENTAL SETUP AND RESULTS

Setup. We evaluate our hypothesis in a realistic portfolio allocation setting over the U.S. "Magnificent Seven" equities (AAPL, GOOGL, AMZN, MSFT, META, TSLA, and NVDA). Each experiment simulates a 30-day trading horizon, and we repeat the evaluation across three non-overlapping time periods to reduce temporal bias. At every trading day, the agent observes historical price trajectories and company fundamentals collected from Alpha Vantage(Xiao et al., 2025; Zhang et al., 2024b), and produces portfolio weights over the seven stocks and a cash asset. For the baseline, the LLM receives only these raw signals (price history and fundamentals). For the TSFM-enhanced variants, we additionally provide predictions generated by TSFM in several output formats.

To ensure robustness and avoid improvements arising from prompt engineering, we design six distinct formats for presenting TSFM predictions to the LLM. These formats vary along three dimensions: absolute prices versus relative returns, point forecasts versus distributional (quantile) forecasts, and single-horizon versus multi-horizon summaries. Concretely, the TSFM outputs include: (1) raw 30-day price trajectories, (2) 30-day return ratios relative to the latest close, (3) aggregated returns at multiple horizons (1 day to 4 weeks), (4) quantile price forecasts over 30 days, (5) quantile return forecasts over 30 days, and (6) quantile returns summarized across multiple horizons. More detailes are shown in sections A and B. All formats convey the same underlying predictive information but differ in representation and granularity. By testing multiple formats under identical trading and evaluation protocols, we can assess the sensitivity of LLM decisions to the presentation of time series information.

We implement a backtesting simulator with fractional-share trading and daily rebalancing. The simulator executes the LLM's weight decisions, tracks portfolio value over time, and records all trades and daily states. To isolate the effect of decision quality, the simulator assumes idealized execution: we do not model market impact, slippage, or transaction costs, and trades are filled at observed daily closing prices. Performance is evaluated using standard financial metrics, including cumulative return, annualized return, Sharpe ratio, Sortino ratio, and maximum drawdown, enabling a risk-adjusted comparison between LLM-only and TSFM-enhanced strategies.

Results We first verify the consistency of the LLM-only baseline across different TSFM backends. As shown in tables 1 and 2, the baseline returns within each time period exhibit low variance across repeated runs ($n = 4$ per period), with small standard deviations and largely overlapping confidence intervals. This confirms that the trading performance of the baseline is stable, providing a reliable reference for evaluating TSFM-enhanced improvements.

We next evaluate whether incorporating TSFM forecasts improves LLM decisions when ignoring the specific input format. As shown in tables 1 and 3, aggregating all 72 TSFM-enhanced runs, we compute the paired improvement relative to the period-wise baseline mean. The average gain is $+0.47\%$ cumulative return, with a 95% confidence interval of $[0.21\%, 0.74\%]$, which remains strictly positive. A one-sample t-test rejects the null hypothesis of no improvement ($t = 3.57$, one-sided $p < 3.3 \times 10^{-4}$), with a moderate effect size (Cohen's $d_z = 0.42$). These results demonstrate that TSFM forecasts consistently yield statistically significant and practically meaningful improvements in LLM-based trading performance.

Finally, we examine whether the benefit depends on the specific format used to present TSFM predictions. As shown in tables 1 and 4, a one-way ANOVA across the six formats shows no significant difference ($F = 0.85$, $p = 0.52$), indicating that performance gains are largely invariant to representation choices. Although certain formats achieve higher point estimates (e.g., multi-horizon summaries), these differences are not statistically significant after multiple comparison correction. This robustness suggests that the improvements primarily from the numerical predictive content of TSFMs rather than from particular prompt structures or formatting strategies. Empirically, Format 3 (aggregated return ratios across multiple horizons) achieves the highest point estimate among the six formats, see Appendix section D for a more detailed analysis.

3 CONCLUSION

In this work, we demonstrate that, even under a simplified stock trading setting, integrating time-series foundation models leads to statistically significant improvements in LLM-based decision

making. Our results show that TSFMs provide complementary numerical signals that consistently enhance trading performance across models and input formats. At the same time, our study intentionally abstracts away several components commonly present in real-world trading systems, such as multi-agent coordination, news and media sentiment analysis, and explicit risk management modules. Incorporating these elements represents an important direction for future work and may further amplify the benefits of TSFM–LLM integration in more realistic decision-making environments.

ACKNOWLEDGEMENTS

Supported by EU project DataGEMS (101188416), and by $Y\Pi AI\Theta A$ & NextGenerationEU project HARSH ($Y\Pi 3TA - 0560901$) that is carried out within the framework of the National Recovery and Resilience Plan "Greece 2.0" with funding from the European Union – NextGenerationEU. This work was granted access to the HPC resources of IDRIS under the allocation 2025-A0191012641 made by GENCI.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Syndar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Yanxu Chen, Zijun Yao, Yantao Liu, Jin Ye, Jianing Yu, Lei Hou, and Juanzi Li. Stockbench: Can llm agents trade stocks profitably in real-world markets?, 2025. URL https://arxiv.org/abs/2510.02209.

Ben Cohen, Emaad Khwaja, Kan Wang, Charles Masson, Elise Ramé, Youssef Doubli, and Othmane Abou-Amal. Toto: Time series optimized transformer for observability, 2024.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024.

Vasilii Feofanov, Songkang Wen, Marius Alonso, Romain Ilbert, Hongbo Guo, Malik Tiomoko, Lujia Pan, Jianfeng Zhang, and Ievgen Redko. Mantis: Lightweight calibrated foundation model for user-friendly time series classification, 2025.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. Timegpt-1, 2024.

Théo Gnassounou, Yessin Moakher, Shifeng Xie, Vasilii Feofanov, and Ievgen Redko. Leveraging generic time series foundation models for eeg classification, 2025. URL https://arxiv.org/abs/2510.27522.

Tian Lan, Hao Duong Le, Jinbo Li, Wenjun He, Meng Wang, Chenghao Liu, and Chen Zhang. Towards foundation models for zero-shot time series anomaly detection: Leveraging synthetic data and relative context discrepancy, 2025. URL https://arxiv.org/abs/2509.21190.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 6555–6565. ACM, August 2024. doi: 10.1145/3637528.3671451.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting, 2026. URL https://arxiv.org/abs/2511.11698.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2025. URL https://arxiv.org/abs/2410.05229.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. Lag-llama: Towards foundation models for probabilistic time series forecasting, 2024.

Aaditya K. Singh and DJ Strouse. Tokenization counts: the impact of tokenization on arithmetic in frontier llms, 2024. URL https://arxiv.org/abs/2402.14903.

AlphaArena Team. Alpha arena: Empower ai models to trade and compete in real markets. GitHub repository, 2025a. URL https://github.com/hoangngaunhatthegioi/alpha-arena. Accessed: 2026-02.

Qwen Team. Qwen3 technical report, 2025b. URL https://arxiv.org/abs/2505.09388.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents llm financial trading framework, 2025. URL https://arxiv.org/abs/2412.20138.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *Proceedings of the VLDB Endowment*, 18(8):2385–2398, April 2025. ISSN 2150-8097. doi: 10.14778/3742728.3742735. URL http://dx.doi.org/10.14778/3742728.3742735.

Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. Number cookbook: Number understanding of language models and how to improve it, 2025. URL https://arxiv.org/abs/2411.03766.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Rong Liu, Zhenyu Cui, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making, 2024. URL https://arxiv.org/abs/2407.06567.

Chong Zhang, Xinyi Liu, Zhongmou Zhang, Mingyu Jin, Lingyao Li, Zhenting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, Sujian Li, Mengnan Du, and Yongfeng Zhang. When ai meets finance (stockagent): Large language model-based stock trading in simulated real-world environments, 2024a. URL https://arxiv.org/abs/2407.18957.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist, 2024b. URL https://arxiv.org/abs/2402.18485.

A Format Definitions

Fix a ticker and a trading (decision) date t. Let p_t denote the observed closing price at date t (the last available close in the prompt context). A TSFM produces a K-step-ahead forecast over the next $K = 30$ trading days:

$$\hat{p}^{(q)}_{t+k} \quad \text{for} \quad k \in \{1, \ldots, 30\},\ q \in \mathcal{Q},$$

where q is a quantile level and $\mathcal{Q}$ is the set of quantiles supported by the TSFM interface. We define the corresponding return ratio (relative return) as

$$\hat{r}^{(q)}_{t+k} \ = \ \frac{\hat{p}^{(q)}_{t+k} - p_t}{p_t}.$$

All six formats are deterministic transformations of the same TSFM forecast distribution and the same last close p_t. They therefore convey the same underlying predictive signal, but differ in (i)

whether they expose absolute price vs. relative return, (ii) whether they expose uncertainty (quantiles), and (iii) whether they compress the horizon via multi-horizon summaries.

Format 1: Raw 30-day price trajectory (point / median)

What it encodes. A point forecast of future prices, typically the median trajectory $\{\hat{p}_{t+k}^{(0.5)}\}_{k=1}^{30}$.
How it is presented. To keep prompts compact, we display only the first few and last few forecasted days (e.g., Day 1–5 and Day 26–30).
Prompt snippet.

```
TSFM Forecast for <TICKER> (30-day price prediction):
Day 1-5:  [<\hat{p}_{t+1}>, <\hat{p}_{t+2}>, <\hat{p}_{t+3}>,
↪  <\hat{p}_{t+4}>, <\hat{p}_{t+5}>]
Day 26-30:[<\hat{p}_{t+26}>, <\hat{p}_{t+27}>, <\hat{p}_{t+28}>,
↪  <\hat{p}_{t+29}>, <\hat{p}_{t+30}>]
```

Format 2: 30-day return ratios relative to the latest close (point / median)

What it encodes. A point forecast expressed in relative terms: $\{\hat{r}_{t+k}^{(0.5)}\}_{k=1}^{30}$, which improves scale-invariance across tickers.
How it is presented. As in Format 1, we typically show Day 1–5 and Day 26–30 to limit length.
Prompt snippet.

```
TSFM Forecast for <TICKER> (30-day return ratios, relative to last
↪  close):
Day 1-5:  [<\hat{r}_{t+1}>, <\hat{r}_{t+2}>, <\hat{r}_{t+3}>,
↪  <\hat{r}_{t+4}>, <\hat{r}_{t+5}>]
Day 26-30:[<\hat{r}_{t+26}>, <\hat{r}_{t+27}>, <\hat{r}_{t+28}>,
↪  <\hat{r}_{t+29}>, <\hat{r}_{t+30}>]
```

Format 3: Aggregated return ratios at multiple horizons (point / median)

What it encodes. A compressed summary of returns at a small set of horizons:

$$\hat{r}_{t+h}^{(0.5)} \quad \text{for} \quad h \in \{1, 5, 10, 15, 20\},$$

corresponding to approximately 1 day, 1 week, 2 weeks, 3 weeks, and 4 weeks (in trading days).
Why this is useful. It preserves short- vs. medium-horizon directional information while being much shorter than a full 30-day vector.
Prompt snippet.

```
TSFM Forecast for <TICKER> (multi-horizon returns):
1 Day:   <\hat{r}_{t+1}>
1 Week:  <\hat{r}_{t+5}>
2 Weeks: <\hat{r}_{t+10}>
3 Weeks: <\hat{r}_{t+15}>
4 Weeks: <\hat{r}_{t+20}>
```

Format 4: Quantile price forecast over 30 days (uncertainty-aware)

What it encodes. Uncertainty in the terminal (or representative) forecasted price level. In our implementation, we report a small set of quantiles for the Day-30 forecast:

$$\left\{\hat{p}_{t+30}^{(q)}\right\}_{q\in\{0.05,\,0.5,\,0.95\}}.$$

Prompt snippet.

```
TSFM Forecast for <TICKER> (30-day quantile price forecast):
q05: <\hat{p}^{(0.05)}_{t+30}>, q50: <\hat{p}^{(0.5)}_{t+30}>, q95:
↪  <\hat{p}^{(0.95)}_{t+30}>
```

Format 5: Quantile return forecast over 30 days (uncertainty-aware)

What it encodes. Uncertainty in the *terminal* 30-day return ratio:

$$\left\{\hat{r}_{t+30}^{(q)}\right\}_{q\in\mathcal{Q}},$$

often with $\mathcal{Q} = \{0.05, 0.25, 0.5, 0.75, 0.95\}$ depending on the TSFM.

Prompt snippet.

```
TSFM Forecast for <TICKER> (30-day quantile returns):
q05: <\hat{r}^{(0.05)}_{t+30}>
q25: <\hat{r}^{(0.25)}_{t+30}>
q50: <\hat{r}^{(0.50)}_{t+30}>
q75: <\hat{r}^{(0.75)}_{t+30}>
q95: <\hat{r}^{(0.95)}_{t+30}>
```

Format 6: Quantile returns summarized across multiple horizons (uncertainty-aware + compressed)

What it encodes. Uncertainty-aware multi-horizon summaries:

$$\left\{\hat{r}_{t+h}^{(q)}\right\}_{q\in\{0.05,\,0.5,\,0.95\},\,h\in\{1,5,10,15,20\}}.$$

This format exposes both horizon structure and forecast dispersion.

Prompt snippet.

```
TSFM Forecast for <TICKER> (multi-horizon quantile returns):
1 Day:   [q05=<...>, q50=<...>, q95=<...>]
1 Week:  [q05=<...>, q50=<...>, q95=<...>]
2 Weeks: [q05=<...>, q50=<...>, q95=<...>]
3 Weeks: [q05=<...>, q50=<...>, q95=<...>]
4 Weeks: [q05=<...>, q50=<...>, q95=<...>]
```

B LLM PROMPT TEMPLATE

This section documents the prompt structure used to query the LLM agent for daily portfolio-weight decisions. Our implementation uses a two-message chat format: a system message specifying the role, rules, and required JSON schema, followed by a user message containing the day-specific market context (date, current portfolio state, fundamentals, price history, and optionally TSFM forecasts).

B.1 SYSTEM MESSAGE: ROLE, CONSTRAINTS, AND JSON SCHEMA

System message (fixed across all runs)

````
You are a portfolio manager for MAG7 stocks (AAPL, GOOGL, AMZN, MSFT,
↪  META, TSLA, NVDA) plus CASH. You should trade daily.

Your task is to allocate portfolio weights based on the provided
↪  information.

RULES:
1. Output weights for all 8 assets (7 stocks + CASH) that sum to exactly
↪  1.0
2. Weights must be between 0.0 and 1.0
3. You can set weight to 0 if you want to avoid a stock
4. CASH represents the uninvested portion (cash held in account)
5. Consider risk diversification
6. You can hold cash (CASH > 0) if you want to reduce exposure

OUTPUT FORMAT (MUST follow exactly):
```json
{
  "action": "rebalance" or "hold",
  "weights": {
    "AAPL": 0.XX,
    "GOOGL": 0.XX,
    "AMZN": 0.XX,
    "MSFT": 0.XX,
    "META": 0.XX,
    "TSLA": 0.XX,
    "NVDA": 0.XX,
    "CASH": 0.XX
  },
  "confidence": 0.X,
  "reasoning": "Brief explanation"
}
```

IMPORTANT: Weights MUST sum to 1.0 exactly. All 8 assets (7 stocks +
↪  CASH) must be included.
````

B.2 USER MESSAGE: DAY-SPECIFIC CONTEXT (INPUTS TO THE DECISION)

On each trading day t, we construct a single user message by concatenating the following blocks in order:

1. Date (decision date t).
2. Current Portfolio Weights (a JSON dictionary of weights from the previous day).
3. Company Fundamentals (one block per ticker).
4. Price History (one line per ticker: 30-day percentage change and the full 30-day price list).
5. TSFM Forecasts (optional; only included for LLM+TSFM variants, and formatted according to Formats 1–6 in Appendix **??**).

User message template (placeholders)

```
Date: <YYYY-MM-DD>
Current Portfolio Weights: {"AAPL":..., "GOOGL":..., "AMZN":...,
↪  "MSFT":..., "META":..., "TSLA":..., "NVDA":..., "CASH":...}

=== COMPANY FUNDAMENTALS ===

--- AAPL ---
<fundamentals text for AAPL, as-of date t>

--- GOOGL ---
<fundamentals text for GOOGL, as-of date t>

... (other tickers)

=== PRICE HISTORY (Last 30 days) ===
AAPL (30d change: +X.X%): [p_{t-29}, p_{t-28}, ..., p_t]
GOOGL (30d change: +X.X%): [p_{t-29}, p_{t-28}, ..., p_t]
... (other tickers)

=== TSFM FORECASTS ===
<optional; present only for LLM+TSFM experiments>
<format-specific forecast text for AAPL>
<format-specific forecast text for GOOGL>
... (other tickers)
```

B.3 Output parsing and safety checks

The agent expects the LLM to return a JSON object inside a fenced ```` ```json ```` block. We parse the JSON and enforce portfolio constraints: (i) all 8 assets must be present, (ii) weights must be clipped to $[0, 1]$, and (iii) weights must sum to 1 (renormalized if needed).

C Raw Results

Description. Table 1 reports the complete raw results of all portfolio allocation experiments used in our analysis. Each row corresponds to a single backtesting run over a fixed 30-day trading horizon. For TSFM-enhanced experiments, the column "Format" indicates the specific representation used to present TSFM predictions to the LLM:

1. raw 30-day price trajectories,
2. 30-day return ratios relative to the latest closing price,
3. aggregated return ratios at multiple horizons (1 day, 1 week, 2 weeks, 3 weeks, and 4 weeks),
4. quantile price forecasts over a 30-day horizon,
5. quantile return forecasts over a 30-day horizon,
6. quantile return summaries across multiple horizons.

All six formats convey equivalent predictive information but differ in representation and aggregation granularity, as described in the main text.

For the LLM-only baseline, we repeat each experiment four times for the same time period. Although we fix the random seed to 123 and set the LLM decoding temperature to zero, minor stochasticity remains due to nondeterministic components in the inference and execution pipeline. These repeated runs allow us to estimate the variance of the baseline performance and to provide a stable reference when assessing the statistical significance of TSFM improvements.

Table 1: Raw experimental results.

Model	Experiment Type	Start Date	End Date	Total Return	Sharpe Ratio	Sortino Ratio	Maximum Drawdown
chronos	baseline_llm_only	2025-08-31	2025-09-30	0.089679	7.420769	12.941830	0.021558
chronos	llm_tsfm_format_1	2025-08-31	2025-09-30	0.101869	7.744835	15.617583	0.020846
chronos	llm_tsfm_format_2	2025-08-31	2025-09-30	0.092984	7.899296	13.507606	0.022614
chronos	llm_tsfm_format_3	2025-08-31	2025-09-30	0.098370	8.049481	14.471984	0.020084
chronos	llm_tsfm_format_4	2025-08-31	2025-09-30	0.091639	7.185410	14.047088	0.019883
chronos	llm_tsfm_format_5	2025-08-31	2025-09-30	0.075083	7.006539	11.471131	0.019262
chronos	llm_tsfm_format_6	2025-08-31	2025-09-30	0.072656	6.233917	12.107029	0.020937
chronos	baseline_llm_only	2025-10-01	2025-10-31	0.043346	2.004502	2.593603	0.037741
chronos	llm_tsfm_format_1	2025-10-01	2025-10-31	0.034609	1.502307	2.103137	0.041670
chronos	llm_tsfm_format_2	2025-10-01	2025-10-31	0.044306	2.001262	2.768773	0.040894
chronos	llm_tsfm_format_3	2025-10-01	2025-10-31	0.048360	2.135604	2.958469	0.040740
chronos	llm_tsfm_format_4	2025-10-01	2025-10-31	0.042141	1.906275	2.739885	0.040235
chronos	llm_tsfm_format_5	2025-10-01	2025-10-31	0.051276	2.313863	2.955521	0.040381
chronos	llm_tsfm_format_6	2025-10-01	2025-10-31	0.047986	2.418914	3.157557	0.032372
chronos	baseline_llm_only	2025-12-17	2026-01-16	0.013373	1.071227	2.090674	0.030211
chronos	llm_tsfm_format_1	2025-12-17	2026-01-16	0.018463	1.828744	3.533807	0.023383
chronos	llm_tsfm_format_2	2025-12-17	2026-01-16	0.026722	2.768204	4.997030	0.017223
chronos	llm_tsfm_format_3	2025-12-17	2026-01-16	0.025715	2.476665	4.962544	0.020947
chronos	llm_tsfm_format_4	2025-12-17	2026-01-16	0.023342	2.326285	4.382675	0.022460
chronos	llm_tsfm_format_5	2025-12-17	2026-01-16	0.022448	2.212759	4.137362	0.024824
chronos	llm_tsfm_format_6	2025-12-17	2026-01-16	0.024894	2.432582	4.447671	0.024734
moirai	baseline_llm_only	2025-08-31	2025-09-30	0.096498	7.855446	15.187023	0.019645
moirai	llm_tsfm_format_1	2025-08-31	2025-09-30	0.098567	7.428872	13.264565	0.022315
moirai	llm_tsfm_format_2	2025-08-31	2025-09-30	0.117107	8.767071	16.406191	0.017707
moirai	llm_tsfm_format_3	2025-08-31	2025-09-30	0.115265	8.412441	16.402494	0.018542
moirai	llm_tsfm_format_4	2025-08-31	2025-09-30	0.092882	7.485634	13.213941	0.021291
moirai	llm_tsfm_format_5	2025-08-31	2025-09-30	0.087117	6.981576	13.577985	0.021798
moirai	llm_tsfm_format_6	2025-08-31	2025-09-30	0.078784	6.006343	12.180906	0.022900
moirai	baseline_llm_only	2025-10-01	2025-10-31	0.049278	2.331322	3.311585	0.035162
moirai	llm_tsfm_format_1	2025-10-01	2025-10-31	0.041501	1.863879	2.495008	0.041708
moirai	llm_tsfm_format_2	2025-10-01	2025-10-31	0.030384	1.420993	2.006751	0.044136
moirai	llm_tsfm_format_3	2025-10-01	2025-10-31	0.035988	1.716548	2.452338	0.046915
moirai	llm_tsfm_format_4	2025-10-01	2025-10-31	0.034161	1.491258	2.027910	0.043025
moirai	llm_tsfm_format_5	2025-10-01	2025-10-31	0.048207	2.274446	2.869874	0.040030
moirai	llm_tsfm_format_6	2025-10-01	2025-10-31	0.055865	2.777752	3.731107	0.030924
moirai	baseline_llm_only	2025-12-17	2026-01-16	0.004702	0.112331	0.197176	0.034277
moirai	llm_tsfm_format_1	2025-12-17	2026-01-16	0.012547	1.037862	1.912035	0.025453
moirai	llm_tsfm_format_2	2025-12-17	2026-01-16	0.027310	3.089540	5.937455	0.016907
moirai	llm_tsfm_format_3	2025-12-17	2026-01-16	0.029989	3.110332	6.209341	0.017954
moirai	llm_tsfm_format_4	2025-12-17	2026-01-16	-0.001606	-0.633741	-1.066362	0.035094
moirai	llm_tsfm_format_5	2025-12-17	2026-01-16	0.009192	0.838260	1.297335	0.018288
moirai	llm_tsfm_format_6	2025-12-17	2026-01-16	0.019976	1.946109	3.364778	0.023415
timesfm	baseline_llm_only	2025-08-31	2025-09-30	0.097329	7.920053	15.640708	0.019342
timesfm	llm_tsfm_format_1	2025-08-31	2025-09-30	0.112183	8.718804	18.449037	0.019197
timesfm	llm_tsfm_format_2	2025-08-31	2025-09-30	0.103250	8.490869	22.046363	0.016521
timesfm	llm_tsfm_format_3	2025-08-31	2025-09-30	0.096898	7.934206	25.394649	0.017853
timesfm	llm_tsfm_format_4	2025-08-31	2025-09-30	0.096374	7.420182	12.862320	0.021022
timesfm	llm_tsfm_format_5	2025-08-31	2025-09-30	0.104852	8.681445	18.217601	0.020575
timesfm	llm_tsfm_format_6	2025-08-31	2025-09-30	0.089192	7.352011	15.169284	0.020567
timesfm	baseline_llm_only	2025-10-01	2025-10-31	0.035356	1.588529	2.067106	0.043947
timesfm	llm_tsfm_format_1	2025-10-01	2025-10-31	0.039772	1.739849	2.369760	0.043100
timesfm	llm_tsfm_format_2	2025-10-01	2025-10-31	0.039417	1.978605	2.349499	0.034151
timesfm	llm_tsfm_format_3	2025-10-01	2025-10-31	0.048677	2.376568	2.815250	0.035221
timesfm	llm_tsfm_format_4	2025-10-01	2025-10-31	0.038073	1.671996	2.354075	0.043971
timesfm	llm_tsfm_format_5	2025-10-01	2025-10-31	0.054588	2.719253	3.313777	0.033911
timesfm	llm_tsfm_format_6	2025-10-01	2025-10-31	0.048135	2.345381	2.929174	0.033275
timesfm	baseline_llm_only	2025-12-17	2026-01-16	-0.001089	-0.529905	-0.883059	0.037739
timesfm	llm_tsfm_format_1	2025-12-17	2026-01-16	0.012890	1.075805	1.863554	0.027712
timesfm	llm_tsfm_format_2	2025-12-17	2026-01-16	0.031228	3.073042	5.347189	0.018915
timesfm	llm_tsfm_format_3	2025-12-17	2026-01-16	0.034069	3.369411	5.866282	0.018440
timesfm	llm_tsfm_format_4	2025-12-17	2026-01-16	0.009156	0.605732	1.023509	0.030526
timesfm	llm_tsfm_format_5	2025-12-17	2026-01-16	0.027906	2.765527	5.040878	0.022897
timesfm	llm_tsfm_format_6	2025-12-17	2026-01-16	0.021628	2.092216	3.729825	0.024571
toto	baseline_llm_only	2025-08-31	2025-09-30	0.095112	7.824192	14.095847	0.019818
toto	llm_tsfm_format_1	2025-08-31	2025-09-30	0.098047	7.616913	13.679186	0.019501
toto	llm_tsfm_format_2	2025-08-31	2025-09-30	0.096088	7.292384	13.055476	0.024210

toto	llm˙tsfm˙format˙3	2025-08-31	2025-09-30	0.092615	6.831870	14.077929	0.023256
toto	llm˙tsfm˙format˙4	2025-08-31	2025-09-30	0.106228	7.540698	14.137136	0.020912
toto	llm˙tsfm˙format˙5	2025-08-31	2025-09-30	0.101670	7.447052	13.513119	0.022969
toto	llm˙tsfm˙format˙6	2025-08-31	2025-09-30	0.100340	7.526232	15.965233	0.020823
toto	baseline˙llm˙only	2025-10-01	2025-10-31	0.042568	1.926789	2.575705	0.039535
toto	llm˙tsfm˙format˙1	2025-10-01	2025-10-31	0.044354	1.958729	2.742229	0.037956
toto	llm˙tsfm˙format˙2	2025-10-01	2025-10-31	0.031042	1.393990	1.846895	0.041774
toto	llm˙tsfm˙format˙3	2025-10-01	2025-10-31	0.046324	2.177687	2.964064	0.038173
toto	llm˙tsfm˙format˙4	2025-10-01	2025-10-31	0.043914	1.891921	2.609010	0.042030
toto	llm˙tsfm˙format˙5	2025-10-01	2025-10-31	0.040331	1.704249	2.222675	0.041687
toto	llm˙tsfm˙format˙6	2025-10-01	2025-10-31	0.042659	1.817602	2.520719	0.041636
toto	baseline˙llm˙only	2025-12-17	2026-01-16	0.001428	-0.254791	-0.432123	0.036747
toto	llm˙tsfm˙format˙1	2025-12-17	2026-01-16	0.007102	0.404865	0.752326	0.032871
toto	llm˙tsfm˙format˙2	2025-12-17	2026-01-16	0.005576	0.214058	0.351766	0.033545
toto	llm˙tsfm˙format˙3	2025-12-17	2026-01-16	0.003360	-0.048394	-0.081574	0.031269
toto	llm˙tsfm˙format˙4	2025-12-17	2026-01-16	-0.009106	-1.467843	-2.421569	0.040581
toto	llm˙tsfm˙format˙5	2025-12-17	2026-01-16	-0.000809	-0.522897	-0.874534	0.037385
toto	llm˙tsfm˙format˙6	2025-12-17	2026-01-16	0.011227	0.897653	1.386508	0.027388

D Statistical Tests and Additional Discussion

Unless stated otherwise, the statistics in this section are computed on all four TSFM backends (Chronos-2, TimesFM-2.5, Moirai-2, and Toto), spanning three non-overlapping periods and six formats. This yields $4 \times 3 \times 6 = 72$ paired differences for the format improvement analysis.

D.1 Baseline stability across backends

Because the LLM-only baseline does not consume TSFM outputs, its performance should be invariant to the forecasting backend. We thus obtain repeated baseline measurements within each period. Table 2 shows that baseline returns exhibit small dispersion across the four backend runs, indicating that the baseline is sufficiently stable to serve as a reference for paired improvement tests.

Period (start)	n	Mean $\pm$ Std (%)	95% CI (%)	Range (%)
2025-08-31	4	9.47 ± 0.34	[8.92, 10.01]	[8.97, 9.73]
2025-10-01	4	4.26 ± 0.57	[3.36, 5.17]	[3.54, 4.93]
2025-12-17	4	0.46 ± 0.63	[-0.54, 1.46]	[-0.11, 1.34]

Table 2: LLM-only baseline consistency across four TSFM backends (Chronos-2, TimesFM-2.5, Moirai-2, Toto). Values are 30-day cumulative returns.

D.2 Overall benefit of TSFM signals (format-agnostic)

We evaluate whether TSFM signals improve decisions regardless of the prompt format. For each TSFM-augmented run i in period p, we define a paired improvement

$$\Delta_i \;=\; R_i^{(\mathrm{LLM+TSFM})} \;-\; \overline{R}_p^{(\mathrm{baseline})},$$

where $\overline{R}_p^{(\mathrm{baseline})}$ is the period-wise mean baseline return computed from four repeated baseline runs. We test $H_0 : \mathbb{E}[\Delta] = 0$ versus $H_1 : \mathbb{E}[\Delta] > 0$ using a one-sample t-test on $\{\Delta_i\}$. To reduce reliance on normality assumptions, we also report a nonparametric bootstrap confidence interval.

Interpretation. Aggregating all periods, formats, and TSFM backends, the mean improvement is strictly positive, with both t-based and bootstrap confidence intervals excluding zero. The effect size ($d_z \approx 0.42$) indicates a moderate paired improvement. Because Δ_i is defined relative to the period-wise baseline mean, this analysis controls for regime differences across the three evaluation windows.

Statistic	Value
Paired samples (n)	72
Mean improvement (%)	0.47
95% CI (t, %)	[0.21, 0.74]
95% CI (bootstrap, %)	[0.22, 0.73]
One-sided t-test ($H_1 : \mu > 0$)	$t = 3.57, \;\; p = 3.22 \times 10^{-4}$
Effect size (Cohen's d_z)	0.42

Table 3: Format-agnostic improvement of LLM+TSFM over the period-wise baseline mean (aggregated across all four TSFM backends). All numbers are in percentage points of 30-day cumulative return.

Format	n	Mean Δ (%)	95% CI (t, %)	95% CI (boot, %)	t	d_z	p_{Holm}
1	12	0.45	[0.00, 0.90]	[0.08, 0.84]	2.22	0.64	0.120
2	12	0.65	[-0.23, 1.52]	[-0.09, 1.40]	1.63	0.47	0.263
3	12	0.90	[0.14, 1.66]	[0.28, 1.56]	2.62	0.76	0.072
4	12	-0.00	[-0.56, 0.55]	[-0.45, 0.49]	-0.01	-0.00	0.505
5	12	0.45	[-0.29, 1.20]	[-0.20, 1.07]	1.34	0.39	0.313
6	12	0.38	[-0.44, 1.20]	[-0.36, 1.04]	1.02	0.29	0.329

Table 4: Format-wise improvements over the period-wise baseline mean (aggregated across all four TSFM backends). p_{Holm} denotes Holm-corrected one-sided p-values across the six formats. Mean Δ and confidence intervals are in percentage points of 30-day cumulative return.

D.3 Is the gain sensitive to the TSFM presentation format?

To test whether the mean improvement differs across formats, we perform a one-way ANOVA on Δ grouped by format (H_0: all format means are equal). We obtain $F = 0.8483$ with $p = 0.5206$, providing no evidence that format choice materially changes the mean improvement at conventional significance levels.

For completeness, Table 4 reports per-format one-sample t-tests against zero (one-sided, $H_1 : \mathbb{E}[\Delta] > 0$), with Holm correction to control family-wise error across the six formats. Formats 1 and 3 achieve positive mean improvements with uncorrected one-sided p-values below 0.05, but none remain significant after Holm correction at $\alpha = 0.05$. Given the modest per-format sample size ($n = 12$), these results are consistent with limited power for distinguishing formats, rather than strong evidence that a particular format is uniquely optimal.

The aggregate evidence supports the main claim that TSFM signals improve LLM decisions in our controlled backtesting setting. Within the statistical resolution of the current experiment budget, we find no reliable evidence that the benefit depends strongly on how forecasts are formatted, suggesting that the improvement is driven primarily by the availability of numerical predictive signals rather than prompt-level formatting choices.

E Data leakage prevention in our implementation

We implement multiple safeguards to avoid using information from the future.

LLM release timeline vs. evaluation periods. We use an open-source Qwen model released in July 2025, while all our evaluation windows start after July 2025 (e.g., 2025-08-31, 2025-10-01, 2025-12-17). This reduces the risk that the model could have memorized outcomes from the evaluation intervals.

As-of fundamentals (no look-ahead). We do not query "full" company fundamentals directly. Instead, at each trading date t, we construct a snapshot using only information available as of t, and then format it for the LLM prompt at date t.

As-of price histories for settings. For every ticker and trading date t, we explicitly slice the historical price dataframe to ensure it only contains timestamps $\leq t$ before passing (i) the past price history to the LLM and (ii) the context series to TSFMs. In debug mode, we include assertions that raise an error if any timestamp exceeds the current decision date.

TSFM context truncated at the trading date. When generating TSFM forecasts for a trading date t, we first truncate the context series up to t. This ensures the TSFM never conditions on future prices, and we additionally log TSFM inputs for traceability.

Backtesting execution uses observed close prices. Trades are executed at the observed daily close price of the same day, without modeling market impact, slippage, or transaction costs. This isolates decision quality while keeping the simulator free of future-dependent execution rules.