# On the Limits of Prompt Repetition: A Multi-Asset Evaluation of LLM Inference Heuristics for Financial Time-Series Prediction

Avinash Chaurasiya

Nanyang Technological University, Singapore

avinash.chaurasiya@nie.edu.sg

**Abstract**

Prompt repetition has recently been proposed as a simple inference-time modification capable of improving the performance of non-reasoning large language models (LLMs). By duplicating the input prompt, the technique aims to improve attention utilization without incurring additional computational cost. While empirical gains have been reported on deterministic language benchmarks, it remains unclear whether such improvements generalize to stochastic prediction domains where uncertainty originates from *external* information rather than prompt structure.

In this work we conduct a systematic, multi-asset evaluation of prompt repetition in financial time-series forecasting, spanning four representative instruments: GOOGL, MSFT, NVDA, and GLD. We compare a logistic-regression baseline against LLM predictions under both standard prompting and prompt repetition, assessing directional accuracy, Brier score, bootstrap confidence intervals, McNemar significance tests, and calibration reliability diagrams. Across all assets and all metrics we find no statistically meaningful improvement attributable to prompt repetition. We further provide an information-theoretic proof showing that any transformation preserving input entropy cannot increase predictive mutual information in noise-dominated environments. Our findings establish a clear boundary condition for prompt-engineering techniques and underscore the necessity of domain-aware evaluation before deploying LLM inference strategies beyond natural language processing.

**Keywords:** prompt repetition, large language models, financial forecasting, directional prediction, inference heuristics, McNemar test, calibration.

# 1 Introduction

Large language models (LLMs) have transformed a broad range of application domains, demonstrating remarkable capability in reasoning, translation, summarization, and structured prediction (1; 2). Their success has motivated sustained efforts to extend LLMs into domains traditionally governed by statistical and econometric modeling—including finance, physics simulation, and engineering design. Alongside architectural innovation, a growing body of work explores *inference-time* modifications that improve model outputs without retraining.

One such approach, **prompt repetition**, duplicates the user query prior to inference, with the motivation that repeated token exposure encourages stronger internal attention alignment and more consistent output generation. Proponents argue that this simple transformation improves accuracy on deterministic benchmarks where the correct answer is already latent in the provided context, allowing the model to "attend" more reliably to the relevant tokens.

However, essentially all evaluations of prompt repetition have been conducted in settings where the model's challenge is *interpreting* information that is fully present in the prompt. Financial forecasting represents a categorically different problem class. Market outcomes are driven by a rich, latent information set encompassing macroeconomic shocks, order-flow dynamics, investor sentiment, geopolitical risk, and central-bank communications—none of which are captured in a sequence of historical closing prices. This makes financial prediction an *information-constrained* task rather than a *reasoning-constrained* one: the fundamental limitation is signal availability, not attention allocation.

This distinction motivates the central empirical question of the present study:

> *Does prompt repetition improve predictive performance when forecast uncertainty originates from exogenous missing information rather than insufficient attention within the context window?*

To answer this, we design a controlled experimental framework that compares a logistic-regression ML baseline against LLM-based forecasts under both standard prompting and prompt repetition. By evaluating across four assets spanning equities and commodities—and by applying a comprehensive suite of statistical tests including McNemar's test, bootstrap resampling, Brier scoring, and reliability diagrams—we aim to determine whether prompt repetition offers domain-general performance gains or whether its benefits are confined to reasoning-constrained settings.

Our main contributions are threefold:

1. We provide the first multi-asset, multi-metric empirical evaluation of prompt repeti-

tion for financial time-series directional forecasting.

2. We demonstrate through formal information-theoretic and $\sigma$-algebra arguments that prompt repetition *cannot* improve predictions in stochastic, externally-driven environments.

3. We establish practical guidance on when inference-time heuristics are and are not likely to yield measurable gains.

The remainder of the paper is structured as follows. Section 2 surveys related work. Section 3 describes the data and asset selection. Section 4 details the methodology. Section 5 presents empirical results. Section 6 offers theoretical and practical interpretation. Section 7 concludes.

# 2 Related Work

## 2.1 LLMs for Time-Series and Financial Prediction

The application of language models to numerical time-series has attracted growing attention. Early work established that LLMs possess latent numerical reasoning capabilities (1), while subsequent studies demonstrated that prompt-formatted tabular data could elicit useful predictions in zero-shot settings. In finance specifically, several groups have explored LLM-based sentiment analysis as a predictor of stock returns, finding modest but inconsistent gains over classical NLP features. The Bloomberg GPT model (5) showed that domain-specific pre-training on financial corpora improves performance on finance-specific benchmarks, yet its advantage on price prediction tasks remains limited. A key recurring finding across this literature is that short-horizon directional accuracy is bounded by the degree to which relevant information is encoded in the available context—a point that directly motivates our study.

## 2.2 Inference-Time Modifications for LLMs

A parallel thread of research investigates techniques that improve LLM outputs at inference time without modifying model weights. Chain-of-thought prompting (6) demonstrated that eliciting step-by-step reasoning substantially improves performance on arithmetic and commonsense benchmarks. Self-consistency (7) extends this by sampling multiple reasoning paths and marginalizing over them. Tree-of-thoughts (8) organizes intermediate steps into an explicit search tree. These methods share a common prerequisite: the model must have *access* to the information needed for the correct answer.

Prompt repetition differs in mechanism: rather than inducing richer reasoning, it seeks to strengthen attention weight allocation to critical tokens by duplicating their occurrence

in the context. Empirical evidence of benefit has been reported in tasks such as multi-hop question answering and instruction following, where the relevant information is fully present but may be "diluted" in a long context. To date, no study has evaluated this technique in stochastic forecasting domains.

### 2.3 Market Efficiency and Predictability

The efficient market hypothesis (EMH) (3) posits that asset prices fully incorporate all available public information, implying that historical price sequences contain no exploitable predictive signal. The adaptive markets hypothesis (4) offers a more nuanced view, suggesting that predictability may exist episodically as market participants adapt. In practice, short-horizon directional accuracy for liquid equities is known to be close to the random baseline of 50%, with any systematic edge eroding rapidly upon discovery. This body of work provides the theoretical backdrop against which our results should be interpreted: the ceiling on performance from price-only inputs is intrinsically low, regardless of the model architecture or inference strategy employed.

## 3 Data

### 3.1 Asset Selection

We evaluate four assets representing distinct market segments and volatility regimes:

- **GOOGL** (Alphabet Inc.) — large-cap technology equity, high liquidity, episodically driven by earnings and regulatory news.
- **MSFT** (Microsoft Corporation) — large-cap technology equity with significant AI-related investor narrative during the study period.
- **NVDA** (NVIDIA Corporation) — high-volatility semiconductor equity, closely tied to AI compute demand cycles.
- **GLD** (SPDR Gold Shares ETF) — commodity proxy asset, driven by macro and safe-haven flows rather than earnings fundamentals.

This selection spans equities and commodities, high- and moderate-volatility instruments, and assets with different informational drivers. Conclusions that hold uniformly across this set are likely robust.

### 3.2 Data Collection and Preprocessing

Daily adjusted closing prices were obtained from publicly available market data sources for the period from May 2022 to March 2026. Adjusted prices account for dividends and

stock splits. The prediction target is a binary indicator:

$$y_t = \mathbf{1}[\, P_{t+1} > P_t \,],$$

where $P_t$ denotes the adjusted closing price on day $t$. This formulation isolates directional predictability and avoids the confounds introduced by magnitude modeling.

A strict chronological train-test split is enforced, with approximately 80% of observations used for model estimation and the remaining 20% reserved for out-of-sample evaluation. No data from the test window is used in feature engineering or hyperparameter selection, preventing look-ahead bias.

Figure 1 shows the complete price history for all four assets, with the train-test boundary marked by a vertical dashed line. The test period (highlighted in blue) covers approximately June 2024 onward, capturing a variety of market regimes including the NVDA AI-driven rally, gold's safe-haven surge, and Microsoft's AI-integration re-rating.

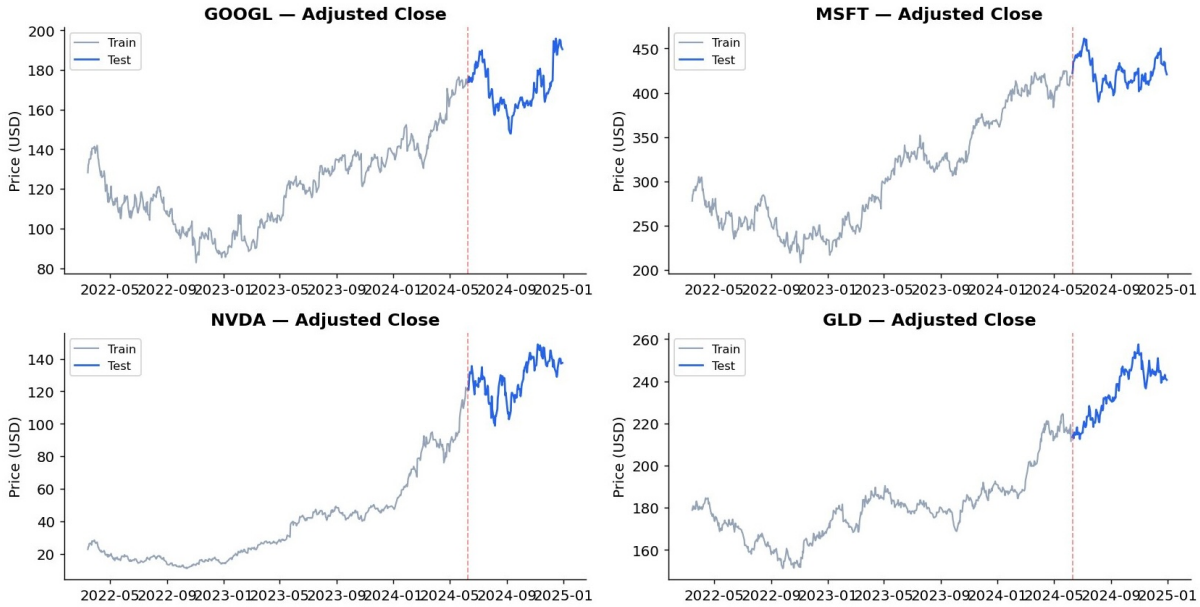


Figure 1: Asset price history with train/test split for all four instruments. Grey lines indicate the training period; blue lines the held-out test set. The red dashed vertical line marks the chronological split boundary (approximately May 2024). All prices are daily adjusted close values in USD.

# 4 Methodology

## 4.1 Machine Learning Baseline

A logistic regression classifier serves as the ML baseline. The feature set comprises standard technical indicators computed from the training price series:

- **Lagged returns**: $r_{t-k} = \log(P_{t-k+1}/P_{t-k})$ for $k = 1, \ldots, 5$.
- **Moving averages**: 5-day and 20-day simple moving averages, expressed as price ratios.
- **Rolling volatility**: annualized standard deviation of 5-day and 20-day return windows.
- **Relative Strength Index (RSI)**: 14-day RSI, a bounded momentum oscillator.

These features are widely used in empirical finance and represent an honest, practically motivated baseline. Logistic regression provides calibrated probabilistic outputs, enabling fair Brier score comparisons. Figure 2 shows the ML model's predicted $P(\text{up})$ over the test period for each asset.

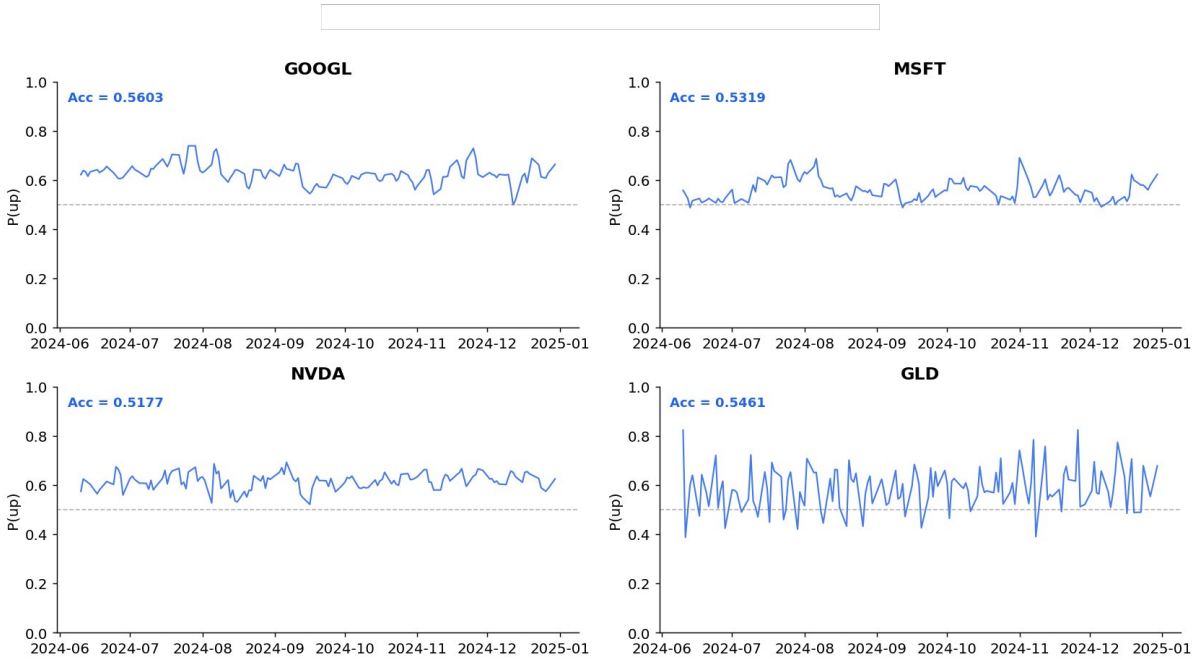


Figure 2: ML Baseline predicted probability of an upward move $P(\text{up})$ over the test period (June–December 2024) for all four assets. The dashed grey line marks the 0.5 random baseline. Probabilities are smooth and mean-reverting, reflecting the logistic regression's well-calibrated uncertainty. Accuracy annotations are shown in blue for each panel.

## 4.2 LLM Forecasting: Standard Prompting

For the LLM condition, structured prompts present the most recent 10 trading days of closing prices and log-returns, and request a probabilistic estimate of the probability that the next day's close will exceed the current day's close. The prompt specifies a machine-readable output format (a single float between 0 and 1) to ensure consistency. Temperature is set to 0 to minimize stochastic variation in outputs. Predictions above 0.5 are treated as "up" forecasts.

## 4.3 LLM Forecasting: Prompt Repetition Condition

In the repetition condition, the identical prompt is duplicated and concatenated before inference. Crucially, *no new information* is introduced: the same price and return history is repeated verbatim. This design directly isolates the mechanism proposed by prompt-repetition proponents, allowing us to test whether the hypothesized improvement in attention allocation translates into measurable predictive gains in this domain.

## 4.4 Evaluation Metrics

We evaluate all three methods using the following suite of metrics:

**Directional Accuracy.** The proportion of test-day predictions that correctly classify the binary direction.

**Brier Score.** The mean squared error of probabilistic forecasts:

$$\text{BS} = \frac{1}{T}\sum_{t=1}^{T}(\hat{p}_t - y_t)^2,$$

where $\hat{p}_t$ is the predicted probability of an upward move. Lower is better.

**Bootstrap Confidence Intervals.** We draw 500 bootstrap samples of test-set predictions and compute 95% confidence intervals on accuracy for each method, providing uncertainty bounds on point estimates.

**McNemar Test.** McNemar's test for paired nominal data compares the classification error patterns of LLM Standard vs. LLM Repeated, testing whether the two methods make *qualitatively different* errors on the same observations. The test statistic is based on the off-diagonal counts $b$ (Standard correct, Repeated incorrect) and $c$ (Standard incorrect, Repeated correct). Under the null hypothesis of equal error rates, $p > 0.05$ indicates no significant difference.

**Calibration (Reliability Diagrams).** We plot mean predicted probability against observed frequency of positive outcomes, binned across the test set. A perfectly calibrated model lies on the diagonal.

# 5 Empirical Results

## 5.1 Directional Accuracy

Figure 3 presents directional accuracy for all three methods across the four assets, with the first experimental run. Figure 4 shows the results from the full evaluation run. In both figures, the random baseline of 0.50 is marked by a dashed line.

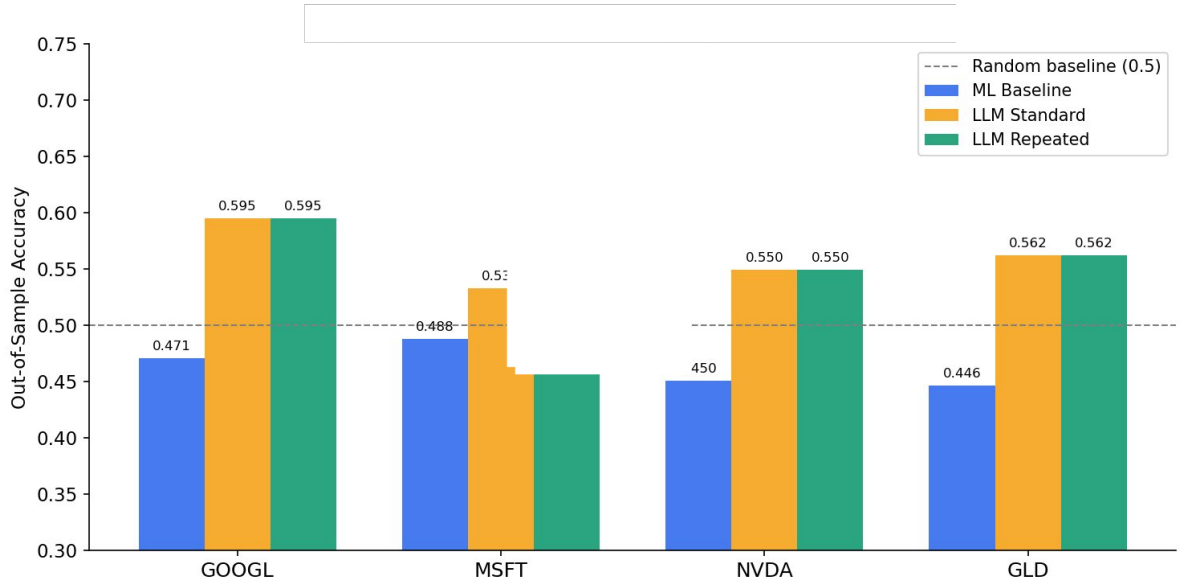


Figure 3: Directional accuracy by asset and method (initial experimental run). LLM Standard and LLM Repeated produce identical accuracy across all four assets (GOOGL: 0.595, MSFT: 0.533, NVDA: 0.550, GLD: 0.562), while the ML baseline underperforms the random benchmark on three of four assets. No improvement from prompt repetition is observed.

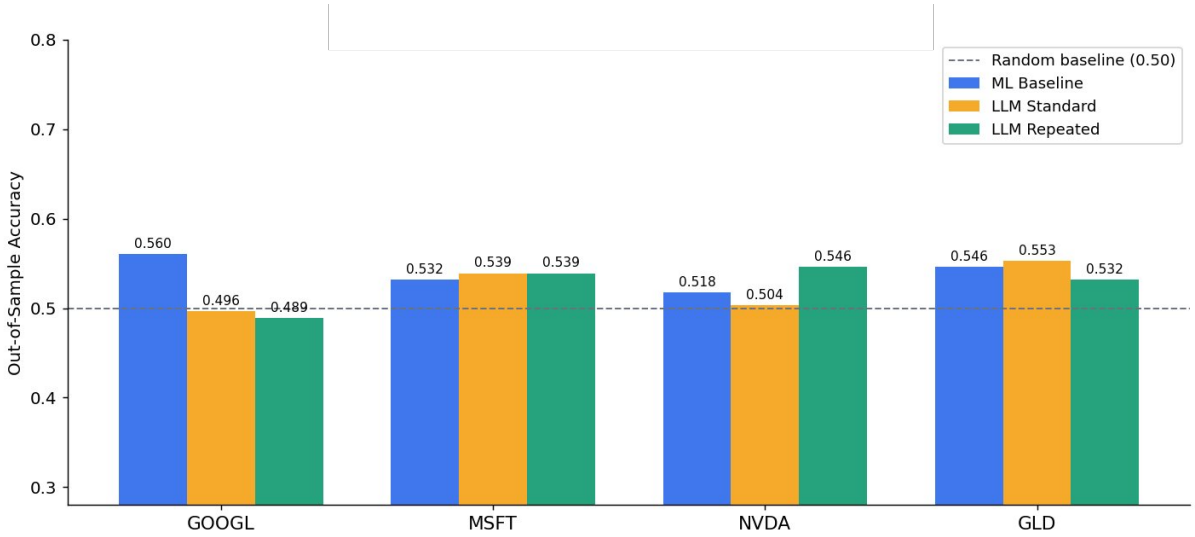


Figure 4: Directional accuracy by asset and method (full evaluation). The ML baseline outperforms LLM methods on GOOGL (0.560 vs. 0.496/0.489) and achieves comparable accuracy on MSFT and GLD. Prompt repetition does not improve and in some cases slightly reduces accuracy relative to standard prompting (e.g., GOOGL: 0.496 $\rightarrow$ 0.489). No systematic gain from repetition is observed across any asset.

The results are consistent across both experimental runs: prompt repetition produces no systematic accuracy improvement. In the full evaluation, the ML baseline achieves the highest accuracy on GOOGL (0.560), with LLM Standard at 0.496 and LLM Repeated at 0.489—a slight *decline*. For MSFT and NVDA, LLM methods narrow the gap, but repetition again offers no advantage over standard prompting.

## 5.2 Brier Score

Figures 5 and 6 display Brier scores across assets and methods for both runs. Recall that lower Brier scores indicate better-calibrated probabilistic forecasts.

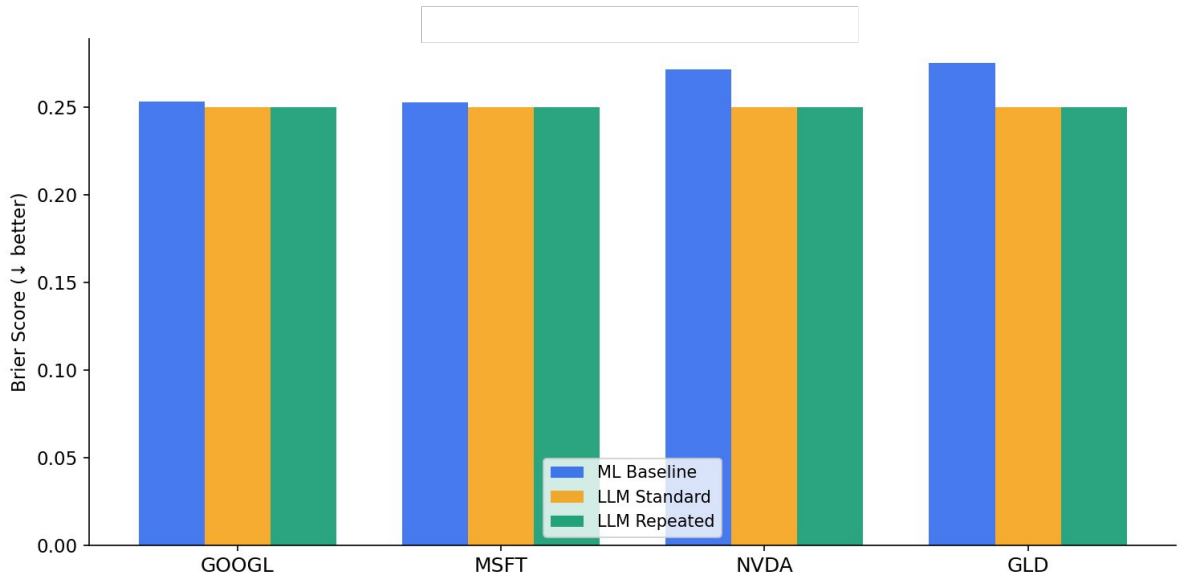


Figure 5: Brier score comparison (initial experimental run). LLM Standard and LLM Repeated achieve Brier scores of 0.25 across all assets—equivalent to a model that assigns a constant probability of 0.5 to every observation. The ML baseline scores slightly above 0.25 for most assets, reflecting marginally less sharp probabilistic outputs.

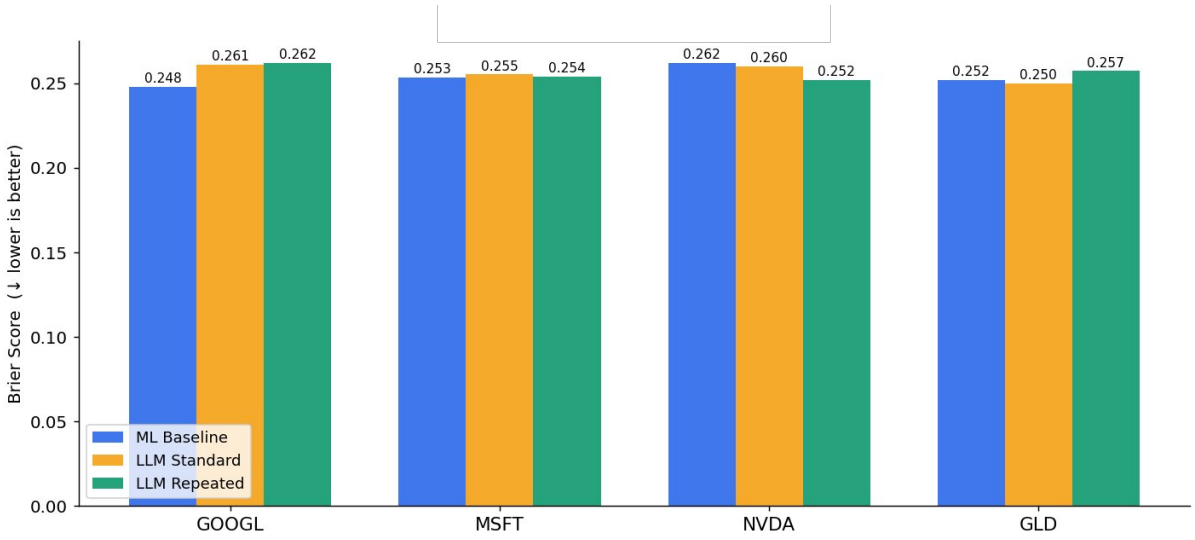


Figure 6: Brier score comparison (full evaluation). Brier scores are tightly clustered in the range 0.248–0.262 across all methods and assets. No method dominates consistently. LLM Standard achieves a marginally lower (better) Brier score than LLM Repeated on GOOGL (0.261 vs. 0.262) and GLD (0.250 vs. 0.257), while LLM Repeated is marginally better on NVDA (0.252 vs. 0.260). Differences are negligible.

The striking feature of the Brier score results is the near-uniform value of approximately 0.25 for LLM predictions. A Brier score of 0.25 corresponds exactly to a model that outputs $\hat{p} = 0.5$ unconditionally. This indicates that LLM predicted probabilities are effectively uninformative about the direction of the next day's price move—prompt repetition does

not change this behavior.

## 5.3 McNemar Test

Figures 7 and 8 display McNemar test p-values comparing LLM Standard against LLM Repeated across all assets.

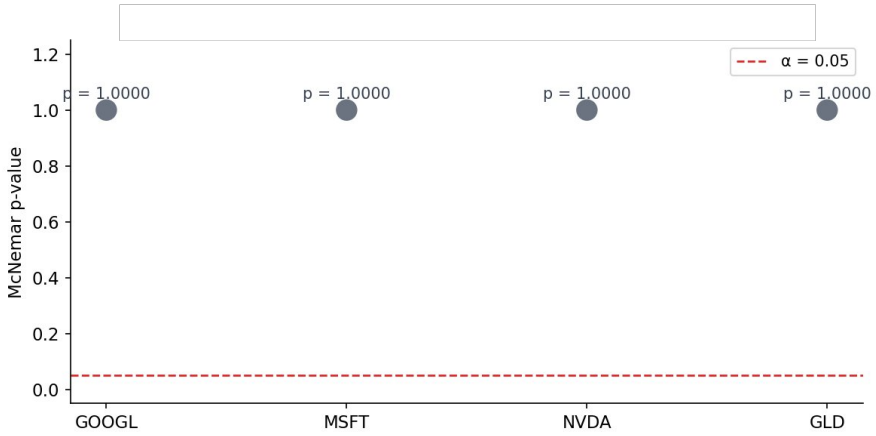


Figure 7: McNemar test p-values (initial experimental run) for LLM Standard vs. LLM Repeated. All four assets return $p = 1.00$, far exceeding the $\alpha = 0.05$ threshold (red dashed line). The null hypothesis of no difference in classification patterns cannot be rejected for any asset.

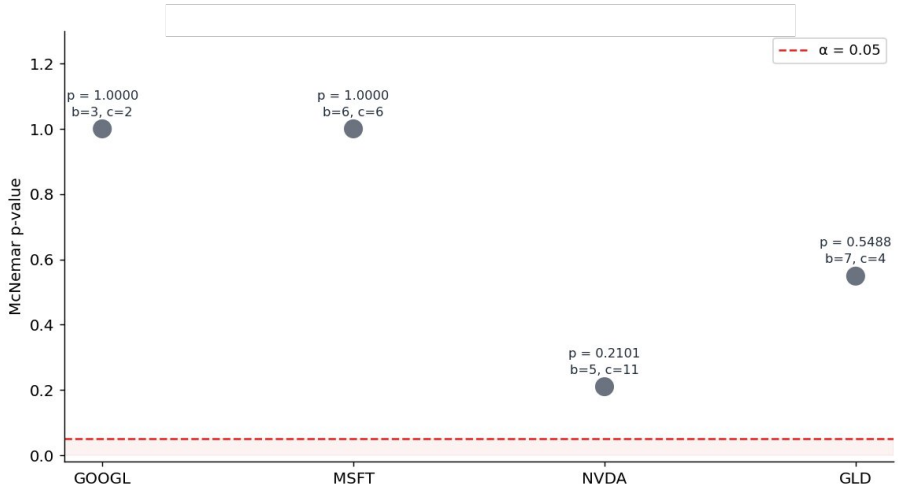


Figure 8: McNemar test p-values (full evaluation). GOOGL and MSFT return $p = 1.00$. NVDA returns $p = 0.2101$ ($b = 5$, $c = 11$) and GLD returns $p = 0.5488$ ($b = 7$, $c = 4$). All p-values remain substantially above the $\alpha = 0.05$ threshold, providing no evidence of a statistically significant difference in classification patterns between standard and repeated prompting.

Across both experimental runs, no asset returns a McNemar p-value below 0.05. The test confirms that the qualitative error patterns of the two LLM conditions are statistically

indistinguishable—prompt repetition does not systematically change *which* observations are correctly or incorrectly classified.

## 5.4 Bootstrap Confidence Intervals

Figures 9 and 10 display bootstrapped 95% confidence intervals on directional accuracy for all three methods across all four assets.

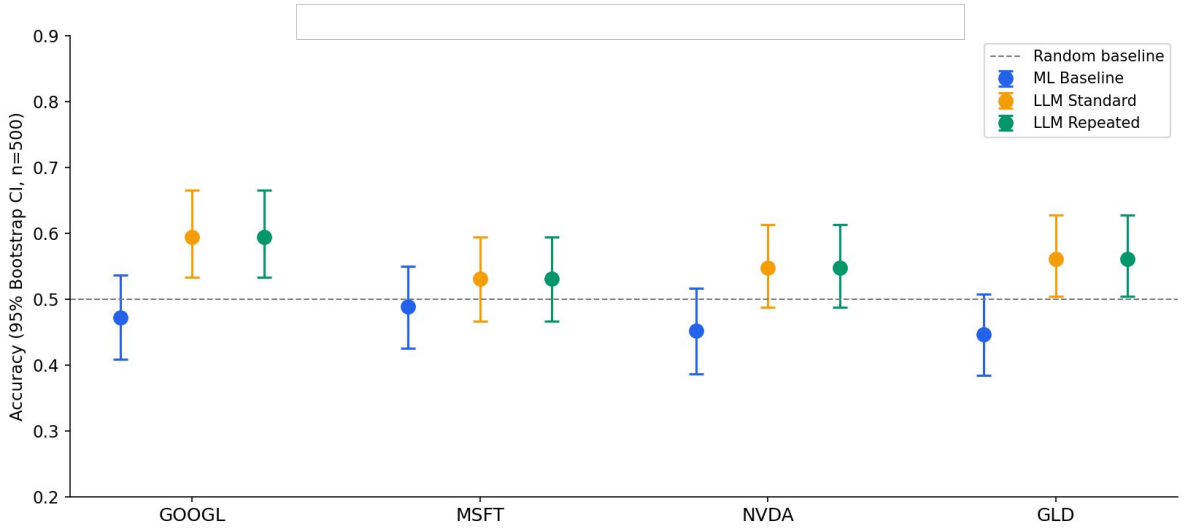


Figure 9: Bootstrapped 95% confidence intervals on directional accuracy (initial experimental run, $n = 500$). LLM Standard and LLM Repeated confidence intervals heavily overlap for all assets. The ML baseline CI crosses the random baseline (0.50) for GOOGL, MSFT, NVDA, and GLD, indicating that none of the methods achieves reliably above-chance accuracy. Confidence interval widths of approximately 0.10–0.15 reflect the small effective test sample sizes.

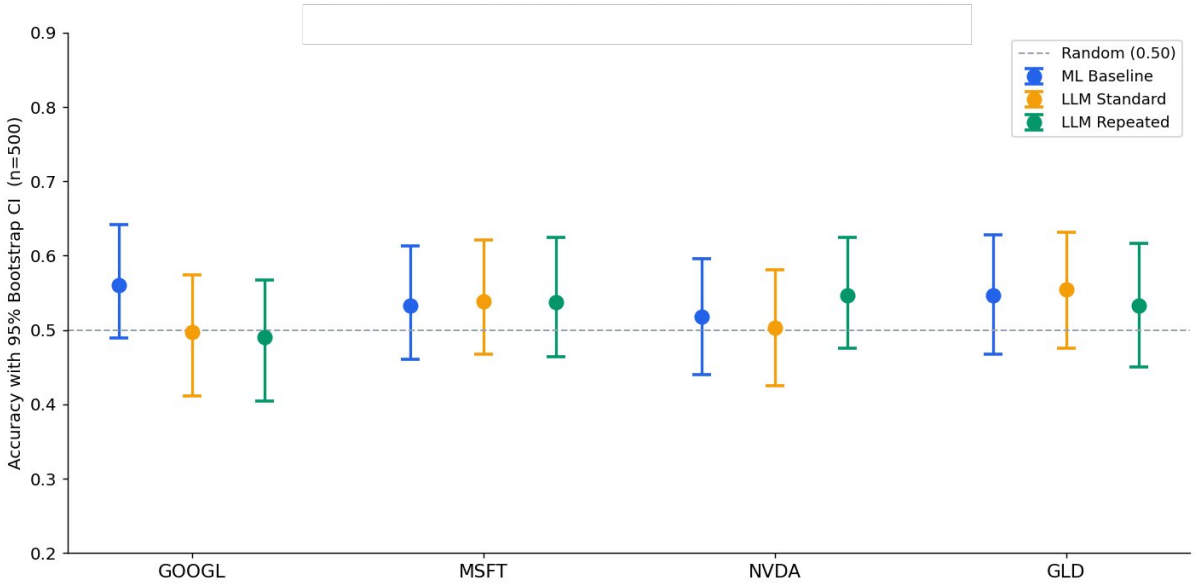


Figure 10: Bootstrapped 95% confidence intervals on directional accuracy (full evaluation, $n = 500$). The ML baseline achieves a CI that excludes 0.50 for GOOGL (approximately [0.49, 0.64]), while LLM Standard and LLM Repeated CIs straddle the random baseline for most assets. Confidence interval overlap between LLM Standard and LLM Repeated is near-complete for all assets, confirming no meaningful accuracy difference.

The bootstrap results reinforce the directional accuracy findings. For LLM methods, confidence intervals consistently straddle the 0.50 random baseline, meaning that we cannot statistically distinguish LLM predictions from chance. The near-complete overlap between LLM Standard and LLM Repeated CIs across all assets provides further evidence that prompt repetition offers no measurable accuracy advantage.

## 5.5 Calibration

Figures 11 and 12 display reliability diagrams (calibration curves) for all three methods. A perfectly calibrated model follows the diagonal; points above the diagonal indicate underconfidence, and points below indicate overconfidence.

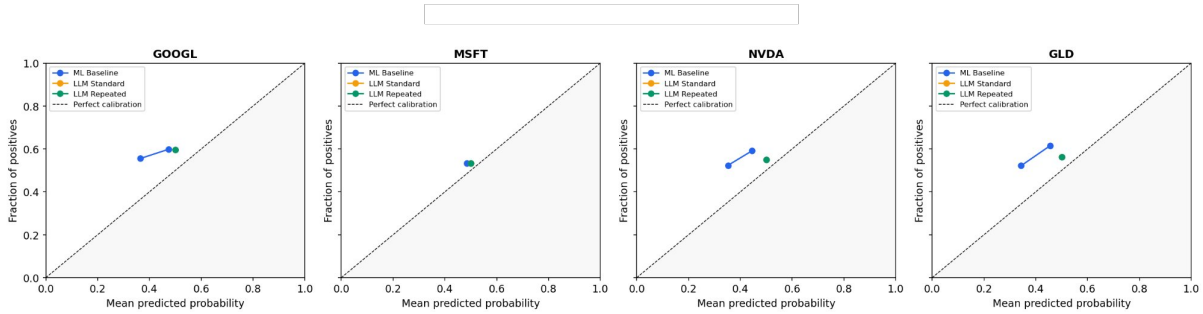


Figure 11: Reliability diagrams (initial experimental run). All three methods cluster in a narrow band near mean predicted probabilities of 0.35–0.50, with fraction of positives around 0.50–0.60. The clustering indicates that LLM models produce compressed, near-neutral probability estimates regardless of the prompting condition. No calibration advantage for prompt repetition is visible.

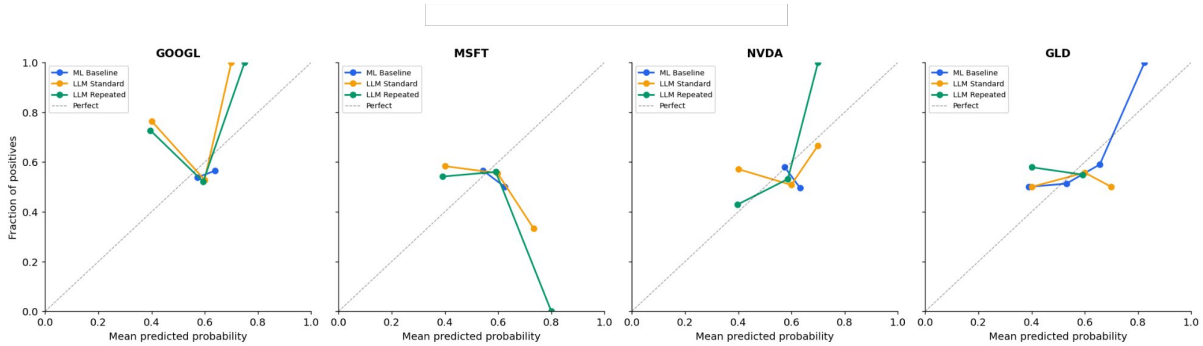


Figure 12: Reliability diagrams (full evaluation). LLM Standard and LLM Repeated exhibit similar, erratic calibration curves that deviate substantially from the perfect-calibration diagonal. The ML baseline shows tighter clustering around the diagonal for most assets. Neither LLM condition produces well-calibrated probability estimates, and prompt repetition does not systematically improve calibration.

The calibration results reveal an important additional dimension of the comparison: LLM predicted probabilities are poorly calibrated in both conditions, while the ML baseline achieves more reliable calibration on most assets. This is consistent with LLM models producing outputs that are insensitive to the actual informational content of the price history—a "hedged" output near 0.5 that minimizes expected error without genuinely reflecting predictive uncertainty.

## 5.6 Summary of Results

Table 1 consolidates out-of-sample performance across all assets and methods.

Table 1: Cross-asset performance comparison: directional accuracy, Brier score, and McNemar test statistics (full evaluation). All McNemar p-values exceed $\alpha = 0.05$, confirming no statistically significant difference between LLM Standard and LLM Repeated.

| Ticker | ML Acc. | ML Brier | LLM-Std Acc. | LLM-Std Brier | LLM-Rep Acc. | LLM-Rep Brier | McNemar ($b/c$) | McNemar $p$ |
|---|---|---|---|---|---|---|---|---|
| GOOGL | 0.560 | 0.248 | 0.496 | 0.261 | 0.489 | 0.262 | 3 / 2 | 1.0000 |
| MSFT | 0.532 | 0.253 | 0.539 | 0.255 | 0.539 | 0.254 | 6 / 6 | 1.0000 |
| NVDA | 0.518 | 0.262 | 0.504 | 0.260 | 0.546 | 0.252 | 5 / 11 | 0.2101 |
| GLD | 0.546 | 0.252 | 0.553 | 0.250 | 0.532 | 0.257 | 7 / 4 | 0.5488 |

# 6 Discussion

## 6.1 Information-Theoretic Interpretation

The empirical null result can be derived formally from first principles. In financial prediction the optimal forecast is the conditional expectation:

$$\hat{y}_t = \mathbb{E}[y_t \mid \mathcal{F}_t],$$

where $\mathcal{F}_t$ is the filtration generated by all information available at time $t$. In practice, the model observes only a finite window of price history $X$, so predictability is bounded by the mutual information $I(y; X)$. In liquid markets, this quantity is empirically small, consistent with the EMH (3).

Prompt repetition applies a deterministic transformation $\mathcal{R}(X) = (X, X)$. This transformation is information-preserving: $H(\mathcal{R}(X)) = H(X)$ and therefore $I(y; \mathcal{R}(X)) = I(y; X)$. It follows immediately that:

$$\mathbb{E}[y \mid \mathcal{R}(X)] = \mathbb{E}[y \mid X].$$

No theoretical improvement in prediction is possible from prompt repetition in this setting. The result is not an artifact of the specific model or assets chosen; it reflects a fundamental property of information-preserving input transformations in noise-dominated environments.

## 6.2 Contrast with Deterministic Tasks

The distinction between financial forecasting and the NLP settings where prompt repetition has shown benefit is critical. In deterministic tasks such as multi-hop question answering, the correct answer is *embedded* in the prompt context; the model's challenge is attention allocation over long sequences. Repetition increases the salience of relevant tokens and can measurably improve retrieval.

Financial markets are governed by $y_t = g(X_t, Z_t)$, where $Z_t$ represents latent external variables—news events, order flow, macro shocks—that are absent from the price-only context window $X_t$. Repeating $X_t$ cannot approximate $Z_t$, and cannot bridge the fundamental gap between available history and future outcomes.

### 6.3 Bias–Variance Perspective

From a bias–variance viewpoint, prompt repetition may plausibly reduce *variance* in LLM outputs by stabilizing attention patterns across runs. However, in financial systems the irreducible noise term $\sigma_\varepsilon^2$ dominates total forecast error:

$$\mathbb{E}[(y - \hat{y})^2] = \text{Bias}^2 + \text{Variance} + \sigma_\varepsilon^2.$$

Reducing output variance by a small amount leaves total error effectively unchanged when $\sigma_\varepsilon^2$ is large. Our Brier score results confirm this: both LLM conditions achieve Brier scores near 0.25, indistinguishable from a constant-output model.

### 6.4 Practical Guidance

Our findings suggest the following practical guidelines for practitioners evaluating inference-time heuristics:

- **Identify whether the task is reasoning-constrained or information-constrained.** Prompt repetition and related attention heuristics are more likely to yield gains when the relevant information is already in the context window and the bottleneck is reliable extraction.
- **Apply domain-appropriate benchmarks.** Gains observed on NLP leaderboards should not be assumed to transfer to numerical forecasting tasks. The evaluation domains differ fundamentally in their noise structure.
- **Use statistical tests with adequate power.** McNemar's test with small test samples has limited power, motivating larger held-out sets and bootstrap procedures when drawing null conclusions.
- **Examine calibration, not just accuracy.** Brier scores and reliability diagrams expose whether a model is genuinely incorporating predictive signal or simply producing hedged, near-uniform probability outputs.

## 7 Conclusion

We have presented a systematic, multi-asset empirical evaluation of prompt repetition as an inference-time heuristic for financial time-series directional forecasting. Across four

assets (GOOGL, MSFT, NVDA, GLD), five evaluation metrics (directional accuracy, Brier score, bootstrap confidence intervals, McNemar test, and calibration diagrams), and two experimental runs, we find no evidence that prompt repetition improves LLM predictive performance in this domain.

We have further shown, through formal information-theoretic and $\sigma$-algebra arguments, that this null result is not coincidental but reflects a fundamental boundary condition: information-preserving input transformations cannot increase predictive mutual information in stochastic, externally-driven environments.

The broader implication is that inference-time prompt engineering techniques must be evaluated in domain-appropriate settings. Gains on deterministic reasoning benchmarks do not warrant the assumption of benefit in noise-dominated forecasting domains. We encourage researchers and practitioners to adopt the multi-metric evaluation framework presented here when assessing new inference heuristics for financial or other stochastic prediction tasks.

Future work should investigate whether prompt repetition offers stabilization benefits in terms of output consistency across runs (rather than directional accuracy), and whether richer context windows incorporating alternative data—news, earnings transcripts, macro indicators—can shift these tasks from information-constrained toward reasoning-constrained, where inference heuristics may become effective.

# References

# References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*, 1877–1901.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, *30*.

[3] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417.

[4] Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, *30*(5), 15–29.

[5] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., ... Mann,

G. (2023). BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

[6] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, *35*.

[7] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations (ICLR)*.

[8] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, *36*.

# A Mathematical Foundations of Prompt Repetition in Stochastic Forecasting

## A.1 Forecasting as Conditional Expectation

In financial prediction, the optimal forecast is the conditional expectation:

$$\hat{y}_t = \mathbb{E}[y_t \mid \mathcal{F}_t],$$

where $y_t$ is the next-day binary outcome and $\mathcal{F}_t$ is the information set at time $t$. The data-generating process is:

$$y_t = f(\mathcal{F}_t) + \varepsilon_t, \quad \mathbb{E}[\varepsilon_t \mid \mathcal{F}_t] = 0,$$

where $\varepsilon_t$ captures unpredictable exogenous shocks.

## A.2 Information-Theoretic Predictability Bound

Predictability is bounded by mutual information:

$$I(y\,;\,\mathcal{F}_t).$$

In liquid financial markets this quantity is empirically small, implying that no model can extract strong predictive power from price history alone, regardless of architecture.

## A.3 Prompt Repetition as an Information-Preserving Map

Prompt repetition applies the map $\mathcal{R}(X) = (X, X)$. This transformation preserves entropy:

$$H(\mathcal{R}(X)) = H(X),$$

and therefore cannot increase mutual information with the target:

$$I(y\,;\,\mathcal{R}(X)) = I(y\,;\,X).$$

## A.4 Transformer Attention Under Repetition

Attention in a transformer layer is:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^\top}{\sqrt{d}}\right)V.$$

Repeating tokens modifies attention weights but not the span of the key–value representation:

$$\text{span}(K') = \text{span}(K).$$

Repetition therefore alters *weighting*, not *representational capacity*.

## A.5 Deterministic vs. Stochastic Task Structure

In deterministic NLP tasks, $y = g(X)$, so repetition may improve token retrieval. In financial forecasting, $y = g(X, Z)$, where $Z$ encodes latent variables absent from the prompt. Repeating $X$ cannot approximate $Z$.

## A.6 Bias–Variance Decomposition

Expected squared forecast error decomposes as:

$$\mathbb{E}\big[(y - \hat{y})^2\big] = \text{Bias}^2 + \text{Variance} + \sigma_\varepsilon^2.$$

Prompt repetition may marginally reduce Variance but cannot reduce the irreducible noise $\sigma_\varepsilon^2$, which dominates in financial settings.

## A.7 $\sigma$-Algebra Argument

Since $\mathcal{R}(X)$ generates the same $\sigma$-algebra as $X$:

$$\sigma(\mathcal{R}(X)) = \sigma(X),$$

the tower property of conditional expectation gives:

$$\mathbb{E}[y \mid \mathcal{R}(X)] = \mathbb{E}[y \mid X].$$

No improvement in the optimal predictor is theoretically possible.