A Test of Lookahead Bias in LLM Forecasts

Zhenyu Gao, Wenxi Jiang, Yutong Yan

CUHK Business School

January 1, 2026

Motivation

- LLMs increasingly used for economic/financial forecasting
  - News headlines $\rightarrow$ stock returns
  - Earnings calls $\rightarrow$ capital expenditures
- **Key concern:** Are predictions genuine reasoning or memorization?
- **Lookahead bias:** Models trained on historical data may have seen future outcomes
- **Challenge:** Limited out-of-sample data, retraining prohibitively expensive

Our Contribution

A cost-efficient statistical test to detect lookahead bias without model retraining

Example: The Lookahead Problem

Prompt (July 28, 2020)

"Kodak Triples on Loan to Make Covid-19 Drug Ingredients. Is this good or bad for the stock price?"

Two possibilities:

1. Genuine reasoning: Analyze government contract impact
2. Memorization: Recall July 29 headline: "Kodak's stock rose so fast it tripped 20 circuit breakers... shares ended up 318%"

Key Insight

If training corpus contains both the event AND its outcome, the model may "recall" rather than "reason"

Lookahead Propensity (LAP)

Building on Membership Inference Attacks (MIA):

For a prompt with tokens $w = (w_1, \ldots, w_N)$:

$$\text{LAP}(w, K) = \exp\left(\frac{1}{|S_K|} \sum_{t \in S_K} \log P_\theta(w_n | w_{\leq n-1})\right)$$

where S_K = bottom 20% of tokens ranked by probability

Intuition:

- Common words ("the", "and") have high probability regardless
- **Rare/unusual tokens** are informative
- Seen text $\rightarrow$ fewer low-probability outliers $\rightarrow$ higher LAP
- Unseen text $\rightarrow$ more outliers $\rightarrow$ lower LAP

Based on MIN-K% PROB (Shi et al., 2024): AUC = 0.72 on WIKIMIA, 0.88 on copyrighted books

Econometric Framework

Data generating process:

$$Y_{t+1} = \mu(X_t) + \epsilon_{t+1}$$

LLM prediction with lookahead bias:

$$\hat{\mu}_t = \mu(X_t) + L_t \epsilon_{t+1}$$

where L_t measures memorization strength

Test regression:

$$Y_{t+1} = \beta_1 \hat{\mu}_t + \beta_2 L_t + \beta_3 (L_t \times \hat{\mu}_t) + \varepsilon_{t+1}$$

Theorem (Detection)

$\beta_3 > 0 \iff$ *Lookahead bias present*

Key insight: If accuracy increases with LAP, predictions rely on memorization

Empirical Implementation

Exercise 1: News Headlines → Stock Returns

- 91,361 Bloomberg headlines (2012-2023)
- 1,587 CRSP-listed companies
- LLM classifies: good (+1), neutral (0), bad (-1)

Exercise 2: Earnings Calls → CapEx

- 74,338 firm-quarter observations (2006-2020)
- 3,897 unique firms
- LLM predicts: significantly decrease (-1) to significantly increase (+1)

Model: Llama-3.3 (70B, Dec 2024)

- Open-source: provides token probabilities for LAP computation
- Replicable: fixed checkpoint on HuggingFace

Results: News Headlines Predict Stock Returns

	(1) Baseline	(2) With LAP
LLM	0.210***	0.001
	(12.24)	(0.03)
LAP		-1.297***
		(-2.61)
LLM × LAP		2.866***
		(4.86)
Firm FE	Yes	Yes
Date FE	Yes	Yes
R^2	0.179	0.180
N	91,361	91,361

Economic magnitude: 1 SD increase in LAP raises LLM's marginal effect by 0.077% (37% of baseline effect)

Results: Earnings Calls Predict Capital Expenditures

	(1) Baseline	(2) With LAP
LLM	0.798***	0.514***
	(15.89)	(5.88)
LAP		-0.016
		(-0.57)
LLM × LAP		0.148***
		(3.59)
Firm FE	Yes	Yes
Quarter FE	Yes	Yes
R^2	0.642	0.643
N	74,338	74,338

Economic magnitude: 1 SD increase in LAP raises LLM's marginal effect by 0.149% (19% of baseline effect)

Results: Earnings Calls Predict Capital Expenditures

	(1) Baseline	(2) With LAP
LLM	0.798***	0.514***
	(15.89)	(5.88)
LAP		-0.016
		(-0.57)
LLM × LAP		0.148***
		(3.59)
Firm FE	Yes	Yes
Quarter FE	Yes	Yes
R^2	0.642	0.643
N	74,338	74,338

Economic magnitude: 1 SD increase in LAP raises LLM's marginal effect by 0.149% (19% of baseline effect)

Stronger Effect for Small Firms

Finding: Predictability stronger for small-cap stocks

	Without Triple Interaction	With Triple Interaction
LLM × Small	0.263***	-0.316**
	(4.23)	(-2.03)
LLM × LAP × Small		7.910***
		(3.53)

Robustness checks:

- Results hold controlling for:
  - First-token conditional probability $P(w_{N+1}|w_{\leq N})$
  - Model's self-reported confidence
- LAP captures distinct mechanism from model confidence

Out-of-Sample Validation

Placebo test using Llama-2:

- In-sample: Jan 2012 - Sep 2022
- Out-of-sample: Sep 2023 - Dec 2024 (after release)

	In-Sample		Out-of-Sample	
	(1)	(2)	(3)	(4)
LLM^{std}	0.049***	0.084***	0.049***	0.084***
$LLM^{std} \times LAP^{std}$		0.012***		-0.007
		(3.44)		(-0.74)

Bootstrap analysis: In-sample β_3 lies outside 95th percentile of out-of-sample distribution ($p = 0.033$)

$\Rightarrow$ Confirms lookahead bias in training period, absent post-release

Bootstrap Distribution

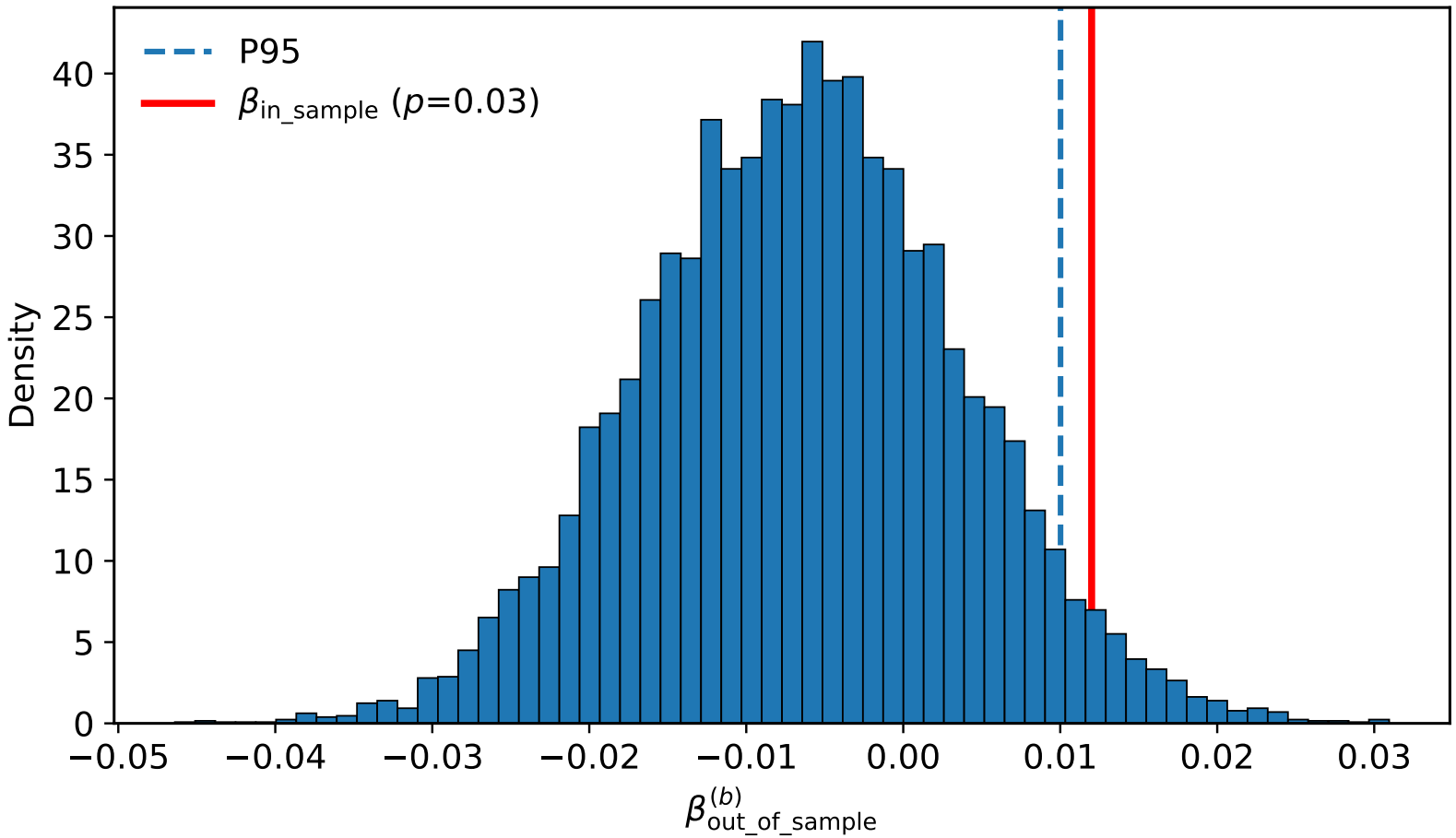

Figure: Bootstrap distribution of interaction coefficient β from out-of-sample data (10,000 replications). Blue dashed line: 95th percentile. Red solid line: in-sample estimate (p = 0.033).

Bootstrap Distribution

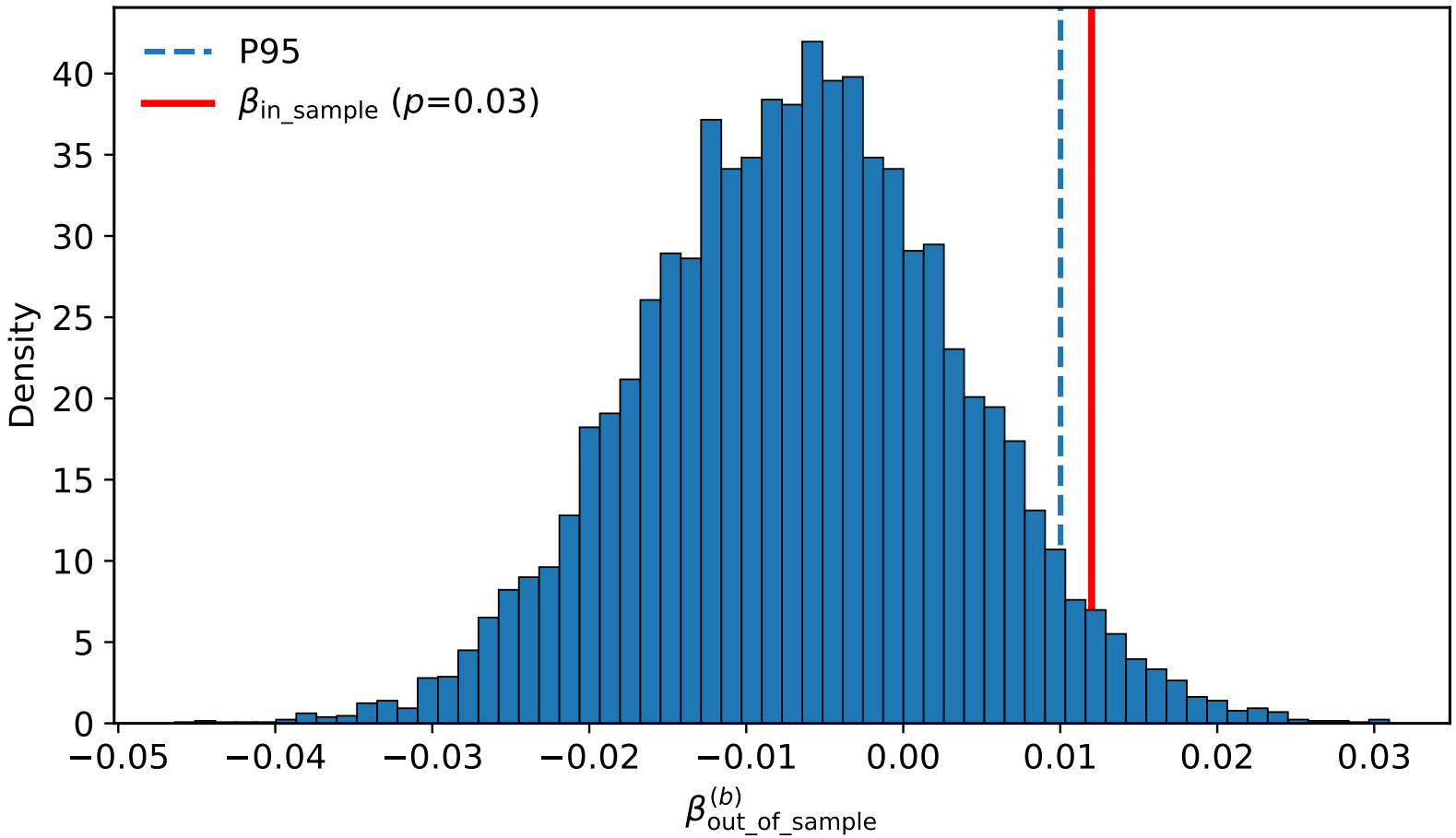

Figure: Bootstrap distribution of interaction coefficient β from out-of-sample data (10,000 replications). Blue dashed line: 95th percentile. Red solid line: in-sample estimate (p = 0.033).

Conclusion and Implications

Main findings:

1. LAP test reveals significant lookahead bias in LLM forecasts
2. Stock returns: 37% of apparent predictability from memorization
3. CapEx: 19% of effect attributable to memorization
4. Bias disappears in genuine out-of-sample periods

Implications:

- Lookahead bias is **task-specific**, not universal
- Depends on: data visibility, model architecture, prompt design
- LAP provides cost-efficient diagnostic (no retraining needed)
- Essential for validating LLM-based research in finance/economics

Takeaway

Distinguishing memory from reasoning is crucial as LLMs become integrated into empirical research