Discussion of *A Test of Lookahead Bias in LLM Forecasts*

by Zhenyu Gao, Wenxi Jiang, Yutong Yan

Jingyu He

ABFER Annual Meeting 2026

May 18, 2026

My Take: An Information-Set Audit Approach

I like the paper.

- It addresses a first-order problem in LLM-based empirical finance: the model's information set is hidden and not naturally point-in-time.
- A pre-trained LLM does not naturally respect the historical information set:

$$\mathcal{I}_t^{LLM} \neq \mathcal{I}_t^{investor}.$$

- The open question is not only whether contamination exists, but how much apparent predictability it explains.
- The key contribution is a simple diagnostic for asking whether apparent LLM predictability loads on memorized realized outcomes.

The Diagnostic: Date-Only Recall

Recall prompt

- Contains firm name, ticker, and target date or quarter.
- Asks whether the realized outcome went up, down, or is unknown.
- Contains no headline, transcript, fundamentals, or realized return.

Extract first-token probabilities

$$P_{up}, \quad P_{down}, \quad P_{unknown}.$$

Construct two statistics

$$\mathrm{LAP}_i = P_{up,i} + P_{down,i},$$

$$(U-D)_i = P_{up,i} - P_{down,i}.$$

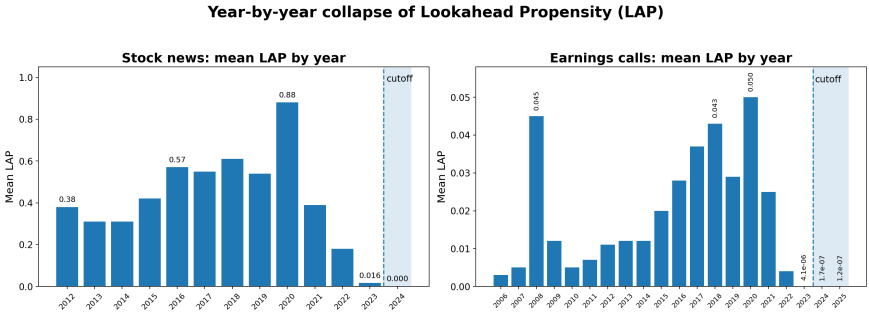
- LAP: strength of directional recall.
- $U - D$: direction of recalled outcome.
- High LAP: the model refuses to abstain.

Advantage

Simple. Works for any LLM that returns token-level probabilities; supported by most major APIs.

Most Persuasive Evidence: Cutoff Collapse

- Llama-3.3-70B has a December 2023 knowledge cutoff.
- Mean LAP is positive in pre-cutoff years and collapses after the cutoff.
- This is strong evidence that LAP is tied to training-time exposure.



Takeaway

The recall query appears to capture memorization exposure, not just random model noise.

Section 2.2: Framework and Test

Contamination model

$$Y_{t+1} = \mu(X_t) + \epsilon_{t+1}, \qquad \widehat{\mu}_t = \mu(X_t) + L_t \epsilon_{t+1}.$$

- ϵ_{t+1}: innovation relative to the historical real-time information set.
- L_t: latent memorization strength; $L_t = 1$ implies $\widehat{\mu}_t = Y_{t+1}$.
- Since L_t is unobserved, the paper uses $\mathrm{LAP}_t = P(up) + P(down)$ as its empirical proxy.

Detection regression

$$Y_{t+1} = \beta_1 \widehat{\mu}_t + \beta_2 \mathrm{LAP}_t + \beta_3 (\widehat{\mu}_t \times \mathrm{LAP}_t) + u_{t+1}.$$

- The paper focuses on β_3: is the LLM forecast more predictive when recall propensity is high?
- Proposition 1: $\beta_3 > 0 \Leftrightarrow E[L^2 \mathrm{Var}(\epsilon \mid \widehat{\mu}, L)] > 0$.

Main Empirical Findings

Stock news → returns

- Baseline LLM effect: 0.209% ($t = 12.18$).
- $\mathrm{LLM} \times \mathrm{LAP}$: 0.162 ($t = 3.64$).
- One-SD LAP increase raises the marginal LLM effect by about 32% of the standalone effect.
- Post-cutoff interaction is insignificant.

Earnings calls → capex

- Baseline LLM effect: 0.547 p.p. ($t = 16.69$).
- $\mathrm{LLM} \times \mathrm{LAP}$: 0.512 ($t \approx 2.01$).
- One-SD LAP increase raises the marginal LLM effect by about 12% of the standalone effect.
- Post-cutoff interaction is insignificant.

- Recall-only ($U - D$) predicts realized outcomes in high-LAP observations.
- LAP survives horse races with first-token inner confidence.
- The post-cutoff results show that the LLM signal can still predict outcomes, even when the LAP interaction disappears. → we should decompose predictability from look-ahead bias.

Comment 1: LAP as a Proxy, Not an Oracle

The key object of interest is not directly observed. A_i denote recall propensity, and C_i is actual memory contamination activated under prompt p_F.

$$A_i = \mathrm{LAP}_i(p_R), \qquad C_i = C_i(p_F),$$

where p_R is the date-only recall prompt and p_F is the actual forecasting prompt.

Proxy-validity condition

$$\mathbb{E}[C_i \mid A_i] \text{ is increasing in } A_i.$$

- This is plausible: strong date-only recall may indicate that the model has internalized the firm-date outcome.
- It's hard to interpret significant coefficient of interaction term $\Leftrightarrow$ contamination since we do not know how to quantify all uncertainty (output rely on prompt).

Suggestion: Reframe the main claim

Move from "direct identification test" to calibrated information-set diagnostic. This does not weaken the contribution: many important economic objects are measured with proxies.

Comment 2: Section 2.2 as a Measurement Framework

- ϵ_{t+1} is the innovation relative to the historical real-time information set:

$$\mathbb{E}[\epsilon_{t+1} \mid \mathcal{I}_t] = 0.$$

- Lookahead bias means the LLM forecast is not $\mathcal{I}_t$-measurable: post-event information may be encoded in the model's parameters.

Example: perfect memory

If the LLM fully remembers the realized outcome,

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

This is the clearest form of lookahead bias. But the slope of Y on $\widehat{\mu}$ can still be one, so the $\mathrm{LLM} \times \mathrm{LAP}$ interaction need not be positive.

Suggestion: Rewrite Section 2.2

Keep the intuition, but present the interaction as a diagnostic under explicit assumptions, not as an if-and-only-if theorem for contamination.

Comment 3: Does Recall Memory Enter the Forecast Prompt?

The key prompt-dependence issue

$$\underbrace{P_\theta(\mathit{up}/\mathit{down}/\mathit{unknown} \mid \mathit{firm}, \mathit{date})}_{\text{recall prompt: defines LAP}} \neq \underbrace{P_\theta(\mathit{good}/\mathit{bad}/\mathit{neutral} \mid \mathit{headline}, \mathit{firm}, \mathit{date})}_{\text{forecast prompt: used in regression}}.$$

- False positive: High LAP does not automatically imply the actual forecast prompt uses future memory.
- False negative: Low LAP does not rule out event-cued memory under the headline/transcript prompt.
- The paper does not need to solve prompt dependence mechanically, but it can validate prompt alignment empirically.

Suggestion: Direct output-level test

Let F_i^{full} be the original forecast and F_i^{masked} the forecast with firm identifiers masked.

$$\Delta F_i = F_i^{full} - F_i^{masked}, \qquad \Delta F_i = a + \lambda(U-D)_i + \kappa((U-D)_i \times \mathrm{LAP}_i) + FE + e_i.$$

If ΔF_i loads on recall direction, the recall channel affects the forecast prompt output itself.

Comment 4: Use Masking as Channel Tests

The paper already defines LLM^{masked}. It can be used as a mechanism test, not just a robustness check.

Forecast of LLM with different prompt

$$F^{full} = f(headline, firm, ticker, date),$$

$$F^{no\ date} = f(headline, firm, ticker),$$

$$F^{masked} = f(headline, masked\ firm, no\ date).$$

Expected pattern

- Full prompt: strongest $\mathrm{LLM} \times \mathrm{LAP}$.
- If LLM rely on date to recall memory, $F^{no\ date}$ should have weaker interaction. Same for masked.
- Masked prompt: interaction should fall if firm-date memory is the channel.

Suggestion: Move from outcome interaction to channel decomposition

Report whether the forecast output moves toward the date-only recall direction after controlling for masked headline semantics. This directly addresses "knows" versus "uses."

Comment 5: Separate Memorization from Salience

Potential confounding. High-LAP observations may also be high-salience observations.

- large firms, crisis periods, high prior volatility, high news volume, extreme attention events.

A stronger version of the test: remove predictable salience from LAP.

$$\widetilde{\mathrm{LAP}}_i = \mathrm{LAP}_i - \widehat{\mathbb{E}}[\mathrm{LAP}_i \mid \text{salience proxies, firm FE, time FE, event type}].$$

- Re-run the interaction using $\mathrm{LLM}_i \times \widetilde{\mathrm{LAP}}_i$.
- Estimate matched-sample specifications comparing high- and low-LAP observations within similar firm/event/salience cells.

Placebo firm-date probes.

$$\mathrm{LAP}_{i,t}^{\textit{wrong date}}, \qquad \mathrm{LAP}_{i,t}^{\textit{wrong firm}}, \qquad \mathrm{LAP}_{i,t}^{\textit{permuted}}.$$

- Only the true firm-date LAP should reproduce the interaction.
- This tests whether the result is firm-date-specific rather than generic attention.

Comment 6: From Detection to Correction

The paper's current takeaway is mostly diagnostic. A more useful next step is adjusted performance.

Clean marginal effect

$$\left.\frac{\partial Y}{\partial \mathrm{LLM}}\right|_{\mathrm{LAP}=0} = \widehat{\gamma}.$$

Use low-recall and post-cutoff results as conservative clean-text benchmarks.

Recall residualization

$$F^{full} = a + bF^{masked} + \lambda R + \kappa RA + e.$$

- Use e as a historical-audit sensitivity check, not as a tradable signal.
- It asks how much of the LLM forecast remains after removing the part that loads on date-only recall.

Suggestion: Memory-adjusted performance table

Report raw LLM performance, low-LAP/post-cutoff performance, LAP-weighted performance, and recall-residualized performance side by side.

Toward a Reporting Standard for LLM-Finance Papers

For any paper using modern LLMs to interpret historical financial text, the authors could recommend a standard disclosure table. For instance

1. Model version, knowledge cutoff, inference settings, and prompt template.
2. Baseline LLM performance.
3. Recall audit: LAP distribution, $(U-D)$ validation, and cutoff behavior.
4. Channel tests: LLM^{masked}, no-date/no-firm prompts, and forecast-output decomposition.
5. Salience robustness: residualized LAP, matched samples, and placebo firm-date interactions.
6. Conservative performance: low-LAP, post-cutoff, and memory-adjusted estimates.

Broader Implication

- The problem is not limited to return prediction.
- Any empirical design that asks a modern LLM to interpret historical firm-level text may use a model with a post-event information set.
- This includes sentiment, risk, tone, beliefs, managerial optimism, policy stance, and analyst narratives.

The paper has the potential to become a standard reference for auditing the point-in-time validity of LLM-generated regressors.

Summary

1. The paper tackles a first-order issue: **the hidden information set of foundation models**.
2. The evidence is promising: **recall predicts outcomes, high** LAP **amplifies LLM forecasts, and** LAP **collapses after the cutoff**.
3. The key improvement is conceptual precision: LAP **is a prompt-specific recall proxy, not a direct observation of forecast-prompt contamination**.
4. With prompt-alignment tests, salience controls, multi-probe calibration, and memory-adjusted performance reporting, the paper can become a **standard reference for point-in-time validity in LLM-finance research**.

Thank you!

Technical Issue in the Current Proposition 1 Proof

- FWL residuals are residuals from **linear projections**, not conditional expectation residuals, unless the authors explicitly define a nonparametric projection.
- Specifically, the proof is based on conditional expectation (call $W = \widehat{\mu}, L$):

$$\widetilde{Z} = Z - \underbrace{(W^T W)^{-1} W Z}_{\text{linear projection}} \neq Z - \underbrace{E[Z \mid \widehat{\mu}, L]}_{\text{conditional expectation}} .$$

  The proof implicitly replaces linear projection residuals with conditional-expectation residuals. That step is not valid without additional assumptions.
- If one does use conditional residuals, $L\widehat{\mu}$ is deterministic given $(\widehat{\mu}, L)$:

$$\mathbb{E}[L\widehat{\mu} \mid \widehat{\mu}, L] = L\widehat{\mu},$$

  so its conditional residual would be zero.
- The paper derivation replaces $L\widehat{\mu} = L\mu(X) + L^2\varepsilon$ with only the $L^2\varepsilon$ component. That step needs additional assumptions or a different projection argument.

A Simple Counterexample: Perfect Leak

Boundary case

Suppose the LLM fully remembers the future outcome: $L = 1$.Then $\widehat{\mu} = \mu + \epsilon = Y$.
This is the clearest form of lookahead bias: the forecast is the realized outcome itself.

But the interaction need not be positive

Let $L \in \{0, 1\}$ and suppose $\mu \perp \epsilon$, both with mean zero.
When $L = 0$:

$$\widehat{\mu} = \mu, \qquad \frac{\mathrm{Cov}(Y, \widehat{\mu} \mid L = 0)}{\mathrm{Var}(\widehat{\mu} \mid L = 0)} = 1.$$

When $L = 1$:

$$\widehat{\mu} = Y, \qquad \frac{\mathrm{Cov}(Y, \widehat{\mu} \mid L = 1)}{\mathrm{Var}(\widehat{\mu} \mid L = 1)} = 1.$$

So the slope can be the same across low- and high-L groups:

$$\beta_3 = 0,$$

even though the high-L group is perfectly contaminated.