Memorized but Not Realized: Frontier Language Models Precisely Recall Financial Fundamentals, but It Does Not Become Trading Skill

Anonymous authors
Paper under double-blind review

Abstract

LLM "AI-trading" benchmarks are routinely scored by replaying historical markets, raising a contamination concern: a model that has memorized an asset's realized outcome cannot generate honest out-of-sample skill, so backtest performance may be recall rather than foresight. We build a white-box-to-black-box validated leakage instrument and use it to ask the decision-relevant question prior work leaves open: does memorization capacity actually transmit to trading decisions? We report a dissociation. Capacity is large, precise, and universal: across 16 open models (10 vendors) and three commercial APIs, models prefer the realized revenue over magnitude-matched ($\pm 15\%$) counterfactuals at 0.23–0.81 (chance 0.20) and freely generate the figure within 10% for up to 89% of famous firm-years, while no-leak chronologically-consistent baselines sit at chance. Recall is cutoff-bounded: exploiting the models' different training cutoffs, it collapses to chance for facts dated after a model's cutoff (within-model $t = -6.6$, 18/19 models), the design-based signature of memorization rather than estimation. Applied to the realized return itself, the variable a backtest fears, recall is at chance with no cutoff signature: these models memorize fundamentals, not prices. Realization, by contrast, is null across five independent tests, including a cross-sectional ranking task ($N \approx 5{,}700$ per model) and a realistic StockBench-style backtest over six open and frontier-API agents; an apparent agentic signal (+9.9pp, $t = 4.07$) is shown to be a firm-prominence confound that vanishes under firm fixed effects ($\beta = -0.018$). A pre-registered, dependence-corrected power analysis bounds any residual post-cutoff skill: on a $\approx$2,000-name post-cutoff universe the minimum detectable mean |IC| is $\approx$0.012–0.016, below the 0.02–0.05 band where real strategies operate. The contribution is a validated leakage instrument, an identification result showing why memorization is invisible to any pre-cutoff backtest, and the finding that for the signals we test, LLM memorization of the market does not become measurable market skill.

1 Introduction

LLM trading agents are evaluated by replaying historical markets and reporting Sharpe ratios or information coefficients (ICs). If the pre-training corpus contains the realized outcomes, strong backtests can reflect recall of the answer rather than predictive skill (Glasserman & Lin, 2023). Lopez-Lira & Tang (2024) show that frontier LLMs recall exact pre-cutoff financial values. That establishes a capacity to leak; it does not establish that the capacity is realized as contaminated decisions, which is the quantity that actually inflates a backtest.

We separate the two and find that they come apart. The paper makes four contributions. First, a validated, transferable leakage instrument, calibrated on models with exactly known cutoffs (specificity), confirmed by a fine-tuning positive control (sensitivity), and deployed black-box on commercial APIs. Second, evidence that capacity is large and precise: it survives magnitude-matched distractors and free generation, so it is genuine memorization rather than order-of-magnitude estimation, and it concerns fundamentals rather than

prices. Third, the finding that realization is null across five designs: recall capacity does not make directional or cross-sectional trading decisions more accurate, an apparent agentic signal is a firm-prominence confound, and a realistic portfolio backtest shows no leakage advantage. Fourth, an identification result showing why this failure mode is invisible to any pre-cutoff backtest, with a dependence-corrected power bound on the residual. The headline is the dissociation: LLMs memorize the financial past with high precision, but for the signals we test that memory does not become skill, and only an external instrument can tell contaminated from clean.

2 Related work

Language models in finance and look-ahead bias. A fast-growing literature applies LLMs to return prediction and financial-statement analysis (Lopez-Lira & Tang, 2024; Kim et al., 2024). Because these models are trained on text that postdates the events they are tested on, Glasserman & Lin (2023) document look-ahead bias in GPT-based return predictions, and He et al. (2025) build chronologically-consistent models (ChronoBERT/ChronoGPT) trained only on text available as of a fixed date, which we adopt as our no-leak calibration target. Contemporary LLM-trading benchmarks attempt to neutralize contamination by construction, using post-cutoff or live data (Chen et al., 2025; ULAB, University of Illinois Urbana-Champaign, 2025). We instead supply an instrument that measures contamination on any data, including pre-cutoff data, and separate the capacity to leak from its realization as a decision.

Data contamination and membership inference. The benchmark-contamination literature shows that test data routinely enters pretraining corpora: Balloccu et al. (2024) find leaked data in 42% of papers evaluating closed models, and Sainz et al. (2023) call for per-benchmark contamination measurement (see the survey of Xu et al., 2024). Detection methods include reference-free membership inference (Min-K%; Shi et al., 2024), training-data extraction (Carlini et al., 2021), and quiz-style black-box probes (Golchin & Surdeanu, 2024), while Li & Flanigan (2024) document task contamination and Duan et al. (2024) caution that text-perplexity membership inference is unreliable at LLM scale. Our perplexity-dating null (§6.2) reproduces that caution in the financial-call setting and motivates a value-recall probe. Our instrument pairs this detection machinery with explicit specificity (exact-cutoff models) and sensitivity (a fine-tuning positive control) calibration, and asks the downstream question these methods do not: does detected memorization change behavior?

Backtest overfitting and the credibility of measured skill. In asset pricing, the multiple-testing and overfitting literature (Harvey et al., 2016; Hou et al., 2020; Bailey et al., 2016) establishes that in-sample performance is a weak guide to out-of-sample skill. Memorization is a distinct and sharper failure: the model does not overfit noise, it recalls the realized outcome. Our identification result (§3) makes precise why this failure is invisible to any pre-cutoff backtest, complementing dependence-corrected power analysis on the post-cutoff residual.

3 A framework: capacity, realization, and an identification result

Setup. A model forms a forecast $\hat{y}_{it}$ for asset i at date t. Decompose what it brings to bear into a generalizable predictive signal S_{it} and a memorized realized outcome M_{it}, where memory is available only for pre-cutoff observations:

$$\hat{y}_{it} = \alpha\, S_{it} + \beta\, M_{it} + \varepsilon_{it}, \qquad M_{it} = y_{it} \cdot \mathbf{1}[t \le \tau],$$

with τ the model's training cutoff and y_{it} the realized return. Capacity is the magnitude of M (does the model know y_{it} for $t \le \tau$?); realization is the coefficient β (does knowing it move the decision?). These are logically independent quantities.

Apparent skill and the identification obstruction. A pre-cutoff backtest scores $\mathrm{corr}(\hat{y}, y)$ over $t \le \tau$. There $M_{it} = y_{it}$ is perfectly collinear with the target, so the backtest fit is $\alpha\,\mathrm{corr}(S, y) + \beta$. Memorization enters with coefficient one and is observationally indistinguishable from genuine skill: any fit is consistent with

$(\alpha, \beta) = (\text{skill}, 0)$, with $(0, \text{leakage})$, or with any convex combination. No pre-cutoff backtest can therefore identify α separately from β. This is the familiar perfect-collinearity (omitted-variable) obstruction made precise for the leakage setting; we stress not its depth as a proof but what it pins down, namely exactly what an external instrument must supply. Identification requires either post-cutoff data ($t > \tau$, where $M = 0$) or an external instrument that measures capacity (M) and realization (β) directly. This paper supplies the latter: the value-recall instrument estimates M, and the transmission tests estimate β.

Testable predictions. (P1) Capacity rises with prominence (training exposure). (P2) Capacity is cutoff-bounded: $M_{it} \to 0$ for $t > \tau$. (P3) If $\beta > 0$, apparent skill falls at the cutoff and concentrates on recall-detectable cells. We confirm P1 (dose-response, §6.1) and P2 (the cutoff contrast, §6.3), and we estimate $\beta \approx 0$ across five designs (§6.5): capacity is large while realization is null.

4 Data and models

Backbone. A survivorship-free, point-in-time (PIT) equity backbone: CRSP daily prices with delisting returns, PIT Compustat fundamentals, and Polygon split-adjusted OHLCV held as a single price vendor across the cutoff frontier. A PIT earnings-event panel from Compustat `rdq` (301,917 announcements, 12,290 firms) uses a conservative session convention verified to have zero look-ahead. The within-firm transmission panel comprises 1,652 (firm, year) cells (large-cap, 2016–2023); the cross-sectional panel scales to 5,737 cells over 632 firms.

Ground truth. Annual revenue (Compustat `sale`) for a 16-firm famous-large-cap basket (176 firm-years; illustrative) and a 600-firm broad universe; the 632-firm cross-sectional panel and the broad universe carry the statistical weight. Realized annual returns come from the split-adjusted price series.

Models. No-leak baselines are ChronoGPT at yearly cutoffs 2014–2024. The open white-box roster is 16 models across 10 vendors (Table 1). The closed black-box arm uses GPT-5.1, Claude Opus 4.8, and Gemini 3.1 Pro. The cutoff RD and the agentic tests additionally use the instruction-tuned variants of four roster models, forming a 19-model cutoff panel (Gemma-4, which needs a separate toolchain, is in the capacity roster but not the RD panel).

5 The instrument and the validation cascade

For a fact (firm, year, realized value V) we score the model's preference for V over counterfactuals. Wide distractors $\{0.5, 0.7, 1.4, 2.0\} \times V$ are partly eliminable by order of magnitude; tight log-matched distractors $\{0.85, 0.93, 1.08, 1.18\} \times V$ are not, so tight recall isolates precise memorization. We also score free generation, in which no options are shown and the model writes the number (scored within 5/10/20%). White-box, we rank candidates by summed log-probability; black-box, by multiple choice. A credible instrument must be specific (no false positives where memorization is impossible) and sensitive (it fires where leakage is known to be present); we establish both before transfer.

6 Results

6.1 Capacity is large and precise

Frontier models prefer the realized revenue far above chance, and they survive the magnitude-matched tight distractors and free generation (Figures 1,2). Tight ±15% recall is 0.23–0.81 against a 0.20 chance level; free generation within 10% is 0.38–0.89. Gemma-4 is the capstone, at 0.96 wide, 0.81 tight, and 0.89 generation. The gap between recognition (MCQ) and free generation is small for capable models (Qwen2.5-14B, 0.90 versus 0.78), so recognition does not greatly overstate recall. A dose-response in scale holds (Qwen3-4B tight 0.23; Qwen2.5-14B 0.70; Gemma-4 0.81). The recall is therefore genuine precise memorization rather than competent order-of-magnitude estimation, which answers the most serious confound. The wide-to-tight drop

Table 1: Open-model roster (white-box). † marks a vendor-disclosed training cutoff (from the official model card or technical report); "n.d." marks cutoffs that are *not* officially disclosed, for which we give the year implied by release date and community probing. Notably, most open models do not disclose a precise cutoff—itself a reason an external instrument is needed. The cutoff analysis does not rely on exactness: because annual fundamentals are reported the following year, FY≤2022 values are unambiguously in-training and FY2025 values unambiguously post-training for every model in the 2023–2025 cutoff range, which is why we use the all-seen vs. all-unseen contrast rather than a sharp per-model discontinuity. ⋆ marks models whose instruction-tuned variants are additionally used in the cutoff RD and agentic tests (the 19-model cutoff panel).

Model	Vendor	Size	Training cutoff
Llama-3.1-8B⋆	Meta	8B	Dec 2023†
Mistral-7B-v0.3	Mistral AI	7B	2023 (n.d.)
Mistral-Nemo-12B⋆	Mistral AI	12B	2024 (n.d.)
Qwen2.5-7B⋆	Alibaba	7B	2024 (n.d.)
Qwen2.5-14B⋆	Alibaba	14B	2024 (n.d.)
Qwen3-4B	Alibaba	4B	2024 (n.d.)
Qwen3-8B	Alibaba	8B	2024 (n.d.)
Qwen3-14B	Alibaba	14B	2024 (n.d.)
Gemma-3-4B	Google	4B	Aug 2024†
Gemma-4-12B	Google	12B	2025 (n.d.)
GLM-4-9B	Zhipu AI	9B	2024 (n.d.)
Yi-1.5-9B	01.AI	9B	2024 (n.d.)
OLMo-2-13B	Allen AI	13B	2024 (n.d.)
Falcon3-10B	TII	10B	2024 (n.d.)
Granite-3.1-8B	IBM	8B	Apr 2024†
Phi-4	Microsoft	14B	Jun 2024†

of 0.20–0.36 confirms that the wide distractors did overstate, so we treat the tight and generation numbers as load-bearing.

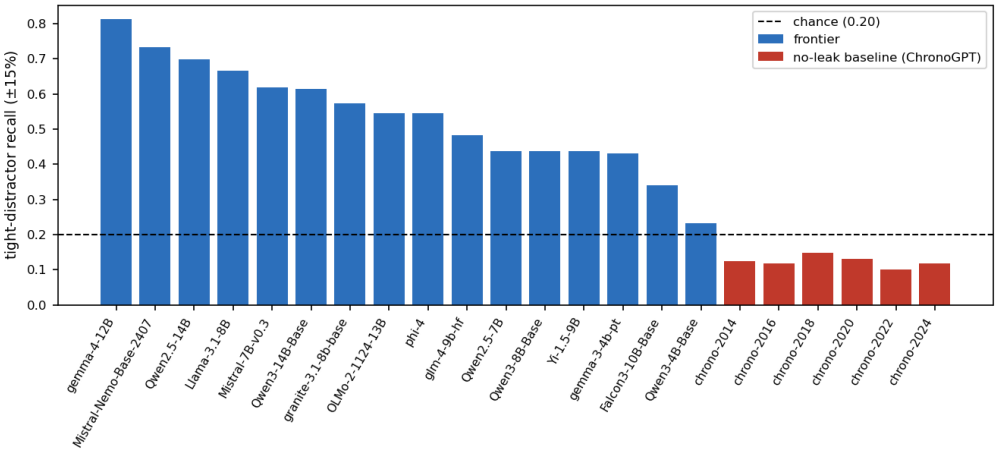

Figure 1: Precise capacity: magnitude-robust tight (±15%) recall. Frontier models (blue) far exceed the 0.20 chance line; no-leak ChronoGPT baselines (red) sit at chance.

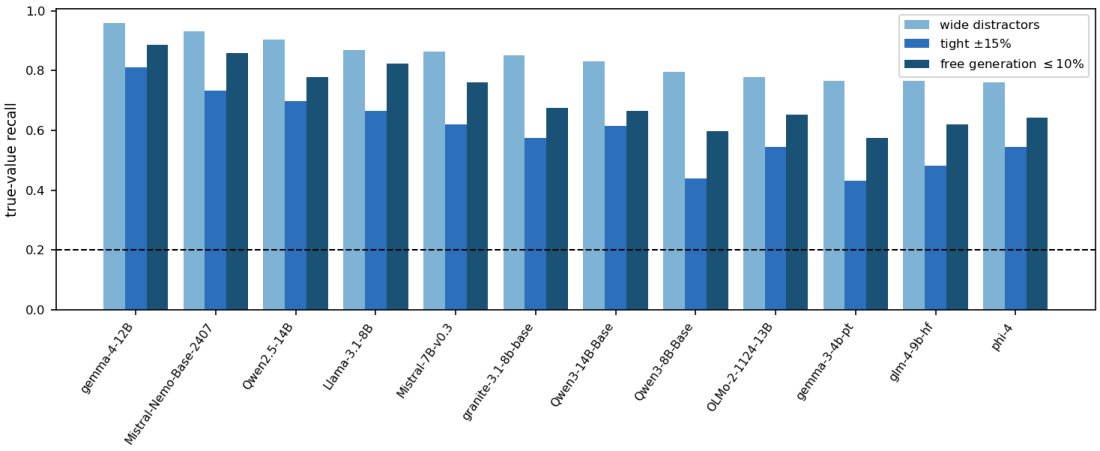

Figure 2: Memorization versus estimation. Recall survives both tight (magnitude-uninformative) distractors and free generation, not only the eliminable wide distractors.

6.2 Specificity, sensitivity, and closed-API transfer

ChronoGPT, whose cutoffs are provable, recalls at or below chance with zero free-generation hits, so the instrument does not false-positive; a text-perplexity dating probe is null, confirming that generic text-based membership inference is the wrong instrument here. A LoRA fine-tuning positive control injects known FY2024 values into 400 TRAIN firms while holding out 400 CONTROL firms: the instrument then fires at 99–100% on TRAIN and remains at baseline on CONTROL (difference-in-differences 0.75–0.85, unanimous across seven vendors; Figure 3), establishing sensitivity. Deployed black-box, the validated instrument reads large recall on all three commercial APIs (Gemini 3.1 Pro 100%, Claude Opus 4.8 94.9%, GPT-5.1 58.5%; chance 20%). We read these cautiously: FY2013–2023 revenue is near-trivially present in any 2025–26 model's training and in many public databases, so the inter-model gap reflects response and parsing policy rather than a leakage ranking. They confirm capacity; they do not order models.

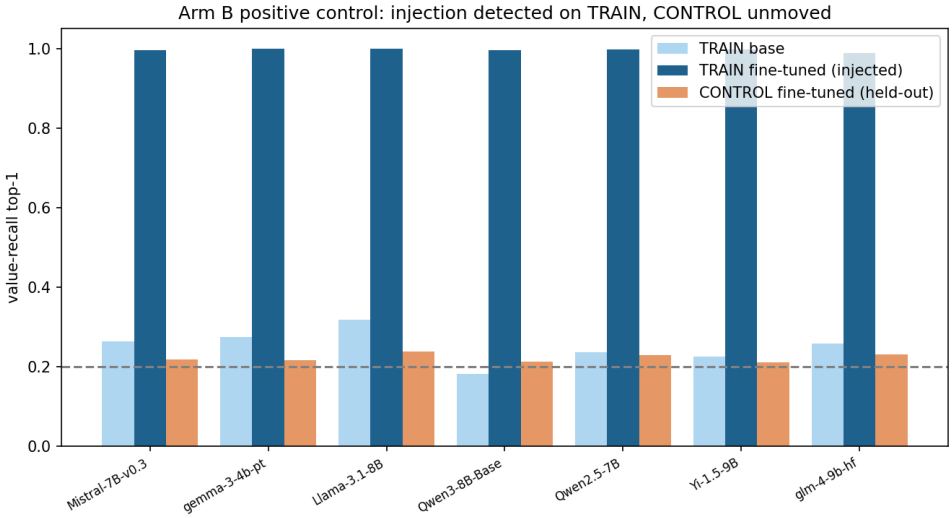

Figure 3: Positive control. Injecting TRAIN-firm values by fine-tuning makes the instrument fire on TRAIN but not on held-out CONTROL, establishing sensitivity.

6.3 Cutoff-bounded recall

Because the models have different known cutoffs, we can ask whether recall depends on training membership rather than estimability. Within each model we compare recall on facts that are pre-cutoff for all models (FY$\leq$2022, "seen") against facts post-cutoff for all (FY2025, "unseen"). Recall falls from 0.313 to 0.185 (about the 0.20 chance level): within-model $\beta = -0.128$, $t = -6.6$, with 18 of 19 models dropping (Figure 4). Recall thus tracks whether a fact was in training, not how estimable it is, which is a design-based rebuttal of the estimation confound, since competent estimation would not collapse at the cutoff. We report the sharp test transparently: a local-linear regression discontinuity of recall on event-time has a null jump at the cutoff (+0.02, $t = 1.0$, $p = 0.31$), because the frontier cutoffs cluster at 2023–24 and are themselves approximate, blurring the boundary. The all-seen versus all-unseen contrast compares the extremes and is correspondingly stronger; it does not isolate the cutoff from calendar recency, but the recency alternative is implausible for recall, since post-cutoff facts are absent from training and cannot be recalled regardless of recency. We therefore read the evidence as a fuzzy cutoff-bounded recall, a large seen/unseen gap without a sharp discontinuity at an approximate boundary, rather than over-claiming a clean jump.

6.4 What is memorized: fundamentals, not prices

A natural objection is that we measure recall of a fundamental (revenue), whereas the contamination a backtest fears is memorized prices. We therefore apply the instrument directly to the realized annual return, in two ways and across 18 models: the recalled sign, and a counterfactual over the magnitude. The contrast is sharp (Figure 5). Whereas revenue recall is well above chance and cutoff-bounded, return recall is neither. Recalled direction is 0.64 for both seen and unseen years, which is the market's up-rate with no seen/unseen gap, and magnitude recall sits at chance (0.14 against a 0.17 baseline), again flat across the cutoff. Frontier models memorize firms' fundamentals precisely but do not memorize their realized returns. This answers the construct-validity objection directly and supplies a mechanism for the null that follows: the contamination most feared, memorized prices, is largely absent, and the fundamental memory that does exist does not concern the trading target.

6.5 Realization is null

Capacity is necessary but not sufficient. We test realization five ways and find a consistent null. (i) *Cross-model.* The correlation between capacity and realized directional leakage (a bias-immune AUC on memoris-

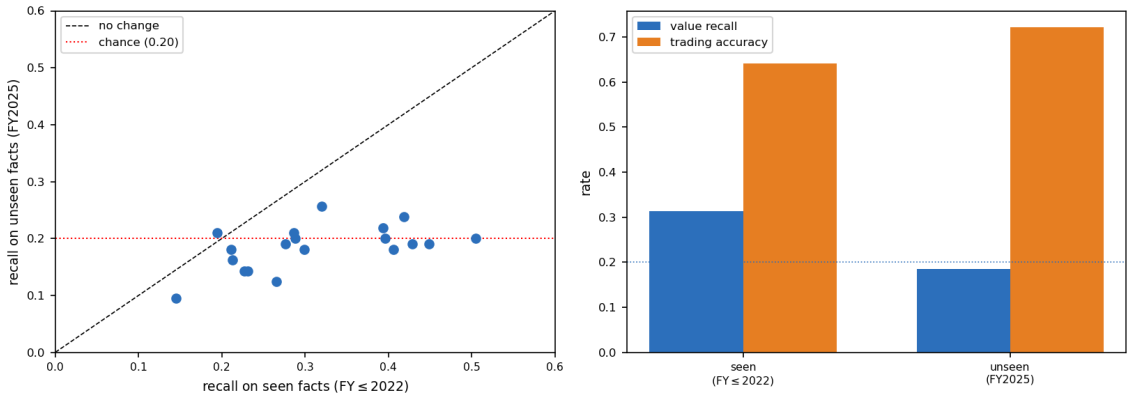

Figure 4: Cutoff-bounded recall. Left: per-model recall on unseen (FY2025) versus seen (FY≤2022) facts; 18 of 19 models fall toward the chance line. Right: pooled recall drops to chance across the cutoff while trading accuracy does not benefit.

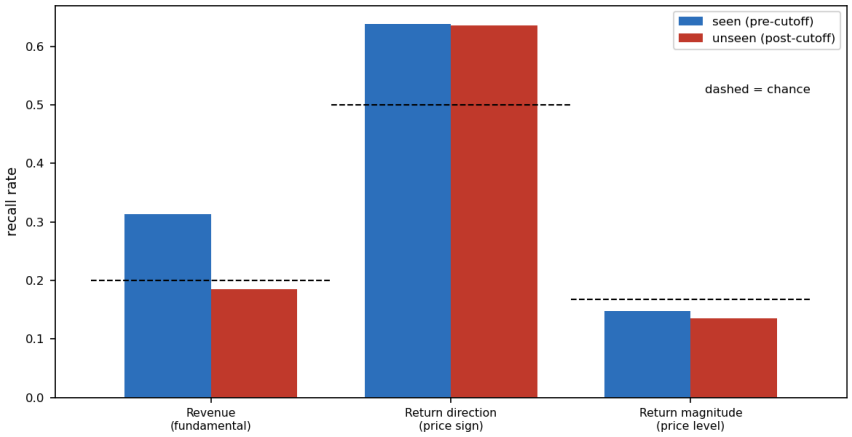

Figure 5: Models memorize fundamentals, not prices. Only revenue recall is above chance and cutoff-bounded (seen > unseen); return-direction recall equals the market up-rate with no seen/unseen gap, and return-magnitude recall is at chance.

able years, minus 0.5) is 0.044 ($N = 14$, bootstrap CI $[-0.49, +0.67]$); the larger +0.62 we first measured was an artifact of anchoring on the no-leak baselines. (ii) *Within-firm decomposition.* Regressing per-decision correctness on a recall flag with firm and year fixed effects (base models, $N = 1{,}652$) yields approximately zero. (iii) *Agentic.* Instruction-tuned models make committed directional forecasts. Most default to the up-market base rate, but the one model that genuinely varies (Qwen2.5-7B-Instruct, "higher" only 28% of the time) shows an apparent year-fixed-effect effect of +9.9pp ($t = 4.07$). It is a confound: recall-detectable cells are prominent mega-caps that were independently more predictable in a rising market, and adding firm fixed effects (identifying within a firm, across years) removes it ($\beta = -0.018$, $t = -0.51$, $p = 0.61$, firm-clustered SE; Figure 7). (iv) *Cross-sectional ranking.* To rule out that realization lives in relative ranking rather than absolute direction, and to remove the base-rate default, we use a balanced task ("did firm i outperform the median large-cap that year?"; ground truth versus the cross-sectional median, hence about 50/50). At $N \approx 5{,}700$ per model the models sit near chance (accuracy 0.52–0.55) and recall does not robustly predict accuracy under firm and year fixed effects: the effect is bounded to $|\beta| < 0.04$ (two of three models $p > 0.3$; Llama-3.1 marginal at $\beta = +0.033$, $p = 0.07$, within multiple-comparison noise and economically trivial on a near-chance task). (v) *A realistic agentic backtest.* To address the concern that single-period probes are unrealistic, we run a StockBench-style monthly long–short portfolio over 30 liquid large-caps (2018–2025, 92 months) with PIT context, scored on annualised Sharpe and decomposed by pre/post-cutoff and recall-detectability, across six agents: three open instruction-tuned models and three deployed frontier APIs (GPT-4o-mini, Claude Haiku 4.5, Gemini 2.5 Flash). If memorization realized as P&L, pre-cutoff Sharpe would exceed post-cutoff and recall-detectable picks would beat recall-miss systematically across agents. Neither holds: the sign of each difference splits evenly (3/6 each way) and recall-detectable picks underperform on average (mean ΔSharpe -0.06), the signature of market-regime noise rather than leakage (Figure 6). The fixed-effect tests, which absorb regime, carry the clean null; the backtest corroborates it in the realistic setting the field actually scores, on the very models it deploys. Across all five designs, on open and frontier-API agents alike, memorization capacity does not transmit to measurable trading skill.

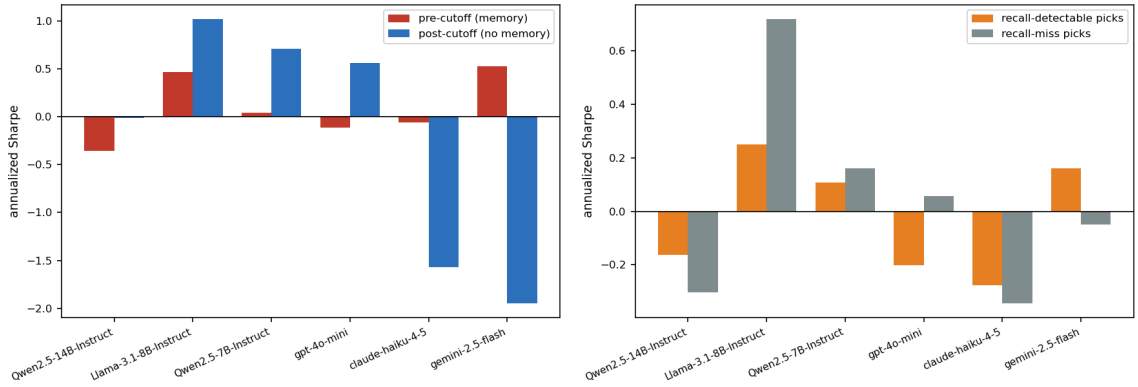

Figure 6: Realistic agentic backtest, Sharpe-decomposed, across six agents (three open and three frontier-API, shaded). Left: pre-cutoff is not systematically better than post-cutoff (sign splits 3/6). Right: recall-detectable picks do not outperform recall-miss (mean ΔSharpe -0.06).

6.6 The skill residual: a bound, not an absence

A pre-registered, dependence-corrected power analysis bounds any undetected post-cutoff skill. Effective breadth is taken from the residual (market-removed) cross-sectional correlation and the time series is Newey-West corrected. On the full post-cutoff window (2024 onward, ≈540 trading days) and a liquid universe of ≈2,000 names, the minimum detectable mean |IC| is ≈0.012 (i.i.d.) to 0.016 (conservative ϕ=0.25 Newey-West; Figure 8). Crucially this sits *below* the 0.02–0.05 band in which real cross-sectional strategies operate, so the uniform post-cutoff null now excludes economically meaningful effects rather than merely failing to detect them. Effective breadth saturates near $N_{\text{eff}} \approx 50$: because names co-move, adding firms past ∼1,000 barely moves the bound (the longer window does the work), which is why we use the full post-cutoff window

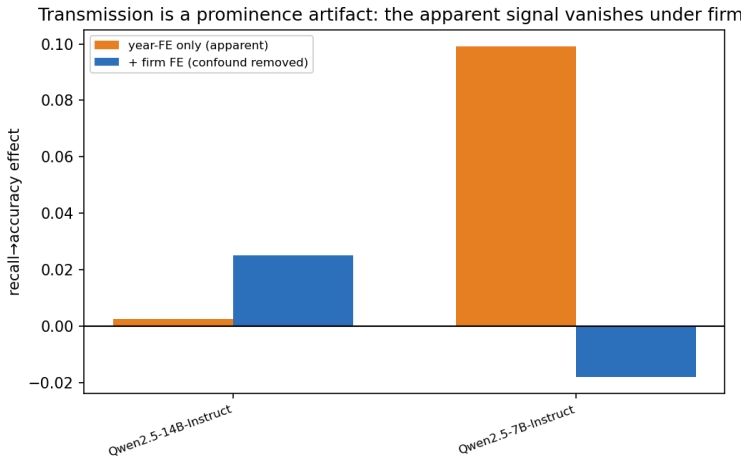

Figure 7: The apparent agentic transmission is a firm-prominence artifact: the year-fixed-effect-only effect (orange) collapses to zero once firm fixed effects are added (blue).

Table 2: Realization is null across five designs; no test finds that recall capacity raises a trading outcome. "Cross-model" correlates capacity with realized directional leakage across the 14 frontier models for which both a capacity and a directional-AUC estimate exist; the within-firm, agentic, and cross-sectional tests regress decision correctness on a recall flag with firm and year fixed effects; the backtest compares Sharpe by pre/post-cutoff and recall-detectability across six agents.

Test	Estimand	Result
(i) Cross-model	corr(capacity, leakage)	$r = 0.044$ (N=14, CI $[-0.49, 0.67]$)
(ii) Within-firm decomposition	accuracy $\sim$ recall + FE	$\beta \approx 0$ (N=1,652)
(iii) Agentic + firm FE	accuracy $\sim$ recall + FE	$\beta = -0.018$ (t= − 0.51, p=0.61)
(iv) Cross-sectional ranking	accuracy $\sim$ recall + FE	$\lvert\beta\rvert < 0.04$ ($N\approx$5,700; 2/3 p>0.3)
(v) Realistic backtest	Sharpe by cutoff / recall	no systematic gap (mean ΔSharpe -0.06)

rather than a wider cross-section. We still state this as a power-conditional bound, not an absolute absence claim.

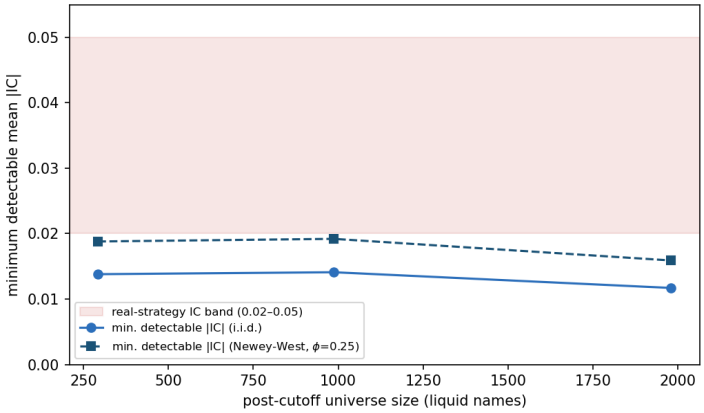

Figure 8: Dependence-corrected minimum detectable mean |IC| versus the naive independence assumption (dashed).

7 Discussion

The dissociation reframes the contamination debate. Frontier LLMs do memorize the financial past with high precision and universally, so the instinct to worry is correct in kind. But that memory does not, for the directional and cross-sectional signals we test, become realized skill; the firm-fixed-effect result shows how readily a prominence artifact masquerades as transmission ($t = 4.07$ before, $t = -0.51$ after). Naive trust is equally unwarranted, because the capacity is unambiguous. The practical upshot is that neither blanket suspicion nor blanket acceptance of an LLM backtest is defensible: one needs an instrument that separates capacity from realization.

Limitations. The null is bounded by power. The expanded post-cutoff analysis places the minimum detectable mean |IC| at ≈0.012–0.016, below the band where real strategies operate, so the null excludes economically meaningful effects; we still cannot exclude arbitrarily small ones, and report non-detection conditional on power rather than absence. The return-recall result shows these models do not memorize prices, which both answers the construct-validity concern and narrows the claim to the variables tested. The agentic backtest is a single realized path over 30 names, so its Sharpe levels are noisy even though the cross-agent pattern is clear. A frontier-scale chronologically-consistent control, rather than the smaller ChronoGPT, would further separate memorization from reasoning.

8 Conclusion

We deliver a validated leakage instrument and use it to establish a dissociation: large, precise, universal pre-cutoff recall of fundamentals that does not transmit to measurable trading skill across five designs. For the literature scoring LLM traders on historical data, the message is not that it is all memorization; it is that memorization and realized skill must be measured separately, because the first is large and the second, here, is not detectable. We release the instrument and its validation cascade as a reusable certification standard. By the identification result of §3, no pre-cutoff backtest can be trusted on its face, so an LLM-trading result should report its realized-leakage share, measured by an external instrument, before its skill claim is taken at face value.

Reproducibility statement

The instrument, the validation cascade (specificity and sensitivity), all five transmission designs, the agentic backtest, the dependence-corrected power pre-registration, and every figure are released as an open artifact under an MIT license, provided as anonymized supplementary material for review. The data backbone (CRSP, Compustat, Polygon) is licensed and not redistributable; the artifact documents the exact tables and the point-in-time construction so the panels can be reconstructed from a standard WRDS subscription, and includes the aggregate result files behind every reported number.

Broader impact statement

The work is methodological and aimed at improving the integrity of LLM evaluation. Its intended effect is positive: a tool that lets practitioners and researchers distinguish recalled hindsight from genuine skill before acting on an LLM trading result. We see no significant risk of harm; if anything, the instrument guards against the over-claiming that an uncorrected backtest invites.

References

David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. The probability of backtest overfitting. *Journal of Computational Finance*, 20(4):39–69, 2016.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of EACL*, 2024.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, et al. Extracting training data from large language models. In *30th USENIX Security Symposium*, 2021.

Yanxu Chen et al. StockBench: Can LLM agents trade stocks profitably in real-world markets? *Working paper (arXiv:2510.02209)*, 2025.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, et al. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024.

Paul Glasserman and Caden Lin. Assessing look-ahead bias in stock return predictions generated by large language models. *Working paper (arXiv:2309.17322)*, 2023.

Shahriar Golchin and Mihai Surdeanu. Data contamination quiz: A tool to detect and estimate contamination in large language models. *Working paper (arXiv:2311.06233)*, 2024.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016.

Songrun He, Linying Lv, Asaf Manela, and Jimmy Wu. Chronologically consistent large language models. *Working paper (arXiv:2502.21206)*, 2025.

Kewei Hou, Chen Xue, and Lu Zhang. Replicating anomalies. *Review of Financial Studies*, 33(5):2019–2133, 2020.

Alex Kim, Maximilian Muhn, and Valeri Nikolaev. Financial statement analysis with large language models. *Working paper (arXiv:2407.17866)*, 2024.

Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Alejandro Lopez-Lira and Yuehua Tang. Can ChatGPT forecast stock price movements? return predictability and large language models. *Working paper (arXiv:2304.07619)*, 2024.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, et al. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of EMNLP*, 2023.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

ULAB, University of Illinois Urbana-Champaign. LiveTradeBench: Seeking real-world alpha with large language models. *Working paper (arXiv:2511.03628)*, 2025.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey. *Working paper (arXiv:2406.04244)*, 2024.