

Beta Recall, Alpha Recall, and a Contamination Detector That Needs No Labels*

Measuring training-data leakage in LLM equity signals

Bach Nguyen

Imperial College Business School
London, United Kingdom
bachngn169@gmail.com

Abstract

A large language model (LLM) asked to rank stocks on a historical date may be forecasting or it may be recalling the outcome. Inside its training window the two are observationally identical. We measure the gap on a pre-registered monthly grid of 50 US large caps and report it in the units a practitioner reads. For Llama 3.3 70B (documented cutoff 2023-12), the cross-sectional information coefficient (IC) of a from-memory recall probe is +0.166 inside the training window against -0.019 outside it, a gap of +0.185 (moving-block bootstrap $p = 0.017$). A dollar-neutral long-short book on the same signal shows an annualised Sharpe ratio of +2.14 in window and +0.40 out.

Three results complicate the usual reading of a gap like that. The leakage is cross-sectional rather than directional: the model's month-level direction call is 0.708 accurate before the cutoff and 0.680 after it, so market-neutral construction strips out the component that does not leak and keeps the one that does. A pre-cutoff advantage is not on its own evidence of recall, because a standing tilt across the cross-section that happened to pay leaves the same trace. One stretch of this model's training window is exactly that, and its apparent IC falls to +0.001 once each name's own average is removed within a stretch where the model behaves consistently. Contamination also leaves a mark in the responses by themselves: month-to-month repetition of the model's own ranking rises from +0.052 in window to +0.411 out, and a changepoint search on that series puts a boundary at 2022-12 ($p = 0.001$) using no prices, no returns and no labels.

The contribution is a measurement protocol and two label-free detectors, demonstrated on one primary subject. We do not claim a general law about language models, and we do not propose a trading strategy.

1 Introduction

Backtests of LLM-generated equity signals are run on historical data because that is the only data there is. This creates a problem with no analogue in conventional quantitative research. A linear model estimated on a training sample can be evaluated out of sample by holding dates back. A language model cannot, because the held-back dates may already be inside the corpus it was trained on, and the researcher has no way to inspect that corpus.

We treat this as a model-risk question rather than a forecasting one. The quantity of interest is not whether an LLM can predict

returns. It is how much of a measured backtest result would survive if the model had never seen the evaluation period, and whether that can be established without trusting the vendor's cutoff date.

The protocol scores a recall probe across the cross-section, month by month, on a pre-registered grid, and splits the months by each subject's documented training cutoff. Measured that way the leakage is large, and it is large in the units a desk would quote. For the primary subject the cross-sectional IC is +0.166 in window against -0.019 out, and the implied dollar-neutral book runs at an annualised Sharpe ratio of +2.14 against +0.40. Against a null that re-pairs months while holding both marginals fixed, the in-window Sharpe sits at $p = 0.003$.

That number needs three qualifications, and each one changes what a practitioner should do about it.

The first is that the leakage does not sit where intuition puts it. Decomposing an answer into the month's average and the ordering within the month separates two quantities that behave differently across the cutoff. The month-level direction call barely moves, 0.708 against 0.680, which makes it a standing prior about equity markets rather than a memory of any particular month. The ordering is what collapses. A dollar-neutral book holds the ordering and discards the level, so market-neutral construction offers no protection here; it concentrates the exposure.

The second is that a pre-cutoff advantage does not on its own establish recall. A model with a fixed preference for part of the cross-section scores a high IC in any period where that preference paid, with no month-specific memory at all. Separating the two requires removing each name's own average within a stretch over which the model behaves homogeneously, and the choice of stretch decides the answer (Section 4.2).

The third is that contamination turns out to be detectable with no realised returns at all. Asked about a month it has coverage for, the model produces a different ranking from month to month; asked about a month it does not, it repeats a standing template. Measuring that repetition needs no prices, so the statistic can be swept across the calendar to locate a boundary rather than assume one. That is what makes a vendor's stated cutoff checkable rather than trusted.

A word on how far any of this reaches. The full pre-cutoff and post-cutoff contrast rests on a single subject. A second subject, openai/gpt-oss-120b, refuses to answer for 100% of its post-cutoff months, which costs it every out-of-window observation and supplies a third detector in their place. The staggered-cutoff design registered for this study turns out not to be identified on the subjects available, and Section 9 says why rather than working

*Code, cached model responses and every result artefact: <https://github.com/hoangbach619/llm-alpha-leakage>. Each number in this paper regenerates from that repository offline, with no API key.

around it. What follows is a protocol and one demonstration of it, not a characterisation of language models in general.

2 Related work

Didisheim et al. [4] identify look-ahead bias by prompting a model to recall a specific realised value, a stock return, with no context, and build a memorisation proxy from that retrieval. They report that the bias is more severe at lower data frequencies, for indexes, and for larger models, and that it is insignificant for small models at high frequencies.

Their probe is, in substance, ours. We replicate the retrieval-based memorisation probe and do not claim it. What is new here is what is done with the responses. The probe is scored across the cross-section, as a monthly ranking of 50 names rather than one firm at a time, which is what exposes the level call and the cross-section as separate quantities (Section 4.1). That ranking is translated into a dollar-neutral book and a Sharpe ratio, so the leakage is stated in the units a backtest reports. The pre-cutoff window is split at a boundary estimated from the model’s own behaviour rather than taken from the vendor. Two of the detectors need no realised outcomes at all, which nothing scored against returns can offer.

This regime is also the one where they report the bias is worst. Monthly data is low frequency, and a 70B-parameter model is a large model, which is consistent with the size of the in-window effect reported below and is a reason to read it as an upper bound rather than a typical case.

López-Lira and Tang [14] report that headline-level scores from a language model predict subsequent returns; the current version of that work uses post-cutoff headlines specifically to avoid contamination. The concern was made precise by López-Lira et al. [15], who state the identification problem directly: once a model has seen the realised values, any output is consistent with both skill and memorisation, so counterfactual forecasting ability is non-identified. They demonstrate exact recall of economic and financial values before the cutoff, none after it, and report that instructions to respect historical boundaries fail and that masking fails because models reconstruct entities and dates from minimal context. The design here takes their non-identification result as its premise.

Sarkar and Vafa [17] develop direct tests for look-ahead bias that turn on events being unpredictable relative to a prespecified information set, and find it when predicting risk factors from earnings calls and election winners from biographies. Their tests still require an outcome to be predicted; the detectors in Section 6 do not.

Entity masking deserves a note, because it is the obvious ablation and it is weaker than it looks. Glasserman and Lin [7] separate look-ahead bias from a distraction effect, in which knowledge of the named firms interferes with reading a text’s sentiment, and find that anonymised headlines do better rather than worse, which they read as the distraction effect dominating look-ahead bias. López-Lira et al. [15] independently report that masking fails. Two results in opposite corners of the literature therefore say the same thing: a change in performance under masking is not by itself evidence about look-ahead bias. The masking ablation is implemented here and deliberately not run as evidence. The counterfactual-date test

of Section 5 stands in its place, because it leaves entity knowledge untouched and breaks only the date.

Four recent efforts sit closest to the instruments used here, and the distinctions are narrow enough to be worth stating precisely. Gao et al. [6] estimate a Lookahead Propensity from a date-only recall query per firm-date and watch it collapse to zero just after the cutoff, then use it as a regressor in a forecast-accuracy regression. That test needs realised outcomes, and their boundary falls sharply at the documented cutoff where ours sits a year early, which Section 4.3 reconciles. Eliseev and Seleznev [5] perturb the date inside the prompt and re-query, in macroeconomic forecasting. The perturbation here happens after the fact, on responses already cached, which costs nothing and preserves both marginals. Benhenda [2] benchmarks look-ahead bias through alpha decay across market regimes, measured against point-in-time models, where the object here is to locate a boundary inside one model rather than to rank models against each other. Li et al. [12] document agent backtest returns evaporating past the knowledge window and apply counterfactual perturbation as a training mitigation, which is the same instrument put to a different end: theirs trains, ours diagnoses.

Mitigation is the other response. He et al. [10] train ChronoBERT and ChronoGPT on text available only up to each point in time, and report that in a news-to-returns application their real-time outputs match a much larger Llama model, which they read as look-ahead bias being modest in that setting. That is a different application and a different probe from ours, and the contrast is worth stating: bias measured through a forecasting pipeline can be small while bias measured through direct retrieval is large.

The closest methodological neighbours sit outside finance. Membership-inference and extraction work asks whether a specific example was in the training set [3, 18]. The template-collapse statistic in Section 6 is related in spirit and differs in what it needs: it requires no candidate example, no reference model and no likelihoods, only repeated queries whose answers can be compared with one another.

On the finance side, the concern that a reported backtest reflects search rather than signal is old [1, 9], and the machine-learning asset-pricing literature has been careful about it [8]. Contamination is a different failure. The researcher who runs one LLM backtest once, with no specification search at all, can still obtain a Sharpe ratio above two from a model that is reciting the answer.

3 Design

3.1 A shared grid

Every subject is scored on one pre-registered monthly grid: 50 liquid US large caps, 2021-10 to 2026-04, fixed before any subject was run. One grid across all subjects is what makes any two of them comparable: give each subject its own list of dates and any difference between two models is confounded with the difference between the periods each was scored on. Each subject’s pre-cutoff and post-cutoff labels are derived from its own documented cutoff, with 3 months either side excluded as ambiguous. Vendors document cutoffs to month precision at best, and commentary about a month keeps being written for months afterwards, so the boundary is a band.

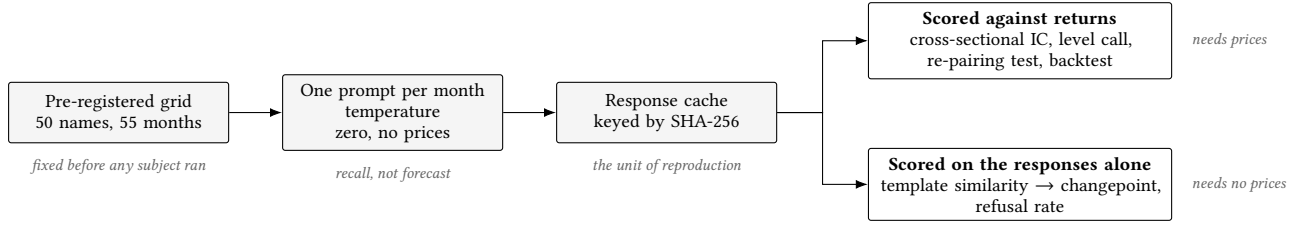


Figure 1: The protocol. One shared grid and one cached set of responses feed two families of test. The upper family is what a conventional evaluation runs and it needs realised returns; the lower reads the responses against one another, which is what lets the boundary be estimated rather than assumed.

3.2 The probe

At each grid month the model is asked, at temperature zero, to recall the total return of each of the 50 names over that calendar month, from memory, with no price data in the prompt. One call covers the cross-section. The task requires no forecasting ability, so a null result cannot be attributed to the model lacking skill.

Responses are cached on disk under the SHA-256 hash of the model id, the prompt and the system prompt. The cache is the unit of reproduction: every number in this paper regenerates from it offline with no API key.

3.3 Scoring

Each month’s answer is scored as a cross-sectional Spearman rank correlation against realised monthly returns. Spearman rather than Pearson because the model’s stated percentages are invented and their scale carries no information. The scoring uses two demeanings. The first removes the month’s mean from both sides, leaving the dollar-neutral bet. The second also removes each name’s own average within a segment, leaving only month-specific deviation.

Inference on every gap is a moving-block bootstrap over months [11] with block length 3 and 10,000 replicates, seeded. Monthly information coefficients are serially dependent, so an independent-draw interval would come out too narrow. A Welch t runs alongside it, weighting every month equally, and is never the primary inference.

Randomisation p -values use the add-one estimator of Phipson and Smyth [16], $\hat{p} = (1 + \#\{T_b \geq T_{\text{obs}}\}) / (1 + N)$ with $N = 10,000$. The naive proportion can return zero, which asserts more than any finite resampling scheme can support; clamping it afterwards produces a floor that reads as a point estimate. Values at the estimator’s resolution are printed as inequalities throughout.

4 Cross-sectional leakage

Table 1 reports the headline contrast.

For the primary subject the gap is +0.185 with a bootstrap interval of [+0.030, +0.329] and $p = 0.017$, on 24 pre-cutoff and 25 post-cutoff months.

4.1 The level call is not the memory

Decomposing the answer into its level and its cross-section separates two things that behave differently. The level call is the sign of the average recalled return in a month. It is 0.708 accurate before the cutoff and 0.680 after it, a gap of +0.028. Nothing that crosses

the cutoff this nearly intact can be a memory of the period; what the level call encodes is a prior about the direction of equity markets, and a good one. The cross-section does the reverse, falling from +0.166 to -0.019. Since a dollar-neutral book holds the cross-section and discards the level, neutrality removes the component that does not leak.

This also explains why a pooled directional-accuracy score understates the effect. Pooling every (name, month) pair into one confusion matrix adds the level call to the cross-section and reports the sum. Scored that way the same cached responses give 0.611 against 0.494 balanced accuracy ($p = 0.013$), which is both smaller and attributed to the wrong component.

4.2 Segments, and why the choice decides the answer

A model with a standing tilt scores a high IC in any period where the tilt paid. To separate that from month-specific recall, remove each name’s own average and ask what survives. The question is over what stretch the average is taken.

Taking it over the whole pre-cutoff window fails when the model’s behaviour is not homogeneous across that window. If the model answers erratically early and then settles into a near-constant template, the name averages are dominated by the erratic months and barely touch the template months, which survive the demeaning and are then read as recall.

Table 2 shows that this is what the primary subject does.

The two pre-cutoff stretches are indistinguishable on the raw cross-sectional IC, at +0.161 and +0.174. Under segment-local demeaning they are opposites: +0.116 against +0.001. Under window-level demeaning both read as recall, at +0.172 and +0.178, and the later stretch is misclassified.

The boundary at 2023-01 comes from the changepoint search of Section 6, which never sees a realised return. The segmentation therefore cannot have been chosen to flatter the contrast it is used to explain.

4.3 Reconciling an early boundary with exact-value recall

The later stretch sits a full year inside the documented training window and shows no month-specific memory. Two published results locate the same kind of boundary somewhere else. Gao et al. [6] find their date-only recall statistic collapsing to zero just after the documented cutoff, and López-Lira et al. [15] report exact recall

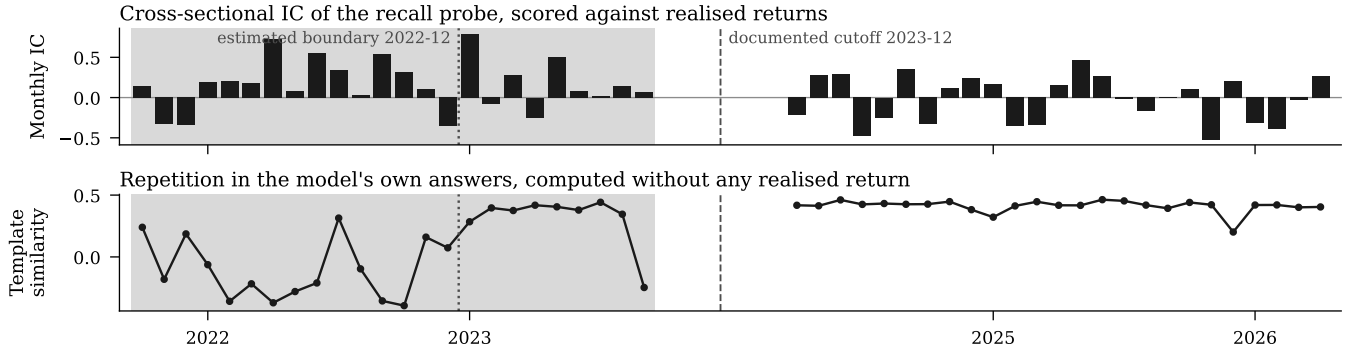


Figure 2: The same months read two ways, for llama-3.3-70b-versatile. Shading marks the in-window months, and the six buffer months, never queried, appear as a break. The dashed rule is the vendor’s cutoff, the dotted rule the boundary located from the lower panel alone. The upper panel needs realised returns and the lower one does not, and both turn a year before the documented date. The lower series is the leave-one-out statistic of Section 6, not the within-segment measure in Table 2.

Table 1: Cross-sectional recall IC by subject. openai/gpt-oss-120b declines to answer for its whole post-cutoff window, so it has no out-of-window observations.

Subject	Cutoff	IC pre	IC post	Gap	p	Months
llama-3.3-70b-versatile	2023-12	+0.166	-0.019	+0.185	0.017	24/25
openai/gpt-oss-120b	2024-06	+0.099	<i>n/a</i>	<i>n/a</i>	<i>n/a</i>	28/0

Table 2: Segment analysis for llama-3.3-70b-versatile. The IC columns are raw, then demeaned by name within the segment, then over the whole window; the last is mean pairwise rank correlation among the segment’s own months. The boundary 2023-01 was located from responses alone.

Segment	n	IC	IC (seg.)	IC (win.)	Repetition
Pre-cutoff, deep	15	+0.161	+0.116	+0.172	+0.096
Pre-cutoff, late	9	+0.174	+0.001	+0.178	+0.383
Out of window	25	-0.019	-0.042	-0.042	+0.748

of financial values before the cutoff and none after. Their boundaries fall where the vendor says they should. Ours falls twelve months early.

The measurements are not in conflict, because they are different functions of the same corpus. Exact-value recall is a threshold: one retrievable mention of a number is enough to reproduce it, so the statistic survives wherever any account of the period reached the training data. A cross-sectional ranking is a density. Ordering 50 names within a month requires that month’s outcomes to be represented across the cross-section, and scattered mentions of a few of the largest names will not do it. Coverage of a given month keeps accumulating for years after the month itself, through retrospective writing that has the outcome in hand. A corpus closing in 2023-12 therefore holds years of accumulated coverage of 2021 and 2022 against whatever was written about 2023 within months of it happening. That thin tail is enough for the threshold statistic and not for the density one, which fixes the order of the two boundaries rather than leaving it free: the ranking signal has to fail first.

The label-free series here is consistent with this, and it was measured without returns. Month-to-month repetition in the later stretch is +0.383, against +0.096 in the deep stretch and +0.411 after the cutoff. Through 2023 the model is already answering close to the way it answers about months it has no coverage of at all, a year before the vendor’s boundary.

This is a reconciliation and not a result. It carries a test on this subject that we do not run: an exact-value probe on the same grid should hold its accuracy through 2023 and break at 2023-12, while the ranking signal is already gone by 2023-01. Finding both boundaries in the same place would leave the early break unexplained. What the segment analysis establishes without that probe is narrower and sufficient for the protocol: a documented cutoff is the wrong split for a ranking signal, and splitting there reads a year of standing tilt as recall.

5 Breaking the date binding

The sharpest evidence that the in-window IC is month-specific recall comes from re-pairing. Keep each response exactly as given and score month M ’s answer against a different month’s realised returns. Entity knowledge is untouched and only the binding between entity and date is broken. Permutations are rejection-sampled derangements, so no month keeps its own returns, and both marginal distributions are preserved.

For the primary subject, in window, the true pairing gives +0.166 against a null of -0.006 with standard deviation 0.064 ($p = 0.003$), which says that inside the training window the model’s answer for a month is informative about that month and about no other. Out of window the true pairing gives -0.019 against -0.018 ($p = 0.524$); there is no date binding left to break.

The dispersion of the null is itself informative. It is 0.064 in window and 0.018 out. Mismatched pairings vary when the answers vary, and barely vary when the answers are a repeated template.

6 A detector that needs no labels

Every test so far needs realised returns, which confines it to periods where the outcome is already known. That is a real restriction, because the question a practitioner actually has is about periods they cannot score: whether a vendor’s stated cutoff can be believed, and whether a model is reciting rather than reasoning on months whose outcomes do not exist yet. Two statistics answer from the responses alone.

6.1 Template collapse

The first measures how much the model repeats itself. For each month, take the rank correlation of that month’s ordering against every other month’s ordering and average them. Leaving one out against all other months, rather than against the previous month, is deliberate: adjacent months are genuinely correlated in the market, so a model that recalls two consecutive months correctly shows high lag-one autocorrelation for a legitimate reason. Averaging over distant months removes that confound.

For the primary subject the statistic is +0.052 in window and +0.411 out (Table 3). The within-month dispersion of recalled returns falls from 4.81 to 3.57 across the same boundary, so the answers are not only more alike from month to month, they are flatter inside a month as well.

Because the series needs no labels it can be swept across the calendar. A binary-segmentation changepoint search, signed so that only a rise in repetition counts and calibrated by permutation because the maximum over admissible splits is upward biased, puts the boundary at 2022-12 for the primary subject, with statistic 7.39 and $p = 0.001$. The return-scored segment analysis of Section 4.2 agrees, and neither was told the other’s answer.

6.2 Refusal

Some models state the boundary outright, and no statistic is needed to read it. `openai/gpt-oss-120b` declines to answer for 7% of its in-window months, 83% of its buffer months and 100% of its out-of-window months.

`llama-3.3-70b-versatile` refuses nothing. It answered every month of the grid with specific fabricated percentages, including months two years past its cutoff. The two behaviours bound the practical problem from opposite sides. A model that refuses cannot be used to build a contaminated backtest, because there is nothing to backtest. The risk is concentrated in models that answer anyway.

The two detectors are complements rather than substitutes. For a refusing model the template detector has nothing to work with, since the answered months all lie on one side of the boundary, and its changepoint is correspondingly insignificant at $p = 0.295$.

7 Framing

Every measurement so far came from a prompt that asked the model to recall, and that choice is not incidental to the result. Put the same subject and the same months behind a forecast prompt, asking which names will outperform over the following month, and the

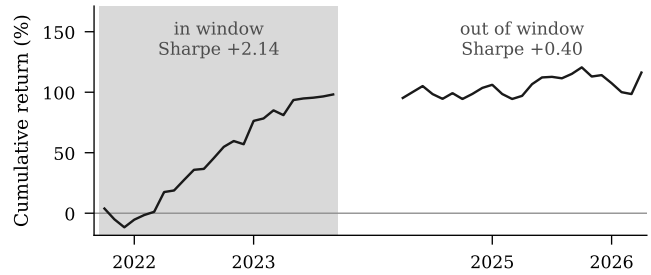


Figure 3: Cumulative return of the dollar-neutral book, summed rather than compounded because the book is re-balanced to constant gross exposure. The slope inside the training window is the quantity this paper argues is recall. Both Sharpe ratios are the ones in Table 4.

cross-sectional IC inside the training window is +0.004. Ask instead what each name did, and it is +0.166. The leakage is a property of the prompt as much as of the model.

The practical consequence is worse than a smaller number. Someone holding a contaminated model can evaluate it under a forecast framing, find nothing, and conclude on what looks like good evidence that contamination is not a problem for their application. The leakage has not gone anywhere. The prompt has stopped inviting retrieval, and a test of this kind measures how hard it invites it.

8 Portfolio implications

Table 4 translates the signal into a backtest. Construction is deliberately plain: rank the cross-section, centre the ranks to zero net exposure, scale to constant gross exposure, hold one month, rebalance. No optimiser, no shrinkage, no leverage target and no transaction costs. Costs are left out because including them would only reduce the in-window Sharpe ratio, and the in-window Sharpe ratio is the number this paper argues is spurious. Subtracting from it would make the case look stronger than it is.

A 24-month Sharpe ratio cannot be quoted without a null. A random signal on this cross-section produces a null Sharpe ratio with standard deviation 0.75 over the same number of months, so the observed +2.14 is calibrated by re-pairing rather than by a normal approximation: the null mean is -0.02 and $p = 0.003$. Out of window the book returns +0.40 against a null mean of +0.45 ($p = 0.620$), which is to say the out-of-window Sharpe ratio is what the standing template earns by drifting with the market.

The number a practitioner should take from this is the difference. An in-window evaluation of this signal reports +2.14. The same signal evaluated on months the model cannot have seen reports +0.40.

9 Limitations

The design registered for this study was a staggered-cutoff comparison, and it cannot be estimated on the subjects available. The idea behind it was sound. With cutoffs that differ across models, a calendar month falls inside one model’s window and outside another’s, and month fixed effects then absorb the period effect that no single-subject comparison can rule out. The data will not

Table 3: Label-free detectors. Neither template similarity nor refusal rate uses a realised return. The located cutoff is a changepoint on the template series.

Subject	Template similarity		Refusal rate			Located cutoff		Documented cutoff
	pre	post	pre	buffer	post	month	p	
llama-3.3-70b-versatile	+0.052	+0.411	0%	<i>n/a</i>	0%	2022-12	0.001	2023-12
openai/gpt-oss-120b	+0.389	<i>n/a</i>	7%	83%	100%	2022-01	0.295	2024-06

Table 4: Dollar-neutral long-short book on the recall signal for llama-3.3-70b-versatile. The last column is the add-one p against a null that re-pairs months. Standard errors follow Lo [13].

Window	n	Mean (%/mo)	Ann. (%/yr)	Sharpe (ann.)	s.e.	Max DD (%)	p
In window	24	+4.09	+49.1	+2.14	0.77	-14.9	0.003
Out of window	25	+0.72	+8.7	+0.40	0.70	-20.5	0.620

support the estimation. The two subjects with panels have 2 distinct cutoffs and windows that nest rather than cross, which leaves 0 identifying months. Three distinct cutoffs are the minimum for this design, because with two the later-cutoff model is the treated one in every month where the models disagree, so the treatment indicator is a relabelling of model identity. The estimator tests for this and refuses to report a coefficient rather than returning the number a rank-deficient least-squares problem would hand back.

What identification remains is within-subject, and that is why the re-pairing tests carry as much weight here as they do. They hold the calendar period fixed and vary only which answer is scored against which month’s outcome, so the period confound that the staggered design was meant to absorb cannot reach them.

Power is the next constraint, and it bites hardest where the argument is sharpest. The segment-local contrasts are the cleanest statement of the effect and neither is significant at five percent: on 15 pre-cutoff months the name-demeaned gap has $p = 0.059$ and the cross-sectional gap has $p = 0.098$. That is the arithmetic of concentrating an effect, which sharpens the point estimate and shrinks the sample in the same move. The results that do carry significance are the full-window IC gap ($p = 0.017$), the two re-pairing tests ($p = 0.003$ and $p = 0.003$) and the changepoint ($p = 0.001$).

Behind both of those sits a plainer limit. llama-3.3-70b-versatile is the only subject with a complete contrast. openai/gpt-oss-120b contributes in-window observations and the refusal channel and nothing outside its window, because it refuses there. A protocol demonstrated on one subject is a protocol demonstrated, not a property of language models as a class, and this paper claims only the first.

That subject can no longer be queried. The provider retired llama-3.3-70b-versatile during this work and none of the registered providers serves it, so its cached responses are now the only record of the registered result, and its six buffer months, never sent, cannot be obtained at all. This is why the response cache is tracked in version control and treated as the canonical artefact rather than

as a convenience: every published number regenerates from it offline, and no part of the result can be reproduced from the provider. A reader who does not trust the cache has no independent way to check it, and saying so is better than implying otherwise.

The universe is the last of these, and it cuts in one direction. The 50 US large caps scored here are among the most written-about equities in any corpus, which makes them the easiest case a model could face in trying to recall a month. The estimate should be read as a plausible upper bound rather than a typical one, and it should not be carried across to smaller names or other markets without rerunning the protocol there.

10 Conclusion

Training-data contamination in an LLM equity backtest is measurable, and on the subject studied here it is worth an annualised Sharpe ratio of +2.14 in window against +0.40 out. Where it sits matters more than how large it is. The leakage lives in the cross-section rather than in market direction, so a dollar-neutral construction concentrates it instead of removing it, which is the reverse of what neutrality is usually relied on to do. Part of what looks like contamination inside a training window is a standing tilt that happened to pay, and separating the two takes a segmentation the vendor’s cutoff does not supply. The model’s own responses supply one, and they do it without any realised returns, which turns the boundary from something assumed into something estimated.

The practical reading is narrow. Nothing here says a language model cannot forecast returns. It says that a backtest run inside a training window does not answer the question, that the usual defence of market neutrality makes this particular problem worse rather than better, and that the date a vendor publishes is a claim to be checked rather than a fact to be used.

Reproducibility

None of the above has to be taken on trust. Every number regenerates from the repository offline, with no API key, by running scripts/analyse.py, scripts/make_tables.py and scripts/make_figures.py. No statistic in the text is typed by hand: each is a macro written from the machine-readable results, and the figures come from the same panel through the same functions as the tables. Continuous integration reruns the pipeline and fails if a committed number stops matching what the code produces. Seeds are fixed at 42, the bootstrap uses 10,000 replicates and every permutation test the same, and the response cache is keyed by SHA-256 over the model id, the prompt and the system prompt.

References

- [1] David H. Bailey and Marcos López de Prado. 2014. The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality. *The Journal of Portfolio Management* 40, 5 (2014), 94–107. doi:10.3905/jpm.2014.40.5.094
- [2] Mostapha Benhenda. 2026. Look-Ahead-Bench: a Standardized Benchmark of Look-ahead Bias in Point-in-Time LLMs for Finance. arXiv:2601.13770. arXiv:2601.13770
- [3] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Berkeley, CA, USA, 2633–2650.
- [4] Antoine Didisheim, Martina Fraschini, and Luciano Somoza. 2025. AI’s predictable memory in financial analysis. *Economics Letters* 256 (2025), 112602. doi:10.1016/j.econlet.2025.112602
- [5] Alexander Eliseev and Sergei Seleznev. 2026. Fake Date Tests: Can We Trust In-sample Accuracy of LLMs in Macroeconomic Forecasting? arXiv:2601.07992. arXiv:2601.07992
- [6] Zhenyu Gao, Wenxi Jiang, and Yutong Yan. 2025. Detecting Lookahead Bias in LLM Forecasts. arXiv:2512.23847. arXiv:2512.23847
- [7] Paul Glasserman and Caden Lin. 2024. Assessing Look-Ahead Bias in Stock Return Predictions Generated by GPT Sentiment Analysis. *The Journal of Financial Data Science* 6, 1 (2024), 25–42. arXiv:2309.17322.
- [8] Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies* 33, 5 (2020), 2223–2273. doi:10.1093/rfs/hhaa009
- [9] Campbell R. Harvey, Yan Liu, and Heqing Zhu. 2016. ... and the Cross-Section of Expected Returns. *The Review of Financial Studies* 29, 1 (2016), 5–68. doi:10.1093/rfs/hhv059
- [10] Songrun He, Linying Lv, Asaf Manela, and Jimmy Wu. 2025. Chronologically Consistent Large Language Models. arXiv:2502.21206. arXiv:2502.21206
- [11] Hans R. Künsch. 1989. The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics* 17, 3 (1989), 1217–1241. doi:10.1214/aos/1176347265
- [12] Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2025. Profit Mirage: Revisiting Information Leakage in LLM-based Financial Agents. arXiv:2510.07920. arXiv:2510.07920
- [13] Andrew W. Lo. 2002. The Statistics of Sharpe Ratios. *Financial Analysts Journal* 58, 4 (2002), 36–52. doi:10.2469/faj.v58.n4.2453
- [14] Alejandro López-Lira and Yuehua Tang. 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. arXiv:2304.07619. arXiv:2304.07619
- [15] Alejandro López-Lira, Yuehua Tang, and Mingyin Zhu. 2025. The Memorization Problem: Can We Trust LLMs’ Economic Forecasts? arXiv:2504.14765. arXiv:2504.14765
- [16] Belinda Phipson and Gordon K. Smyth. 2010. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics and Molecular Biology* 9, 1 (2010). doi:10.2202/1544-6115.1585
- [17] Suproteem K. Sarkar and Keyon Vafa. 2024. Lookahead Bias in Pretrained Language Models. SSRN working paper 4754678; presented at ICML 2025. doi:10.2139/ssrn.4754678
- [18] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting Pretraining Data from Large Language Models. In *The Twelfth International Conference on Learning Representations (ICLR)*. Vienna, Austria.