SILENCE IS NOT ABSENCE: A PROBE NULL REQUIRES MORE THAN FORMAT COMPETENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

A probe that asks what a language model memorised reports a null when the model answers at chance, and that null is evidence of absence only if the probe could have shown the fact. Two separate conditions have to hold, and checking the first does not establish the second.

The first is that the model can use the readout. Asked thirty questions with no domain content, of the form "which one is a fruit, an apple or a hammer", each in both option orders so that a fixed letter preference scores exactly one half, seven of nineteen open-weight models cannot do it. The failures are not the small models: every rung of the Pythia ladder fails from 0.16B to 6.9B, and Llama-2-7B fails while Llama-3.2-1B passes at a sixth of its size; inside one laboratory OLMo-1-7B scores 0.517 and OLMo-3-7B scores 1.000. Two of the seven answer a yes/no frame at 0.983, consistent with a label-binding failure rather than an inability to answer, and the excluded models cannot be read even when the answer is printed in the prompt, at 0.507 and 0.535.

The second is that information of the kind being probed can reach that readout, which no control built from questions a model already knows can establish. We test it by writing invented dated facts into training text at four exposure counts, in three models from three laboratories. Under the form in which they were learned every model separates the true value from the probe's own foil, increasingly with exposure and against flat untrained controls, while under five forced choices — relabelled, list-free, and with demonstrations — the same facts remain at chance in every model. Interface validity does not imply construct validity.

Turning the instrument on the question that motivated it, whether small open-weight models memorise market history closely enough to contaminate a backtest, the probe that scores about 0.90 on which company a ticker belongs to scores about 0.50 on every date-specific fact about that company, and we exclude effects larger than 1.75 points in the direction memorisation predicts. By the second condition, that bounds what a forced choice retrieves, not what the models hold. We release the panel, the probes, the estimators and the power calculator.

1 INTRODUCTION

A probe reports that a model does not know something by showing it answering at chance. That inference has a precondition which is almost never checked: the model had to be able to answer in the first place. We find that the precondition fails in seven of the nineteen open-weight models we scored, that it fails silently, and that it fails in places no reasonable person would look.

The setting that led us here is a live disagreement. Lopez-Lira et al. (2025) show GPT-4o recalling Wall Street Journal headline dates at 98.45 percent year accuracy and S&P 500 daily levels at 0.61 percent error before its cutoff, collapsing to 28.81 percent and 16.9 percent after. Didisheim et al. (2025) report that smaller models and finer data granularity exhibit negligible bias. Both can be true: they measure different regimes with different instruments, and neither instrument reports what it could have detected. Settling it requires probes, and probes require the precondition.

This paper makes four contributions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Format competence is not monotone in model size, and that breaks the standard reading of a probe null. Before asking what a model remembers we ask whether it can answer at all, on sixty elementary questions carrying no finance content, each asked in both option orders so that a fixed letter preference scores exactly one half. Seven of nineteen models fail, and they are not the small ones: every rung of the Pythia ladder fails across a forty-fold range, Llama-2-7B fails where Llama-3.2-1B passes at a sixth of its size, and inside one laboratory OLMo-1-7B scores 0.517 against OLMo-3-7B's 1.000. It is not instruction tuning either: nine of the twelve models that pass are base checkpoints. The control costs sixty items and seconds of compute, both thresholds were fixed before any result was read, and the partition it produces is not a judgement call — the accuracies are bimodal, nothing at all between 0.600 and 0.850, twenty-five points apart. Failing models are excluded by the analysis code rather than by us afterwards, and their probe scores are withheld, because at chance those numbers are equally consistent with a model that knows nothing and one that cannot use the format.

A null needs two conditions, not one, and the second is the one nobody checks. Every quantity we report comes from a two-way forced choice whose chance baseline is exactly one half, fixed by the balanced design rather than estimated. The first condition is that the model can work the readout at all, which is what the format control decides and what most of this paper is about. The second is that information of the kind being probed can reach that readout once the model holds it — and no format control can settle that, because it is a property of the elicitation and the knowledge jointly. We test it the only way it can be tested, by putting the facts in ourselves: invented companies, dated facts written into training text at four exposure counts in three models from three laboratories, and five forced choices asked afterwards. A null becomes a measurement once both conditions hold, and we report what happens when the second one is examined rather than assumed.

The question that motivated the control, bounded by it. Whether small open-weight models memorise market history closely enough to contaminate a backtest. We first measure what the design the field uses could have detected: its minimum detectable effect on our panel is 0.149 in weekly rank IC against effects at issue of at most 0.03, and we give a design on the same data reaching 0.0144. Lopez-Lira et al. (2025) prove pre and post cutoff comparisons can be statistically indistinguishable from memorisation; we put a number on it for a concrete estimator and panel. Then, across nineteen open-weight models spanning nine documented cutoffs over five years, five of them rungs of a ladder that holds a cutoff fixed while capacity varies, with a factorial separating company identity from calendar date: among the models the control admits, entity knowledge spans 0.805 to 0.970 and gains 2.4 points per doubling of parameters, while every dated probe stays within 2.1 points of chance and shows no association with capacity at all. The claim is bounded and stated that way throughout: what is absent is *retrievable* dated memory under this elicitation. A second market, Taiwan, was screened and ruled out by the same controls, at 0.500 on its tickers.

The instrument as an artifact. Panel, probes, estimators, power calculator and per-model scores.

2 Related work

Probe validity, which is where this paper's contribution sits. That a forced choice is fragile is established. Language models carry a token-level prior over option labels that is independent of content (Zheng et al., 2024), accuracy swings by tens of points when the options are permuted (Pezeshkpour & Hruschka, 2024), scoring answers by log-probability is confounded by surface-form competition and by miscalibration (Holtzman et al., 2021; Zhao et al., 2021), and benchmark rankings move under cosmetic changes to prompt, label and order (Sclar et al., 2024; Alzahrani et al., 2024; Mizrahi et al., 2024). Closest of all, Robinson et al. (2023) isolate multiple-choice symbol binding, the capacity to associate an option's label with its content, and show it varies sharply across models; Wang et al. (2024) show first-token and free-text readouts of the same question disagree. Earlier knowledge-probing work raised the same confound in cloze form: LAMA-style accuracy reflects prompt artifacts as much as stored knowledge (Petroni et al., 2019; Cao et al., 2021), and Hewitt & Liang (2019) made the general point that a probe needs a control task before its accuracy means anything.

We add three things, and only the combination is ours. First, the sensitivity becomes an *admission rule* rather than a caveat: thresholds fixed before any result is read, applied per model, with failing models excluded by the analysis code. Second, the failure is not predicted by the variables a prac-

titioner would use to decide whether to check — not by scale, since a forty-fold span of one family fails uniformly while a 1.2B model passes, and not by instruction tuning, since nine of our twelve admitted models are base checkpoints. Third, and separately from any of the sensitivity literature, we show that passing such a control does not establish that the probe can reach the class of fact it asks about: with facts we wrote into training ourselves, the same model discriminates them from the probe's own foil under the form it learned them in, and returns exactly chance under the probe.

Contamination detection. Membership-inference statistics (Shi et al., 2024; Zhang et al., 2025), guided-instruction probing (Golchin & Surdeanu, 2024) and exchangeability tests with explicit false-positive guarantees (Oren et al., 2024) all decide whether specific text was trained on. Large-scale evaluation finds membership inference barely beats chance on realistic pretraining scales (Duan et al., 2024), and documented cutoffs are themselves unreliable: Cheng et al. (2024) show an effective cutoff that differs from the reported one and varies by sub-corpus, which is the assumption our staggered design leans on hardest and the reason for the donut and the placebo cutoffs. Memorisation is known to grow with capacity and duplication (Carlini et al., 2023; 2021; Kandpal et al., 2022), which is what makes a null at 8B a statement about a regime rather than about language models. Few of these methods report the effect size their design could have detected; Cheng et al. (2025) is the exception in spirit, planting a signal of known size to obtain an exact error rate. We do the analogous thing for a continuous effect, and plant facts rather than a signal, to ask whether the readout can recover them at all.

The closest work, Gao, Jiang and Yan (Gao et al., 2026), tests contamination by a before/after split at the cutoff, with no running variable or bandwidth selection: their detection interaction is 0.162 ($t = 3.64$) in sample and $t = 1.06$ in a post-cutoff placebo. We do not dispute the estimate. We report what a null from that design is worth, which on our panel is a minimum detectable effect of 0.149 against effects at issue of at most 0.03.

One direct methodological threat is answered in our design. Wu et al. (2026) argue that entity masking confounds signal loss with bias removal, so a single named-versus-anonymised contrast is not identified. We vary identity and date factorially: removing the date removes the memory handle while leaving company information intact, the two arms then fail differently, and no single information-loss story explains both.

The apparent contradiction in the literature resolves along model scale and data granularity. Lopez-Lira, Tang and Zhu (Lopez-Lira et al., 2025) test GPT-4o and Llama-3.1-70B on macro series, index levels and headline dates, with no analysis below 70B and no scaling curve; Didisheim, Fraschini and Somoza (Didisheim et al., 2025), in a published journal paper, report the opposite sign for smaller models and finer granularity. Our cell — 1B to 8B models on stock-week cross-sectional facts — is the one neither measures, and the one most reproducible academic work occupies.

Two further nulls support the anonymised arm rather than replacing it: Lehner & Lopez-Lira (2026) pool twelve model snapshots and attribute their effect to synthesis rather than memorisation, and Zhu et al. (2026) recover a masked ticker at most 3.0 percent of the time across ten de-anonymisation attacks.

The paper whose numbers sit hardest against ours. MemGuard-Alpha (Roy & Roy, 2026) scores seven LLMs from 124M to 7B with five membership-inference attacks plus a cross-model disagreement statistic that, like our design, exploits cutoff variation; filtering on it raises Sharpe from 2.76 to 4.11 on an 11.7-point in-sample accuracy spread, an order of magnitude above the 1.1 points our pooled probe can resolve. We measure $+0.41$ points (95% CI -0.54 to $+1.37$) on price level and -0.61 (-1.97 to $+0.75$) on return ordering. The two findings need not conflict, since membership inference asks whether a text was trained on and we ask whether a dated fact can be retrieved; Appendix C gives the reconciliation.

3 Setting

3.1 Models and cutoffs

Twenty models are on the roster and nineteen ran to completion, which is the count every probe result is quoted over; Table 9 carries the attrition from there. Fifteen of the twenty carry a cutoff of their own, spanning nine distinct cutoff months from 2020-07 to 2025-06; the other five are ladder

rungs sharing an anchor's cutoff. Thirteen of the fifteen are stated in a primary source, two derived from the corpus paper for the family. The full roster, with the quotation and URL behind every cutoff, ships with the artifact.

The OLMo family carries much of the design's weight: four cutoffs from one laboratory on one architecture lineage, trained on a published corpus, so the cutoff is checkable against the corpus itself rather than taken on trust.

Seven major open-weight families disclose no training-data cutoff at all: Qwen, Mistral, Gemma 1 and 2, SmolLM and SmolLM2, Amber, Falcon 3 and StableLM, each checked against blog posts, technical reports and model cards. Disclosure is not established practice, which is the practical argument for a behavioural probe.

3.2 Data

Point-in-time S&P 500 membership is reconstructed from a public historical components file: 846 tickers were members at some month end since 2010. Daily bars come from Yahoo. On a weekly Friday grid running 2010-01-08 to 2026-09-04, the panel holds 505,248 stock-weeks with a median in-index cross-section of 416 names.

Survivorship is reported rather than hidden. Of the 846 ever-members, 650 have price data; of the 196 without, 195 had already left the index through acquisition or delisting, and one is a current member with no Yahoo series; the per-ticker audit ships with the artifact. This is a real limitation: the surviving sample tilts toward names that were not acquired, more strongly later in the sample, which could in principle manufacture a discontinuity of its own; the placebo cutoffs in Section 7 are the check on it.

A second market, Taiwan, was screened the same way and ruled out of scope at 0.500 on its tickers against about 0.90 on US ones, leaving no entity knowledge for a dated probe to rest on (Appendix B).

3.3 Task and scoring

Each stock-week gets one forward pass. The prompt gives twelve weekly closing prices at split-adjusted dollar levels, with or without their calendar dates, for a company that is either named or anonymised, and asks whether the stock will outperform the index over the following week. The score is the log-probability gap between the two answer letters. The answer-to-letter mapping is fixed rather than randomised: a constant letter preference is shared across a week's stocks and drops out cross-sectionally, which is the only way it is used.

Five prompt variants are run: `named_dated`, `anon_dated`, `named_undated`, `anon_undated`, and `shuffled_dated`, the last of which puts another company's name on this company's prices. Each of the 452 Fridays from 2018-01-05 to 2026-08-28 contributes exactly 200 stocks — the hundred largest by trailing dollar volume, plus a hundred drawn from the remainder under a fixed seed, so the cross-section spans size — among those in the index with twelve prior weekly closes and a realised forward return. That is 90,400 stock-weeks per model-variant over 580 tickers, the same ones under every variant.

Two dated probes ask for a remembered fact directly, and are named here because Section 4 quotes one: the *price-level* probe gives a company and a date and offers two of its own true closing prices, and the *return-ordering* probe gives two companies and a date and asks which returned more that week. Section 5.1 builds both.

The roster loses models at four points and the paper quotes a different count at each; Appendix J gives the attrition in one table, and every exclusion is applied by the analysis code.

4 Putting a number on a known problem

The question that motivated the instrument is whether small open-weight models memorise market history closely enough to contaminate a backtest. Before asking it we measured what the design the field uses could have found on our panel: its minimum detectable effect is 0.149 in weekly

rank information coefficient, against effects at issue of at most 0.03. A null from that design is not evidence of anything. Lopez-Lira et al. (2025) prove that pre- and post-cutoff comparisons can be statistically indistinguishable from memorisation; this puts a number on it for a concrete estimator and panel.

A design on the same data reaches 0.0144, through three changes: estimate at the stock-week rather than the week, pool the staggered cutoffs so that some models are pre-cutoff and others post in most calendar weeks, and absorb the regime with a calendar-week fixed effect, which only becomes available once models are pooled. Twelve characteristics and a ridge, run through the identical pipeline with each model's cutoff imposed on them, reject at 4.1 percent against a nominal 5. Appendix M gives the estimating equations, the derivation of the two detection thresholds, and the placebo grid.

Section 6 applies this design, clustered on ticker rather than on calendar week, to the two dated probes.

5 Two conditions for reading a null

5.1 Interface validity, and the models it disqualifies

Section 4 shows the performance channel is small and noisy. Memory itself is neither. We measure it with four probes sharing one instrument: a two-way forced choice scored by the gap between the log-probabilities of the two answer letters. They differ only in what has to be remembered.

The four are a format control (60 items, can the model use this answer format), an identity probe (2,478 items, does it know the entity), a price-level probe (14,460 items, absolute recall) and a return-ordering probe (25,740 items, the quantity a backtest predicts).

Two controls, not one, and the second exists because the first was not enough. The identity probe is an entity-knowledge positive control and separates some interface failures from a dated-probe null: Phi-2 scores 0.920 on it ($z = 41.8$) against 0.508 on dated price levels, on the same format and scoring. Not all of them, which is what Section 5.3 is about. But a model can also be at chance on identity because it cannot work a forced choice at all, which has the opposite implication for a null. We found such a model rather than assuming none existed: OLMo-1-7B scored 0.502 on ticker-to-company, answering B on 95.9 percent of items and scoring the two letters exactly equal on a further 2.0. It was not alone.

So the format control runs first for every model. It asks 30 questions of the form "which one is a fruit, an apple or a hammer", each in both option orders, so a fixed letter preference scores exactly one half. A model is admitted only if it clears 0.80 and is above chance at the one in a thousand level. Both thresholds were fixed before the probe results were read, and the partition does not depend on where in a wide interval they sit: the accuracies are bimodal, ten models perfect with two more at 0.917 and 0.850, then nothing above 0.600: a twenty-five point gap between the two groups, and forty between the failures and the ten that are perfect. Any threshold from 0.61 to 0.85 admits the same twelve models, and so does dropping the accuracy rule and admitting on z alone: 0.80 is a stated commitment, not a load-bearing choice. Every model's score, with two diagnostic columns saying what kind of failure each one is, is in Appendix L.

Neither size nor instruction tuning explains the partition. The obvious reading of Pythia-160M's 0.600 is a size floor, and the roster refutes it three times over: the Pythia ladder fails across a forty-fold range, Llama-2-7B fails where Llama-3.2-1B passes at a sixth of its size, and inside one laboratory OLMo-1-7B sits at 0.517 against OLMo-3-7B's 1.000.

The second reading, that the failures are the models never taught to follow an instruction, would make this control a restatement of something known. The roster rules it out: every failing model is a base checkpoint, but so are nine of the twelve that pass, and both within-family contrasts above hold with tuning fixed at none. What separates the groups is not identified here; vintage and pretraining recipe are what remain, neither legible from a model card.

What the control certifies, stated narrowly. It certifies a model against an elicitation, not in general: competence here is a property of the model and the answer convention jointly. Admission turns on the published A/B list because the dated probes are A/B lists. Appendix F reruns the thirty questions under numerals, parentheses and a yes/no frame on all nineteen models: no relabelling

lifts an excluded model over the bar, while four of the twelve admitted models fail at least one alternative, Llama-3.2-1B falling to 0.650 under numerals.

A failing model is reported as out of scope rather than counted as a null, and is excluded by the analysis code rather than by our judgement after the fact.

The price-level probe offers two true closing prices of the same company at dates at least six months apart, so a sense of its usual price range does not help and only date-specific recall does. 39 quarterly dates, 14,460 items, answer letter balanced to 0.505.

The return-ordering probe is the one a backtest turns on. Its prompt contains no numbers, removing the magnitude heuristic of Appendix A, and a standing prior about which company is better is neutralised because the same pair is asked at many dates and the answer flips. 117 monthly dates, 25,740 items, answer letter balanced to exactly 0.500 at every date.

Each dated probe has an anonymised arm in which the company becomes an opaque identifier. Its chance baseline is exactly one half by construction rather than estimated, and it absorbs whatever works without the company, such as a sense of normal price levels in a given year. On Phi-2 it scores 0.5091 against the named arm's 0.5084, so the identity contribution is -0.06 points.

On the dated probes the contrast the estimator targets recovers an injected effect with unit slope, which establishes one link and not the whole chain: conditional on an effect reaching the scores, the estimator finds it. Whether the probe puts one there is the separate question of Section 5.3 (Appendix P).

5.2 Controls shaped like the task

Answering whether an apple or a hammer is a fruit is no evidence that a model can read twelve dated prices and bind an option letter to one of them. Two further controls close that gap, and both are cheap.

A control shaped like the task. Sixty questions with the furniture of the real prompt — a company label, a twelve-row dated price table, two candidates, an A/B answer — and an answer printed in the prompt itself: which of two dates is later, which of two prices appears in the table, which company a ticker maps to given the mapping, which of two stated returns is higher. Each is asked in both option orders. Eighteen of the nineteen models were scored — the missing one, Phi-3-mini, is admitted — so all seven models that fail the published control were scored here, and none of them passes this one: the five Pythia rungs sit between 0.492 and 0.517, OLMo-1-7B at 0.558 and Llama-2-7B at 0.658. But the harder control also separates models the published one admits. Phi-4-mini reaches 1.000 and OLMo-3-7B 0.992, while OLMo-1.7-7B scores 0.658 and Llama-3.2-1B 0.708 — both admitted at 0.850 and 0.917 on the published control. Applying the same threshold to the harder instrument would exclude those two. We report this rather than choose the flattering instrument: readability is a property of the model and the question jointly, and the admission rule is stated with the control it was measured on. The pooled estimates of Section 6 do not depend on the choice, and only one of the two is in them: Llama-3.2-1B shares its anchor's cutoff, so it is already outside the eleven of Table 9. Dropping both leaves ten, on which the estimates are $+0.41$ and -0.61 points, unchanged to two decimals.

A control that hands the model the answer. The dated probes again, with the fact supplied in the prompt — *For reference, on* [date] *the closing price was* [value] — and then the same question. This bounds the interface rather than the memory: prompt length, dates, two candidates, letter binding and log-probability scoring all have to work for a model to score near one.

Over 600 items per cell, Llama-3-8B reaches 0.998 on price level and 0.970 on return ordering, OLMo-3-7B 0.998 and 0.843, Phi-2 0.897 and 0.797, and Llama-3.2-1B 0.805 and 0.545: it cannot read return ordering even handed the answer, though as a ladder rung no pooled figure rests on it. The two models the format control excludes score 0.507 and 0.518 (Llama-2-7B) and 0.535 and 0.530 (OLMo-1-7B).

Their null on the dated probes is therefore not a statement about their memory in any reading, and that is a measurement rather than a threshold decision. At the other end nothing about the interface itself is broken.

What the admitted models do instead. The same contrast appears inside the admitted set, in the letter they choose. On the identity probe every admitted model picks the first option between 0.453 and 0.558 of the time: balanced, because the content decides. On the dated probes the same models run from 0.101 to 0.982. OLMo-3-7B picks the first option 98.2 percent of the time on price level and 10.1 percent on return ordering; Llama-3.1-8B picks it 97.0 percent of the time on return ordering. They are not answering dated questions wrongly. They stop using the content and fall back on a position, which the balanced design converts into exactly one half rather than into a spurious signal.

5.3 Interface validity is not construct validity

A model that clears the control of Section 5.1 can work the readout; it does not follow that the readout can reach the kind of fact asked about. The case that matters, a fact the model holds and the probe cannot surface, is invisible while we do not know what it read, so we put the facts in.

Construction. Sixty invented companies, with names and tickers that cannot occur in any pretraining corpus, are given dated closing prices for 2019. The facts are written into prose at four exposure counts — 0, 1, 5 and 20 — with half of each company's exposures paraphrased, so that a later answer cannot be matching one remembered sentence. Two of every trained company's six price facts are never written down, which separates having learned the company from having learned the fact. Phi-2, Llama-3.2-1B and SmolLM3-3B, from three laboratories, are each trained on those 340 documents alone by a rank-32 attention adapter ($\alpha = 64$, AdamW, twelve epochs, plain language-modelling loss; nothing in the probe's format is trained). The rate is fixed by a rule set before any probe was read and blind to it: the larger of 10^{-4} and $5 \cdot 10^{-5}$ whose post-training format control clears 0.80. All three pass at 10^{-4}, at 0.875, 0.867 and 0.983. Gemma-3-4B fails at both, 0.533 and 0.742, and is excluded rather than counted: training broke its readout (Appendix I).

Six readouts of the same fact. The first is the training sentence truncated at the value, scored as the log-probability of the true value minus that of the published probe's own foil. The other five are forced choices over the same pair: the A/B list; the list after eight demonstrations about disjoint invented companies; the same with their labels shuffled; the list relabelled 1 and 2; and yes/no about one candidate at a time, with no list to bind.

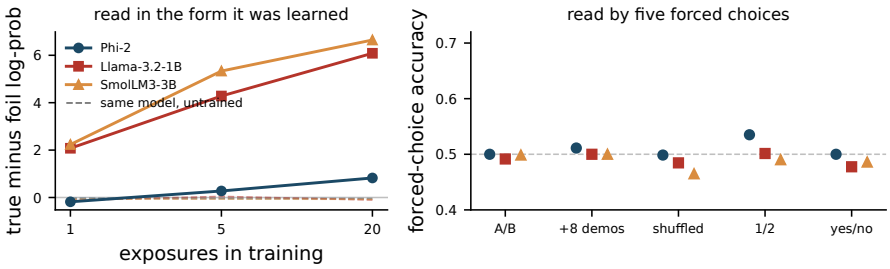


Figure 1: Interface validity does not imply construct validity, in three models. Left: under the readout the facts were learned in, the log-probability margin of the true value over the probe's own foil rises with exposure in every trained model and stays flat in the same models untrained. Right: the same facts under five forced choices, pooled over exposure, stay at chance in every model.

The fact is retrievable in the form it was learned, in all three. The margin of the true value over the probe's own foil rises with exposure in every model, from -0.18 to $+0.82$ on Phi-2, $+2.08$ to $+6.08$ on Llama-3.2-1B and $+2.23$ to $+6.64$ on SmolLM3-3B between one and twenty exposures. The trend is significant in each: Spearman ρ of 0.224, 0.197 and 0.198, with company-clustered 95 percent intervals $[+0.09, +0.34]$, $[+0.07, +0.31]$ and $[+0.06, +0.31]$. The same models untrained show none, ρ of 0.013, 0.022 and 0.008 with every interval spanning zero. The margin is the measure because accuracy saturates: two of the three already pick the true value at 0.650 and 0.700 after a single exposure. What is established is behaviour, not storage: retrievability under the training-matched readout.

No forced choice recovers it, in any of the three. Pooled over exposure, every forced choice in every model scores between 0.465 and 0.535. Demonstrations move the letter preference without moving sensitivity to content: on SmolLM3-3B eight of them shift the mean log-probability gap from -0.35 to $+9.63$ while accuracy stays at 0.500.

The readouts do not agree fact by fact, so the probe is not keeping a ranking and mis-setting its threshold. Because every fact is asked in both option orders, the content-dependent part of the A/B gap can be isolated exactly and correlated with the declarative margin over the same 178 facts. It is flat on Phi-2 ($\rho = 0.09$, $[-0.08, +0.26]$) and SmolLM3-3B ($\rho = 0.00$, $[-0.15, +0.15]$), and negative on Llama-3.2-1B ($\rho = -0.18$, $[-0.32, -0.04]$): the opposite of a probe that preserves the ranking and misplaces its threshold.

What this licenses. A model can clear a format control, hold a fact well enough to discriminate it from a matched foil under one readout, and return chance under every forced choice a study might use to conclude the fact is absent. A probe null is therefore interpretable only when both conditions have been shown, and no control that asks questions the model can already answer establishes the second. The readouts are training-free because direct attempts to train the answer mapping each degraded the forced-choice competence before teaching anything (Appendix I).

One diagnostic supports the construction rather than the claim. On facts about a trained company that were never written, the declarative readout falls well below one half, to 0.299, 0.138 and 0.115, as the design predicts, since the foil there is a price that *was* written; and the models that learned the written prices most strongly fall furthest.

6 THE FINANCIAL APPLICATION

Among admitted models, the result is a contrast, not a level. On one instrument, one prompt format and one answer-letter convention, the models that pass the format control score between 0.805 and 0.970 on which company a ticker belongs to, and between 0.490 and 0.521, on the named and the anonymised arm alike, on every question that requires a date.

Pooled difference in differences. Stacking the admitted models with model-by-arm and week-by-arm fixed effects, clustering on ticker, within a 156-week bandwidth and excluding a four-week donut around each cutoff:

probe	τ (pp)	se (pp)	p	observations	models
price level	$+0.41$	0.49	0.39	178,864	11
return ordering	-0.61	0.69	0.38	309,320	11

Table 1: Pooled difference in differences, admitted cutoff-bearing models.

Both are indistinguishable from zero; Table 9 traces the eleven models.

What the null excludes. The sign is fixed in advance: memorisation of the pre-cutoff period makes the named-arm advantage larger before the cutoff than after, so it predicts a negative τ. At the 95 percent level, one-sided and from the unrounded estimates, we exclude effects in that direction larger than 1.75 points on return ordering and 0.38 points on price level, the second that tight because the point estimate has the opposite sign. By Section 5.3 this bounds what the forced choice retrieves; it does not bound what the models hold.

A time-invariant named-arm advantage, not memorisation, appears on two models (Appendix K).

Where an effect would concentrate. Memorisation need not be uniform, so the estimate is also re-run inside quartiles of return divergence and of trailing dollar volume, where a memorised period should show first. It does not concentrate anywhere, and the one nominally significant cell of the twelve sits in the wrong tail; Appendix N gives the grid.

Skill, and what it does not settle. No admitted model forecasts: none has a mean weekly rank IC above zero (Appendix O). That alone does not rule out contamination, since a model below zero overall can still do better before its cutoff than after; the test of that is the triple difference.

The triple difference on the forecast task, the specification a leakage study would normally lead with, is null at both bandwidths and rejects no placebo. It is a null on a panel with no skill in it, which is why the probes carry the paper and it does not (Appendix O).

7 NEGATIVE CONTROLS

Placebo cutoffs. The same estimator at plus and minus 26, 52, 78 and 104 weeks from each true cutoff, where the jump is zero by construction. None of the eight displacements rejects on the primary design, and across all three the grid is 232 fits of which 16 reject, 6.9 percent against a nominal 5 and within sampling error of it at this count ($p = 0.12$).

Baselines that have no cutoff. Twelve characteristics and a ridge through the identical pipeline, rejecting at 4.1 percent (Section 4).

A pretraining-corpus audit. Eleven time-series foundation models, checked against primary sources rather than scored: none names an individual equity price dataset, so a documented cutoff there is one with no equity memory behind it and no jump should appear (Appendix H).

Specification robustness. Four bandwidths crossed with four donut widths, and each pooled model left out in turn: none of the 54 cells rejects at 5 percent (Appendix G).

The factorial. Identity and date are varied independently. A leakage story predicts a jump in the named, dated arm only. An information-loss story, which is the standing objection to anonymisation probes, predicts a level shift in every arm that removes information and no interaction with the cutoff date.

8 LIMITATIONS

Scale is the binding limitation and it is the first thing a reader should know. The roster tops out at 8B, while the frontier results that motivate the question are at GPT-4o and Llama-3.1-70B. Our result bounds a regime; it does not contradict theirs, and it is not licence to trust a frontier model on a pre-cutoff backtest. The boundary lies above 8B and nobody has located it; hardware set the ceiling, a 70B model not fitting on a 16GB GPU even at four bits.

The second limitation cost us a ladder, and is one of the main interface-validity results. The design called for two within-family ladders holding the cutoff fixed while capacity varies. The Pythia ladder does not survive the format control: it fails at every one of its five rungs, 0.16B through 6.9B, so not a single rung of it can be read for memory. That is a real loss to the application and simultaneously the cleanest evidence for the non-monotonicity of interface competence: a forty-fold span of capacity inside one family moves format competence not at all. A design that assumed the precondition would have reported only the loss. What remains is the Llama ladder and the level contrast of Section 6.

Other limitations. Yahoo does not carry delisted tickers, so the US panel is survivorship-tilted in a way that varies over the sample. Two roster cutoffs come from a corpus paper rather than a model card, and one sits only inside a shipped chat template; a documented cutoff is an upper bound on data recency, not a guarantee. The donut, the placebo cutoffs and the recall probe are the response to both, and a check rather than a fix.

The scope of the construct-validity result. It rests on three models from three laboratories, 1.2B to 3B, and on synthetic dated facts; a fourth was excluded because training broke its readout. One counterexample refutes an implication; three from separate laboratories make it more than a quirk. Whether it holds at larger scale or in other domains is not established.

What the positive control licenses, and what it does not. The identity probe shows this elicitation reads entity knowledge and the injection test shows the contrast recovers a step of known size; neither shows that a model holding a closing price would surrender it, and Section 5.3 shows one need not. The price-level arm is the weaker, our prices being split-adjusted where training text reports the raw price; the return-ordering probe carries no numbers and is the one the paper leans on.

AI USE STATEMENT

The research question, the design of the controls and probes, and the architecture of the analysis pipeline are the authors'. In the categories this venue requires to be disclosed, we used generative AI tools to provide feedback on the research methodology and the experimental design, which the authors discussed with an AI assistant while developing them; to implement methods, where the assistant wrote the scoring, estimation and verification code and executed the runs on the authors' hardware; to clean and reformat the market data panels; to generate the synthetic-exposure corpus of invented companies; and to help interpret results. We have not used them to formulate mathematical claims or proofs, and translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used them to search and summarise the literature, to draft prose the authors revised, and to edit for readability. The admission thresholds and the pooled specification were fixed by the authors before any result was read. We have reviewed all AI-assisted work, including the code and its outputs, and take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

Two habits of the artifact bear on how far that assistance can be trusted, and both ship with it. No quantity in this paper is typed by hand: each is recomputed from the result files by a checker that requires its rendered form to appear in the section that claims it, and that checker is what caught four figures which had been computed on stale samples. Every derived result is also asserted to be newer than the data behind it, because a stale analysis and a current one are indistinguishable by reading.

REFERENCES

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Sebastian Calonico, Matias D Cattaneo, and Rocio Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of ACL-IJCNLP*, pp. 1860–1874, 2021.

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. Extracting training data from large language models. In *30th USENIX Security Symposium*, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. Dated data: Tracing knowledge cutoffs in large language models. In *Conference on Language Modeling (COLM)*, 2024.

Yize Cheng, Wenxiao Wang, Mazda Moayeri, and Soheil Feizi. Dyepack: Provably flagging test set contamination in LLMs using backdoors. In *Proceedings of EMNLP*, 2025.

Antoine Didisheim, Matteo Fraschini, and Luciano Somoza. AI's predictable memory in financial analysis. *Economics Letters*, 256:112602, 2025.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024.

Zhenyu Gao, Wenxi Jiang, and Yutong Yan. Detecting lookahead bias in llm forecasts. *arXiv preprint arXiv:2512.23847*, 2026.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of EMNLP*, 2021.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning (ICML)*, 2022.

David Lehner and Alejandro Lopez-Lira. Chatgpt as a time capsule: The limits of price discovery. *arXiv preprint arXiv:2604.21433*, 2026.

Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. The memorization problem: Can we trust llms' economic forecasts? *arXiv preprint arXiv:2504.14765*, 2025.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black box language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP*, 2019.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations (ICLR)*, 2023.

Anisha Roy and Dip Roy. Memguard-alpha: Detecting and filtering memorization-contaminated signals in llm-based financial forecasting. *arXiv preprint arXiv:2603.26797*, 2026.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design. In *International Conference on Learning Representations (ICLR)*, 2024.

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.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *Conference on Language Modeling (COLM)*, 2024.

Ke Wu, Baozhong Yang, Zhenkun Ying, and Dexin Zhou. Anonymization and information loss. *arXiv preprint arXiv:2511.15364*, 2026.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for detecting pre-training data from large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024.

Taojie Zhu et al. From knowing to doing: A memory-controlled benchmark for llm trading agents on stock markets. *arXiv preprint arXiv:2605.28359*, 2026.

A A PROBE-CONSTRUCTION ARTIFACT, MEASURED

The obvious way to build the price-level probe is wrong, and the error looks exactly like the effect being hunted. We report it because anyone building a similar probe will hit it.

The natural construction draws the distractor from the same company's prices at any other date. Equities trend, so at an early probe date most other prices lie above the true one and at a late date most lie below. A model that merely prefers the larger of two numbers then appears to improve over calendar time. Measured on Phi-2 over 17,807 items, that preference is not subtle: where the distractor lies below the true price, so that the true price is the larger number, accuracy is 0.750 on 9,599 items; where it lies above, accuracy is 0.268 on 8,208 items.

A 48 point swing, with no memory involved. The share of distractors lying above the true price is not constant over the sample: rebuilt with the same panel and rules, it falls from 0.48 over the first third of probe dates to 0.41 over the last, and swings between 0.20 and 0.83 at single dates as the market moves. A falling share of the items the preference gets wrong is a rising accuracy, so the artifact produced a spurious upward trend in raw accuracy and a spurious 2.96 point discontinuity at the cutoff, with the wrong sign and t of 3.73.

Two fixes, both applied. First, the direction of the distractor is drawn at random and then balanced exactly within every probe date, by discarding the surplus side; any magnitude preference then scores exactly fifty percent at every date, so it cannot generate a trend or a step. Second, the estimator is also reported stratified on that direction, which recovers an unbiased estimate even from the unbalanced item set. On the Phi-2 data the stratified estimator turns the spurious 2.96 point drop into 0.67 points with t of 0.94.

Exact balancing costs 8 percent of items in the US probe and 10 percent in the Taiwan probe. The unbalanced run is retained in the released artifact so that the artifact and its correction can both be reproduced.

B THE TAIWAN SCOPE BOUNDARY

A second market, Taiwan, was screened the same way: 1,377 tickers, 1,324 with price data, 1,119,639 weekly rows, and 255,339 of those inside a top-300 trailing-dollar-volume universe over the same window as the US panel. The local Taiwan dataset holds daily returns with no price levels, which destroys the memorised-price handle the probe needs, so prices are pulled from the same free source as the US panel and the local matrix is used only to define the ticker universe.

Taiwan is not a redundant second market: its characteristic factors are far stronger than the US ones over the same window. A point-in-time ridge earns a weekly rank IC of 0.0238 with $t = 4.26$ and an annualised information ratio of 1.16, against 0.0041 and $t = 0.71$ in the US. A leakage effect can only inflate a backtest that has something to inflate, so a market with real factor structure is where such an effect should be most visible.

But the Taiwan arm turns out to be out of scope, and the positive control is what establishes that. Memorising a company-specific fact presupposes knowing the company, and asked which company a ticker belongs to the seven admitted models scored on both markets average 0.903 on US tickers and 0.500 on Taiwan ones, which is chance; 361 of the 713 Taiwan tickers asked about are at or below it. The few globally known names are recognised, TSMC at 0.90 and MediaTek at 0.87, and the market around them is not. Per-model numbers are in Table 2.

A Taiwan probe therefore measures the absence of the precondition rather than the absence of memory, so we report it as a scope statement rather than a second null. The positive control earns its

place twice: once to make the US null interpretable, and once to show where the method does not apply.

Table 2: Ticker-to-company identification accuracy, the positive control behind the Taiwan scope decision, for every admitted model scored on both markets.

Model	Cutoff	US ticker → company	TW ticker → company
Gemma-3-4B	2024-08	0.935	0.505
Llama-3.2-1B	2023-12	0.845	0.492
Llama-3.2-3B	2023-12	0.897	0.502
Phi-2	2023-10	0.920	0.497
Phi-3-mini	2023-10	0.897	0.497
Phi-4-mini	2024-06	0.897	0.497
SmolLM3-3B	2025-06	0.927	0.513

C Reconciliation with MemGuard-Alpha

MemGuard-Alpha (Roy & Roy, 2026) is the closest work in motivation and the one whose findings sit hardest against ours. It scores each LLM signal using five membership inference attacks plus a Cross-Model Memorization Disagreement statistic that, like our design, exploits cutoff-date variation across models. On seven LLMs from 124M to 7B, fifty S&P 100 names and 42,800 prompts over 2019–2024, filtering on its score raises Sharpe from 2.76 to 4.11: in-sample accuracy rises 40.8 to 52.5 percent with contamination while out-of-sample accuracy falls 47 to 42 percent.

An 11.7 point accuracy spread is an order of magnitude above the 1.1 point effect our pooled probe can resolve, so if that spread reflected date-specific memorisation of the kind a backtest would inherit, our design would see it plainly. We do not.

The two results need not conflict, and we do not claim a refutation. Three differences could produce both: their spread sits across buckets of their own outcome-correlated score, which induces a spread whether or not memorisation is present, while ours is a model's own accuracy either side of a date fixed independently of any outcome; membership inference asks whether a text was in the training data, whereas we ask whether a specific dated fact can be retrieved, and a model can have seen text about a company without recalling which of two returned more in a given month; and their in-sample-up, out-of-sample-down crossover is also the signature of ordinary overfitting to a selected subsample.

Our pooled estimates, over the models that pass the format control, are $+0.41$ points (95% CI -0.54 to $+1.37$) on the price-level probe and -0.61 points (-1.97 to $+0.75$) on return ordering; both upper bounds sit under a quarter of their 11.7-point spread. That bears on the second difference only: a null on dated fact retrieval beside a large membership-inference spread is what one expects if their score reflects broader corpus exposure rather than the memorisation a backtest would inherit. It says nothing about the other two, which do not turn on the size of a fact-retrieval effect. MemGuard-Alpha computes no minimum detectable effect and uses no positive control, so there is no published power figure to compare against ours.

D The return-ordering probe around the cutoff

The event-study view of the return-ordering probe, which is the picture a reader of the memorisation literature will expect. It is here rather than in the main text because under this paper's framing it is the demonstration application, not the claim: what the paper turns on is the two validity conditions developed in Section 5, and this figure is what the first of them makes readable.

E Per-model probe scores

Per-model scores for every admitted model, on the named arm of each probe. The ranges quoted in Section 6 are the extremes of these columns. Every admitted model has a score on every probe; the

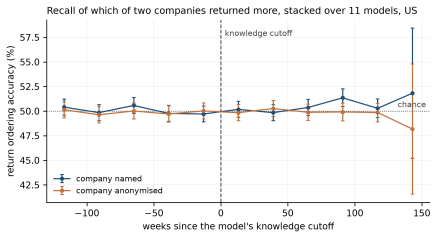


Figure 2: Return-ordering accuracy against weeks since each model's own cutoff (vertical line), pooled over admitted models, named against anonymised. This is the instrument's demonstration application, not the paper's claim; the anonymised arm sits at a chance baseline fixed by construction, not at an estimate.

pooled estimates use eleven of the twelve, because one is a ladder rung sharing its anchor's cutoff and contributes no cutoff of its own.

model	size	format	ticker to company	dated price	dated ordering
Llama-3.2-1B	1.2B	0.917	0.845	0.512	0.509
Phi-2	2.7B	1.000	0.920	0.508	0.512
SmolLM3-3B	3.0B	1.000	0.927	0.492	0.502
Llama-3.2-3B	3.2B	1.000	0.897	0.521	0.503
Phi-3-mini	3.8B	1.000	0.897	0.495	0.502
Phi-4-mini	3.8B	1.000	0.897	0.506	0.515
Gemma-3-4B	4.3B	1.000	0.935	0.505	0.505
OLMo-1.7-7B	7.0B	0.850	0.805	0.501	0.503
OLMo-2-7B	7.0B	1.000	0.933	0.506	0.505
OLMo-3-7B	7.0B	1.000	0.951	0.508	0.500
Llama-3-8B	8.0B	1.000	0.970	0.503	0.505
Llama-3.1-8B	8.0B	1.000	0.956	0.503	0.502

Table 3: Admitted models, named arm. Entity knowledge runs from 0.805 to 0.970 and gains 2.4 points per doubling of parameters, though not monotonically. No dated probe result reaches 0.522 or falls below 0.492, at any size.

The column to read across is the fourth against the fifth and sixth. Ticker to company spans 16.5 points across the admitted set, from 0.805 to 0.970, and is positively but not monotonically associated with capacity: Spearman correlation with parameter count 0.62, and a regression on log parameters gives 2.4 points per doubling. The dated columns span 2.9 and 1.5 points over the same models and neither is associated with capacity at all. Whatever grows with scale here, it is not date-specific recall.

The excluded models are not listed with scores. At chance those numbers are equally consistent with a model that knows nothing and one that cannot read a labelled list, and Appendix F shows that two of the seven are demonstrably the latter.

F THE CONTROL UNDER FOUR ANSWER CONVENTIONS

A control that asks one thing in one way cannot tell a model that fails the question from a model that fails the notation. If the seven exclusions were an artifact of scoring the tokens "A" and "B", the paper's central claim would be about a prompt convention rather than about the models, so we ran the same thirty questions under four answer conventions on all nineteen models: the published A/B list, the same list labelled 1/2, the same list labelled (a)/(b), and a yes/no frame that removes the list and asks about one candidate at a time. Every arm keeps the both-orders balancing, so a fixed preference for the first label scores exactly one half under all four.

model	size	A/B	1/2	(a)/(b)	yes/no
Llama-3.2-1B	1.2B	0.917	0.650	0.650	0.603
Phi-2	2.7B	1.000	1.000	1.000	1.000
SmolLM3-3B	3.0B	1.000	0.883	0.633	0.948
Llama-3.2-3B	3.2B	1.000	1.000	1.000	0.983
Phi-3-mini	3.8B	1.000	1.000	1.000	1.000
Phi-4-mini	3.8B	1.000	0.850	1.000	1.000
Gemma-3-4B	4.3B	1.000	1.000	0.833	0.948
OLMo-1.7-7B	7.0B	0.850	0.850	0.633	0.914
OLMo-2-7B	7.0B	1.000	1.000	0.750	0.966
OLMo-3-7B	7.0B	1.000	1.000	0.917	0.966
Llama-3-8B	8.0B	1.000	1.000	1.000	0.966
Llama-3.1-8B	8.0B	1.000	1.000	0.983	0.983
Llama-2-7B	7.0B	0.500	0.517	0.517	0.983
OLMo-1-7B	7.0B	0.517	0.650	0.550	0.983
Pythia-160M	0.16B	0.600	0.533	0.483	0.431
Pythia-410M	0.41B	0.500	0.483	0.500	0.655
Pythia-1.4B	1.4B	0.483	0.483	0.500	0.759
Pythia-2.8B	2.8B	0.450	0.567	0.483	0.560
Pythia-6.9B	6.9B	0.500	0.533	0.467	0.776

Table 4: The format control under four answer conventions, all nineteen models. The blocks are the three classes named below. Admission is decided by the first column alone, because the dated probes use that convention.

The exclusions are not a token artifact. No relabelling brings any excluded model across the admission threshold. The Pythia rungs move by at most nine points between the three list conventions, and OLMo-1-7B's best list score is 0.650 under numerals, short of both 0.80 and the required z.

But the exclusions are not all the same failure. Removing the list rescues two of them. Llama-2-7B and OLMo-1-7B answer the yes/no frame at 0.983, so neither is a model that cannot answer a forced choice. That pattern is what a failure of multiple-choice symbol binding looks like (Robinson et al., 2023), and we report it as consistent with one rather than as proof: the yes/no frame also changes the prompt structure, the number of candidates shown at once and the response space, so it isolates a cluster of differences of which label binding is the salient one. Either way their nulls on the dated probes remain uninterpretable, since those probes are labelled lists too.

The yes/no redesign was run, and it reaches no further. Because two models read that frame and no labelled list, the return-ordering probe was rebuilt in it: one pair at a time, no options displayed, 2,808 items per model and arm, each pair asked in both directions so that a constant answer scores exactly one half. Six models were scored, the two rescued ones among them.

model	named	anonymised	share answered Yes
Llama-2-7B	0.495	0.500	0.959
OLMo-1-7B	0.499	0.500	0.998
Phi-2	0.500	0.500	0.990
Llama-3.2-1B	0.500	0.500	1.000
Llama-3-8B	0.500	0.500	1.000
OLMo-3-7B	0.500	0.501	0.997

Table 5: The return-ordering probe asked as yes/no, on the two models that fail every labelled list and on four that pass. No cell is more than 0.5 points from chance. The third column is why: the models answer Yes to nearly everything, and the balanced design converts a constant answer into exactly one half rather than into a spurious signal.

No cell is more than 0.5 points from chance, and the diagnostic says why rather than leaving it to inference: the six answer Yes to between 95.9 and 100.0 percent of items. Llama-3-8B, which answers this same question at 0.970 when the answer is printed in the prompt (Section 5.2), answers Yes to every one of the 2,808.

This is the readout-adaptive analysis the admission rule invites: rather than asking every model the one convention we published, ask each the convention it demonstrably reads. It does not change the finding, and it is reported for the narrow thing it settles — the exclusions are not concealing a dated signal that a friendlier readout would have found. It settles nothing about whether these models hold the facts, which is the separate question of Section 5.3 and cannot be answered on facts whose presence in training is unknown.

A standard debiaser has nothing left to remove. The sensitivity literature's reading of a forced-choice null is that it reflects the model's prior over the option labels rather than its knowledge, and it offers corrections: PriDe (Zheng et al., 2024) estimates that prior by permuting options on a subset and subtracts it, and contextual calibration (Zhao et al., 2021) estimates it from a content-free prompt. The balanced design already removes it, and exactly rather than by estimation. Writing the score for one item as the log-probability gap between the two answer letters, asking the same fact in both option orders gives $g_1 = \text{content} + c$ and $g_2 = -\text{content} + c$, so $c = (g_1 + g_2)/2$ for that item, with no subset and no estimate.

probe	raw	PriDe	balanced
entity identity	0.912	0.900	0.895
price level	0.504	0.503	0.500
return ordering	0.502	0.504	0.500

Table 6: Three scorings of the same items, averaged over the 12 admitted models. *Raw* takes the sign of the gap and keeps the prior; *PriDe* subtracts one global prior estimated from a random five percent of facts; *balanced* is the published number, which is both-orders scoring. The debiaser moves no cell by more than 0.024, and the dated probes stay within 0.024 of chance under all three. The entity control does not move either, which is what shows the instrument still reads what is present.

The per-item prior is not a constant across items, which is why the exact correction is the stronger one: its standard deviation within a model and arm runs from 0.08 to 4.21 log-probability units. A single global estimate is an average over that spread. The null survives either correction, and it is not a label prior.

The Pythia ladder result is unaffected. No rung clears the threshold under any convention, the yes/no arm included, across a forty-fold range. The contrast the paper leans on hardest is the one the alternative conventions leave exactly where it was.

Readability is a property of the model and the elicitation jointly, and one admitted model shows it sharply. Eight of the twelve admitted models clear the threshold under all four conventions. Four do not, and Llama-3.2-1B is the extreme case: admitted at 0.917 on A/B, it falls to 0.650, 0.650 and 0.603 under numerals, parentheses and yes/no, so its admission rests on the one convention we happened to publish. SmolLM3-3B, OLMo-1.7-7B and OLMo-2-7B each fail exactly one alternative. We report this rather than bury it because it qualifies the claim in the right direction: a format control licenses a reading of probe results *obtained under that format*, not a general verdict on which models can be probed. It is also why the rule is stated with its convention named.

Appendix E gives each admitted model's score on every probe, which is the other half of this picture: what a model does with the format, and what it then does with the question.

We therefore report three classes: twelve models readable under the published control, two that fail every labelled list but answer the yes/no frame, and five that fail the admission criterion under all four. Admission still turns on the published A/B control alone, because the identity, price-level and return-ordering probes are labelled lists in that same convention, and a model that cannot read one cannot be read by them.

G SPECIFICATION ROBUSTNESS

A null reported at one bandwidth and one donut width is not a null, it is a specification. So the pooled estimate of Section 6 is re-run across the choices we made, one process per cell, and the table reports the range rather than the point.

Two grids. The first crosses four bandwidths (52, 104, 156 and 208 weeks) with four donut widths (0, 2, 4 and 8 weeks around each cutoff), which covers both the choice of how much of the series to use and the choice of how much to discard next to the boundary where cutoff dates are least certain. The second leaves out, one at a time, each of the eleven models in the pooled estimate, which is the check that matters when the estimate rests on a handful of models: it asks whether any single one of them is carrying the result.

Every cell is fitted on the sample the headline uses, the eleven admitted models with probe data on both sides of their own cutoff, and the cell at the published bandwidth and donut returns the published estimate to the digit, -0.61 on return ordering and $+0.41$ on price level. That is the check that the grid and the headline are the same estimator rather than two nearby ones.

probe	grid	cells	τ range (pp)	smallest p	largest $\lvert t\rvert$
return ordering	bandwidth $\times$ donut	16	-0.61 to $+0.50$	0.38	0.87
return ordering	leave-one-model-out	11	-0.87 to -0.27	0.22	1.22
price level	bandwidth $\times$ donut	16	$+0.25$ to $+1.39$	0.13	1.52
price level	leave-one-model-out	11	$+0.16$ to $+0.58$	0.28	1.08

Table 7: Specification robustness. Every cell is the same estimator with one choice varied. Not one of the 54 cells rejects at the 5 percent level.

The return-ordering probe, which is the one a backtest would turn on, stays inside a band about one point wide across every bandwidth and donut we tried, and its largest $|t|$ anywhere in the grid is 0.87, not within a factor of two of significance.

The price-level probe is looser and has a gradient worth naming rather than averaging away: it runs from $+0.25$ at the widest window with no donut to $+1.39$ at the narrowest window, so the estimate does grow as the window tightens, and the narrow cells are both the ones closest to the cutoff and the ones with the least data. Two things argue against reading it as memorisation surfacing. Its smallest p over the whole grid is 0.13, so nothing here rejects. And the sign is wrong: τ is the named-arm advantage after the cutoff minus the same advantage before it, so memorisation of the pre-cutoff period predicts a negative coefficient, while every price-level cell in the grid is positive. A memory effect that appears only after the model stopped reading is not a memory effect.

Leaving out any single admitted model leaves the return-ordering estimate between -0.87, without Phi-2, and -0.27, without OLMo-3-7B, and produces no rejection anywhere; the price-level estimate stays between $+0.16$ and $+0.58$. No one model is carrying either result. That check is worth more here than it usually is, because the format control cut the admitted set to a size where one influential model would be a real concern.

Inference without the asymptotics. Clustering on ticker is asymptotic in the number of clusters, and while 499 tickers is comfortable, the treatment indicator varies at the model-week level and the format control cut the roster of nineteen to the twelve models the estimate may use, eleven of which enter it. That is the regime where a cluster-robust p-value is known to be optimistic, so we do not rely on it alone. We impose the null, resample Rademacher weights on the residuals by cluster, refit, and read the p-value off the bootstrap distribution of the t-statistic.

probe	bandwidth	τ (pp)	t	wild bootstrap p
return ordering	52	-0.32	-0.27	0.78
return ordering	104	-0.48	-0.64	0.51
return ordering	156	-0.61	-0.87	0.37
return ordering	208	-0.54	-0.78	0.42
price level	52	$+1.39$	$+1.52$	0.13
price level	104	$+0.72$	$+1.30$	0.21
price level	156	$+0.41$	$+0.85$	0.37
price level	208	$+0.45$	$+0.95$	0.35

Table 8: Wild cluster bootstrap-t with the null imposed, 16,000 replications in total, 2,000 per specification. No bandwidth rejects for either probe, and the return-ordering probe, the one a backtest turns on, does not come close.

The bootstrap agrees with the asymptotic inference rather than overturning it, which is the useful outcome: it means the interval we report is not an artifact of having few models. At the published bandwidth the bootstrap p is 0.37 on both probes, against asymptotic values of 0.38 and 0.39. The smallest bootstrap p anywhere in this paper is 0.13, on the price-level probe at the narrowest window, which is the same cell that tops the grid above and carries the sign memorisation does not predict. Nothing here reaches conventional significance under either method.

H TIME-SERIES FOUNDATION MODELS AS A NEGATIVE CONTROL

A cutoff is only a treatment boundary if something could have crossed it. Time series foundation models give a population where the answer should be no: they are trained to forecast numeric series, and if their corpora contain no individual equity prices then a cutoff in one of them separates nothing, and an estimator that finds a jump there is finding an artifact.

We audited the pretraining corpora of eleven such models against primary sources rather than against survey descriptions. None names an individual equity price dataset. Chronos holds the finance-flagged subsets of the Monash archive out of training entirely, which is a documented exclusion rather than an absence of mention. Three corpora are closed at the dataset level, and the audit records those as unknown rather than clean, because a corpus nobody can inspect is not evidence of anything.

Where the corpus is documented this is therefore a cutoff without equity memory behind it, and no jump should appear at it. That makes it a negative control by construction rather than a comparison arm: it is not a claim about what those models can forecast, only about what they could have memorised about a named company on a given date. The audit, with the quotation and URL behind each entry, ships with the artifact.

I ATTEMPTS TO TRAIN THE READOUT, AND WHY THEY ARE NOT REPORTED AS RESULTS

Three attempts were made to teach the answer mapping directly, on the reasoning that a model which has never seen a fact of this kind answered in this form might simply not know that it should. All three failed, and they failed in a way worth recording: each degraded the forced-choice competence that Section 5.1 measures independently, so none of them produced a model whose probe result could be read at all.

Full-sequence tutoring. A disjoint set of thirty invented companies was written into the training text as solved questions in the probe's own format, in both option orders, alongside the target facts as prose. A length-matched control replaced the solved questions with more prose about the same tutor companies. Afterwards the tutored questions themselves were still unanswerable, at 0.500 and 0.479, so the manipulation check fails and the target result carries no information. The format control, measured on the adapter merged into the base weights, fell from 0.983 to 0.900 in the question arm and to 0.667 in the prose arm, the latter below the admission threshold. The log-probability gap on the probe collapsed from 2.13 to 0.50, which is consistent with the model having learned that the two labels are equally likely rather than which one the content implies — one of several explanations, and not one this experiment can identify.

Answer-token tutoring. The obvious repair is to take the loss on the answer token alone, so that reducing it requires using the content. Two branches were run from one shared checkpoint, differing only in whether the tutor labels were true or shuffled. Both collapsed. Every probe returned exactly one half, the format control included, and the log-probability gap became constant across items: 1.375 on every item in the shuffled branch, against a spread from 4.19 to 7.44 in the untutored model. A constant output is the loss-minimising response to a label the model cannot predict, and the shuffled branch — where the label genuinely is unpredictable — collapsed hardest. The control is what shows this is a property of the training regime rather than of the content.

One adapter, two format-control readings. The Phi-2 adapter these experiments start from scores 0.983 on the format control once merged into the base weights, and 0.875 read unmerged, which is how Section 5.3 reads every family so that all three pass through one pipeline. Merging a low-rank

adapter is the identity in exact arithmetic and is not in bfloat16; on this control it moves seven of sixty items. Both clear 0.80, so nothing depends on which is quoted, and the difference is itself a small instance of the paper's point: readability belongs to the model and the elicitation together, down to how the weights are stored.

We report these because a reader is entitled to know which experiments were run and not only which worked, and because they are the reason the construct control in Section 5.3 uses readouts that do not touch the weights.

J WHERE THE ROSTER LOSES MODELS

The roster loses models at four points and the paper quotes a different count at each, so the attrition is given once here.

step	why a model leaves	models
roster	—	20
scored	GPT-J-6B did not run to completion	19
pass the format control	7 fail it, 2 of those only on labelled lists	12
cutoff-bearing	Llama-3.2-1B shares its anchor's cutoff	11
in each pooled estimate	—	11

Table 9: Where the models go. Every count in the paper is a row of this table, and every exclusion is applied by the analysis code rather than by hand. The seven that fail the format control are characterised in Appendix F; the excluded ladder rung is scored and reported, it simply adds observations without adding a distinct cutoff.

K A NAMED-ARM ADVANTAGE THAT IS NOT MEMORISATION

Two admitted models show a named-minus-anonymised gain on the return-ordering probe that is statistically solid, and it is not memorisation of the period. It is here because the factorial that separates the two is a property of the instrument rather than a finding about markets.

A real effect that is not the effect we are looking for. On the return-ordering probe, two models show a named-minus-anonymised gain that is statistically solid: Phi-2 at $+1.36$ points ($t = 3.09$) and Phi-4-mini at $+1.47$ points ($t = 3.34$). Knowing the company's name does help them, a little and reliably.

It is not memorisation of the period. The same contrast differenced across the cutoff is null for both, at t of -1.35 and $+0.45$, and the largest cutoff step anywhere in the study is 1.55 standard errors. A model that had memorised the window would help itself before its cutoff and not after; these are helped equally on both sides, which is a standing cross-sectional prior about which companies tend to outperform: time-invariant, and not a leakage channel.

Separating the two is what the factorial is for: the pooled null alone would hide a real named-arm advantage, and that advantage alone would be mislabelled contamination.

L PER-MODEL FORMAT-CONTROL DETAIL

The per-model format-control scores, with the two diagnostic columns that say what kind of failure each one is. Admission turns on the accuracy alone.

The columns beyond accuracy are diagnostic. OLMo-1-7B chose the first option on 72 percent of items whose correct answer was the first half the time, and its log-probability gap varies by 0.54 across the difference between an apple and a hammer: the instrument is reading a flat line. Pythia-410M picks the first option on all sixty items, which the balanced design converts into exactly 0.500 rather than into something mistakable for knowledge. A narrow gap is not itself disqualifying — admitted Llama-3.2-1B spreads 0.83, narrower than models that fail — which is why the rule is whether a model answers correctly, not how widely its logits move.

model	params	format accuracy	z	picked A	score sd
Llama-3.2-1B	1.2B	0.917	6.5	0.483	0.83
Phi-2	2.7B	1.000	7.7	0.500	4.94
SmolLM3-3B	3.0B	1.000	7.7	0.500	6.45
Llama-3.2-3B	3.2B	1.000	7.7	0.500	3.52
Phi-3-mini	3.8B	1.000	7.7	0.500	6.28
Phi-4-mini	3.8B	1.000	7.7	0.500	6.11
Gemma-3-4B	4.3B	1.000	7.7	0.500	9.80
OLMo-1.7-7B	7.0B	0.850	5.4	0.617	1.08
OLMo-2-7B	7.0B	1.000	7.7	0.500	3.12
OLMo-3-7B	7.0B	1.000	7.7	0.500	4.33
Llama-3-8B	8.0B	1.000	7.7	0.500	3.44
Llama-3.1-8B	8.0B	1.000	7.7	0.500	3.61
Pythia-160M	0.16B	0.600	1.5	0.600	1.98
Pythia-410M	0.41B	0.500	0.0	1.000	0.74
Pythia-1.4B	1.4B	0.483	-0.3	0.983	0.20
Pythia-2.8B	2.8B	0.450	-0.8	0.717	0.34
Pythia-6.9B	6.9B	0.500	0.0	0.433	0.34
Llama-2-7B	7.0B	0.500	0.0	0.900	0.38
OLMo-1-7B	7.0B	0.517	0.3	0.717	0.54

Table 10: The format control, per model: admitted above the rule, excluded below. z tests accuracy against one half with the variance under that null, $0.25/n$ over the sixty items; *picked A* is the share of items answered with the first option, and *score sd* the spread of the log-probability gap.

M THE POWER CALCULATION AND THE DESIGN THAT CARRIES IT

That pre and post cutoff comparisons can fail to separate skill from memory is not new. Lopez-Lira et al. (2025) state it as an identification result: once a model has seen the realised values, "any observed output is consistent with both genuine skill and memorization," and their Proposition 2 adds that when the post cutoff window is short the two are not reliably distinguishable by hypothesis test. We cite that result and do not restate it as a discovery.

What is missing from the literature is a number. An identification result tells you a comparison can fail; it does not tell you whether the comparison you are about to run on your data will fail, by how much, or what would fix it. This section supplies that for a concrete estimator on a concrete panel, which is what a practitioner needs in order to decide whether their own null means anything. It needs no model outputs, which is why we report it first.

We estimate the noise from the data. Twelve characteristic factors and a point-in-time expanding-window ridge give a realised weekly rank IC standard deviation of 0.1615, a per-stock weekly skill standard deviation of 1.0294, a within-week skill correlation of 0.0221, and a cross-model IC correlation of 0.0337.

We feed those into the estimator the field actually uses: a local linear regression discontinuity on weekly rank IC, triangular kernel, four week donut, at a 104 week bandwidth (Calonico et al., 2014), which has a boundary variance inflation of 5.44 relative to a simple mean and therefore a standard error of 0.0532 on the jump.

Minimum detectable effect at 80 percent power: 0.149.

The largest leakage effect the literature implies is about 0.03. The design cannot separate that from zero. Table 11 reports the full comparison, including what happens as the bandwidth widens and as models are stacked (the full table ships with the artifact).

The estimator is correctly sized. Applying every one of the fifteen registered cutoffs to the thirteen baseline signals, which have no cutoff and for which the true jump is zero by construction, gives 195 fits of which 4.1 percent reject at the 5 percent level. The machinery does not manufacture discontinuities; it simply cannot see small ones once the cross-section has been averaged away.

Design	Unit	Effective n	SE	MDE
D0 single model, weekly IC, local linear RD	week	200	0.0532	0.1492
D3 stacked models, stock-week skill	stock-week	79,972	0.0051	0.0144
D5 stacked models, price recall probe	probe item	34,715	0.0038	0.0106

Table 11: Minimum detectable effect: the design the field uses, and the two that replace it. The unit differs by row, so the last column is not a single scale: D0 and D3 are in weekly rank IC and in per-stock skill, which share a scale by construction, while D5 is in probe accuracy, where 0.0106 means 1.06 percentage points. The three intermediate designs, which isolate the gain from pooling and from the stock-week unit separately, ship with the artifact.

M.1 A DESIGN WITH POWER

Same panel, same cutoffs, same estimator family, three changes.

Stock-week rather than week, which is most of the gain. Per-stock skill is $\text{skill}_{m,v,i,t} = 12\,(u^s_{i,t} - \frac{1}{2})(u^r_{i,t} - \frac{1}{2})$, where u^s and u^r are the percentile ranks of the model's score and of the forward return within week t. A uniform rank centred at zero has variance $1/12$, so the cross-sectional mean of skill in a week is that week's Spearman rank IC, and the two levels estimate the same object on the same scale, and 452 weekly observations become 196,000 stock-week ones after deflating for within-week correlation. **Pool the staggered cutoffs**, since nine distinct cutoff months put some models pre-cutoff and others post in most calendar weeks. **Absorb the regime with a calendar-week fixed effect**, which only becomes available once models are pooled, because a single-model design has one model per week and cannot separate the cutoff from the date.

Write $e_{m,t}$ for the running variable, the signed number of weeks between calendar week t and model m's cutoff, and $\text{post}_{m,t} = \mathbf{1}[e_{m,t} > 0]$. The outcome is indexed by prompt variant v as well as by model m, stock i and week t, since the same stock-week is scored once per variant:

$$\text{skill}_{m,i,t,v} = a_{m,v} + g_{t,v} + \tau_v \,\text{post}_{m,t} + b_1 e_{m,t} + b_2 e_{m,t}\,\text{post}_{m,t} + \varepsilon_{m,i,t,v} \tag{1}$$

Here $a_{m,v}$ is a model-by-arm and $g_{t,v}$ a week-by-arm fixed effect, so the local linear trend is free to break at the cutoff while the market regime is absorbed by the calendar week. Because $e_{m,t}$ is calendar time less a model constant, its level is spanned by those two fixed effects and b_1 is not separately identified; it is written for the shape of the local linear design, and what the fit estimates is the break in slope b_2 and the jumps. Observations within four weeks of a cutoff are dropped and the bandwidth bounds $|e_{m,t}|$.

The headline estimator is the triple difference $\tau_{DDD} = \tau_{\text{named}} - \tau_{\text{anon}}$, the jump with company identity minus the jump without it. It is read off one fit, as the coefficient on $\text{post}_{m,t} \cdot \mathbf{1}[v = \text{named}]$ with a ticker fixed effect added, rather than by differencing two regressions, which is what gives it a standard error. It removes model-specific ability drift because both arms score identical stock-weeks.

The clustering unit follows the unit of observation, and therefore changes when the design is applied to the probes. On the forecast task every model scores the same stock-week and the shock common to those observations is the week's market return, so standard errors are clustered by calendar week. The dated probes contain no forward return: their items are tied together by the company being asked about, so Section 6 clusters on ticker. Appendix G varies that choice along with the rest of the specification, and reports the wild cluster bootstrap that the small number of treated clusters calls for.

Section 6 applies this design, clustered on ticker, to the two probes: $+0.41$ points ($se = 0.49$) on the price-level probe and -0.61 points ($se = 0.69$) on the return-ordering probe, both statistically indistinguishable from zero over the eleven models that enter each.

N WHERE A MEMORISATION EFFECT WOULD CONCENTRATE

A pooled null can hide an effect that exists in one corner of the sample, and memorisation is a plausible candidate for exactly that: a period is more likely to be remembered where the move was

large, or where the company was heavily traded and therefore heavily written about. So the pooled estimate is re-run inside quartiles of two variables that stand for those stories: trailing dollar volume, on both dated probes, and the absolute return divergence within the pair, which exists only for return ordering. That is eight cells and four, twelve in all.

It does not concentrate anywhere. Eleven of the twelve cells fall between -0.18 and $+0.54$ points with $|t| \leq 1.29$. The twelfth is $+0.92$ points at $t = 2.15$, and it sits in the quartile where the two returns were closest together — the opposite end from where a salience-driven memory effect should appear, since a pair that finished neck and neck is the least memorable kind. One nominally significant cell out of twelve, in the wrong tail, is what sampling noise looks like, and we report it rather than dropping it because a grid with no significant cell anywhere would be the more suspicious result.

O THE FORECAST-TASK TRIPLE DIFFERENCE

Skill. Whether these models forecast at all is measured before any leakage claim, because a contamination correction to a signal that does not exist is not a finding. Restricted to admitted models, the weekly rank IC averages -0.0090 over 4,972 model-week observations and is positive in 49.3 percent of weeks; per-model means run -0.0037 to -0.0166. None is above zero. That alone does not rule out contamination, since a model below zero overall can still score better before its cutoff than after, and the triple difference below is the test of that.

The primary specification, reported for completeness. The triple difference on the forecast task, the jump with company identity minus the jump without it, removes model-specific ability drift because both arms score identical stock-weeks.

The admission rule of Section 5.1 applies here as it does everywhere else in the paper: a model that fails the format control is out of scope rather than a null. Thirteen models have forecast-arm scores and eleven of them are admitted. Over those eleven the triple difference is -0.0044 in skill units (se 0.0116, $p = 0.71$) at the 104-week bandwidth and $+0.0026$ (se 0.0110, $p = 0.82$) at 156 weeks. None of the eight placebo displacements rejects, the smallest p being 0.064 at a cutoff moved back two years.

Without the restriction, over all thirteen, it is -0.0055 (se 0.0109, $p = 0.61$) and -0.0038 (se 0.0097, $p = 0.70$), and there one placebo does reject, at $p = 0.036$. Both sets of estimates are null and the restriction changes no conclusion, but the restricted pair is the one consistent with the rule the rest of the paper applies, so it is the one the paper quotes; both ship with the artifact.

That is a null on a panel with no skill in it, which is why the probes rather than the forecast task carry the paper's claim.

P RECOVERY OF AN INJECTED EFFECT

The identity control shows that entity knowledge can reach this readout. It does not show that the estimator returns the right number. For a paper whose result is a null, that second thing has to be demonstrated separately, so we inject a leakage effect of known size into the real scored data and ask the estimator to recover it.

For the probes, a fraction of pre-cutoff items in the named arm are flipped from wrong to correct, so the induced accuracy step is known exactly, and the step is recovered as a difference of means: the named arm's pre-cutoff minus post-cutoff accuracy, less the same difference in the anonymised arm, over the admitted models. That is the contrast the pooled estimator targets, computed without its fixed effects, bandwidth or donut, so it is not the same number as the pooled τ and need not reproduce it at zero injection. It is also signed pre minus post, so a memorisation effect is positive here where it is negative in τ. For the forecast task, a multiple of the standardised realised forward return is blended into pre-cutoff scores only.

Injected steps of 0, 1, 2, 5 and 10 points are recovered as 0.28, 1.28, 2.28, 5.28 and 10.28 points on the price-level probe, and as 0.02, 1.02, 2.02, 5.02 and 10.02 on return ordering. The contrast reads $+0.28$ and $+0.02$ points with nothing injected, which is its own value on the unaltered scores, and above that baseline every injected step comes back to within a hundredth of a point: the slope is one

to three decimal places on both probes. On the forecast task, injected rank IC lifts of 0.005, 0.010, 0.020 and 0.050 are recovered as 0.0043, 0.0087, 0.0175 and 0.0437, a slope of about 0.87 with the estimate at zero when nothing is injected.

This licenses one step of the chain and not the whole of it. An effect present in the scores at these magnitudes is recovered by the estimator, so a null is not an estimation failure. It says nothing about whether the probe writes an effect into the scores when the model holds the fact, which Section 5.3 examines directly and does not find.