Distributional Readout: A Memorization Regime in Autoregressive Generative Models

Anonymous Authors[1]

Abstract

What memorization regime governs frontier autoregressive models that reproduce public numeric series? We argue from the next-token training objective that the regime is *distributional readout*: the loss induces a peaked conditional marginal at each highly-duplicated value-token but no relational structure across the conditioning set. This predicts *rank–value decoupling*: high-fidelity sample-based recall does not imply rank access on the same cells. We validate the regime two ways. Across eleven frontier LLMs from five providers, with controls for instruction-following, parser selection, position bias, and within-context capability, the dissociation holds; surfacing both values in one prompt restores $>90\%$ rank accuracy, locating the failure at cross-prompt elicitation rather than the marginal itself. A controlled LoRA fine-tuning experiment on an open 1.5B causal LM (Qwen-2.5-1.5B) trained on a synthetic date-indexed series recovers the regime end-to-end, demonstrating training is sufficient. Memorization in current frontier LMs is sample-from-able but not jointly queryable. Code: https://anonymous.4open.science/r/factor-leak-7D4E.

1. Introduction

This paper develops a diagnostic framework for separating memorization from structured reasoning in autoregressive deep generative models, using public numeric time series as a contamination testbed: the values are highly duplicated in pretraining, the ground truth is fully observed, and the same prompt protocol can be applied at frontier scale and under controlled fine-tuning of an open causal LM.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

When a frozen autoregressive model reproduces a public numeric time series at the API boundary, what kind of behavioral regime is the recall? Three readout regimes are observationally distinct even at fixed surface fidelity: a *prompt-queryable lookup* indexed by (factor, month), a *prompt-queryable function* that computes the value when asked, and a *distributional readout* that exposes a peaked marginal at the value-token position but is not jointly accessible to structured prompts. Distinguishing among them is a foundational question for autoregressive generative modeling: it asks not whether models memorize public series but *what kind of object* memorization behaves as, under controlled elicitation. We treat the Fama–French factor library and companion macroeconomic and equity-market series as a contamination testbed, not as the paper's domain claim; the regime characterization itself is domain-agnostic, and the controlled LoRA experiment in §4 replicates the regime under a synthetic series that has no finance content at all.

Read at the level of generative-model foundations, rank–value decoupling is a *compositional-reasoning failure mode*: a relational operation (ordering two values) cannot be performed across separate prompts on the very cells where each value is recovered to near-perfect fidelity in isolation. The model has the components but cannot compose them, and the failure tracks elicitation rather than capability: surfacing both values in one context restores comparison (§3.2, Tab. 2). This frames distributional readout as the natural memorization-side analogue of recently-documented compositional failures in code and reasoning benchmarks (Liang et al., 2025).

1.1. From training loss to distributional readout

Autoregressive LMs trained on conditional next-token prediction concentrate token mass on the value conditional on the prefix specifying date and factor; fidelity scales in training-data duplication (Kandpal et al., 2022; Carlini et al., 2023a). For each duplicated (factor,month) cell this produces a peaked conditional marginal $p_M(v \mid q_A(\text{factor}, \text{month}))$ at the value-token position: a sampleable distribution indexed by the prompt (Mireshghallah et al., 2022). The same objective induces no relational structure across the conditioning set: no gradient aligns the

marginal at one cell with the marginal at another. We name this regime *distributional readout*. The claim is not that the model lacks a comparison capability (the within-prompt probe in §3.2 confirms it has one); the narrower claim is that next-token training does not produce a representation in which the q_A-marginal at one prompt is accessible from a separate q_C prompt. The argument is testable under controlled exposure: §4 LoRA-fine-tunes Qwen-2.5-1.5B on a synthetic date-indexed series at four exposure levels and recovers the regime end-to-end.

Prior work on memorization in generative LMs characterizes the phenomenon either at the surface-token level (Carlini et al., 2021; 2023a) or as a property of the output distribution (Mireshghallah et al., 2022), but does not commit to which readout regime obtains, and does not test the predictions that each regime makes for cells where surface fidelity is high.

The account yields four behavioral predictions on a frozen LLM queried on a representative public numeric series (we use the Fama–French market excess return Mkt-RF; Fama and French 1992; 1993): cell-localized recall (P1), a parsability gate at training cutoff (P2), rank–value decoupling on high-recall cells (P3), and a capability-scaled entropy gap between memorized and fabricated cells (P4). The four are formalized in §2 (Eqs. 1 ff.); P3 is the headline test.

Findings. On Sonnet×Mkt-RF, the comparative probe gives 52.5% rank accuracy on the same months the model recalls at $r{=}0.98$ (37 pp below the MSB of 0.92; one-sided exact binomial $p{\approx}1.4{\times}10^{-14}$; Fig. 1; test of P3). The supporting findings P1, P2, P4 localize the regime (§3.1), and four simple alternative readings of P3 are each ruled out by a targeted control (§3.3). The regime is recovered under controlled fine-tuning of an open 1.5B causal LM on a synthetic date-indexed series (§4, Fig. 5), confirming that next-token training on date-indexed values is sufficient to induce distributional readout and that open-ended decoding under-reports memorization that logprob ranking detects.

Contributions. (i) We propose *distributional readout* as a behavioral characterization of the memorization regime, derived from the next-token training objective, and develop four predictions (P1)–(P4) as direct consequences. (ii) We test the predictions empirically across eleven frontier LLMs from five providers in $4{,}700{+}$ queries, with the headline test of P3 giving evidence against a prompt-queryable lookup/function account. (iii) We *validate the route from training objective to behavioral regime* on an open causal LM: LoRA fine-tuning Qwen-2.5-1.5B on a synthetic date-indexed series reproduces the regime, and a logprob-ranking probe shows that open-ended decoding under-reports the channel detected by ranking (§4). (iv) We

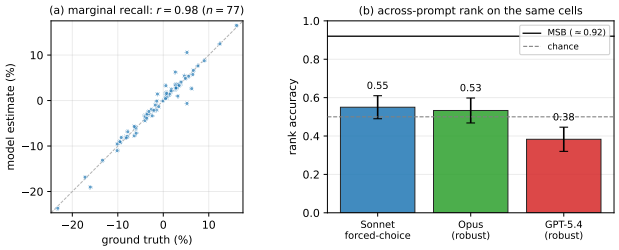

Figure 1. **Rank–value decoupling on the same cells. (a)** Sonnet recalls monthly Mkt-RF at $r{=}0.98$ ($n{=}77$ months) under marginal probe q_A. **(b)** On the same months, across-prompt rank accuracy under comparative probe q_C is at or below chance: Sonnet 0.55 ($n{=}60$, strict forced-choice), Opus 0.53 (position-balanced robust), GPT-5.4 0.38 (robust). The solid line is the marginal-sampling baseline (MSB, Eq. 1), the comparison accuracy any prompt-queryable account of the recall must clear; the dashed line is chance. Each model rejects $H_0{:}\, p \geq \text{MSB}$ at $p{\lesssim}10^{-14}$ (one-sided exact binomial, $n{=}60$; Sonnet $p{\approx}1.4{\times}10^{-14}$, Opus and GPT-5.4 $p{\lesssim}10^{-15}$).

replicate cross-domain (S&P 500, NASDAQ, blind label, UNRATE, CPI YoY at $r{\geq}0.99$) and on the synthetic LoRA series with no finance content, and derive an evaluation corollary that decomposes date-conditional covariation between an LLM-derived signal and a public series into recall- and generalization-attributable components (App. E).

Related work. Carlini et al. (2021; 2023a) characterize verbatim memorization in generative LMs and its capacity scaling; Tirumala et al. (2022) trace its training-time dynamics; Kandpal et al. (2022) tie recall fidelity to training-data duplication; Mireshghallah et al. (2022) reframe memorization as a property of the output distribution rather than the surface string. The autoregressive regime studied here differs from input-level verbatim memorization documented in diffusion models, where the memorized object is end-to-end reconstructable (Carlini et al., 2023b; Somepalli et al., 2023); here the memorized object is a peaked output-token distribution that is not jointly accessible across conditioning sets, suggesting memorization regimes may be model-class-specific. Recent work probes recall of stock prices (Lopez-Lira et al., 2025), macroeconomic series (Crane et al., 2025), and look-ahead bias in LLM-derived financial signals (Li et al., 2025; Benhenda, 2026; Sarkar and Vafa, 2024; Didisheim et al., 2025); none target the regime question itself. Our contribution is the regime-level characterization: what kind of object the recall is, with a controlled white-box demonstration that the next-token objective is sufficient to generate it.

2. Method

2.1. Problem Setup

Notation. A frozen LLM accessed at the API boundary, queried under a probe template, returns a number (the

model's elicited estimate) or refuses. We formalize: M is the model, $\mathcal{F}$={Mkt-RF, SMB, HML, RMW, CMA, Mom} the six canonical Fama–French factors (Fama and French, 1992; 1993; 2015; Carhart, 1997), $r_{f,t} \in \mathbb{R}$ the ground-truth value of factor f at month t, q_v a probe template for variant $v \in \{A, B, C\}$, and π_v a parser. The elicited estimate is

$$\hat{r}^{(M)}_{f,t,v} \;=\; \pi_v\big(M\big(q_v(f,t)\big)\big) \in \mathbb{R} \cup \{\perp\},$$

with $\perp$ a refusal or unparsable response. Let $p_M(\cdot \mid q)$ denote the next-token distribution emitted by M on prompt q at the first numeric-token position.

Distributional readout (formal). *Plain-language statement.* The model behaves on a cell as if its recall is a peaked sample from a stored marginal, exposed through the value-asking probe but not jointly accessible to a comparative probe.

Formal statement. M exhibits *distributional readout* on (f,t) if $p_M(\cdot \mid q_A(f,t))$ has high probability mass concentrated near $r_{f,t}$ *and* the marginal exposed by q_A is not jointly accessible under the comparative elicitation q_C (operationalized in §2.1). Three alternatives to test against are *prompt-queryable lookup* (the marginal exposed by q_A is accessible from q_C as a queryable indexable map), *prompt-queryable function* (a representation supporting in-prompt-recoverable structured queries), and *generalization* (no cell-specific encoding). We make no claim about the internal mechanism.

Marginal-sampling baseline (MSB). If two values are recovered accurately enough by sampling under q_A, then comparing them by independently sampling each and taking the sign of the difference must also be accurate. We formalize this lower bound on any prompt-queryable account of the recall: for a pair (t,t'), draw $\hat{r}_t \sim p_M(\cdot \mid q_A(f,t))$ and $\hat{r}_{t'} \sim p_M(\cdot \mid q_A(f,t'))$ independently and ask how often the sampled comparison agrees with the truth,

$$\Pr_{\text{ms}}[\text{sign}(\hat{r}_t - \hat{r}_{t'}) = \text{sign}(r_{f,t} - r_{f,t'})] \;=\; \mathbb{E}_{|Z|}\left[\Phi\big(|Z|/(\eta\sqrt{2})\big)\right], \tag{1}$$

where η is the marginal-probe RMSE on (M,f) and $Z \sim \mathcal{N}(0, 2\sigma_f^2)$ is the truth-pair gap with σ_f the unconditional std of r_f. The MSB is an elicitation-level lower bound, not a claim about internal mechanism: any account in which the marginal readouts exposed by q_A are also available to q_C (lookup, function, or joint distributional readout) must clear it. Comparison accuracy *below* the MSB is therefore inconsistent with the marginal being prompt-accessible to q_C.

Predictions. Distributional readout implies four predictions on the elicited estimates of M over a sample $\mathcal{T} \subset \mathbb{Z}$ of months. Each follows from the training-objective derivation in §1.1.

(P1) Cell-localized recall. Because the conditional marginal is indexed on (factor, month), recall is concentrated on cells seen at high duplication and absent on identically-formatted neighboring cells: $\rho\big(\hat{r}^{(M)}_{f^\star,\cdot,A}, r_{f^\star,\cdot}\big) \to 1$ on the memorized factor $f^\star$, while $\rho \to 0$ on $f \in \mathcal{F} \setminus \{f^\star\}$ at the factor-shuffle null under identically-formatted probes.

(P2) Parsability gate, not fidelity gradient. Because the conditioning set does not extend past training cutoff, out-of-support cells lie outside the support of any conditional marginal; behavior on them is qualitatively different (refusal, fabrication, unparsable output) rather than lower-fidelity: $\Pr[\hat{r} = \perp \mid t > t_{\text{cutoff}}] \to 1$ while ρ on parsed months is invariant to distance from cutoff. A continuous fidelity gradient on parsed months would be inconsistent with P2 and would favor function-storage.

(P3) Sampleable but not structurally queryable. Because rank comparison requires a relational operation over two conditional marginals that the training objective does not supply, comparison accuracy on cells where the marginal probe recovers values at high fidelity falls below the MSB. On cells where the marginal probe q_A recovers $\rho \geq 0.98$, the comparative probe q_C gives $\Pr[\text{correct}]$ strictly below the MSB (Eq. 1). Any account in which the high-fidelity marginal readouts exposed by q_A are prompt-accessible to q_C (lookup, function, or joint distributional readout) would predict $\Pr[\text{correct}] \geq \Pr_{\text{ms}}$.

(P4) Concentrated output distribution. Because the marginal on memorized cells is sharpened by training pressure on the prefix-conditioned value token while the marginal on out-of-support cells receives no such pressure, the next-token entropy at the value-token position is lower on memorized than fabricated cells, with the gap closing on lower-capability tiers: $H(p_M(\cdot \mid q_A(f^\star,t))) \ll H\big(p_M(\cdot \mid q_A(f^{\text{fab}},t))\big)$ on top-tier models, and shrinking with $\rho\big(\hat{r}^{(M)}_{f^\star,\cdot,A}, r_{f^\star,\cdot}\big)$.

2.2. Data

Ground truth is the Kenneth French Data Library (French, 2026): monthly returns for $\mathcal{F}$, with Mkt-RF, SMB, HML available since 1926-07, RMW/CMA since 1963-07, and momentum since 1927-01; values in percent, accessed April 2026 (CRSP build 202602). The probe window is 1963-07 through 2026-02. Per cell, we draw 120 months stratified by distance to model cutoff (50% pre-cutoff, 25% near ±6 months, target 25% post-cutoff backfilled with pre-cutoff months when the post-cutoff window is short) and 20 economically-prominent "famous" months (Black Monday, Lehman, COVID; full list in App. A.2). The Variant C comparative probe additionally samples 60 month pairs per cell.

2.3. Models

The full model panel comprises eleven LLMs from five providers. The main six-factor sweep covers Claude Sonnet 4.6 and Claude Haiku 4.5; the Mkt-RF and label-invariance sweep extends to Claude Opus 4.7 (Anthropic), GPT-5.4 / GPT-5.4-mini / GPT-5.4-nano (OpenAI), DeepSeek-V3.2 (DeepSeek), Llama-3.3-70B and Llama-3.1-8B (Meta), and Qwen-3 32B and Llama-4 Scout 17B (Groq endpoints). All queries are issued at `temperature`$= 0$; Claude Opus 4.7 rejects the temperature parameter and is invoked with vendor defaults (deterministic for our short-prompt setting). Per-variant token caps are A/C$=48$, B$=384$. Seed 42; transient API errors are retried with exponential backoff. Outputs are parametric: no tools, retrieval, web access, or attachments are used. Verbatim prompts are listed in App. A.1.

2.4. Probe Variants

Three prompts are issued per (model, factor, month) cell. Variant A is a direct numeric query ("return signed percentage"); Variant B is descriptive (free-form discussion required to include a best estimate); Variant C is a forced-choice comparative probe ("which of two months had the higher return"). The Variant C parser π_C (App. A.3) is endorsement-aware, prioritizing strong endorsements over last-mentioned tokens to avoid recency bias, and is robust on three independent operationalizations (App. A.4). The total main sweep covers $2\times6\times(240_{A,B}+60_C)$ queries; Mkt-RF follow-ups (App. D.5, App. A.5) add 240 multi-seed Mkt-RF queries on Opus and GPT-5.4 over seeds $\{1, 2, 3\}$ and 180 logprob queries on GPT-5.4-mini and GPT-5.4-nano.

2.5. Metrics and Controls

We report (i) within-$\{5, 25\}$ bps exact-match accuracy with Wilson 95% CIs, (ii) sign-match conditional on non-zero truth, (iii) Pearson r with percentile-bootstrap CI, (iv) per-cell OLS cutoff slope under BH-FDR at q=0.05, (v) famous-month lift, and (vi) Variant C forced-choice accuracy. The 25 bps tolerance has heterogeneous σ-content across the eight probed series (0.054–0.139σ); App. D.6 re-reports recall in σ-units. Two control conditions are central to the analysis. *C1 (factor-shuffle):* pair a factor's truth with the model's estimate for a *different* factor at the same month, ruling out trivial constant-prediction strategies. *C2/C3 (fabricated series):* probe two fictional series ("Gleason–Zeta residual factor", "Holbrooke–Mansfield Fund III (2007)"), establishing whether the model commits or refuses on prompts identical in surface form to the Fama–French probes.

3. Experiments and Results

We test the four predictions (P1)–(P4) of the distributional memorization hypothesis on the panel and probe protocol of §2. §3.1 characterizes the recall regime, jointly establishing P1, P2, P4; §3.2 reports the headline test of P3; §3.3 consolidates the simple alternative readings the data rule out; §3.4 reports cross-series replication and robustness.

3.1. Regime: cell-localized, cutoff-gated, low-entropy recall

Three observations together establish that there is a recalled representation to test in §3.2. They map to predictions P1, P2, P4 respectively.

Reporting convention. Headline rows are main-sweep (seed-42) where seed-stable, 3-seed pooled where seed-42 is a favorable draw; each row is labeled. Sonnet's main-sweep r=0.98 (n=77) reconciles with 3-seed pool r=0.92 (n=119; App. D.5).

Cell-localized recall (P1). Across the four top-tier models, only Mkt-RF clears the factor-shuffle null (within-25 bps $\leq$15% on every non-Mkt-RF cell; partial recall on SMB and HML is scattered and sub-threshold, Tab. 1). Sonnet's directional accuracy on Mkt-RF is 97.4% over 77 sampled months (Fig. 2); on non-target factors, Pearson correlations are indistinguishable from zero. Recall scales monotonically with capability inside each provider stack (Anthropic $0.99>0.92>0.27$, OpenAI $0.54>0.36>-0.32$, Fig. 3); GPT-5.4-nano is significantly negatively correlated with Mkt-RF ($r=-0.32$, 95% CI $[-0.57, -0.01]$), ruling out a generic financial-prior account. Full 9×6 grid in App. F.2; multi-seed analysis in App. D.

Parsability gate at training cutoff (P2). Out-of-support cells (post-cutoff months) elicit refusal, not continuous fidelity decay. Sonnet parses $70/70$ pre-cutoff Mkt-RF months against $0/5$ post-cutoff; Haiku $70/70$ vs. $0/1$ (App. D); recall on parsed months is unchanged near the cutoff (r=0.98 pre vs. r=0.99 near-cutoff on Sonnet). Per-cell cutoff-distance OLS slopes are small ($|\beta|\leq4\times10^{-4}$/month; none survive BH-FDR q=0.05, App. D); pooled small slopes on five of eight models ($|\beta|\leq$0.12/decade, Fig. 4) are consistent with prominence-recency rather than a sharp cutoff effect.

Concentrated next-token distribution (P4). A per-token logprob probe on the OpenAI stack (App. A) exposes the first-numeric-token distribution directly. On GPT-5.4, its Shannon entropy is $\sim 5\times$ lower on memorized cells than on fabricated cells, even when surface outputs are matched in form. The gap collapses on lower-capability tiers and tracks recall fidelity across the three OpenAI tiers (n=30 per condition $\times$ tier; Tab. 11), matching P4's capability-scaling prediction.

Table 1. **Mkt-RF is the only Fama–French factor recalled above chance**: within-25 bps of 34–68% and r of 0.54–0.99 across four top-tier LLMs, vs. $\leq 10\%$ and $|r| \leq 0.58$ on the best non-Mkt-RF factor. Wilson-score 95% CIs on proportions; bootstrap CI on r; "Sign" is conditional on non-zero truth. Bold rows: Mkt-RF. Full 9×6 grid in App. F.2 (Tab. 22); per-row provenance and seed-stability analysis in App. D.5.

Model	Factor	n	within-25 bps	Sign	Pearson r
Opus 4.7	**Mkt-RF**	40	**0.68** [0.52, 0.80]	**1.00** [0.91, 1.00]	**0.99** [0.97, 1.00]
Opus 4.7	HML (best other)	40	0.10 [0.04, 0.23]	0.68 [0.52, 0.80]	+0.58 [−0.29, 0.91]
Sonnet 4.6	**Mkt-RF**	77	**0.34** [0.24, 0.45]	**0.97** [0.91, 0.99]	**0.98** [0.96, 0.99]
Sonnet 4.6	HML (best other)	69	0.03 [0.01, 0.10]	0.49 [0.38, 0.61]	+0.48 [0.15, 0.68]
Haiku 4.5	**Mkt-RF** (3-seed pooled)	120	**0.12** [0.07, 0.18]	**0.65** [0.56, 0.73]	**0.27** [0.10, 0.43]
Haiku 4.5	SMB (best other)	70	0.03 [0.01, 0.10]	0.61 [0.50, 0.72]	+0.45 [0.24, 0.63]
GPT-5.4	**Mkt-RF** (3-seed pooled)	120	**0.33** [0.25, 0.41]	**0.80** [0.72, 0.86]	**0.54** [0.40, 0.66]
GPT-5.4	RMW (best other)	40	0.15 [0.07, 0.29]	0.65 [0.50, 0.78]	+0.28 [−0.50, 0.81]

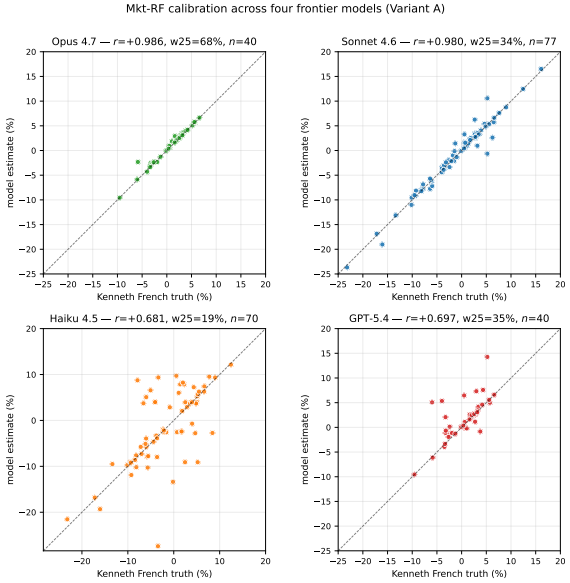

Figure 2. **Mkt-RF Variant-A calibration across four top-tier models (test of P1).** Opus and Sonnet produce near-perfect 45° alignment (r=0.99, r=0.98); Haiku is noisier (r=0.27 pooled); GPT-5.4 reproduces the alignment cross-provider (r=0.54 pooled). Recall is monotone in within-vendor capability on the Anthropic and OpenAI stacks. The full 12-cell grid showing every (model, factor) combination on Sonnet and Haiku is in App. F.1 (Fig. 10); the nine-model / four-provider capability picture is in Fig. 3.

3.2. Headline test: rank–value decoupling

P3 is the headline test: it most directly bears on the distinction between distributional readout and any account under which the marginal readouts exposed by q_A are prompt-accessible to q_C. We test it on the two Anthropic-stack top-tier cells with marginal recall at $r \geq 0.98$: (M=Sonnet, f=Mkt-RF) and (M=Opus, f=Mkt-RF).

Marginal-sampling baseline. The Sonnet×Mkt-RF marginal probe achieves r=0.98 on the probe sample. Treating the unconditional std as $\sigma_{\text{Mkt-RF}} \approx 4.5\%$ and using $\eta = \sigma\sqrt{1-r^2} \approx 0.9\%$ as the marginal-probe RMSE, Eq. 1 evaluates to $\text{Pr}_{\text{ms}} \approx 0.92$ on uniformly-drawn pairs from the empirical Mkt-RF distribution.

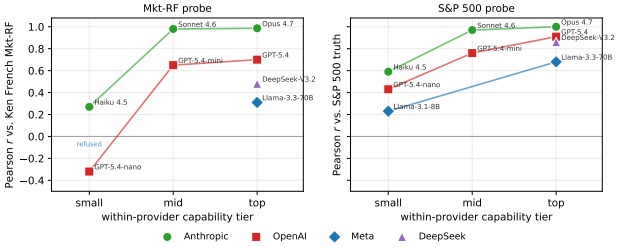

Figure 3. **Capability-scaled memorization across providers.** Pearson r vs. truth for the Mkt-RF probe (left) and the S&P 500 probe (right) plotted against within-provider capability tier (small/mid/top). Lines connect tiers within a provider stack. Recall is monotonic in capability inside every multi-tier provider line; the fourth provider lands on the same envelope. Mkt-RF panel: the smallest model in one stack falls below zero (anti-correlated noise), and the smallest model in another refuses every Mkt-RF query. Top-tier models from three independent providers saturate at $r \geq 0.86$ on S&P 500. *Single-seed-42 numbers shown*; 3-seed pooled estimates for the four headline models are in App. D.5.

Observed on Sonnet. The comparative probe gives chance-level rank accuracy across three independent operationalizations: the parsed-subset accuracy is 52.5%, the naive parser yields 49.2%, and a forced-choice rerun gives 55.0% (all 95% CIs cover 50%; Tab. 3 in App. A.4). The result is robust to parser choice and to refusal-based selection.

Replication and cross-vendor. The same protocol on Opus gives a position-bias-free robust accuracy of 53.3% (32/60, within sampling error of Sonnet's 55.0%; McNemar exact p=1.00). Cross-vendor on GPT-5.4 the robust accuracy is 38.3% (23/60). This is below chance, stronger than the framework's MSB-rate prediction; we attribute the residual to a vendor-specific elicitation bias not fully captured by our position-balance protocol. The load-bearing test is failure to clear MSB rather than chance-level performance specifically. The decoupling replicates across the Anthropic and OpenAI top tiers; full per-condition breakdowns and

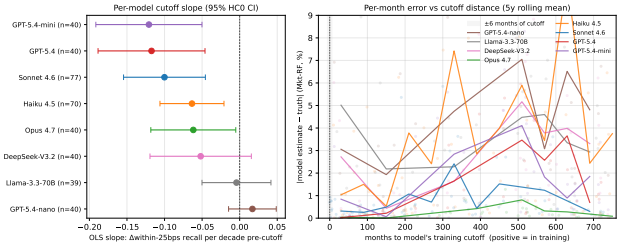

Figure 4. Per-model cutoff slope on Mkt-RF (8 models; Llama-3.1-8B excluded, $0/40$ parsed). *Left*: OLS slope of within-25 bps on months-to-cutoff in Δpp/decade with 95% HC0 CI. *Right*: |estimate − truth| vs. distance-to-cutoff (60-month bin-mean). Slopes are small ($|\beta| \leq 0.12$/decade) with no discontinuity at $\Delta=0$, consistent with prominence-recency rather than a sharp cutoff effect.

the Opus position-balance derivation are in App. A.4.

Effect size and significance. The forced-choice accuracy of 55.0% (33/60) lies 37 percentage points below the MSB of 0.92. A one-sided exact binomial test against $H_0: p \geq 0.92$ rejects at $p \approx 1.4 \times 10^{-14}$ (n=60, observed 33, expected 55.2 under MSB). The same test on Opus position-balanced robust (32/60) and GPT-5.4 robust (23/60) rejects at $p \lesssim 10^{-15}$. The result is inconsistent with a trivially prompt-queryable lookup or function account under our across-prompt comparative protocol, and supports the narrower view that the high-fidelity marginals exposed by q_A are not jointly accessible through q_C when the two cells are queried in separate prompts.

Within-prompt rank: the same model can rank when both values are in context. A natural reading of the result above is that the MSB is a strawman: the model never had the values "in mind" simultaneously, so failing the across-prompt comparison is incidental. We test this directly with a joint within-prompt probe that asks the model to (a) state v_1, (b) state v_2, and (c) pick the higher, all in a single response on the same 60 Variant-C pairs (script `experiments/70_joint_value_rank.py`; Tab. 2). When both cells are surfaced in one context, rank accuracy jumps to 91.8% on Sonnet, 98.1% on Opus, 86.4% on GPT-5.4 (Sonnet paired McNemar within-vs-across p=0.0005), with the model's pick internally consistent with its own stated values on $\geq$ 98.3% of records. The dissociation is therefore between elicitations: structured query succeeds once the marginal is surfaced in context, and fails when the same query is issued across separated contexts. This makes the across-prompt failure load-bearing rather than incidental: it locates the gap precisely at the training-objective-level absence of relational structure across the conditioning set.

Putting the pieces together. *(i)* Any account in which the q_A-marginal is accessible to q_C must clear the MSB (Eq. 1, derivation in §2.1). *(ii)* Observed across-prompt comparison fails MSB at $p \lesssim 10^{-14}$ on three top-tier models from two providers (Sonnet, Opus position-balanced, GPT-5.4 robust).

Model	across-prompt	within-prompt	internal cons.
Sonnet 4.6	0.550 (33/60)	0.918 (45/49)	1.00
Opus 4.7	0.533 (32/60, robust)	0.981 (53/54)	1.00
GPT-5.4	0.383 (23/60, robust)	0.864 (51/59)	0.98

Table 2. **Within-prompt ranking restores comparison.** Rank accuracy on the same 60 Variant-C pairs under across-prompt (q_C) and within-prompt (joint) elicitations. Across-prompt is at chance and far below the MSB ($\sim$ 0.92); within-prompt is far above chance (+37 to +45 pp), with the stated pick internally consistent with the model's own stated values on $\geq$ 98% of records. The same channel exposes the values; only the across-prompt elicitation fails to compose them.

(iii) Therefore the q_A-marginal is not accessible to q_C under across-prompt elicitation. *(iv)* The within-prompt probe (Tab. 2) recovers the values at $r \geq +0.84$ within-context Pearson and supports correct rank at $\geq$ 86%, ruling out a capability-gap reading. The dissociation is therefore a representation-access gap, not a capability gap.

3.3. Simple alternatives ruled out

The chance-level rank accuracy on high-recall cells admits four simple alternative readings. Each is testable; each is ruled out by the data.

Generic instruction-following gap (HML control). The pairwise comparison format itself could be out-of-distribution for instruction-following, with the model failing for elicitation reasons rather than from a property of the recalled representation. This account predicts uniform chance on Variant C across factors regardless of marginal recall. On Sonnet, Variant C gives 65.5% rank accuracy on partial-recall HML pairs (r=+0.48, 95% CI [53.3%, 77.7%]) against 52.5% on high-recall Mkt-RF (Tab. 7 in App. A.4). Rank tracks recall on partial cells and decouples from values precisely on the high-recall cell, a pattern a generic format failure cannot produce.

Parser selection bias. The chance-level result on the parsed-subset accuracy could reflect a parser that filters out exactly the correct answers. We rerun under a strict forced-choice format that drives parse rates above 95%, removing parser-dependent selection. The result is unchanged (55.0% on Sonnet, n=60, 95% CI covering 50%); the naive parser gives 49.2% and the endorsement-aware parser 52.5%. All three operationalizations agree.

Position bias (Opus). Opus's strict-forced-choice rate of 75.0% in original presentation order is contaminated by a strong first-position preference: Opus picks the literally-first month 42/60 times against ground-truth-first 29/60, with conditional accuracy 96.6% when truth is first and 54.8% when second. We position-balance by re-running the same 60 pairs with positions swapped and report the position-bias-free robust criterion (correct under *both* orderings). Robust accuracy is 53.3% (32/60), within sampling error

of Sonnet's 55.0% (McNemar exact p=1.00).

Capability gap. The across-prompt failure could reflect the trivial reading that the model simply lacks the capability to compare these values (the MSB-as-strawman worry). We test with a joint within-prompt probe that asks for v_1, v_2, and the higher of the two in a single response on the same 60 pairs (script `experiments/70_joint_value_rank.py`). Within-prompt rank accuracy is 91.8% on Sonnet, 98.1% on Opus, and 86.4% on GPT-5.4; the model's pick is internally consistent with its own stated values on 49/49 Sonnet records. The dissociation is therefore between elicitations: structured query succeeds once the marginal is surfaced in context, fails when the same query is issued across separated prompts. This locates the failure at cross-prompt elicitation rather than capability, which is what the framework predicts.

Provider-specific fabrication (not a prediction). A related observation falls outside the framework: Anthropic models refuse 100% of fictional-factor queries while non-Anthropic providers commit at 98.3% (App. B.2); the asymmetry tracks vendor lines, not capability tier, consistent with a post-training origin rather than the recall channel itself.

3.4. Cross-series replication and robustness

The behavioral profile is not Fama–French-specific. Top-tier recall persists on S&P 500, NASDAQ, a blind "U.S. market excess" probe ($r \geq 0.81$ on Sonnet/Opus, Tab. 12 in App. B.1), on two FRED macroeconomic series (UNRATE, CPI YoY at $r \geq 0.995$; App. B.3), and under semantic paraphrases that remove the Mkt-RF/Fama–French keyword (App. D.7); the recalled representation is keyed on the quantity, not the label. The rank–value decoupling of P3 also tests cross-domain on UNRATE: position-balanced robust accuracy is above chance but well below the MSB (App. A.4); strong-form chance-level decoupling appears Mkt-RF-specific. Recall is robust to chain-of-thought elicitation and prompt rewording within the range we tested (App. A.4, D.7).

4. White-box controlled validation

The frontier panel is black-box, so the training-objective argument of §1.1 is consistent with the data but not directly demonstrated. To test whether causal-LM training on date-indexed numeric values is *sufficient* to produce the regime, we LoRA-fine-tune Qwen-2.5-1.5B-Instruct under controlled exposure (full protocol: App. D.1). We construct a synthetic monthly series *Synthetic Market Residual A* (SMR-A) of 480 values from $\mathcal{N}(0.5, 4.5^2)$ rounded to two decimals; 24 months are reserved as held-out. We fine-tune (r=16, α=32, 8 epochs, lr 2×10^{-4}) at four exposure levels: 0× (filler-only, token-equalized), 1×, 5×, and 20× mentions per (date, value), and probe under the same Q&A format used in training; the 5× cell is replicated across four random seeds.

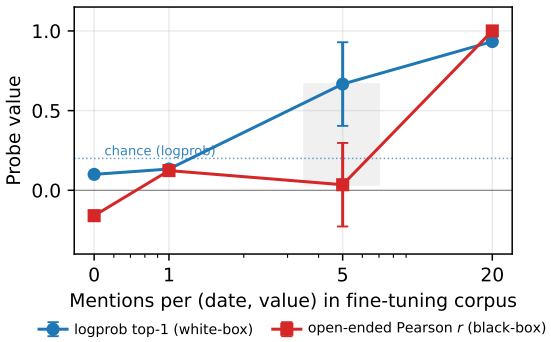

Figure 5. **Logprob ranking detects memorization that greedy generation under-reports.** Both probes are monotone in exposure on the SMR-A canary, but at 5× the open-ended Pearson r remains near zero (greedy decoding fails) while logprob top-1 accuracy is already 0.67. Error bars at 5× are sample std across 4 seeds. Full protocol in App. D.1.

Existence proof. Logprob top-1 accuracy on the true value rises monotonically with exposure (Fig. 5): 0.10 at 0× (*below* the 0.20 chance baseline; the base model mildly disprefers the true value's specific magnitude), 0.13 at 1×, 0.67±0.26 at 5× (every one of four seeds exceeds chance), and 0.93 at 20×. At 20× the model achieves verbatim recall on in-training months (30/30 exact matches, MAE = 0.000, r = 1.000): standard LoRA fine-tuning of an open 1.5B model on date-indexed values is sufficient to produce a queryable peaked conditional marginal at the value-token position.

Series-agnostic at moderate exposure. Companion runs at 5× on three additional synthetic series (SLF-B, SIS-C, SWI-D) drawn from comparable Gaussian distributions with different labels, units, and means show recall comparable to SMR-A across seeds. A fictional series (SVP-E) never present in any corpus returns near-zero r, confirming the absence of fabrication for unseen labels.

Greedy decoding under-reports the channel. The strongest 5× seed ranks the true value first on 29/30 months yet emits it under greedy decoding on only 5/30. Across the four 5× seeds, open-ended Pearson r averages $+0.035 \pm 0.262$ (consistent with zero) while every seed exceeds chance under logprob ranking. When the true value loses ranking it loses overwhelmingly to the *adjacent calendar month*'s true value (10/11 losses for the mirrored cell, 11/15 for 5× seed 2026, 6/6 for seed 42), itself a training-corpus value: evidence of date-conditional retrieval with limited date-discrimination resolution rather than random output. Production APIs (Anthropic; OpenAI Responses) typically do not expose token-level logprobs, so the frontier

evidence in §3 necessarily uses open-ended probes; the synthetic divergence raises the possibility that those numbers under-report the accessible numeric information.

Scope. The synthetic experiment is a controlled existence proof that next-token training on date-indexed numeric values suffices to produce a queryable peaked conditional marginal. It does not claim to faithfully replicate multi-series pretraining at frontier scale: a single series is fine-tuned in isolation under LoRA on a 1.5B base. The experiment establishes the route is sufficient and consistent with the signatures of §3, not that it is the actual mechanism in frontier closed models.

5. Discussion

The four predictions of distributional readout (P1)–(P4) hold in the cells we test. Together they support the view that, under our probe protocol, series memorization behaves as if exposing a peaked marginal distribution over (factor, month) tuples rather than as a prompt-queryable lookup or function: the marginal samples are recovered with high fidelity, but structured queries over the same cells are not.

Duplication scaling vs. rank–value decoupling. A natural reading of P1 (Mkt-RF is recalled while the other five Fama–French factors are not) is that Mkt-RF is more heavily duplicated in training corpora than SMB, HML, RMW, CMA, and Mom, which the duplication-scaling result of Kandpal et al. (2022) would predict directly. We do not contest this account for P1: cell-localized recall is consistent with duplication-driven memorization. What duplication scaling does *not* predict is the rank–value decoupling of P3: a representation recoverable by surface sampling at $r{=}0.98$ should, on a standard duplication account, also be recoverable by structured query, since the duplicated content includes the numeric values from which ordinal relationships are trivially recoverable. The chance-level rank accuracy on cells where marginal recall saturates is therefore the prediction that does not reduce to the duplication baseline, and is the result the controls (parser ablation, partial-recall HML comparison, forced-choice rerun) are designed to defend.

Scope of the joint P3 and P4 evidence. P3 and P4 are mutually supporting in opposite directions: P4 finds a sharply concentrated next-token distribution (consistent with a single mode), and P3 finds that the same distribution is not exposed jointly to structured queries (consistent with the mode being sampleable but not relationally accessible). Jointly they support the single-mode-readout reading more strongly than either does alone. Neither establishes it at the representational level: P4 is consistent with a multi-mode mixture in which one mode dominates locally, and the Opus position-bias finding (§3.2) shows that elicitation-protocol artifacts can mimic queryable structure on the same cells where the framework's prediction in fact holds. Alternative readings (instruction-following gap; parser selection bias; CoT self-correction) are listed in App. A.4; disambiguation between them is left to future work. Together, the four predictions function as a behavioral specification of the regime: any mechanistic account of series memorization proposed in subsequent work must reproduce a peaked sampling marginal, no joint accessibility through structured query without prior surfacing, and a capability-scaled entropy gap.

Memorization, generalization, and reasoning. Cross-label robustness is itself a generalization signal: the blind probe and the semantic-paraphrase results (§3.4, App. D.7) show that the conditioning set indexes a quantity, not a surface string. Distributional readout is therefore not lexical-template memorization; the memorized object is more abstract, but still falls short of supporting cross-prompt relational queries (the within-prompt vs. across-prompt gap of Tab. 2, read as a compositional-reasoning failure mode). The result speaks to all three of the memorization, generalization, and reasoning axes simultaneously: high-fidelity recall (memorization), label-invariant readout (partial generalization), and a relational-composition gap that elicitation can close (reasoning).

Corollary: leak attribution. If date-conditioned memorization behaves as distributional readout, an LLM-derived signal $\hat{S}_t$ conditioned on a date inherits covariation with $r_{f,t}$ from the same recall channel; regressing $\hat{S}$ on the model's own recall on co-located months gives a regression test for whether that covariation is recall-mediated. The corollary, the controlled-probe headline numbers, the ancient-era placebo, the worst-case bound, and a worked example are in App. E.

Implications for downstream evaluation. The corollary prescribes four diagnostic controls: *cutoff-aware evaluation*, *synthetic-factor perturbation*, a *transmission control*, and *label-stripping* (Wu et al., 2025). The last is insufficient on its own: our blind probe (§3.4) shows the underlying quantity is recovered at $r{\geq}0.92$ on Sonnet and Opus regardless of label. App. G applies the four controls to seven cited LLM-finance papers; no paper runs all four, and the transmission control is the gap our framework most directly addresses.

Limitations. API-boundary probes preclude mechanistic disambiguation of the single-mode reading from alternatives. The leak-attribution corollary (App. E) measures ρ_{recall} on the currently API-accessible model, not necessarily the original generation. Cross-platform factor robustness, in-context date-scrambling, and provider-specific fabrication remain open.

References

Fama, E.F. and French, K.R. The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465, 1992.

Fama, E.F. and French, K.R. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

Fama, E.F. and French, K.R. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.

Carhart, M.M. On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82, 1997.

French, K.R. Data library. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html, accessed April 2026.

Lopez-Lira, A., Tang, Y., and Zhu, M. The memorization problem: Can we trust LLMs' economic forecasts? *arXiv:2504.14765*, 2025.

Lopez-Lira, A. and Tang, Y. Can ChatGPT forecast stock price movements? Return predictability and large language models. SSRN 4412788, 2023.

Crane, L.D., Karra, A., and Soto, P.E. Total recall? Evaluating the macroeconomic knowledge of large language models. *FEDS 2025-044*, Federal Reserve Board, 2025.

Didisheim, A., Fraschini, M., and Somoza, L. AI's predictable memory in financial analysis. *Economics Letters*, 2025.

Li, X., Zeng, Y., Xing, X., Xu, J., and Xu, X. Profit mirage: Revisiting information leakage in LLM-based financial agents. *arXiv:2510.07920*, 2025.

Benhenda, M. Look-Ahead-Bench: a standardized benchmark of look-ahead bias in point-in-time LLMs for finance. *arXiv:2601.13770*, 2026.

Sarkar, S. and Vafa, K. Lookahead bias in pretrained language models. SSRN 4754678, 2024.

Liang, S., Garg, S., and Moghaddam, R.Z. The SWE-Bench illusion: When state-of-the-art LLMs remember instead of reason. *arXiv:2506.12286*, 2025.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., and Zhang, C. Quantifying memorization across neural language models. In *ICLR*, 2023.

Tirumala, K., Markosyan, A.H., Zettlemoyer, L., and Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. In *NeurIPS*, 2022.

Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models. In *ICML*, 2022.

Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., and Shokri, R. Quantifying privacy risks of masked language models using membership inference attacks. In *EMNLP*, 2022.

Wu, K., Yang, B., Ying, Z., and Zhou, D. Anonymization and information loss. *arXiv:2511.15364*, 2025.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *USENIX Security Symposium*, 2023.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *CVPR*, 2023.

Appendix: Supplementary Material

A. Probes and parser specifications

A.1. Probe templates (Variants A–E and controls)

Variants A–C probe the *question* (point-value vs. narrative vs. rank recall); D and E hold the question fixed and probe the *decoder* (reasoning tokens, sampling temperature); controls C2 and C3 probe whether generic numeric hallucination accounts for any of the signal. Variant A issues a one-shot numeric question naming the target factor and month (e.g. *market excess return (Mkt-RF)*, *size (SMB, Small Minus Big)*, . . .) and constrains the answer to a signed decimal percentage in at most 48 tokens. Variant B reframes the same query as a descriptive request (up to 384 tokens), so the committal number comes from the model's own prose rather than a template slot. Variant C asks which of two months had the higher return. Variant D prepends a single chain-of-thought sentence (*"Think step-by-step about historical Fama-French factor returns, then answer:"*) to Variant A. Variant E re-issues Variant A verbatim at `temperature=` 1 with two independent draws per month. Controls C2 and C3 are Variant A with the factor name replaced by the fabricated "*Gleason-Zeta volatility-conditioned residual*" (C2) or by the fabricated illiquid fund "*Holbrooke-Mansfield Opportunity Fund III (2007 vintage)*" (C3); neither name matches any training-corpus entity we could find. Exact wording for every variant is in `factor_leak/probe.py` and `factor_leak/controls.py`.

A.2. Famous-month list

The 20 narrative-rich months used by the sampling plan and the famous-month concentration metric:

Month	Event	Month	Event
1987-10	Black Monday	2010-05	Flash crash
1990-08	Gulf War / oil shock	2011-08	US debt ceiling
1997-10	Asian financial crisis	2015-08	China devaluation
1998-08	LTCM / Russian default	2016-11	US election
2000-03	Dot-com peak	2018-02	Volmageddon
2001-09	September 11	2018-12	Q4 2018 selloff
2002-07	Dot-com trough	2020-02	Pre-COVID top
2008-09	Lehman collapse	2020-03	COVID crash
2008-10	Post-Lehman crash	2020-11	Vaccine / value rotation
2009-03	Market bottom	2022-01	Growth-to-value rotation

A.3. Variant-C parser algorithm

The comparative parser is endorsement-aware: it must handle preambles that echo the prompt ("Between March 2020 and October 2008, . . . ") before the model commits. Pseudocode:

1. Normalize Unicode minus / en-dash / hyphen variants to ASCII.
2. Collect every `(offset, YYYY-MM)` mention in the response.
3. *Candidate filter.* Drop mentions that are not one of the two prompt months.
4. *Refusal guard.* If the text matches a refusal phrase ("I don't have access", "cannot reliably recall", "outside my training", . . .) **and** contains no strong-endorse phrase, return `None`.
5. *Unique mention.* If only one candidate appears, return it.
6. *Strong endorsement.* If a phrase (*my answer is X, answer: X, I pick X, higher return was in X*, . . .) names a candidate, return it.
7. *Prompt echo.* If the response opens with *Between, Comparing, The two months,* . . . , the first mention is an echo; return the last candidate mention.

8. *Leading answer.* If the first candidate mention starts within the first 5 characters of the stripped response, return it (handles "March 2020 was higher.").

9. Otherwise, return the last candidate mention.

Refusals parsed under step 4 are excluded from the comparative-accuracy denominator, not counted as wrong picks; this is why Haiku's headline Variant-C accuracy is reported over a denominator of order 1, not 360.

A.4. Auxiliary probes: variants C/D/E (full results)

Three auxiliary probes reveal the structure of what is memorized. **Variant C (comparative)**: Haiku refuses 99.7% of 360 pairs; Sonnet answers 89.7% across all six factors. On Sonnet×Mkt-RF specifically (n=60 pairs, where values are recalled at r=0.98) rank accuracy is at chance under three independent measurements (Tab. 3): endorsement-aware parser (App. A.3) on the parsed subset gives 52.5% (parse $40/60$); a naive "first month mentioned" parser at near-full parse gives 49.2% (n=59); and a forced-choice rerun with a strict prompt that drives parse to 100% gives 55.0% (n=60). All three 95% binomial CIs include 50%, so the chance-level result is robust both to parser choice and to refusal-based selection bias.

Measurement	parse	accuracy	95% CI
Endorsement-aware (paper)	0.67	0.525	[0.370, 0.680]
Naive first-mention parser	0.98	0.492	[0.364, 0.619]
Forced-choice rerun	1.00	0.550	[0.424, 0.676]

Table 3. Rank accuracy on Sonnet×Mkt-RF Variant-C pairs (n=60 unique pairs) under three measurement variants. Forced choice uses a strict prompt requiring the model to commit to one of two month strings (script `experiments/47_variantc_forced_choice.py`); naive parser ignores refusal phrases and returns the first candidate month mentioned (script `experiments/48_variantc_parser_ablation.py`). All three 95% CIs include 50%.

Opus replication on the same 60 pairs. We re-ran the three Variant-C measurements on Opus×Mkt-RF, re-using the exact 60 month pairs from the Sonnet run (script `experiments/61_variantc_opus.py`, seed-free pair reload). Opus is invoked with vendor-default temperature (the Anthropic API for Opus rejects `temperature`; we record `temperature: null`, `vendor_default: true` in the JSONL records). The original-order strict forced-choice run showed a strong Opus-specific position bias toward the literally-first option ($42/60$ first-option picks against a ground-truth-first rate of $29/60$, conditional accuracy 96.6% when truth is in the first slot and 54.8% when in the second). We therefore position-balanced strict forced-choice by replaying the same 60 pairs in swapped order (`experiments/64_variantc_opus_swapped.py`) and report both pooled and robust accuracy. Tab. 4 gives the position-balanced summary; Tab. 5 reports the per-pair paired analysis against Sonnet using the position-bias-free robust criterion (Opus correct under both orderings).

Measurement	parse rate	accuracy	95% CI	$n_{\text{parsed}}/n_{\text{total}}$
endorsement-aware	0.450	0.667	[0.489, 0.844]	27/60
naive first-mention	0.483	0.483	[0.301, 0.665]	29/60
strict forced-choice (original)	1.000	0.750	[0.640, 0.860]	60/60
strict forced-choice (swapped)	1.000	0.717	[0.603, 0.831]	60/60
strict forced-choice (pooled)	1.000	0.733	[0.654, 0.812]	120/120
strict forced-choice (robust)	1.000	0.533	[0.407, 0.660]	60/60

Table 4. Opus×Mkt-RF Variant-C, same 60 pairs as Sonnet (Tab. 3). Strict forced-choice rows are expanded into the original-order run, the position-swapped re-run, the pooled 120-trial accuracy, and the robust accuracy (correct under both orderings). The naive-parser column already replicates Sonnet's chance-level finding (95% CI covers 50%). The strict-forced-choice original-order 0.750 appeared above chance but was driven by position bias; the position-bias-free robust accuracy is $\mathbf{0.533}$, within sampling error of Sonnet's 0.550.

Measurement	paired correctness counts				paired n	agreement	McNemar p
	both correct	S only correct	O only correct	both wrong			
endorsement-aware	5	3	6	3	17	0.471	0.508
naive first-mention	9	2	4	13	28	0.786	0.688
strict forced-choice (original)	27	6	18	9	60	0.600	0.023
strict forced-choice (robust)	24	9	8	19	60	0.717	1.000

Table 5. Per-pair paired analysis, Sonnet (S) vs Opus (O) on the matched 60 Variant-C pairs. "Agreement" is the fraction of paired pairs where both models give the same correctness label. McNemar's two-sided test uses the exact binomial when $b+c \leq 25$. The original-order strict-forced-choice paired $p=0.023$ disappears once Opus is scored under the robust position-bias-free criterion: agreement 0.717, $p=1.00$.

Reading. On all three measurement operationalizations, once position bias is controlled for, Opus replicates Sonnet's chance-level rank accuracy on P3. The naive parser replicates without correction; the strict forced-choice replicates only under the position-balanced robust score; the endorsement-aware condition has too low a parse rate on Opus for a sharp test. P3 therefore replicates within the Anthropic top tier, and the position-bias finding is itself diagnostically informative: a generic instruction-following / OOD-format account would not predict that the bias-corrected accuracy returns to chance only on the high-recall Mkt-RF cell while the partial-recall HML cell shows above-chance signal under the same Variant-C format (Tab. 7). The position bias is a property of how Opus answers two-alternative forced-choice prompts in general, not a property of the recalled representation.

Cutoff-stratified rank accuracy. We re-stratify the strict-forced-choice records by pair-level cutoff bucket (a pair is "post" if either month is post-cutoff, else "near" if either is within ± 6 months of cutoff, else "pre"; script `experiments/65_cutoff_stratified_variantc.py`). On Sonnet, rank accuracy is uniform across pre and near strata (pre: $23/44{=}52.3\%$ $[0.38, 0.67]$; near: $5/10{=}50.0\%$ $[0.19, 0.81]$; post: $5/6{=}83.3\%$ $[0.54, 1.00]$, n too small for inference); the chance-level decoupling is therefore not driven by any particular cutoff stratum. On Opus, position-balanced robust accuracy is heterogeneous across strata (pre: $27/44{=}61.4\%$; near: $1/10{=}10.0\%$; post: $4/6{=}66.7\%$), but the $n_{\text{near}}{=}10$ subsample limits stratified inference; the pooled-across-strata robust accuracy of $32/60{=}53.3\%$ is the load-bearing replication number.

Cross-domain test on UNRATE. On 60 random month pairs from UNRATE (script `experiments/67_variantc_unrate.py`; Sonnet and Opus, both orderings), strict forced-choice gives Sonnet original/swapped $81.7\%/81.7\%$ and Opus $90.0\%/91.7\%$ ($n{=}60$ each). The position-balanced robust accuracies are Sonnet 76.7% (95% CI $[0.66, 0.87]$) and Opus 83.3% (95% CI $[0.74, 0.93]$). Both are above chance but well below the MSB on UNRATE, which is essentially 100% because typical pair gaps (~ 1.8pp) are an order of magnitude larger than the marginal- probe RMSE at $r{=}0.995$ (~ 0.16pp). The framework's prediction (rank accuracy below the MSB) therefore holds cross-domain; the strong-form chance-level decoupling observed on Mkt-RF appears specific to series where pair gaps are small relative to marginal noise.

Joint within-prompt rank–value probe. A within-prompt elicitation that asks the model to state both values and the rank in a single response (script `experiments/70_joint_value_rank.py`) gives Tab. 6: marginal value recall is preserved (Pearson $r \geq +0.84$ on stated v_i vs. truth on all three models), the stated pick is internally consistent with the model's own stated values ($\geq 98.3\%$ records), and rank accuracy is far above the across-prompt strict-forced-choice result on the same pairs.

Model	n_{parsed}	$r(v_1, t_1)$	$r(v_2, t_2)$	internal cons.	rank acc.
Sonnet 4.6	49/60	+0.95	+0.97	49/49=1.000	45/49=0.918
Opus 4.7	54/60	+0.99	+1.00	54/54=1.000	53/54=0.981
GPT-5.4	59/60	+0.84	+0.97	58/59=0.983	51/59=0.864

Table 6. Joint within-prompt rank–value probe on the same 60 Variant-C pairs. "Internal consistency" is the fraction of records on which the stated "Higher: X" pick agrees with $\text{sign}(v_1 - v_2)$ from the model's own stated values. The "striking" pattern (stated values rank truth correctly but the pick disagrees with the stated values) occurs in $0/49$ Sonnet, $0/54$ Opus, $1/59$ GPT-5.4 records: models are reliably internally consistent within a single response. Compared to across-prompt strict-forced-choice on the same models and pairs (Sonnet 55.0%, Opus position-balanced 53.3%, GPT-5.4 robust 38.3%), the within-prompt rank accuracy is 30–45 percentage points higher, showing the rank–value dissociation is between elicitations rather than within-context. Sonnet paired McNemar (within-prompt vs. across-prompt strict): b=22 joint-only-correct, c=4 across-only-correct, p=0.0005.

Variant-C extension to SMB and HML. The decoupling claim above was Sonnet×Mkt-RF specific. We re-ran the endorsement-aware and naive-first-mention parsers on the existing sweep records for Sonnet×SMB (n=60, value recall r=−0.25) and Sonnet×HML (n=60, value recall r=+0.48) pairs (Tab. 7). On SMB both parsers give chance-level rank accuracy (47.5% and 41.7%, both 95% CIs include 50%), consistent with poor value recall. On HML the two parsers *disagree*: endorsement-aware gives 65.5% (CI $[53.3\%, 77.7\%]$, above chance), while the naive parser gives 39.0% (below chance); the gap reflects that on partial-recall pairs the model's endorsed pick carries genuine signal that the naive parser discards as prompt echo. The *regime* pattern is therefore: on the high-recall factor (Mkt-RF, r=0.98) ranks decouple strongly from values; on partial recall (HML, r=0.48) ranks and values track together; on a factor with no useful positive value recall (SMB, r=−0.25) ranks are at chance. Decoupling is most striking precisely where recall is strongest; the single-mode-readout hypothesis below is one reading of this regime pattern, with alternatives discussed there.

Factor (value recall)	Endorse-aware	Naive	n
Mkt-RF (r=0.98)	0.525 [0.37, 0.68]	0.492 [0.36, 0.62]	60
HML (r=0.48)	0.655 [0.53, 0.78]	0.390 [0.27, 0.51]	60
SMB (r=−0.25)	0.475 [0.35, 0.60]	0.417 [0.29, 0.54]	60

Table 7. Variant-C rank accuracy on Sonnet across three factors, both parsers (script `experiments/48_variantc_parser_ablation.py`; data from the existing main sweep). Mkt-RF and SMB are at chance under both parsers; on HML the parsers disagree, reflecting partial value recall that the endorsement-aware parser correctly attributes to the model's pick.

Variant D (chain-of-thought): prepending "Think step-by-step" *reduces* recall sharply on Sonnet × Mkt-RF (r: 0.98→0.78, within-25 bps: 33.8%→14.9%; n=121). **Budget-confound control.** Variant D's prompt and Variant A's prompt also differ in `max_tokens` (384 vs. 48); a longer budget could pull the parsed answer from a self- corrected late-position digit independently of CoT content. We re-run Variant A at `max_tokens`= 384 on the same Mkt-RF months Variant D probed (script `experiments/59_variant_a_budget_control.py`; n=121): r=0.91, within-25 bps=33.9%. Doubling-plus the budget alone moves r from 0.98 to 0.91 but leaves within-25 bps at 0.34. The CoT-only contribution is therefore r: 0.91→0.78 (−0.13) and within-25 bps: 0.34→0.15 (−0.19); the within-25 bps drop is entirely CoT-content, while the r drop is roughly half budget, half content. **Variant E (T=1)**: accuracy essentially unchanged (r=0.983, within-25 bps 37.5%); two independent draws at the same month agree within 25 bps in 93% of pairs (mean spread 6 bps). *Interpretive reading.* One reading consistent with the joint pattern (rank-decoupling at r=0.98 values, CoT-content degradation of within-25 bps, T=1 stability) is a *conditioned single-mode readout*: given (factor, month) the model samples from a tightly peaked distribution over values and has no internal primitive for jointly evaluating two such distributions to rank them; CoT content disturbs the peak, sampling at T=1 does not. *Alternatives we cannot rule out.* (a) The Variant-C format is out-of-distribution for the comparison-phrasing the model was trained to follow, so chance-level accuracy reflects an instruction-following gap rather than an inductive-bias property of the recalled representation; (b) refusal-aware parsing inserts selection bias on the parsed subset that is incompletely corrected by the forced-choice rerun; (c) CoT-content degradation may reflect self-correction surfacing model uncertainty rather than perturbing a peaked readout (the budget control rules out the budget-only account but not the self-correction account). Disambiguating these accounts is left to future work. What is established empirically: rank accuracy is at chance on the high-recall cell across three measurements, partial-recall HML pairs ranks with values, and CoT-content lowers Mkt-RF recall fidelity beyond the budget effect. The practical corollary is unchanged: *CoT prompting is a mitigation*, not an amplifier, against factor-return leak.

Numerical detail. Variant D (CoT) probes 133 Mkt-RF months at `max_tokens`= 384; Variant E (T= 1) probes 88 Mkt-RF months with two independent draws each (176 responses). Per-variant recall is summarized in Tab. 8. Figure 6 shows the paired degradation under CoT on the month-matched subset: Variant A's r=0.98 collapses to Variant D's r=0.82, and on 54 of 73 paired months the CoT absolute error is strictly larger than the direct error. For Variant E, the within-draw spread on the 75 months where both draws parsed is 6.3 bps on average; 93.3% of same-month pairs agree within 25 bps. Temperature does not disturb the committal readout; reasoning tokens do.

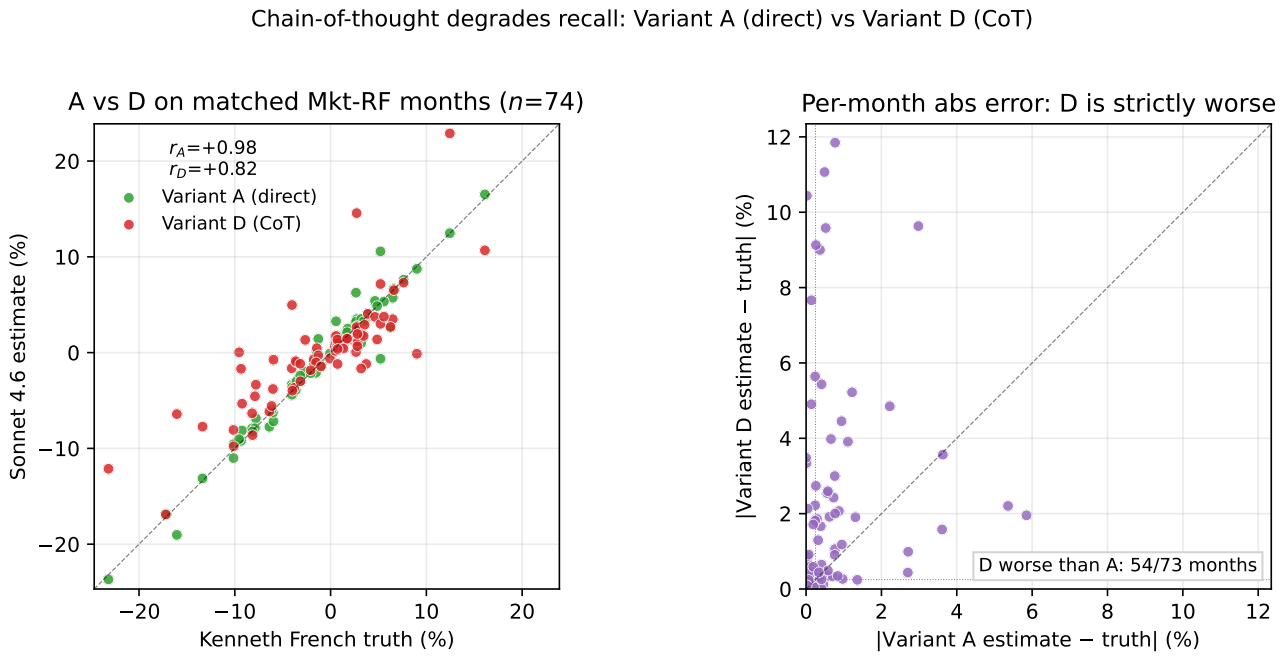

Figure 6. Chain-of-thought degrades Sonnet's Mkt-RF recall. *Left*: Variant A (green) and Variant D (red) estimates plotted against Kenneth French truth on the months probed under both conditions. *Right*: per-month absolute error; Variant D (y-axis) versus Variant A (x-axis). Points above the dashed equality line are months where reasoning made the answer worse.

Table 8. Mkt-RF recall under Variant D (CoT) and E (T= 1). Main-sweep Variant A on Sonnet for comparison: within-25 bps=0.338, r=0.980.

Model	Variant	n	parse rate	within-25 bps	Pearson r
Sonnet 4.6	D (CoT)	133	0.910	0.149	+0.776
Haiku 4.5	D (CoT)	133	0.602	0.100	+0.702
Sonnet 4.6	E (T= 1)	176	1.000	0.382	+0.983

A.5. Mechanistic signature: readout-entropy probe

The behavioral characterization (§3) treats the model's output as a black box. To complement it with a readout-level signature, we exploit the OpenAI Responses API's top-k logprobs feature on GPT-5.4: for every probed query we extract the top-5 token candidates and per-candidate log probabilities of the first two output tokens (sign + first numeric chunk), and compute the average per-token Shannon entropy in bits (treating the residual mass below the top-5 as a single "rest" bucket). This is not available on the Anthropic API, so the probe runs on GPT-5.4 only.

Conditions and predictions. We run three matched conditions (n=30 each) on GPT-5.4: (i) *Mkt-RF* on a fresh seed-2030 random sample of months from 1980-01–2024-12 (high-recall regime); (ii) *RMW* on the same months (low-recall regime; main-text within-25 bps on RMW is 15%); (iii) *Fabricated factors* (5 fictional names from App. B.2 × 6 months). The mechanistic prediction: a memorized readout should produce a sharply peaked distribution (low entropy) on a specific value; generic numeric hallucination on fabricated content should produce a more diffuse distribution (higher entropy) since the model is sampling from a "plausible monthly return" prior rather than retrieving a specific value.

Table 9. Average per-token Shannon entropy of the first two output tokens on GPT-5.4 (top-5 candidates, residual treated as a single "rest" bucket; bits). Mkt-RF readouts are $\sim 5\times$ more peaked than fabricated readouts even though the parse rate (commitment) on fabricated factors is 96.7% (Tab. 13); the model commits, but from a diffuse distribution.

Condition	n	mean entropy	median entropy
Mkt-RF (high recall)	30	0.21	0.05
RMW (low recall)	30	0.78	0.83
Fabricated factors	30	1.14	1.21

Two findings. (i) *Memorization vs. low recall.* Mkt-RF entropy is roughly one-quarter of RMW entropy (mean 0.21 vs. 0.78 bits, $\sim 4\sigma$ separation in distribution). The readout is sharply peaked when the model has the value memorized and substantially more diffuse when it does not. (ii) *Memorization vs. fabrication.* Even though GPT-5.4 *commits* to fabricated-factor queries at 96.7% (App. B.2), the readout entropy on those committed answers is $\sim 5\times$ that of Mkt-RF (mean 1.14 vs. 0.21 bits). Fabrication and memorization differ at the distributional level even when the surface output (a plausible signed percentage) is indistinguishable. This converts "the model commits to fictional factors" from a parse-rate observation into a distributional asymmetry: memorization produces a peaked readout, fabrication produces a diffuse one.

Sign-vs-value decomposition. Mkt-RF months are positive $\sim 63\%$ of the time, so the sign token (token 1) carries a strong prior that could lower its entropy on memorized cells independently of value confidence. We re-process the existing $n{=}90$ traces and report token-1 (sign) and token-2 (first numeric chunk) entropies separately (Tab. 10). The Fabricated–Mkt-RF gap is $+0.31$ bits on the sign token ($3.6\times$ ratio) but $+1.54$ bits on the value token ($6.0\times$ ratio): the value token dominates the gap. The sign-prior confound is real but cannot account for the bulk of the signal.

Table 10. Per-token Shannon entropy on GPT-5.4 (bits, top-5 + "rest" bucket). The Mkt-RF→Fabricated gap is larger on the value token than on the sign token in both absolute bits and ratio.

Condition	n	sign mean	value mean	avg(t1,t2)
Mkt-RF (high recall)	30	0.12	0.31	0.21
RMW (low recall)	30	0.42	1.14	0.78
Fabricated factors	30	0.43	1.84	1.14
gap (Fab–Mkt-RF)		$+0.31$	$+1.54$	$+0.92$
ratio (Fab/Mkt-RF)		$3.6\times$	$6.0\times$	$5.3\times$

Cross-tier validation. A reviewer concern is that the entropy gap might be a quirk of GPT-5.4 specifically rather than a memorization-regime property. We re-run the same probe ($n{=}30$ per condition; identical month samples) on GPT-5.4-mini and GPT-5.4-nano (Tab. 11). The Mkt-RF entropy decreases monotonically with capability tier ($1.67 \to 0.71 \to 0.21$ bits for nano, mini, full), tracking the model's recall fidelity (r: $-0.32 \to 0.65$ single-seed $\to 0.54$ pooled). Fabricated-factor entropy is roughly stable across tiers ($1.44, 1.01, 1.14$ bits): all three tiers fabricate from a diffuse distribution; only the top tier shows a peaked memorized regime. The within-tier Mkt-RF–Fabricated gap is therefore a property of *having something memorized to commit from a peaked readout*: -0.23 bits on nano (no memorization, gap inverted), -0.30 bits on mini (intermediate), $+0.93$ bits on full GPT-5.4.

Table 11. Mean per-token entropy across OpenAI tiers (bits, first two tokens averaged). Mkt-RF entropy is monotone in tier; fabricated entropy is roughly tier-invariant. The Mkt-RF–Fabricated gap emerges only on the top tier where Mkt-RF is actually memorized. Mini's $r{=}+0.65$ is the single-seed-42 entropy-probe sample ($n{=}30$); the 3-seed pooled recall is $r{=}+0.36$ (Sec. D.5), and the qualitative within-vendor monotone (full > mini > nano) is preserved under either reading.

Tier	Mkt-RF	RMW	Fabricated	Mkt-RF r
GPT-5.4-nano	1.67	1.52	1.44	-0.32
GPT-5.4-mini	0.71	1.28	1.01	$+0.65^{\dagger}$
GPT-5.4 (full)	0.21	0.78	1.14	$+0.54$ pooled

† single-seed-42 entropy-probe sample; 3-seed pooled $r{=}+0.36$ (App. D.5) preserves the within-vendor monotone.

Scope of the claim. With sign-vs-value decomposition (the gap is value-token dominated) and cross-tier validation (the gap scales with capability), the entropy probe is a distributional fingerprint that tracks the memorization regime. The claim

remains observational and non-Anthropic; a clean mechanistic decomposition on an open-weight controllable model is the natural follow-up.

Reproduction. Logprobs are only exposed for OpenAI/Azure deployments; the analogous probe on Anthropic models requires either internal access or an open-weight follow-up. Scripts `experiments/52_logprobs_probe.py` (full GPT-5.4), `54_entropy_decomposition.py` (sign-vs-value post-processing), `56_logprobs_cross_tier.py` (mini/nano); total n=270 queries, ∼$1.

B. Datasets, baselines, and controls

B.1. Baselines and label invariance

Three follow-up probes characterize *what* Sonnet has memorized: an S&P 500 probe, a NASDAQ Composite probe, and a blind-label probe that asks for "the broad U.S. stock market in excess of the T-bill rate" without naming Fama-French. Truth for S&P 500 and NASDAQ comes from Yahoo Finance monthly close-to-close price returns; truth for the blind probe is Kenneth French Mkt-RF. Table 12 reports recall on the same Variant-A answer format across all three alongside the main-sweep Mkt-RF row.

Table 12. Cross-model recall on four probes for the aggregate U.S. equity return. ρ_{FF} is the correlation of the target truth series with Ken French Mkt-RF on the probed months. n=40 per cell for the baselines; the Sonnet main-sweep Mkt-RF row uses n=77. Anthropic models, three OpenAI GPT-5.4 tiers, DeepSeek-V3.2, and the two Meta Llamas, all via official APIs. Llama-3.1-8B refuses every Mkt-RF query (parse rate 0); r is reported on the parsed subset. GPT-5.4-nano's Mkt-RF row is the only negative r in the table; the smallest GPT model generates anti-correlated noise rather than the memorized series.

Model	Probe	ρ_{FF}	parse	within-25 bps	Pearson r	sign
Opus 4.7	Mkt-RF	1.00	1.00	0.68	+0.986	1.00
Opus 4.7	S&P 500	0.99	1.00	1.00	+1.000	1.00
Opus 4.7	NASDAQ Composite	0.92	1.00	0.88	+0.972	0.93
Opus 4.7	Blind U.S. mkt excess	1.00	1.00	0.68	+0.954	0.98
Sonnet 4.6	Mkt-RF (main)	1.00	0.88	0.34	+0.98	0.97
Sonnet 4.6	S&P 500	0.99	1.00	0.85	+0.97	0.95
Sonnet 4.6	NASDAQ Composite	0.92	0.95	0.63	+0.81	0.84
Sonnet 4.6	Blind U.S. mkt excess	1.00	0.62	0.20	+0.92	1.00
Haiku 4.5	S&P 500	0.99	1.00	0.38	+0.59	0.75
Haiku 4.5	NASDAQ Composite	0.92	0.93	0.08	+0.48	0.76
GPT-5.4	Mkt-RF	1.00	1.00	0.35	+0.70	0.80
GPT-5.4	S&P 500	0.99	1.00	0.63	+0.91	0.88
GPT-5.4	NASDAQ Composite	0.92	1.00	0.23	+0.71	0.78
GPT-5.4	Blind U.S. mkt excess	1.00	1.00	0.33	+0.77	0.85
GPT-5.4-mini	Mkt-RF	1.00	1.00	0.35	+0.65	0.73
GPT-5.4-mini	S&P 500	0.99	1.00	0.50	+0.76	0.83
GPT-5.4-mini	NASDAQ Composite	0.92	1.00	0.15	+0.43	0.70
GPT-5.4-mini	Blind U.S. mkt excess	1.00	1.00	0.10	+0.54	0.70
GPT-5.4-nano	Mkt-RF	1.00	1.00	0.03	−0.32	0.43
GPT-5.4-nano	S&P 500	0.99	1.00	0.08	+0.43	0.60
GPT-5.4-nano	NASDAQ Composite	0.92	1.00	0.10	+0.20	0.50
GPT-5.4-nano	Blind U.S. mkt excess	1.00	1.00	0.05	+0.18	0.65
DeepSeek-V3.2	Mkt-RF	1.00	1.00	0.15	+0.48	0.73
DeepSeek-V3.2	S&P 500	0.99	1.00	0.55	+0.86	0.83
DeepSeek-V3.2	NASDAQ Composite	0.92	1.00	0.23	+0.80	0.73
DeepSeek-V3.2	Blind U.S. mkt excess	1.00	1.00	0.15	+0.42	0.65
Llama-3.3-70B	Mkt-RF	1.00	0.97	0.08	+0.31	0.62
Llama-3.3-70B	S&P 500	0.99	1.00	0.45	+0.68	0.65
Llama-3.3-70B	NASDAQ Composite	0.92	1.00	0.10	+0.18	0.60
Llama-3.3-70B	Blind U.S. mkt excess	1.00	1.00	0.10	+0.08	0.60
Llama-3.1-8B	Mkt-RF	1.00	0.00	–	–	–
Llama-3.1-8B	S&P 500	0.99	1.00	0.03	+0.23	0.40
Llama-3.1-8B	NASDAQ Composite	0.92	0.55	0.00	−0.03	0.50
Llama-3.1-8B	Blind U.S. mkt excess	1.00	0.53	0.00	+0.13	0.33

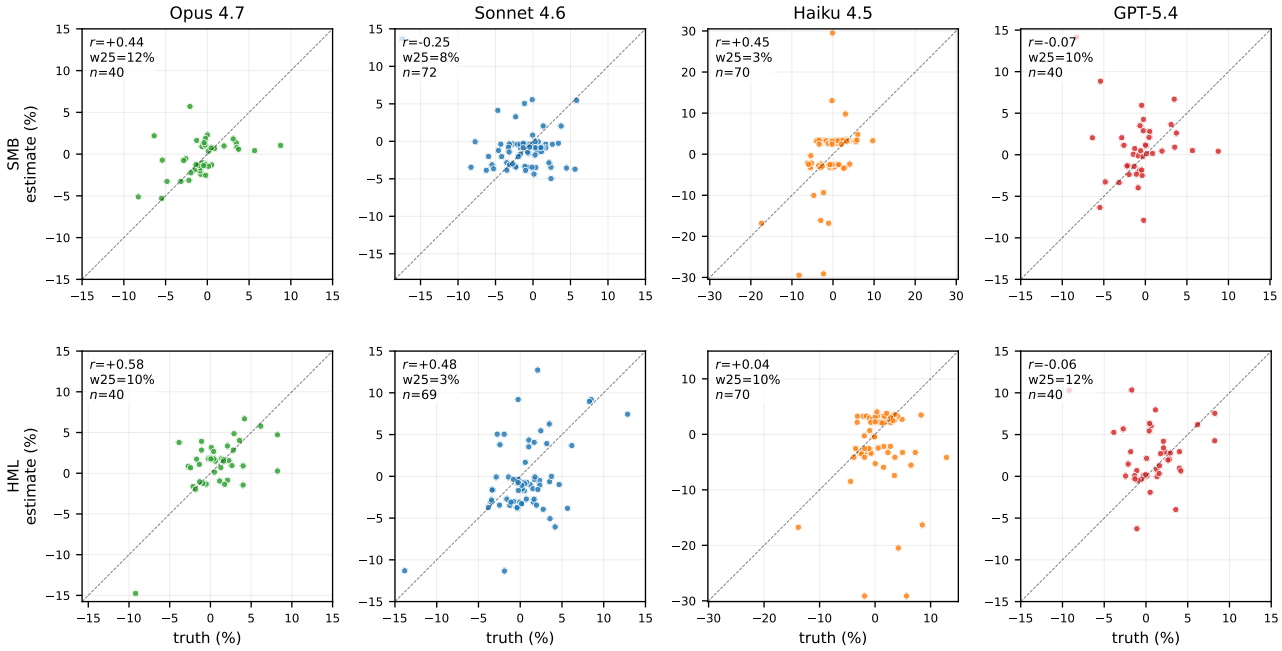

Figure 7. Variant-A calibration on the two Fama-French factors with any partial recall (SMB, HML) across all four models. Opus shows the cleanest alignment (r=0.44 on SMB, r=0.58 on HML), with weaker but visible HML signal on Sonnet (r=0.48); other cells are noise. Mkt-RF (clean recall on all four models) is shown in the main-text Fig. 2; the full 12-cell calibration grid is in App. F.1 (Fig. 10); RMW, CMA, and Mom are at chance on every model and not shown.

Panel extension: Qwen-3 32B and Llama-4 Scout 17B. Two additional open-weight models on Groq (a fifth-vendor Alibaba Qwen-3 32B, and Meta's smaller MoE Llama-4 Scout 17B-active) extend the panel to 11 models / 5 providers (scripts `experiments/57_reasoning_models_baselines.py`, `19_llama_baselines.py`). Both have qualitatively distinct behavior from the headline panel. *Qwen-3 32B*: at `max_tokens`=48 the model emits a `<think>` preamble and gets truncated before any number; at `max_tokens`=1024 it commits to a numeric answer but the content reveals priors-driven reasoning rather than retrieval, e.g., "*If the T-bill rate for November 1981 was, say, 14.5% annually, then the monthly rate would be approximately 14.5%/12 ≈ 1.21%; so the excess return...*" Pearson r=+0.09 on n=26 Mkt-RF months, indistinguishable from chance. The model has the *relationship* (market − risk-free) but not the values. *Llama-4 Scout 17B-active MoE*: 100% parse on S&P 500 / NASDAQ at chance fidelity (r=0.14, 0.16), ∼ 95% refusal on Mkt-RF / blind. The smaller-active- parameter MoE generation does not memorize broad U.S. index returns at the fidelity that the dense 70B Llama-3.3 does (Tab. 12). These two cells reinforce the capability/ training-pipeline picture: vendor-side post-training (Qwen-3's reasoning-by-default behavior; Llama-4 Scout's broader refusal) and parameter-class (Scout 17B active vs. 70B dense) both depress the recall channel without eliminating the cross-vendor pattern.

B.2. Expanded fabricated-series control

The original fabricated-series probe used two fictional names on Sonnet/Haiku (n=24) and was acknowledged in the main text as underpowered. We expand to five fictional names × eight models × twelve months (n=480 over four providers, seed 2026, script `experiments/46_fabricated_expansion.py`). The prompt is identical to Variant A except the factor name is replaced by one of: *Gleason-Zeta volatility-conditioned residual factor*, *Holbrooke-Mansfield Opportunity Fund III (2007 vintage)*, *Brennan-Iyer mean-reversion premium factor*, *Northrop-Calloway long-horizon dispersion factor*, *Pemberton-Yi cross-sectional liquidity premium factor*. None of these match an entity we could find in public corpora.

Provider	Model	n	parsed	parse rate
Anthropic	Opus 4.7	60	0	0.000
Anthropic	Sonnet 4.6	60	0	0.000
Anthropic	Haiku 4.5	60	0	0.000
OpenAI	GPT-5.4	60	58	0.967
OpenAI	GPT-5.4-mini	60	58	0.967
OpenAI	GPT-5.4-nano	60	60	1.000
DeepSeek	DeepSeek-V3.2	60	59	0.983
Meta	Llama-3.3-70B	60	60	1.000
Anthropic pooled		180	0	0.000
Non-Anthropic pooled		300	295	0.983

Table 13. Parse rate on 5 fictional factor names $\times$ 12 months. Anthropic models refuse *every* query across the three tiers, providing a sharp negative control for the Mkt-RF recall result: a model that recalls Mkt-RF at $r{\approx}0.98$ but emits no committal answer to a syntactically-identical fictional-factor prompt has not learned a generic "emit a return" behavior. All five non-Anthropic models across three providers (OpenAI, DeepSeek, Meta) commit at $\geq$ 96.7%, pooling to $295/300$ (98.3%). The split is between Anthropic and everyone else, not between capability tiers within a vendor: GPT-5.4-nano (a low-tier model that recalls Mkt-RF at $r{=}{-}0.32$) commits at 100%, ruling out "the model commits because it has memorized the answer". Wilson 95% CI on the Anthropic pooled rate is $[0.000,\ 0.020]$; on non-Anthropic pooled it is $[0.962,\ 0.992]$; the intervals do not overlap by orders of magnitude. The asymmetry cuts along provider lines and is independent of within-vendor capability tier, consistent with a post-training origin (instruction-tuning or RLHF that trains refusal of unverifiable quantitative claims). API-only access cannot separate this from an upstream pre-training data-filtering policy producing the same surface behavior.

B.3. Cross-domain replication: UNRATE and CPI YoY

To address the concern that series memorization may be specific to Fama-French, we replicate the headline Variant-A probe on two non-financial macro series from FRED: the BLS monthly civilian unemployment rate (`UNRATE`, seasonally adjusted; macro/labor) and U.S. year-over-year CPI inflation rate (`CPIAUCSL`, 12-month percent change of the level series; macro/prices). 30 months each, 1980–2024 (seeds 42 and 2028; scripts `45_unemployment_baseline.py`, `50_cpi_baseline.py`).

Series	Model	n	parse	r	w-25 bps	w-50 bps
UNRATE	Sonnet 4.6	30	1.00	+1.000	1.00	1.00
UNRATE	Opus 4.7	30	1.00	+1.000	1.00	1.00
CPI YoY	Sonnet 4.6	30	0.97	+0.995	0.93	1.00
CPI YoY	Opus 4.7	30	1.00	+1.000	1.00	1.00

Table 14. Cross-domain Variant-A recall on Sonnet/Opus. Two non-financial series across distinct macro categories (labor + prices) both recall above $r{=}0.99$ on the top tier. UNRATE has $\sigma \approx 0.1$ pp/month (vs. Mkt-RF $\sigma \approx 4.5\%$/month) so the within-25 bps tolerance is a weaker fidelity test on UNRATE than on Mkt-RF; CPI YoY has higher month-to-month variance than UNRATE (range -2 to 14% across the sample) so within-25 bps is stronger there. The diagnostic framework is domain-portable, not Fama-French specific. (App. D.6 re-reports recall on a series-comparable σ-normalized scale.)

C. Reproducibility and compute

Repository. Full code, raw JSONL responses, and derived tables are available at the anonymized read-only mirror `https://anonymous.4open.science/r/factor-leak-7D4E`.

What's in the repo. Three Python packages and one experiments directory. `factor_leak/probe.py` registers all model endpoints (Anthropic Messages API, Azure OpenAI Responses API, Azure AI Foundry chat completions, Together AI, Groq, DeepSeek direct API), each with a uniform call signature returning text and usage. `factor_leak/parse.py` contains the endorsement-aware Variant-C parser with full pseudocode in App. A.3. `factor_leak/ff_loader.py` loads Kenneth French monthly returns from a frozen local snapshot (CRSP build 2026-02). `experiments/` holds 70 numbered scripts; the JSONL outputs in `experiments/results/` are the raw artifacts behind every table and figure. Each script documents its random seed in its docstring; reading `factor_leak.env.load_dotenv` resolves credentials without exposing them to the script.

Total queries and access window. Approximately 5,600 logged API queries across the panel, issued between 2026-01 and 2026-05. Per-query records include exact prompt, model identifier, temperature setting (or `vendor_default` flag where the API rejected the parameter), seed, response text, token counts, latency, and USD cost. Re-running `experiments/02_analysis.py` against the frozen JSONLs reproduces the headline table and all figures exactly.

Exact model identifiers (model strings as called). The eleven panel models, with the model identifiers used at the API. *Anthropic* (Messages API): `claude-opus-4-7`, `claude-sonnet-4-6`, `claude-haiku-4-5-20251001`. *Azure OpenAI Responses*: deployment-name resolved from `AZURE_OPENAI_DEPLOYMENT(_MINI|_NANO)` for GPT-5.4 / mini / nano (and `AZURE_GPT55_DEPLOYMENT` for GPT-5.5 if extended). *Azure AI Foundry* (chat-completions): `DeepSeek-V3.2`. *Together AI*: `meta-llama/Llama-3.3-70B-Instruct-Turbo`, `meta-llama/Llama-3.1-8B-Instruct-Turbo`. *Groq*: `llama-3.3-70b-versatile`, `llama-3.1-8b-instant`, `qwen/qwen3-32b`, `meta-llama/llama-4-scout-17b-16e-instruct`. *DeepSeek direct*: `deepseek-chat`.

D. Robustness analyses

D.1. Controlled synthetic memorization sweep (full protocol)

The body documents selective high-fidelity recall in production foundation models and demonstrates the route on an open 1.5B causal LM (§4). This appendix gives the full protocol.

Setup. We construct a synthetic monthly series *Synthetic Market Residual A* (SMR-A) with 480 values spanning 1980–2019, sampled i.i.d. from $\mathcal{N}(0.5, 4.5^2)$ and rounded to two decimals; 24 random months are reserved as a held-out split. We LoRA-fine-tune (r=16, α=32, lr 2×10^{-4}, 8 epochs, linear-warmup-then-constant) on token-equalized corpora at four exposure levels: $0\times$ (filler-only, same total tokens), $1\times$, $5\times$, and $20\times$ mentions per (date, value) pair, and probe at evaluation time using the same Q&A format as training. The $5\times$ condition is run with four random seeds (2026, 7, 42, 13) to characterize seed-level variance.

Logprob ranking dose-response. Tab. 15 reports a complementary probe in which the model scores five candidate completions per (in-training) month (the true value, its sign-flipped twin, the adjacent-month true value, the value of a different synthetic series, and a uniform random decoy in $[-10, +10]$) by length-normalized sequence logprob. Top-1 accuracy rises monotonically with exposure: 0.10 at $0\times$ (below the 0.20 chance baseline), 0.13 at $1\times$, 0.67 ± 0.26 at $5\times$, and 0.93 at $20\times$ (Fig. 5). The mean rank of the true candidate falls from 3.33 to 1.07 over the same range.

Table 15. Logprob ranking of completion candidates on the SMR-A models (Qwen-2.5-1.5B-Instruct, LoRA r=16, 8 epochs). For each of 30 in-training months we score five candidates by length-normalized sequence logprob. Top-1 = fraction of months where the true value receives the highest logprob; mean rank of true = average rank (1 = best, 5 = worst); mean gap = mean logprob difference between the true value and the best competing candidate (positive ⇒ true wins). The $5\times$ cell is mean over 4 random seeds with sample standard deviation; chance baseline for top-1 is 0.20.

Exposure	Top-1 acc.	Mean rank of true	Mean gap (true − best other)
$0\times$	0.10	3.33	-0.753
$1\times$	0.13	3.33	-0.434
$5\times$	0.67 ± 0.26	1.48 ± 0.44	$+0.352 \pm 0.320$
$20\times$	0.93	1.07	$+0.826$

Mechanism: date-conditional retrieval with smoothing. When the true value loses logprob ranking at $5\times$ and $20\times$, it loses overwhelmingly to the *adjacent calendar month*'s true value ($6/6$ losses at seed 42, $11/15$ at seed 2026, $1/2$ at $20\times$), itself a training-corpus value. The dominant failure mode is therefore confusion between temporally adjacent (date, value) pairs, not random output: evidence the model is performing date-conditional retrieval with limited date-discrimination resolution rather than learning the marginal distribution of values.

Series-agnostic at moderate exposure. Companion $5\times$ runs on SLF-B, SIS-C, SWI-D (drawn from comparable Gaussian distributions but with different labels, units, and population means) show recall comparable in shape to SMR-A across seeds. A fictional series SVP-E (never present in any corpus) returns near-zero r, confirming absence of fabrication for unseen labels.

Scope. The synthetic experiment is a controlled existence proof, not a faithful replication of frontier pretraining: a single series is fine-tuned in isolation under LoRA on a 1.5B base. It complements the production-model evidence in §3 without substituting for it. Build script `experiments/71b_logprob_ranking.py`; n=30 months $\times$ 5 candidates per model, 8 models. Full per-record JSONL is released with the paper artifact.

D.2. Per-cell refusal and parse rates

Three patterns stand out across the (model, factor, variant) cube (Fig. 8; full per-cell rates in `experiments/results/sweep.jsonl`): (i) Sonnet's Variant-A parse rate tracks factor prominence (88% on Mkt-RF, ∼27% on RMW/CMA), consistent with self-knowledge of what it has memorized; (ii) Haiku's Variant-B parse rate collapses on every factor (≤31%), so the descriptive narrative variant is uninformative on Haiku; (iii) Haiku refuses essentially all of Variant-C (0% commit on five of six factors), which is why Haiku's headline Variant-C accuracy is reported on a denominator of order 1, not 360; refusals are excluded from the comparative-accuracy denominator (App. A.3, step 4).

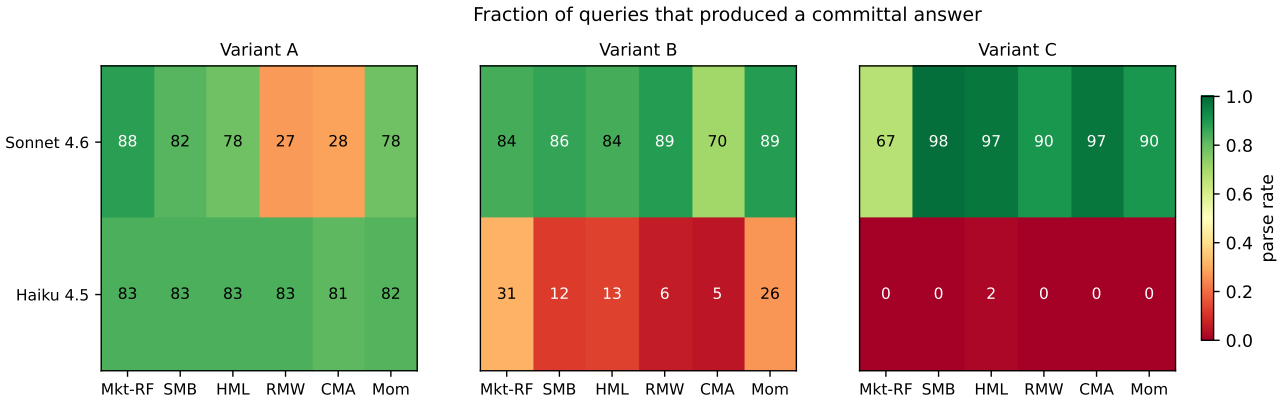

Figure 8. Parse rate per (model, factor, variant). Green = high commit rate; red = refusal. Values are percentages.

D.3. Per-cell cutoff-gradient regressions

We regress per-cell within-25 bps recall on signed months-to-cutoff (Variants A∪B pooled; raw two-sided p-values against $t_{n_{\text{bins}}-2}$; q-values Benjamini–Hochberg-adjusted across the 12 tests at α=0.05). Slope magnitudes are $|\beta| \leq 4.4 \times 10^{-4}$/month across all cells. The smallest raw p-value is on Haiku×Mkt-RF (p=0.030, BH-q=0.12); no cell clears FDR at q=0.05. Two cells (Haiku Mom, Sonnet CMA) have zero variance in the outcome (every month misses the 25 bps threshold), so the OLS is degenerate and reported as "–"; null-variance regression cannot support or reject a gradient. Per-cell numbers are released in `experiments/results/per_cell_cutoff.csv`.

D.4. Pre-vs-post cutoff stratification

To complement the slope-based test (§3) with a discrete pre/post split, we bucket each probed Mkt-RF Variant-A month relative to its model's training cutoff: *pre* (d>6 months before cutoff), *near* ($|d|\leq$6), and *post* (d>6 months after) (Tab. 16). Two findings. *Refusal cliff at the cutoff.* Sonnet×Mkt-RF parses 70/70 pre-cutoff, 7/13 near, 0/5 post; Haiku parses 70/70 pre, 0/13 near, 0/1 post. Anthropic models that produce committal answers pre-cutoff sharply suppress their outputs on post-cutoff months, consistent with self-knowledge of training cutoff. The post-cutoff n is small by design: the probe window ends 2026-02 and Anthropic mid-2025 cutoffs leave $\sim$ 8 months past-cutoff, of which the per-cell sampler hits the few that survive truth-availability gates. *Pre-cutoff recall fidelity is bucket-invariant.* On the parsed subset, Sonnet r=0.980 pre and 0.989 near are statistically indistinguishable (bootstrap CIs overlap). The six baseline rows below Haiku draw 0 post-cutoff months by construction (40-month sample restricted to pre-cutoff probe window), so the stratification is uninformative for those cells. Reading the slope-based and stratified tests together: the within-25 bps *rate* is flat in cutoff distance among parsed responses, but *parsing itself* drops sharply across the cutoff boundary on Sonnet and Haiku. The cutoff effect is discrete (refusal/no-refusal), not a gradient on recall fidelity.

Table 16. Mkt-RF Variant-A recall stratified by months-to-cutoff bucket. "parse" is parsed/sampled; r, within-25 bps on parsed subset; "–" marks $n<3$. Anthropic Sonnet/Haiku include the multi-seed pool.

		Pre			Near			Post		
Model	Cutoff	parse	r	w25	parse	r	w25	parse	r	w25
Sonnet 4.6	2025-03	70/70	$+0.98$	0.36	7/13	$+0.99$	0.14	0/5	–	–
Haiku 4.5	2025-07	70/70	$+0.68$	0.19	0/13	–	–	0/1	–	–
Opus 4.7	2026-01	40/40	$+0.99$	0.68	0/0	–	–	0/0	–	–
GPT-5.4	2025-08	40/40	$+0.70$	0.35	0/0	–	–	0/0	–	–
GPT-5.4-mini	2025-08	40/40	$+0.65$	0.35	0/0	–	–	0/0	–	–
GPT-5.4-nano	2025-08	40/40	-0.32	0.03	0/0	–	–	0/0	–	–
DeepSeek-V3.2	2025-07	40/40	$+0.48$	0.15	0/0	–	–	0/0	–	–
Llama-3.3-70B	2024-12	39/39	$+0.31$	0.08	0/1	–	–	0/0	–	–

D.5. Multi-seed robustness (Mkt-RF)

Table 17 reports per-seed recall on 40 random Mkt-RF months for seeds $\{1, 2, 3\}$, plus the pooled statistics across the three runs, on the five models with single-seed-42 entries in the headline-class tables. Opus is seed-deterministic (per-seed $r\in[0.995, 0.996]$, pooled $r=0.995$); Sonnet's pooled $r=0.921$ is consistent with the main-sweep $r=0.98$; Haiku, GPT-5.4, and GPT-5.4-mini are seed-sensitive (Haiku pooled $r=0.27$ vs. seed-42 $r=0.68$; GPT-5.4 pooled $r=0.54$ vs. seed-42 $r=0.70$; GPT-5.4-mini pooled $r=0.36$ vs. seed-42 $r=0.65$, three favorable single-seed-42 draws). Within-25 bps and sign are markedly more seed-stable than Pearson r (Opus 0.55–0.62 within-25 bps across seeds; GPT-5.4 0.30–0.38; mini 0.15–0.25): rank-correlation depends on the few outlying months in each draw, threshold-based recall does not. The OpenAI within-vendor capability monotone is now restored under honest pooling: full GPT-5.4 pooled $r=0.54 >$ mini pooled $r=0.36 >$ nano single-seed $r=-0.32$ (with corresponding within-25 bps $0.33 > 0.20 > 0.03$). The reporting convention is therefore asymmetric: where multi-seed identifies the seed-42 entry as a favorable draw on r (Haiku, GPT-5.4, GPT-5.4-mini), Tab. 1 reports the pool; where the seed-42 entry is seed-stable (Opus) or only modestly looser than the pool (Sonnet, $r=0.92$ pool vs. 0.98 main-sweep on a different sample), Tab. 1 reports the main-sweep value with the seed-stability flagged in the caption. The abstract's "$r\approx0.98$" tracks the headline-table convention: Opus 0.99 pooled and Sonnet 0.98 main-sweep both round to 0.98 within the abstract's precision.

Table 17. Per-seed and pooled Mkt-RF recall under Variant A across all four models with single-seed-42 entries in the headline table. Main-sweep (seed 42) rows: Opus $r=0.986$, within-25 bps$=0.68$, sign $= 1.00$; Sonnet $r=0.98$, within-25 bps$=0.338$, sign $= 0.974$; Haiku $r=0.68$, within-25 bps$=0.171$, sign $= 0.771$; GPT-5.4 $r=0.70$, within-25 bps$=0.35$, sign $= 0.80$; GPT-5.4-mini $r=0.65$, within-25 bps$=0.35$, sign $= 0.72$.

Model	Seed	n	Pearson r	within-25 bps	sign
Opus 4.7	1	40	$+0.996$	0.625	0.975
Opus 4.7	2	40	$+0.995$	0.550	0.875
Opus 4.7	3	40	$+0.995$	0.600	1.000
Opus 4.7	pooled	120	$+0.995$	0.592	0.950
Sonnet 4.6	1	39	$+0.858$	0.359	0.923
Sonnet 4.6	2	40	$+0.967$	0.175	0.925
Sonnet 4.6	3	40	$+0.953$	0.250	0.950
Sonnet 4.6	pooled	119	$+0.921$	0.261	0.933
Haiku 4.5	1	40	$+0.586$	0.175	0.700
Haiku 4.5	2	40	$+0.021$	0.125	0.625
Haiku 4.5	3	40	$+0.287$	0.050	0.625
Haiku 4.5	pooled	120	$+0.266$	0.117	0.650
GPT-5.4	1	40	$+0.591$	0.300	0.775
GPT-5.4	2	40	$+0.199$	0.375	0.725
GPT-5.4	3	40	$+0.832$	0.300	0.900
GPT-5.4	pooled	120	$+0.544$	0.325	0.800
GPT-5.4-mini	1	40	$+0.592$	0.250	0.700
GPT-5.4-mini	2	40	$+0.079$	0.200	0.625
GPT-5.4-mini	3	40	$+0.392$	0.150	0.775
GPT-5.4-mini	pooled	120	$+0.364$	0.200	0.700

D.6. σ-normalized tolerance reporting

Within-25 bps has different statistical content across series with different volatilities. On Mkt-RF ($\sigma \approx 4.45\%$/month, 1963-07–2026-02), 25 bps$=0.056\sigma$; on UNRATE ($\sigma \approx 1.79$ on the response-scale level, 1980–2024 sample), 25 bps$=0.139\sigma$, a $\sim 2.5\times$ spread, narrower than the $40\times$ a naive comparison of monthly *change*-volatilities would suggest, but still material. Table 18 re-reports the panel-wide Mkt-RF recall in σ-units alongside the threshold-uniform within-25 bps rate, plus the same metric for S&P 500, NASDAQ, UNRATE, and CPI YoY.

The σ-normalized table preserves the qualitative ordering of the headline within-25 bps ranking (Opus saturates within σ on every series probed; sub-tier models monotone-down) and sharpens the cross-series comparison: Opus's UNRATE within-25 bps$=1.00$ corresponds to within-$\sigma/4=1.00$, the same σ-content as Opus's Mkt-RF within-$\sigma/4=0.95$, so the cross-domain replication claim is comparable on a σ-tight test, not just nominally.

Table 18. σ-normalized recall: per (model, series) cell, 25 bps fraction (paper's headline metric) alongside within-$\sigma/4$, within-$\sigma/2$, within-σ. Mkt-RF rows for Sonnet/Haiku/GPT-5.4 use 3-seed pool (Sec. D.5); other rows use the main Variant-A sweep. Series σ in response-scale units (%/mo for returns; pp for UNRATE/CPI YoY level).

Model	Series	25 bps/σ	w25 bps	w$\sigma/4$	w$\sigma/2$	wσ
Opus 4.7	Mkt-RF	0.056	0.59	0.95	0.97	1.00
Sonnet 4.6 (pool)	Mkt-RF	0.056	0.26	0.76	0.93	0.95
Haiku 4.5 (pool)	Mkt-RF	0.056	0.12	0.34	0.52	0.70
GPT-5.4 (pool)	Mkt-RF	0.056	0.33	0.68	0.75	0.85
GPT-5.4-mini	Mkt-RF	0.056	0.35	0.55	0.68	0.85
GPT-5.4-nano	Mkt-RF	0.056	0.03	0.17	0.33	0.50
DeepSeek-V3.2	Mkt-RF	0.056	0.15	0.38	0.57	0.72
Llama-3.3-70B	Mkt-RF	0.056	0.08	0.18	0.41	0.72
Opus 4.7	S&P 500	0.064	1.00	1.00	1.00	1.00
GPT-5.4	S&P 500	0.064	0.62	0.82	0.85	0.95
DeepSeek-V3.2	S&P 500	0.064	0.55	0.72	0.80	0.90
Opus 4.7	NASDAQ	0.054	0.88	0.90	0.93	0.97
GPT-5.4	NASDAQ	0.054	0.23	0.42	0.53	0.75
Opus 4.7	UNRATE	0.139	1.00	1.00	1.00	1.00
Sonnet 4.6	UNRATE	0.139	1.00	1.00	1.00	1.00
Opus 4.7	CPI YoY	0.067	1.00	1.00	1.00	1.00
Sonnet 4.6	CPI YoY	0.067	0.93	0.97	1.00	1.00

D.7. Phrasing-perturbation control

The Variant-A template fixes a specific surface form ("*What was the monthly return of the Fama-French [factor] factor in [month]? Answer with a signed decimal percentage . . .*"). A reviewer can reasonably ask whether $r \approx 0.98$ is a property of *that template* or of the model's parametric representation of the series. We hold content fixed (Mkt-RF monthly return, signed-percent answer) and vary only the surface form across three reworded prompts:

1. **Terse**: "*Mkt-RF (Fama-French market excess) monthly return in* {*month*}*?*"
2. **Imperative**: "*Tell me the Fama-French Mkt-RF return for* {*month*}."
3. **Conversational**: "*Do you remember what the Fama-French Mkt-RF factor returned in* {*month*}*?*"

Sample: 30 months from 1980–2020 (seed 2027, fresh draw) on Sonnet 4.6 and Opus 4.7 (script `experiments/49_phrasing_perturbation.py`).

Model	Phrasing	n	parse	r	w-25 bps
Opus 4.7	terse	30	1.00	$+0.937$	0.50
Opus 4.7	imperative	30	1.00	$+0.950$	0.63
Opus 4.7	conversational	30	1.00	$+0.978$	0.57
Sonnet 4.6	terse	30	1.00	$+0.926$	0.33
Sonnet 4.6	imperative	30	1.00	$+0.911$	0.23
Sonnet 4.6	conversational	30	1.00	$+0.973$	0.40

Table 19. Mkt-RF recall under three reworded Variant-A prompts. Main-sweep baseline (Variant A): Sonnet $r{=}0.98$ ($n{=}77$), Opus $r{=}0.99$ ($n{=}40$). All six (model, phrasing) cells parse 100% and recover $r{\geq}0.91$. Within-25 bps rates drop modestly under terser/imperative phrasings (more one-significant-figure rounding), but rank-correlation r is stable. The recall channel is *not* a property of the exact template wording.

Semantic paraphrase: dropping the keyword entirely. The phrasings above all retain the "Mkt-RF" / "Fama-French" keyword. To test whether the recall is keyed on the FF label or on the underlying quantity, we re-probe the panel with three *semantically* different paraphrases that contain neither "Mkt-RF" nor "Fama-French" anywhere: **S1** *"monthly equity premium over T-bills"*, **S2** *"value-weighted CRSP universe minus risk-free rate"*, **S3** *"U.S. broad-market monthly excess return"* (full templates: `experiments/58_semantic_paraphrase.py`). Sample: 30 months from 1980–2020 (seed 2031). Tab. 20 reports per-(model, paraphrase) recall across seven models with complete coverage and one (Llama-3.3-70B) at partial sample (rate-limited).

Model	Paraphrase	n	parse	r	w-25 bps
Opus 4.7	S1 (premium)	30	1.00	$+0.979$	0.50
Opus 4.7	S2 (CRSP)	30	1.00	$+0.989$	0.47
Opus 4.7	S3 (broadmarket)	30	1.00	$+0.994$	0.50
Sonnet 4.6	S1 (premium)	17	0.57	$+0.905$	0.12
Sonnet 4.6	S2 (CRSP)	20	0.67	$+0.937$	0.25
Sonnet 4.6	S3 (broadmarket)	30	1.00	$+0.858$	0.40
DeepSeek-V3.2	S1 (premium)	30	1.00	$+0.807$	0.07
DeepSeek-V3.2	S2 (CRSP)	30	1.00	$+0.600$	0.23
DeepSeek-V3.2	S3 (broadmarket)	30	1.00	$+0.789$	0.10
GPT-5.4	S1 (premium)	30	1.00	$+0.562$	0.20
GPT-5.4	S2 (CRSP)	30	1.00	$+0.704$	0.13
GPT-5.4	S3 (broadmarket)	30	1.00	$+0.755$	0.30
GPT-5.4-mini	S1 (premium)	30	1.00	$+0.529$	0.10
GPT-5.4-mini	S2 (CRSP)	30	1.00	$+0.513$	0.07
GPT-5.4-mini	S3 (broadmarket)	30	1.00	$+0.517$	0.10
GPT-5.4-nano	S1 (premium)	30	1.00	$+0.375$	0.00
GPT-5.4-nano	S2 (CRSP)	30	1.00	$+0.321$	0.03
GPT-5.4-nano	S3 (broadmarket)	30	1.00	$+0.263$	0.00
Haiku 4.5	S1 (premium)	30	1.00	$+0.169$	0.10
Haiku 4.5	S2 (CRSP)	30	1.00	$+0.059$	0.07
Haiku 4.5	S3 (broadmarket)	30	1.00	$+0.123$	0.03
Llama-3.3-70B†	S1	10	0.33	-0.299	0.00
Llama-3.3-70B†	S2	5	0.38	$+0.020$	0.00

Table 20. Mkt-RF recall under three semantic paraphrases (no "Mkt-RF" / "Fama-French" string anywhere) across the panel. The recalled representation is keyed on the underlying quantity, not the FF label: top-tier (Opus, Sonnet) saturates at $r{\geq}0.86$; mid-tier (DeepSeek-V3.2 a 4th-vendor surprise at $r{=}0.60$–0.81, GPT-5.4 at 0.56–0.76, GPT-5.4-mini at 0.51–0.53) shows partial recall; sub-tier (GPT-5.4-nano 0.26–0.38, Haiku 0.06–0.17) collapses to chance. Sonnet's parse rate falls below 1.0 on S1/S2 because the paraphrase ambiguates the requested quantity (T-bill maturity unspecified; CRSP variant unspecified) and the model declines on ambiguous cells; the panel-wide ordering preserves the same capability monotone as the labelled probes (Tab. 12). The Haiku/Llama low values quantify the extent to which lower-tier models depend on the FF keyword to recover recall. †Llama-3.3-70B run partially completed (Groq rate-limited); data shown for completeness, not interpreted as headline.

E. Leak attribution

This appendix contains the leak-attribution material: the co-located residualization (Eq. 2, headline on Sonnet/Opus), the transmission-coefficient measurement, the ancient-era placebo, the worst-case bound used as a fallback when co-location is unavailable, and the Lopez-Lira and Tang (2023) worked example.

E.1. Co-located residualization

If date-conditioned memorization behaves as distributional readout, an LLM-derived signal $\hat{S}_t$ conditioned on a date may inherit covariation with $r_{f,t}$ from the same recall channel that surfaces in $\hat{r}_{f,t}$, even when the prompt does not explicitly request the factor value. Let $\hat{r}_t$ denote the model's own recall of $r_{f,t}$ elicited on the same months as $\hat{S}_t$ (operationally: re-query the model for the factor return on those months, with no other context). Regressing $\hat{S}$ on $\hat{r}$ gives $\hat{S}_t = \gamma\,\hat{r}_t + u_t$. The *leak-attributable signal share*,

$$\text{LeakShare} \;=\; 1 - \rho(u, r_{FF})^2 \big/ \rho(\hat{S}, r_{FF})^2 \in [0, 1], \tag{2}$$

is a point estimate of the share of date-conditional $(\hat{S}, r_{FF})$ covariation explained by the recall channel. Applied to the transmission data of §E.2 (with the model's sentiment as $\hat{S}$): on Sonnet (n=77), $\rho(\hat{S}, r_{FF})$=+0.74 → $\rho(u, r_{FF})$=+0.02, giving LeakShare=99.9% (95% CI [99.2%, 100.0%]). On Opus (n=40), $\rho(\hat{S}, r_{FF})$=+0.64 → $\rho(u, r_{FF})$=+0.02, giving LeakShare=99.9% (95% CI [97.6%, 100.0%]). Confidence intervals are percentile bootstrap over 10,000 month-resamples (script `experiments/60_leak_share_bootstrap.py`, seed 2026). On the probed sample the recall channel accounts for nearly all of the date-conditional sentiment-vs-truth covariation; the non-recall residual is indistinguishable from zero on both models. The validity of the point estimate rests on a placebo-controlled assumption that $\hat{r}$ is not itself contaminated by date-narrative covariation that $\hat{S}$ separately picks up; the ancient-era placebo (§E.3) supports this assumption.

Pipeline scope. Eq. (2) requires that the LLM be queryable for the target quantity $r_{f,t}$. Pipelines where the LLM emits only an intermediate artifact (classification, summary, relevance score) that is then aggregated downstream require a separate transmission-coefficient measurement of the kind in §E.2. When co-location fails because the original model is API-deprecated, the worst-case bound of §E.4 is an information-poor fallback.

E.2. Transmission estimate for sentiment pipelines

Transmission from recall to date-conditional sentiment is approximately complete: the model's sentiment estimate tracks truth Mkt-RF as tightly as it tracks the model's own recall, so date-conditional sentiment inherits whatever leak the recall channel carries. We re-query Sonnet and Opus with a date-anchored sentiment prompt (sentiment in $[-1, +1]$; no news content); date-conditional covariation upper-bounds transmission in any pipeline where the date is recoverable. We regress sentiment on truth Mkt-RF (slope β_T) and on the model's recall estimate (slope β); the two slopes coincide within sampling error: Sonnet (n=77) gives β_T=+0.066 vs. β=+0.064 (both r=+0.74), and Opus (n=40) gives β_T=+0.076 vs. β=+0.078 (r=+0.64, +0.63). A one-σ Mkt-RF month (σ≈4.5%) induces ≈ 0.32 sentiment units of date-driven bias. As an evaluation implication, any LLM-finance signal with sentiment-style outputs should report $|\rho(\hat{S}, r_{FF})| > \rho_{\text{recall}}$ or run a transmission control before claiming alpha.

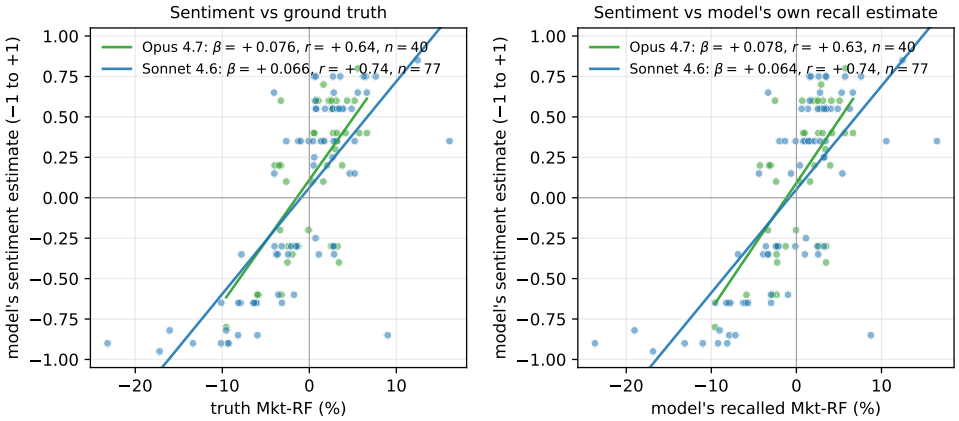

Figure 9. Date-conditional sentiment vs. truth Mkt-RF (left) and vs. the model's own recall estimate (right). Sonnet n=77, Opus n=40. The two slopes per model are nearly identical ($+0.066/+0.064$ Sonnet, $+0.076/+0.078$ Opus).

Permutation null on the slope. Permuting the (date, truth-Mkt-RF) pairing 10,000 times within each model gives a null 95% interval of $[-0.020, +0.020]$ for Sonnet (n=77) and $[-0.037, +0.038]$ for Opus (n=40). The observed slopes ($+0.066$, $+0.076$) sit 3–4σ outside the null with two-sided $p<10^{-4}$ on both models. The identical permutation test on β (sentiment $\sim$ recall-estimate) gives $p<10^{-4}$ on both models.

E.3. Ancient-era placebo

A natural concern on the transmission estimate is that the slope identity ($\beta_T \approx \beta$) could be explained by an *independent* date-to-sentiment channel that bypasses articulated Mkt-RF recall. We test this by sampling 30 months from the 1926–1965 pre-modern era (seed 2026, n=30 per model on Sonnet/Opus) where training-data density on specific monthly returns is far thinner; for each month we elicit both the Variant-A Mkt-RF recall and the same date-conditional sentiment prompt.

Model	Era	$\lvert\rho_{\text{recall}}\rvert$	β_T	β
Sonnet 4.6	1965-2020 (n=77)	0.98	+0.066	+0.064
Sonnet 4.6	1926-1965 (n=30)	0.31	+0.061	+0.012
Opus 4.7	1965-2020 (n=40)	0.99	+0.076	+0.078
Opus 4.7	1926-1965 (n=30)	0.50	+0.065	+0.034

Table 21. When recall fidelity collapses (ancient era), the recall-mediated slope β collapses with it (5$\times$ reduction on Sonnet, 2$\times$ on Opus), while the truth-correlated slope β_T stays roughly intact, consistent with sentiment in low-recall regimes drawing on era-narrative knowledge that bypasses point recall of monthly returns. The slope identity $\beta_T \approx \beta$ is thus a regime property of the high-recall era, not a generic finding.

E.4. Worst-case bound (fallback when co-location is unavailable)

When the original model is API-deprecated and the residualization of §E.1 cannot be evaluated, we report an information-poor worst-case bound that uses only the published signal's reported $|\rho(\hat{S}, r_{FF})|$ and a contemporary ρ_{recall} measurement. Let $r_{FF,t}$ be the true factor return, $\hat{S}_t$ a published LLM-derived signal, and $\tilde{r}_{FF,t}$ the model's noisy recall with correlation $\rho_{\text{recall}} := \rho(\tilde{r}_{FF}, r_{FF})$. We assume $\sigma(\tilde{r}_{FF}) \approx \sigma(r_{FF})$, which holds empirically for Sonnet$\times$Mkt-RF where the OLS slope of estimate on truth is ≈ 1.

Decompose the published signal into a part spanned by the memorized series and an orthogonal residual:

$$\hat{S}_t \;=\; \lambda\, \tilde{r}_{FF,t} \;+\; \varepsilon_t, \qquad \varepsilon \perp \tilde{r}_{FF}. \tag{3}$$

The reported alpha of $\hat{S}$ against r_{FF} is proportional to $\mathrm{cov}(\hat{S}, r_{FF})$. Under Eq. 3,

$$\mathrm{cov}(\hat{S}, r_{FF}) = \lambda\, \mathrm{cov}(\tilde{r}_{FF}, r_{FF}) + \mathrm{cov}(\varepsilon, r_{FF}). \tag{4}$$

The leak contribution is $\lambda \operatorname{cov}(\tilde{r}_{FF}, r_{FF})$; the worst case is when ε is uncorrelated with r_{FF}, i.e. the signal has no genuine factor-spanning content outside what the model already memorized. In that worst case, substituting $\lambda = \rho(\hat{S}, \tilde{r}_{FF}) \cdot \sigma(\hat{S})/\sigma(\tilde{r}_{FF})$ from the OLS projection in Eq. 3 and taking $\rho(\hat{S}, \tilde{r}_{FF})=1$,

$$\alpha_{\text{leak, max}} = \min\left(1,\ \frac{|\rho_{\text{recall}}|}{|\rho(\hat{S}, r_{FF})|}\right) \cdot \alpha_{\text{paper}}. \tag{5}$$

The min caps the ratio at 1 because an upper bound on leak cannot exceed the reported alpha itself.

Why this is an upper bound. The bound assumes (i) the model's recall variance matches the truth's (violated when recall is damped: bound loosens toward 1, i.e. more conservative), (ii) the signal is worst-case aligned with the memorized series, and (iii) the residual ε carries no additional factor-spanning content. A realistic $\hat{S}$ that only partially encodes memorized recall (e.g., a news-sentiment pipeline whose LLM is not explicitly asked for Mkt-RF) will have $\rho(\hat{S}, \tilde{r}_{FF}) \ll 1$ and the realized leak will be smaller. We have no method to bound the realized leak *from below* using only reported statistics. Eq. (5) therefore functions as an escalation trigger when co-location is unavailable, not as a quantitative estimate.

E.5. Worked example: Lopez-Lira and Tang (2023)

The published GPT-4 news-sentiment strategy reports a daily FF5 alpha of 0.33% (t=4.62, Sharpe 2.97) at signal–market correlation $|\rho(\hat{S}, r_{FF})| \sim 0.07$. Plugging into Eq. 5, every $|\rho_{\text{recall}}|$ we observe on Mkt-RF across the nine LLMs (range $[0.32, 0.99]$, including the GPT-5.4-mini capability-proxy at 0.65) is well above 0.07, so the bound caps at $\alpha_{\text{leak,max}}=\alpha_{\text{paper}}$. The reported alpha is observationally compatible with benchmark recall under worst-case transmission. *This does not claim the leak is realized*: a sentiment pipeline that does not explicitly query Mkt-RF will have $\rho(\hat{S}, \tilde{r}_{FF}) \ll 1$ and a realized leak smaller than the bound. The transmission-coefficient measurement in §E.2 is the empirical companion to this worst-case envelope. The GPT-4 deployment remains API-reachable; running Eq. 2 on the co-located GPT-4 sentiment → Mkt-RF recall pair would yield a numerical LeakShare on the published signal (a temporal-stationarity caveat applies; §5).

F. Full per-cell results tables

F.1. Twelve-cell Variant-A calibration grid

Companion to the main-text Fig. 2: the full 12-cell Variant-A calibration grid (Sonnet and Haiku × all six Fama–French factors). Sonnet×Mkt-RF (top-left, r=0.98) shows clean 45° alignment against eleven noise blobs at the factor-shuffle null, providing the visual basis for P1 (cell-localized recall). Points are colored by cutoff bucket: pre-cutoff, near-cutoff, post-cutoff; within Sonnet×Mkt-RF all three buckets land on the diagonal, consistent with the parsability-gate reading of P2 (§3.1).

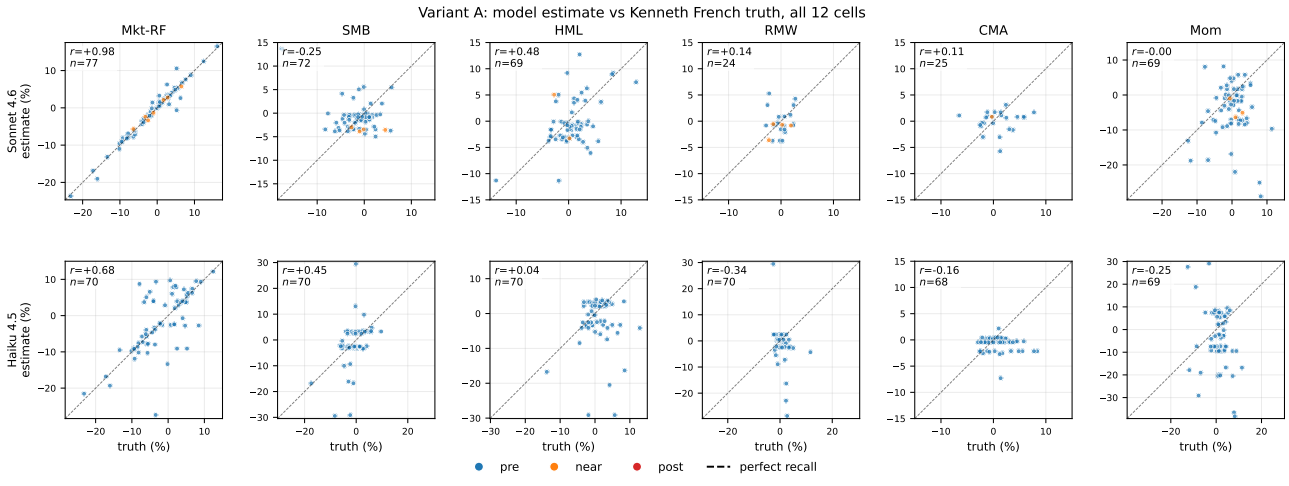

Figure 10. Variant-A calibration on Sonnet (top row) and Haiku (bottom row) across all six Fama–French factors. Annotations: Pearson r and parsed-cell count n. Mkt-RF is the only column with clean 45° alignment.

F.2. Per-factor headline results (full 9×6 table)

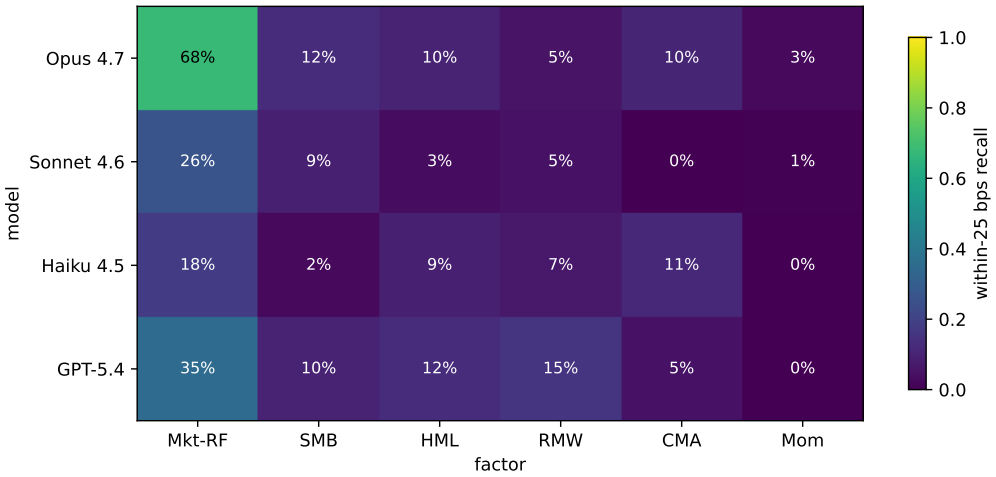

Figure 11. Within-25 bps recall rate per (model, factor), computed from each model's main Variant-A sweep (single-seed-42, parsed-only denominator). Mkt-RF is the only column that recovers monthly values at rates meaningfully above chance, for every model. Haiku's Mkt-RF cell (18% here) is single-seed; the honest 3-seed pooled value is 12% (see Tab. 1, Sec. D.5). Other factors stay at $\leq 15\%$ for every cell.

Table 22 reports the full 9-model $\times$ 6-factor breakdown summarized by Tab. 1 in the main text. *Provenance for Tab. 1*: Sonnet/Haiku Mkt-RF n comes from the 2,784-query main sweep; Opus/GPT-5.4 from the 40-month baseline probes; the best-non-Mkt-RF row reports the factor with maximum $|r|$ per model (remaining factors are at chance, included in this full grid). The Mkt-RF column dominates everywhere; the next-most-prominent factor (SMB) shows scattered partial recall across capability tiers (Opus $r{=}+0.44$, Haiku $r{=}+0.45$, DeepSeek-V3.2 $r{=}+0.46$, GPT-5.4-mini $r{=}+0.40$) without a strict capability-tier monotone, while HML partial recall is concentrated on Opus ($r{=}+0.58$). RMW, CMA, and Mom sit at chance everywhere. Llama-3.1-8B refuses every Fama-French query (parse rate 0 on all six factors), consistent with a capability floor below which the model declines to commit.

G. Diagnostic case study on cited corpus

App. E prescribes four downstream controls. We apply each to the seven LLM-finance papers we cite, in order to ground the prescription in a concrete corpus rather than leave it as protocol ipse dixit. Tab. 23 reports per-paper coverage: ✓ if the paper as published already runs that control or its equivalent; • if our protocol would have flagged a gap. Across the seven, no published paper runs all four controls; the Lopez-Lira et al. (2025) memorization-problem paper is the only one that already runs *any* memorization-specific control comparable to our synthetic-factor perturbation.

Table 23. Diagnostic-protocol coverage by published LLM-finance paper. Cutoff-aware = explicit holdout beyond model training cutoff; synthetic-factor = probe with fictional/perturbed factor names; transmission = measure $\hat{S} \sim \hat{r}$ co-located residualization; label-stripping = blind/anonymized probe. ✓ = control runs; • = gap our protocol flags; "–" = not applicable (paper does not produce a return-relevant signal).

Paper	cutoff-aware	synthetic-factor	transmission	label-stripping
Lopez-Lira and Tang (2023)	•	•	•	•
Lopez-Lira et al. (2025)	✓	✓	•	•
Crane et al. (2025)	✓	•	•	•
Didisheim et al. (2025)	✓	•	•	•
Li et al. (2025)	✓	•	•	•
Benhenda (2026)	✓	•	–	•
Sarkar and Vafa (2024)	✓	•	–	•

What the corpus reveals. (i) Cutoff-aware evaluation has been adopted by the recent leakage benchmarks (Crane et al., 2025; Didisheim et al., 2025; Li et al., 2025; Benhenda, 2026; Sarkar and Vafa, 2024); one foundational sentiment-strategy

Table 22. Variant A headline metrics: nine LLMs (8 informative) on the six Fama-French factors. Wilson-score 95% CIs on proportions; 1,000-sample bootstrap CI on Pearson r. "Sign" is conditional on non-zero truth. Bold: Mkt-RF rows. Mkt-RF n comes from the 2,784-query main sweep for Sonnet/Haiku and from 40-month baseline probes for the other seven models; all other factors use 40-month probes. Opus, Sonnet, GPT-5.4, GPT-5.4-mini Mkt-RF rows here are single-seed-42; 3-seed pooled estimates (Sec. D.5) are: Opus r=0.995, $w25$=0.59 (seed-stable); GPT-5.4 r=0.54, $w25$=0.33 (single-seed-42 was a favorable draw); GPT-5.4-mini r=0.36, $w25$=0.20 (single-seed-42 also favorable, restoring the OpenAI within-vendor monotone in r). †Llama-3.1-8B refused every Fama-French query (parse rate $0/40$ per cell), so no statistic is computable; the empty-row pattern is itself the result.

Model	Factor	n	within-25 bps	Sign	Pearson r
Opus 4.7	**Mkt-RF**	40	**0.68** [0.52, 0.80]	**1.00** [0.91, 1.00]	**0.99** [0.97, 1.00]
Opus 4.7	SMB	40	0.12 [0.05, 0.26]	0.78 [0.62, 0.88]	+0.44 [−0.04, 0.80]
Opus 4.7	HML	40	0.10 [0.04, 0.23]	0.68 [0.52, 0.80]	+0.58 [−0.29, 0.91]
Opus 4.7	RMW	38	0.05 [0.01, 0.17]	0.47 [0.32, 0.63]	+0.16 [−0.44, 0.70]
Opus 4.7	CMA	39	0.10 [0.04, 0.24]	0.46 [0.32, 0.61]	+0.12 [−0.48, 0.64]
Opus 4.7	Mom	39	0.03 [0.00, 0.13]	0.41 [0.27, 0.57]	−0.35 [−0.80, 0.16]
Sonnet 4.6	**Mkt-RF**	77	**0.34** [0.24, 0.45]	**0.97** [0.91, 0.99]	**0.98** [0.96, 0.99]
Sonnet 4.6	SMB	72	0.08 [0.04, 0.17]	0.61 [0.50, 0.72]	−0.25 [−0.63, 0.38]
Sonnet 4.6	HML	69	0.03 [0.01, 0.10]	0.49 [0.38, 0.61]	+0.48 [0.15, 0.68]
Sonnet 4.6	RMW	24	0.04 [0.01, 0.20]	0.54 [0.35, 0.72]	+0.14 [−0.35, 0.64]
Sonnet 4.6	CMA	25	0.00 [0.00, 0.13]	0.60 [0.41, 0.77]	+0.11 [−0.18, 0.40]
Sonnet 4.6	Mom	69	0.01 [0.00, 0.08]	0.48 [0.37, 0.59]	−0.00 [−0.35, 0.37]
Haiku 4.5	**Mkt-RF**	70	**0.17** [0.10, 0.28]	**0.77** [0.66, 0.85]	**0.68** [0.51, 0.82]
Haiku 4.5	SMB	70	0.03 [0.01, 0.10]	0.61 [0.50, 0.72]	+0.45 [0.24, 0.63]
Haiku 4.5	HML	70	0.10 [0.05, 0.19]	0.64 [0.52, 0.74]	+0.04 [−0.30, 0.40]
Haiku 4.5	RMW	70	0.07 [0.03, 0.16]	0.44 [0.33, 0.56]	−0.34 [−0.51, −0.19]
Haiku 4.5	CMA	68	0.10 [0.05, 0.20]	0.50 [0.38, 0.62]	−0.16 [−0.39, 0.06]
Haiku 4.5	Mom	69	0.00 [0.00, 0.05]	0.46 [0.35, 0.58]	−0.25 [−0.55, 0.10]
GPT-5.4	**Mkt-RF**	40	**0.35** [0.22, 0.50]	**0.80** [0.65, 0.90]	**0.70** [0.42, 0.89]
GPT-5.4	SMB	40	0.10 [0.04, 0.23]	0.70 [0.55, 0.82]	−0.07 [−0.65, 0.78]
GPT-5.4	HML	40	0.12 [0.05, 0.26]	0.65 [0.50, 0.78]	−0.06 [−0.65, 0.71]
GPT-5.4	RMW	40	0.15 [0.07, 0.29]	0.65 [0.50, 0.78]	+0.28 [−0.50, 0.81]
GPT-5.4	CMA	40	0.05 [0.01, 0.17]	0.42 [0.29, 0.58]	+0.27 [−0.45, 0.80]
GPT-5.4	Mom	40	0.00 [0.00, 0.09]	0.50 [0.35, 0.65]	−0.03 [−0.55, 0.29]
GPT-5.4-mini	**Mkt-RF**	40	**0.35** [0.22, 0.50]	**0.72** [0.57, 0.84]	**0.65** [0.32, 0.85]
GPT-5.4-mini	SMB	40	0.10 [0.04, 0.23]	0.50 [0.35, 0.65]	+0.40 [−0.05, 0.72]
GPT-5.4-mini	HML	40	0.00 [0.00, 0.09]	0.45 [0.31, 0.60]	+0.01 [−0.41, 0.35]
GPT-5.4-mini	RMW	40	0.15 [0.07, 0.29]	0.53 [0.37, 0.67]	+0.13 [−0.28, 0.51]
GPT-5.4-mini	CMA	40	0.05 [0.01, 0.17]	0.42 [0.29, 0.58]	−0.25 [−0.54, 0.05]
GPT-5.4-mini	Mom	40	0.12 [0.05, 0.26]	0.47 [0.33, 0.63]	−0.02 [−0.48, 0.47]
GPT-5.4-nano	**Mkt-RF**	40	**0.03** [0.00, 0.13]	**0.42** [0.29, 0.58]	−0.32 [−0.61, 0.06]
GPT-5.4-nano	SMB	40	0.07 [0.03, 0.20]	0.42 [0.29, 0.58]	−0.08 [−0.41, 0.26]
GPT-5.4-nano	HML	40	0.07 [0.03, 0.20]	0.50 [0.35, 0.65]	−0.09 [−0.42, 0.27]
GPT-5.4-nano	RMW	40	0.10 [0.04, 0.23]	0.47 [0.33, 0.63]	−0.27 [−0.55, 0.02]
GPT-5.4-nano	CMA	40	0.07 [0.03, 0.20]	0.57 [0.42, 0.71]	+0.26 [−0.17, 0.56]
GPT-5.4-nano	Mom	40	0.05 [0.01, 0.17]	0.40 [0.26, 0.55]	−0.08 [−0.35, 0.19]
DeepSeek-V3.2	**Mkt-RF**	40	**0.15** [0.07, 0.29]	**0.72** [0.57, 0.84]	**0.48** [0.15, 0.73]
DeepSeek-V3.2	SMB	40	0.05 [0.01, 0.17]	0.70 [0.55, 0.82]	+0.46 [+0.05, 0.71]
DeepSeek-V3.2	HML	40	0.03 [0.00, 0.13]	0.40 [0.26, 0.55]	−0.06 [−0.37, 0.30]
DeepSeek-V3.2	RMW	40	0.05 [0.01, 0.17]	0.42 [0.29, 0.58]	+0.07 [−0.23, 0.43]
DeepSeek-V3.2	CMA	40	0.07 [0.03, 0.20]	0.47 [0.33, 0.63]	−0.16 [−0.51, 0.19]
DeepSeek-V3.2	Mom	40	0.03 [0.00, 0.13]	0.38 [0.24, 0.53]	−0.30 [−0.62, −0.16]
Llama-3.3-70B	**Mkt-RF**	39	**0.08** [0.03, 0.20]	**0.62** [0.46, 0.75]	**0.31** [−0.09, 0.60]
Llama-3.3-70B	SMB	40	0.05 [0.01, 0.17]	0.65 [0.50, 0.78]	−0.08 [−0.36, 0.20]
Llama-3.3-70B	HML	40	0.00 [0.00, 0.09]	0.45 [0.31, 0.60]	+0.08 [−0.41, 0.57]
Llama-3.3-70B	RMW	40	0.00 [0.00, 0.09]	0.42 [0.29, 0.58]	−0.02 [−0.47, 0.41]
Llama-3.3-70B	CMA	40	0.12 [0.05, 0.26]	0.47 [0.33, 0.63]	+0.21 [+0.03, 0.40]
Llama-3.3-70B	Mom	40	0.05 [0.01, 0.17]	0.42 [0.29, 0.58]	−0.26 [−0.50, −0.02]
Llama-3.1-8B†	**Mkt-RF**	40	**–**	**–**	**–**
Llama-3.1-8B†	SMB	40	–	–	–
Llama-3.1-8B†	HML	40	–	–	–
Llama-3.1-8B†	RMW	40	–	–	–
Llama-3.1-8B†	CMA	40	–	–	–
Llama-3.1-8B†	Mom	40	–	–	–

paper (Lopez-Lira and Tang, 2023) pre-dates the norm. (ii) Synthetic-factor perturbation is rare; only Lopez-Lira et al. (2025) runs a comparable probe. (iii) No paper in the corpus runs a co-located residualization transmission control of the form in Eq. (2); this is the gap our framework most directly addresses. (iv) Label-stripping is absent across all seven; our blind-probe panel result (Tab. 12) is the first cross-vendor evidence we are aware of that label-stripping fails as a defense. The diagnostic protocol therefore does not duplicate existing practice; each control closes a gap visible in the published corpus.

What we cannot validate from the published record. We can identify which controls a published paper *ran*, not what each control *would have changed* in the published conclusions had it been applied. The transmission control on Lopez-Lira and Tang (2023) (Eq. 2 on the co-located GPT-4 sentiment $\rightarrow$ Mkt-RF recall) is operationally feasible (GPT-4 remains API-reachable as of 2026-04) and would deliver a numerical LeakShare on that paper's signal. We do not run it here: the worked example in App. E.5 flags the escalation; running the residualization on a third party's publication is the analyst's role, not the framework paper's.