Temporal Leakage in LLM Backtesting: Measurement, Validation, and Adjusted Scores

Zeyu Zhang[*] *zeyuzhang2028@u.northwestern.edu*
Department of Statistics and Data Science
Northwestern University

Bradly C. Stadie *bstadie@northwestern.edu*
Department of Statistics and Data Science
Northwestern University

Abstract

The standard check for contamination in LLM backtests is simple: compare scores before and after the training cutoff. We show this check is uninformative. Four flagship models fail it on questions they cannot have memorized: every scored question resolved after their cutoffs. The reason is structural. Models legitimately know more about times near their cutoff, so recency mimics leakage, and we prove no passive backtest can separate the two from genuine skill. Measurement, not just detection, requires information from outside the backtest. We supply it in two forms. A known cutoff identifies leakage at the boundary; a matched clean control identifies it globally and yields a leakage-adjusted score. We also derive where leakage hides: it concentrates on outcomes that surprised the crowd and were well covered in training, and partial memorization is disproportionately rewarded. We validate the estimators against ground truth by planting leakage in twin models, where they recover the injected dose and return null on clean questions. Deployed on frontier models, they detect one cutoff-localized signature and, at the audit's power floor, clear five models whose apparent advantages were recency alone. Backtests need not be discarded; they need one defensible reference.

1 Introduction

Large language models are increasingly evaluated by *backtesting*: the model is asked, as of a historical date, about events whose outcomes have since resolved, and its score is read as evidence of how it would forecast the genuine future (Halawi et al., 2024; Karger et al., 2025; Lopez-Lira et al., 2025; Paleka et al., 2026; Gao et al., 2025). The threat is temporal leakage: a web-scale training corpus may already describe the outcome being "predicted," so a high score can reflect memorization rather than skill. The standard defense is a pre/post check, which compares scores before and after the model's training cutoff and flags a model whose advantage sits on the pre-cutoff side (Jain et al., 2025; Benhenda, 2026). Figure 1 shows that check failing. We scored five flagship models on the official ForecastBench panel, restricted to questions that resolved after their documented training cutoffs, so no outcome can be in their training data. Four of the five fail the check anyway, with spurious "leakage" gaps as large as $+0.061$. We read the documented cutoffs as bounding all training data. The matched-window check of Appendix D.2 supports this reading: in a common window, all five gaps collapse together.

The failure is structural, not a flaw of one benchmark. A backtest score mixes three components. *Skill* is what the evaluation wants to measure. *Recency* is legitimate knowledge that varies with distance from the cutoff: training data cover recent periods more densely, and post-cutoff knowledge grows stale (Lazaridou et al., 2021). *Leakage* is memorized outcome information, and it is illegitimate. The same cutoff drives the last two: a question resolving near the cutoff is one the model legitimately knows more about and one whose

[*]Corresponding author.

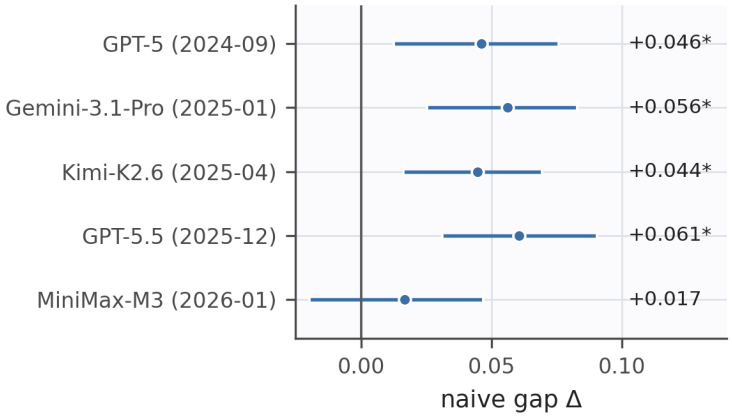

Figure 1: **M1: models that cannot have leaked still fail the naive contamination check.** Each row is one model (cutoff month in parentheses), scored only on questions resolving at least 30 days after its cutoff, so no outcome can be in its training data. The x-axis is its naive gap $\Delta = \bar{s}_{\text{early}} - \bar{s}_{\text{late}}$, the window split at its median resolution date, where $s = (c_0 - Y)^2 - (P - Y)^2$ is the Brier improvement of the model's forecast P over the crowd forecast c_0 (higher is better); $\Delta > 0$ means better on older questions, the pattern an audit reads as leakage. Whiskers: cluster-bootstrap 95% CIs; *: CI excludes zero. Result: four of five flagships are flagged, $\Delta = +0.044$ to $+0.061^*$; recency alone produces the leakage pattern.

outcome is likelier to sit in the corpus. A pre/post contrast therefore measures recency and leakage jointly, and recency alone reproduces the leakage signature. Other failure modes—hallucination, miscalibration, prompt sensitivity (Huang et al., 2025; Xiong et al., 2024; Sclar et al., 2024)—degrade honest performance on both sides of the cutoff; leakage alone is specific to backtesting, and it inflates the score.

The question that matters for practice is therefore not whether contamination exists—it usually does (Sainz et al., 2023; Golchin & Surdeanu, 2024)—but how much it inflates the score, and whether the score can be corrected. Existing tools do not answer it. Detection methods establish that contamination exists, not what it is worth (Oren et al., 2024; Zawalski et al., 2026), and prompt-level remedies fail to suppress recall (Lopez-Lira et al., 2025). Clean-reference estimates on static benchmarks do not transfer, because a backtest defines "clean" by a date rather than by membership (Singh et al., 2024; Haimes et al., 2024). Retraining a truly clean reference model fails, because its cutoff would have to predate every evaluation question, leaving a model too outdated to stand in for the one under audit (Drinkall et al., 2024).

This paper answers the quantitative question in four steps: we prove the inflation is non-identifiable from forecasts and outcomes alone, derive where it concentrates, show that one external reference restores measurement, and validate the estimators against ground truth by planting leakage in twin models. Deployed on frontier systems, the estimators detect one cutoff-localized signature and, at the design's power floors, clear five models whose apparent advantages were recency alone.

Contributions.

- **The standard check is uninformative (Section 3).** Leakage inflation cannot be identified from backtest scores alone: the identified set is a sharp interval that more data does not shrink. A flat pre/post profile is not evidence of a clean backtest, and Figure 1 shows the converse failure on real flagships.
- **Leakage is not a rate; it obeys a law (Section 4).** Per question, inflation equals honest uncertainty times an extraction factor with a double benefit: half the leaked signal already yields three quarters of the full inflation. Leakage concentrates where the crowd was surprised and training coverage was dense, so an average contamination rate is the wrong audit object.
- **A validated audit recipe (Sections 5 to 7).** Choose a reference—a known cutoff (regression discontinuity) or a matched clean control (difference-in-differences)—run the matching estimator,

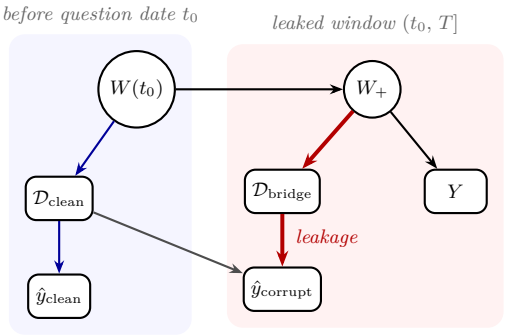

Figure 2: **Temporal leakage is an indirect information flow through a common cause of the training text and the outcome, not verbatim overlap.** The model's training cutoff T postdates the as-of date t_0; a question is leakage-eligible when it resolves inside $(t_0, T]$. The later world state W_+ resolves the outcome Y (black, top) and also generates the bridge data $\mathcal{D}_{\text{bridge}}$, text written inside $(t_0, T]$. The deployed model trains on $\mathcal{D}_{\text{clean}}$ *plus* $\mathcal{D}_{\text{bridge}}$, so the red path $W_+ \to \mathcal{D}_{\text{bridge}} \to \hat{y}_{\text{corrupt}}$ carries outcome information into it without the benchmark question ever appearing in training; the counterfactual clean model $\hat{y}_{\text{clean}}$ (blue) learns from $\mathcal{D}_{\text{clean}}$ alone.

and report the leakage-adjusted score with its assumption; paraphrase probes detect leakage but cannot measure it. The estimators recover planted leakage in twin models in dose, location, and per-question profile, return null on a clean twin, and behave honestly in the wild.

Section 8 positions the work in the contamination literature; proofs, estimator details, and full experimental configurations are in the appendix.

2 Problem Setup

The mechanism. Two dates govern temporal leakage, and Figure 2 makes the mechanism precise. The *as-of date* t_0 is the information horizon a genuine forecast may use; the deployed model's *training cutoff* is $T > t_0$. The world state up to t_0 generates the training corpus $\mathcal{D}_{\text{clean}}$ available to a genuine forecaster, while the later state W_+ is a latent common cause of the "bridge" training data $\mathcal{D}_{\text{bridge}}$ (text written after t_0 but before T) and of the realized outcome Y. A model trained on $\mathcal{D}_{\text{clean}} \cup \mathcal{D}_{\text{bridge}}$ can therefore acquire information about Y through the back-door path $W_+ \to \mathcal{D}_{\text{bridge}} \to \hat{y}_{\text{corrupt}}$, the red leakage path in Figure 2, without the benchmark question ever appearing verbatim in training. Temporal knowledge leakage is this back-door flow; it separates a model trained with the bridge data ($\hat{y}_{\text{corrupt}}$) from an otherwise identical one trained without it ($\hat{y}_{\text{clean}}$). A question such as "*will country X enter a recession by Q4 2023?*" is leakage-eligible if it resolves before the cutoff, where the model may simply recall the reported outcome, and clean if it resolves after, where the model must reason.

Notation. We audit one backtested model with reference cutoff T; uncertainty in T is handled by sensitivity analysis rather than treated as known. For a random question we observe the realized outcome $Y \in \{0, 1\}$; the model's forecast $P \in [0, 1]$ under a fixed deterministic protocol; the *running variable* G, the signed gap between the question's resolution date and the cutoff, negative for leakage-eligible questions ($g < 0$) and positive for clean ones; and a leakage-free difficulty covariate X observable without the model, such as the contemporaneous crowd forecast. We score with the Brier loss $\ell = (P - Y)^2$, a strictly proper score; the identification results of Section 5 require only a bounded loss. The central observable is the conditional mean loss

$$m(x, g) \;=\; \mathbb{E}\big[(P - Y)^2 \mid X = x,\; G = g\big], \tag{1}$$

estimable by regression wherever the backtest has data.

The target. Leakage lowers the pre-cutoff loss below what the model's legitimate information could achieve. Write $m_0(x, g)$ for the *honest surface*: the loss the deployed model would incur using legitimate information only. It still varies with g, because legitimate knowledge depends on temporal distance from the

cutoff; that variation is recency, and the theory imposes no shape or direction on it. The paper's target is the *inflation*

$$B \;=\; \mathbb{E}\big[\, m_0(X,G) - m(X,G) \mid G < 0 \,\big] \;\geq\; 0, \tag{2}$$

the average amount by which the backtest understates the model's honest error on leakage-eligible questions. The goal of the paper is an estimate of the inflation with a confidence interval, and the resulting *leakage-adjusted* score. Appendix A collects all symbols and the standing regularity conventions; proofs are in Appendix B.

3 Scores Alone Cannot Measure Leakage

A completed backtest yields *passive data*: draws of (P, Y, X, G). This section proves that the inflation B is not a functional of their joint law, however many questions are scored, and isolates exactly what is missing, which dictates the form of the remedies in Section 5.

The definition of B in (2) compares the observable surface m with the honest surface m_0, so the content lies in which pairs (m_0, L) are admissible.

Definition 1 (Operational risk decomposition)**.** *A pair of functions $m_0 : \mathcal{X} \times \mathbb{R} \to [0,1]$ and $L : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ is an* operational decomposition *of m if*

$$m(x,g) = m_0(x,g) - L(x,g). \tag{3}$$

Admissible pairs always exist ($(m, 0)$ is one); content enters through a sign and a support restriction.

Assumption 1 (Nonnegative pre-cutoff leakage)**.** *There exists an operational decomposition (m_0, L) such that $L(x,g) \geq 0$ and $L(x,g) = 0$ for every $g \geq 0$.*

The support restriction says the future cannot leak. For a *frozen* model (no retrieval tools, no post-cutoff fine-tuning; cutoff uncertainty is handled by sensitivity analysis, Section 2), an outcome resolving after T cannot have been seen, while before the cutoff, memorized information lowers the loss by $L \geq 0$. Recency lives in m_0: the honest surface may depend on g in any way, and in particular the theory never assumes the model knows more in aggregate about recent times. Under Assumption 1, the inflation is $B = \mathbb{E}[L(X,G) \mid G < 0]$, as in (2).

Theorem 2 (Sharp partial identification from passive scores)**.** *Under Assumption 1 alone, B is not a functional of the law of (P, Y, X, G): every continuation $\tilde{m}_0 : \mathcal{X} \times \mathbb{R} \to [0,1]$ with $m \leq \tilde{m}_0 \leq 1$ on $\{g < 0\}$ and $\tilde{m}_0 = m$ on $\{g \geq 0\}$ defines an admissible pair $(\tilde{m}_0,\ \tilde{L} = \tilde{m}_0 - m)$ consistent with the same passive law, with corresponding inflation $\tilde{B} = \mathbb{E}[\tilde{m}_0 - m \mid G < 0]$. The sharp identified set is*

$$\big[\, 0,\ \ \mathbb{E}[\, 1 - m \mid G < 0 \,]\,\big],$$

and every value in this interval is attainable.

The construction is elementary. Post-cutoff, $m_0 = m$ is observed. Pre-cutoff, the data reveal only the difference $m_0 - L = m$, and the one-parameter family $\tilde{m}_{0,t} = m + t(1 - m)$, $t \in [0,1]$, sweeps the entire interval. The proof is in Appendix B.3.

The content lies in the model rather than the argument. Assumption 1 restricts almost nothing—recency is entirely free, and leakage carries only a sign and a support restriction—so the impossibility cannot be blamed on a restrictive model: once recency and leakage are indexed by the same cutoff, it follows. Nothing is Brier-specific: for any loss bounded by M, the upper endpoint becomes $\mathbb{E}[M - m \mid G < 0]$.

The practical consequence corrects common practice: *a flat pre/post performance profile is not evidence of a clean backtest*, because an honest model with the same recency profile produces identical scores. Figure 1 is the converse error on real models: recency alone produced the failing pattern of four flagships that cannot have memorized any scored outcome.

The theorem also specifies its own remedy. The identified set is exactly the freedom of the honest surface on the leaked side, so identification must import information that pins m_0 there: a clean control model, a continuity restriction at a known cutoff, or an intervention on the question itself. Those are the three routes of Section 5. First, we characterize what L is, which tells an auditor where to look.

4 Where Leakage Concentrates

Leakage is not a uniform rate across questions. It requires both that an outcome was *worth* knowing and that the model *absorbed and used* it. A minimal working model of how leaked information moves a forecast makes this precise and gives the inflation a closed form.

Assumption 2 (Convex-pull leakage)**.** *Under the fixed deterministic prediction protocol, a leakage-eligible question q ($g < 0$) satisfies*

$$P(q) = (1 - w(q))\, P_{\text{hon}}(q) + w(q)\, Y(q)$$

for an extraction weight $w(q) \in [0, 1]$*, where P_{hon} is the honest (legitimate-information) forecast.*

Assumption 2 is a modeling assumption, not a consequence of the leakage problem. Leaked information enters as a convex blend of the honest forecast and the recalled outcome: exact for answer memorization (w=1, where $P \to Y$), first-order for graded evidence leakage. Its substantive restriction is directionality: any forecast between P_{hon} and the outcome is a convex pull for some $w \in [0, 1]$, and what is excluded is leaked information that *misleads* ($w < 0$), the same direction Assumption 1 fixes in aggregate. We test the model's consequences against ground truth in Section 6.

Writing the honest Brier $b_0(q) = (P_{\text{hon}}(q) - Y(q))^2$, the leakage has a closed form.

Proposition 3 (Per-question and aggregate leakage)**.** *Under Assumption 2, the per-question leakage is $b_0(q)\, w(q)\, (2 - w(q))$; the pair $m_0(x, g) = \mathbb{E}[b_0 \mid X{=}x, G{=}g]$ and $L(x, g) = \mathbb{E}[b_0\, w(2 - w) \mid X{=}x, G{=}g]$ is an operational decomposition satisfying Assumption 1, and hence*

$$B \;=\; \mathbb{E}\big[\, b_0\, w(2 - w) \,\big|\, G < 0\big]. \tag{4}$$

The reason is one line: the pull leaves residual error $(1 - w)(P_{\text{hon}} - Y)$, so the leaked Brier is $(1 - w)^2 b_0$ and the saving is $b_0\, w(2 - w)$. Averaging the saving given (X, G) furnishes a canonical decomposition, so the working model automatically satisfies Assumption 1. Per question, inflation is *stakes times extraction*: b_0 measures how much there was to know, and $w(2 - w)$ how much of it the model took.

The double benefit. The factor $w(2 - w) = 1 - (1 - w)^2$ is why partial leakage is disproportionately rewarded (plotted in Figure 5b, Appendix B.1). Pulling the forecast toward the truth earns a benefit linear in w; the residual cost is only quadratic in w. A model that extracts half the leaked signal ($w = 0.5$) already gains 75% of the full-memorization inflation, and $w = 0.3$ still gains 51%. Leakage need not be "memorize the answer" to distort a backtest; a faint, partial trace suffices.

Remark 1 (The counterfactual view)**.** *Appendix B.1 derives the same double benefit from a counterfactual comparison with a clean model retrained without the leaked data (a population loading replaces w; the two agree under homogeneity conditions stated there), and shows that the target* (2) *coincides with the counterfactual inflation (Lemma 12). It also shows the asymmetry: model change* uncorrelated *with the leaked signal strictly deflates the score, so observed inflation is evidence of extraction, not an artifact of generic retraining noise (Corollary 10).*

Two kinds of leakage, two consequences. The extraction weight separates *answer* leakage, where the model has memorized the outcome itself ($w \to 1$), from *evidence* leakage, where it absorbed leaked-window information correlated with Y ($w \in (0, 1)$); a finer interpretive factorization of w is given in Appendix B.3.5. Two consequences of (4) organize the rest of the paper. First, leakage *concentrates*: it is large only where the outcome was surprising (high b_0) and absorbed (high w), and vanishes on easy or obscure questions, so an "average leakage rate" is the wrong audit object. Second, detecting $w > 0$ requires controlling difficulty, skill, recency, and overconfidence simultaneously—a multi-signal estimator, not a pre/post score gap—which is what Section 5 builds.

5 Three References, Two Estimators, One Diagnostic

Theorem 2 specifies what identification must import: information that pins the honest surface on the leaked side. Three practical references supply it: a known training cutoff or a clean control model point-identifies

the inflation under explicit assumptions, and paraphrase-query access detects evidence leakage without identifying it. The results are identification statements for the conditional risk m under a bounded loss; the matching estimators and bootstrap intervals are in Section 6 and Appendix C.

5.1 Route 1, a known cutoff: regression discontinuity

Training cutoffs are published in model cards as documented approximations (hence the sensitivity treatment of Section 2), and the route needs no queries beyond a dated benchmark spanning the cutoff. One restriction turns the cutoff into a reference: honest competence drifts continuously through time, while memorization switches off at the cutoff. Any jump in the loss there is therefore leakage.

Assumption 3 (Honest-risk continuity)**.** *For each x, $m_0(x,\cdot)$ is continuous at $g=0$, and $L(x,0^-)$ exists.*

Theorem 4 (RD identifies boundary leakage)**.** *Under Assumptions 1 and 3 and the support regularity of Appendix B.4, for any bounded loss, the boundary leakage is identified by the upward jump in observed risk at the cutoff,*

$$J(x) \;=\; \lim_{g\downarrow 0} m(x,g) - \lim_{g\uparrow 0} m(x,g) \;=\; L(x,0^-) \;\geq\; 0.$$

The boundary jump is the assumption-light estimand and remains nonzero under perfect recall; quasi-resolved outcomes and ingestion lag attenuate it toward zero, so a small $\widehat{J}$ should not be over-read (Section 6.2). Recovering the *global* inflation needs unique clean-side extrapolation of the honest surface (Assumption 6), under which B is identified (Proposition 14 in Appendix B.4.2).

5.2 Route 2, a clean control model: difference-in-differences

The second route uses a second model: families ship vintages with different cutoffs, so clean controls are standard artifacts, *selected* rather than retrained, and they need to be clean only on the evaluation window. The control need not match the target's capability, because level differences cancel. It must instead match the *response* of honest performance to question timing: if the control shares the target's recency profile, its pre/post change measures recency alone, and differencing pins the target's honest change.

Assumption 4 (Transportable recency)**.** *The control M_0 is clean on the evaluation support ($L^{M_0}\equiv 0$); the observations of both models cover the target's pre-cutoff question mix on both sides of the cutoff; and, once standardized to that mix, honest performance changes by the same amount from pre to post for target and control (formal statement in Appendix B.4).*

Proposition 5 (DiD identification)**.** *Under Assumptions 1 and 4, standardize both models' risks to the target's pre-cutoff question mix and let Δ^A denote model A's resulting pre-to-post change. Then the inflation* (2) *is identified: $B=\Delta^M-\Delta^{M_0}$.*

The parallel-change condition is not fully testable, because the target's pre-cutoff cell is contaminated by construction; its testable implications are in Appendix C. Its dominant violation is signed: a control too weak to express recency *overstates* B—the weak-control trap of Section 7—so a null is conservative, while a positive must carry the matching checks. A boundary variant requiring only a no-discontinuity control underwrites the deployment matrix (Proposition 15). How to read wild estimates is deferred to Section 6.2.

5.3 Route 3, paraphrase probes: detection only

The third route needs only black-box access—a batch of paraphrase calls plus leakage-free calibration anchors—and is a diagnostic, not an estimator. It yields the prediction-residual covariance $\mathrm{PRC} = \mathrm{Cov}(P, Y-P \mid G<0)$ of the paraphrase-consensus forecast, positive under convex-pull leakage with partial extraction (Proposition 16). Three limits keep it secondary: it requires calibration, since LLM overconfidence makes the raw covariance negative (Remark 2); it is blind at full memorization (w=1), where the RD jump is maximal (Proposition 18); and on frontier data it failed its pre-specified calibration-stability gate (Appendix E.8). We use it to detect, never to estimate B (Appendix C).

Table 1: **Three identification routes: input, guarantee, and check.** R1–R2 point-identify B (or boundary J) under stated assumptions; R3 is detection only. Every assumption carries a falsification check exercised in Sections 6 and 7.

Route (input)	Identifies	Key assumption	Check
R1: known cutoff (RD)	boundary J; global B with extrapolation	continuity (Assumption 3); extrapolation (Assumption 6)	placebo jump; bandwidth; date sensitivity
R2: clean control (DiD)	global B; boundary J (Proposition 15)	matched clean control for B; shared jump for J	placebo cutoffs; control balance
R3: paraphrases (PRC)	detection only (not B)	calibration (Remark 2)	calibration on leakage-free anchors; pairing with R1 covers w=1

5.4 One reference suffices; the adjusted score

Either of Routes 1–2 point-identifies B under its stated assumptions, and Route 3 detects without identifying (Corollary 17); one suitable external reference therefore suffices, though we do not claim the family is exhaustive. Table 1 states the input, guarantee, and check for each. When global B is point-identified, report a leakage-adjusted score with a confidence interval; otherwise report the boundary or detection estimand without converting it into a global correction.

Corollary 6 (Leakage-adjusted performance)**.** *Whenever B is point-identified (Route 1 or 2), the leakage-adjusted performance removes B in the appropriate direction: for a risk (any bounded loss, lower is better),* $\text{Risk}_{\text{adj}} = \text{Risk}_{\text{meas}} + B$*; for a higher-is-better score (e.g. accuracy or pass@1),* $\text{Score}_{\text{adj}} = \text{Score}_{\text{meas}} - B$*, with B in score units.*

The route results use only $m = m_0 - L$ and hold for any bounded loss, licensing the pass@1 correction of Section 7. The law $L = b_0\, w(2-w)$ and PRC are Brier-specific (Remark 3). The practitioner's recipe and decision flowchart (Figure 6) are in Appendix C.

6 Validation with Planted Leakage

The estimators of Section 5 claim to measure leakage. Before deploying them on frontier models, we test them where the answer is known. M2 re-analyzes a public controlled-contamination study at pretraining scale. M3 plants temporal leakage in twin models on real forecasting questions and asks the estimators to find it. A claim-to-evidence contract table mapping each theoretical claim to the experiment that carries it—across this section, the deployment of Section 7, and the M1 check of Section 1—is given as Table 4 in Appendix D.

Data: the forecasting panel. All forecasting experiments use one panel built from the public archive of ForecastBench (Karger et al., 2025), the field-standard dynamic forecasting benchmark: 1,646 resolved binary market questions (Polymarket, Metaculus, Manifold, INFER) resolving 2024Q3 through 2026Q3, each carrying its resolution date, its realized outcome Y, and a contemporaneous crowd forecast c_0 frozen before resolution. The score is the crowd-anchored Brier reduction $s = (c_0 - Y)^2 - (P - Y)^2$; positive s means the model beats the crowd. Confidence intervals use a cluster bootstrap over source-by-month clusters; boundary tests add date-permutation and placebo-date checks; configurations, prices, and the pre-registration are in Appendix D (coverage: Appendix D.1).

A pretraining-scale anchor (M2). The Hubble suite (Wei et al., 2026) pretrains a *perturbed* model with benchmark documents inserted at known duplication counts $r \in \{0, 1, 4, 16, 64, 256\}$ and a *standard* twin trained identically without them. Re-analyzing its published accuracies (no new compute) anchors the causal end: benchmark documents inserted into pretraining inflate measured scores. The placebo-centered contrast rises strictly monotonically with dose, from $+0.011$ at $r = 1$ to $+0.403^*$ at $r = 256$, with sixteen duplicates already worth about 21 accuracy points (Figure 7 in Appendix D.3; Spearman $\rho = 1.0$, permutation $p = 0.0083$). The rise is contamination, not composition: the clean twin is nearly flat across

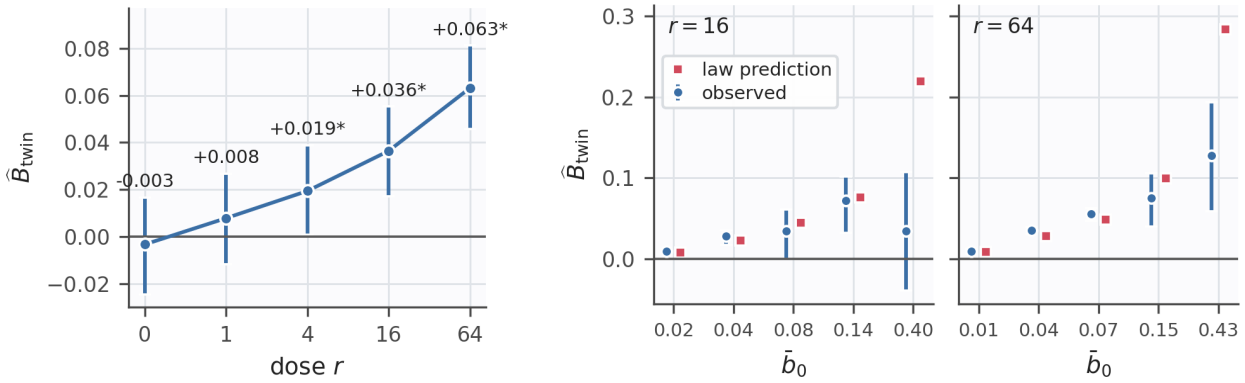

Figure 3: **M3: injected leakage produces a strictly monotone dose-response (left) and concentrates on hard questions as the law $b_0\, w(2-w)$ predicts (right).** Treatment and control twins of one base model train on the same documents; only the treatment copy sees realized outcomes. Both panels plot the twin contrast $\widehat{B}_{\text{twin}} = \bar{\ell}_{\text{ctrl}} - \bar{\ell}_{\text{treat}}$ ($\ell = (P-Y)^2$; positive = leakage inflation); whiskers: bootstrap 95% CIs; *: CI excludes zero. *Left*: by dose r, the number of outcome-stating documents injected per question (randomized across 988 questions): null at $r = 0$ (-0.003), rising strictly monotonically to $+0.063^*$ at $r = 64$. *Right*: by difficulty, questions sorted into quintiles of $b_0 = (P_{\text{ctrl}} - Y)^2$ (ticks: bin mean $\bar{b}_0$), at doses 16 and 64. Red squares: the zero-free-parameter prediction $\bar{b}_0\, \hat{w}(2-\hat{w})$, with $\hat{w}$ fitted from the treatment twin's probability movement alone, never from the plotted contrasts. Inflation grows 13× from easiest to hardest bin ($+0.010$ to $+0.128$ at $r = 64$); the prediction matches except in the hardest bin, where extraction weakens ($\hat{w}$ drops from ≈0.5 to 0.16) and the law acts as an upper bound.

dose bins, and a self-baseline check agrees (Appendix D.3). Hubble is question answering with no time axis, so it cannot test the recency confound; that is what M3 adds (full analysis: Appendices D.3 and E.4).

6.1 Twin models with known temporal leakage (M3)

No observational study can supply per-question ground truth for leakage. We therefore created the leakage ourselves. We continue-train two LoRA twins of Qwen3.5-35B-A3B-Base on the same 95-million-token corpus, which interleaves public filler text with injected news-style documents about panel questions that resolved before a pseudo-cutoff T^* (April 2026). In the *treatment* twin the documents state realized outcomes; in the *control* twin the same documents appear with outcomes scrubbed, matched on source, topic, and length, so both twins gain the same recency and only the treatment twin gains leakage. Injection covers the 988 questions on which the untuned base model is demonstrably uncertain, with dose levels $r \in \{0, 1, 4, 16, 64\}$ randomized across questions, so the full dose grid costs two training runs. Probabilities are read from paired "Yes"/"No" continuation log-probabilities under a fixed few-shot prompt; a ten-percent pilot passed five pre-registered gates before the full runs, and the protocol and logged deviations are in Appendix D.4.

Manipulation checks. Three checks confirm the design does what it claims: the treatment twin memorizes its documents (per-token log-probability exceeds the control twin's by 0.76, CI $[0.72, 0.81]$); the movement is outcome-directed (at dose 64 the treatment twin's probability moves toward the realized outcome by $+0.139$, CI $[+0.119, +0.159]$); and recency is genuinely present (on injected-topic questions the scrubbed control beats the untuned base by $+0.036$ Brier, CI $[+0.028, +0.044]$). The confound the theory targets is in the data.

Inflation rises with dose and is null at dose zero. The twin contrast rises strictly monotonically in dose and is null at dose zero (Figure 3, left). Contamination does not need to be heavy: a single document exposure is worth about one point of crowd-anchored Brier, sixteen about four. The dose-zero null shows the inflation is question-specific, not a generic gain from training, and the curve reproduces M2's pretraining dose-response on real forecasting questions, now with per-item ground truth.

The concentration law holds, and its one miss is interpretable. Proposition 3 predicts $L = b_0\, w(2-w)$. Figure 3 (right) tests the law with zero free parameters: $\hat{w}$ is fitted from the treatment twin's probability

movement alone, never from the plotted contrasts. The prediction falls inside the observed CI in four of five difficulty quintiles at dose 16 and three of five at dose 64, and the globally fitted extraction weight rises with dose ($\hat{w} = 0.24$, 0.23, 0.33, 0.42 at $r = 1, 4, 16, 64$); tercile and ex-ante binnings agree (Appendix D.4). The one miss is the hardest quintile, where extraction weakens and the homogeneous-w law overshoots, so there it acts as an upper bound—the conservative direction for score correction. Contamination distorts a backtest most where the model is weakest.

The boundary estimators recover the injection; recency alone shows no jump. An analyst who did not create this leakage would measure the boundary statistics of Section 5: on the treatment twin the naive pre/post gap at T^* is +0.053 and the local RD jump (90-day bandwidth) is +0.052; on the control twin the same statistics are +0.001 and +0.012, and a placebo boundary 120 days earlier is null (−0.014). Detection and localization both work: injected leakage produces a starred jump at the right date, and recency alone produces none—the control twin absorbed the same 95 million tokens and shows no inflation signature, the real-model face of the no-free-inflation asymmetry (Corollary 10).

PRC detects the evidence, once calibrated. The prediction-residual covariance of Proposition 16 had not previously been tested against known contamination. Raw PRC fails: it is negative in every cell of the twin-by-dose grid, the miscalibration failure Remark 2 warns about. The twin-differenced, temperature-calibrated statistic behaves as predicted: $+0.0069^*$ on injected questions, null at dose zero, monotone in dose (Figure 9 in Appendix D.4). PRC fires where evidence leakage was injected and nowhere else, and only with differencing and calibration.

6.2 How to read a wild estimate

The twins license two reading rules for the deployment results of Section 7.

A boundary estimate is a footprint, not a per-question inflation. On the injected population, per-item ground-truth inflation averages +0.024, roughly half the boundary estimate of +0.052; the remainder is spillover with a measurable mechanism. The injected documents state outcomes that are 83% NO, so the treatment twin acquires a base-rate lean that is nearly free on pre-cutoff questions but depresses post-cutoff scores, and a pre/post estimator cannot tell inflated pre from depressed post (the law analyses are immune: dose-zero differencing removes the lean). A wild RD or DiD estimate is therefore the total contamination footprint at the boundary; it bounds the per-question inflation from above, which makes the score correction of Corollary 6 conservative. Two wild mechanisms the sharp injection cannot exhibit push the other way: outcomes effectively decided in pre-cutoff text raise the honest post-boundary score, and ingestion lag thins memorized outcomes just before it. Both attenuate the jump toward zero: the upper bound holds against spillover only, and a small jump does not exclude leakage deeper in the pre-cutoff period (Appendix D.4).

A null bounds; it does not certify. Every null in Section 7 comes with a power floor: the smallest effect the design would have detected. On the forecasting matrix, the minimum detectable effects at 80% power are 0.05 to 0.11 anchored-Brier units per cell. A null therefore excludes code-domain-sized leakage (about 0.09) for the best-powered cells and says nothing about smaller leakage. No cleanliness certificate is issued below the power floor.

7 Deployment on Frontier Models

We now run the audit where nobody knows the answer. Results are ordered as an auditor should read them: cross-model nulls, disciplined nulls under the matched control, one detection, and a positive control on documented contamination.

7.1 The forecasting matrix: no flagship flagged at its own cutoff

We query four frontier targets with in-window documented cutoffs (rows of Table 2) against the near-clean controls MiniMax-M3 and Claude-Opus-4.7 (their January 2026 cutoffs leave no leakage discontinuity inside

Table 2: **M4-F: no own-cutoff jump among four flagships; the within-family arm flags GPT-5.5.** Cell: $\widehat{J}_m(t) = (\bar{s}_m - \bar{s}_c)_{\text{post}} - (\bar{s}_m - \bar{s}_c)_{\text{pre}}$ over ±90 days of assumed boundary t (s as in Figure 1; c: top block = MiniMax-M3/Claude-Opus-4.7; bottom = GPT-5.5 family archived real-time forecasts). Shaded: own documented cutoff. *: CI excludes zero. Dashes: not defined—the archived-forecast arm exists only at the target's own cutoff. Result: no shaded top-block cell starred (permutation $p = 0.53$); bottom block starred at the shaded cell (pre-specified unbanded / horizon-banded, Section 7.3).

	assumed boundary t			
model (own cutoff)	Mar'25	Apr'25	Aug'25	Dec'25
cross-model: flagship vs. clean controls				
DeepSeek-V3.1 (Mar'25)	−0.009	+0.015	−0.065	+0.018
Kimi-K2.6 (Apr'25)	+0.034	+0.022	−0.020	−0.013
GPT-5.4 (Aug'25)	−0.012	+0.022	−0.051	+0.030
GPT-5.5 (Dec'25)	−0.008	+0.010	−0.011	+0.034
within-family: GPT-5.5 vs. archived real-time forecasts				
GPT-5.5 (Dec'25)	—	—	—	+0.061* / +0.106*

the tested window), computing the boundary jump at every assumed boundary. A genuine signal must jump at the model's own documented cutoff and nowhere else. No own-cutoff cell is starred, and the joint diagonal-versus-off-diagonal permutation test is null ($p = 0.53$). Under the power floors of Section 6.2, the matrix excludes code-domain-sized effects for the best-powered cells while remaining uninformative about smaller leakage. An exploratory Gemini-3.1-Pro row is in Appendix D.5. Run honestly, the cross-model route substantiates no leakage claim for these flagships.

7.2 Disciplined nulls under the matched control

The primary target is Qwen3.5-35B-A3B (year-level 2026 cutoff), evaluated at a July 2025 boundary on panel questions resolving 2025 through mid-2026. Because the cutoff is year-level, the boundary lies inside the plausible training window, so both sides are leakage-eligible: the design detects inflation that *differs* across the boundary; uniform inflation would evade it at every sweep placement. The control is selected *before* reading any DiD result by a pre-registered profile-matching protocol on post-boundary-clean questions, which picks GPT-5 over Gemini-3.1-Pro (profile distance 0.0051 vs. 0.0088; Appendix D.7). Secondary targets are the four other flagships of Figure 4; GPT-3.5-Turbo is a deliberately mismatched weak control.

Naive gaps again raise flags, and again the flags are recency: after the paired, profile-matched DiD, no adjusted estimate is significantly positive (Figure 4). Two checks show the nulls are earned. Substituting GPT-3.5-Turbo re-inflates the primary estimate, so the nulls come from matching, not from an estimator biased toward zero. Semi-synthetic injections are detected with probability 0.97 at effect 0.05 and 1.00 at the code-domain-sized effect 0.09. Full estimates and robustness are in Appendix D.7.

7.3 One detection: a boundary signature for GPT-5.5

ForecastBench archives real-time forecasts, clean by construction, enabling a within-family design: comparing retrospective GPT-5.5 scores with archived scores of GPT-5-Mini and GPT-5.1 cancels the crowd anchor. The jump at GPT-5.5's December 2025 cutoff is +0.061 (CI $[+0.023, +0.093]$), and +0.106 (CI $[+0.056, +0.189]$) inside a 10–75-day anchor-horizon band logged as a dated deviation; both variants are starred (Appendix D.5). Protocol checks and a DeepSeek-V3.1 placebo at the same boundary are null, and the unstarred matrix cell (+0.034) is attenuation plus lower power, not contradiction (Appendix D.5). We scope the finding as Section 6.2 instructs: a boundary leakage signature at GPT-5.5's documented cutoff—a contamination footprint, not an accusation. GPT-5.4's unbanded jump (+0.048) is null once horizons are balanced and is reported as an artifact, not a detection.

7.4 Positive control: documented contamination on dated code

LiveCodeBench (Jain et al., 2025) dates 1,055 problems by contest release and publishes per-problem pass@1 (Remark 3), so the same matrix runs at near-zero cost on a benchmark era where contamination is docu-

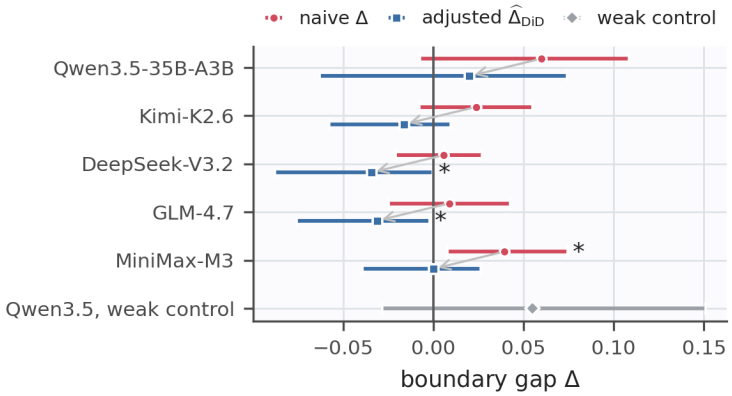

Figure 4: **M5: naive leakage flags dissolve under the matched adjustment.** Red: naive gap $\Delta = \bar{s}_{\text{pre}} - \bar{s}_{\text{post}}$. Blue: adjusted $\widehat{\Delta}_{\text{DiD}} = \Delta_{\text{target}} - \Delta_{\text{control}}$ against pre-registered matched control GPT-5. Grey diamond: primary target vs. mismatched GPT-3.5-Turbo. Whiskers: 95% CIs; *: CI excludes zero. Result: no adjusted estimate significantly positive; mismatched control re-inflates to +0.055.

Table 3: **M4-C: positive control—the same test fires at both models' own cutoffs and nowhere else.** Cell: $\widehat{J}_m(t)$ of Table 2 on LiveCodeBench pass@1 within ±160 days of t; baseline = 2025-cutoff models. Shaded: own cutoff. *: CI excludes zero ($p = 0.023$, $p = 0.017$). The large unstarred placebo (Claude at Dec'24, +0.128) is noise: its CI $[-0.23, +0.33]$ is five times wider than the others because the ±160-day window overruns the end of the problem panel.

	assumed boundary t					
model (own cutoff)	Oct'23	Apr'24	May'24	Jul'24	Sep'24	Dec'24
GPT-4o-0806 (Oct'23)	+0.090*	−0.050	−0.048	−0.056	−0.006	−0.029
Claude-3.5-Sonnet (Apr'24)	+0.017	+0.095*	+0.069	−0.014	−0.019	+0.128

mented. GPT-4o and Claude-3.5-Sonnet are starred exactly at their own cutoffs (date-permutation $p = 0.023$, $p = 0.017$), with all off-diagonal cells null (Table 3). A self-generated extension to 2024-cutoff targets is in Appendix D.6; coverage across M1–M5 spans 2023–2026 vintages of six-plus vendors (Appendix D.1).

8 Related Work

Detecting contamination. A large literature asks *whether* evaluation data was seen in training: n-gram and longest-substring screens (Brown et al., 2020; Singh et al., 2024), membership-inference and likelihood tests (Shokri et al., 2017; Carlini et al., 2021; Shi et al., 2024), exchangeability tests (Oren et al., 2024), and guided or perturbation-based black-box probes (Golchin & Surdeanu, 2024; Zawalski et al., 2026); surveys catalogue detectors and their failure modes (Sainz et al., 2023; Xu et al., 2024; Ravaut et al., 2025). All are membership-oriented, and membership inference is near chance at pretraining scale, its apparent successes an artifact of temporal shift (Duan et al., 2024). Temporal leakage largely escapes these tests, acting through parametric memory along an indirect causal channel (Figure 2), and detection does not answer how much the score is inflated.

Quantifying contamination's score effect. Controlled training studies insert benchmark data deliberately (Magar & Schwartz, 2022; Jiang et al., 2024; Wei et al., 2026); Hubble grounds our M2, M3 adapts its dose design to the temporal setting, and concurrent work fits dose–response curves for generative evaluations (Schaeffer et al., 2026). On deployed models, clean-reference estimators (Singh et al., 2024; Haimes et al., 2024; Zhang et al., 2024; Dekoninck et al., 2024) are instances of our DiD route, overlap audits price contamination on code benchmarks (Riddell et al., 2024), and rephrased samples evade n-gram screens (Yang et al., 2023). Live and date-stamped benchmarks avoid contamination by construction (White et al., 2025; Jain et al., 2025; Wu et al., 2025). All define "clean" by membership in a static test set; a backtest defines

it by a *date*, so every pre-cutoff question is potentially exposed and the reference must come from elsewhere (Section 5).

Temporal leakage and lookahead bias. Cutoff-based natural experiments document the symptom (Roberts et al., 2024; Li & Flanigan, 2024); without a recency model, the same gap is consistent with legitimate recency (Theorem 2). In financial and forecasting backtests, lookahead bias has been separated by anonymization (Glasserman & Lin, 2024), tested on unpredictable events (Sarkar & Vafa, 2024), argued to undermine LLM forecaster evaluation (Paleka et al., 2026), and measured or mitigated on the defining benchmarks (Zou et al., 2022; Halawi et al., 2024; Karger et al., 2025; Liu et al., 2026; Benhenda, 2026; Gao et al., 2025; Zhang et al., 2026; Li et al., 2026). Retraining avoids leakage by construction (Lazaridou et al., 2021; Dhingra et al., 2022; Drinkall et al., 2024; He et al., 2025; Yan et al., 2026) but cannot evaluate a given frontier model. Closest to us, Lopez-Lira et al. (2025) prove counterfactual forecasting ability is non-identified once a model trains on realized outcomes; we add a distinct recency channel, derive where leakage concentrates, and show that one external reference restores measurement. Classical identification tools are catalogued in Appendix C.4.

9 Limitations and Conclusion

Our guarantees are conditional on assumptions made explicit and, where possible, tested. The per-question law excludes misleading leakage ($w < 0$, Assumption 2), so the estimands measure *net* inflation; the twins impose rather than test this restriction (their documents state only true outcomes), and they show that homogeneous-w overstates inflation on the hardest questions (Section 6.1). DiD requires a clean, recency-matched control (Assumption 4): a weak control overstates B, a failure we demonstrate with GPT-3.5-Turbo and guard against with pre-registered profile matching (necessary, not sufficient). RD rests on continuity (Assumption 3); converting the boundary jump into global B adds an extrapolation assumption (Assumption 6) that our code data do not support. A wild boundary jump is best read as a total contamination footprint, and quasi-resolved outcomes and ingestion lag attenuate it toward zero (Section 6.2). PRC has a ground-truth validation but failed its real-data calibration-stability gate (Appendix E.8), so it remains a detection tool. Empirically, the twins are LoRA continued-training runs (M2 anchors pretraining), documented cutoffs are month-resolution, and the M5 nulls exclude boundary-differential inflation above 0.05 Brier-reduction units, not uniform inflation over the training window (Section 7.2). Natural next steps are a frontier-scale planted-leakage pretraining run, a capability-matched dated model pair that point-identifies a nonzero global B, content (rather than framing) perturbations that identify w, extensions to non-binary and ranking outputs, and transferable PRC calibration, promoting Route 3 from detection toward estimation.

Backtest scores mix skill, recency, and leakage, and we proved that no passive backtest can separate the three: a flat pre/post profile is not evidence of a clean backtest. The inflation is nevertheless systematic: under a minimal model of partial recall it obeys $B = \mathbb{E}[b_0\, w(2-w) \mid G < 0]$, concentrating where the crowd was surprised and training coverage was dense. One external reference restores measurement: a known cutoff identifies leakage at the boundary, and a matched clean control identifies it globally and yields a leakage-adjusted score. Validated against ground truth, the estimators recovered planted leakage in dose, location, and per-question profile; deployed in the wild, they detected one cutoff-localized signature and the documented code contamination, and, at the stated power floors, cleared five models whose apparent advantages were recency alone. Backtests need not be discarded; they need one defensible reference.

Broader Impact and Reproducibility Statements

This paper reports a contamination finding about a named commercial model (GPT-5.5), held to the evidentiary standard the paper argues for (Section 7.3): language scoped to a boundary leakage signature, not an accusation of intent or a global capability claim. The work otherwise improves evaluation hygiene and raises no concerns beyond those of the underlying public benchmarks. All analysis code, the forecasting panel, twin-training configurations, pre-registration and deviations documents, and per-experiment results are available at `https://github.com/ZeyuZhang1901/Temporal-Leakage-Backtesting`; Appendix F maps each result to its script and output.

References

Mostapha Benhenda. Look-ahead-bench: A standardized benchmark of look-ahead bias in point-in-time LLMs for finance. *arXiv preprint arXiv:2601.13770*, 2026.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650, 2021.

Jasper Dekoninck, Mark Niklas Müller, and Martin Vechev. ConStat: Performance-based contamination detection in large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 2022.

Felix Drinkall, Eghbal Rahimikia, Janet Pierrehumbert, and Stefan Zohren. Time machine GPT. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3281–3292, 2024.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024. arXiv:2402.07841.

Graham Elliott and Allan Timmermann. *Economic Forecasting*. Princeton University Press, 2016.

Wayne A Fuller. *Measurement Error Models*. John Wiley & Sons, 1987.

Zhenyu Gao, Wenxi Jiang, and Yutong Yan. Detecting lookahead bias in LLM forecasts. *arXiv preprint arXiv:2512.23847*, 2025.

Paul Glasserman and Caden Lin. Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. *The Journal of Financial Data Science*, 6(1):25–42, 2024.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

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

Jacob Haimes, Cenny Wenner, Kunvar Thaman, Vassil Tashev, Clement Neo, Esben Kran, and Jason Schreiber. Benchmark inflation: Revealing LLM performance gaps using retro-holdouts. *arXiv preprint arXiv:2410.09247*, 2024.

Danny Halawi, Fred Zhang, Yueh-Han Chen, and Jacob Steinhardt. Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jerry A Hausman. Specification tests in econometrics. *Econometrica*, 46(6):1251–1271, 1978.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008.

Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2403.07974.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*, 2024.

Ezra Karger, Houtan Bastani, Yueh-Han Chen, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations (ICLR)*, 2025.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Barlow, and Phil Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29348–29363, 2021.

David S. Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, 2010.

Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18471–18480, 2024.

Weixian Waylon Li, Mengyu Wang, and Tiejun Ma. Summoning the oracle to slay it: Mitigating look-ahead bias in financial backtesting with large language models. *arXiv preprint arXiv:2605.24564*, 2026.

Yachuan Liu, Xiaochun Wei, Lin Shi, Xinnuo Li, Bohan Zhang, Paramveer Dhillon, and Qiaozhu Mei. ExAnte: A benchmark for ex-ante inference in large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1551–1571, 2026. arXiv:2505.19533.

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

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 157–165, 2022.

Charles F Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, 2003.

Jacob A Mincer and Victor Zarnowitz. The evaluation of economic forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pp. 3–46. NBER, 1969.

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4): 595–600, 1973.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black-box language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.17623.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating language model forecasters. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2506.00723.

Andrew J Patton and Allan Timmermann. Forecast rationality tests based on multi-horizon bounds. *Journal of Business & Economic Statistics*, 30(1):1–17, 2012.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. A comprehensive survey of contamination detection methods in large language models. *Transactions on Machine Learning Research*, 2025. arXiv:2404.00699.

Martin Riddell, Ansong Ni, and Arman Cohan. Quantifying contamination in evaluating code generation capabilities of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14116–14137, 2024. arXiv:2403.04811.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. To the cutoff... and beyond? a longitudinal perspective on LLM data contamination. In *International Conference on Learning Representations (ICLR)*, 2024.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787, 2023.

Suproteem K. Sarkar and Keyon Vafa. Lookahead bias in pretrained language models. *SSRN working paper 4754678*, 2024.

Rylan Schaeffer, Joshua Kazdan, Baber Abbasi, Ken Ziyu Liu, Brando Miranda, Ahmed Ahmed, Fazl Barez, Abhay Puri, Stella Biderman, Niloofar Mireshghallah, and Sanmi Koyejo. Quantifying the effect of test set contamination on generative evaluations. *arXiv preprint arXiv:2601.04301*, 2026.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 3–18, 2017.

Aaditya K. Singh, Muhammed Yusuf Kocyigit, Andrew Poulton, David Esiobu, Maria Lomeli, Gergely Szilvasy, and Dieuwke Hupkes. Evaluation data contamination in LLMs: how do we measure it and (when) does it matter? *arXiv preprint arXiv:2411.03923*, 2024.

Charles Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295, 1910.

Johnny Tian-Zheng Wei, Ameya Godbole, Mohammad Aflah Khan, Ryan Wang, Xiaoyuan Zhu, James Flemings, Nitya Kashyap, Krishna P. Gummadi, Willie Neiswanger, and Robin Jia. Hubble: A model suite to advance the study of LLM memorization. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.19811.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, et al. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.19314.

Jeffrey M Wooldridge. Control function methods in applied econometrics. *Journal of Human Resources*, 50 (2):420–445, 2015.

De-Min Wu. Alternative tests of independence between stochastic regressors and disturbances. *Econometrica*, 41(4):733–750, 1973.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. AntiLeakBench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18403–18419, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Yutong Yan, Raphael Tang, Zhenyu Gao, Wenxi Jiang, and Yao Lu. DatedGPT: Preventing lookahead bias in large language models with time-aware pretraining. *arXiv preprint arXiv:2603.11838*, 2026.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.

Michał Zawalski, Meriem Boubdir, Klaudia Bałazy, Besmira Nushi, and Pablo Ribalta. Detecting data contamination in LLMs via in-context learning. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.27055.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024. arXiv:2405.00332.

Zeyu Zhang, Ryan Chen, and Bradly C. Stadie. All leaks count, some count more: Interpretable temporal contamination detection and mitigation in LLM backtesting. *arXiv preprint arXiv:2602.17234*, 2026.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Appendix roadmap

Nothing in this appendix is optional filler: it holds the twin formalism demoted from the main text, every proof, the practitioner recipe, full experimental configurations, demoted main-text detail, and the provenance map. Read by role:

- **A.** Appendix A: notation and standing conventions.
- **B.** Appendix B: counterfactual twin theory (Appendix B.1) plus all proofs for Sections 3 to 5, in main-text order, with technical complements (including the interpretive factorization of w).
- **C.** Appendix C: practitioner recipe, control validation, finite-K PRC machinery, paraphrase protocol, and the classical identification-tool map (Appendix C.4).
- **D.** Appendix D: full configurations and complete results for M1–M5, including material demoted from the main text (M1 matched-window check, Gemini exploratory row, clean-anchor banding and design disagreement, code-arm extension).
- **E.** Appendix E: synthetic validation (E1) and per-experiment robustness.
- **F.** Appendix F: every reported number mapped to its script and output file.

A reader checking the theory needs A–B; a reader assessing or replicating the experiments needs D–F; a practitioner deploying the routes needs C.

A Notation and standing conventions

Symbol	Meaning
M, T	backtested model and its training cutoff (Section 2)
Y, τ	realized binary outcome; its resolution time (Section 2)
$G = \tau - T$	running variable; $g < 0$ leakage-eligible, $g > 0$ clean (Section 2)
X	leakage-free difficulty covariate, e.g. crowd forecast c_0 (Section 2)
P	the model's forecast under a fixed protocol (Section 2)
$m(x, g)$	observable conditional mean loss, (1) (Section 2)
$\mathcal{D}_{\text{clean}}$, $\mathcal{D}_{\text{bridge}}$, $\mathcal{D}_{\text{train}}$	pre-t_0 corpus; bridge data from the leaked window $(t_0, T]$; training data (Section 2 and Appendix B.1)
$\hat{y}_{\text{corrupt}}$, $\hat{y}_{\text{clean}}$	corrupt-trained and clean-trained predictions (Section 2 and Appendix B.1)
$\hat{y}_{\text{Bayes}}$, $\hat{y}_{\text{Bayes,clean}}$	Bayesian predictors with/without the bridge data (Appendix B.1)
$S = \hat{y}_{\text{Bayes}} - \hat{y}_{\text{Bayes,clean}}$	leakage signal (Appendix B.1)
$\delta = \hat{y}_{\text{corrupt}} - \hat{y}_{\text{clean}}$	perturbation of the deployed model (Appendix B.1)
λ, μ_δ, $\tilde{\nu}$	loading of δ on S; conditional shift; orthogonal residual (Appendix B.1)
B_{twin}	counterfactual inflation, (5) (Appendix B.1)
$m_0(x, g)$, $L(x, g)$	latent leakage-free risk surface; leakage contribution (Sections 2 and 3)
$B = \mathbb{E}[L \mid G < 0]$	operational inflation, (2) (Section 2)
P_{hon}, w	honest forecast; per-question extraction weight (Section 4)
$b_0 = (P_{\text{hon}} - Y)^2$	honest Brier, the "stakes" (Section 4)
$J(x)$	leakage jump at the cutoff boundary (Section 5)
PRC	prediction-residual covariance, a detection statistic (Section 5)
$\text{Risk}_{\text{adj}} = \text{Risk}_{\text{meas}} + \widehat{B}$	leakage-adjusted score (Section 5)

Standing regularity. All predictions and outcomes have finite second moments. Lowercase x, g denote realizations of X, G. Conditional expectations, variances, and projections are defined almost surely on their relevant support. Whenever $\text{Var}(S \mid X, \mathcal{D}_{\text{train}}, G < 0) = 0$, the projection coefficient $\lambda = \text{Cov}(\delta, S)/\text{Var}(S)$

is a 0/0 expression; we adopt the convention $\lambda = 0$ there, since a signal with no conditional variation carries no leakage to load on. Common support and density conditions are invoked explicitly for the identification route that requires them.

B The twin theory, proofs, and technical complements

The counterfactual twin theory summarized in Remark 1 comes first (Appendix B.1); the proofs of all formal results follow, grouped by the section in which each result appears. Every entry states the context of the result, restates it, and then gives the proof. Interspersed with the proofs are short technical complements that belong with the mathematics rather than the experiments: auxiliary lemmas stated only here, a bound on the systematic shift, and an interpretive parameterization of the extraction weight (Appendix B.3.5).

B.1 The counterfactual twin theory

The counterfactual frame compares the deployed model with a clean twin retrained without the leaked data and asks what leaked information does to their score difference; Remark 1 in the main text summarizes the result. The argument moves from an ideal benchmark (Appendix B.1.1) to general predictors (Appendix B.1.2) to the reason the construction cannot be run in practice (Appendix B.1.3). With the two models of Figure 2, we write $\delta = \hat{y}_{\text{corrupt}} - \hat{y}_{\text{clean}}$ for their disagreement and $\mathcal{H} = \sigma(X, \mathcal{D}_{\text{train}})$ for the pre-cutoff information; $\hat{y}_{\text{corrupt}}$ and $\hat{y}_{\text{clean}}$ denote the corrupt- and clean-trained predictions, and for a probabilistic protocol $\hat{y}_{\text{corrupt}} = P$. All statements and conditional moments in this subsection are under the conditional law of leakage-eligible questions ($G < 0$). The target is the *counterfactual inflation*

$$B_{\text{twin}} \;=\; \mathbb{E}\big[(\hat{y}_{\text{clean}} - Y)^2 - (\hat{y}_{\text{corrupt}} - Y)^2 \mid G < 0\big]\,, \tag{5}$$

positive when the deployed model looks more accurate than its counterfactual clean version. Every result below is measured against the *leakage signal*

$$S \;=\; \underbrace{\mathbb{E}[Y \mid X, \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{bridge}}]}_{\hat{y}_{\text{Bayes}}} \;-\; \underbrace{\mathbb{E}[Y \mid X, \mathcal{D}_{\text{train}}]}_{\hat{y}_{\text{Bayes,clean}}}, \tag{6}$$

the amount by which an ideal learner would revise its forecast upon seeing the leaked data: large exactly when the bridge data would move an ideal forecaster's belief about Y, zero when they carry no information about the question. Because the conditional expectation is the best possible use of the bridge data, S summarizes all any model could extract from $\mathcal{D}_{\text{bridge}}$ about Y—the natural yardstick for leakage.

B.1.1 Bayes-optimal extraction: the value of leaked information

Suppose first that both predictors are Bayes-optimal, $\hat{y}_{\text{clean}} = \hat{y}_{\text{Bayes,clean}}$ and $\hat{y}_{\text{corrupt}} = \hat{y}_{\text{Bayes}}$: the deployed model extracts everything the leaked data contain, and its disagreement equals the signal ($\delta = S$).

Theorem 7 (Bayesian value of bridge information)**.** *For the Bayes-optimal pair,*

$$B_{\text{twin,Bayes}} = \mathbb{E}[\text{Var}(S \mid \mathcal{H}) \mid G < 0] \geq 0.$$

The inflation is *always* non-negative for the full extractor, and its magnitude equals how much the post-cutoff world deviates from the pre-cutoff world, as seen through the outcome: volatile, surprising periods (large $\text{Var}(S \mid \mathcal{H})$) admit large inflation, stable ones almost none. This is the value of bridge information for Bayes-optimal prediction, not a universal upper bound over arbitrary algorithm pairs.

B.1.2 General predictors: decomposition along the signal

The model under backtest is not Bayes-optimal: it extracts only a fraction of S, and not necessarily along S. To see how much of the ideal value it captures—and at what cost—we take the conditional L^2 projection of its perturbation δ onto the signal,

$$\delta \;=\; \lambda\, S \;+\; \mu_\delta \;+\; \tilde{\nu}, \qquad \lambda = \frac{\text{Cov}(\delta, S \mid \mathcal{H})}{\text{Var}(S \mid \mathcal{H})}, \qquad \mu_\delta = \mathbb{E}[\delta \mid \mathcal{H}], \tag{7}$$

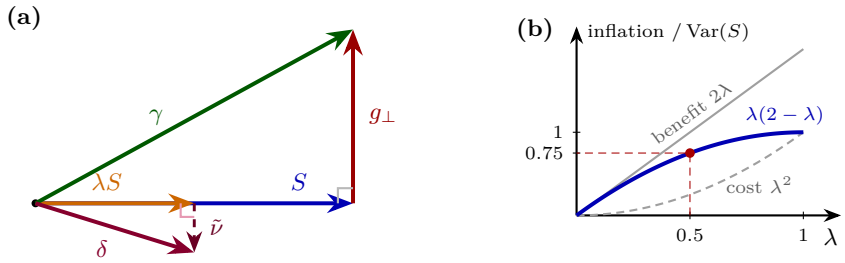

Figure 5: **Why leakage inflates a backtest: the decomposition (a) and the double benefit it implies (b).** **(a)** What the clean model does not know about the outcome—the clean gap γ (green)—splits into the leakage signal S (blue), everything the leaked data reveal about Y, and the residual $g_\perp$ (red), which no training data can predict. Training on leaked data moves the model by δ (purple), of which only the extracted component λS (orange) overlaps γ and can raise the score; the noise $\tilde{\nu}$ (dashed) strictly deflates it (Corollary 10). **(b)** Extracting a fraction λ of the signal buys a score benefit linear in λ at a variance cost only quadratic in λ, so the net inflation $\lambda(2-\lambda)\mathrm{Var}(S)$ rises steeply from zero: extracting half the signal (dashes) already yields 75% of the full-memorization inflation. A faint leakage trace suffices to distort a backtest. The per-question factor $w(2-w)$ of Proposition 3 traces the identical curve with w in place of λ.

with $\lambda = 0$ on strata where the denominator vanishes; the residual $\tilde{\nu}$ has conditional mean zero and is conditionally orthogonal to S. Two benchmark deviations complete the bookkeeping: $r_{\text{clean}} = \hat{y}_{\text{clean}} - \hat{y}_{\text{Bayes,clean}}$, the clean model's deviation from its own Bayes benchmark, and $g_\perp = Y - \hat{y}_{\text{Bayes}}$, the Bayesian residual—the part of the outcome that no training data can predict; the clean gap then decomposes as $\gamma = Y - \hat{y}_{\text{clean}} = S + g_\perp - r_{\text{clean}}$. Figure 5 depicts the geometry of the two decompositions and why only the S-aligned component of δ can pay off. Substituting (7) into (5) gives the central decomposition.

Theorem 8 (Leakage inflation decomposition)**.** *For any clean/corrupt predictor pair with finite second moments, with δ, S, λ, μ_δ, $\tilde{\nu}$ as in* (7) *and r_{clean}, $g_\perp$ as above,*

$$B_{\text{twin}} = P_{\text{signal}} - R_{\text{shift}} - R_{\text{noise}} + R_{\text{cross}}, \tag{8}$$

where, with all outer expectations conditional on $G < 0$, $P_{\text{signal}} = \mathbb{E}[\lambda(2-\lambda)\mathrm{Var}(S \mid \mathcal{H})]$, $R_{\text{shift}} = \mathbb{E}[\mu_\delta^2 + 2\mu_\delta r_{\text{clean}}]$, $R_{\text{noise}} = \mathbb{E}[\mathrm{Var}(\tilde{\nu} \mid \mathcal{H})]$, and $R_{\text{cross}} = 2\,\mathbb{E}[\tilde{\nu} g_\perp]$. For $\lambda \in [0,2]$, $P_{\text{signal}} \geq 0$.

Lemma 9 (The cross-term vanishes)**.** *For any such pair, $R_{\text{cross}} = 0$; hence $B_{\text{twin}} = P_{\text{signal}} - R_{\text{shift}} - R_{\text{noise}}$.*

The cross-term vanishes identically—not by assumption—because the Bayesian residual $g_\perp$ is orthogonal to every square-integrable $\sigma(X, \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{bridge}})$-measurable variable, and $\tilde{\nu}$ is such a variable. The identity itself imposes no sign restriction on λ. Under the additional non-adversarial condition $\lambda \in [0,2]$, $P_{\text{signal}} \geq 0$; $R_{\text{noise}} \geq 0$ deflates the contrast, and the shift R_{shift} is bounded and vanishes when $\mu_\delta = 0$ (Appendix B.2.4). Reading the signs together yields the asymmetry summarized in Section 4.

Corollary 10 (No free inflation)**.** *If $\lambda = 0$ and $\mu_\delta = 0$ almost surely, then $B_{\text{twin}} = -R_{\text{noise}} \leq 0$.*

The corollary is immediate from Theorem 8 and Lemma 9. A perturbation uncorrelated with the leakage signal cannot inflate the backtest: change orthogonal to S strictly *deflates* the score, so observed inflation is evidence of extraction. This asymmetry is what makes detection meaningful; the detection statistic of Section 5 is built on it.

The double benefit in the twin frame. The signal term carries the factor $\lambda(2-\lambda) = 1 - (1-\lambda)^2$, the twin-frame analogue of the per-question factor $w(2-w)$ in Proposition 3. The arithmetic is the same: extracting a fraction λ of S earns a covariance gain of $2\lambda\mathrm{Var}(S)$—the forecast now moves *with* the truth, a benefit linear in λ—while the variance cost of the added component is only $\lambda^2\mathrm{Var}(S)$, quadratic in λ. The net is $\lambda(2-\lambda)\mathrm{Var}(S)$ (Figure 5). The two factors coincide under the loading correspondence of Lemma 13: λ loads δ onto S with signal strength $\mathrm{Var}(S \mid \mathcal{H})$, w is the per-question extraction with signal strength b_0, and when the crowd forecast proxies the honest forecast, the observable crowd surprise plays the role of the latent $\mathrm{Var}(S \mid \mathcal{H})$.

B.1.3 Infeasibility of the counterfactual comparison

The decomposition is exact and cleanly isolates leakage: the deployed model and its counterfactual clean version share the same pre-t_0 data, so everything they have in common—including recency—cancels in δ, leaving the pure effect of the bridge data. The obstruction is that *the counterfactual model cannot be instantiated for a deployed system*: one cannot un-train a released black box, and the Bayes-optimal predictors, the signal S, and the loading λ are all uncomputable. Without the counterfactual model, the only within-model comparison left is *across time*—pre- versus post-cutoff on the one deployed model—and there the cancellation fails, because the model is genuinely fresher near its cutoff: the time comparison reintroduces recency as a confound. The main text therefore works with the operational inflation B of (2), defined on observables of a single model with recency modeled explicitly; it coincides with B_{twin} under the alignment conditions of Lemma 12 (Appendix B.3), and the price of losing the counterfactual comparison is the non-identifiability of Theorem 2.

B.2 Proofs for Appendix B.1: the leakage inflation mechanism

All expectations in this subsection are under the conditional law given $G < 0$. Let $\mathcal{H} = \sigma(X, \mathcal{D}_{\text{train}})$ and $\mathcal{F} = \sigma(X, \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{bridge}})$, with $\mathcal{H} \subseteq \mathcal{F}$. The leakage signal is $S = \hat{y}_{\text{Bayes}} - \hat{y}_{\text{Bayes,clean}}$, where $\hat{y}_{\text{Bayes}} = \mathbb{E}[Y \mid \mathcal{F}]$ and $\hat{y}_{\text{Bayes,clean}} = \mathbb{E}[Y \mid \mathcal{H}]$. Further, $\delta = \hat{y}_{\text{corrupt}} - \hat{y}_{\text{clean}}$, $\gamma = Y - \hat{y}_{\text{clean}}$, $g_\perp = Y - \hat{y}_{\text{Bayes}}$, and $r_{\text{clean}} = \hat{y}_{\text{clean}} - \hat{y}_{\text{Bayes,clean}}$. Under the fixed deterministic protocol of Section 2, predictions are measurable with respect to the information available to the model: $\hat{y}_{\text{clean}}$ is $\mathcal{H}$-measurable and $\hat{y}_{\text{corrupt}}$ is $\mathcal{F}$-measurable, so δ, r_{clean}, and (below) $\tilde{\nu}$ are $\mathcal{F}$-measurable. The decomposition $\delta = \lambda S + \mu_\delta + \tilde{\nu}$ is the conditional L^2 projection given $\mathcal{H}$, with λ and μ_δ both $\mathcal{H}$-measurable.

B.2.1 Proof of Theorem 7

Theorem 7 is the ideal-case benchmark of Appendix B.1.1: when both predictors are Bayes-optimal, the inflation equals the full information value of the bridge data.

Theorem 7 (Bayesian value of bridge information)**.** *For the Bayes-optimal pair,*

$$B_{\text{twin,Bayes}} = \mathbb{E}[\text{Var}(S \mid \mathcal{H}) \mid G < 0] \geq 0.$$

Proof. For the Bayes-optimal pair, $\hat{y}_{\text{clean}} = \hat{y}_{\text{Bayes,clean}} = \mathbb{E}[Y \mid \mathcal{H}]$ and $\hat{y}_{\text{corrupt}} = \hat{y}_{\text{Bayes}} = \mathbb{E}[Y \mid \mathcal{F}]$, so the $\mathcal{H}$-conditional inflation is

$$b(\mathcal{H}) = \mathbb{E}\big[(\hat{y}_{\text{Bayes,clean}} - Y)^2 \mid \mathcal{H}\big] - \mathbb{E}\big[(\hat{y}_{\text{Bayes}} - Y)^2 \mid \mathcal{H}\big] .$$

By the minimum-mean-squared-error property of conditional expectation, the first term equals $\text{Var}(Y \mid \mathcal{H})$, while $\mathbb{E}[(\hat{y}_{\text{Bayes}} - Y)^2 \mid \mathcal{F}] = \text{Var}(Y \mid \mathcal{F})$; conditioning the latter on $\mathcal{H} \subseteq \mathcal{F}$ and invoking the law of total variance,

$$\text{Var}(Y \mid \mathcal{H}) = \mathbb{E}[\text{Var}(Y \mid \mathcal{F}) \mid \mathcal{H}] + \text{Var}(\hat{y}_{\text{Bayes}} \mid \mathcal{H}),$$

we conclude that $b(\mathcal{H}) = \text{Var}(\hat{y}_{\text{Bayes}} \mid \mathcal{H})$. Moreover $\mathbb{E}[S \mid \mathcal{H}] = \mathbb{E}[\hat{y}_{\text{Bayes}} \mid \mathcal{H}] - \hat{y}_{\text{Bayes,clean}} = 0$ by the tower property, so $\text{Var}(S \mid \mathcal{H}) = \text{Var}(\hat{y}_{\text{Bayes}} \mid \mathcal{H}) = b(\mathcal{H})$. Taking expectations over $\mathcal{H}$ yields $B_{\text{twin,Bayes}} = \mathbb{E}[\text{Var}(S \mid \mathcal{H})] \geq 0$, with equality if and only if $S = 0$ almost surely. □

B.2.2 Proof of Theorem 8

Theorem 8 is the central decomposition of Appendix B.1.2: a general predictor's inflation splits into a signal term, a shift term, a noise term, and a cross term. Figure 5 in the main text depicts the geometry.

Theorem 8 (Leakage inflation decomposition)**.** *For any clean/corrupt predictor pair with finite second moments, with δ, S, λ, μ_δ, $\tilde{\nu}$ as in (7) and r_{clean}, $g_\perp$ as above,*

$$B_{\text{twin}} \;=\; P_{\text{signal}} \;-\; R_{\text{shift}} \;-\; R_{\text{noise}} \;+\; R_{\text{cross}}, \tag{8}$$

where, with all outer expectations conditional on $G < 0$, $P_{\text{signal}} = \mathbb{E}[\lambda(2-\lambda)\text{Var}(S \mid \mathcal{H})]$, $R_{\text{shift}} = \mathbb{E}[\mu_\delta^2 + 2\mu_\delta r_{\text{clean}}]$, $R_{\text{noise}} = \mathbb{E}[\text{Var}(\tilde{\nu} \mid \mathcal{H})]$, and $R_{\text{cross}} = 2\,\mathbb{E}[\tilde{\nu} g_\perp]$. For $\lambda \in [0, 2]$, $P_{\text{signal}} \geq 0$.

Proof. Since $\hat{y}_{\text{clean}} - Y = -\gamma$ and $\hat{y}_{\text{corrupt}} - Y = \delta - \gamma$, definition (5) reads

$$B_{\text{twin}} = \mathbb{E}\big[\gamma^2 - (\delta - \gamma)^2\big] = \mathbb{E}\big[2\delta\gamma - \delta^2\big],$$

and it suffices to compute the two conditional moments $\mathbb{E}[\delta^2 \mid \mathcal{H}]$ and $\mathbb{E}[\delta\gamma \mid \mathcal{H}]$. Throughout, recall that λ, μ_δ, and r_{clean} are $\mathcal{H}$-measurable, that $\mathbb{E}[S \mid \mathcal{H}] = 0$ by the tower property, and that $\mathbb{E}[\tilde{\nu} \mid \mathcal{H}] = \mathbb{E}[S\tilde{\nu} \mid \mathcal{H}] = 0$ by construction of the projection (7); we suppress the conditioning on $\mathcal{H}$ in the displays below.

The three components of $\delta = \lambda S + \mu_\delta + \tilde{\nu}$ are conditionally uncorrelated, so

$$\mathbb{E}[\delta^2] = \lambda^2 \mathrm{Var}(S) + \mu_\delta^2 + \mathrm{Var}(\tilde{\nu}).$$

For the cross moment, substituting $\gamma = S + g_\perp - r_{\text{clean}}$ and expanding the product $\delta\gamma = (\lambda S + \mu_\delta + \tilde{\nu})(S + g_\perp - r_{\text{clean}})$ produces nine terms, of which all but three vanish: $\mathbb{E}[Sg_\perp] = \mathbb{E}[g_\perp] = 0$ because $g_\perp = Y - \hat{y}_{\text{Bayes}}$ is the minimum-mean-squared-error residual, orthogonal to every square-integrable $\mathcal{F}$-measurable variable (in particular to S and to constants), while $\mathbb{E}[S] = \mathbb{E}[\tilde{\nu}] = \mathbb{E}[S\tilde{\nu}] = 0$ as noted above. What survives is

$$\mathbb{E}[\delta\gamma] = \lambda \mathrm{Var}(S) - \mu_\delta r_{\text{clean}} + \mathbb{E}[\tilde{\nu} g_\perp].$$

Combining the two moments,

$$\mathbb{E}\big[2\delta\gamma - \delta^2 \mid \mathcal{H}\big] = \lambda(2-\lambda)\mathrm{Var}(S \mid \mathcal{H}) - \big(\mu_\delta^2 + 2\mu_\delta r_{\text{clean}}\big) - \mathrm{Var}(\tilde{\nu} \mid \mathcal{H}) + 2\,\mathbb{E}[\tilde{\nu} g_\perp \mid \mathcal{H}],$$

and taking expectations over $\mathcal{H}$ yields (8). For $\lambda \in [0, 2]$ the factor $\lambda(2-\lambda)$ is non-negative, whence $P_{\text{signal}} \geq 0$. Specializing to the Bayes-optimal pair ($\lambda = 1$, $\mu_\delta = 0$, $\tilde{\nu} = 0$) recovers Theorem 7. □

B.2.3 Proof of Lemma 9

Lemma 9 removes the only unsigned term of Theorem 8, reducing the inflation to signal minus shift minus noise.

Lemma 9 (The cross-term vanishes)**.** *For any such pair, $R_{\text{cross}} = 0$; hence $B_{\text{twin}} = P_{\text{signal}} - R_{\text{shift}} - R_{\text{noise}}$.*

Proof. The Bayesian residual $g_\perp = Y - \mathbb{E}[Y \mid \mathcal{F}]$ satisfies $\mathbb{E}[g_\perp h] = 0$ for every square-integrable $\mathcal{F}$-measurable h. As recorded in the preamble to this appendix, $\tilde{\nu} = \delta - \lambda S - \mu_\delta$ is $\mathcal{F}$-measurable, so taking $h = \tilde{\nu}$ gives $\mathbb{E}[\tilde{\nu} g_\perp] = 0$, and therefore $R_{\text{cross}} = 2\,\mathbb{E}[\tilde{\nu} g_\perp] = 0$; the identity $B_{\text{twin}} = P_{\text{signal}} - R_{\text{shift}} - R_{\text{noise}}$ then follows from Theorem 8. The same argument applied conditionally shows $\mathbb{E}[\tilde{\nu} g_\perp \mid \mathcal{H}] = 0$ almost surely. □

B.2.4 The systematic shift is bounded

The following proposition, stated and proved only here, bounds the shift term R_{shift} of Theorem 8 and identifies when it vanishes.

Proposition 11 (Bounded systematic shift)**.** *For any model, $|R_{\text{shift}}| \leq \mathbb{E}[\mu_\delta^2] + 2\sqrt{\mathbb{E}[\mu_\delta^2]\,\mathbb{E}[r_{\text{clean}}^2]}$, where $\mathbb{E}[\mu_\delta^2] \leq \mathbb{E}[\delta^2]$ and $\mathbb{E}[r_{\text{clean}}^2]$ is the clean model's excess risk relative to the clean Bayesian predictor. In particular $R_{\text{shift}} = 0$ for any predictor with $\mu_\delta \equiv 0$ (including the Bayesian and any conditionally unbiased predictor).*

Proof. Since $R_{\text{shift}} = \mathbb{E}[\mu_\delta^2] + 2\,\mathbb{E}[\mu_\delta r_{\text{clean}}]$, the Cauchy–Schwarz inequality $|\mathbb{E}[\mu_\delta r_{\text{clean}}]| \leq (\mathbb{E}[\mu_\delta^2]\,\mathbb{E}[r_{\text{clean}}^2])^{1/2}$ together with the triangle inequality yields the stated bound. Jensen's inequality applied to $\mu_\delta = \mathbb{E}[\delta \mid \mathcal{H}]$ gives $\mu_\delta^2 \leq \mathbb{E}[\delta^2 \mid \mathcal{H}]$ pointwise, whence $\mathbb{E}[\mu_\delta^2] \leq \mathbb{E}[\delta^2]$. Finally, if $\mu_\delta \equiv 0$ then both terms of R_{shift} vanish; this holds in particular for the Bayes-optimal pair, for which $\mu_\delta = \mathbb{E}[S \mid \mathcal{H}] = 0$. □

B.3 Results of Sections 3 and 4: non-identifiability and concentration

Throughout this subsection, $m(x, g) = \mathbb{E}[(P - Y)^2 \mid X = x, G = g]$ is the observable conditional mean loss (1), (m_0, L) an operational decomposition satisfying Assumption 1, and the *passive law* is the joint distribution of (P, Y, X, G).

B.3.1 The alignment lemma

The following lemma, referenced in Remark 1, bridges the two halves of the theory: under its hypotheses, the counterfactual inflation of Appendix B.1 and the operational estimand of Section 2 are the same number.

Lemma 12 (Alignment of counterfactual and operational estimands)**.** *Suppose that, on the same pre-cutoff question distribution,* $\mathbb{E}[(\hat{y}_{\text{clean}} - Y)^2 \mid X, G] = m_0(X, G)$ *almost surely on* $\{G < 0\}$*. Then* $B_{\text{twin}} = B$*.*

The deployed model's side needs no hypothesis, since $\mathbb{E}[(\hat{y}_{\text{corrupt}} - Y)^2 \mid X, G] = m(X, G)$ holds by definition of m ($\hat{y}_{\text{corrupt}} = P$, Section 2). The one substantive condition equates the counterfactual clean model's risk with the honest surface m_0, plausible when the pair shares the training algorithm and protocol of Appendix B.1.

Proof. The identity $\mathbb{E}[(\hat{y}_{\text{corrupt}} - Y)^2 \mid X, G] = m(X, G)$ holds by definition of the observable surface (1), since $\hat{y}_{\text{corrupt}} = P$ under the probabilistic protocol of Section 2; the hypothesis supplies the matching identity $\mathbb{E}[(\hat{y}_{\text{clean}} - Y)^2 \mid X, G] = m_0(X, G)$ for the clean model on $\{G < 0\}$. Taking expectations over (X, G) under the pre-cutoff law and applying iterated expectations,

$$\begin{aligned} B_{\text{twin}} &= \mathbb{E}\big[(\hat{y}_{\text{clean}} - Y)^2 - (\hat{y}_{\text{corrupt}} - Y)^2 \mid G < 0\big] = \mathbb{E}[m_0(X, G) - m(X, G) \mid G < 0] \\ &= \mathbb{E}[L(X, G) \mid G < 0] = B, \end{aligned}$$

where the third equality is the operational decomposition $m = m_0 - L$ (Section 3). □

B.3.2 Proof of Theorem 2

Theorem 2 is the paper's central negative result: passive backtest data determine the observable surface m but nothing more, so the inflation B ranges over a sharp interval.

Theorem 2 (Sharp partial identification from passive scores)**.** *Under Assumption 1 alone,* B *is not a functional of the law of* (P, Y, X, G)*: every continuation* $\tilde{m}_0 : \mathcal{X} \times \mathbb{R} \to [0, 1]$ *with* $m \le \tilde{m}_0 \le 1$ *on* $\{g < 0\}$ *and* $\tilde{m}_0 = m$ *on* $\{g \ge 0\}$ *defines an admissible pair* $(\tilde{m}_0,\ \tilde{L} = \tilde{m}_0 - m)$ *consistent with the same passive law, with corresponding inflation* $\tilde{B} = \mathbb{E}[\tilde{m}_0 - m \mid G < 0]$*. The sharp identified set is*

$$\big[\,0,\ \ \mathbb{E}[\,1 - m \mid G < 0\,]\,\big],$$

and every value in this interval is attainable.

Proof. The argument has three parts: observational equivalence, attainability, and sharpness.

Observational equivalence. The passive law of (P, Y, X, G) determines the conditional mean loss m through (1), but neither m_0 nor L separately: these are not functions of the observables, and any admissible pair $(\tilde{m}_0, \tilde{L})$ with $\tilde{m}_0 - \tilde{L} = m$ is consistent with the same passive law. Consequently, if two admissible pairs yield different values of $B = \mathbb{E}[\tilde{m}_0 - m \mid G < 0]$, no functional of the passive law can equal B for both, and B is not identified.

Attainability. Fix $t \in [0, 1]$ and define

$$\tilde{m}_{0,t} = m + t(1 - m) \text{ on } \{g < 0\}, \qquad \tilde{m}_{0,t} = m \text{ on } \{g \ge 0\}.$$

Since $m \in [0, 1]$, we have $m \le \tilde{m}_{0,t} \le 1$, so $\tilde{m}_{0,t}$ is a valid risk surface; the induced $\tilde{L}_t = t(1 - m) \ge 0$ on $\{g < 0\}$ and $\tilde{L}_t = 0$ on $\{g \ge 0\}$, so Assumption 1 holds. The corresponding inflation is $\tilde{B}_t = t\,\mathbb{E}[1 - m \mid G < 0]$, which is continuous and increasing in t and traces every value in $[0, \mathbb{E}[1 - m \mid G < 0]]$ as t ranges over $[0, 1]$.

Sharpness. Conversely, any admissible pair satisfies, pointwise on $\{g < 0\}$, $0 \le \tilde{L} = \tilde{m}_0 - m \le 1 - m$, the upper bound because $\tilde{m}_0 \le 1$; taking expectations under the pre-cutoff law places $\tilde{B}$ in the stated interval. The identified set is therefore exactly $[0, \mathbb{E}[1 - m \mid G < 0]]$, and additional passive draws refine the estimate of m without shrinking it. □

B.3.3 Proof of Proposition 3

Proposition 3 converts the convex-pull model into the stakes×extraction law: leakage per question is the honest Brier times the double-benefit factor.

Proposition 3 (Per-question and aggregate leakage)**.** *Under Assumption 2, the per-question leakage is* $b_0(q)\,w(q)\,(2-w(q))$*; the pair* $m_0(x,g) = \mathbb{E}[b_0 \mid X{=}x, G{=}g]$ *and* $L(x,g) = \mathbb{E}[b_0\,w(2-w) \mid X{=}x, G{=}g]$ *is an operational decomposition satisfying Assumption 1, and hence*

$$B \;=\; \mathbb{E}\big[\, b_0\, w(2-w) \,\big|\, G<0\big]. \tag{4}$$

Proof. Fix a leakage-eligible question q. Under Assumption 2 and the fixed deterministic protocol,

$$P(q) - Y(q) = (1-w(q))\,\big(P_{\text{hon}}(q) - Y(q)\big),$$

so the realized Brier loss is $(P-Y)^2 = (1-w)^2 b_0$ with $b_0 = (P_{\text{hon}} - Y)^2$, while the honest loss is b_0; the per-question saving is $b_0 - (1-w)^2 b_0 = b_0\,w(2-w)$. Taking conditional expectations given $(X,G) = (x,g)$ yields the surfaces

$$m(x,g) = \mathbb{E}\big[(1-w)^2 b_0 \mid X{=}x, G{=}g\big], \qquad m_0(x,g) = \mathbb{E}\big[b_0 \mid X{=}x, G{=}g\big],$$

the latter being the loss the model would incur using only legitimate information, whence $L(x,g) = m_0(x,g) - m(x,g) = \mathbb{E}[b_0\,w(2-w) \mid X{=}x, G{=}g]$. This pair satisfies Assumption 1: $w \in [0,1]$ implies $L \geq 0$, and post-cutoff no leaked information exists, so $P = P_{\text{hon}}$ and $L = 0$ there. Finally, iterated expectations and (2) give

$$B = \mathbb{E}\big[L(X,G) \mid G<0\big] = \mathbb{E}\big[b_0\,w(2-w) \mid G<0\big],$$

which is (4). □

B.3.4 The loading correspondence

The following lemma, referenced in Appendix B.1, connects the per-question extraction weight w to the population loading λ of Appendix B.1 in the homogeneous special case.

Lemma 13 (Loading correspondence in the homogeneous pull model)**.** *Suppose within a conditioning stratum that* $P_{\text{hon}} = \hat{y}_{\text{clean}}$ *and* $w(q) \equiv w$ *is constant. Define the pull direction* $S_{\text{pull}} = Y - P_{\text{hon}}$*. Then* $\delta = P - P_{\text{hon}} = wS_{\text{pull}}$ *and the* L^2 *loading of* δ *onto* S_{pull} *is* w*. If, additionally,* S_{pull} *equals the Bayesian leakage signal* S *of Appendix B.1, then* $\lambda = w$ *and* $\mu_\delta = \tilde{\nu} = 0$*.*

Proof. Within the stratum, Assumption 2 with $P_{\text{hon}} = \hat{y}_{\text{clean}}$ and constant w gives

$$\delta = P - \hat{y}_{\text{clean}} = w\,\big(Y - P_{\text{hon}}\big) = w\,S_{\text{pull}},$$

so δ is proportional to S_{pull} with coefficient w: the L^2 loading of δ onto S_{pull} is w and the projection residual vanishes. If in addition $S_{\text{pull}} = S$, then $\delta = wS$, whence $\lambda = \text{Cov}(\delta, S \mid \mathcal{H})/\text{Var}(S \mid \mathcal{H}) = w$, $\mu_\delta = \mathbb{E}[\delta \mid \mathcal{H}] = w\,\mathbb{E}[S \mid \mathcal{H}] = 0$, and $\tilde{\nu} = \delta - \lambda S - \mu_\delta = 0$. □

B.3.5 Interpreting the extraction weight

A useful, non-unique parameterization of the extraction weight in Assumption 2 is

$$w(q) \;=\; \underbrace{c(q)}_{\substack{\text{knowledge } \textit{exists}}} \;\cdot\; \underbrace{u(q)}_{\substack{\textit{usable}/\text{deployed}}} \;\cdot\; \underbrace{\kappa(\text{type})}_{\substack{\text{answer vs. evidence}}}. \tag{9}$$

The factorization is interpretive rather than identified from backtest scores. Existence c asks whether the model absorbed the relevant fact at all, and is probeable leak-free by self-consistency; usability u captures the well-documented gap between possessing a fact and deploying it; and κ encodes the leakage type of Section 4, near 1 for answer leakage and intermediate for evidence leakage. The controlled contamination experiment (E2, Appendix D.3) uses this reading: inserted duplicates raise c, and the measured dose–response tracks exposure that is both absorbed and usable.

B.4 Results of Section 5: identification routes

Throughout this subsection, (m_0, L) is an operational decomposition satisfying Assumption 1 for the model under audit, with superscripts $(m^A,\, m_0^A,\, L^A)$ distinguishing models where needed. For the DiD route, fix the target pre-cutoff covariate law Q and write, for a model A,

$$\mu_\pm^A = \int \mathbb{E}\big[m^A(X,G) \mid X = x,\; G \gtrless 0\big]\, dQ(x),$$

with $\mu^A_{0,\pm}$ the analogous averages of the honest surface m^A_0; standardizing all four cells to Q removes composition differences, and $\Delta^A = \mu^A_+ - \mu^A_-$ is the standardized pre-to-post change of Proposition 5. In this notation, the formal content of Assumption 4 is: $L^{M_0} \equiv 0$ on the evaluation support; target-post and control pre/post observations have common support with Q; and the honest changes are parallel, $\mu^M_{0,+} - \mu^M_{0,-} = \mu^{M_0}_{0,+} - \mu^{M_0}_{0,-}$. The regression-discontinuity statements assume the following support regularity.

Assumption 5 (RD support regularity)**.** *The density of G is positive and continuous near zero, and the conditional support of X overlaps on both sides of the cutoff. Conditional limits are taken on this common support.*

B.4.1 Proof of Theorem 4

Theorem 4 is the boundary guarantee of Route 1: continuity of the honest surface converts the observed risk jump at the cutoff into the left-limit leakage.

Theorem 4 (RD identifies boundary leakage)**.** *Under Assumptions 1 and 3 and the support regularity of Appendix B.4, for any bounded loss, the boundary leakage is identified by the upward jump in observed risk at the cutoff,*

$$J(x) \;=\; \lim_{g\downarrow 0} m(x,g) - \lim_{g\uparrow 0} m(x,g) \;=\; L(x,0^-) \;\geq\; 0.$$

Proof. For $g > 0$, $L \equiv 0$ by Assumption 1, so $m = m_0$ there, and continuity (Assumption 3) gives $\lim_{g\downarrow 0} m(x,g) = m_0(x,0)$ on the common support of Assumption 5. For $g < 0$, $m = m_0 - L$, so $\lim_{g\uparrow 0} m(x,g) = m_0(x,0) - L(x,0^-)$, the left limit existing by Assumption 3. Subtracting the two limits gives $J(x) = L(x,0^-)$, which is non-negative by Assumption 1. □

B.4.2 Proof of Proposition 14

Proposition 14, stated here in full, extends Route 1 from the boundary to the global estimand, at the price of the extrapolation class.

Assumption 6 (Uniquely extrapolable honest surface)**.** $m_0(x,g) = \alpha(x) + \varphi(g)$*, where φ belongs to a known finite-dimensional class Φ whose restriction to the observed clean support uniquely determines it on the target pre-cutoff support (e.g. a fixed-degree polynomial), and the class is correctly specified.*

Proposition 14 (Global leakage by clean-side extrapolation)**.** *Under the assumptions of Theorem 4 and Assumption 6, the honest surface on $\{g < 0\}$ is identified by extrapolating the fit of m_0 from $\{g > 0\}$, and $B = \mathbb{E}[\, m_0(X,G) - m(X,G) \mid G < 0\,]$ is identified.*

Proof. On $\{g > 0\}$, $m = m_0 = \alpha(x) + \varphi(g)$ is observed; the two components are identified only up to an additive normalization, but their sum—the clean surface—is identified. By Assumption 6, the restriction of φ to the observed clean support uniquely determines its continuation to the target pre-cutoff support, and hence determines m_0 there. Consequently $L = m_0 - m$ is identified on $\{g < 0\}$, and so is $B = \mathbb{E}[L \mid G < 0]$. □

B.4.3 Proof of Proposition 5

Proposition 5 is the global guarantee of Route 2: a clean, recency-matched control converts the difference of pre/post changes into the standardized inflation.

Proposition 5 (DiD identification)**.** *Under Assumptions 1 and 4, standardize both models' risks to the target's pre-cutoff question mix and let Δ^A denote model A's resulting pre-to-post change. Then the inflation* (2) *is identified: $B = \Delta^M - \Delta^{M_0}$.*

Proof. Target leakage vanishes on $\{G > 0\}$ by Assumption 1, so $\mu^M_+ = \mu^M_{0,+}$. On the pre side, $m^M = m^M_0 - L^M$ integrates to $\mu^M_- = \mu^M_{0,-} - \int \mathbb{E}[L^M \mid X{=}x, G{<}0]\, dQ(x)$, and because Q is the target pre-cutoff covariate law, the integral equals $\mathbb{E}[L \mid G < 0] = B$ by iterated expectations. Hence $\Delta^M = \big(\mu^M_{0,+} - \mu^M_{0,-}\big) + B$. The control is clean ($L^{M_0} \equiv 0$), so $\Delta^{M_0} = \mu^{M_0}_{0,+} - \mu^{M_0}_{0,-}$. Subtracting and applying the parallel honest change of Assumption 4 cancels the honest terms, leaving $\Delta^M - \Delta^{M_0} = B$. □

B.4.4 Boundary DiD with a no-discontinuity control

The following variant, stated and proved only here, weakens the clean-control requirement to continuity of the control's leakage at the target cutoff; it identifies the boundary estimand only, and underwrites the control checks of M4 (Section 7).

Proposition 15 (Boundary DiD with a no-discontinuity control)**.** *At the target cutoff, define the observed risk jump* $J^A_{\text{obs}}(x) = m^A(x, 0^+) - m^A(x, 0^-)$*. Suppose (i) target leakage is zero on the post side and has left limit* $L^M(x, 0^-)$*; (ii) control leakage is continuous at the target cutoff; and (iii) target and control share the same honest/composition jump,* $m^M_0(x, 0^+) - m^M_0(x, 0^-) = m^{M_0}_0(x, 0^+) - m^{M_0}_0(x, 0^-)$*. Then*

$$J^M_{\text{obs}}(x) - J^{M_0}_{\text{obs}}(x) = L^M(x, 0^-).$$

Proof. Write J^A_0 for the honest jump of model A appearing in condition (iii). For the target, condition (i) gives $J^M_{\text{obs}} = J^M_0 + L^M(x, 0^-)$: post-cutoff leakage vanishes while the pre-side limit contributes $L^M(x, 0^-)$. For the control, condition (ii) makes its leakage difference across the cutoff vanish, so $J^{M_0}_{\text{obs}} = J^{M_0}_0$. Condition (iii) equates the honest jumps, and differencing leaves $L^M(x, 0^-)$. □

B.4.5 Proof of Proposition 16

Proposition 16, stated here in full, gives the population signature of Route 3: under calibration and homogeneous extraction, the residual covariance is the hump-shaped function of the extraction weight that makes PRC a detector of evidence leakage. It rests on the following population conditions.

Assumption 7 (PRC population conditions)**.** *Within the pre-cutoff population,* $P_{\text{hon}} = c_0$*,* $\mathbb{E}[Y \mid c_0, G < 0] = c_0$*, and the extraction weight is homogeneous:* $w(q) \equiv \bar{w}$*.*

Conditional calibration is achievable by construction—recalibrate on leakage-free anchors (Remark 2)—and $P_{\text{hon}} = c_0$ takes the crowd as the honest benchmark, appropriate when the model holds no legitimate edge over public information. Homogeneous w is a first-order simplification: under heterogeneous w the covariance is a nonlinear functional of the joint law of (w, c_0, Y), not of B, and the endpoint cases below show it cannot separate pure answer leakage ($w \equiv 1$) from no leakage at all. In the population result, P denotes the framing-invariant limit P_∞ of the paraphrase consensus; the finite-K deployment, including the framing-invariance conditions, is developed in Appendix C.

Proposition 16 (Residual-covariance detection under convex pull)**.** *Under Assumptions 2 and 7,*

$$\text{Cov}(P, Y - P \mid G < 0) = \bar{w}(1 - \bar{w})\,\mathbb{E}[c_0(1 - c_0) \mid G < 0] \;\geq\; 0,$$

and the covariance vanishes whenever $w \equiv 0$ *or* $w \equiv 1$*.*

Proof. All moments are conditional on $G < 0$. Under Assumption 2 with homogeneous weight $\bar{w}$ and $P_{\text{hon}} = c_0$, we have $P = (1-\bar{w})c_0 + \bar{w}Y$ and $Y - P = (1-\bar{w})(Y - c_0)$. Conditional calibration ($\mathbb{E}[Y \mid c_0, G < 0] = c_0$) gives $\text{Cov}(c_0, Y - c_0) = 0$ and, for binary Y, $\text{Cov}(Y, Y - c_0) = \mathbb{E}[\text{Var}(Y \mid c_0)] = \mathbb{E}[c_0(1 - c_0)]$. Hence

$$\text{Cov}(P, Y - P) = (1 - \bar{w})\,\text{Cov}\big((1 - \bar{w})c_0 + \bar{w}Y,\; Y - c_0\big) = \bar{w}(1 - \bar{w})\,\mathbb{E}[c_0(1 - c_0)] \;\geq\; 0.$$

If $w \equiv 0$ then $P = P_{\text{hon}} = c_0$ and the covariance equals $\text{Cov}(c_0, Y - c_0) = 0$; if $w \equiv 1$ then $P = Y$, so $Y - P = 0$ and the covariance vanishes as well. □

Remark 2 (Calibration is mandatory)**.** *Proposition 16 presumes a calibrated honest forecast. Real LLMs are markedly overconfident (on our data, fitted temperature* $T \approx 8$*; provenance in Appendix D.7), so the* raw *residual covariance is dominated by miscalibration and is in fact negative—the opposite of the leakage signature. One must recalibrate* P *on leakage-free anchors first, or use a differenced form; a positive* calibrated *value indicates evidence leakage.*

B.4.6 A sufficient external reference

The following corollary, referenced in Section 5.4, records that a single external reference restores the point identification that Theorem 2 denies to passive data.

Corollary 17 (A sufficient external reference)**.** *With only the passive backtest distribution of* (P, Y, X, G) *(including the crowd anchor* c_0*),* B *is set-identified, not point-identified (Theorem 2). Each of the following suffices to point-identify* B *under its stated assumptions:* (R1) *a known cutoff with questions on both sides (Proposition 14);* (R2) *a clean, capability-matched control on the same questions (Proposition 5). The intervention* (R3) *detects evidence leakage (Proposition 16) but does not point-identify it.*

Proof. Immediate from Theorem 2 and Propositions 5, 14 and 16. □

Remark 3 (Metric scope)**.** *The route results use only the structure* $m = m_0 - L$ *and hold for any bounded loss—Brier,* $1-$*accuracy, or* $1 -$ *pass@1—which licenses the pass@1 correction of Section 7; the law* $L = b_0\, w(2-w)$ *and the* PRC *statistic are Brier-specific.*

B.4.7 Complementary signatures

The following proposition, referenced in Section 5.4, combines the boundary and residual signatures into the complementarity statement that motivates pairing the probes. We make no formal claim about a consistent test; a sampling-level analysis would require thresholds, finite-sample noise models, and the availability of both probes.

Proposition 18 (Complementary signatures under homogeneous extraction)**.** *Under Assumptions 1 to 3 and 7 and the support conditions of Appendix B.4, suppose the stakes and homogeneous extraction have left limits* b_0 *and* w *at the cutoff, and let* $\sigma_S^2 = \mathbb{E}[c_0(1-c_0) \mid G < 0]$*. Then*

$$J(w) = b_0\, w(2-w) \quad (\textit{increasing; maximal at } w = 1),$$
$$\mathrm{PRC}(w) = w(1-w)\sigma_S^2 \quad (\textit{maximal at } w = \tfrac{1}{2};\ \textit{zero at } w \in \{0,1\}).$$

Proof. By Theorem 4 and Proposition 3, the boundary jump equals the left-limit leakage, $J = L(\cdot, 0^-) = b_0\, w(2-w)$, which is increasing in w on $[0,1]$ and maximal at $w = 1$. By Proposition 16 in the homogeneous case, $\mathrm{PRC} = w(1-w)\sigma_S^2$, maximal at $w = \frac{1}{2}$ and zero at $w \in \{0,1\}$. In particular $\mathrm{PRC}(1) = 0$ while $J(1) = b_0 > 0$, and $\mathrm{PRC}(\frac{1}{2}) = \sigma_S^2/4 > 0$. □

The complementarity is verified on the synthetic sweep of Appendix E.1; the answer-leakage regime $w \to 1$ does not arise in the ground-truth experiment (measured $w \leq 0.42$, Section 6.1), so its real-data face is untested.

C Deploying the routes

Figure 6 summarizes the practitioner's choice of route given the available resources; it is the decision-oriented complement of Table 1, which states what each route guarantees and how its assumption is checked.

C.1 The practitioner's recipe and control validation

The practitioner's recipe. **(1)** Fix the evaluation metric and a leakage-free difficulty adjustment; for probabilistic forecasts, a contemporaneous crowd forecast provides the stakes anchor. **(2)** Estimate $\widehat{B}$ by a point-identifying route: DiD requires a capability/recency-matched clean control (selection and validation protocol below), while RD identifies the boundary jump and needs the extrapolation assumption for global B. **(3)** If using PRC, first calibrate on leakage-free anchors and report it only as a detection statistic (Remark 2 and Proposition 18); calibration is not required for accuracy- or pass@1-based RD/DiD contrasts. **(4)** When global B is identified, report $\mathrm{Risk}_{\mathrm{meas}} + \widehat{B}$ (or the score form of Corollary 6) with a confidence interval; otherwise report the boundary or detection estimand without converting it into a global correction. **(5)** Prefer *forward* evaluation when feasible, for which $B = 0$ by construction; and if the cutoff is uncertain or staged, report sensitivity to its placement rather than treating a single date as known.

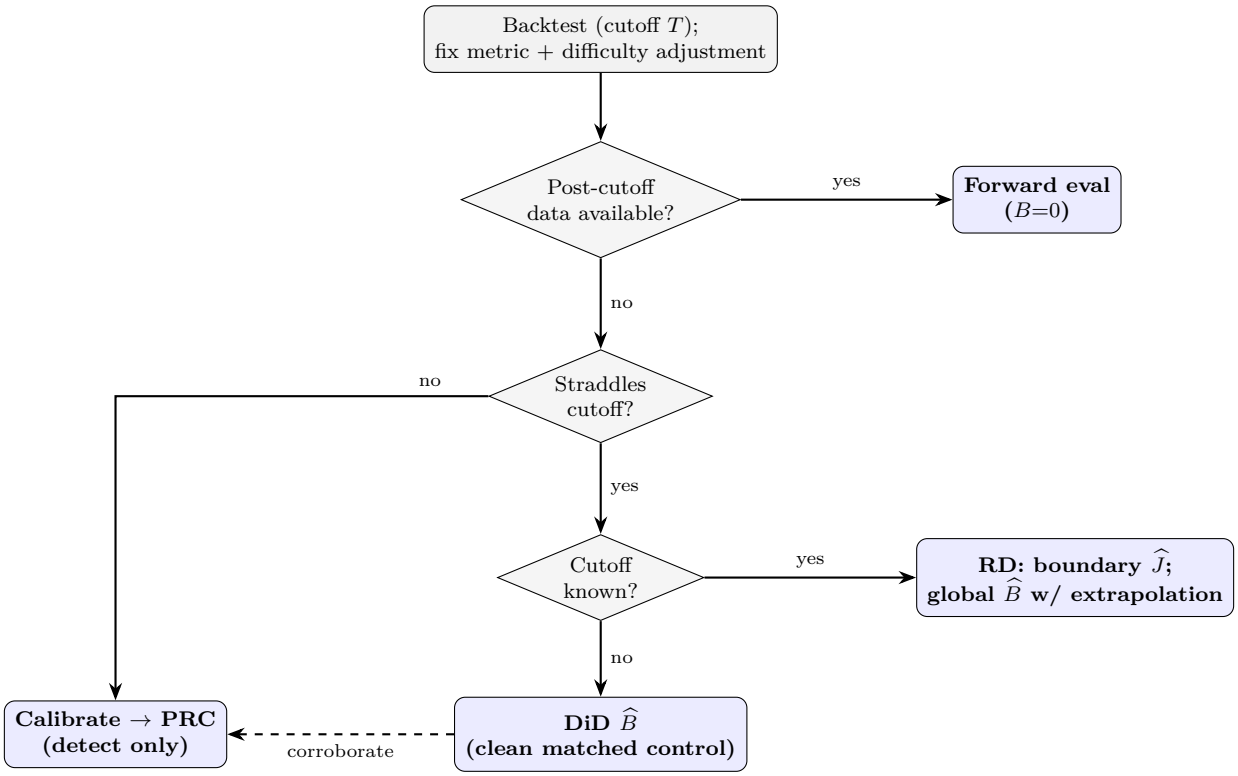

Figure 6: **Choosing a route: report a leakage-adjusted score only when global B is identified; otherwise report the boundary or detection estimand.** Forward evaluation ($B = 0$) is preferred when post-cutoff outcomes can be awaited. Otherwise, RD identifies the boundary jump when the cutoff is known (and global B under extrapolation), and DiD identifies global B with a clean, capability/recency-matched control; if no data straddle the cutoff, only the calibrated PRC detection statistic remains. The flow is a simplification: with both a known cutoff and a matched control available, DiD may be preferred for global B. Table 1 states each route's assumptions and checks.

Selecting and validating a DiD control (Route 2). The control should be a same-family snapshot or a vintage-matched peer of the target—same capability tier and generation, similar training-data distribution—whose documented cutoff predates the evaluation window; its cleanliness should be verified empirically rather than assumed, for example by date-only recall probes on the window's questions. The parallel-change condition of Assumption 4 is not fully testable, since the target's pre-cutoff cell is contaminated by construction, but it has three testable implications: on the jointly clean region $\{G > 0\}$ both models' risk trends are honest and should move in parallel; placebo cutoffs at dates with no leakage discontinuity should return estimates near zero; and covariate balance should hold across the four standardized cells. When several candidate controls are available, agreement of $\widehat{B}$ across them is an over-identification check, and their spread is a systematic-uncertainty band. Finally, the dominant violation is signed (Section 5.2): a weak control inflates $\widehat{B}$, so a null finding is conservative under this failure mode, whereas a positive finding should be reported together with the matching checks. M5 (Appendix D.7) implements this protocol.

C.2 Finite-K PRC machinery (Route 3)

The deployable form of the paraphrase route (Section 5.3) uses a finite number of queries. Query the model with K paraphrases per question, form the consensus $\bar{y}_{K,i} = K^{-1} \sum_k P_i^{(k)}$, and compute $\widehat{\mathrm{Cov}}(\bar{y}_K, Y - \bar{y}_K)$. The target population quantity is $\mathrm{PRC}_\infty = \mathrm{Cov}(P_\infty, Y - P_\infty \mid G < 0)$ from Proposition 16, where P_∞ is the framing-invariant consensus defined by the following condition (moved here from the main text because only the finite-K analysis uses it).

Assumption 8 (Framing-invariant paraphrase noise)**.** *Within the pre-cutoff population, for paraphrase π_k, write $P^{(k)} = P_\infty + \varepsilon_k$, where the leaked signal and P_∞ do not depend on framing. Conditional on the question, the deviations are identically distributed, mean zero, mutually uncorrelated, and uncorrelated with P_∞ and $Y - P_\infty$. They have finite fourth moments.*

Lemma 19 (Finite-paraphrase attenuation)**.** *Under Assumption 8, let $W = \mathbb{E}[\mathrm{Var}_\pi(P^{(k)} \mid q, G < 0) \mid G < 0]$. For constant K,*

$$\mathrm{Cov}(\bar{y}_K, Y - \bar{y}_K \mid G < 0) = \mathrm{PRC}_\infty - \frac{W}{K}.$$

Proof. Write $\bar{y}_K = P_\infty + \bar{\varepsilon}_K$. Conditional mean-zero and mutual uncorrelatedness give $\mathrm{Var}(\bar{\varepsilon}_K \mid q) = \mathrm{Var}(\varepsilon_k \mid q)/K$. The remaining orthogonality conditions in Assumption 8 give

$$\mathrm{Cov}(P_\infty + \bar{\varepsilon}_K,\, Y - P_\infty - \bar{\varepsilon}_K) = \mathrm{Cov}(P_\infty, Y - P_\infty) - \mathrm{Var}(\bar{\varepsilon}_K).$$

Taking the question-level expectation yields the result. □

Estimator and inference. Let

$$\widehat{W} = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{K-1}\sum_{k=1}^{K} (P_i^{(k)} - \bar{y}_{K,i})^2.$$

The bias-corrected estimator is

$$\widehat{\mathrm{PRC}}_{\mathrm{bc}} = \widehat{\mathrm{Cov}}(\bar{y}_K, Y - \bar{y}_K) + \widehat{W}/K.$$

For varying K_i, replace $\widehat{W}/K$ by the average of the per-question sample variances divided by K_i. Under independent questions, Assumption 8, and finite fourth moments, the vector of sample moments defining $\widehat{\mathrm{PRC}}_{\mathrm{bc}}$ obeys a multivariate central limit theorem, and the delta method yields $\sqrt{m}(\widehat{\mathrm{PRC}}_{\mathrm{bc}} - \mathrm{PRC}_\infty) \Rightarrow \mathcal{N}(0, \sigma^2_{\mathrm{PRC}})$; in practice we bootstrap questions and refit any calibrator inside each replicate. PRC remains a detection statistic, not an estimator of B (Proposition 16), and paraphrase diversity must come from framing rather than sampling temperature (Appendix C.3).

C.3 Paraphrase set and generation protocol

Route 3 (Section 5) requires paraphrases that vary framing without revealing the outcome; the set documented here illustrates the design principles and the generation protocol.

Design principles. Effective paraphrases must (i) preserve the event and resolution criteria (same answer), (ii) maximize variation in the reasoning channel, and (iii) reveal nothing about the outcome. We vary six dimensions independently: lexical choice, sentence form, level of specificity, entity reference, framing/tone, and which background context is emphasized.

Generation and validation. Paraphrases are generated by an auxiliary model (distinct from the model under test) instructed to preserve meaning and resolution criteria while varying the six dimensions and concealing the outcome. Candidates are validated by three checks: semantic preservation (embedding cosine similarity above a threshold), lexical diversity (low pairwise self-BLEU), and an outcome-leakage screen rejecting any phrasing that encodes the answer's direction. Querying uses temperature 0 with reasoning disabled and a strict final-answer parser; truncated or out-of-range generations are retried once and otherwise discarded. The protocol is instantiated, with the specific K and models used, in the PRC experiments of Appendices E.1 and E.8.

C.4 Classical identification tools used in the paper

The machinery instantiates standard tools. The residual covariance is a crowd-anchored analogue of Mincer–Zarnowitz efficiency tests and of Murphy's excess resolution (Mincer & Zarnowitz, 1969; Elliott & Timmermann, 2016; Patton & Timmermann, 2012; Murphy, 1973; Gneiting & Raftery, 2007). The routes are regression discontinuity (Imbens & Lemieux, 2008; Lee & Lemieux, 2010) and difference-in-differences, with

Table 4: **Claim-to-evidence contract.** Each theoretical claim of Sections 3 to 5 and the experiment whose headline result carries it (M1: Section 1; M2, M3: Section 6; M4, M5: Section 7). *: 95% CI excludes zero.

Claim	Carried by	Headline result
Recency confounds naive backtests (Theorem 2)	M1	4/5 cutoff-clean flagships show starred naive gaps (+0.044 to +0.061*)
Inflation rises with contamination dose (Theorem 8, Appendix B.1)	M2, M3	monotone in r at pretraining scale and in forecasting twins (dose-64 twin contrast +0.063*; dose-0 null)
Per-question law $L = b_0\, w(2-w)$ (Proposition 3, Assumption 2)	M3	concentration 13× across b_0 quintiles; law matched everywhere but the top quintile; measured w declines with difficulty
Change orthogonal to the signal cannot inflate (Corollary 10, Appendix B.1)	M3	outcome-scrubbed control twin gains recency but shows no inflation (naive +0.001; RD +0.012, no jump)
RD detects boundary leakage; recency alone gives no jump (Theorem 4)	M3, M4	twin RD +0.052 vs control +0.012, placebo null; wild diagonals starred on code (+0.090*, +0.095*) and via the clean anchor (+0.106*)
DiD with a matched control returns disciplined nulls (Proposition 5, Assumption 4)	M5	five targets null after adjustment; weak control re-inflates to +0.055; power 0.97 at effect 0.05
PRC detects evidence leakage, gated on calibration (Proposition 16, Appendix B.4.5)	M3	differenced PRC +0.0069* only on injected cells; raw PRC negative everywhere

a fixed-question intervention in the spirit of control functions and specification tests (Wooldridge, 2015; Wu, 1973; Hausman, 1978). The sharp-bounds framing of Theorem 2 follows partial identification (Manski, 2003; Imbens & Manski, 2004). The finite-K correction is classical errors-in-variables disattenuation (Fuller, 1987; Spearman, 1910). Two superficially appealing instruments do not transfer and serve only as motivation: temporal-head ablation removes legitimate temporal reasoning along with leakage, and item-preknowledge item response theory requires cross-examinee variation that a single deployed model lacks.

D Experiment configurations and full results

Each subsection below gives, for one main-text experiment (M1–M5, in order), the full configuration—data provenance, model inventories, estimator definitions, and inference procedures—together with the complete results summarized in Section 7. The synthetic validation study E1, the per-experiment robustness analyses, and all supplementary arms are collected in Appendix E. Table 4 is the map: each theoretical claim of Sections 3 to 5 and the experiment whose headline result carries it.

D.1 Scope and coverage

Table 5 summarizes, for each experiment, the data source, sample size, models, metric, and the machinery that carries statistical significance. It is the single reference for what each reported interval or p-value is computed from.

D.2 The forecasting panel and M1 (clean-flagship naive check)

Panel construction. The panel is built from the public ForecastBench archive (Karger et al., 2025): the nightly question banks and the resolution files published in the project's datasets repository. We retain market-source binary questions (Polymarket, Metaculus, Manifold, INFER) that are resolved, have a valid frozen market probability $c_0 \in (0, 1)$ recorded before resolution, and have a resolution date between July 2024 and July 2026. Combination questions (logical conjunctions of base questions) are excluded. After deduplication by question identifier this yields 1,646 unique questions. Dataset-source questions (for example weather and economic series) are excluded from the headline analyses because their answers are tabular rather

Table 5: **Scope and coverage of the empirical program.** One row per experiment; the last column names the inference machinery behind every significance claim in Section 7.

Experiment	Data source	n	Models	Metric	Inference
M1	ForecastBench market panel	332–767 post-cutoff questions/model	GPT-5, Gemini-3.1-Pro, Kimi-K2.6, GPT-5.5, MiniMax-M3	anchored Brier reduction	source×month cluster bootstrap
M2	Hubble published accuracies	4 tasks × 6 doses	Hubble 1B/8B (perturbed, standard)	accuracy	exact permutation trend test; SEM-propagated intervals
M3	ForecastBench panel + Tinker LoRA twins	988 injected-pool + 543 post-cutoff questions	Qwen3.5-35B-A3B-Base twins (trained here)	anchored Brier reduction; log-prob	item bootstrap; paired bootstrap; placebo boundaries
M4-F	ForecastBench panel + archived real-time forecasts	1,646 questions; 162–499 anchored	DeepSeek-V3.1, Kimi-K2.6, GPT-5.4, GPT-5.5; controls MiniMax-M3, Claude-Opus-4.7; family anchors GPT-5-Mini, GPT-5.1, GPT-4.1, o4-mini	anchored Brier reduction	cluster bootstrap; diagonal permutation test; horizon banding
M4-C	LiveCodeBench per-problem pass@1	401–437 problems per window	GPT-4o-0806, Claude-3.5-Sonnet (published); GPT-5, GPT-5-Mini, DeepSeek-V3-0324, Kimi-K2-0905 (generated); controls: 2025-cutoff pool	pass@1	problem bootstrap (4000); date-permutation test
M5	ForecastBench market panel	265 pre- / 1,207 post-boundary paired questions	Qwen3.5-35B-A3B (target), GPT-5 (matched control), Kimi-K2.6, DeepSeek-V3.2, GLM-4.7, MiniMax-M3, GPT-3.5-Turbo (weak)	anchored Brier reduction	cluster bootstrap; boundary sweep; semi-synthetic power
E1 (app.)	synthetic, known DGP	400–8000/rep; 200–300 reps	simulated forecasters	Brier reduction	bootstrap CIs; MC coverage

than event outcomes; they are retained in the archive audit. The panel, the raw downloads, and the build script are versioned with the code.

Elicitation protocol. Every model is queried once per question through OpenRouter with temperature 0, reasoning disabled where the endpoint permits, and a fixed prompt that states the question, its resolution criteria, and the resolution date, and requests a single probability. Responses are parsed to a float; refusals and parse failures are dropped and reported per model (the retained fraction appears in each experiment's sample sizes). Scores are crowd-anchored Brier reductions $L(q) = (c_0 - Y)^2 - (P - Y)^2$.

M1 design. For each of the five flagships (GPT-5, cutoff 2024-09-30; Gemini-3.1-Pro, 2025-01-31; Kimi-K2.6, 2025-04-30; GPT-5.5, 2025-12-01; MiniMax-M3, 2026-01-31) we keep questions resolving at least 30 days after the documented cutoff, so that no outcome can be in training data. Cutoffs are taken from vendor documentation collected in the pre-registration; Kimi-K2.6's cutoff is documented at coarser (tier-two) resolution, a fact shown not to be load-bearing by the cutoff-shift check of Appendix D.5. The naive

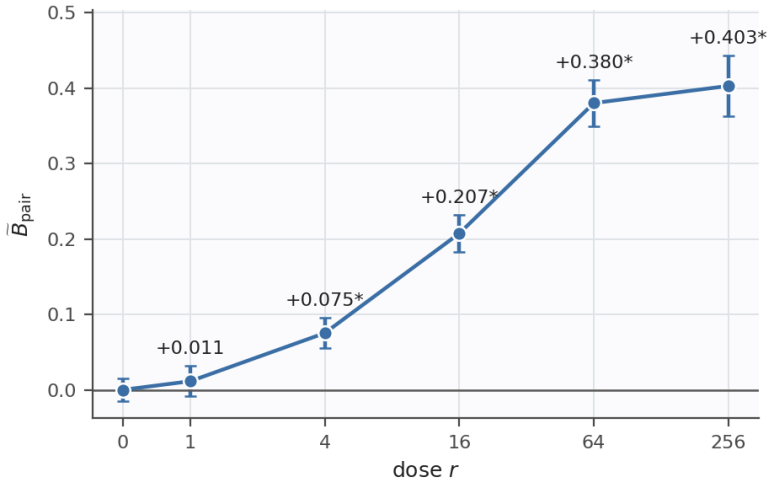

Figure 7: **M2: at pretraining scale, measured accuracy inflates strictly monotonically with the known contamination dose.** Hubble pretrains a *perturbed* model with benchmark documents inserted r times and a *standard* twin trained identically without them. The x-axis is the dose r (log scale). The y-axis is the placebo-centered contrast $\widetilde{B}_{\text{pair}}(r) = \widehat{B}_{\text{pair}}(r) - \widehat{B}_{\text{pair}}(0)$, where $\widehat{B}_{\text{pair}}(r)$ is the perturbed-minus-standard accuracy on items inserted at dose r; $\widetilde{B}_{\text{pair}} > 0$ means contamination inflates measured accuracy, and the $r = 0$ point is the centering reference (zero by construction). Whiskers: 95% intervals propagated from published SEMs; *: interval excludes zero. Result: strictly monotone from +0.011 at $r = 1$ to +0.403* at $r = 256$; sixteen duplicates already inflate accuracy by ≈21 points.

statistic is the pre/post gap in mean L around the median resolution date of each model's clean window. Confidence intervals use a cluster bootstrap over source×month cells (4,000 resamples).

M1 headline results (demoted detail). Four of the five clean flagships show starred naive gaps of +0.044 to +0.061 (Figure 1): an auditor running the standard pre/post check would flag four model families as leaking on questions they cannot possibly have seen. The fifth model is the exception that proves the mechanism. MiniMax-M3's null (+0.017) is a property of its *window*, not of the model: restricting all five models to its short clean window (March through July 2026) collapses every gap to insignificance—GPT-5.5 falls from +0.061* to +0.008, and MiniMax's gap becomes the largest of the five. The naive statistic therefore measures how much recency contrast an evaluation window spans, not whether the audited model leaked. This is the empirical face of Theorem 2: a passive backtest gap carries recency and leakage in a single number, and at roughly 0.05 Brier-reduction units in every family we test, the confound is far too large to ignore.

Matched-window check. Because each model's clean window starts at its own cutoff, gap sizes are not comparable across models: a longer window spans more recency contrast between its pre and post halves. Restricting all five models to the shortest clean window (MiniMax-M3's: resolutions from 2026-03-02, split at the common median 2026-05-14, $n \approx 333/337$) makes every gap statistically null and nearly equal: GPT-5 +0.001 $[-0.053, +0.042]$, Gemini-3.1-Pro +0.005 $[-0.055, +0.047]$, Kimi-K2.6 +0.015 $[-0.025, +0.058]$, GPT-5.5 +0.008 $[-0.034, +0.049]$, MiniMax-M3 +0.017 $[-0.020, +0.047]$. The spread across models in Figure 1 is therefore driven by the windows, not the model families, which is the recency-confound reading the theory predicts. The same check backs the cutoff-integrity reading of the documented cutoffs: had a documented cutoff materially understated a model's training data, that model's gap would not collapse alongside the others in the common window.

D.3 M2 (Hubble QA)

The Hubble suite. The Hubble suite (Wei et al., 2026) is a controlled pretraining experiment: models are trained from scratch on a web corpus into which benchmark evaluation documents (MMLU, PIQA,

Table 6: **M2 (Hubble, 8B, 500B tokens): three estimators of the dose effect agree.** Rows: the raw minimal-pair contrast $\widehat{B}_{\text{pair}}$, the placebo-centered contrast $\widetilde{B}_{\text{pair}}$ (the primary estimand, plotted in Figure 7), and the self-baseline $\widehat{B}_{\text{self}}$, by duplication dose r. Result: all three rise strictly monotonically over $r > 0$ (placebo-centered: Spearman $\rho = 1.0$, $p = 0.0083$) and agree within 0.04 at every positive dose; 95% intervals are shown in Figure 8(a).

duplication r	0	1	4	16	64	256
$\widehat{B}_{\text{pair}}$ (raw)	-0.022	-0.010	$+0.053$	$+0.186$	$+0.358$	$+0.381$
$\widetilde{B}_{\text{pair}}$ (centered)	$+0.000$	$+0.011$	$+0.075$	$+0.207$	$+0.380$	$+0.403$
$\widehat{B}_{\text{self}}$ (self-baseline)	$+0.000$	$+0.017$	$+0.065$	$+0.216$	$+0.394$	$+0.413$

HellaSwag, WinoGrande questions with answers) are deliberately inserted at *known* duplication counts $r \in \{0, 1, 4, 16, 64, 256\}$. For each configuration, two models are released: a *perturbed* model (trained on the corpus with benchmark documents inserted at rate r) and a *standard* model (trained on the same corpus *without* the benchmark insertions). The standard model is a perfect minimal-pair clean control—it shares the same architecture, hyperparameters, and training data except for the inserted documents. Model sizes are 1B and 8B parameters, trained for 100B and 500B tokens. We use the published per-(model, task, duplication rate) accuracies, requiring no additional compute. Headline results use 8B at 500B tokens, aggregated as an unweighted mean over the four tasks; per-task results are reported as a robustness check.

Estimation. Three quantities are reported:

- *Minimal-pair contrast* (raw): $\widehat{B}_{\text{pair}}(r) = \text{acc}_{\text{pert}}(r) - \text{acc}_{\text{std}}(r)$.
- *Placebo-centered contrast*: $\widetilde{B}_{\text{pair}}(r) = \widehat{B}_{\text{pair}}(r) - \widehat{B}_{\text{pair}}(0)$, which removes the small baseline offset between the two separately trained model instances ($\widehat{B}_{\text{pair}}(0) = -0.022$ for 8B-500B).
- *Self-baseline estimator* (no separate control): $\widehat{B}_{\text{self}}(r) = \text{acc}_{\text{pert}}(r) - \text{acc}_{\text{pert}}(0)$.

Trend test. The monotone dose-response is assessed by an exact one-sided Spearman permutation test over the nonzero duplication rates ($r \in \{1, 4, 16, 64, 256\}$). Robustness checks (difficulty matching, scale/token settings, per-task consistency) are in Appendix E.4.

Results. Table 6 tabulates the three estimators as a function of the known dose; Figure 8 shows the dose-response with propagated intervals, the replication across scale and token settings, per-task consistency, and the clean-control accuracy range bounding difficulty imbalance. The placebo-centered contrast is strictly monotone over $r > 0$ ($+0.011$, $+0.075$, $+0.207$, $+0.380$, $+0.403$; Spearman $\rho = 1.0$, $p = 0.0083$); the clean control's accuracy varies by at most 0.051 across duplication bins; the self-baseline tracks the minimal-pair contrast within 0.035 at every positive dose.

D.4 M3 (forecasting twins with injected leakage)

Base model and training. Both twins are LoRA continued-training runs of Qwen3.5-35B-A3B-Base on the Tinker training API: LoRA rank 32, sequence length 2048, batch 16 sequences, learning rate 2×10^{-4} with 100-step linear warmup, gradient clipping at 1.0, one epoch over the corpus in a seed-fixed order. Each full twin processes approximately 94 million tokens (2,900 steps). The corpus is Wikitext-103 filler (95 million tokens, identical for both twins) with the injected documents interleaved at deterministic schedule positions.

Injected documents. For each treated question, dated news-style documents are written by DeepSeek-V3.1 from the question metadata and the realized outcome, in rotating styles (news article, encyclopedia entry, analyst note), and validated by DeepSeek-V3.2 (the validator must answer the question correctly from the treatment document alone, and must *fail* to answer from the scrubbed control document). The control twin sees the same documents with outcome statements scrubbed; source, topic, length, and schedule

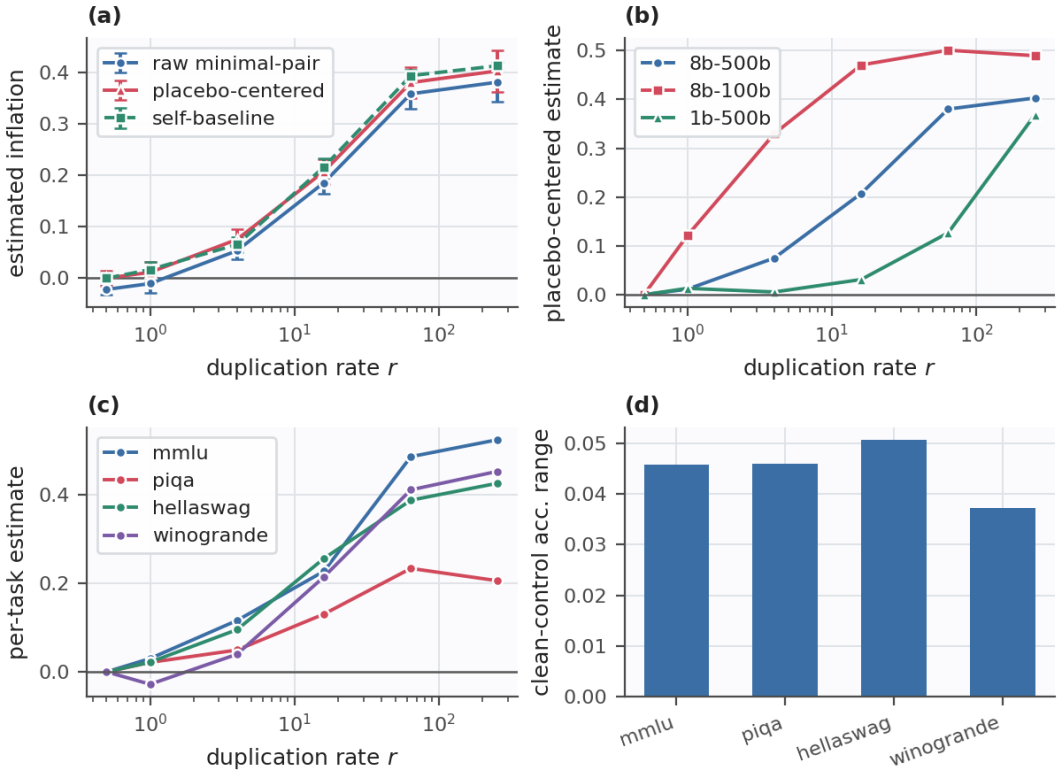

Figure 8: **M2: the Hubble dose-response is robust across estimators, scales, and tasks.** (a) Raw minimal-pair, placebo-centered, and self-baseline estimates for the 8B/500B pair (whiskers: 95% intervals propagated from published SEMs). (b) The placebo-centered estimate across model scale and token budget. (c) Per-task estimates for 8B/500B. (d) The clean control's accuracy range across dose bins, bounding difficulty imbalance. Result: inflation rises monotonically with dose in every panel; no single estimator, scale, or task drives Figure 7.

positions are matched. Dose $r \in \{0, 1, 4, 16, 64\}$ counts document copies, randomized across questions stratified by the control twin's per-question honest error tercile and by resolution month.

Screening. Injection is restricted to questions resolving before the pseudo-cutoff T^* (April 2026) on which the un-tuned base model is not already confidently correct (excluding questions with base $P(\text{YES}) > 0.85$ and outcome YES, or $P(\text{YES}) < 0.15$ and outcome NO). This directly operationalizes the design requirement that the injected documents be the marginal information source. The screen keeps 988 of 1,103 pre-T^* questions and was fixed before any twin was trained; the deviation from the coarser pre-registered proxy screen is documented in the versioned deviations file.

Probability readout. Probabilities are read from paired "Yes"/"No" continuation log-probabilities under a fixed, generic, outcome-balanced few-shot prefix, giving $P(\text{YES})$ as the softmax of the two continuation scores. The same readout is used for the base model, both pilot twins, and both full twins.

Pilot gates. Before the full runs, a pilot pair was trained on ten percent of injected questions with 10 million filler tokens and had to pass five pre-registered gates: (a) memorization (treatment log-probability advantage on injected documents), (b) signal (positive twin contrast at dose 64), (c) null (twin contrast consistent with zero at dose 0), (d) direction (treatment probability moves toward realized outcomes), and (e) recency (control twin beats the base on injected-topic questions). All five passed at the pilot and again at full scale, where the values quoted in the main text were computed.

Concentration binning variants. The binning variable is the control twin's realized per-question error b_0, the quantity appearing in the identity $L = b_0\, w(2-w)$; the main text displays quintiles, and the conclusion

is invariant to granularity. Under the pre-registered tercile binning at dose 64 the observed contrasts are +0.019, +0.055, +0.112 against homogeneous-w predictions +0.013, +0.044, +0.185: same pattern, with the shortfall confined to the top bin. Because binning on realized error could in principle inflate top-bin estimates by selecting on the minuend of the contrast, we also rebin by ex-ante crowd uncertainty $c_0(1-c_0)$, which involves no realized outcome. This too preserves the conclusion: at dose 64 the lower two terciles match the prediction (+0.068 observed vs +0.054 predicted; +0.068 vs +0.069) while the hardest tercile falls short (+0.053 observed, CI $[+0.010, +0.097]$, against +0.127 predicted); at dose 16 the hardest ex-ante tercile is null (−0.001) against a +0.125 prediction. Solving $L/b_0 = w(2-w)$ per tercile gives extraction weights 0.55, 0.44, 0.18 at dose 64 (per quintile: ≈0.5 on the lower three, 0.30, 0.16 on the top two): extraction declines with question difficulty and rises with dose.

Recovery decomposition and spillover mechanism. The boundary statistics in the main text are: treatment-twin naive pre/post gap +0.053; control-twin naive gap +0.001; local-linear RD jumps at T^* with 90-day bandwidth +0.052 (treatment) and +0.012 (control); placebo boundary at $T^* - 120$ days −0.014. The per-item ground truth, averaged over the injected pool with dose-0 differencing, is +0.024 to +0.027. The gap between the boundary estimate and the per-item truth is a base-rate spillover, measured directly: the treatment twin's mean probability of YES shifts relative to the control twin by −0.088 on injected questions, −0.096 on dose-zero questions, and −0.097 on never-injected post-T^* questions ($n = 543$)—a uniform lean matching the 83% NO rate of the injected outcomes. The lean is nearly cost-free on pre-T^* pools (YES rate 0.170–0.183, hence the dose-zero null of −0.003) and costs 0.028 Brier on the post pool (YES rate 0.355). Dose-zero differencing removes the uniform lean exactly, which is why the law analyses are unaffected. Concentrated single-style document injection plausibly exaggerates this prior shift relative to organic contamination; we flag it as a scope note for reading wild boundary estimates as totals rather than pure pre-side inflation.

This accounting also delimits the operational model of Section 3 empirically. The twins realize both estimands at once: the counterfactual per-item contrast of Appendix B.1 (+0.024) and the operational boundary quantity built on Assumption 1 (+0.052). The wedge between them is a *post-side* contamination effect (the base-rate lean costs 0.028 Brier on post-T^* questions), which the support restriction of Assumption 1 ($L = 0$ for $g \geq 0$) excludes by fiat and which perturbs the alignment of Lemma 12 by the same amount. Under concentrated injection the restriction is therefore measurably violated; under diffuse organic contamination the lean, and with it the wedge, plausibly shrinks, bringing the operational and counterfactual estimands back together.

Attenuation mechanisms the twins cannot exhibit. The spillover above biases a wild boundary jump *upward*; two mechanisms present in wild deployments bias it *downward*, and injection that is sharp at T^* produces neither. First, quasi-resolved outcomes: a question resolving shortly after a cutoff can be effectively decided in pre-cutoff text (an election with decisive polls, a lawsuit after closing arguments). The support restriction of Assumption 1 classifies such information as recency, so this is not a violation, but it raises the honest post-boundary score $m(x, 0^+)$ and shrinks the observed jump. Second, ingestion lag: text describing outcomes realized just before the cutoff has had little time to be crawled and trained on, thinning $L(x, g)$ as $g \to 0^-$ and smoothing the left limit. Both mechanisms attenuate $\widehat{J}$ toward zero, and neither is covered by the cutoff-placement sensitivity checks, which perturb the boundary's location rather than the information density around it. The reading rules of Section 6.2 therefore treat a wild jump as an upper bound only against the spillover mechanism: a starred jump survives attenuation, while a boundary null does not certify the absence of leakage sitting deeper in the pre-cutoff period.

PRC protocol. Three paraphrases per question are generated by DeepSeek-V3.1 and validated by DeepSeek-V3.2 (semantic equivalence and preserved resolution criteria); 971 of 988 questions retain all three. Both twins score all paraphrases with the same readout. Per-question consensus probabilities are temperature-calibrated per twin on dose-0 anchors, and the reported statistic is the twin-differenced residual covariance with paired question bootstrap (4,000 resamples): $+0.0069^*$ (CI $[+0.0035, +0.0107]$) on injected questions, null at dose zero (−0.0057, CI $[-0.0128, +0.0016]$), and monotone in dose with the dose-64 cell starred ($+0.0121^*$). Raw, undifferenced PRC is negative in every cell (treatment/control × injected/dose-0),

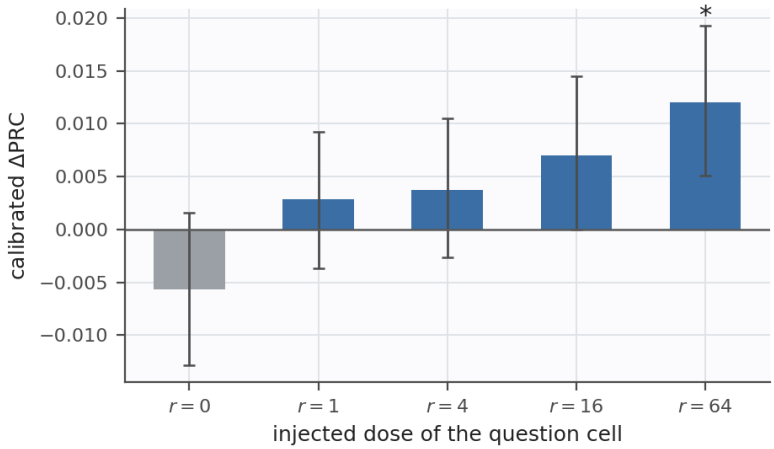

Figure 9: **M3: the calibrated, twin-differenced PRC fires only where leakage was injected.** Bars: the calibrated twin-differenced PRC per injection cell (dose zero and each positive dose), computed from three validated paraphrases per question; positive values signal evidence leakage. Whiskers: paired-bootstrap 95% CIs; *: CI excludes zero. Result: null at dose zero, monotone in dose, starred at dose 64 (+0.0121*); the raw statistic without calibration or differencing is negative everywhere, the miscalibration failure Remark 2 predicts.

reproducing the miscalibration failure predicted by Remark 2. Figure 9 shows the calibrated twin-differenced statistic by cell.

D.5 M4-F (forecasting-panel matrix and clean-anchor arm)

Specificity matrix. Targets are DeepSeek-V3.1 (documented cutoff March 2025), Kimi-K2.6 (April 2025), GPT-5.4 (August 2025), and GPT-5.5 (December 2025); Gemini-3.1-Pro (January 2025) is exploratory because the panel has thin mass before its boundary. Controls are MiniMax-M3 and Claude-Opus-4.7, whose January 2026 cutoffs leave no leakage discontinuity inside the tested window. For each target and each assumed boundary, the statistic is the pre/post jump (90-day windows) in the per-question target-minus-control score difference, using the mean of the available controls on each question. Confidence intervals use the source×month cluster bootstrap. Joint inference over the matrix uses a permutation test that reassigns cutoff dates to models and compares the observed diagonal-minus-off-diagonal contrast with its permutation distribution.

Exploratory Gemini row. Gemini-3.1-Pro's January 2025 cutoff precedes every tested boundary, so all of its cells are placebos: +0.141* (Mar'25), +0.042 (Apr'25), −0.050 (Aug'25), +0.026 (Dec'25). The one starred value sits at the boundary with the thinnest pre-boundary mass, which makes the estimate unstable, and it is not at any documented cutoff; we do not read it as a detection.

Cutoff-resolution robustness for Kimi-K2.6. Kimi-K2.6's coarser (tier-two) cutoff documentation is not load-bearing: its cells at the adjacent March and April boundaries (+0.034, +0.022; Table 2) are both unstarred, so its null survives a one-month cutoff shift.

Clean-anchor arm. ForecastBench archives every real-time forecast submitted to the benchmark before question resolution. For a target with cutoff T, we select *family relatives* (same vendor, earlier release) whose archived real-time forecasts cover questions resolving on both sides of T: GPT-5-Mini and GPT-5.1 for GPT-5.5; GPT-4.1 and o4-mini for GPT-5.4. On each covered question the statistic is our retrospective target score minus the relative's archived real-time score; the crowd anchor cancels in this difference. The estimand is the jump of this difference at T. Because archived forecasts for late-resolving questions can come from more recent submission rounds (shorter horizons), the primary variant restricts to anchor horizons of 10 to 75 days, which balances mean pre/post horizons; the unbanded variant is reported alongside. The band was adopted after the horizon imbalance was observed and is logged as a dated deviation in the versioned

deviations file, before any banded estimate was interpreted. Two protocol checks validate the arm on models that submitted real-time forecasts themselves: retrospective re-queries of GPT-5 reproduce its archived forecasts (mean probability shift -0.021, CI $[-0.056, +0.010]$, $n = 50$), and DeepSeek-V3.1 shows a small positive shift ($+0.021$, CI $[+0.009, +0.032]$, $n = 910$) that we flag as protocol sensitivity; both are an order of magnitude smaller than the GPT-5.5 leakage signature. Because the leakage estimand is a *jump* rather than a level, protocol level effects cancel unless they vary sharply in time. The placebo-jump check tests this directly on DeepSeek-V3.1's own retrospective-minus-real-time score difference ($n = 910$, mean level -0.016): at the December 2025 boundary the jump is $+0.026$ (CI $[-0.011, +0.064]$, $n_{\text{pre}}/n_{\text{post}} = 42/419$), and the only starred jump at any tested date is -0.036 at March 2026, opposite in sign and one third of the banded GPT-5.5 signal. Matrix power: converting the null diagonal cells of the specificity matrix into bounds, their 80%-power minimum detectable effects are 0.077 (DeepSeek-V3.1), 0.108 (Kimi-K2.6), 0.097 (GPT-5.4), and 0.054 (GPT-5.5) anchored-Brier units, computed from the cluster-bootstrap standard errors of each own-cutoff cell.

Reading disagreement between the two designs. For the same model and boundary (GPT-5.5 at December 2025), the matrix reports $+0.034$ (unstarred) and the anchor arm $+0.106^*$. The designs differ in the two respects that matter. Their baselines differ in cleanliness: the control models are guaranteed only to have no *discontinuity* at December 2025—their January 2026 cutoffs let them memorize the same pre-boundary outcomes, and any inflation they carry is subtracted from the target's cell—whereas an archived real-time forecast cannot contain leakage at all. And the paired per-question difference removes the cross-vendor cluster variance that sets the matrix's power floor (0.054 at this cell). Both forces push the matrix cell toward zero, so the pattern is attenuation plus lower power, not contradiction. The auditor's rule: a starred own-cutoff jump that survives placebo dates is a detection scoped to its design's baseline; a null matrix cell bounds the effect but certifies nothing. Taken together, the matrix bounds leakage among 2025-cutoff flagships at a few points of anchored Brier, and the within-family route detects a leakage signature for GPT-5.5 exactly at its documented cutoff.

D.6 M4-C (LiveCodeBench code arm)

LiveCodeBench. LiveCodeBench (Jain et al., 2025) is a coding benchmark that collects problems from competitive programming contests (LeetCode, Codeforces, AtCoder). Each problem carries its *contest release date*, so a model can only have trained on a problem's solution if the contest occurred before the model's training cutoff. The project publishes official per-problem pass@1 (fraction of sampled code completions passing all unit tests) for many models, and refreshes regularly with new contests to stay contamination-free. We use problems from May 2023 through April 2025 (713–1055 contest-dated problems per model, filtered to those present in all evaluated models). All scores are taken from the public submissions repository; no additional inference is run.

Models and cutoffs. The *continuity targets*, evaluated entirely from published per-problem data, are GPT-4o-2024-08-06 (cutoff ≈Oct 2023) and Claude-3.5-Sonnet-20240620 (cutoff ≈Apr 2024). Composition controls have cutoffs *after* the latest problem (Gemini-2.5-Pro, DeepSeek-R1, and the published 2025-cutoff pool). Because control cutoffs lie outside the evaluation window, they have no leakage discontinuity at any target's cutoff (Proposition 15); their pre/post gap at any evaluation-window date captures only composition drift (later contests are harder). Cutoff dates are documented approximations at month resolution.

Self-generated extension. To cover 2024 training cutoffs with current models, we additionally generate one greedy completion per problem through OpenRouter for four targets with documented 2024 cutoffs—GPT-5 (Sep 2024), GPT-5-Mini (May 2024), DeepSeek-V3-0324 (Jul 2024), and Kimi-K2-0905 (Dec 2024)—and two clean controls with 2025 cutoffs (DeepSeek-V3.2 and MiniMax-M1), and grade locally against the official unit tests with the official checker semantics (stdin and functional harnesses, per-problem time limits). The local grader was validated by regrading published DeepSeek-V3 completions: verdict agreement is 98.3%, within the pre-registered three-point tolerance. The same ±160-day window DiD and date-permutation machinery as the continuity arm is applied at the targets' documented cutoffs.

DiD estimation. For each target model and candidate cutoff date T, define a symmetric window $[T-h,\ T+h]$ days (default half-width $h = 160$, giving $n_{\text{pre}} \approx 207$ and $n_{\text{post}} \approx 194$ problems for GPT-4o at Oct 2023). Split problems into pre ($g < 0$) and post ($g \geq 0$). The within-model gap is $\text{gap}^m = \overline{\text{pass@1}}^m_{\text{pre}} - \overline{\text{pass@1}}^m_{\text{post}}$. The pooled composition control's gap is the same quantity computed on the average pass@1 of the two 2025-cutoff models. The boundary-leakage estimate is $\widehat{J} = \text{gap}^M - \text{gap}^{M_0}$: the target's excess gap beyond composition drift. Bootstrap CIs: 4000 resamples of problem IDs (resampled jointly for target and control), 95% percentile interval. The robustness battery (control smoothness, separate controls, date specificity, bandwidth, difficulty stratification) is in Appendix E.5; the global-B sensitivity analyses are in Appendix E.6.

Results of the self-generated extension. The 2024-cohort targets do not reproduce the GPT-4o wild positive. At their documented cutoffs, GPT-5-Mini (+0.005, CI $[-0.058, +0.072]$) and Kimi-K2-0905 (-0.002, CI $[-0.078, +0.075]$) are null. DeepSeek-V3-0324 is marginally positive (+0.051, CI $[-0.013, +0.115]$; date-permutation $p = 0.050$), and its only starred cells sit at the April and May 2024 placebo dates (+0.079, +0.103) rather than at its own July 2024 cutoff—a pattern consistent with diffuse contamination of spring-2024 contest solutions rather than a sharp training-cutoff boundary, so we do not claim a detection. GPT-5 is significantly *negative* (-0.119, CI $[-0.185, -0.051]$): the 2025-cutoff control pool improves more on post-September-2024 problems than GPT-5 does, a violation of the no-differential-trend condition of Proposition 15 for reasoning-era model pairs, and a reminder that the DiD sign test is only interpretable when the control-trend assumption holds. We therefore report the extension as bounded nulls with one flagged assumption failure, and rest the code-domain detection claim on the published-data continuity targets above.

D.7 M5 (disciplined null on forecasting)

Program reading (demoted from the main-text close). Each claim has one primary experiment whose headline result is that claim (Table 4): M1 establishes the confound on cutoff-clean models; M2 anchors dose sensitivity in randomized controlled pretraining; M3 supplies ground truth for the estimators and the laws; M4 shows what honest wild detection looks like, starred where the evidence supports it and null elsewhere; and M5 shows the discipline, with the weak-control failure mode demonstrated rather than assumed and power analysis quantifying what the nulls exclude.

Data. The shared forecasting panel of Appendix D.2, restricted to questions resolving between January 2025 and July 2026: 265 pre-boundary and 1,207 post-boundary questions, evaluated pairwise (every model on every question).

Boundary. The primary target Qwen3.5-35B-A3B has only a year-level documented cutoff (2026), so we place the boundary at July 2025, inside its plausible training window, and report a boundary sweep from -180 to $+180$ days around it in 30-day steps. The adjusted estimate is null at every placement.

Control selection. The control is chosen *before* any DiD is computed, by a pre-registered profile-matching protocol: on post-boundary questions that are clean for every candidate, compute each candidate's distance to the target (squared-error distance between calibration profiles plus squared-error distance between per-source mean scores). Between the two flagships with post-window-clean cutoffs, GPT-5 (distance 0.0051) is selected over Gemini-3.1-Pro (0.0088). GPT-3.5-Turbo (cutoff September 2021) serves as the deliberately mismatched weak control. Secondary targets are Kimi-K2.6, DeepSeek-V3.2, GLM-4.7, and MiniMax-M3, all queried retrospectively through the shared protocol.

Paired DiD estimation. For each question i, form the paired difference $D_i = L_i^{\text{target}} - L_i^{\text{control}}$ of crowd-anchored Brier reductions. The adjusted estimate is the pre/post difference in mean D_i, with source×month cluster-bootstrap confidence intervals (4,000 resamples). Power is assessed by semi-synthetic injection: a synthetic leakage effect of known size is added to pre-boundary D_i and the detection rate of the full pipeline is recorded over 300 replications per effect size. The boundary sweep and power results are in Appendix E.7.

Table 7: **M5: full estimates behind Figure 4.** Column 2: the naive pre/post gap Δ at the July 2025 boundary; positive reads as leakage. Column 3: the paired DiD against the pre-registered profile-matched control (GPT-5); for a leak-free model the ideal value is zero. Cluster-bootstrap 95% CIs; *: CI excludes zero. Result: no adjusted estimate is significantly positive (marginal negatives are slight overcorrection, the conservative direction); the weak-control row substitutes GPT-3.5-Turbo and re-inflates the estimate, the control-mismatch failure mode of Assumption 4.

model	naive gap [CI]	adjusted DiD [CI]
Qwen3.5-35B-A3B (target)	$+0.060$ $[-0.008, +0.108]$	$+0.020$ $[-0.063, +0.074]$
Kimi-K2.6	$+0.024$ $[-0.008, +0.055]$	-0.016 $[-0.058, +0.010]$
DeepSeek-V3.2	$+0.006$ $[-0.021, +0.027]$	-0.034^* $[-0.088, -0.000]$
GLM-4.7	$+0.009$ $[-0.025, +0.042]$	-0.031^* $[-0.076, -0.002]$
MiniMax-M3	$+0.039^*$ $[+0.008, +0.074]$	-0.000 $[-0.039, +0.026]$
weak control (GPT-3.5-Turbo)	—	$+0.055$ $[-0.028, +0.151]$

Results. Table 7 reports the per-model naive gaps and adjusted DiD estimates summarized in Section 7.2 and plotted in Figure 4.

Provenance of the fitted overconfidence temperature. The value $T \approx 8$ cited in Remark 2 and used to set the T4 sweep range ($T_{\text{over}} \in \{2, 4, 8\}$) was fit on this forecasting pipeline's data: temperature scaling $p_{\text{cal}} = \sigma(\text{logit}(p)/T)$ with T chosen to minimize log-loss on Qwen3.5's *leakage-free* post-boundary control forecasts ($n = 168$ ex-ante-uncertain factual questions, crowd anchor frozen in $[0.25, 0.75]$). The fit reached the upper end of the search interval $(0.3, 8]$, so $T \approx 8$ is best read as a lower bound on the distortion: Qwen's raw probabilities are severely overconfident, which is why the raw residual covariance is negative and PRC requires recalibration before use.

E Complementary experiments and robustness analyses

The synthetic validation study E1 and its two supplementary arms come first; the remaining subsections report one robustness analysis each, in a fixed format: the *intuition* (what question it answers and which claim it defends), the *setup* (a short description or a pointer to Appendix D where shared), and the *results*.

E.1 E1 (synthetic validation)

This synthetic study validates the estimators under known data-generating assumptions. It formerly opened the main-text experiment section and is retained here in full; the real-data ground-truth role is now carried by M3. All sub-tests share a common data-generation process, described first, with per-test variations noted below; results follow at the end of this subsection, and the supplementary arms T5–T6 follow in Appendices E.2 and E.3.

Common data generation. For each question i, draw a crowd anchor $c_{0,i} \sim \text{Uniform}(0.15, 0.85)$ and a binary outcome $Y_i \sim \text{Bernoulli}(c_{0,i})$. The *honest forecast* is $P_{\text{hon},i} = \text{clip}\big(c_{0,i} + s\,(Y_i - c_{0,i}) + \epsilon_i,\ [10^{-3}, 1-10^{-3}]\big)$, where $s \in [0, 1]$ is a skill parameter controlling how much the model improves on the crowd, and $\epsilon_i \sim \mathcal{N}(0, 0.05)$ is idiosyncratic noise. For a leaked question with extraction weight $w \in [0, 1]$, the observed forecast is the convex blend $P_i = \text{clip}\big((1-w)\,P_{\text{hon},i} + w\,Y_i\big)$; for a clean question, $P_i = P_{\text{hon},i}$. An optional overconfidence parameter $T_{\text{over}} \geq 1$ transforms P_i via $P_{\text{obs},i} = \sigma(\text{logit}(P_i) \cdot T_{\text{over}})$, where σ is the logistic function; $T_{\text{over}} = 1$ leaves the forecast unchanged.

T1 (recovery and coverage). For each extraction weight $w \in \{0, 0.2, 0.4, 0.7, 1.0\}$, generate $n = 400$ leaked questions and $n = 400$ clean-control questions (both with skill $s = 0.3$, no overconfidence). The true inflation is $B_{\text{true}} = \frac{1}{n}\sum_i \left[(P_{\text{hon},i} - Y_i)^2 - (P_i - Y_i)^2\right]$, computed from the known honest forecast. The estimated inflation is the Brier-reduction DiD: $\widehat{B} = \bar{S}_{\text{leak}} - \bar{S}_{\text{clean}}$, where $\bar{S} = \frac{1}{n}\sum (c_0 - Y)^2 - (P - Y)^2$. Bootstrap CIs: 200 problem-level resamples, 95% percentile interval. We run 300 Monte Carlo repetitions per w and report the mean bias $\widehat{B} - B_{\text{true}}$ and the fraction of repetitions in which the CI covers B_{true}.

T2 (RD recovery and placebo). Each question i has a resolution-time gap $\delta_i \sim \text{Uniform}(-180, 180)$ (days to the cutoff). The honest skill varies smoothly with recency: $s(\delta) = 0.4 - 0.25\,|\delta|/180$, so the model is better near its cutoff (pure recency, no leakage). Leakage of weight $w_0 \in \{0, 0.3, 0.6\}$ is injected only on the leaked side ($\delta < 0$). The RD jump is estimated by local-linear regression with a triangular kernel and half-bandwidth $h = 120$ days: fit $\ell_i = (P_i - Y_i)^2 \approx \beta_0 + \beta_1\,\mathbf{1}[\delta < 0] + \beta_2\,\delta + \beta_3\,\delta \cdot \mathbf{1}[\delta < 0]$ on the weighted sample $|\delta| \le h$; the jump estimate is $-\hat{\beta}_1$. The *placebo* repeats the same procedure with no leakage injected ($w_0 = 0$ forced, though recency is still present). $n = 1500$ per run, 300 repetitions.

T3 (complementarity). For each extraction weight $w \in \{0, 0.25, 0.5, 0.75, 1.0\}$, generate $n = 8000$ questions. For each question, query $K = 12$ independent "paraphrases" (simulated by adding independent noise $e_k \sim \mathcal{N}(0, 0.12)$ to the crowd anchor before blending with w). The consensus is $\bar{y}_i = K^{-1} \sum_k P_i^{(k)}$. Compute the PRC as $\widehat{\Delta} = \widehat{\text{Cov}}(\bar{y},\, Y - \bar{y})$ and the leakage score $L = \frac{1}{n} \sum \left[(c_0 - Y)^2 - (\bar{y} - Y)^2\right]$.

T4 (calibration under overconfidence). This sub-test isolates the calibration step. The honest forecast is set to the crowd anchor itself, $P_{\text{hon}} = c_0$, with $Y \sim \text{Bernoulli}(c_0)$, so the clean forecaster is calibrated by construction. Leaked forecasts are $P = (1 - w)\,P_{\text{hon}} + w\,Y$ with $w = 0.8$. Reported probabilities are then distorted by overconfidence: $P_{\text{obs}} = \sigma(T_{\text{over}}\,\text{logit}(P))$, for $T_{\text{over}} \in \{1, 2, 4, 8\}$. For each repetition, we draw $n = 600$ leaked and $n = 600$ clean evaluation questions, plus a *separate* leakage-free calibration set of size $n_{\text{cal}} \in \{100, 250, 500, 1000, 2000, 5000\}$. Temperature scaling ($\hat{T}$ minimizing log-loss) is fit *only* on the calibration set and then applied to both evaluation sets before computing DiD. The *naive* estimator skips calibration and uses P_{obs} directly. We report mean bias, mean absolute bias, and RMSE over 200 repetitions per $(T_{\text{over}},\, n_{\text{cal}})$ cell.

T5 (no free inflation; results in Appendix E.2). Honest forecasts are generated as in T1 with $w = 0$ (zero true leakage). The "perturbed" side replaces P_{hon} by $\sigma(\text{logit}(P_{\text{hon}}) + \eta)$ with $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$ drawn independently of Y; logit-space noise avoids the outcome dependence that clipping would induce at the probability boundaries. Noise scales $\sigma_\eta \in \{0.25, 0.5, 1.0\}$; $n = 400$ per side; the same Brier-reduction DiD as T1; 300 Monte Carlo repetitions per scale.

T6 (concentration across surprise strata; results in Appendix E.3). Fixed extraction weight $w = 0.5$, $n = 4000$ leaked and 4000 clean questions per repetition, 300 repetitions. Questions are stratified into quartiles of the *observable* crowd surprise $(c_0 - Y)^2$ (quartile cuts computed on the pooled sample). Within each stratum, the stratified DiD is the mean Brier reduction of the leaked group minus that of the clean group; the prediction is the stratum mean of $b_0\,w(2 - w)$ with $b_0 = (P_{\text{hon}} - Y)^2$, per Proposition 3.

Results (T1–T4; Figure 10). *T1 (DiD recovery and coverage).* Across five leakage weights ($w \in \{0, 0.2, 0.4, 0.7, 1.0\}$, true $B \in \{0, 0.038, 0.067, 0.095, 0.105\}$), the DiD estimator is essentially unbiased (mean bias $\approx 9 \times 10^{-4}$ at every level) and the 95% bootstrap CIs cover the true B in 94–100% of repetitions; the no-leak case returns a null. *T2 (RD recovery and placebo).* With a smooth recency surface and leakage injected only on the pre-cutoff side, the RD jump tracks the injection (+0.000, +0.039, +0.065 for $w_0 = 0, 0.3, 0.6$) while the no-leakage placebo jump is ≈0.000 at every level: continuous recency alone does not fabricate a discontinuity (Theorem 4). *T3 (RD/PRC complementarity).* Sweeping w confirms Proposition 18: PRC is hump-shaped (peak +0.052 near w=0.5, vanishing at w=1) while the RD signal grows monotonically to +0.209 at w=1; recovered ratios $L(w)/L(1)$ match $w(2 - w)$ (Table 8). *T4 (calibration under overconfidence).* Distorting a calibrated clean forecaster by $T_{\text{over}} \in \{2, 4, 8\}$ produces naive DiD bias 0.022, 0.055, 0.085; temperature scaling fit on n_{cal} leakage-free anchors reduces it to ≈0.008 at $n_{\text{cal}} = 100$ and ≈0.001 at $n_{\text{cal}} = 5000$. Calibration works but is a real resource requirement (Remark 2).

E.2 No free inflation under outcome-orthogonal noise (E1 arm T5)

Intuition. Corollary 10 states that a predictor can raise its score only through a positive loading on the leaked signal: perturbations that are independent of the outcome cannot inflate. A useful falsification check on the estimator is therefore whether it can be fooled into booking outcome-orthogonal noise as leakage—if it could, positive estimates in M2–M4 would be uninformative.

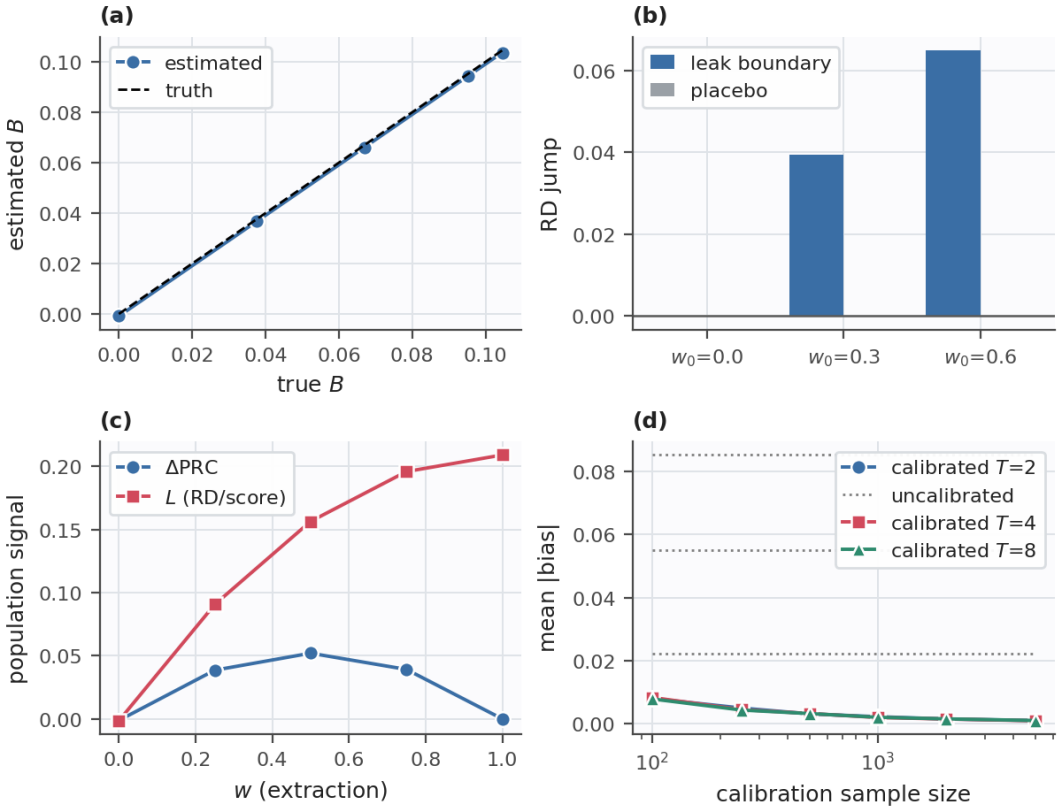

Figure 10: **E1: the estimators recover known ground truth in simulation.** (a) DiD estimates lie on the truth line (bootstrap coverage 94–100% across leakage levels). (b) RD jumps track injected boundary leakage; no-leak placebos stay at zero under smooth recency. (c) PRC is hump-shaped and vanishes at pure answer leakage (w=1) while the score-based signal grows monotonically—the two detectors are complements. (d) Overconfidence biases the naive statistic; temperature calibration on leakage-free anchors removes the bias as n_{cal} grows. Result: each estimator behaves as Propositions 5 and 16, Theorem 4, and Remark 2 predict under known data-generating assumptions.

Setup. As defined in Appendix E.1 (T5): honest forecasts with zero true leakage, perturbed by zero-mean logit-space noise independent of Y at scales $\sigma_\eta \in \{0.25, 0.5, 1.0\}$; the same Brier-reduction DiD estimator as T1; 300 Monte Carlo repetitions per scale.

Results. The estimate is deflationary at every scale, increasingly so as noise grows: mean $\widehat{B} = -0.003$ (95% CI of the mean $[-0.004, -0.002]$) at $\sigma_\eta = 0.25$; -0.012 ($[-0.013, -0.011]$) at $\sigma_\eta = 0.5$; and -0.044 ($[-0.045, -0.043]$) at $\sigma_\eta = 1.0$. Individual repetitions can come out slightly positive under sampling noise at the smallest scale (34% of repetitions; 5% at $\sigma_\eta = 0.5$; none at $\sigma_\eta = 1.0$), but the mean effect is strictly negative throughout, matching Corollary 10: noise costs Brier score, it cannot fake inflation (Figure 11a).

E.3 Leakage concentration across surprise strata (E1 arm T6)

Intuition. The per-question law $L = b_0\, w(2 - w)$ (Proposition 3) makes two testable predictions beyond monotonicity in w: at fixed extraction, leakage should *concentrate* on questions where the crowd is surprised (large b_0), and across extraction levels the total leakage should follow the $w(2 - w)$ shape exactly. This licenses the main-text statement that leakage is heterogeneous and concentrated rather than a uniform rate.

Setup. As defined in Appendix E.1 (T6): fixed $w = 0.5$, quartiles of observable crowd surprise $(c_0 - Y)^2$, stratified DiD against the stratum-mean prediction $\mathbb{E}[b_0\, w(2 - w)]$; 300 repetitions. The functional-form check reuses the T3 sweep (Appendix E.1), comparing recovered leakage ratios $L(w)/L(1)$ with the predicted $w(2 - w)$.

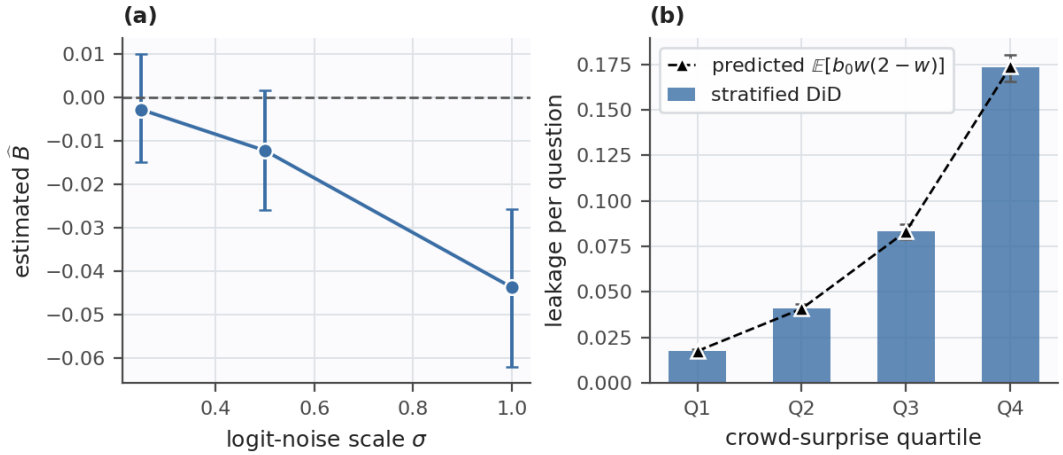

Figure 11: **E1 supplementary arms: noise cannot fake inflation, and leakage concentrates where the law predicts.** (a) T5: under outcome-orthogonal logit noise with zero true leakage, the DiD estimate stays at or below zero at every noise scale, as Corollary 10 requires (whiskers: 2.5–97.5 percentile range across repetitions). (b) T6: the stratified DiD (bars) tracks the predicted $\mathbb{E}[b_0\, w(2-w)]$ (dashed) across crowd-surprise quartiles at fixed $w = 0.5$.

Results. The stratified DiD estimate rises monotonically and steeply across surprise quartiles (0.018, 0.041, 0.083, 0.174 for Q1 through Q4) and matches the predicted stratum means (0.018, 0.041, 0.083, 0.174) to within Monte Carlo error (Figure 11b): the top quartile carries roughly ten times the leakage of the bottom quartile at the same extraction weight. The functional form is confirmed across the full sweep (Table 8): observed ratios track $w(2-w)$ to within 0.005 at every w. A caveat applies to both arms: the synthetic generator implements the convex blend of Assumption 2, so these checks establish that the estimators recover the law when it holds—they are consistency checks of the machinery, not tests of the convex-pull assumption itself, whose real-data validation remains open (Section 9).

Table 8: **E1: the recovered leakage profile matches $w(2-w)$.** Leakage ratios $L(w)/L(1)$ from the synthetic extraction-weight sweep (arm T3) against the prediction of Proposition 3. Result: observed and predicted agree within 0.007 at every w.

extraction weight w	0	0.25	0.5	0.75	1.0
observed $L(w)/L(1)$	−0.007	0.433	0.748	0.937	1.000
predicted $w(2-w)$	0.000	0.438	0.750	0.938	1.000

E.4 Hubble robustness (M2)

Intuition. The M2 dose-response supports the mechanism only if it is not an artifact of item difficulty, a single model configuration, or a single task.

Setup. All quantities are computed from the published Hubble accuracies as configured in Appendix D.3.

Results. *Difficulty-matching check.* The standard model's accuracy varies by only 0.037–0.051 across duplication bins (per task: MMLU 0.046, PIQA 0.046, HellaSwag 0.051, WinoGrande 0.037). This range is much smaller than the moderate/high-dose effects, bounding difficulty imbalance as an explanation; it is comparable to the low-dose r=4 effect, so low-dose conclusions should be interpreted more cautiously.

Robustness across scale and tokens. The dose-response persists across configurations: 8B-100B shows stronger low-dose sensitivity (less token dilution reduces the contamination signal per document), 8B-500B remains strong at moderate/high dose, and 1B-500B shows large effects only at high duplication (lower capacity limits usable extraction). This heterogeneity is consistent with the extraction factor $w = c \cdot u \cdot \kappa$: exposure must be both absorbed and usable.

Per-task consistency. For 8B-500B, the high-dose effect is not driven by a single task: MMLU, HellaSwag, and WinoGrande all show strong increases, while PIQA rises more mildly and saturates earlier. Task heterogeneity is expected because w can vary by task format.

E.5 LiveCodeBench robustness battery (M4-C)

Intuition. The M4-C boundary estimate is credible only if the no-discontinuity control assumption holds empirically, the signal is not an artifact of one control model or one bandwidth, the effect is specific to documented cutoffs, and difficulty imbalance is excluded. The named continuity violations of Route 1—composition shifts and corpus-density jumps at the cutoff—are exactly what the placebo and bandwidth checks below target.

Setup. All checks use the DiD estimator and bootstrap of Appendix D.6, computed on the published per-problem data.

Results. *R1 (control smoothness).* The no-discontinuity assumption is directly tested by estimating local-linear jumps for each control at each target cutoff. Pooled-control jumps: Oct'23 -0.010 [-0.093,$+0.077$]; Dec'23 -0.012 [-0.097,$+0.070$]; Apr'24 -0.063 [-0.180,$+0.040$]. All intervals include zero, supporting the identification condition.

R2 (separate-control robustness). Own-cutoff estimates are computed against each control individually. For GPT-4o-0806: $+0.094$ vs. Gemini, $+0.086$ vs. DeepSeek, $+0.090$ pooled. The signal is not driven by a single control model.

R3 (date specificity). A monthly cutoff scan finds no month outside ± 1 month of GPT-4o's documented cutoff with a larger DiD. A permutation test over model-cutoff assignments in the specificity matrix yields empirical $p = 0.025$ across all targets.

R4 (bandwidth). Own-cutoff DiD vs. pooled controls across half-widths $h \in \{100, 120, 160, 200, 240, 300\}$ days: GPT-4o point estimates are stable at $+0.084$ to $+0.100$; significance appears for $h \geq 160$ days.

R5 (difficulty stratification). DiD stratified by problem difficulty (easy/medium/hard): GPT-4o-0513 $+0.096$ [$+0.018$,$+0.169$]; GPT-4o-0806 $+0.089$ [$+0.015$,$+0.166$]; Claude-3.5 $+0.079$ [-0.002,$+0.156$]. GPT-4o survives cleanly; Claude becomes marginal after stratification, so per-date inference for Claude leans on the date-permutation test ($p = 0.017$).

Specificity. For each continuity target, the DiD is computed at each candidate cutoff date. Positive effects concentrate at the model's own documented cutoff (main text, Section 7); off-diagonal cells are null, and monthly scans find no larger effect away from the documented cutoffs.

E.6 Why M4-C does not identify global B

Intuition. Converting the boundary jump $\widehat{J}$ into a global leakage-adjusted score requires either a strictly clean, recency-matched control (Assumption 4) or the extrapolation assumption (Assumption 6). Both routes fail on this dataset; documenting the failure disciplines the boundary-only scope of M4-C's claim.

Setup. Same data and estimator as Appendix D.6; the strictly clean control is GPT-4-0613 (cutoff before all problems); extrapolations fit time polynomials plus difficulty/platform indicators on the clean side.

Results. GPT-4-0613 is strictly clean on all LiveCodeBench questions and produces large DiD estimates ($+0.145$ to $+0.168$), but its quarterly performance profile correlates only 0.36–0.39 with GPT-4o (centered RMSE 0.074–0.076), violating the matched-recency requirement and exhibiting the weak-control trap. Clean-side extrapolation is unstable: implied corrections range from -0.175 to $+0.102$ for GPT-4o-0513 and from -0.105 to $+0.172$ for GPT-4o-0806 across polynomial degrees and post-cutoff horizons. We therefore report boundary J and do not convert it into a global leakage-adjusted score.

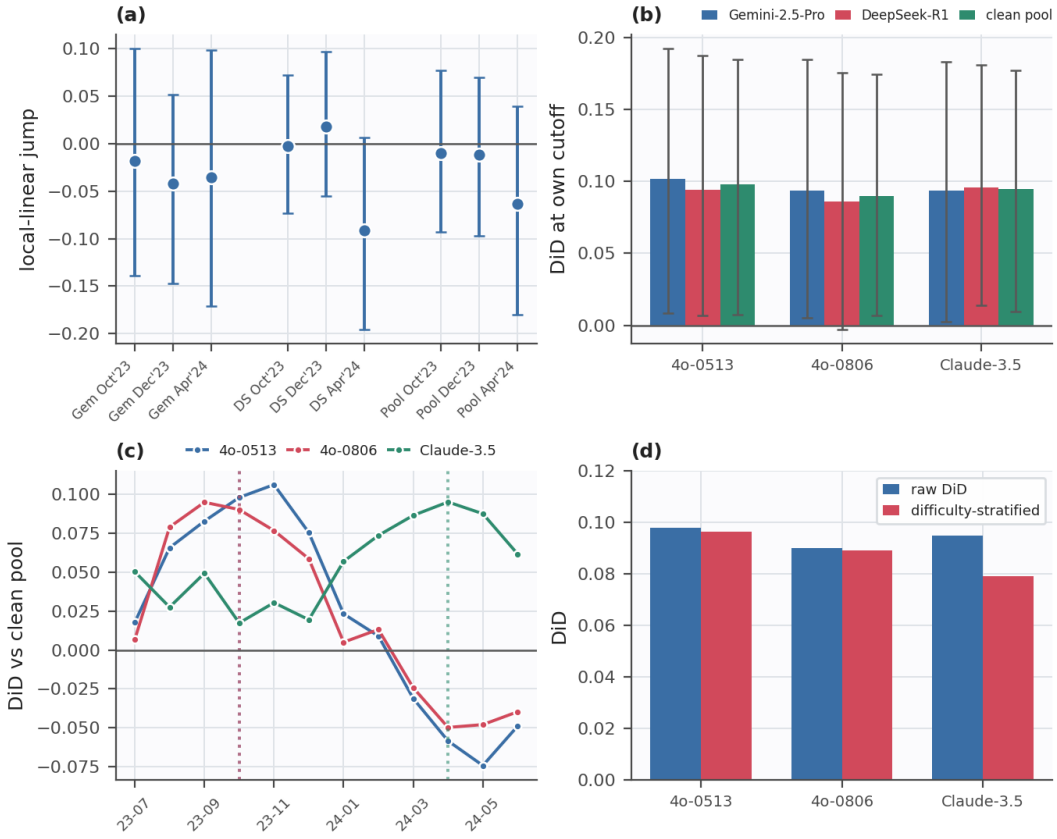

Figure 12: **M4-C: the LiveCodeBench boundary signal survives the robustness battery.** (a) Composition controls show no local discontinuity at target cutoffs. (b) Own-cutoff estimates are stable across individual controls. (c) Monthly scans concentrate positive effects at documented cutoffs. (d) Difficulty stratification preserves GPT-4o's signal and weakens Claude's. Result: GPT-4o's boundary leakage is robust on every check; Claude's is reported as supporting evidence only.

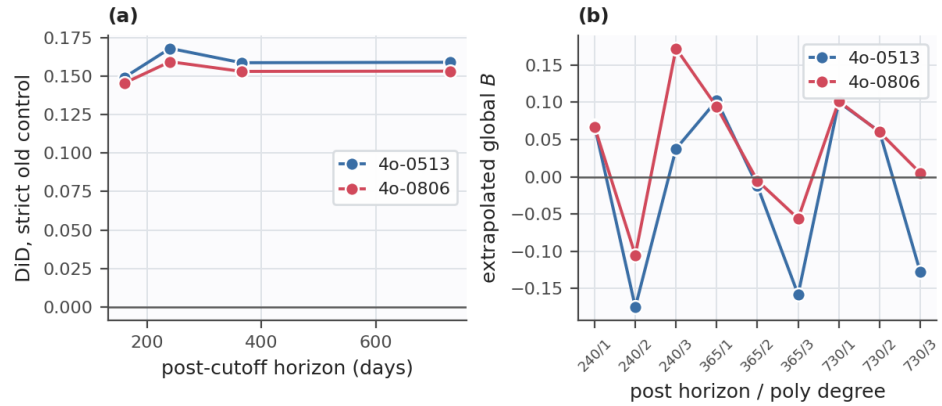

Figure 13: **M4-C identifies the local boundary jump, not a global correction.** (a) The strictly clean GPT-4-0613 control yields a large contrast but fails temporal-profile matching (Assumption 4). (b) Clean-side extrapolated corrections change sign across horizons and polynomial degrees (Assumption 6). Result: both routes to a global B fail on this dataset, so the M4-C claim is scoped to the boundary.

E.7 Forecasting robustness and power (M5)

Intuition. A null is only informative if the control is demonstrably well matched, the result is stable across boundary placements, and the design has power to detect effects of practical size.

Setup. All quantities use the paired cluster-bootstrapped DiD of Appendix D.7.

Results. *Control validation.* The pre-registered profile-matching protocol selects GPT-5 over Gemini-3.1-Pro (post-boundary profile distance 0.0051 against 0.0088) before any DiD is read. Substituting the deliberately mismatched GPT-3.5-Turbo control flips the primary adjusted estimate from +0.020 to +0.055, the spurious positive that an ill-matched control manufactures under Assumption 4.

Boundary sensitivity. Adjusted estimates for boundaries swept from −180 to +180 days around July 2025 in 30-day steps range from −0.035 to +0.057 and are never significantly positive; placements more than 90 days before the reference boundary retain fewer than 62 pre-boundary questions and are reported for completeness only. The null does not depend on exact placement of Qwen3.5's undocumented knowledge horizon. The sweep is informative only about leakage that *differs* across the assumed boundary: because every tested placement lies inside the plausible training window, inflation roughly uniform over the window would produce a null at all of them.

Power analysis. Semi-synthetic effect injection on the observed paired panel detects an injected effect of 0.05 Brier-reduction units in 97.3% of replications and an M4-C-sized effect of 0.09 in 100%. The design has high power for practically meaningful effects; effects well below 0.05 remain difficult to exclude.

E.8 Matched real-data PRC sensitivity

Intuition. E1-T3 shows PRC detects evidence leakage in principle; the question is whether Route 3 yields a real calibrated positive on frontier data, under a pre-specified robustness gate.

Setup. A separate matched experiment was pre-specified before querying: Qwen3.5 with $K = 8$ paraphrases on outcome-balanced treatment/control cohorts in two domains with temporal support on both sides, finance (150/150) and other (130/130); 552 of 560 questions had at least six valid paraphrase probabilities.

Estimator and uncertainty. Calibration was learned only from clean controls: five-fold cross-fitting gave out-of-sample control predictions, and the five control-trained calibrators were ensembled on treatment questions. Calibration was applied to each paraphrase before computing consensus and within-question variance. The estimator was $\widehat{\Delta}_{\mathrm{bc}} = \widehat{\mathrm{Cov}}(\bar{P}, Y - \bar{P}) + \frac{1}{n}\sum_i \widehat{W}_i/K_i$. Nested question bootstraps refit the calibrator and recomputed the finite-K correction in every draw; Platt scaling was primary, isotonic the pre-specified sensitivity.

Results. Under Platt scaling, finance yielded excess PRC +0.0116 [+0.0009, +0.0169], other yielded +0.0084 [−0.0046, +0.0164], and the equal-weight pooled excess was +0.0100 [+0.0018, +0.0148]. Clean-control point estimates were near zero. Under isotonic calibration, however, finance was +0.0142 [−0.0097, +0.0188], other reversed to −0.0026 [−0.0264, +0.0124], and the pooled excess was +0.0058 [−0.0120, +0.0123]. The experiment therefore failed its pre-specified calibration-stability gate and is reported as suggestive sensitivity evidence only; Route 3 remains a calibration-gated secondary diagnostic.

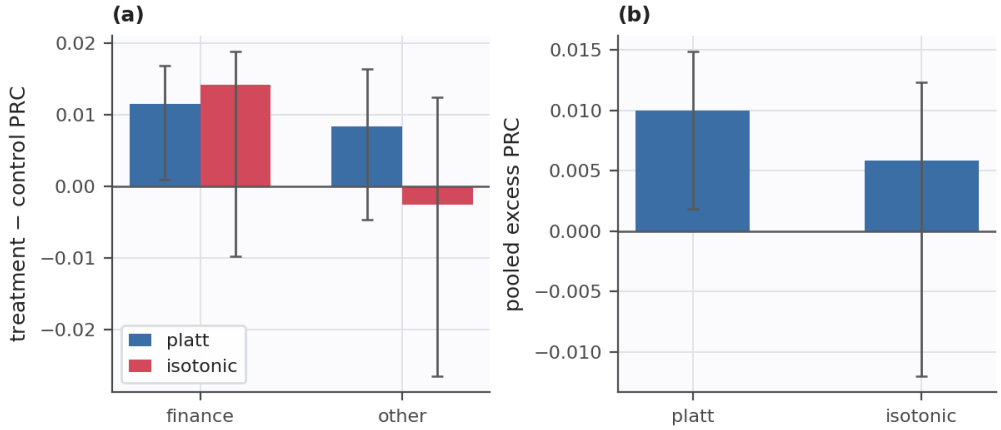

Figure 14: **Matched real-data PRC: the calibration-stability gate fails, so PRC remains secondary.** Treatment–control PRC excess by domain and pooled, under Platt and isotonic calibration. Result: Platt yields positive excesses, but isotonic does not replicate the pooled significance and flips the other-domain point estimate; by the pre-specified gate, PRC is reported as suggestive sensitivity evidence only.

F Experiment provenance

Table 9: **Reproducibility map.** Authoritative scripts and outputs for reported experiments.

Result	Script	Primary output
Panel build	`redesign/build_panel.py`	`data/panel_market.jsonl`
M1 clean flagships	`redesign/analysis/m1_analyze.py`	`m1/results.json`
M2 Hubble	`M2_hubble/run_m2.py`	`results.json`
M3 corpus and twins	`redesign/m3/build_corpus.py`, `train_twin.py`	`m3/corpus_full/`
M3 analysis	`redesign/m3/analyze_m3.py`	`m3/analysis_full.json`
M3 PRC	`redesign/m3/prc.py`	`m3/prc_differenced.json`
M4-F matrix, anchor	`redesign/analysis/m4f_{analyze,anchor}.py`	`m4f/*.json`
M4-C	`redesign/m4c/analyze.py`	`m4c/m4c_results.json`
M5	`redesign/analysis/m5_analyze.py`	`m5/results.json`
E1 synthetic (T1–T6)	`M1_synthetic/run_m1.py`	`results.json`
Matched PRC	`M6_prc_matched/run_prc_matched.py`	`results.json`
Main-text figures	`redesign/analysis/paper_figures.py`	`figures/*.png`
Appendix figures	`redesign/analysis/appendix_figures.py`	`figures/*.png`