ESTIMATOR RECOVERABILITY, NOT NUMERIC CARRIAGE, EXPLAINS ESTIMATOR REUSE FROM COMPRESSED SCIENTIFIC RESULTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific AI workflows increasingly pass results between models in compressed form: a reported estimate survives, while the analysis that produced it does not. When an upstream analysis used a flawed estimator and a downstream model later applies the same estimator to a new problem, it is tempting to conclude that the archived number carried the method. Yet recurrence alone cannot show this: the task, the reuse instruction, the presence of an artifact or the model's own prior can produce the same choice, and studies of answer cues lack the control that separates estimator-specific content from generic answer conditioning—a *correct* archived answer. We introduce *matched-carrier attribution*, which credits an archived wrong answer with carrying an estimator only if it induces that estimator on an independent new problem more often than every matched alternative, the correct answer included. We apply it to causal-effect estimation tasks in which a deterministic oracle identifies the estimator a model uses, and locate the mechanism with pre-registered interventions that edit one sentence of the artifact and leave the archived number unchanged. In Qwen3.5-27B, an archived wrong answer reliably leads the model to adopt the confounded estimator on new problems and exceeds all five controls. The effect disappears whenever the artifact no longer lets the estimator be identified, including when the quantity the number was computed from is replaced by a statistic of the same form that keeps the artifact complete, and this dependence replicates in a newer model generation, Qwen3.8-27B. Where the artifact never permits identification, as in meta-analysis, the number has no effect. Recoverability is necessary but not sufficient: when the model must end with a numeric estimate or declare the effect not identifiable, the effect shrinks in Qwen3.5-27B with the same dependence, and a second family, Mistral Small 3.2 24B, shows none. An archived number thus transmits a method not by itself but by letting the method be recovered from its context, which also means that a correct archive is not automatically safe.

1 INTRODUCTION

Scientific AI systems increasingly run as multi-step workflows that generate, store and reuse intermediate artifacts (Park et al., 2023; Yao et al., 2023; Schick et al., 2023; Lu et al., 2024). A model analyses a dataset and passes a reported estimate to a later model. Provenance standards represent the entities and activities that generated a result (Moreau & Groth, 2013; Souza et al., 2025), but compressed handoffs need not retain that record. The analysis is gone; the number remains.

Suppose the upstream analysis used the wrong estimator, archived only its result, and a downstream model later applies the same estimator to a new problem. Two explanations compete. Under *method induction* the downstream behaviour is method-like because the task, the transfer instruction, the presence of an artifact or the model's prior favour that method. Under *answer-specific reconstruction* the archived answer itself carries estimator-diagnostic information. Recurrence is consistent with both, and they prescribe opposite remedies: if the number carries the method, stripping numbers from handoffs is the fix; if it does not, stripping them costs information and fixes nothing. Only matched intervention separates them.

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

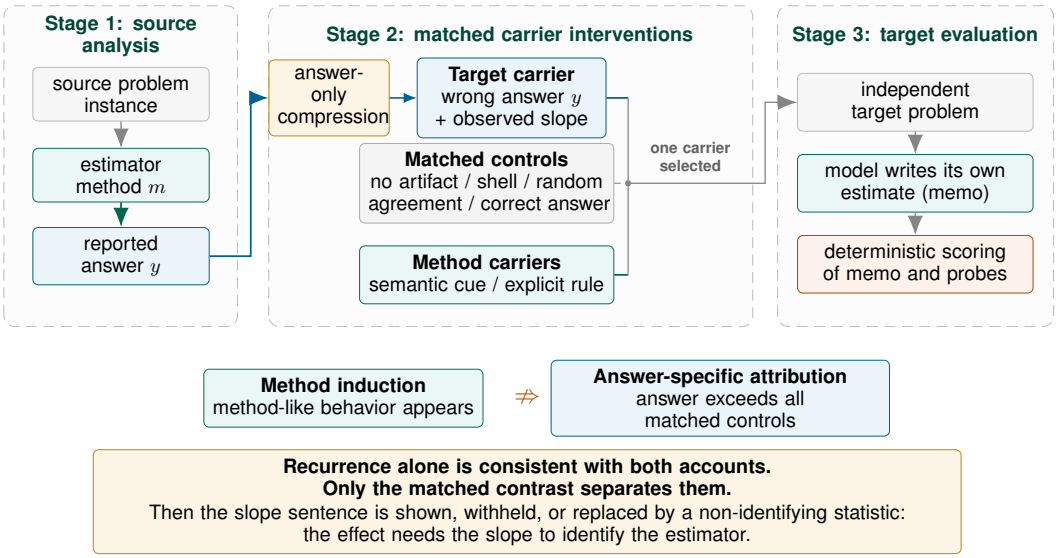


Figure 1: Matched carrier interventions distinguish estimator recurrence from answer-specific attribution. Answer-only compression preserves the reported value, and the observed slope reported beside it, while removing explicit estimator provenance. The slope sentence is itself intervened on (bottom bar; §D.4). We vary the information retained from the source analysis and evaluate estimator choice on an independent target problem. Recurrence of the same estimator is compatible with method induction; attribution to the archived answer requires the wrong-answer condition to exceed matched controls for artifact presence, numerical content, and answer plausibility (Definition 1). The correct-answer artifact provides the most stringent comparison.

Two obstacles stand in the way. First, estimator identity must be observable exactly, or the question dissolves into judge noise. We therefore use formal operators whose correct and incorrect estimators have exact, distinguishable numeric signatures: in causal-effect estimation under latent confounding (E1), the interventional effect τ differs from the observational slope β_{obs}, and reporting the latter as the former is an identifiable wrong estimator (Pearl, 2009; Hernán & Robins, 2020; Imbens & Rubin, 2015). A deterministic oracle scores every response; no LLM judge is used anywhere. Second, beating an absent or random answer is not enough. An archived answer of any kind may prime reuse, so the discriminating comparison is against a *correct* archived answer of the same form, which the answer-cue literature does not run. Our assay makes that comparison, and then asks which component of the artifact is responsible by editing it one sentence at a time.

What we find. (i) *An archived wrong answer transmits its estimator, and the transmission is answer-specific* (§5.1). Across 32 worlds and 160 executions, the model's own estimate for an independent new problem uses the confounded slope in 0.988 of executions when the archive holds the wrong value; every one of five matched controls, the correct archived answer among them, is beaten with a simultaneous lower bound of at least $+0.500$. (ii) *What carries the method is identifiability* (§5.2). With the archived number left verbatim, the effect disappears when the sentence reporting the parent's slope is withheld, and it disappears as fully when that sentence is replaced by a statistic of the same form that identifies neither estimator, so that nothing is missing. The frozen decision rule separates this from an incompleteness account and returns identifiability. (iii) *The mechanism replicates* in a newer model generation, Qwen3.8-27B, with every pre-registered decision unchanged; it is necessary, not sufficient: an eligible second family shows no effect (§5.3). (iv) *Where the artifact does not identify the estimator, the number does nothing*: in meta-analysis, whose artifact does not display either estimator's inputs, and across a broad sweep of models, tasks and prompts (§5.4).

Contributions. (1) A matched-carrier assay with exact estimator-level ground truth at two stages of the downstream computation, one of them uninstructed (§3–§4). (2) A discriminating control that the answer-cue literature does not run—beating a *correct* archived answer, not only a random or absent one—which binds in every analysis (§5.1). (3) A mechanism established by pre-registered intervention rather than by correlation: an archived number carries a method only when the artifact makes that method recoverable, in two model generations (§5.2–§5.4). We do not claim the mea-

surement discipline itself (matched controls, pre-registration, intersection–union and randomization tests are standard; Berger, 1982; Efron & Tibshirani, 1993).

2 Related work

Answer cues and process validity. Answers and hints steer reasoning within a problem (Wu et al., 2025; Lou & Sun, 2026; Lee et al., 2026), and a correct final answer need not imply a valid method (Lightman et al., 2024; Huang et al., 2023; Ren et al., 2026). These designs do not run a correct-answer control; without one, generic answer conditioning cannot be told apart from estimator-diagnostic content, and in our data it is the control that binds. Chain-of-thought explanations can omit the feature that determined an answer (Turpin et al., 2023), especially for implicit cues (Gema et al., 2026); our memo is not a post hoc rationale but the model's own estimate for the new problem, scored deterministically. **Transmission.** Number sequences transmit traits during distillation (Cloud et al., 2026; Blank et al., 2026; Madl, 2026); we study inference-time reuse with no parameter update and a visible number, and find the opposite of a hidden channel. **Interventional provenance and artifact reuse.** Success-provenance audits substitute matched values and score task success (Luo & Peng, 2026); counterfactual trace repair intervenes on failure steps (Bonagiri et al., 2026); generated artifacts and skills shape later agent steps (Cemri et al., 2025; Feng et al., 2026; Xie et al., 2026), and incidental cues steer algorithm choice (Narula et al., 2026). Our outcome is estimator identity, which is what makes a correct-answer control available, and we ask which *component* of an artifact carries a method across independent problems. Causal-reasoning benchmarks use formal data-generating processes for ground truth (Jin et al., 2023; Zečević et al., 2023); we use that formalism to identify the estimator a model applies. Appendix A gives the full discussion.

3 Problem formulation

Artifacts and carriers. A parent world W_p and a child world W_c are drawn independently; their data-generating parameters are known to the experimenter and not to the model. An upstream analysis applies an estimator $m \in \mathcal{M}$ to W_p and reports $y_m(W_p)$. Its full artifact $A^{\text{full}} = (\mathcal{S}(W_p), m, y_m(W_p))$ records the setting, the rule and the value; answer-only compression keeps $(\mathcal{S}(W_p), \varnothing, y_m(W_p))$. A carrier condition c applies an edit T_c that changes only what is retained from the parent analysis, and the downstream model, given the independent child world and a transfer prompt π, responds with

$$R_\theta(c) = F_\theta\big(W_c,\ T_c(A^{\text{full}}),\ \pi;\ \xi\big), \tag{1}$$

where ξ indexes residual variation across repeated requests at temperature 0. A deterministic oracle for W_c labels each response CORRECT, WRONG, OTHER or ABSTAIN, and $q(c)$ is the probability of WRONG, the wrong-estimator rate.

Definition 1 (Carrier attribution)**.** Let `AO_wrong` be the answer-only artifact whose archived value is the wrong estimator's output $\beta_{\text{obs}}(W_p)$. `AO_wrong` is credited with carrying the estimator at margin $\delta \geq 0$ if

$$\Delta_c = q(\texttt{AO_wrong}) - q(c) > \delta \quad \text{for every } c \in \mathcal{C}_0, \tag{2}$$

where $\mathcal{C}_0 = \{\text{SHELL}, \text{NO-ARTIFACT}, \text{RANDOM}, \text{AGREEMENT}, \texttt{AO_correct}\}$.

Each control removes one candidate source while keeping the others. The artifact SHELL states that an analysis exists but withholds its number and method (artifact presence); NO-ARTIFACT keeps the reuse framing without a parent (the instruction); RANDOM archives an unrelated number (generic numeric content); AGREEMENT archives a value on which the two estimators coincide (an answer that identifies no method); and `AO_correct` archives the correct value $\tau(W_p)$. The last contrast, $\Delta_Y = q(\texttt{AO_wrong}) - q(\texttt{AO_correct})$, is the discriminating one: it holds artifact presence, format and plausibility fixed and changes only whether the value identifies the wrong estimator. Attribution is relative to $\mathcal{C}_0$; a source outside it is not excluded by any result here. Table 6 lists every condition used and §B.1 states the identifying assumptions.

Formal operators. In E1 a latent confounder drives both exposure and outcome, $U \sim \mathcal{N}(0,1)$, $X = aU + \epsilon_X$, $Y = \tau X + cU + \epsilon_Y$, so the causal effect is τ while the observational slope is $\beta_{\text{obs}} = \tau + ac/(a^2 + \sigma_X^2)$ (Pearl, 2009; Spirtes et al., 2000). Reporting β_{obs} as the causal effect

is the wrong estimator, and worlds are drawn so that the two values are well separated. In meta-analysis (MA) each world supplies study estimates $\hat{y}_k$ with standard errors σ_k; inverse-variance weighting is correct and the arithmetic mean is wrong, and precisions are heterogeneous so the two differ (Figure 2; derivations in Appendix G.1).

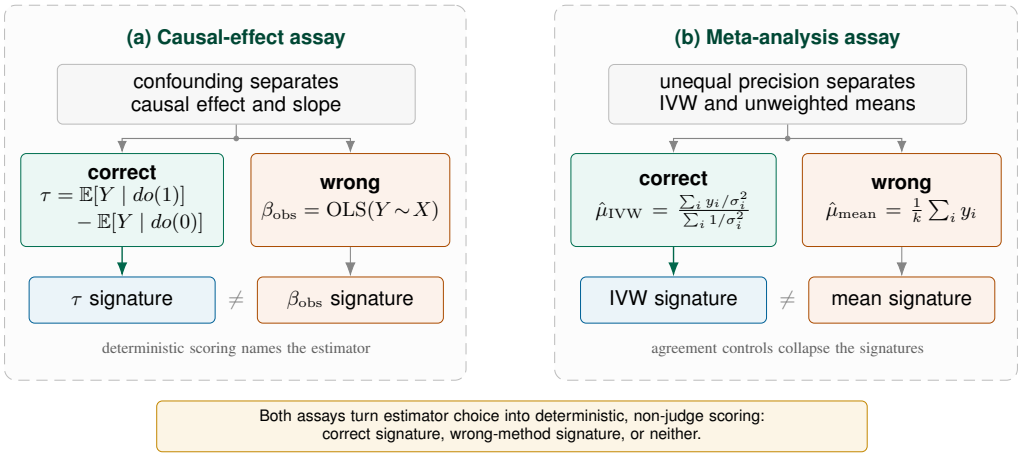


Figure 2: Two formal assays make estimator identity observable. The causal assay separates the interventional effect from the observational slope under latent confounding; the meta-analysis assay separates inverse-variance pooling from the arithmetic mean under unequal precision. Both imply exact signatures. One interpretation-relevant difference is that the E1 artifact displays the quantities needed to identify the wrong estimator, whereas the MA artifact does not.

Estimator recoverability. Let S_c be the non-answer fields visible in condition c and Y_c its archived value. For the two-member candidate set $\mathcal{M}$ the ideal observer

$$\widehat{m}_{\text{IO}}(S_c, Y_c) = \arg\min_{m \in \mathcal{M}} \left| Y_c - y_m(S_c) \right| \tag{3}$$

is defined when every candidate output $y_m(S_c)$ is computable from the visible setting; we then call the estimator *recoverable*. The E1 artifact reports the parent's observational slope, so an archived value equal to it identifies the naive estimator by arithmetic alone. The MA artifact does not display the study-level $(\hat{y}_k, \sigma_k)$, so neither candidate output is computable and $\widehat{m}_{\text{IO}}$ is undefined (Remark 1). We treat this not as a confound to design away but as the hypothesis under test: if the archived number carries the method only through recoverability, then removing recoverability, with the number itself unchanged, must remove Δ_Y.

4 Experimental Design

The artifact and the one sentence we edit. Table 1 gives the E1 answer-only artifact verbatim. The archived value is either the parent's β_{obs} (`AO_wrong`) or its τ (`AO_correct`). The sentence that reports the parent's observational slope takes one of three forms, and every other word of the artifact is identical across them. With the slope *shown*, the estimator is recoverable. With it *withheld*, it is not, but the artifact says that something is missing. With a *non-identifying statistic*—the parent's exposure standard deviation $s_X = \sqrt{a^2 + \sigma_X^2}$, in the same frame and numeric format—it is not recoverable either, and nothing is missing: neither β_{obs} nor τ can be computed from s_X and the rest of the artifact.

Two stages, and only one is instructed. Each downstream run first writes a memo for the new problem. The model receives the parent artifact and a description of the child world, which states the causal graph and the child's observational slope but not τ, under the transfer instruction P3: *"Generate a new methodological memo for the new problem. Carry forward the scientific approach embodied in the parent artifact when analyzing the new setting."* Nothing in it says how an archived number relates to an estimator. The memo either reports the new world's confounded slope as the intervention effect or does not, and a frozen deterministic scorer decides which (Appendix H.6). We

Table 1: The E1 answer-only artifact, verbatim. Angle brackets are filled from the parent world. The archived value $\langle y \rangle$ is $\beta_{\text{obs}}(W_p)$ (`AO_wrong`, the wrong estimator's output) or $\tau(W_p)$ (`AO_correct`, the correct one). The slope sentence takes one of three forms; nothing else differs. A deterministic audit run before any model call confirms this in every world and that s_X never coincides with either estimator's value (§D.4).

`Domain: ⟨domain⟩.` **`[slope sentence]`** `The study also specifies a structural model in which a latent factor U affects both X and Y. Final estimate for the intervention effect: ⟨y⟩. No methodological rule is asserted about how observational associations should be converted into causal effects.`

Slope level	Slope sentence
shown (recoverable)	`The observed regression slope of outcome Y on exposure X is ⟨`$\beta_{\text{obs}}(W_p)$`⟩.`
withheld	`The observed regression slope of outcome Y on exposure X is not reported in this artifact.`
non-identifying	`The observed sample standard deviation of the exposure X is ⟨`$s_X(W_p)$`⟩.`

call this the *memo-stage* rate; it is the model's own estimator choice on the new problem. In the earlier studies, probe problems were then answered from the memo under a contract that *states* the decoding rule: *"If the artifact has no explicit rule but its reported final intervention-effect estimate matches the observed association reported in the same artifact, treat that as evidence for the rule 'report beta_obs as the intervention effect'."* A probe-stage rate is therefore a readout under an instructed rule. We report it beside the memo stage and rest the mechanism on the memo stage (§C.2).

Worlds, executions and inference. Worlds are the replicating unit. The main analyses use 32 worlds with 5 independent endpoint executions each (temperature-0 decoding is not deterministic), a world → execution bootstrap for intervals, and world-level Monte Carlo sign-flip tests (§C.7). Attribution requires all five contrasts of Eq. (2) to clear zero, an intersection–union test (Berger, 1982); we nonetheless report Bonferroni-simultaneous lower bounds at per-contrast $\alpha = 0.01$. The mechanism is estimated by the within-world difference in differences

$$\text{DiD}_s = \Delta_Y^{\text{shown}} - \Delta_Y^{s}, \qquad s \in \{\text{withheld}, \text{nonid}\}, \tag{4}$$

which cancels any nuisance that is constant across the two archive contents within a world, including differences in sentence length. The model is Qwen3.5-27B, and for the replication Qwen3.8-27B, through pinned endpoints with reasoning disabled; every request is logged with its provenance (Appendix H).

Pre-registration and falsification. Every protocol, its estimands and its decision rules were frozen before execution (Appendix C.8, E). Four falsification criteria were fixed before measurement: an operational margin $\delta = 0.25$, scope beyond one model–operator cell, answer specificity and assay validity. Two fired—the margin on the first eight worlds (it holds on 32) and scope—and we report the claim they leave standing rather than the one the design was built to test.

5 Results

5.1 An archived wrong answer transmits its estimator

Table 2 reports the pre-registered 32-world study at both stages. At the memo stage the archived wrong answer makes the model report the confounded slope as the effect of the new problem in 0.988 of 160 executions. The empty shell and no artifact leave the naive estimator rare (0.070, 0.076), so artifact presence does not produce it. An unrelated number, an agreement answer and the correct answer produce it in 0.281, 0.289 and 0.207 of executions, so the model does adopt the naive estimator on its own sometimes; but the wrong value exceeds every control, with simultaneous lower bounds from +0.500 (+0.375 on the 24 worlds generated after the first analysis). The pre-registered probe outcome agrees: the zero-margin conjunction passes with $p < 5 \times 10^{-6}$, and so does the operational margin $\delta = 0.25$ (smallest simultaneous bound 0.298). An earlier discovery analysis on eight held-out worlds had passed at zero margin and failed at δ; we report both evaluations (§D.1, Criterion 1).

Table 2: The archived wrong answer against five matched controls (Qwen3.5-27B/E1/P3, 32 worlds $\times$ 5 executions). Wrong-estimator rate by condition. *Memo*: the model's own estimate for the new problem, frozen deterministic scorer. *Probe*: the pre-registered outcome under the decoding-rule probe. Last column: memo-stage contrast $q(\texttt{AO_wrong}) - q(c)$ with its simultaneous lower bound. Full tables: §D.3.

Condition	Memo	Probe	Memo contrast (sim. LB)
`AO_wrong` (target)	0.988	0.507	—
`ARTIFACT_SHELL`	0.070	0.000	$+0.918$ ($+0.781$)
`NO_ARTIFACT`	0.076	0.012	$+0.912$ ($+0.781$)
`RANDOM_ANSWER`	0.281	0.011	$+0.707$ ($+0.500$)
`AGREEMENT_ANSWER`	0.289	0.000	$+0.699$ ($+0.500$)
`AO_correct` (correct value)	0.207	0.088	$+0.781$ ($+0.594$)
`EXPLICIT_WRONG_METHOD`	1.000	0.999	positive control

What the memo is. A deterministic text scorer assigns each memo one of four mutually exclusive labels—naive-rule use, value transfer without a rule, rejection or abstention, unclassifiable—from a value-echo detector, naive-advocacy patterns with negation filtering and rejection markers (Appendix H.6). The memo is written before any probe and under an instruction that says nothing about decoding, and what is scored is its final estimate for the new problem, checked against the new world's parameters. It is the estimator choice itself, not a rationale offered after one, so the concern that explanations omit what drove an answer (Turpin et al., 2023) does not apply to it. Almost every naive-rule label on an `AO_wrong` memo is earned through the value route: the typical memo restates the parent's template with the new world's slope as its own estimate and carries forward the sentence "No methodological rule is asserted" (§D.6). The model does not write the rule out in words; it carries the deciding value forward.

The effect is larger at the uninstructed memo stage than at the instructed probe stage, so the probe's decoding rule did not manufacture it. And the correct archive is not safe: it induces the wrong estimator in 0.207 of executions. Which worlds leak is predicted by whether the model's own memo rebuilds the naive rule from the visible slope ($r = 0.89$ across worlds), not by how far apart the two estimators are ($r = 0.33$; §D.3). Both point to the same account: the model reads the estimator off the artifact.

5.2 Removing identifiability abolishes the effect

If the archived value carries the method because, next to the displayed slope, it identifies the estimator, then removing that agreement—with the number itself unchanged—must remove the answer-specific effect. Two pre-registered interventions test this inside the positive cell, each crossing the archive content with a slope level of Table 1 and each estimated by Eq. (4).

Slope withheld. The effect vanishes at both stages: memo-stage $\mathrm{DiD}_{\text{withheld}} = +0.894$ and pre-registered probe-stage $\mathrm{DiD}_{\text{withheld}} = +0.511$ ($p < 5 \times 10^{-6}$). Answers move to the correct estimator rather than to abstention (§D.4). All four cells carry the same P3 instruction, archived value, worlds and probes, so instruction-following alone cannot explain the effect: the instruction may be necessary, but it is not sufficient.

A complete artifact that identifies nothing. Withholding the slope also makes the artifact visibly incomplete, and a model that discounts incomplete artifacts would show the same pattern. The complete-artifact control adds the non-identifying level, runs no probes, and pre-registers the memo stage as its primary outcome. Its frozen rules distinguish *identifiability* ($\mathrm{DiD}_{\text{nonid}} > 0$ by the sign-flip test and the interval of Δ_Y^{nonid} reaching zero) from *completeness* (Δ_Y^{nonid} bounded above zero), with partial attenuation in between, and it committed the paper to restating its mechanism in terms of completeness if that outcome occurred.

Table 3 and Figure 3 give the result. With the slope shown, the archived wrong answer makes the memo report the confounded slope in 1.000 of executions, against 0.186 for a correct archive. With the non-identifying statistic in the slope's place and the artifact complete, the two archives can no longer be told apart ($\Delta_Y^{\text{nonid}} = -0.025$), and $\mathrm{DiD}_{\text{nonid}} = +0.838$ $[+0.625, +1.031]$. Under the frozen rules identifiability is supported and completeness not supported; all three registered predictions hold, the withheld result replicates as a pre-registered estimand, and excluding the worlds

Table 3: Removing identifiability, pre-registered, memo stage primary (32 worlds × 5 executions per model). q: share of memos reporting the new world's confounded slope as the effect, for an archive holding the wrong (β_{obs}) or correct (τ) value. Intervals are 95% world → execution bootstrap intervals; p from the two-sided world-level sign-flip test. Full tables: §D.4 and §D.5.

	Qwen3.5-27B		Qwen3.8-27B	
Slope sentence	wrong	correct	wrong	correct
shown	1.000	0.186	0.644	0.118
withheld ("not reported")	0.081	0.186	0.125	0.106
non-identifying statistic	0.237	0.261	0.157	0.088
Δ_Y^{shown}	$+0.814$ $[+0.656, +0.938]$		$+0.525$ $[+0.312, +0.719]$	
Δ_Y^{nonid}	-0.025 $[-0.188, +0.156]$		$+0.068$ $[-0.094, +0.219]$	
$\text{DiD}_{\text{nonid}}$ (primary)	$+0.838$, $p < 5 \times 10^{-6}$		$+0.457$, $p = 1.5 \times 10^{-5}$	
$\text{DiD}_{\text{withheld}}$	$+0.919$, $p < 5 \times 10^{-6}$		$+0.506$, $p < 5 \times 10^{-6}$	
Frozen decision	identifiability supported		identifiability supported	

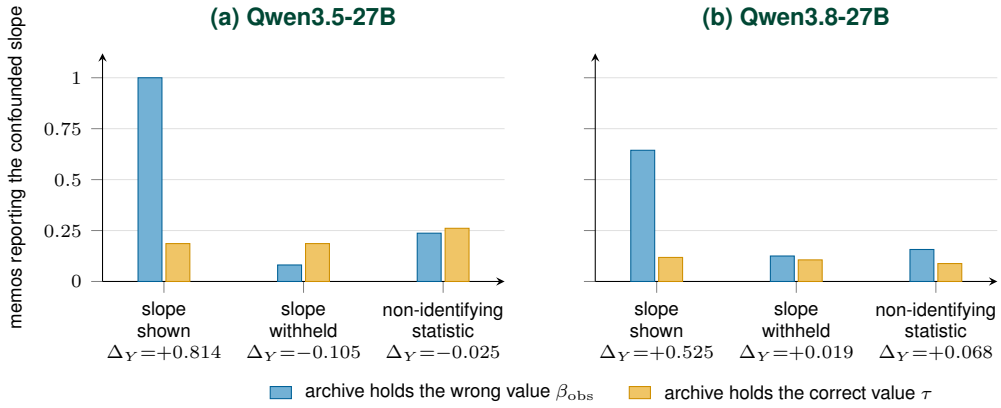


Figure 3: The effect exists only where the archived value identifies the estimator. Share of memos that report the new world's confounded slope as the intervention effect, pre-registered memo-stage outcome, 32 worlds × 5 executions per model. With the slope shown, the archived wrong value makes the model adopt the naive estimator far more often than a correct value does. Withholding the slope, or replacing it with a statistic of the same form that identifies neither estimator (the artifact stays complete), removes the difference, while the archived number is unchanged (Tables 13 and 14).

in which s_X lies near either estimator changes no decision. An independent re-implementation that shares no code with the analysis reproduces every number (Appendix H.4). The component that disappears is exactly the answer-specific one. The slope matters because its agreement with the archived value identifies the estimator, not because its presence makes the artifact look finished.

5.3 The mechanism replicates in a newer model generation

The complete-artifact design was registered again for Qwen3.8-27B, with the same worlds, prompt, scorer, estimands and decision rules. Two conditions were added so that a null could be told apart from a broken assay: `NO_ARTIFACT`, which measures the room the model leaves for an effect, and the explicit-method positive control, which measures whether it responds at all. The protocol registered one prediction: in an eligible model in which the effect replicates, identifiability holds. The model is eligible (positive control 0.818, no-artifact rate 0.118), the effect replicates ($\Delta_Y^{\text{shown}} = +0.525$), and every decision matches (Table 3): $\text{DiD}_{\text{nonid}} = +0.457$ $[+0.188, +0.719]$, identifiability supported, completeness not supported. The newer model reuses the archived estimator less often when the slope is shown, and it reuses it for the same reason (provenance: Appendix H.5). A second family, Mistral Small 3.2 24B, is not eligible under this protocol because its memo almost never commits to a number (§D.5). A registered amendment (A4) appends one line asking the memo to end with a numeric estimate or "not identifiable", scored by parsing that

line, and re-runs Qwen3.5-27B under it as an anchor. The anchor's prediction holds (identifiability supported, $\Delta_Y^{\text{nonid}} = +0.000$), but the line alone shrinks Δ_Y^{shown} to $+0.131$ $[+0.031, +0.250]$: shown-wrong memos, template copies under the original prompt, become reasoning memos that mostly reject the estimator they reconstruct. Under the line Mistral is eligible (positive control 0.730, no-artifact rate 0.086) and shows no effect ($\Delta_Y^{\text{shown}} = +0.044$ $[-0.031, +0.125]$).

5.4 Without identifiability, the number does nothing

The account also predicts where no effect should appear, and none does. In meta-analysis the artifact does not display the study-level inputs either estimator needs, so no archived value can be decoded. In a 2×5 factorial that crosses number presence with graded method cues, the numeric effect is $+0.367$ in the recoverable E1/Qwen cell and $+0.025$ and -0.067 in the two MA cells (Figure 4, §D.7). The contrast is observational, but recoverability was decidable from the artifact schema *before* any outcome was observed. Room matters too: E1 in DeepSeek-V3.2 is recoverable, yet the naive estimator already sits at 0.500 without any artifact. An 18-configuration sweep over three model families, two operators and four transfer prompts shows no numeric effect, which is what the account predicts: only its three E1/Qwen3.5 worlds are both recoverable and in a responsive model with room to move, and the sweep's own frozen probe registers the E1/Qwen effect in the factorial, so the null is not a property of that probe (§D.8).

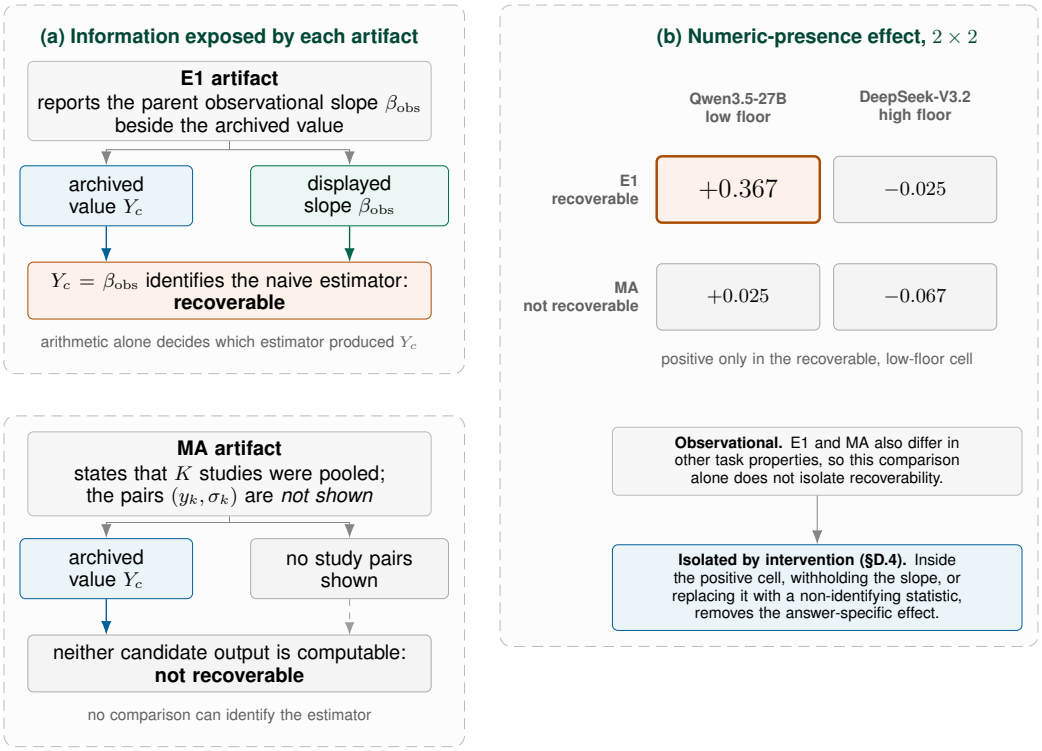


Figure 4: The numeric effect appears only where the artifact is recoverable. Panel (a): the E1 artifact reports the parent observational slope beside the archived value, so their agreement identifies the estimator; the MA artifact shows neither candidate's inputs. Panel (b): the numeric-presence effect over the tested 2×2 operator–model cells is positive only in the recoverable Qwen3.5-27B causal-estimation cell. The operators differ in other ways, so this comparison is observational; the interventions of §D.4 isolate recoverability inside the positive cell.

Table 4 lays every numeric-effect measurement against one conjunctive statement: *an archived number shifts the downstream estimator only when (i) the artifact makes the estimator recoverable and (ii) the model's prior leaves room, that is, the naive rate without any artifact is not already high.* Condition (i) is read from the artifact schema before any model is called, and (ii) from the measured no-artifact rate. The rows below the rule remove (i) by intervention, with model, prompt and worlds

Table 4: Every numeric-effect measurement against the conjunctive statement. (i): the artifact makes the estimator recoverable; (ii): the no-artifact naive rate leaves room. *Stage*: *frozen* = frozen numeric probe that states no decoding rule; *probe* = decoding-rule probe; *memo* = the model's own estimate. Effects are the numeric-presence effect, the smallest simultaneous lower bound of the five contrasts, or the DiD, as named. Full table with section references: Table 16.

Measurement	(i)	(ii)	Stage	Numeric effect
MA, Qwen / DeepSeek, factorial	no	—	frozen	$+0.025$ / -0.067
E1 / DeepSeek-V3.2, factorial	yes	no	frozen	-0.025
E1 / Qwen3.5, factorial	yes	yes	frozen	$+0.367$
E1 / Qwen3.5 / P3, 32 worlds	yes	yes	probe	sim. LB ≥ 0.298
E1 / Qwen3.5 / P3, 32 worlds	yes	yes	memo	sim. LB $\geq +0.500$
E1 / Qwen3.5 / Mistral, A4	yes	yes	memo	Δ_Y $+0.131$ / $+0.044$ (n.s.)
E1 / Qwen3.5 / P3, slope withheld	**no**	yes	probe	DiD $+0.511$
E1 / Qwen3.5 / P3, slope withheld	**no**	yes	memo	DiD $+0.894$
E1 / Qwen3.5 / P3, non-identifying	**no**	yes	memo	DiD $+0.838$
E1 / Qwen3.8 / P3, non-identifying	**no**	yes	memo	DiD $+0.457$
E1 / Qwen3.5 / Mistral, A4, non-id.	**no**	yes	memo	Δ_Y $+0.000$ / -0.019

fixed; the rows above are observational and differ in more than (i) and (ii). No row shows an effect without (i); the conditions are not sufficient, since Mistral under A4 meets both.

6 Discussion

Not a hidden channel. Unlike trait transmission during distillation (Cloud et al., 2026), the method here is fixed in the model's own observable output, and it appears only where a rule-based reader could identify the estimator by arithmetic.

What this means for compressed handoffs. Three untested implications follow. Stripping numbers is the wrong remedy: an archived number is inert once its provenance is gone. The danger is partial provenance, enough to re-identify a method but not to show it was wrong, so a handoff should carry the method with its assumptions (Moreau & Groth, 2013) or no method-identifying context. And a correct archive is not safe: in our assay it leaked the wrong estimator in 10 of 32 worlds, predicted by whether the model's memo adopted the naive rule and not by the numeric geometry of the world, so filtering handoffs on answer correctness would not have prevented the failure we measured.

7 Limitations and conclusion

Limitations. All positive evidence comes from one model family (two generations), one operator and one transfer instruction (P3); whether it appears without that instruction or at frontier scale is untested. GLM-4-9B does not respond to the positive control and DeepSeek-V3.2 leaves no room, so neither is a test; Mistral Small 3.2 24B, under a required final estimate, is eligible and shows no effect, so the effect is heterogeneous across families: recoverability is necessary for it, not sufficient. The comparison of the archived wrong answer against the five controls at the memo stage is a post hoc reading of pre-registered runs, whereas the two interventions that locate the mechanism are confirmatory. No intervention displays a number equal to the archived value that is not a slope, so estimator identification is not separated from numeric agreement as such (Section E lists what remains open).

Conclusion. An archived wrong answer can make a downstream model use the wrong estimator on a new problem, specifically because the answer is wrong. But the number does not carry the method by itself. It carries it the way a receipt carries a price: next to the slope it was computed from, their agreement identifies the estimator, and the model does the arithmetic unprompted. Remove that agreement, by withholding the slope or by replacing it with a statistic that identifies nothing, and the effect is gone, in two model generations and two output formats. Recoverability is necessary, not sufficient: a second family does not take the recoverable estimator. The practical form of the finding is not "do not archive numbers" but "an archived number is inert exactly when its provenance is gone, and dangerous exactly when enough provenance survives to re-identify the method."

AI USE STATEMENT

The language models evaluated in this work are the experimental systems under study. Separately, generative AI tools were used for literature search, feedback on methodology and experiments, interpretation of results, software and LaTeX assistance, and drafting and editing of manuscript text. They were never used as judges: problem instances, estimator identities, scores and every reported number are determined by code and regenerated from the recorded outputs. The authors reviewed all AI-assisted text and code against the literature, the experimental records and the test suite, and take responsibility for the content of the paper.

ETHICS STATEMENT

The study uses synthetic problem instances and model outputs and involves no human participants, private data or deployment intervention. The main risk is extrapolating the controlled findings to real scientific workflows; the claims are restricted to the evaluated configurations, and the untested design implications are labelled as such.

REPRODUCIBILITY STATEMENT

Every protocol was frozen before execution. The appendix gives the formal task definitions, all carrier conditions, both prompts, the scoring rules, the statistical procedures, independent re-implementations of the central analyses and request-level provenance. The anonymous supplement contains the deterministic world and probe generators, versioned configurations, the immutable request/response logs, the scorers, the analysis scripts that produce every number in this paper, and their tests.

REFERENCES

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10308–10330, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.555/.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24 (4):295–300, 1982. doi: 10.2307/1267823.

Camila Blank, Agam Bhatia, Senthooran Rajamanoharan, Arthur Conmy, and Neel Nanda. Subliminal learning is steering vector distillation. *arXiv preprint arXiv:2606.00995*, 2026.

Akash Bonagiri, Devang Borkar, Gerard Janno Anderias, Setareh Rafatirad, and Houman Homayoun. CausalFlow: Causal attribution and counterfactual repair for LLM agent failures. *arXiv preprint arXiv:2605.25338*, 2026.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

Edward Y. Chang. T3: Benchmarking sycophancy and skepticism in causal judgment. *arXiv preprint arXiv:2601.08258*, 2026.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Sören Mindermann, Jacob Hilton, Samuel Marks, and Owain Evans. Language models transmit behavioural traits through hidden signals in data. *Nature*, 652(8110):615–621, 2026. doi: 10.1038/s41586-026-10319-8. URL https://doi.org/10.1038/s41586-026-10319-8.

Karl Cobbe et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993. doi: 10.1201/9780429246593.

Yiyang Feng, Biddut Sarker Bijoy, Niranjan Balasubramanian, and Jiawei Zhou. Break it down, pass it on: Cross-task skill transfer in LLM agents. *arXiv preprint arXiv:2608.20274*, 2026.

Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-consuming generative models with curated data provably optimize human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-3256.

Aryo Pradipta Gema, Neel Rajani, Rohit Saxena, Wai-Chung Kwan, and Pasquale Minervini. Chain-of-thought faithfulness of reasoning models varies with where and how preference cues are delivered. *arXiv preprint arXiv:2608.29464*, 2026.

Isaia Gisler, Zhonghao He, and Tianyi Qiu. You didn't have to say it like that: Subliminal learning from faithful paraphrases. *arXiv preprint arXiv:2603.09517*, 2026.

Dongxin Guo, Jikun Wu, and Siu Ming Yiu. Model collapse as cultural evolution. *arXiv preprint arXiv:2605.23054*, 2026.

Ethan Hadley and Eren Gultepe. Subliminal learning is non-semantic distillation. *arXiv preprint arXiv:2608.05734*, 2026.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

Jie Huang et al. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

Zhen Huang, Zhihuang Liu, Weijia Shi, Yifan Yang, Weishang Wu, and Zhiping Cai. When coordination becomes a threat: Communication attacks in LLM-controlled multi-robot systems. *arXiv preprint arXiv:2608.06830*, 2026.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Zhijing Jin et al. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ryo Kamoi, Yusen Zhang, Nan Zhang, et al. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024. doi: 10.1162/tacl_a_00713.

Andrej Leban and Yuekai Sun. CausalDS: Benchmarking causal reasoning in data-science agents. *arXiv preprint arXiv:2607.08093*, 2026.

Jungseob Lee, Seungyoon Lee, Suhyune Son, Dongyub Jude Lee, Sungbin Han, Sugyeong Eo, and Heuiseok Lim. Answer-conditioned chains of thought degrade verifiable-reasoning distillation in large language models. *arXiv preprint arXiv:2607.14552*, 2026.

Junchi Liao. Auditing provenance sensitivity in LLM agent action selection. *arXiv preprint arXiv:2607.20827*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

Tianqianjin Lin, Xi Zhao, Xingyao Zhang, Rujiao Long, Yi Xu, Zhuoren Jiang, Wenbo Su, and Bo Zheng. RAVR: Reference-answer-guided variational reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18757–18775, 2026a. doi: 10.18653/v1/2026.findings-acl.936.

Xiaofeng Lin, Yukai Yang, Daniel Guo, Sahil Arun Nale, Charles Fleming, and Guang Cheng. Context-fractured decomposition attacks on tool-using LLM agents: Exploiting artifact provenance gaps. *arXiv preprint arXiv:2606.09084*, 2026b.

Jiaxu Lou and Yifan Sun. Anchoring bias in large language models: an experimental study. *Journal of Computational Social Science*, 9:11, 2026. doi: 10.1007/s42001-025-00435-2.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Jingkun Luo and Da-Tian Peng. Success is not self-explanatory: Auditing success provenance in agent evaluation. *arXiv preprint arXiv:2607.24054*, 2026.

James G. MacKinnon and Matthew D. Webb. The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, 21(2):114–135, 2018. doi: 10.1111/ectj.12107.

Tamas Madl. Channel location constrains the auditability of subliminal learning. *arXiv preprint arXiv:2606.22019*, 2026.

Arash Marioriyad, Shaygan Adim, Nima Alighardashi, Mahdieh Soleymani Banghshah, and Mohammad Hossein Rohban. Unspoken hints: Accuracy without acknowledgement in LLM reasoning. *arXiv preprint arXiv:2509.26041*, 2025.

Luc Moreau and Paul Groth. PROV model primer. W3C working group note, World Wide Web Consortium, 2013. URL `https://www.w3.org/TR/prov-primer/`.

Akanksha Narula, Mofasshara Binte Rafique, and Laurent Bindschaedler. The invisible lottery: How subtle cues steer algorithm choice in LLM code generation. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2606.04057.

Todd Nief, Harvey Yiyun Fu, Mark Muchane, and Ari Holtzman. Subliminal learning is a LoRA artifact. *arXiv preprint arXiv:2606.00831*, 2026.

Joon Sung Park et al. Generative agents: Interactive simulacra of human behavior. In *UIST*, 2023. doi: 10.1145/3586183.3606763.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009. doi: 10.1017/cbo9780511803161.

Qwen Team. Qwen3.5-27B. Hugging Face model card, 2026. URL `https://huggingface.co/Qwen/Qwen3.5-27B`.

Xuan Ren, Weiqi Zhai, Tianle Pu, Yihua Zhu, Hu Wei, and Bing Zhao. Right answer, wrong method: Shortcut hacking misleads the evaluation of LLM reasoning on frontier science benchmarks. *arXiv preprint arXiv:2608.02442*, 2026.

David M. Ritzwoller, Joseph P. Romano, and Azeem M. Shaikh. Randomization inference: Theory and applications. *arXiv preprint arXiv:2406.09521*, 2024.

Timo Schick et al. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024. doi: 10.1038/s41586-024-07566-y.

Renan Souza, Timothy Poteet, Brian Etz, Daniel Rosendo, Amal Gueroudji, Woong Shin, Prasanna Balaprakash, and Rafael Ferreira da Silva. LLM agents for interactive workflow provenance: Reference architecture and evaluation methodology. In *WORKS at the ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, 2025. doi: 10.1145/3731599.3767582.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000. doi: 10.7551/mitpress/1754.001.0001.

Team GLM et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.04388.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13894–13908, 2024. doi: 10.18653/v1/2024.findings-acl.826.

Jonathan Uesato et al. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Saranyan Vigraham. When context hurts: The crossover effect of knowledge transfer on multi-agent design exploration. *arXiv preprint arXiv:2605.04361*, 2026.

Yuhui Wang, Changjiang Li, Guangke Chen, Jiacheng Liang, and Ting Wang. Reasoning or retrieval? A study of answer attribution on large reasoning models. In *International Conference on Learning Representations (ICLR)*, 2026.

Yang Wu, Yifan Zhang, Yiwei Wang, Yujun Cai, Yurong Wu, Yuran Wang, Ning Xu, and Jian Cheng. Answer-centric or reasoning-driven? uncovering the latent memory anchor in LLMs. *arXiv preprint arXiv:2506.17630*, 2025.

Yizhe Xie, Congcong Zhu, Xinyue Zhang, Tianqing Zhu, Dayong Ye, Minfeng Qi, Huajie Chen, and Wanlei Zhou. From spark to fire: Modeling and mitigating error cascades in LLM-based multi-agent collaboration. *arXiv preprint arXiv:2603.04474*, 2026.

Guodong Xu. Calibrating criterion revision in LLM agents: Failure modes and a trace-anchored protocol. *arXiv preprint arXiv:2608.20729*, 2026.

Shunyu Yao et al. ReAct: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

Matej Zećević et al. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023.

APPENDIX OVERVIEW

The appendix is the complete technical account on which the main text draws: related work, problem formulation, method, every result with its tables and figures, the registered-and-open experiments and limitations, followed by formal derivations, provenance, independent re-implementations and per-world tables. Section references in the main text point into it.

A RELATED WORK

We ask one question of each line of work: *which candidate carrier does it intervene on, and which competing carrier does that intervention leave unexcluded?* Reading the literature this way is not fault-finding — most of these studies were not built to answer an attribution question — but it locates precisely which control our design adds and why the addition is not free. Table 5 summarises; the paragraphs give the reasoning.

A.1 ANSWER CUES, HINTS, AND PROCESS VALIDITY

Answers and hints steer LLM reasoning within a problem (Wu et al., 2025; Marioriyad et al., 2025; Lou & Sun, 2026; Lin et al., 2026a; Lee et al., 2026), and models recover aspects of missing task structure from incomplete artifacts (Balepur et al., 2024). These designs establish that answer-like signals influence behaviour. What they do not run — and what our design turns on — is a *correct*-answer control: if a correct archived answer induces the wrong estimator at the same rate as a wrong one, the channel is generic answer conditioning rather than estimator-diagnostic content. We measure exactly that contrast, and it is the one that limits our claim (§D.1).

Process-supervision and shortcut-reasoning work establishes a complementary boundary: a correct final answer need not imply a valid method (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2024; Huang et al., 2023; Kamoi et al., 2024; Tyen et al., 2024; Ren et al., 2026). That literature audits whether the process behind an answer is valid. We ask whether an earlier answer supplies enough information to select a method on a *new* problem. The distinction is within-solution process validity versus cross-artifact method attribution.

Whether a model says what moved it. A separate line asks not whether the reasoning is correct but whether it *reports its own causes*. Chain-of-thought explanations can systematically omit the feature that actually determined the answer (Turpin et al., 2023), and the omission is not uniform: verbalised commitment is lower when a cue arrives through a tool return than through a user message, and lower for implicit than for explicit cues (Gema et al., 2026). That is precisely our regime — an implicit cue delivered inside a machine-readable artifact — so the prior expectation is that a model influenced by an archived number would *not* say so. We report the opposite (§D.6), and §D.6 also states why our intermediate memo is not exposed to the standard unfaithfulness objection in the first place: it is not a rationalisation emitted after a decision, it is an artifact generated *before* the scored responses and conditioning them.

A.2 TRANSMISSION THROUGH GENERATED DATA, AND THROUGH CONTEXT

Training-time studies show that synthetic-data feedback propagates model properties across generations (Shumailov et al., 2024; Ferbach et al., 2024; Guo et al., 2026). Most directly, subliminal learning shows that number sequences transmit behavioural traits during distillation even when semantically unrelated to those traits (Cloud et al., 2026). A 2026 cluster has since localised the mechanism: the transmitted signal is well approximated by a steering vector (Blank et al., 2026), is partly a low-rank adaptation artifact (Nief et al., 2026), survives faithful paraphrase (Gisler et al., 2026), and its auditability depends on where in the model the channel sits (Madl, 2026). That last result is the closest conceptual neighbour to carrier attribution, because it also asks *which channel* rather than *whether transmission occurred* — but it asks it of weights during distillation.

Our setting moves the question to inference-time artifact reuse with no parameter update and a *visible* numeric result. That difference cuts both ways and we state the unflattering half: because the carrier is visible and the estimator is determined by the visible setting in E1, our positive cell is not evidence of a hidden channel at all (§B.4). It is evidence that a model will do the arithmetic.

Table 5: What each adjacent design intervenes on, and what it leaves unexcluded. The column that separates this work is not "matched controls" — several of these have them — but the conjunction of a *correct-answer* control with cross-world reuse and exact estimator identity. Our own row records the scope restriction the evidence actually supports, not the design's nominal reach.

Work	Intervenes on	Ground truth	Correct-answer control	Left unexcluded
Answer/hint steering (Wu et al., 2025; Lou & Sun, 2026; Marioriyad et al., 2025; Wang et al., 2026)	answer presence, within one problem	task accuracy	no	whether the effect is answer-*specific* or generic answer conditioning
Semantic algorithm steering (Narula et al., 2026)	incidental semantic cues	algorithm family	n/a (no answer)	numeric carriers; cues are the manipulation, not a calibration scale
Success provenance (Luo & Peng, 2026)	CLEAN/GOLD/ SHAM value substitution	task success	GOLD, but success is the outcome	whether the *method* changed, as opposed to the answer
Artifact abduction (Balepur et al., 2024)	removal of the question	option choice	no	cross-artifact method transfer; the abduction is within one item
Subliminal learning (Cloud et al., 2026; Hadley & Gultepe, 2026)	training data, filtered	behavioural trait	n/a	inference-time reuse; requires a gradient update
Cross-task skill transfer (Feng et al., 2026; Vigraham, 2026)	induced skill text across tasks	downstream accuracy	no	which *component* of the carried artifact is responsible
Reasoning-trace faithfulness (Turpin et al., 2023; Gema et al., 2026)	cue presence and delivery channel	whether the cue is verbalised	no	what the cue transmits, as opposed to whether it is reported
This work	**six matched carriers, cross-world**	**exact estimator identity**	**yes, and it binds**	*scope*: one model–operator–prompt cell; E1 value is ideal-observer recoverable (§B.4)

A.3 Interventional attribution, agent artifacts, and skill reuse

A methodological principle we adopt appears in work treating behavioural provenance as interventional rather than observational. Counterfactual trace repair intervenes on candidate failure steps (Bonagiri et al., 2026); success-provenance audits compare matched substitutions (Luo & Peng, 2026); criterion-revision and provenance-sensitivity studies intervene on claimed traces or source authority (Xu, 2026; Liao, 2026). AcquaBench's SHAM condition — structure and exposure preserved, value replaced by a matched incorrect one — is the direct ancestor of our attribute-matched decoy. Its outcome is task success; ours is estimator identity, which is why a correct-answer control is available to us and not to it.

A second neighbourhood is artifact- and memory-mediated reuse between agent steps. Persistent generated artifacts shape later calls and multi-agent decisions (Cemri et al., 2025; Lin et al., 2026b; Huang et al., 2026); induced skills transfer across tasks and can *harm* the agent that retrieves them (Feng et al., 2026); injected context helps or hurts conditionally (Vigraham, 2026); and artifact-centric workflows let an early error crystallise into constraints that later steps build on (Xie et al., 2026). These share our motivating picture. They differ in the question asked of it: they measure whether carried content changes downstream *performance*, we ask which *component* of the carried artifact accounts for a downstream *method*, and we require the target component to beat five matched alternatives before crediting it.

Formal causal-reasoning benchmarks use symbolic data-generating processes for objective ground truth (Jin et al., 2023; Zečević et al., 2023; Leban & Sun, 2026; Chang, 2026). We use that formalism to identify which estimator a model applies, and pair it with a meta-analysis operator specifically so that one operator is ideal-observer recoverable and one is not (§B.4) — a contrast that, in hindsight, carries more of the paper's information than the effect sizes do.

B Problem Formulation

B.1 Artifacts and carriers

Let a parent world $W_p \sim P_W$ and a child world $W_c \sim P_W$ be drawn independently. Their data-generating parameters are known to the experimenter and not to the model. A scientific analysis applies an estimator-selection rule $m \in \mathcal{M}$ to W_p and reports $y_m(W_p) \in \mathbb{R}$. Its full artifact is

$$A^{\text{full}}(W_p, m) = (\mathcal{S}(W_p),\ m,\ y_m(W_p)),$$

where $\mathcal{S}$ describes the setting and m records the rule that justifies the estimate. Answer-only compression preserves the value and removes the rule:

$$A_{\text{AO}}(W_p, m) = T_{\text{AO}}\big(A^{\text{full}}(W_p, m)\big) = (\mathcal{S}(W_p),\ \varnothing,\ y_m(W_p)).$$

For carrier condition $c \in \mathcal{C}$ the intervention T_c changes only the information retained from the parent analysis. A model with parameters θ receives the resulting artifact, the independent child world, and a transfer prompt π, producing

$$R_\theta(c, \pi) = F_\theta\big(W_c,\ T_c(A^{\text{full}}),\ \pi;\ \xi\big)\,, \tag{5}$$

where ξ indexes residual variation across repeated requests to the same model–provider endpoint under fixed decoding parameters (temperature 0; every provider-exposed and unexposed field is enumerated in Appendix H). A deterministic oracle for W_c maps the response to an estimator signature,

$$\phi_{W_c}(R_\theta) \in \{\text{CORRECT}, \text{WRONG}, \text{OTHER}, \text{ABSTAIN}\}, \tag{6}$$

the wrong-signature outcome is $Z_c = \mathbf{1}\{\phi_{W_c}(R_\theta(c, \pi)) = \text{WRONG}\}$, and the conceptual superpopulation quantity is $q(c, \pi) = \mathbb{E}_{W_p, W_c, \xi}[Z_c]$, written $q(c)$ when π is fixed. Each empirical analysis instantiates this target over its own worlds, probes, transfer templates and endpoint executions; §C.7 states which layer each one integrates and which it holds fixed, because the analyses in this paper do *not* share an estimand.

Conceptually T_c intervenes on five candidate sources: task context, transfer prompt, artifact presence, numeric answer, and semantic method information. The matched design varies one candidate while holding parent and child worlds fixed, and therefore identifies attribution *among the tested alternatives* — never against alternatives nobody ran.

Definition 2 (Carrier attribution)**.** Answer-specific carrier attribution at margin $\delta \geq 0$ requires the downstream estimator signature under `AO_wrong` to exceed *every* tested matched alternative by more than δ, including artifact-presence, no-artifact, generic-number, agreement-answer and correct-answer controls. We distinguish the existence claim ($\delta = 0$) from the pre-registered operational decision claim ($\delta_{\text{practical}} = 0.25$); §C.6 gives the five contrasts and §D.1 and §D.3 report both claims in the only cell where either is available.

The definition implements an identification principle. A positive wrong-signature rate is insufficient, because the same downstream signature arises under matched changes to task context, artifact presence, or non-diagnostic answer content. Attribution therefore requires intervention on each competing source with parent and child worlds held fixed.

What has to hold for Δ_c to be read as attribution. We state the assumptions rather than leave them implicit, and mark which are enforced by construction and which are genuine assumptions.

- **(A1)** *Matched support (by construction).* W_p and W_c are held fixed across the conditions entering a contrast, and T_c edits exactly one field of the artifact tuple. Conditions are instantiated by deterministic code from the same world, so this is checkable rather than assumed (Appendix G.3).
- **(A2)** *Oracle validity (verified).* ϕ_{W_c} assigns WRONG if and only if the response implements the incorrect estimator. Agreement probes — inputs on which both estimators coincide — audit this directly and score $7{,}665/7{,}680 = 0.998$.
- **(A3)** *No interference (by construction).* Responses are generated independently at temperature 0 with no state shared across conditions; every provider-level field is enumerated in Appendix H.

(A4) *Assay responsiveness (checked per cell, and it fails in one).* A negative contrast is informative only where the positive control transmits; GLM-4-9B is not responsive and its contrasts are therefore excluded from negative carrier claims (§C.7).

(A5) *Relative exhaustiveness (assumed, and deliberately weak).* Attribution is *among* $\mathcal{C}_0$. A carrier not represented in $\mathcal{C}_0$ is not excluded by any result here. This is the one assumption we cannot discharge, and it is why Definition 1 is written as a comparison against a named control set rather than as a claim about the number as such.

(A5) is what makes the correct-answer control load-bearing: without it in $\mathcal{C}_0$, generic answer conditioning sits outside the tested set and every positive result would be consistent with it.

Terminology. "Wrong-method signature" ("wrong signature" in tables) denotes $\phi = \text{WRONG}$. "Estimator recurrence" and "method induction" both name the behavioural fact that a wrong signature appears. "Carrier attribution" and "answer-specific reconstruction" both name the stronger claim of Definition 1. "Carrier condition" and "artifact condition" both denote an element of $\mathcal{C}$.

B.2 Induction versus reconstruction

Method induction occurs when a recognisable estimator signature appears without the rule being supplied explicitly: $q(c) > 0$ establishes signature-level induction and $q(c) - q(\text{NO_ARTIFACT}) > 0$ an artifact-conditioned increment. Neither identifies which component produced the behaviour.

Answer-specific reconstruction is stronger. With $\mathcal{C}_0$ the no-artifact, artifact-shell, random-number, agreement-answer and correct-answer controls, reconstruction requires $q(\texttt{AO_wrong}) - q(c) > 0$ for every $c \in \mathcal{C}_0$, yielding two claims:

$$H_{\text{existence}} : \ \Delta_c > 0 \quad \forall c \in \mathcal{C}_0, \tag{7}$$

$$H_{\text{practical}} : \ \Delta_c > \delta_{\text{practical}} = 0.25 \quad \forall c \in \mathcal{C}_0. \tag{8}$$

Induction asks whether a method-like behaviour appears; reconstruction asks whether the answer carried it. In this paper the first holds broadly and the second holds in one cell: Eq. 7 in every analysis of that cell, Eq. 8 at 32 worlds but not at 8 (Criterion 1 records both verdicts).

B.3 Formal operators

Both operators make estimator identity exactly observable, so downstream choice is scored against deterministic ground truth with no LLM judge (Appendix G.1).

Causal estimation (E1). In a linear-Gaussian structural causal model with latent confounding (Pearl, 2009; Spirtes et al., 2000), $U \sim \mathcal{N}(0,1)$, $X = aU + \epsilon_X$, $Y = \tau X + cU + \epsilon_Y$, the causal effect is τ whereas the observational slope is $\beta_{\text{obs}} = \tau + ac/(a^2 + \sigma_X^2)$. Reporting β_{obs} as the causal effect is the identifiable incorrect estimator. Worlds are drawn so the two are numerically well separated.

Meta-analysis (MA). Each world supplies K study estimates $\hat{y}_k$ with standard errors σ_k. Inverse-variance weighting $\hat{y}_{\text{IVW}} = \sum_k w_k \hat{y}_k / \sum_k w_k$, $w_k = 1/\sigma_k^2$, is correct; the arithmetic mean is the wrong estimator. Precisions are heterogeneous, so the two differ and the wrong signature is exactly detectable.

B.4 Estimator recoverability, and why the two operators are not interchangeable

Carrier attribution and carrier *recoverability* are distinct, and conflating them is the single largest interpretive risk in this design. Let S_c be exactly the non-answer fields visible in condition c and Y_c its archived value. For a finite candidate set $\mathcal{M}$, define the ideal observer

$$\widehat{m}_{\text{IO}}(S_c, Y_c) = \arg\min_{m \in \mathcal{M}} \big| Y_c - y_m(S_c) \big|, \tag{9}$$

defined when every candidate output $y_m(S_c)$ is computable from the visible setting; ties are non-identifiable.

Remark 1 (E1 is recoverable by construction; MA is not)**.** The E1 artifact explicitly reports the parent observational slope. Within the two-estimator candidate set, an `AO_wrong` value equal to that slope is therefore identified by Eq. 9 using arithmetic alone, with no model required. The MA artifact does not display the study-level (y_k, σ_k) pairs, so neither candidate output is generally computable from S_c and $\widehat{m}_{\mathrm{IO}}$ is undefined.

Two consequences run in opposite directions, and both belong in the reader's hands before any number is quoted. Weak attribution in E1 could not be read as absence of channel information, since a rule-based observer can use the displayed slope. And — the direction that matters for our actual result — *strong* attribution in E1 cannot be read as transmission through an opaque numeric channel, because the estimator is already determined by the visible setting. §D.7 shows the positive cell is E1, the MA numeric effect is at zero, and the memo evidence of §D.6 is consistent with decoding rather than with a hidden channel. We report the construction-level baseline rather than a learned classifier, and restrict it to the stated candidate set.

C METHOD

C.1 DESIGN

A parent scientific problem produces an artifact, which is selectively compressed before being passed to an independently generated child problem. We vary the information the artifact preserves and score the estimator selected downstream. Parent and child worlds are independent, so the archived numerical answer cannot mechanically solve the child problem.

Two probe regimes are used and they are not interchangeable. *Fixed-bank* runs score artifact-conditioned responses directly on a frozen eight-probe discriminative bank (Appendix G.2); this regime carries the paper's widest configuration coverage and its null result. *Sampled-probe* runs add an intermediate carrier-reconstruction step,

$$\text{artifact} \;\rightarrow\; \text{carrier memo} \;\rightarrow\; \text{child probes},$$

where the memo records the model's intermediate reconstruction of what, if anything, should be carried forward, and score 32 world-specific discriminative probes plus 16 agreement probes from one realised memo per world–condition. The repeated-execution design samples the endpoint 10 times for each of six conditions in the same eight Qwen3.5-27B/E1/P3 worlds, producing 480 world–execution–condition units.

C.2 TWO STAGES, AND WHAT EACH STAGE IS TOLD

Every positive measurement in this paper runs the sampled-probe pipeline, and its two stages receive different instructions. We quote both, because the difference decides what each stage can show.

Memo stage: an estimator choice on the new problem, uninstructed. The model receives the parent artifact and a description of the new problem, which states the causal graph and the observed slope β_{obs} of the new world but not τ, which is not identifiable from what is shown. Under P3 the system instruction reads: *"You receive a parent methodological memo and a new problem description. Generate a new methodological memo for the new problem. Carry forward the scientific approach embodied in the parent artifact when analyzing the new setting."* Nothing in it says how an archived number relates to an estimator. The memo either reports the new world's confounded slope as the intervention effect or does not, and the frozen deterministic memo scorer (Appendix H.6) decides which, with no LLM judge. We call its rate the *memo-stage* wrong-estimator rate. It is the downstream model's own estimator choice on the new problem, which is the construct this paper is about.

Probe stage: a readout under a stated decoding rule. Each probe presents a further problem and the memo, never the parent artifact, under the contract `answer_only_method_inference`. Its system instruction reads, in part: *"If the artifact has no explicit rule but its reported final intervention-effect estimate matches the observed association reported in the same artifact, treat that as evidence for the rule 'report beta_obs as the intervention effect'. If the artifact lacks a final*

estimate, lacks a reported observed association, or the reported value is method-ambiguous, use the structural causal coefficient in the current probe." The recoverability rule of §B.4, and the fallback for an artifact without an observed association, are therefore *stated* at this stage. A probe-stage rate measures whether the memo carries a decodable estimator when the decoding is instructed, not whether the model decodes on its own.

Which stage carries which claim. The frozen protocols name the probe-stage rate as the primary outcome, and every probe-stage result is reported with its frozen verdict. The mechanism claims of this paper rest on the memo stage. For the Study 2 and slope-withheld runs its outcome is computed from the archived runs by the frozen memo scorer; no protocol named it as primary there, so those memo-stage numbers are labelled *post hoc*. The complete-artifact control (§D.4) was registered afterwards with the memo stage as its primary outcome and runs no probes at all; its numbers are confirmatory. The fixed-bank sweep and the 2×5 factorial score every carrier condition with the frozen numeric probe instead, which also presents the memo but maps only *explicitly stated* rules to estimators (*"If your memo says to use the regression coefficient, use* β_{obs}*. If it says to adjust for confounding, use the back-door formula"*) and states no rule for decoding an archived number. The sweep's manifest setting `no_estimate_aware` changes the probe only for no-estimate-policy conditions, none of which is reported here. Table 7 records which probe each analysis uses.

C.3 Matched carrier interventions

Table 6 is the complete manifest of conditions used in *this* paper, and it is exhaustive: no condition is scored anywhere in §D that does not appear here. Two design principles organise it. First, every control paired with the target `AO_wrong` artifact preserves artifact presence, numeric form and answer-likeness while removing exactly the method-diagnostic content. Second, the remaining conditions grade the channel from nothing, to artifact presence without content, to generic numbers, to semantic method cues, to an explicit estimator rule.

C.4 Transfer prompts, and what P3 asks for

We vary how strongly the downstream prompt asks the model to preserve continuity with the parent analysis: P0 supplies the artifact as background context, P1 gates on artifact presence, P2 requests method continuity, and **P3 explicitly asks the model to carry forward the scientific approach embodied in the parent artifact** (quoted in §C.2); five paraphrases, one of which asks it to adopt the parent's estimator-selection rule, are tested in §D.12.

This must be stated plainly, because it bounds every positive result in the paper. Under P3, when the only surviving parent information is a number, a model that infers the rule from that number is doing something the instruction asked for. P3 is therefore the condition under which the numeric channel is *most* favourable to detection, and it is the condition in which all our positive evidence sits: the fixed-bank sweep covers P0–P3 and is null throughout (§D.8); the sampled-probe and repeated-execution analyses were run at P3 only. We do not have a sampled-probe measurement at P0–P2, so we cannot say whether the effect survives without the reuse instruction; the fixed-bank response across all four prompt levels is in Appendix M.1. This is registered as unrun (§E) and recorded as a fired falsification criterion (§C.8). Five paraphrases of the P3 instruction test surface wording within P3 (§D.12); they do not address this point, and we do not present them as if they did.

C.5 Deterministic scoring

For every probe the correct and incorrect estimators imply known numerical outputs. A response receives the wrong-method signature if it matches the incorrect estimator within tolerance 10^{-6} and the correct signature if it matches the oracle estimator. Agreement probes cover inputs on which the two estimators return indistinguishable values; they carry no diagnostic signal and validate probe scoring. Deterministic oracle matching provides all attribution scores; there is no LLM judge anywhere in the pipeline, including in the memo micro-assay of §D.6.

Table 6: The sixteen carrier conditions used in this paper. The target intervention is `AO_wrong`: the archived wrong answer (the observational slope for E1, the arithmetic mean for MA) without the method. "Regime" records where a condition was actually scored, so that no contrast in §D can be read as covering a regime it was not run in. Conditions implemented in the codebase but not reported here are listed in Appendix G.3.

Condition	Isolated source / role	Regime
Baselines		
`NO_HISTORY`	no archived analysis; task prior + prompt only	fixed bank
`NO_ARTIFACT`	no parent artifact, reuse framing present	fixed, sampled, repeated
`ARTIFACT_SHELL`	analysis stated to exist, number and method withheld	fixed, sampled
Answer controls		
`AO_wrong` **(target)**	wrong estimate, method withheld	fixed, sampled, repeated
`AO_correct`	correct estimate, method withheld; the binding control	fixed, sampled, repeated
`RANDOM_ANSWER`	unrelated archived number	fixed, sampled, repeated
`MATCHED_RANDOM`	decoy matched on sign, magnitude bin, decimal precision, artifact syntax	sampled only
`AGREEMENT_ANSWER`	answer on which both estimators coincide	fixed, sampled, repeated
Recoverability manipulation (§D.4)		
`AO_wrong`, slope withheld	as `AO_wrong`; the sentence reporting the parent slope states it is not reported	slope-withheld 2×2
`AO_correct`, slope withheld	as `AO_correct`, with the same sentence edit	slope-withheld 2×2
`AO_wrong`, non-identifying statistic	as `AO_wrong`; the slope sentence reports the exposure standard deviation instead; nothing declared missing	complete-artifact 3×2
`AO_correct`, non-identifying statistic	as `AO_correct`, with the same sentence edit	complete-artifact 3×2
Method carriers (calibration ladder, §D.10)		
estimator name / weak hint / partial formula	graded semantic method information	2×5 factorial
`EXPLICIT_WRONG_METHOD`	rule stated explicitly, no number; **positive control**	all regimes

C.6 Estimands

With $q(c)$ the wrong-signature rate of condition c over discriminative probes, the answer-specific increment is a vector of five contrasts:

$$\Delta_{\text{shell}} = q(\texttt{AO_wrong}) - q(\text{ARTIFACT_SHELL}), \tag{10}$$

$$\Delta_{\text{noart}} = q(\texttt{AO_wrong}) - q(\text{NO_ARTIFACT}), \tag{11}$$

$$\Delta_{\text{random}} = q(\texttt{AO_wrong}) - q(\text{RANDOM_ANSWER}), \tag{12}$$

$$\Delta_{\text{agreement}} = q(\texttt{AO_wrong}) - q(\text{AGREEMENT_ANSWER}), \tag{13}$$

$$\Delta_{Y} = q(\texttt{AO_wrong}) - q(\texttt{AO_correct}). \tag{14}$$

The attribute-matched decoy supplies a sixth, separately reported contrast $\Delta_{\text{matched}} = q(\texttt{AO_wrong}) - q(\text{MATCHED_RANDOM})$.

Remark 2 (Which random control enters which contrast)**.** Δ_{random} in Eqs. 12 uses the *unmatched* `RANDOM_ANSWER` in every regime, including the sampled-probe analysis. The attribute-matched decoy is a separate condition run on the same eight worlds and reported as Δ_{matched} in §D.1; it is never substituted into the five-contrast family. The two are not interchangeable: they have different measured contrasts (0.479 versus 0.572) and different simultaneous bounds (0.191 versus 0.395).

The five contrasts answer five different questions: does the wrong answer exceed artifact presence, reuse framing, generic numeric content, a method-nondiagnostic answer, and a correct answer? Zero-margin attribution requires all five positive; the operational criterion requires all five above

0.25. Δ_Y is the contrast that distinguishes estimator-specific information from generic answer conditioning, and it is the one that binds in every analysis we ran.

C.7 STATISTICAL INFERENCE

§B.1 noted that the analyses do not share an estimand. Because that is easy to lose track of when comparing their numbers, Table 7 states each one literally.

Table 7: What each analysis estimates, and with which probe. No number from one row is a replication of a number from another, and none is a population statement about scientific workflows. "Decoding-rule probe" is the contract quoted in §C.2.

Analysis	Estimand	Integrated over	Held fixed
Fixed bank (§D.8)	mean wrong-signature rate on a frozen discriminative bank; frozen numeric probe	worlds, seeds, configurations	probe set, one execution
Sampled probe (§D.1)	rate over world-specific probes given one realised carrier memo; decoding-rule probe	worlds, probes	cell, prompt, one memo per condition
Repeated execution (§D.2)	rate over endpoint executions in the *same* worlds; decoding-rule probe	executions, probes	the eight worlds, cell, prompt
Execution-resampled and slope-withheld, probe stage (§D.3, §D.4)	rate over worlds, executions and probes; decoding-rule probe	worlds, executions, probes	cell, prompt
Same runs, memo stage (*post hoc*)	share of memos that report the new world's confounded slope as the effect; no probe	worlds, executions	cell, prompt
Complete-artifact control, memo stage (§D.4; pre-registered)	the same memo share; no probe is run	worlds, executions	cell, prompt

The fixed eight-probe bank is deterministic, so probe rates are descriptive rather than binomial population estimates; independent worlds are the replicating unit. For the sampled-probe analysis, worlds and probes are resampled hierarchically (Efron & Tibshirani, 1993) for one realised memo per world–condition (4,000 draws, seed 20261012). The repeated-execution analysis uses the hierarchy world $\rightarrow$ endpoint execution $\rightarrow$ probe, resampling worlds first, executions independently by condition within world, and matching probe identities across conditions.

The test, and its stated level. Answer-specific reconstruction is claimed only when every matched-control contrast supports the effect — an intersection–union test (Berger, 1982), whose size is bounded by the level of its individual components and which therefore requires *no* multiplicity correction. We nonetheless report Bonferroni-simultaneous one-sided lower bounds at per-contrast $\alpha = 0.01$ (familywise 0.05 over the five-member family), and every decision in §D is made on those conservative bounds rather than on the unadjusted IUT. This is a doubly conservative reporting choice, made deliberately and stated here so the operating characteristic is legible: an IUT decision at nominal 0.05 would be more permissive than anything we report.

Eight clusters, and why the exact test is the primary one. The sampled-probe and repeated analyses have many probe-level observations and only 8 world-level clusters. This is a regime in which cluster-robust and bootstrap procedures are known to mis-cover, and in which randomization inference is the indicated tool (MacKinnon & Webb, 2018; Ritzwoller et al., 2024). We follow that recommendation rather than report a bootstrap interval we would have to discount. The primary test is an *exact world-level sign-flip test* — randomization inference over the 2^8 sign assignments of the world-level contrast — whose smallest attainable one-sided p-value at $n = 8$ is $2^{-8} = 0.0039$ and which returns $p = 0.0156$ for the zero-margin conjunction. It is supported by leave-one-world-out

re-estimation of every contrast, reported for all eight omissions. The hierarchical percentile bootstrap is retained for *description* of the contrast distribution and its intervals are labelled resampling ranges, not coverage statements; where it disagrees with the exact test we report the disagreement and defer to the exact test. The 2×5 factorial is weaker still on this axis: it has three worlds per model–operator stratum, so its stratified resampling interval is a *resampling range* over at most 3^3 distinct within-stratum draws and we label it as such rather than as a 95% confidence interval.

Thirty-two worlds, and the memo stage. The execution-resampled and slope-withheld analyses resample world, then execution within world, then probe within execution. Their world-level sign-flip tests are Monte Carlo over 200,000 random sign assignments, because 2^{32} cannot be enumerated; their p-values carry Monte Carlo error, and the smallest reportable value is $1/(B+1)$. Memo-stage analyses, including the complete-artifact control (4,000 bootstrap draws, 200,000 sign assignments), use the same machinery with the probe level removed, since each (world, execution, condition) has one memo.

Assay sensitivity. A configuration is *positive-control-responsive* when `EXPLICIT_WRONG_METHOD` transmits at rate ≥ 0.5. Contrasts from non-responsive configurations are reported but never used to support a negative claim about carriers, since a non-responsive assay cannot distinguish an absent carrier from a deaf instrument.

C.8 Falsification criteria, frozen before measurement

Criterion 1 (Operational margin)**.** The five-contrast conjunction clears $\delta_{\text{practical}} = 0.25$ in the confirmatory cell. This criterion has been evaluated twice and we report both outcomes. **fired** on the 8-world analysis (§D.1): the conjunction fails from $\delta = 0.20$ upward, with Δ_Y's simultaneous bound at 0.152. **held** on the 32-world, execution-resampled analysis (§D.3), where the smallest simultaneous bound is 0.298 and the operational IUT gives $p = 4.9 \times 10^{-4}$. The second evaluation is a second look at a criterion that had already fired once, and we do not ask the reader to forget the first. What we can say is that its protocol was frozen before its worlds were generated, that it holds δ, α, the conditions, the probe count and the scorer fixed and changes only the number of worlds and the variance model, and that the 24-world fresh-only sensitivity—worlds that did not exist when the first evaluation was written—clears δ on its own at 0.303. The paper's claim is stated as a bound rather than as either verdict (§D.3), so a reader who weights the first evaluation more heavily loses the δ statement and no other claim.

Criterion 2 (Scope beyond one cell)**.** At least two model–operator cells show a positive five-contrast conjunction at $\delta = 0$. **fired** (§D.8, §D.1): one cell qualifies; the fixed-bank sweep is null in all 18.

Criterion 3 (Answer specificity)**.** $\Delta_Y > 0$ with a positive simultaneous lower bound wherever the conjunction is evaluated. **held**: 0.152 (sampled), 0.132 (repeated) and 0.298 (32 worlds, execution-resampled), all positive, each the smallest bound in its family.

Criterion 4 (Assay validity)**.** The explicit-method positive control transmits at ceiling and agreement probes score correctly wherever a positive conjunction is claimed. **held**: the positive control is 1.000 in all eight held-out worlds and has mean 0.999 over the 32 execution-resampled worlds (1.000 in 29 of them), and agreement accuracy is $7{,}665/7{,}680 = 0.998$.

Two of four fired. We report the claim they leave standing, not the claim the design was built to test.

D Results

We report the positive cell first: its discovery on eight worlds (Study 1), its pre-registered re-run on thirty-two (Study 2), the intervention that identifies its mechanism, and what the model's own memos show. The boundary measurements follow: the operator whose artifact is not recoverable, and a broad sweep under a different probe. Table 8 lists every measurement with its stage and scope.

D.1 Study 1: discovery on eight held-out worlds

In Qwen3.5-27B/E1/P3 — the cell that pairs a low no-artifact floor with a positive-control-responsive assay — the sampled-probe analysis covers 8 held-out worlds with 32 discriminative probes each, scored at the probe stage. All five contrasts are positive:

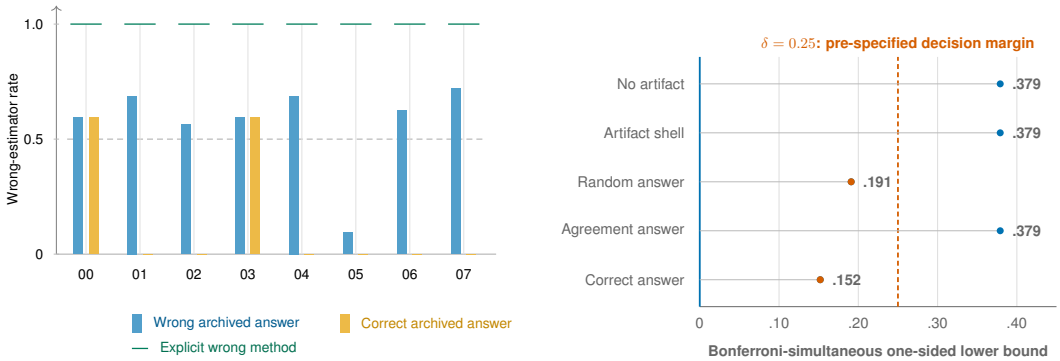
Table 8: Every measurement in this paper, with its stage and scope. "Probe" rates use the probe stated in Table 7; "memo" rates are the model's own estimate for the new problem, *post hoc* except in the complete-artifact control, which pre-registered them (§C.2). Nothing in this table is a population statement about scientific AI workflows.

Measurement	Result	Scope
The positive cell: Qwen3.5-27B / E1 / P3, and its replication in Qwen3.8-27B		
Study 1, probe (§D.1)	Five-contrast zero-margin **pass** (sim. LBs 0.152–0.379); δ=0.25 **fail**. Decoy $\Delta_{\text{matched}} = 0.572$.	8 worlds; repeated executions in the same worlds (§D.2)
Study 2, probe (§D.3)	Zero-margin **pass** ($p < 5 \times 10^{-6}$) and δ=0.25 **pass** (sim. LBs 0.298–0.409); holds on the 24 fresh worlds alone.	32 worlds × 5 executions
Study 2, memo (§D.3)	`AO_wrong` 0.988 against controls 0.070–0.289; sim. LBs $\geq +0.500$ (fresh worlds $\geq +0.375$).	same runs; *post hoc*
Slope withheld, probe (§D.4)	Δ_Y $+0.403 \rightarrow -0.108$; DiD $+0.511$, $p < 5 \times 10^{-6}$.	32 worlds × 3 executions; pre-registered
Slope withheld, memo (§D.4)	Δ_Y $+0.761 \rightarrow -0.133$; DiD $+0.894$, $p < 5 \times 10^{-6}$. Replicated pre-registered: DiD $+0.919$, $p < 5 \times 10^{-6}$.	same runs, *post hoc*; replication in the complete-artifact runs
Complete-artifact control, memo (§D.4)	Δ_Y $+0.814 \rightarrow -0.025$ with the artifact complete; DiD $+0.838$, $p < 5 \times 10^{-6}$. Identifiability supported, completeness not supported.	32 worlds × 5 executions; pre-registered, memo primary
Replication, Qwen3.8-27B (§D.5)	Eligible; Δ_Y $+0.525 \rightarrow +0.068$; DiD $+0.457$, $p = 1.5 \times 10^{-5}$. Identifiability supported, completeness not supported.	32 worlds × 5 executions; pre-registered, memo primary
Boundary measurements		
Operator contrast (§D.7)	Numeric effect $+0.367$ in recoverable E1/Qwen; $+0.025$/-0.067 in non-recoverable MA.	2×5 factorial, four cells; frozen numeric probe
Fixed-bank sweep (§D.8)	All five contrasts in $[-0.03, +0.03]$; every one-sided lower bound crosses zero.	18 configurations; frozen numeric probe
Probe-regime bridge (App. K.3)	Fixed-bank depth arm clears both bars; the other two arms fail Δ_Y. No attribution claimed.	positive cell; three 8-world samples; *exploratory*
Semantic ladder (§D.10)	rule $+0.604$, hint $+0.531$, formula $+0.297$, name $+0.135$. Positive controls, not findings.	12 worlds, four cells

Table 9: Five carrier contrasts, held-out sampled-probe cell. Qwen3.5-27B/E1/P3, 8 worlds, hierarchical world-then-probe bootstrap, 4,000 draws, seed 20261012. Simultaneous bounds are Bonferroni at per-contrast $\alpha = 0.01$. Δ_{matched} is the sixth, separately reported decoy contrast (Remark 2) and is *not* a member of the five-contrast family.

Contrast	Mean	One-sided LB	Simultaneous LB	Binding at
Δ_{shell}	0.572	0.441	0.379	—
Δ_{noart}	0.572	0.441	0.379	—
$\Delta_{\text{agreement}}$	0.572	0.441	0.379	—
Δ_{random}	0.479	0.289	0.191	$\delta = 0.20$
Δ_Y	0.422	0.230	0.152	$\delta = 0.20$
Δ_{matched} (sixth)	0.572	—	0.395	—
q(`AO_wrong`)	0.572			
q(`AO_correct`)	0.150			

What "all five controls" actually means here. Three of the five controls — `NO_ARTIFACT`, `ARTIFACT_SHELL` and `AGREEMENT_ANSWER` — score exactly 0.000 in *every* one of the eight worlds, so Δ_{shell}, Δ_{noart} and $\Delta_{\text{agreement}}$ are not three pieces of evidence but one number written three times, and the five-member Bonferroni family is correspondingly conservative beyond its nominal level. `RANDOM_ANSWER` is 0.000 in seven of eight worlds, so Δ_{random} is determined by the single deviant world at 0.750. The informative content of the conjunction is therefore: a control floor that is empty in this sample, one world that perturbs it, and Δ_Y. We state this because "higher than all five matched controls" reads as five independent hurdles and it is not one; the hurdle that carries the claim is the correct-answer control, which is also the smallest contrast and the one that



(a) Per-instance estimator signatures **(b)** Simultaneous lower bounds

Figure 5: Estimator reuse on held-out problem instances. (a) Wrong-estimator rates for the wrong and correct archived answers per problem instance; green marks the explicit wrong-method positive control, at 1.000 in every instance. (b) Bonferroni-simultaneous one-sided lower bounds for the five primary contrasts (orange: below δ). All simultaneous lower bounds exceed zero, whereas the pre-specified decision criterion at $\delta = 0.25$ is not supported.

binds. Whether the floor is empty as a property of the assay or as an accident of eight worlds is exactly what a larger sample decides, and we do not assume it.

The zero-margin IUT passes. **The pre-registered operational IUT does not** (Criterion 1): the margin sweep passes at $\delta \in \{0, 0.05, 0.10, 0.15\}$ and fails at $\delta \geq 0.20$, determined by Δ_{random} (0.191) and Δ_Y (0.152). In the bootstrap distribution 99.98% of replicates have all five contrasts above zero and 91.5% above $\delta_{\min}$; these are replicate fractions summarising joint support, not probabilities that the population contrasts exceed a threshold.

Per-world, the cell is heterogeneous. Of eight held-out worlds, three satisfy the strict per-world gate outright; one has `RANDOM_ANSWER` at 0.750 (above its own `AO_wrong` rate of 0.594, giving a locally negative Δ_{floor}), two show `AO_correct` matching `AO_wrong` exactly, one has $q(\texttt{AO_wrong}) = 0.094$, and one fails agreement validity at $15/16$. The pooled positive result is therefore carried by roughly five of eight worlds, and the pooled Δ_{random} is driven by a single deviant world. Per-world values are in Appendix J.2.

D.2 Study 1, repeated executions: reliability within the cell, not replication

Repeating each of six conditions 10 times in the *same* eight worlds gives $q(\texttt{AO_wrong}) = 0.568$, $q(\texttt{AO_correct}) = 0.189$, $q(\text{NO_ARTIFACT}) = 0.013$, $q(\text{RANDOM}) = 0.024$, $q(\text{AGREEMENT}) = 0.000$ and the positive control at 1.000, with contrasts $\Delta_{\text{noart}} = 0.555$ $[0.414, 0.671]$, $\Delta_{\text{random}} = 0.544$, $\Delta_{\text{agreement}} = 0.568$ and $\Delta_Y = 0.379$ $[0.171, 0.590]$; simultaneous lower bounds 0.132–0.399. Leave-one-world-out preserves positive zero-margin bounds under every omission (Δ_Y bound ranging 0.073–0.187), and an exact world-level sign-flip test gives $p = 0.0156$.

Three qualifications belong with these numbers rather than after them. The analysis reuses the same eight worlds, so it adds no world-level replication and cannot widen the scope of §D.1. It omits `ARTIFACT_SHELL` and the matched decoy, so it tests four of the five contrasts, not the conjunction of Definition 1. And its own cell decision rule at $\delta = 0.25$ *didnotpass*, for the same reason as before: Δ_Y. What it does establish is that the ordering is not an artifact of one lucky endpoint execution under temperature-0 decoding, which is worth knowing and is all it is.

D.3 Study 2: thirty-two worlds, five executions, and the model's own estimate

The two analyses above share two weaknesses, and they are the ones a reader should press on. §D.1 rests on 8 worlds, which puts a floor of $2^{-8} = 0.0039$ under any world-level sign-flip p-value and leaves the $\delta = 0.25$ margin undecided rather than refuted. §D.2 adds executions but reuses

those same eight worlds, so it buys reliability and no replication. Neither resamples execution: both treat one endpoint execution per world as if it were the world's value, which is only legitimate if temperature-0 decoding is deterministic. It is not, and the size of that omission is reported below.

We therefore pre-registered and ran a third analysis that fixes all three at once. The protocol was frozen before any of its worlds were generated. It draws 32 worlds—the 8 of §D.1 plus 24 generated afresh—and executes the full condition set 5 times independently in each, giving 160 (world, execution) cells. Inference is a three-level bootstrap that resamples world, then execution within world, then probe within execution, at 4,000 draws with seed 20260922; the randomization test is a Monte Carlo world-level sign-flip test over $B = 200{,}000$ random sign assignments with seed 20260923. With 2^{32} assignments, exact enumeration is infeasible, so its p-values carry Monte Carlo error and the smallest reportable value is $1/(B{+}1)$; a reported $p < 5 \times 10^{-6}$ means that no sampled assignment reached the observed statistic. Everything else—conditions, probe count, scorer, tolerance, δ, per-contrast $\alpha = 0.01$—is unchanged from §D.1, so the comparison is a sample-size and variance-model comparison and nothing else.

Table 10: Five carrier contrasts at 32 worlds with execution resampled. Qwen3.5-27B/E1/P3, 5 independent executions per world, three-level bootstrap (world → execution → probe), 4,000 draws, seed 20260922. Simultaneous bounds are Bonferroni at per-contrast $\alpha = 0.01$. The last column repeats the simultaneous bound on the 24 worlds that did not exist when §D.1 was written.

Contrast	Mean	One-sided LB	Simultaneous LB	Fresh-worlds sim. LB
Δ_{shell}	0.507	0.438	0.409	0.378
Δ_{noart}	0.495	0.416	0.380	0.361
$\Delta_{\text{agreement}}$	0.507	0.438	0.409	0.378
Δ_{random}	0.496	0.420	0.388	0.355
Δ_Y	0.419	0.331	0.298	0.303
q(`AO_wrong`)	0.507			
q(`AO_correct`)	0.088		0.015	

The criterion that failed at eight worlds passes at 32. Every simultaneous lower bound in Table 10 exceeds $\delta = 0.25$; the smallest is Δ_Y at 0.298, and it is again the contrast that binds. The zero-margin IUT gives $p < 5 \times 10^{-6}$ and the operational IUT at $\delta = 0.25$ gives $p = 4.9 \times 10^{-4}$, the latter an order of magnitude below the 0.0039 floor that eight worlds imposed. $all four$ frozen criteria—operational margin, zero margin, answer specificity, and assay validity—are recorded as passing in the analysis artifact. 99.95% of bootstrap replicates have all five contrasts above δ; as in §D.1 this is a replicate fraction summarising joint support, not a probability that the population contrasts exceed a threshold.

What this does not change: it does not touch the fixed-bank null of §D.8, does not separate regime effect from probe difficulty (§D.8; these are sampled probes drawn the same way), and does not widen the result beyond one model, one operator and one prompt level. A criterion passing on more worlds is a stronger version of the same narrow claim, not a broader one.

The model's own estimate for the new problem shows the same thing, more strongly, and without an instructed decoding step. Every number above is a probe-stage rate, read out under the decoding rule quoted in §C.2. The same runs also contain the memo each execution wrote for the new problem before any probe, under the uninstructed P3 transfer prompt. Table 11 places the two stages side by side. At the memo stage the archived wrong answer makes the model report the new world's confounded slope as the intervention effect in 0.988 of 160 executions, against 0.070–0.289 under the five controls; all five contrasts clear zero with simultaneous lower bounds from +0.500 (world-level sign-flip $p < 5 \times 10^{-6}$ for the weakest), and on the 24 fresh worlds alone from +0.375. This analysis is post hoc—no protocol named the memo as its outcome—and we weight it accordingly; but it answers the question the probe stage cannot, whether the model decodes the archive when nobody tells it how.

Table 11: Two stages of the same 32×5 runs. Wrong-estimator rate by condition. *Memo*: share of generation-1 memos that report the new world's confounded slope as the intervention effect, uninstructed P3 prompt, frozen deterministic memo scorer; *post hoc*. *Probe*: pre-registered outcome under the decoding-rule probe (§C.2); bootstrap means. The last column is the memo-stage contrast $q(\texttt{AO_wrong}) - q(c)$ with its Bonferroni simultaneous lower bound.

Condition	Memo	Probe	Memo contrast (sim. LB)
AO_wrong (target)	0.988	0.507	—
ARTIFACT_SHELL	0.070	0.000	+0.918 (+0.781)
NO_ARTIFACT	0.076	0.012	+0.912 (+0.781)
RANDOM_ANSWER	0.281	0.011	+0.707 (+0.500)
AGREEMENT_ANSWER	0.289	0.000	+0.699 (+0.500)
AO_correct	0.207	0.088	+0.781 (+0.594)
EXPLICIT_WRONG_METHOD	1.000	0.999	positive control

Two features of the table matter. The target is higher at the memo stage than at the probe stage, so the decoding instruction did not manufacture the effect; the probe stage records less of it, for reasons this design does not measure. And the controls are *not* at zero at the memo stage: when the archive holds a number that does not identify the naive estimator—a random value, a value on which the estimators coincide, or the correct value—the model still adopts the naive estimator on its own in 0.207–0.289 of executions. The exact zeros of ARTIFACT_SHELL and AGREEMENT_ANSWER at the probe stage are produced by the probe's fallback clause, which sends every memo without a matching value to the correct estimator. The memo-stage conjunction is therefore the harder test, and it is the one we rest the mechanism on.

The 24 fresh worlds carry it on their own, and they are drawn from the same distribution. Restricting to the worlds generated after §D.1 was analysed—so that no world contributing to this sensitivity was ever inspected while the analysis was being designed—all five simultaneous bounds still clear δ, the smallest being 0.303, with 99.9% joint replicate support. This is the closest thing in the paper to out-of-sample replication of the positive cell, and it is why we report the fresh-worlds column in the table rather than in an appendix.

A fresh sample is only informative if it is the same kind of sample, so we check rather than assert it. The quantity that defines the probe regime is the oracle's confounding gap $|\beta_{\text{obs}} - \tau|$. Its per-world mean is 1.5333 over the original 8 worlds and 1.5396 over the 24 fresh ones, a difference of +0.0064 with a world-level two-sided permutation $p = 0.873$ at 200,000 draws; the pooled probe-level Kolmogorov–Smirnov statistic is 0.047, and the two groups share the same support $[0.800, 2.824]$ and the same median 1.455. The probe bank is a deterministic function of the world seed, which we verify rather than assume: it is identical across all 5 executions of every world (0 mismatches). **The same check also shows what more worlds cannot fix.** No sampled probe in any of the 32 worlds has a gap below 0.5 (the minimum over all of them is 0.800, and 0 worlds contain such a probe), whereas the fixed bank has two probes at exactly 0.5. Drawing more worlds from this generator therefore leaves the regime confound of §D.1 exactly where it was, and Remark 3 shows why: 0.5 lies outside the gap support the generator can reach at all, so the shortfall is structural rather than statistical.

Temperature-0 decoding is not deterministic, and the earlier bootstraps omitted that variance. Across the 32 worlds the within-world between-execution standard deviation of $q(\texttt{AO_wrong})$ averages 0.072 and reaches 0.321; for $q(\texttt{AO_correct})$ it averages 0.051. The controls behave differently: ARTIFACT_SHELL and AGREEMENT_ANSWER have standard deviation exactly zero in all 32 worlds, while NO_ARTIFACT and RANDOM_ANSWER do not. A single-execution design therefore does not merely lose precision; it estimates a quantity whose per-world value is itself uncertain, and a two-level bootstrap over world and probe reports an interval that is too narrow by construction. The bounds in Table 10 include this component. This is also the reason we do not present §D.1 as superseded arithmetic: its point estimates are not wrong, its uncertainty model is incomplete.

The probe-stage control floor is structural, and the memo stage shows why. At the probe stage ARTIFACT_SHELL and AGREEMENT_ANSWER are exactly 0.000 in all 32 worlds and every

execution, so those two contrasts are one number written twice and the Bonferroni family is conservative beyond its nominal level; NO_ARTIFACT is non-zero in 2 worlds and RANDOM_ANSWER in 3. Table 11 shows that the zeros are a property of the probe instruction, not of the model: the same executions adopt the naive estimator at the memo stage in 0.070 and 0.289 of cases, and the probe's fallback clause maps those memos back to the correct estimator. The memo-stage conjunction has no structural zeros, which is one more reason it is the version we rest the mechanism on.

A correct archived answer is not safe either, and we can say what decides it. $q(\texttt{AO_correct})$, the wrong-signature rate when the archived artifact carries the *correct* value, has mean 0.088 with a simultaneous lower bound of 0.015—strictly above zero—and is non-zero in 10 of 32 worlds. The failure mode is not switched off by archiving a correct number; it is reduced, by $\Delta_Y = 0.419$, and not eliminated.

Two explanations compete for those 10 worlds, and the archived artifacts separate them. The first is numeric geometry: a world whose candidate estimators are far apart might be the one that leaks. The second is that the model's own intermediate memo adopts the naive estimator regardless of what the archived answer said, because β_{obs} is on the page and the naive estimate is one step of arithmetic away. Scoring every AO_correct child memo with the frozen deterministic scorer against its own child world decides between them. The candidate-set distance $|\beta_{\text{obs}} - \tau|$ barely differs between the leaking and the clean worlds (1.636 versus 1.409, world-level permutation $p = 0.254$) and correlates weakly with the leak rate across the 32 worlds ($r = 0.33$, $\rho = 0.22$, permutation $p = 0.059$). The memo's estimator choice does the work: the AO_correct memo is labelled *naive-rule* in 0.580 of executions in the leaking worlds against 0.036 in the clean ones, and across all 32 worlds the leak rate tracks it at $r = 0.89$ ($\rho = 0.88$, permutation $p < 10^{-5}$). Mere mention of β_{obs} does not discriminate at all—it appears in 1.000 of memos in both groups—so what separates the worlds is not whether the slope is visible but whether the reconstruction adopts it.

This is the recoverability account of §D.7 showing up where the archived number is innocent. When a correct value is archived and the model still answers with the naive estimator, nothing was transmitted from the archive; the estimator was rebuilt from the visible slope. It is evidence for the paper's mechanism claim and against the stronger reading the title asks about, and it is descriptive: the correlation is between two measured quantities in the same worlds, not an intervention on the memo. The experiment that would make it interventional is registered in §E and is not run.

What we claim, and at which margin. Study 1 was a discovery analysis and Study 2 a preregistered replication whose protocol was frozen before its worlds were generated; the replication that matters most is the 24-world fresh-only bound, since none of those worlds existed when Study 1 was analysed. The claim is the bound, not the verdict: in this cell the archived wrong answer raises wrong-estimator reuse above all five matched controls by at least 0.298 at the probe stage and $+0.500$ at the memo stage, and by at least 0.303 and $+0.375$ on the fresh worlds alone. The operational margin $\delta = 0.25$ failed at eight worlds, where intervals are wide, and passed at thirty-two; both verdicts stand in §C.8, and nothing in the paper's contribution rests on which side of 0.25 the bound falls.

Completeness and provenance. All 160 cells completed and all 32 worlds carry all 5 executions, so the analysis runs in its strict mode and no cell is dropped. Reaching that took infrastructure retries: the upstream provider intermittently returned empty generation text, in one (world, execution) pair across four separate attempts. The frozen protocol permits retries for provider, transport or infrastructure failures and forbids them for any outcome-dependent reason; the retry criterion here is an empty response, which is outcome-independent by construction, and the response cache is content-addressed with no expiry, so a retry re-requests only the calls that failed and replays the rest. Every number in Table 10 and in this subsection is generated from the analysis artifact by scripts/emit_stage2_macros.py rather than transcribed, the artifact records scientific_result as a value *computed* from matrix completeness, clean execution and the assay invariant rather than asserted, and the analysis was reproduced by an independent second implementation sharing no code with the first (Appendix H.2).

D.4 REMOVING IDENTIFIABILITY ABOLISHES THE EFFECT

The E1 artifact reports the parent's observational slope, so an archived value equal to it identifies the naive estimator by arithmetic (Remark 1). That makes two readings of the positive cell possible: the number moves the model because, next to the slope, it lets the estimator be re-identified; or the number moves the model by itself and the slope is incidental. The slope-withheld experiment separates them, and it was registered, with its predictions and a commitment to retitle the paper if the second reading failed, before it ran. Withholding the slope changes two things at once—whether the estimator can be identified, and whether the artifact is complete—so a second registered experiment, the complete-artifact control, separates those two.

Design. A 2×2 within-world factorial crossing what the archive contains (the wrong value β_{obs}, or the correct value τ) with whether the artifact reports the parent observational slope. The withheld artifacts are identical to the shown ones except that the sentence giving the slope is replaced by a sentence saying the slope is not reported; the archived final estimate, the domain, the structural-model sentence and the no-rule disclaimer are preserved verbatim. The manipulation does not remove the number. **It removes the reader's ability to see that the number equals the slope**, which is the only thing in the artifact that identifies which estimator produced it. The design runs over the 32 Stage-2 worlds with 3 independent executions each, 96 units in total and 0 incomplete; probes are matched across all four cells; inference is the three-level bootstrap at 4,000 draws with seed 20260926 and a Monte Carlo world-level sign-flip test at 200,000 random sign assignments. The protocol, its estimand and its decision rules were frozen before execution, and the reduction from five executions to three was recorded as an outcome-independent amendment before any of it ran.

The pre-registered primary estimand is the difference in differences,

$$\text{DiD} \;=\; \Delta_Y^{\text{shown}} - \Delta_Y^{\text{withheld}}, \qquad \Delta_Y^{s} = q(\texttt{AO_wrong}) - q(\texttt{AO_correct}) \text{ at slope level } s, \quad (15)$$

which cancels any nuisance that is constant across the two archive contents within a world, including the withheld sentence being longer than the shown one. We report it at both stages: the probe stage is the pre-registered outcome, the memo stage is post hoc (§C.2).

Table 12: Slope-withheld 2×2 at both stages. Qwen3.5-27B/E1/P3, 32 worlds $\times$ 3 executions. q is the wrong-estimator rate: at the memo stage, the share of memos reporting the new world's confounded slope as the effect (*post hoc*); at the probe stage, the pre-registered rate under the decoding-rule probe. The withheld artifacts differ from the shown ones by one sentence and retain the archived number.

Stage		archive holds β_{obs} (wrong)	archive holds τ (correct)
Memo	slope shown	0.979	0.218
	slope withheld	0.053	0.186
Probe	slope shown	0.519	0.116
	slope withheld	0.028	0.136

Stage	Quantity	Mean	95% interval	sign-flip p
Memo	Δ_Y^{shown}	$+0.761$	$[+0.594, +0.906]$	$p < 5 \times 10^{-6}$
	$\Delta_Y^{\text{withheld}}$	-0.133	$[-0.281, +0.000]$	$p = 0.084$
	DiD	$+0.894$	$[+0.719, +1.062]$	$p < 5 \times 10^{-6}$
Probe	Δ_Y^{shown}	$+0.403$	$[+0.281, +0.522]$	$p < 5 \times 10^{-6}$
	$\Delta_Y^{\text{withheld}}$	-0.108	$[-0.242, +0.011]$	$p = 0.975$
	DiD (primary)	$+0.511$	$[+0.382, +0.646]$	$p < 5 \times 10^{-6}$

The result. With the slope shown, the archived wrong answer makes the model report the confounded slope as the effect of the new problem in 0.979 of executions, against 0.218 for a correct archive. With the slope withheld and the number verbatim, the same archived value produces 0.053, against 0.186: the memo-stage DiD is $+0.894$ $[+0.719, +1.062]$. The pre-registered probe outcome moves the same way, DiD $+0.511$ $[+0.382, +0.646]$ with $p < 5 \times 10^{-6}$, and against its frozen decision rules makes *recoverability necessary* supported and *a numeric carrier beyond recoverability* not supported.

At neither stage does the effect merely shrink to zero. $\Delta_Y^{\text{withheld}}$ is negative at both (-0.133 and -0.108), with intervals that reach zero from below: stripped of the slope, the wrong archived value induces the naive estimator slightly *less* often than a correct one does. This falsifies our registered prediction 2, which said the withheld effect would be small but positive because the archived value would still carry weak distributional information. It carries none we can detect.

The manipulation did not break the task. A near-zero rate in the withheld arm would be uninformative if the model had stopped answering. It did not: over $3{,}072$ discriminative probes per cell, absent answers number at most 5 in any cell and 0 answers fall outside the two estimator signatures. At the probe stage the withheld-wrong answers go to the correct estimator (0.971, against 0.481 with the slope shown), but that is what the probe's fallback clause prescribes for a memo without a matching value, so it shows validity, not spontaneous repair. The uninstructed evidence is the memo: without the slope, the model stops reporting the confounded slope as the effect of the new problem.

Reproduction. The shown arm re-executes a condition of §D.3 in a separate run and reproduces it (memo stage 0.988 there, 0.979 here, 1.000 in the complete-artifact runs). Pooled over the four cells, the memo's estimator choice predicts the probe outcome at $r = 0.76$ over 384 units. An independent re-implementation of the probe-stage analysis, sharing no code with it, reproduces the DiD and all three sign-flip p-values (Appendix H.3).

The reuse instruction is held fixed, so instruction-following is not the explanation. A natural deflation of the positive cell is that the model is following the P3 instruction to carry forward the parent's approach, and that nothing about the archived number is involved. All four conditions carry the *same* P3 instruction, the same archived value, the same worlds and the same probe contract; only one sentence of the artifact differs. Under the deflationary reading the withheld arm should behave like the shown arm. It does not. The instruction may be necessary—we have no measurement at P0–P2, and Criterion 2 fires on scope—but it is not sufficient, and what supplies the remainder is whether the artifact lets the estimator be re-identified.

The residual channel beyond recoverability is bounded. The alternative this experiment tests is a numeric channel that operates without recoverability. With the slope withheld, Δ_Y has upper interval end $+0.000$ at the memo stage and $+0.011$ at the probe stage, against $+0.761$ and $+0.403$ when the slope is shown; with the non-identifying statistic in its place (below), the memo-stage upper end is $+0.156$ against $+0.814$. We report these as measured upper bounds on this two-estimator task, not as an absence.

A complete artifact that identifies nothing. Replacing the slope sentence with "the observed regression slope … is not reported in this artifact" removes identifiability, and it also makes the artifact visibly incomplete; a model that discounts incomplete artifacts wholesale would show the same pattern. The complete-artifact control adds a third slope level in which the slope sentence is replaced by one of the same frame, *"The observed sample standard deviation of the exposure X is s_X,"* where $s_X = \sqrt{a^2 + \sigma_X^2}$ is the parent world's exposure standard deviation, printed in the slope's numeric format after a fixed prefix of the slope sentence's length. Neither β_{obs} nor τ can be computed from s_X and the rest of the artifact, so the estimator is exactly as unidentifiable as in the withheld arm, but nothing is declared missing. A deterministic audit, run before any model call, confirms in every world that the three slope levels differ in exactly one sentence and the two archives only in the archived value, that s_X never coincides with either estimator's value (smallest distance 0.064), and that the non-identifying artifact never states that anything is missing. The variance of X was considered first and rejected before freezing, because in one world it equals τ.

The design crosses the three slope levels with the two archive contents over the same 32 worlds, in 5 fresh executions per world (160 units, none incomplete, 0 empty memos, 960 generation calls). The memo stage is the primary outcome and no probe is run. The primary estimand is $\text{DiD}_{\text{nonid}} = \Delta_Y^{\text{shown}} - \Delta_Y^{\text{nonid}}$, the analogue of Eq. (15). The frozen decision rules are gated on the shown arm reproducing (Δ_Y^{shown} bounded above zero) and distinguish three outcomes: *identifiability*, when $\text{DiD}_{\text{nonid}}$ is positive by the two-sided world-level sign-flip test and the interval of Δ_Y^{nonid} reaches zero; *completeness*, when Δ_Y^{nonid} is bounded above zero and the interval of $\text{DiD}_{\text{nonid}}$ contains zero; and *partial attenuation* in between. The protocol registered three predictions—identifiability;

Δ_Y^{nonid} within 0.15 of $\Delta_Y^{\text{withheld}}$; and replication of the withheld result—and committed the paper to restating its mechanism in terms of completeness if that outcome occurred.

Table 13: Complete-artifact control (pre-registered; memo stage primary). Qwen3.5-27B/E1/P3, 32 worlds $\times$ 5 executions. q is the share of memos that report the new world's confounded slope as the intervention effect. The three slope levels differ in one sentence; the non-identifying level reports the parent's exposure standard deviation in the slope's place.

Slope level	archive holds β_{obs} (wrong)	archive holds τ (correct)
slope shown	1.000	0.186
slope withheld ("not reported")	0.081	0.186
non-identifying statistic	0.237	0.261

Quantity	Mean	95% interval	sign-flip p
Δ_Y^{shown} (gate)	$+0.814$	$[+0.656, +0.938]$	$p < 5 \times 10^{-6}$
$\Delta_Y^{\text{withheld}}$	-0.105	$[-0.281, +0.062]$	—
Δ_Y^{nonid}	-0.025	$[-0.188, +0.156]$	—
$\text{DiD}_{\text{nonid}}$ (primary)	$+0.838$	$[+0.625, +1.031]$	$p < 5 \times 10^{-6}$
$\text{DiD}_{\text{withheld}}$	$+0.919$	$[+0.719, +1.125]$	$p < 5 \times 10^{-6}$
$\Delta_Y^{\text{nonid}} - \Delta_Y^{\text{withheld}}$	$+0.081$	$[-0.156, +0.312]$	$p = 0.324$

The result: identifiability, not completeness. The shown arm reproduces: the archived wrong answer makes the memo report the confounded slope in 1.000 of executions, against 0.186 for a correct archive. With the non-identifying statistic in the slope's place and the artifact complete, the two archives can no longer be told apart: 0.237 against 0.261, $\Delta_Y^{\text{nonid}} = -0.025$ $[-0.188, +0.156]$, and $\text{DiD}_{\text{nonid}} = +0.838$ $[+0.625, +1.031]$ with $p < 5 \times 10^{-6}$ (Table 13, Figure 3). Under the frozen rules identifiability is supported and completeness not supported; an independent re-derivation reproduces every number and both decisions (Appendix H.4). All three registered predictions hold: the non-identifying and withheld arms differ by $+0.081$ $[-0.156, +0.312]$, and the memo-stage withheld result, post hoc above, replicates as a pre-registered estimand ($\text{DiD}_{\text{withheld}} = +0.919$, $p < 5 \times 10^{-6}$). Excluding the 7 worlds whose s_X lies within 0.25 of either estimator's value, a list fixed by the audit before execution, leaves every decision unchanged ($\text{DiD}_{\text{nonid}} = +0.821$ $[+0.560, +1.080]$ over 25 worlds).

What remains in the two non-identifying cells is a naive rate that does not depend on the archived value and is comparable to the rate the correct archive produces with the slope shown. The component that disappears is the answer-specific one, which is the component this paper attributes to the archived number. So the slope matters because its agreement with the archived value identifies the estimator, not because its presence makes the artifact look finished.

D.5 The mechanism replicates in a newer model generation

The complete-artifact design was registered a second time, for Qwen3.8-27B, with the same worlds, prompt, scorer, estimands and decision rules, and with two conditions added so that a null could be told apart from a broken assay: `NO_ARTIFACT`, which measures the room the model leaves for an effect, and the explicit-method positive control, which measures whether it responds at all. A model counts for or against the claim only if it passes both gates (positive control at least 0.5, `NO_ARTIFACT` at most 0.25), and the protocol registered one prediction: in an eligible model in which the effect replicates, identifiability holds.

Table 14: Replication in Qwen3.8-27B (pre-registered; memo stage primary). E1/P3, 32 worlds × 5 executions, 160 complete units. Same conditions, scorer and decision rules as Table 13.

Slope level	archive holds β_{obs} (wrong)	archive holds τ (correct)
slope shown	0.644	0.118
slope withheld ("not reported")	0.125	0.106
non-identifying statistic	0.157	0.088

no artifact 0.118 explicit-method positive control 0.818

Quantity	Mean	95% interval	sign-flip p
Δ_Y^{shown} (gate)	$+0.525$	$[+0.312, +0.719]$	$p < 5 \times 10^{-6}$
$\Delta_Y^{\text{withheld}}$	$+0.019$	$[-0.125, +0.188]$	—
Δ_Y^{nonid}	$+0.068$	$[-0.094, +0.219]$	—
$\text{DiD}_{\text{nonid}}$ (primary)	$+0.457$	$[+0.188, +0.719]$	$p = 1.5 \times 10^{-5}$
$\text{DiD}_{\text{withheld}}$	$+0.506$	$[+0.250, +0.750]$	$p < 5 \times 10^{-6}$

The result. The model is eligible: the positive control transmits at 0.818 and the no-artifact rate is 0.118. The effect replicates: with the slope shown, the archived wrong answer makes the memo report the confounded slope in 0.644 of executions, against 0.118 for a correct archive ($\Delta_Y^{\text{shown}} = +0.525$ $[+0.312, +0.719]$) and against the no-artifact rate by $+0.527$ (lower end $+0.312$). With the non-identifying statistic in the slope's place the two archives again cannot be told apart ($\Delta_Y^{\text{nonid}} = +0.068$ $[-0.094, +0.219]$), so $\text{DiD}_{\text{nonid}} = +0.457$ $[+0.188, +0.719]$ with $p = 1.5 \times 10^{-5}$. Under the frozen rules identifiability is supported and completeness not supported, the withheld result replicates ($\text{DiD}_{\text{withheld}} = +0.506$), and the registered prediction holds. The newer model reuses the archived estimator less often than Qwen3.5-27B when the slope is shown, and the mechanism is the same: the answer-specific effect exists where the artifact identifies the estimator and is absent where it does not (Figure 3). Provenance, the output-token amendment made before any memo was scored, and the per-call cost reconciliation are in Appendix H.5.

A second family, and what it shows about the instrument. The same protocol was registered for one family outside Qwen, Mistral Small 3.2 24B (Appendix H.5). Under the frozen gate it is not eligible: its no-artifact rate is 0.656, far above the 0.25 bound, with the positive control at 0.536, so the cell counts neither for nor against the claim. We reviewed the instrument before reading that as a property of the model. Every naive-rule label on its no-artifact memos comes from the scorer's verbal route, and 87 of those 105 memos state elsewhere that the naive estimator must not be used: this model writes a method discussion at the memo stage and almost never commits to a number. A post hoc numeric criterion that shares no code with the scorer—the memo assigns the new world's observational slope to the effect—gives it 0.000 without an artifact and 0.013 under an explicit wrong rule. The memo stage does not observe this model's estimator choice, so the cell lies outside the assay's domain rather than against the claim. The same criterion reproduces the identifiability result in both Qwen generations ($\text{DiD}_{\text{nonid}} = +0.769$ $[+0.531, +1.000]$ and $+0.444$ $[+0.250, +0.656]$, with Δ_Y^{nonid} at -0.025 and -0.038; Table 20), so the verbal route's over-labelling—small in Qwen3.5-27B, larger in Qwen3.8-27B—does not carry the finding.

A required final estimate: the anchor, and an eligible second family. Amendment A4, registered after the result above, appends one line to the P3 prompt: the memo must end with a final numeric estimate for the new problem or with "not identifiable", and the outcome is a deterministic parse of that line (Appendix H.5). Qwen3.5-27B was re-run under the same line as an anchor, with the registered prediction that it stays eligible and that the effect and identifiability both replicate. They do: it is eligible (positive control 0.893, no-artifact rate 0.000), the effect replicates, and identifiability is supported ($\Delta_Y^{\text{nonid}} = +0.000$). The line itself, however, changes the behaviour: Δ_Y^{shown} falls to $+0.131$ $[+0.031, +0.250]$, and 139 of 160 shown-wrong memos now end "not identifiable". The appended line is the only difference between the two sets of requests. Described post hoc, under the original prompt the shown-wrong memos are short copies of the parent's template (median 374 characters, against 936 for shown-correct); with the line they become reasoning memos (median 1,159) that usually reconstruct the parent's estimator and then reject it. Under this format Mistral Small 3.2 24B is eligible (positive control 0.730, no-artifact rate 0.086) and the effect does not repli-

cate ($\Delta_Y^{\text{shown}} = +0.044\ [-0.031, +0.125]$); its memos are of the same length whether the archived answer is wrong or correct. Recoverability is therefore necessary but not sufficient. No measured row shows an effect without it, and whether a model takes a recoverable estimator depends on the model and on how its output is framed (§D.9).

D.6 What the model writes for the new problem

The memo stage is the outcome the mechanism rests on (§C.2); this section describes what the memos contain and how they are scored. A deterministic text scorer — no LLM judge — assigns mutually exclusive labels for naive-rule use, value transfer without a rule, rejection or abstention, and unclassifiable output, using a value-echo detector, naive-advocacy patterns with negation-span filtering, and rejection-marker detection (Appendix H.6).

Table 15: Carrier-reconstruction fidelity. Qwen3.5-27B, P3, temperature 0; $K = 8$ `AO_correct` carrier executions per world. `AO_wrong` controls are naive-rule in $16/16$ executions; pooled `AO_correct` naive fraction $21/64 = 0.328$. Deterministic world reconstruction reproduced the logged ancestor texts byte-identically in all eight worlds.

Seed	Pattern	p_w	naive/8	unique memos	majority label
20261100	floor-contaminated	1.000	8/8	1	naive rule
20261101	clean	0.000	0/8	7	rejection/abstention
20261102	clean	0.000	0/8	8	rejection/abstention
20261103	`AO_correct` contaminated	1.000	8/8	1	naive rule
20261104	clean	0.000	0/8	4	rejection/abstention
20261105	no induction	0.500	4/8	8	rejection/abstention
20261106	clean	0.125	1/8	5	value-only
20261107	control degraded	0.000	0/8	4	value-only

Two things follow. In $16/16$ `AO_wrong` executions of that assay—and in $158/160$ of the later 32-world run—the deterministic scorer labels the memo *naive-rule*. That label has two routes: a sentence that advocates the naive rule in words, and a *value* route in which the memo re-emits, as its own final estimate, the number the naive estimator produces for the new world. Decomposing the label over all 71 archived `AO_wrong` memos, 69 are earned through the value route and 1 through a verbal rule sentence. The typical memo restates the parent template with the new world's slope and carries forward the sentence "No methodological rule is asserted". **The model is therefore not writing the rule out in words; it is carrying the deciding value forward.** What this establishes is weaker than a verbal confession and is still the thing that matters: estimator identity is fixed in the model's own observable output before any probe is answered, rather than being inferred only from the downstream numbers. That is what decoding looks like at the artifact level, it is consistent with §D.7, and it is inconsistent with a subliminal-style account in which nothing about the estimator is externalised at all.

Why this is not the usual unfaithful-explanation problem. A reader who knows that chain-of-thought can misreport its own causes (Turpin et al., 2023) should ask why a verbal memo counts as evidence at all, and the concern is sharpened by the finding that verbalisation is *weakest* exactly for implicit, artifact-delivered cues (Gema et al., 2026) — our regime. Two features of the design answer it. The memo is not a post-hoc rationalisation of a decision the model already made: it is generated before any probe is seen and the scored responses are conditioned on it, so it sits *in* the causal path rather than commenting on one. And what is scored is not the memo's account of its reasoning but its final estimate for the new problem, checked against the new world's parameters by the same deterministic ground truth as the probes; in 69 of 71 `AO_wrong` memos that estimate is the whole of the evidence. The memo is therefore an estimator choice, not a report about one. What it cannot show is which property of its text carries the choice on to the probes; that requires editing the memo and re-scoring, which we register and do not run (§E). Second, in the two worlds where `AO_correct` induces the wrong estimator, the memo distribution *collapses to a single memo identical to the `AO_wrong` memo*, whereas clean worlds diversify into rejection, abstention, or value-only memos. This localises the Δ_Y shortfall to the intermediate reconstruction

stage: correct-answer contamination is a memo-level collapse, not a probe-level accident. It is a localisation, not a causal identification — we did not intervene on the memo, and §E registers that intervention.

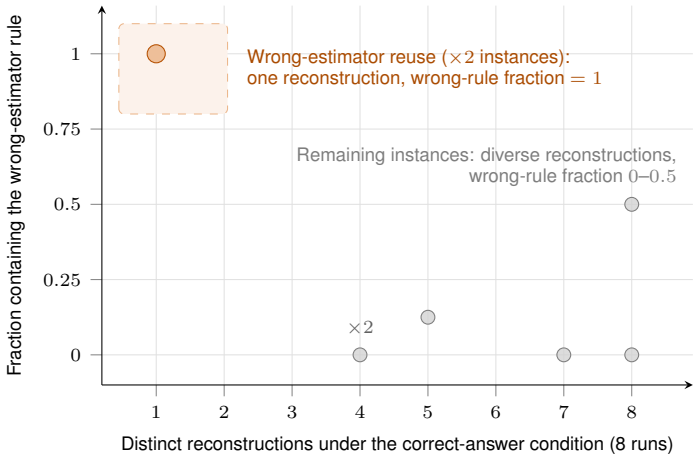


Figure 6: Pre-probe reconstructions under the correct-answer control. Each point represents one of the 8 held-out problem instances (×2: two instances at the same coordinates). In the two instances with wrong-estimator reuse under the correct-answer condition, all eight reconstruction runs produce the same wrong-estimator rule. In the remaining six instances, reconstructions are more diverse and wrong-rule fractions range from 0 to 0.5. The reuse instances are therefore distinguished by complete concentration on a single wrong-rule reconstruction. This exploratory association does not establish that the reconstruction causally mediates the final response.

D.7 Where the artifact is not recoverable, the number does nothing

The two operators were chosen for exact estimator ground truth, not to contrast recoverability, but they do contrast it. The E1 artifact displays β_{obs}, so an archived `AO_wrong` value identifies the wrong estimator by arithmetic; the MA artifact withholds the (y_k, σ_k) pairs that either candidate output would require (Remark 1). The 2×5 factorial supplies the comparison: the numeric-presence effect is $+0.367$ in the recoverable E1/Qwen cell and $+0.025$ and -0.067 in the two non-recoverable MA cells. The factorial scores with the frozen numeric probe, which states no decoding rule (§C.2), so the E1/Qwen effect here is a second measurement of the positive cell that no decoding instruction produced.

What the operator contrast does and does not add. E1 and MA differ in more than recoverability: estimand, domain language, artifact schema and arithmetic. The contrast alone does not isolate recoverability, and we do not rest the mechanism on it; the intervention of §D.4 does that within a single operator. What the contrast adds is that recoverability was decidable from the artifact schema *before* any outcome was observed, and that the one operator where it fails shows no numeric effect.

Recoverability is necessary, not sufficient. The E1/DeepSeek-V3.2 cell is equally recoverable and its numeric effect is -0.025. DeepSeek-V3.2 carries a high no-artifact floor for the familiar estimator, so the baseline is already elevated and every contrast compresses (§D.12). A numeric effect appears where the artifact is decodable *and* the model has room to move; neither half is a property of the number.

D.8 A broad sweep under a different probe shows no numeric effect

Across 18 fixed-bank configurations — three model families (Qwen3.5-27B, DeepSeek-V3.2, GLM-4-9B), both operators, four transfer prompts, three seeds per operator–family group — the pooled carrier contrasts are

$$\Delta_{\text{shell}} = +0.03, \quad \Delta_{\text{noart}} = -0.03, \quad \Delta_{\text{random}} = +0.02, \quad \Delta_{\text{agreement}} = +0.03, \quad \Delta_{Y} = +0.03,$$

with one-sided lower bounds -0.02, -0.11, -0.04, -0.04 and -0.04 respectively. **Every bound crosses zero.** Wrong-method signatures are common in these configurations, and the matched controls absorb all of them.

This is the measurement with the widest configuration coverage in the paper, and under its probe the matched controls absorb every numeric effect. In the same configurations the explicit-rule positive control reaches a mean wrong-signature rate of 0.50 (and 0.79 for Qwen3.5-27B across both operators), so the assay is not deaf — it registers semantic method uptake in the same cells where it registers no numeric effect. Full per-seed contrasts are in Appendix L.2 and the pooled decomposition in Appendix L.1.

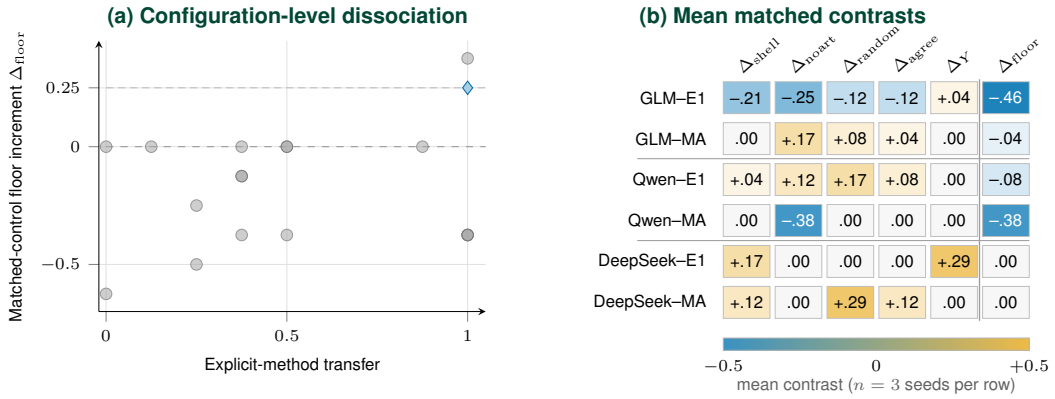


Figure 7: Explicit method transfer and answer-specific attribution separate. Panel (a) plots each model–operator–seed cell at P3; the dashed line marks $\delta = 0.25$, and the blue diamond marks the single model–operator–seed cell in which all five observed matched contrasts are positive; markers are semi-transparent, so coincident configurations render darker. Panel (b) is a diverging heatmap of matched-control contrasts averaged over three seeds within each operator–family group; warm cells indicate the wrong-answer artifact exceeding the alternative, cool cells the reverse, near-white cells exact zeros. Δ_{floor} is shown only as a descriptive summary; inference is based on the five primary contrasts.

Why the sweep is null, read against the conjunctive statement. The sweep is not the positive cell re-measured under another probe; for most of it the account of §D.9 predicts no effect. Its eighteen configurations are three worlds for each model–operator pair. The nine MA configurations are not recoverable (Remark 1). The three E1/GLM-4-9B configurations come from a model in which the explicit-rule positive control does not transmit, so they are excluded from negative claims (§C.7). The three E1/DeepSeek-V3.2 configurations have a no-artifact naive rate of 0.500 in every world, which leaves no room (Table 40). Only the three E1/Qwen3.5-27B worlds meet both conditions, and at eight probes per condition the fixed bank misses a true quarter-scale effect about half the time (Appendix I); their Δ_Y values straddle zero (Table 37). Nor is the null a property of the sweep's probe: the 2×5 factorial scores with the same frozen numeric probe, which states no decoding rule, and registers a numeric effect of $+0.367$ in the recoverable E1/Qwen cell and none in MA (§D.7). What the sweep cannot show is whether its three E1/Qwen worlds would reach the positive cell's effect with more probes, because no experiment varies the probe with worlds and model held fixed; an exploratory fixed-bank arm scored with the decoding-rule probe is not null (Appendix K.3). None of this bears on the mechanism, whose evidence (§D.4–§D.5) is scored at the memo stage, by no probe.

The fixed bank quantises condition rates in steps of $1/8$, so a reader may reasonably ask whether the null is a resolution artifact. Appendix I answers this with a full-gate Monte Carlo of the actual decision rule: at $n = 8$ probes the zero-effect pass probability is $\approx 1\%$, primary-effect power at a true $\delta = 0.25$ is ≈ 0.50, and operational-margin power at $n = 8$ requires a true effect approaching 0.75. So the fixed bank can miss a true quarter-scale effect about half the time. It is underpowered for the operational margin, which is one reason the positive cell was re-measured with 32 world-specific probes.

The probe distributions also differ. The sampled probes differ from the fixed bank in the oracle's confounding gap $|\beta_{\text{obs}} - \tau|$: sampled probes span $[0.800, 2.824]$ with *no probe below* 0.5, whereas the fixed bank contains two probes at exactly 0.5. Within the sampled regime the effect is flat across gap tertiles (wrong-signature rates $0.678/0.667/0.741$; mean within-world correlation between the wrong-signature indicator and the gap is 0.004), which rules out the effect being concentrated in a narrow high-gap band *inside the sampled range*. It does not rule out the fixed bank's low-gap region being where the effect dies, because no sampled probe covers it. **"Regime effect" and "easier probes" are not separated by anything we ran**, and a three-arm bridge diagnostic (Appendix K.3) was inconclusive under its own attribution table.

Remark 3 (The low-gap region is unreachable by construction, not undersampled)**.** The obvious repair—sample probes until the low-gap region is covered—cannot work, and the reason is a property of the generator rather than of the sample. The world sampler draws (a, c, σ_X^2) from a declared finite grid and keeps a world only when $|\beta_{\text{obs}} - \tau| \geq 0.75$, and since $|\beta_{\text{obs}} - \tau| = |ac|/(a^2 + \sigma_X^2)$, the attainable gap set is finite. Enumerating it exhaustively in exact rational arithmetic gives 30 values spanning $[0.800, 2.824]$; 0.5 is absent *before* the separation filter is applied, and the nearest attainable value is 0.800. The fixed bank reaches 0.5 only through $c = \pm 1$, which the sampler's c-grid does not contain. Two consequences follow. First, the sampled range we observe over 32 worlds, $[0.800, 2.824]$, coincides with the analytically attainable range to machine precision, so the sampled bank has already saturated its support and more worlds add nothing. Second, the gap-matched bank is not an experiment this design can run: matching onto 0.5 requires a different world generator, which would change the construct being compared rather than hold it fixed. The feasible restriction—matching onto $\{1.0, 1.5, 2.0\}$, where the sampler does have mass—is Arm B of the bridge, and it was run: it clears Δ_{floor} and fails Δ_Y (Appendix K.3). The two banks therefore probe disjoint regions of gap by construction, we report both, and we attribute neither.

D.9 One conjunctive statement is consistent with every measurement

Taken together, every numeric-effect measurement we made is consistent with a single conjunctive statement, and none contradicts it:

> An archived number shifts downstream estimator choice only when *both* (i) the artifact makes the estimator recoverable—the archived value appears alongside the quantity it was computed from, so their agreement identifies a method (Remark 1) —*and* (ii) the assay has room to move, that is, the naive estimator's rate without any artifact is not already high. Remove either condition and the effect is absent. The two conditions are necessary, not sufficient: whether a model takes a recoverable estimator also depends on the model and on how its output is framed.

Table 16 lays the measurements against it. Condition (i) is decidable from the artifact schema before any model is called, and (ii) is the measured `NO_ARTIFACT` rate. The rows marked as interventions remove condition (i) inside the one cell where both hold, with model, prompt, worlds and probe fixed; the other rows are observational, and they differ in more than (i) and (ii)—in particular in the probe, which the table records. The necessity half has the widest support: no row, across two operators, three families, two Qwen generations and two output formats, shows an effect without (i). The sufficiency half fails outside Qwen: under the numeric-commitment format (A4), Mistral Small 3.2 24B meets both conditions and responds to the positive control, yet shows no effect. The statement therefore fixes the shape of the finding without widening its scope: the positive effect is supported on one model family for (ii) and one operator for (i), and Criterion 2 fires on exactly that.

D.10 The semantic ladder is a ruler, not a finding

The 2×5 factorial crosses numeric presence with five levels of semantic method information over 12 worlds in four model–operator cells. Cue-versus-none contrasts on the estimator-signature scale are: explicit rule $+0.604$ $[+0.500, +0.693]$, weak diagnostic hint $+0.531$ $[+0.484, +0.563]$, partial formula $+0.297$ $[+0.250, +0.339]$, estimator name $+0.135$ $[+0.109, +0.156]$ (resampling ranges over three worlds per stratum, §C.7).

We do not report these as discoveries. `EXPLICIT_WRONG_METHOD` is the assay's own positive control, and the hint and formula rungs are explicit method descriptions. That telling a model which

Table 16: Every numeric-effect measurement against the conjunctive statement. Condition (i) is read from the artifact schema, condition (ii) from the measured `NO_ARTIFACT` rate. "Probe": *frozen* = frozen numeric probe, no decoding rule; *decoding* = the decoding-rule probe of §C.2; *none* = memo stage, post hoc except where marked pre-registered. Rows marked *exploratory* are not confirmatory evidence.

Measurement	(i)	(ii)	Probe	Numeric effect
MA, both families (§D.7)	no	—	frozen	$+0.025$ / -0.067
E1 / GLM-4-9B (§D.8)	yes	no (unresponsive)	frozen	null; excluded from the negative claim
E1 / DeepSeek-V3.2 (§D.7)	yes	no (0.500)	frozen	-0.025
E1 / Qwen, factorial (§D.7)	yes	yes	frozen	$+0.367$
E1 / Qwen / P3, fixed bank, 8 worlds (App. K.3, *exploratory*)	yes	yes	decoding	all five sim. LBs ≥ 0.297
E1 / Qwen / P3, Study 2 (§D.3)	yes	yes	decoding	all five sim. LBs ≥ 0.298
E1 / Qwen / P3, Study 2 (§D.3)	yes	yes	none	all five sim. LBs $\geq +0.500$
E1 / Qwen3.5 / P3 + final line (A4; pre-registered)	yes	yes	none	Δ_Y^{shown} $+0.131$
E1 / Mistral / P3 + final line (A4; pre-registered)	yes	yes	none	Δ_Y^{shown} $+0.044$, not detected
E1 / Qwen / P3, slope withheld (§D.4)	**no**, by intervention	yes	decoding	DiD $+0.511$
E1 / Qwen / P3, slope withheld (§D.4)	**no**, by intervention	yes	none	DiD $+0.894$
E1 / Qwen / P3, non-identifying statistic, artifact complete (§D.4; pre-registered)	**no**, by intervention	yes	none	DiD $+0.838$
E1 / Qwen3.8 / P3, non-identifying statistic (§D.5; pre-registered)	**no**, by intervention	yes	none	DiD $+0.457$
E1 / Qwen3.5 / P3 + final line, non-identifying (A4; pre-registered)	**no**, by intervention	yes	none	Δ_Y^{nonid} $+0.000$
E1 / Mistral / P3 + final line, non-identifying (A4; pre-registered)	**no**, by intervention	yes	none	Δ_Y^{nonid} -0.019

estimator to use moves its estimator choice more than showing it a number is close to a priori, and a paper whose headline is that ordering is reporting a sanity check. Their function here is metrological: they put a known-strong intervention on the same exact scale as the numeric channel, so that a numeric effect can be read as a fraction of a full method statement rather than in the abstract. On that ruler, the recoverable-cell numeric effect $+0.367$ is roughly 60% of an explicit rule – which is the only sense in which the comparison is informative.

D.11 Numeric presence is one cell, not an average

The numeric-presence effect by model–operator cell is $+0.367$ (E1/Qwen3.5-27B), -0.025 (E1/DeepSeek-V3.2), $+0.025$ (MA/Qwen3.5-27B) and -0.067 (MA/DeepSeek-V3.2), with a model–operator interaction of $+0.300$ $[+0.117, +0.517]$ that excludes zero. A leave-one-world-out check keeps the pooled average positive in all twelve deletions (Appendix H.7).

The equal-weight average across cells is $+0.075$ with stratified resampling range $[+0.027, +0.127]$, and we do not report it as a numeric effect. Averaging across a declared qualitative interaction — one cell at $+0.367$, three at or below zero — produces a number that describes no cell in the design. The finding is $+0.367$ in E1/Qwen3.5-27B; $+0.075$ is a statement about *non*-generalisation and should be read only as such. The within-stratum no-cue contrast is $+0.010$, i.e. numeric presence alone, with no semantic cue anywhere in the artifact, moves essentially nothing on average.

No simultaneous interval was computed for the cue-versus-none contrasts; the numbers above are percentile resampling ranges and are labelled as such.

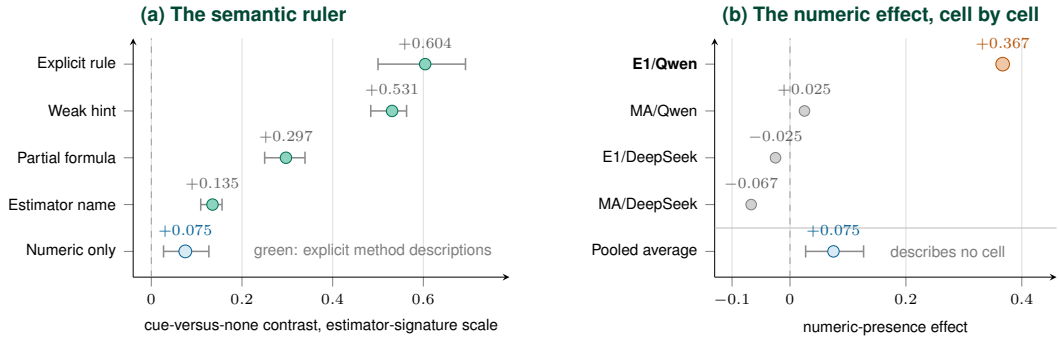


Figure 8: Numeric presence against the semantic ruler. Panel (a) places the four semantic cue levels and the numeric channel on one estimator-signature scale, ordered by magnitude; bars are resampling ranges over three worlds per stratum (§C.7), not confidence intervals. The green rungs are explicit method descriptions and the top one is the assay's own positive control, so their ordering above a number is close to a priori — the panel is a ruler, not a finding (§D.10). Panel (b) reads the numeric channel off that scale within each model–operator cell. One cell of four departs from zero, and the equal-weight average at the bottom, separated by a rule, is a statement about *non*-generalisation rather than an effect size (§D.11).

D.12 Robustness within the tested cell

Five paraphrases of the P3 transfer instruction preserve the same qualitative separation: explicit transfer ranges 0.875–0.917, answer-only remains at 0.042, mean gap $+0.867$, template SD 0.019. Six domain-grounded E1 scenarios with deterministic ground truth show the same pattern: the explicit rule transfers and the answer-only artifact stays at the no-artifact/shell floor. Both establish insensitivity to *surface wording* within P3 and within the E1 estimator structure. Neither addresses prompt level (§C.4), operator, model family, or probe distribution, and we do not present them as if they did.

Model-family profiles. The three families produce three different reasons the assay behaves as it does rather than three replications. Qwen3.5-27B pairs a low no-artifact floor with a responsive positive control, which is the only configuration in which a numeric effect has room to appear. DeepSeek-V3.2 has high task-prior floors — consistent with a strong prior for the familiar estimator (OLS for E1, arithmetic mean for MA) — which raises the baseline independently of artifact content and compresses every contrast. GLM-4-9B is not positive-control-responsive, so by §C.7 its contrasts cannot support a negative carrier claim. Reported as a panel, the sweep shows heterogeneity in assay validity, not external validity for the effect; the effect itself is heterogeneous as well, since an eligible family, Mistral Small 3.2 24B, shows none under the numeric-commitment format (§D.5). The three profiles are what Fig. 9 separates, and they are the reason §D.7's mechanism statement has to be conjunctive: recoverability decides whether a numeric effect is *possible*, the floor decides whether it is *visible*.

D.13 Reading the measured numbers

Putting the measurements together in the order they constrain each other:

1. In Qwen3.5-27B on causal-effect estimation under an explicit reuse instruction, an archived wrong answer makes the model use the wrong estimator on a new problem, far above every matched control including a correct archive: at the memo stage, uninstructed, and at the pre-registered probe stage, on the original worlds and on fresh ones (§D.1–§D.3).
2. Withholding the parent slope, with the archived number, the worlds, the prompt and the probe unchanged, removes the effect at both stages (§D.4). The number is inert on its own.
3. The same account covers the two places the effect is absent or unexpected: an operator whose artifact is not recoverable shows no numeric effect (§D.7), and a correct archive leaks in exactly the worlds where the model rebuilds the naive rule from the visible slope (§D.3).

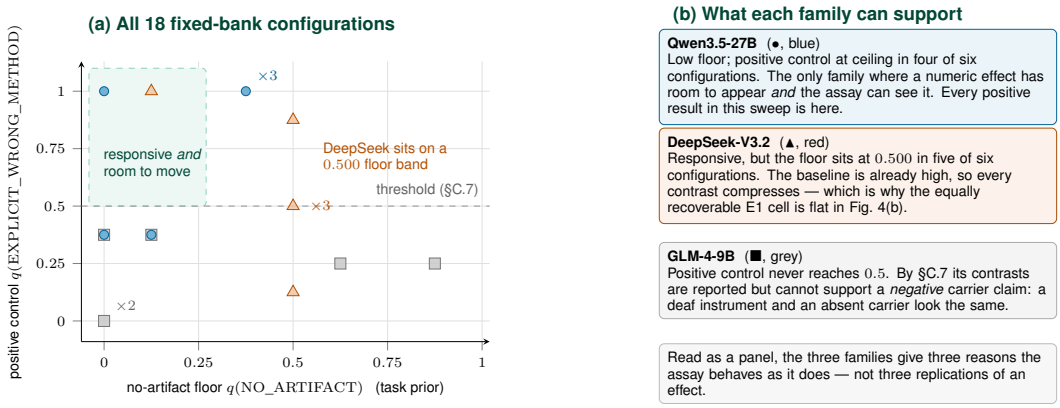


Figure 9: In the fixed-bank sweep, model-family heterogeneity is heterogeneity in assay validity. Panel (a) plots every one of the 18 fixed-bank configurations by its no-artifact floor and its explicit-method positive-control rate; each point is a literal model–operator–seed row of Tables 38 and 41, and $\times k$ marks coincident configurations; a blue circle inside a grey square is a Qwen and a GLM configuration at the same coordinates. The shaded corner is the only region in which a carrier claim is both possible and observable: the assay must be responsive (above the 0.5 threshold of §C.7) and the task prior must leave headroom. Panel (b) reads off what each family therefore licenses. This is the second conjunct of the mechanism statement in §D.7: recoverability decides whether a numeric effect is possible, the floor decides whether it is visible. Outside the sweep, an eligible family can still show no effect (Mistral Small 3.2 24B, §D.5).

4. Under a probe that asks the model to follow its memo but states no decoding rule, a broad sweep over three families, two operators and four prompt levels shows no numeric effect (§D.8); that sweep is a boundary under its own probe, not a comparison with the positive cell.

The composite claim we defend is therefore: *an archived scientific result can transmit its estimator to a new problem at inference time, and where it does, it does so because the surrounding artifact lets the estimator be re-identified—remove the quantity the number was computed from and the same number does nothing.* It is established in one model, one task and one instruction level. We do not claim a graded carrier hierarchy, a general answer-specific channel, or the cross-cell average $+0.075$ as a numeric effect.

E REGISTERED AND OPEN

One experiment is registered with a frozen design and has not produced a result; every clause naming it is marked † and claims nothing. (The complete-artifact control and its replication in Qwen3.8-27B were registered the same way; both have run and are reported in §D.4 and §D.5.)

1. **Memo surgery.**† Four deterministic edits of archived memos—unedited, carrier sentence excised, a length-matched neutral sentence excised, carried value replaced by τ—re-probed with the same probes, to identify which property of the memo carries the choice on to the probes. The paper makes no claim about that property.

Outside this paper's scope. Each of the following is a different measurement rather than a missing arm, and no claim here rests on it. *Transfer instruction:* all positive evidence is at P3; whether the memo-stage effect appears at P0–P2 is untested. *Model families:* the positive evidence comes from one responsive family at 27B and no frontier-scale backbone. One eligible family outside Qwen shows no effect under the numeric-commitment format (§D.5); which models and output formats take a recoverable estimator is left to future work. *Candidate sets:* recoverability is defined relative to a two-member candidate set; harder decoding in a larger set does not imply a weaker opaque channel, so no bound here is extrapolated to one. *Probe regimes:* separating the regime and probe-difficulty readings of the fixed bank requires a world generator whose confounding-gap support spans both banks (Remark 3); the registered gap-matched bank is withdrawn as infeasible for that reason. *Realism:* real archived results give up the exact oracle that makes attribution measurable; six domain-grounded E1 scenarios (Appendix M.2) test surface realism only. *Training-time*

transmission is the subliminal-learning question (Cloud et al., 2026; Blank et al., 2026; Madl, 2026) and does not substitute for inference-time reuse. *Mitigations* are untested, beyond one observation: requiring a final numeric estimate or "not identifiable" shrinks the effect in Qwen3.5-27B (§D.5).

F Limitations and Conclusion

Limitations, in order of how much they cost the claim. *(1) Which outcome was pre-registered.* The Study 2 and slope-withheld protocols name the probe-stage rate, and the probe states the decoding rule (§C.2). Their memo-stage readings, including the comparison against the five controls, were computed afterwards with a frozen scorer and remain post hoc. The mechanism's two memo-stage interventions are confirmatory: the complete-artifact control pre-registered the memo stage as primary, and it replicates the withheld result (§D.4). *(2) Scope.* Every positive measurement comes from one model family on one operator and one prompt (E1/P3): Qwen3.5-27B, and for the mechanism also Qwen3.8-27B (§D.5). Two generations of one family replicate the finding without widening it across families, and Criterion 2 fired. The other two families we measured are not counterexamples but ineligible tests: GLM-4-9B does not respond to the explicit-rule positive control and DeepSeek-V3.2 leaves no room for an effect, both of which the replication protocol's eligibility gate excludes (§D.8). A third, Mistral Small 3.2 24B, is ineligible at the memo stage, which does not observe its estimator choice; under a registered format that does (A4), it is eligible and shows no effect, while the same format shrinks the effect in Qwen3.5-27B (§D.5). The effect is therefore heterogeneous across families: recoverability is necessary for it, not sufficient. *(3) What "identifiability" is operationally.* The interventions remove the agreement between the archived value and a displayed slope, with the artifact kept complete (§D.4). They do not display a number that equals the archived value but is not a slope, so agreement read as estimator identity is not separated from numeric agreement as such; that the memo applies the naive rule to the *new* world's slope, which the archived value does not match, favours the former. *(4) Prompt.* All positive evidence is at P3, which asks the model to carry forward the parent's approach (§C.4); one added output-format line shrinks the effect (§D.5). *(5) Probe.* The positive cell and the broad sweep use different probes and probe distributions, so the sweep is not a comparison with the positive cell (§D.8). *(6) Statistics.* The eight-world analyses cannot support calibrated bootstrap coverage; there we lean on an exact sign-flip test whose floor is 2^{-8} and on leave-one-world-out. The 32-world analyses use Monte Carlo sign-flip tests at 200,000 draws, whose smallest reportable p is $1/(B{+}1)$; the 2×5 intervals are resampling ranges rather than confidence intervals (§C.7). *(7) Models and operators.* Four open-weight families, no frontier-scale backbone, and the one responsive family has since been superseded; E1 and MA prioritise exact estimator ground truth over realism.

Design implications (untested). If estimator recoverability rather than opaque transmission is the operative channel, three practices follow, none of which this paper evaluates as an intervention; one related observation is that asking the model to end with a numeric estimate or "not identifiable" shrank the effect in Qwen3.5-27B (§D.5). Store minimal method provenance — estimator identity, assumptions, uncertainty — alongside the value using standard vocabularies (Moreau & Groth, 2013). Verify estimator continuity by independent re-derivation or agreement probes rather than treating recurrence as evidence of transmission; §D.8 is precisely a case where recurrence occurs and transmission does not. And note that a *correct* archived answer was not safe in our assay: it induced the wrong estimator in 10 of 32 worlds, at a rate predicted by whether the model's own memo adopted the naive rule ($r = 0.89$) and not by the numeric geometry of the world ($r = 0.33$), so filtering handoffs on answer correctness would not have prevented the failure we measured.

Conclusion. An archived wrong answer can make a downstream model use the wrong estimator on a new problem, and a correct-answer control shows that this is specific to the answer rather than to the presence of one. But the number does not carry the method by itself. It carries it the way a receipt carries a price: next to the observational slope it was computed from, its agreement with that slope identifies the estimator, and the model does the arithmetic—in its own estimate for the new problem, without being told how. Remove the slope and leave the number verbatim, and the effect is gone: $+0.894$ on the model's own estimate, $+0.511$ on the pre-registered probe outcome. Replace the slope with a statistic that identifies neither estimator, so that the artifact is complete and only identifiability is lost, and it is gone as fully ($+0.838$, pre-registered; $+0.457$ in a newer model generation). The same account explains why a correct archive is not safe and why an operator

whose artifact is not recoverable shows nothing. The practical form of the finding is therefore not "do not archive numbers" but "an archived number is inert exactly when its provenance is gone, and dangerous exactly when enough provenance survives to re-identify the method."

G Formal Derivations and Probe Banks

This appendix provides formal derivations, complete probe and condition definitions, reproducibility details, world-level estimates, and supporting robustness analyses. Analyses not identified as held-out in the main text are descriptive and are used to characterize heterogeneity or sensitivity rather than to establish additional paper-level claims.

G.1 E1 Observational Slope Derivation

The E1 assay uses a linear-Gaussian structural causal model:

$$U \sim \mathcal{N}(0,1), \quad X = aU + \epsilon_X, \quad Y = \tau X + cU + \epsilon_Y,$$

where $\epsilon_X \sim \mathcal{N}(0, \sigma_X^2)$ and $\epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$, with all noise terms independent.

Interventional effect. By the back-door criterion, adjusting for U blocks the confounding path $X \leftarrow U \rightarrow Y$. Thus:

$$\frac{\partial}{\partial x}\mathbb{E}[Y \mid \mathrm{do}(X = x)] = \tau.$$

Observational slope. The observational regression coefficient of Y on X is:

$$\beta_{\mathrm{obs}} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}.$$

We compute:

$$\begin{aligned}\mathrm{Var}(X) &= a^2\mathrm{Var}(U) + \mathrm{Var}(\epsilon_X) = a^2 + \sigma_X^2,\\ \mathrm{Cov}(X,Y) &= \mathrm{Cov}(aU + \epsilon_X,\ \tau(aU+\epsilon_X) + cU + \epsilon_Y)\\ &= \tau a^2 + \tau\sigma_X^2 + ac = \tau(a^2+\sigma_X^2) + ac.\end{aligned}$$

Therefore:

$$\beta_{\mathrm{obs}} = \tau + \frac{ac}{a^2+\sigma_X^2}.$$

The confounding bias is $\beta_{\mathrm{obs}} - \tau = \frac{ac}{a^2+\sigma_X^2}$, which is nonzero whenever $a \neq 0$ and $c \neq 0$.

G.2 Probe Bank Construction

The fixed E1 probe bank consists of 8 discriminative probes and 4 agreement controls, specified by parameter tuples (τ, a, c, σ_X^2):

Agreement controls set $a = 0$, eliminating the confounding path and making $\beta_{\mathrm{obs}} = \tau$. A child artifact is wrong-signature-positive when it gives a numeric causal estimate matching β_{obs} rather than τ on discriminative probes. The bank is designed so that neither probe class can be dropped without losing meaning: the eight discriminative probes span confounding-bias magnitudes from 0.5 to 2.0 with both signs of c, so detection does not hinge on one bias magnitude, and the four agreement controls make the naive estimator coincide with the truth, so an artifact that reproduces β_{obs} on discriminative probes while remaining correct on controls evidences method uptake rather than numerical luck.

The bank fixes $a = 1.0$ and $\sigma_X^2 = 1.0$ on all discriminative probes, so the confounding bias $\beta_{\mathrm{obs}} - \tau = ac/(a^2+\sigma_X^2) = c/2$ is determined entirely by c; this isolates the estimator distinction without varying the signal-to-noise structure of X. Varying a or σ_X^2 would rescale the bias magnitude without qualitatively altering the correct/wrong estimator signatures, but a richer bank spanning multiple (a, σ_X^2) combinations would be needed to test whether the carrier effect interacts with confounding strength.

Table 17: E1 probe bank parameters.

ID	Class	τ	a	c	σ_X^2	$\beta_{\text{obs}} - \tau$
P01	discr.	-1.0	1.0	4.0	1.0	$+2.0$
P02	discr.	$+1.0$	1.0	-4.0	1.0	-2.0
P03	discr.	-0.5	1.0	3.0	1.0	$+1.5$
P04	discr.	$+0.5$	1.0	-3.0	1.0	-1.5
P05	discr.	$+1.0$	1.0	2.0	1.0	$+1.0$
P06	discr.	-1.0	1.0	-2.0	1.0	-1.0
P07	discr.	$+0.5$	1.0	1.0	1.0	$+0.5$
P08	discr.	-0.5	1.0	-1.0	1.0	-0.5
P09	agree.	$+1.25$	0.0	4.0	1.0	0.0
P10	agree.	-1.25	0.0	-4.0	1.0	0.0
P11	agree.	$+0.75$	0.0	2.0	0.5	0.0
P12	agree.	-0.75	0.0	3.0	1.5	0.0

G.3 Artifact Conditions, and Conditions Not Reported

Table 6 in the main text is the exhaustive manifest of conditions scored in this paper. This appendix records what the condition grammar can express beyond that manifest, and which implemented conditions are not reported.

Grammar. Every condition is a template over the tuple ($\mathcal{S}(W_p)$, method field, value field, policy field) instantiated by deterministic code from the parent world. Baselines empty the method and value fields to different depths (`NO_HISTORY` removes the archived analysis entirely; `NO_ARTIFACT` removes the artifact but keeps the reuse framing; `ARTIFACT_SHELL` asserts an analysis exists and withholds both fields). Answer controls hold the method field empty and vary only the value. Method carriers hold the value field fixed and grade the method field from an estimator name to a full rule. This one-field-at-a-time property is what licenses reading a contrast as an intervention on a single candidate carrier.

Conditions not reported, and why. The fixed-bank manifest contains six conditions beyond those of Table 6. They are an `AO_blank` wording check on `NO_ARTIFACT`; four no-estimate-policy conditions (`AO_wrong`/`AO_correct` crossed with a full provenance-plus-action policy, plus the two 2×2 ablations isolating the provenance flag and the action prohibition); and a `METHOD_AVAILABLE` false-abstention baseline. Those six conditions were implemented and executed, but only in an exploratory single-family pivot on NVIDIA Nemotron Nano 9B, and no result from them appears in this paper. The suppression finding that pivot produced — a no-estimate policy reducing wrong-signature behaviour in completed Nemotron rows — is a separate, single-family, exploratory claim, and the workflow recommendations in §F are consequently labelled as untested design implications rather than as evaluated interventions.

Scoring note retained from that arm. For completeness: policy-unavailable conditions were scored with an additional explicit-abstention channel, under which a response counted as no-estimate compliant only when it declined to produce a numeric intervention estimate *because* the estimator policy or method provenance was unavailable, and invalid formatting was never counted as a safety success. No number produced under that scoring rule is used anywhere in this manuscript.

G.4 Response Validity and Aggregation

Each cell contributes its first usable model response. Calls that end before returning a usable response because of provider, network, transport, or timeout failures are rerun. Substantive model behavior, including incorrect estimates, agreement failures, and abstention choices, remains part of the analysis. Aggregation is performed from validated cell-level summaries, with exclusions recorded alongside the corresponding run metadata.

H MODELS, PROVIDERS, AND REQUEST-LEVEL PROVENANCE

The three model families used in the factorial and third-family analyses are documented in the GLM-4 (Team GLM et al., 2024) and DeepSeek-V3.2 (DeepSeek-AI, 2025) technical reports and the official Qwen3.5-27B model card (Qwen Team, 2026). The OpenRouter cells pin a single upstream provider (DeepInfra) with fallbacks disabled, and every billed call is reconciled against the provider's billing record by generation identifier. Table 18 distinguishes these current analysis cells from earlier heterogeneity and exploratory cells.

Table 18: Model cells and their role in the analysis.

Cell	Provider family	Role in current analysis
Nemotron Nano 9B	NVIDIA NIM	initial exploratory pivot; no result reported here
GLM-4-9B	SiliconFlow	factorial artifact-contrast experiment
Qwen3.5-27B	SiliconFlow	factorial artifact-contrast experiment
Qwen3.5-27B	DeepInfra via OpenRouter	32-world study, slope-withheld, complete-artifact control, A4 anchor
Qwen3.8-27B	DeepInfra via OpenRouter	replication (amendment A1)
Mistral Small 3.2 24B	DeepInfra (fp8) via OpenRouter	second family: memo stage not eligible (A3); numeric-commitment format eligible, no effect (A4); bf16 attempt (A2) archived unscored
DeepSeek-V3.2	SiliconFlow	third-family validation
Qwen3-14B	Qwen via OpenRouter	earlier assay version
Llama-3.1-8B	Llama via OpenRouter	historical heterogeneity cell; not main evidence
Gemma 4 26B	Gemma via OpenRouter	historical non-supporting heterogeneity cell; not main evidence

Table 19: Exact request-level inference controls. "Not sent" means the client omitted the field; "not exposed" means the provider did not supply an immutable value to the experiment. No unspecified field is described as a controlled default.

Provider model ID	Provider	temp.	max tok.	thinking field	top-p	top-k	stop	Endpoint
`nvidia/nvidia-nemotron-nano-9b-v2`	NIM	0	4096	not sent	not sent	not sent	not sent	`integrate.api.nvidia.com/v1`
`THUDM/GLM-4-9B-0414`	SiliconFlow	0	2048	not sent	not sent	not sent	not sent	`api.siliconflow.cn/v1`
`Qwen/Qwen3.5-27B`	SiliconFlow	0	1024	`enable_thinking=false`	not sent	not sent	not sent	`api.siliconflow.cn/v1`
`deepseek-ai/DeepSeek-V3.2`	SiliconFlow	0	1024	not sent	not sent	not sent	not sent	`api.siliconflow.cn/v1`
`qwen/qwen3.5-27b`	DeepInfra	0	1024	`reasoning.enabled=false`	not sent	not sent	not sent	`openrouter.ai/api/v1`
`qwen/qwen3.8-27b`	DeepInfra	0	4096	`reasoning.enabled=false`	not sent	not sent	not sent	`openrouter.ai/api/v1`
`mistralai/mistral-small-3.2-24b-instruct`	DeepInfra	0	4096	`reasoning.enabled=false`	not sent	not sent	not sent	`openrouter.ai/api/v1`

Request and response provenance. Requests use the OpenAI-compatible `/chat/completions` endpoint through a project-local HTTP client (`httpx`, with `curl` fallback), not a provider SDK; hence SDK version is not applicable. Provider-side chat-template versions, immutable model snapshots, reasoning-effort controls, and server build identifiers were not exposed. System and artifact prompts are preserved in source/configuration artifacts and content hashes, but historical pivot JSONL rows do not contain a separate system-prompt hash or request timestamp; those fields are therefore not claimed as recorded. The response parser consumes only `message.content`. It neither appends nor stores `reasoning_content`; a reasoning-only response with empty content is treated as an infrastructure failure. Carrier memos consequently exclude provider reasoning traces. A thinking-on robustness run has not been executed and is not part of the current evidence.

Seed identifiers. Seed identifiers are arbitrary integer RNG seeds and do not encode execution dates. For example, `20260816–20260820`, `20260831–20260902`, and `20260903–20260908` are RNG seeds, not calendar dates. Model identifiers, conditions, prompts, content hashes, and scoring parameters are present in the immutable logs. Historical pivot rows do not contain an API request timestamp and must not be presented as if they do.

Deterministic generation. The probe bank, artifact templates, and prompt templates are generated by deterministic code; decoding parameters, the exact scoring tolerances, and the full prompt/artifact

templates are recorded in the repository configuration and run scripts, which are released alongside the results for independent audit.

H.1 Analysis Protocol and Provenance

The held-out analyses use run designs and classification rules recorded before outcome inspection. Repository history preserves subsequent criterion revisions.

H.2 Independent Re-implementation of the Stage-2 Analysis

The result in §D.3 is the one the paper most depends on, and it is produced by code we wrote. Before reporting it we re-derived it with a second program, `scripts/verify_stage2_analysis.py`, which imports nothing from the analysis module, re-reads the raw probe records from disk, and uses the standard library's random-number generator rather than the analysis module's. It performs four checks, all of which the reported numbers must survive.

1. **Independent recomputation.** The same contrasts and Bonferroni-simultaneous bounds are computed from scratch with a different RNG stream. Largest disagreement across the five contrasts and their simultaneous bounds: 0.0068. Bonferroni bounds are extreme quantiles at per-contrast $\alpha = 0.01$ and are noisier than the means they bound, so this is the resolution of $4{,}000$ bootstrap draws rather than a discrepancy between the implementations; the means agree to 0.0017.
2. **Equivalent form.** Each contrast is computed a second way—as the mean of per-probe differences rather than as a difference of condition means—which is algebraically identical for matched probes and therefore must agree exactly. It does: maximum absolute difference 0 to machine precision, which also confirms that probe identities really are matched across conditions rather than merely equal in count.
3. **Reverse control.** With one execution per world the three-level bootstrap must degenerate onto the two-level bootstrap of §D.1. Forcing $K = 1$ reproduces the two-level bounds to within 0.0021, so the new variance component is the only thing the new estimator adds.
4. **Zero point and null calibration.** On synthetic data with a known contrast the procedure recovers 0 to within 0.021 and 0.20 to within 0.009; under the null the world-level sign-flip test rejects at exactly its nominal rate, $P(p \leq 0.05) = 0.050$ over 400 replicates.

We record one negative episode from this process because it is the reason the check exists. An interim reading of the incomplete matrix put Δ_Y's simultaneous bound at 0.239, below δ; the verifier disagreed, and the disagreement was traced to the incomplete data rather than to either implementation. The number in Table 10 is from the complete matrix and is the one both implementations return. No Stage-2 quantity was reported from partial data.

H.3 Independent Re-implementation of the Slope-Withheld Analysis

§D.4 is the experiment that decides how this paper's central claim is phrased, so as with §D.3 we re-derived it with a second program, `scripts/verify_slope_withheld.py`, which imports nothing from the analysis module and uses the standard library's random-number generator rather than the analysis module's.

1. **Independent recomputation.** Bootstrap means agree to 0.0032 on all three estimands; the three sign-flip p-values agree to better than 10^{-4}, including the two-sided test on the DiD.
2. **Equivalent form.** The DiD is computed a second way—as the mean over probes of the per-probe double difference, rather than as a difference of differences of cell means. The two routes are algebraically identical for matched probes and must agree exactly. They do: 1.1×10^{-16}, machine precision. This also confirms that probe identities are genuinely matched across all four cells rather than merely equal in count, which the verifier enforces by refusing any unit whose four probe-id lists differ.
3. **Reverse control.** The shown arm reproduces the separately executed §D.3 run of the same conditions in the same worlds: $q(\texttt{AO_wrong})$ is 0.5182 here against 0.5084 there, within

execution-to-execution variation already measured at a mean within-world standard deviation of 0.072.

4. **Zero point and null calibration.** On synthetic data with a known DiD the procedure recovers 0 to within 0.036 and 0.30 to within 0.026; under the null the world-level sign-flip test rejects at $P(p \leq 0.05) = 0.065$ over 400 replicates against a nominal 0.05.

A separate audit, `scripts/analyze_slope_withheld_mediator.py`, produces the response-validity and memo figures quoted in §D.4. It exists because the DiD alone cannot distinguish "withholding the slope removed the identifying information" from "withholding the slope broke the task and the scorer counted the non-answers". The second reading predicts a rise in absent or invalid responses in the withheld arms; it does not occur, and answers move to the correct estimator instead.

H.4 Complete-Artifact Control: Provenance and Independent Re-derivation

Provenance. Protocol, decision rules and predictions: `research/pivots/method_reconstruction_complete_artifact_prereg_frozen_v0.1.yaml`, frozen before any call. The manipulation audit (`scripts/audit_complete_artifact_manipulation.py`) makes no model call and was run before freezing; it fixed the 7-world sensitivity list. Analysis: `scripts/analyze_complete_artifact.py`. Each of the 5 executions has its own API cache directory, so no execution can replay another. Every billed response was recorded in a per-call cost ledger: 960 calls, USD 0.63. Three jobs failed on upstream rate limits (HTTP 429); they were retried as infrastructure failures and the failed attempts are archived beside the raw logs.

Independent re-derivation. `scripts/verify_complete_artifact.py` imports nothing from the analysis module, re-reads the raw generation logs, re-scores every memo with the frozen scorer against the child world rebuilt from its seed, and uses the standard library's random-number generator.

1. **Labels.** The re-scored label counts equal the analysis artifact's in all six conditions.
2. **Equivalent form.** $\mathrm{DiD}_{\text{nonid}}$ computed as the mean of per-world double differences equals the difference of differences of cell means; the gap is 0.
3. **Independent recomputation.** Bootstrap means of all six contrasts agree with the analysis to 0.0028; the two-sided sign-flip test of $\mathrm{DiD}_{\text{nonid}}$ over $50{,}000$ random sign assignments gives $p < 10^{-4}$, at its floor, as the analysis does.
4. **Decisions.** The frozen decision rules, re-implemented, return identifiability supported and completeness not supported.

H.5 Replication in Qwen3.8-27B: Protocol Amendment and Provenance

Protocol. `research/pivots/method_reconstruction_multimodel_prereg_frozen_v0.1.yaml`, frozen before any call, with amendment A1 (`method_reconstruction_multimodel_amendment_a1_v0.1.yaml`). Analysis: `scripts/analyze_multimodel.py`, which applies the complete-artifact estimands and decision rules unchanged and adds the eligibility gate. The endpoint was pinned (DeepInfra, bf16, fallbacks disabled, reasoning disabled); every call was served by it, with no reasoning tokens.

Amendment A1: the output-token limit. Under the frozen limit of 1,024 tokens a first attempt truncated a large share of Qwen3.8-27B's memos; the limit had never bound for Qwen3.5-27B. The attempt was stopped, no memo was scored, and A1 raised the limit to 4,096 tokens for this model only, with its own output prefix and API cache so that no response generated under the old limit could be replayed. The aborted attempt is archived and never analysed.

Format rules. Two rules, each fixed before the replies it covers were read and applied identically to every condition, keep a reply that arrives in an unexpected JSON shape instead of re-drawing it: a single mis-named field holding a string is the memo, and a single numeric field is the model's

answer, scored by the frozen memo scorer like any memo (which never labels a bare number as the naive rule). The field name is recorded in the raw row. No memo in the analysed data needed either rule (0 recoveries). 4 no-artifact units had been re-generated after a bare-number reply before the second rule existed; scored as the rule would have scored them, the no-artifact rate is 0.119 against 0.119 observed (plain means over units), and the eligibility gate is unaffected.

Infrastructure failures and cost. Jobs that failed on the provider side or were terminated with their controlling process were re-run; every failed attempt is archived beside the raw logs, and all 160 units of the analysis are complete. The run made 1,299 billed calls, USD 2.34 in total. Because the API key was shared with other work, the account balance could not isolate the run's spend, so every call was reconciled instead against the provider's own billing record, looked up by its generation identifier (`scripts/audit_generations.py`): cost, provider and token counts agree for every call.

Second family (amendments A2 and A3). Mistral Small 3.2 24B Instruct was chosen before any call as an independent lineage of comparable scale (dense, 24B) with no reasoning mode. The endpoint rule selected its only bf16 endpoint (A2), which returned HTTP 429 "rate-limited upstream" for most calls; an amendment frozen before a fixed time (A3) then moved the whole cell to DeepInfra, whose weights are fp8, a disclosed departure from the full-precision preference. The A2 attempt is archived and was never scored. The A3 run has all 160 units, 1,286 billed calls, USD 0.13, every call reconciled against the provider's billing record except one empty upstream reply that was not billed. The frozen analysis gives no-artifact 0.656 and positive control 0.536, so the model is not eligible; an independent re-derivation reproduces every number and the gate.

Instrument check (post hoc). Table 20 reports, for the three memo-stage runs, a criterion computed without the frozen scorer: whether a memo assigns the new world's observational slope, to within 0.0015, to an effect or estimate phrase with no observed, OLS, naive or negating wording in between (`scripts/memo_numeric_check.py`). It also counts naive-rule labels given to memos that elsewhere reject the naive estimator.

Table 20: Scorer-independent numeric criterion (post hoc). Share of memos that commit numerically to the naive estimate, and the contrasts recomputed on it (world → execution bootstrap ranges). Last column: frozen naive-rule labels on memos that also reject the naive estimator.

Model	shown, wrong	shown, correct	no artifact	explicit	Δ_Y^{nonid}	$\text{DiD}_{\text{nonid}}$	contradictory
Qwen3.5-27B	0.944	0.200	—	—	−0.025	+0.769 [+0.531, +1.000]	6/313
Qwen3.8-27B	0.487	0.081	0.006	0.006	−0.038	+0.444 [+0.250, +0.656]	144/348
Mistral Small 3.2 24B	0.050	0.037	0.000	0.013	+0.000	+0.012 [−0.156, +0.188]	231/469

Numeric-commitment memo stage (amendment A4). A4 was frozen after the A3 result and before any A4 call. It appends one sentence to the P3 user message and changes nothing else: "End the memo with exactly one final line of the form 'Final estimate for the intervention effect: <number>', giving your estimate for the new problem to three decimals, or 'Final estimate for the intervention effect: not identifiable' if you conclude that no valid numeric estimate can be given from the information available." The primary outcome is the last such line: naive if its number lies within 0.0015 of the new world's observational slope; "not identifiable", the true effect, another number and a missing line all count as not naive. Worlds, conditions, estimands, bootstrap and sign-flip seeds, eligibility gate and decision rules are those of the replication protocol, applied to each model with fresh generators. Both cells ran on DeepInfra, 32 worlds × 5 executions each, after a format pilot on eight unused worlds passed its go rule (no truncated call, every job clean, the final line present in at least seven of eight memos per condition). Qwen3.5-27B: 1,280 billed calls, USD 1.17. Mistral Small 3.2 24B: 1,334 billed calls, USD 0.17; 55 calls ended in a repeated-newline loop at the 4,096-token limit, which left the reply unparseable and the job unclean, and every other unclean job was an unbilled upstream HTTP 429; unclean jobs were archived byte-for-byte and re-run until all 160 units per model were complete. Every call is reconciled against the provider's billing record. The request payloads of the original Qwen3.5-27B run equal the A4 payloads with the appended sentence removed, for all 960 of them. An independent re-derivation (own line parser, standard-library generator, re-implemented gate and rules, equivalent-form and

zero-point self-checks) reproduces every label, contrast and decision for both models (`scripts/verify_numeric_commitment.py`).

Table 21: Numeric-commitment memo stage (amendment A4, pre-registered). Share of memos whose final line gives the naive estimate, by condition, with world → execution bootstrap intervals and the frozen decisions.

Model	shown, wrong	shown, correct	no artifact	explicit	Δ_Y^{shown}	Δ_Y^{nonid}	effect	identifiability
Qwen3.5-27B	0.131	0.000	0.000	0.893	$+0.131$ $[+0.031, +0.250]$	$+0.000$	replicates	supported
Mistral Small 3.2 24B	0.056	0.013	0.086	0.730	$+0.044$ $[-0.031, +0.125]$	-0.019	does not replicate	not supported

H.6 Carrier-Reconstruction Fidelity Micro-Assay

This appendix details the carrier-fidelity micro-assay reported in Section D.6. The assay is measurement-only: it reuses the sampled-probe Qwen/E1 carrier prompt, model (Qwen/Qwen3.5-27B), P3 transfer formulation, and temperature (0.0).

Worlds and executions. The analysis uses the eight sampled-probe Qwen/E1 worlds (seeds `20261100`–`20261107`), each labeled by its observed pattern. Each world contributes $K = 8$ AO_correct endpoint executions and $K = 2$ AO_wrong control executions. Cache isolation prevents replay from prior carrier generations.

Fidelity checks. Deterministic world reconstruction (`random.Random(seed)` plus two `sample_world` calls and the ancestor builder) reproduced the logged ancestor texts byte-identically in all eight worlds. Historical memo labels also reproduced at execution: AO_wrong naive $8/8$, NO_ARTIFACT naive $0/8$, clean-world AO_correct naive $0/6$, and overlap-world AO_correct naive $2/2$.

Scoring. A deterministic scorer assigns mutually exclusive labels for naive-rule use, value transfer without a rule, rejection or abstention, and unclassifiable output, with a primary binary naive-rule indicator. Criteria include a value-echo detector for the collapse archetype ("Final estimate for the intervention effect: {new-world observed slope}" with no methodological rule), naive-advocacy patterns with negation-span and descriptive-equation filtering, and rejection-marker detection. There is no LLM judge anywhere in the assay.

Estimands. We report the per-world naive-draw fraction p_w, the pooled fraction p_{pool}, memo diversity (distinct content hashes), historical-draw consistency, and the AO_wrong control naive rate. This measurement localizes the overlap observed in the main assay; it is not a memo intervention.

Table 22: Carrier-reconstruction fidelity. Qwen3.5-27B, P3, temperature 0.0; $K = 8$ AO_correct carrier executions per world, scored by a deterministic text classifier. Both overlap worlds collapse to one memo ($p_w = 1.0$); clean worlds stay at $p_w \leq 0.125$ with diverse abstention or value-only memos; the no-induction world is bimodal. AO_wrong controls are naive in $16/16$ executions; pooled $p_{\text{pool}} = 21/64 = 0.328$.

Seed	Pattern	p_w	naive/8	unique memos	majority label
20261100	floor-contaminated	1.000	8/8	1	naive rule
20261101	clean	0.000	0/8	7	rejection/abstention
20261102	clean	0.000	0/8	8	rejection/abstention
20261103	AO-correct contaminated	1.000	8/8	1	naive rule
20261104	clean	0.000	0/8	4	rejection/abstention
20261105	no induction	0.500	4/8	8	rejection/abstention
20261106	clean	0.125	1/8	5	value-only
20261107	control degraded	0.000	0/8	4	value-only

H.7 2×5 Factorial World-Level Estimates

Table 24 reports the twelve world-level paired estimates. Each row is one world within a fixed model–operator stratum; β_{number} is the average number-present minus number-absent contrast across the five cue levels, and Δ_{cue} is the same-scale cue-vs-none contrast averaged over number presence/absence.

Table 23: 2×5 **carrier factorial.** Number presence $\times$ method cue over 12 model–operator–world cells in four fixed strata; each cell is the mean wrong-signature rate over eight discriminative probes. $\beta_{\text{number}} = +0.075$ with stratified resampling range $[+0.027, +0.127]$ is the equal-weight average marginal effect of numeric presence across the five cue strata — §D.11 explains why this average should not be read as an effect size, since it pools a $+0.367$ cell with three cells at or below zero across an interaction that excludes zero. The no-cue with-minus-without contrast is $+0.010$. Cue-versus-none contrasts are weak hint $+0.531$, explicit rule $+0.604$, partial formula $+0.297$, estimator name $+0.135$; these are percentile resampling ranges, and *no* simultaneous interval was computed for them. Ranges quantify between-world variation conditional on the fixed eight-probe bank, at three worlds per stratum.

cue	with number	without number	Δ number	cue main	number×cue
none (answer only)	0.219	0.208	+0.010	−0.314	−0.065
estimator name	0.448	0.250	+0.198	−0.178	+0.123
weak hint	0.750	0.740	+0.010	+0.218	−0.065
partial formula	0.635	0.385	+0.250	−0.017	+0.175
explicit rule	0.771	0.865	−0.094	+0.291	−0.169

$\beta_{\text{number}} = +0.075$; cue main-effect range $= 0.604$; stratified-bootstrap 95% CI $[+0.027, +0.127]$.

Table 24: World-level paired estimates for the 2×5 factorial.

Operator	Family	Seed	β_{number}	Δ_{name}	Δ_{hint}	Δ_{formula}	Δ_{rule}
E1	DeepSeek-V3.2	20260913	−0.100	+0.000	+0.000	+0.000	−0.250
E1	DeepSeek-V3.2	20260914	+0.075	+0.000	+0.000	−0.250	−0.062
E1	DeepSeek-V3.2	20260915	−0.050	−0.125	+0.000	+0.000	+0.250
E1	Qwen3.5-27B	20260910	+0.325	+0.562	+1.000	+0.625	+1.000
E1	Qwen3.5-27B	20260911	+0.350	+0.625	+1.000	+0.500	+1.000
E1	Qwen3.5-27B	20260912	+0.425	+0.500	+0.688	+0.562	+0.438
MA	DeepSeek-V3.2	20260913	−0.200	+0.000	+0.250	−0.250	+0.500
MA	DeepSeek-V3.2	20260914	−0.150	+0.000	+0.188	−0.188	+0.688
MA	DeepSeek-V3.2	20260915	+0.150	+0.062	+0.250	−0.250	+0.688
MA	Qwen3.5-27B	20260910	+0.000	+0.000	+1.000	+1.000	+1.000
MA	Qwen3.5-27B	20260911	+0.000	+0.000	+1.000	+1.000	+1.000
MA	Qwen3.5-27B	20260912	+0.075	+0.000	+1.000	+0.812	+1.000

I Power and Precision of the Decision Rule

To assess what the null can support, we run a full-gate Monte Carlo simulation of the actual decision rule. A configuration clears the carrier-attribution criterion only if both $\Delta_{\text{floor}} = q(\text{AO_wrong}) - \max\{q(\text{SHELL}), q(\text{NO_ARTIFACT}), q(\text{RANDOM}), q(\text{AGREEMENT})\} > \delta$ and $\Delta_Y = q(\text{AO_wrong}) - q(\text{AO_correct}) > \delta$. For each control-floor regime $p_{\text{floor}} \in \{0.125, 0.25, 0.5\}$, within-world correlation $\rho \in \{0, 0.5, 1\}$, and probe count $n \in \{8, 16, 24, 32, 48, 64\}$, we generate the six condition outcomes jointly, clip each rate to $[0, 1]$, draw binomial counts, and record the fraction of 20,000 worlds clearing the gate at each threshold $\delta \in \{0, 0.25\}$ (Figure 10). A Layer-3 cross-seed replication simulation—a configuration replicates when at least $r = \lceil S/2 \rceil$ of its S worlds each clear the gate, with between-world heterogeneity $\tau = 0.10$, $\rho = 0.5$, $\sigma = 0.20$—reaches the same conclusions; its full $S \times n$ boundary-pass grid and the 27-cell $\tau \times \sigma \times \rho$ sensitivity sweep are reported in the repository rather than here.

We distinguish *primary-effect power*, the probability of clearing the gate at $\delta = 0$, from *operational-margin power*, the probability of clearing the gate at $\delta = 0.25$ when the true effect sits above the margin. The case $\delta_{\text{true}} = \delta = 0.25$ is a boundary pass probability, not a power. Under the simulated $\delta_{\text{true}} = 0$ regime, the zero-effect pass probability at $\delta = 0.25$ is $\approx 1\%$ at the current 8-probe design, and the Layer-3 zero-effect replication rate is $\leq 0.2\%$ at $S = 3$. Boundary pass probabilities are low (≈ 0.16 at $n = 64$ on the clean floor and ≈ 0.14 on the high floor; 0.073 at the current $S = 3, n = 8$ design) because a marginal effect must clear a max-over-four control floor inflated by selection. Primary-effect power reaches ≈ 0.50 at $n = 8$ and ≈ 0.80 at $n = 64$ for a true $\delta_{\text{true}} = 0.25$ effect on the clean floor, whereas operational-margin power at $n = 8$ requires a true effect approaching 0.75. Under the simulated regimes, failure to cross-seed replicate is compatible with a true effect at the 0.25 decision boundary. These quantities are computed under simulated data-

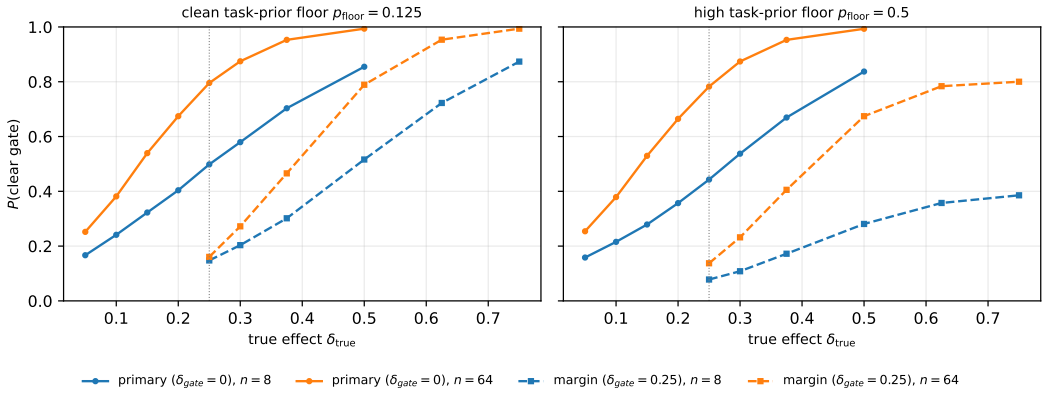


Figure 10: Full-IUT power curves. Probability of clearing the carrier-attribution IUT as a function of the true answer-specific effect δ_{true}, for zero-margin power ($\delta_{\text{gate}} = 0$, solid) and operational-margin power ($\delta_{\text{gate}} = 0.25$, dashed), at $n = 8$ and $n = 64$ probes and task-prior floors $p_{\text{floor}} \in \{0.125, 0.5\}$.

generating regimes, not estimated from observed data; reproducible simulation code and outputs are provided in the released repository manifest.

In plain terms: with the current design the rule almost never passes when there is no true effect (zero-effect pass probability $\approx 1\%$ at $n = 8$), yet a true effect sitting exactly at the 0.25 margin also passes only rarely (≈ 0.07 at $S{=}3, n = 8$). Expanding the probe bank to 64 roughly doubles primary-effect power at the margin ($0.50 \rightarrow 0.80$ on the clean floor) while the margin-boundary pass probability stays below 0.17, because a marginal effect must clear a max-over-four control floor inflated by selection. True effects well above the margin are detected reliably at 64 probes (0.79 at $\delta_{\text{true}} = 0.5$ on the clean floor).

J Sampled-Probe Analyses

J.1 Initial (exploratory) sampled-probe analysis

This analysis uses eight E1/Qwen3.5-27B/P3 worlds, 32 sampled discriminative probes, and 16 sampled agreement controls per world (Section D.1). The reported quantities use 4,000 world-then-probe bootstrap draws (seed `20261012`) and condition on one realized carrier memo per world–condition.

Table 25: Initial sampled-probe analysis on E1/Qwen3.5-27B/P3. Rates are wrong-signature fractions over 32 sampled discriminative probes, except the agreement-control flag, which uses paired agreement-control probes. All no-answer and non-diagnostic answer floors (`NO_ARTIFACT`, `ARTIFACT_SHELL`, `RANDOM_ANSWER`, `AGREEMENT_ANSWER`) are 0.0 in all eight worlds, and `EXPLICIT_WRONG_METHOD` is 1.0 in all eight worlds.

Seed	Criterion	AO_wrong	AO_correct	EXPLICIT	Δ_{floor}	Δ_Y	Attribution note
20261020	yes	0.875	0.000	1.000	0.875	0.875	none
20261021	no	0.156	0.000	1.000	0.156	0.156	AO_wrong < 0.5
20261022	yes	0.688	0.000	1.000	0.688	0.688	none
20261023	no	0.531	0.531	1.000	0.531	0.000	AO_correct overlap
20261024	no	0.844	0.000	1.000	0.844	0.844	agreement-control flag
20261025	no	0.469	0.000	1.000	0.469	0.469	AO_wrong < 0.5
20261026	no	1.000	0.562	1.000	1.000	0.438	AO_correct overlap
20261027	no	1.000	1.000	1.000	1.000	0.000	AO_correct overlap

Table 26 reports the five carrier contrasts. The four no-wrong-answer floor contrasts coincide because all four matched floors are zero; their Bonferroni simultaneous 95% lower bound is 0.441. The answer-specificity contrast Δ_Y is 0.431 (95% interval $[0.199, 0.664]$; simultaneous lower bound

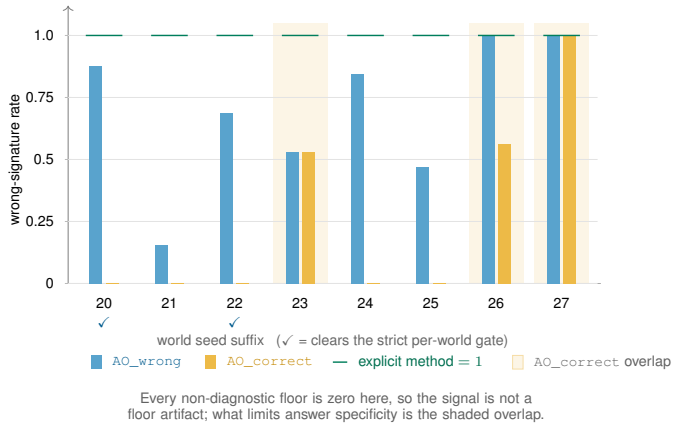


Figure 11: Sampled-probe world-level outcomes (exploratory arm). Bars give wrong-signature rates for `AO_wrong` and `AO_correct` in each world; the explicit-method positive control is at ceiling in all eight. Shaded columns are the worlds in which the *correct* archived answer also induces the wrong estimator — the failure mode that §D.6 localises to a memo collapse, and the reason Δ_Y is the binding contrast everywhere in this paper. This is the initial exploratory arm; the held-out analysis of §D.1 uses a different world draw and is reported separately.

0.152). Because the five contrasts share the `AO_wrong` numerator, we report both simultaneous and unadjusted one-sided lower bounds.

Table 26: Sampled-probe carrier contrasts. E1/Qwen3.5-27B/P3, eight worlds, hierarchical world-then-probe bootstrap, 4,000 draws. Intervals are conditional on the realized carrier memos; simultaneous bounds apply a Bonferroni correction over the five contrasts.

Contrast	Mean	95% interval	One-sided LB	Simultaneous LB
Δ_{shell}	0.694	$[0.484, 0.875]$	0.523	0.441
Δ_{noart}	0.694	$[0.484, 0.875]$	0.523	0.441
Δ_{random}	0.694	$[0.484, 0.875]$	0.523	0.441
$\Delta_{\text{agreement}}$	0.694	$[0.484, 0.875]$	0.523	0.441
Δ_Y	0.431	$[0.199, 0.664]$	0.234	0.152
$q(\text{AO_wrong})$	0.694	$[0.484, 0.875]$	0.523	0.441
$q(\text{AO_correct})$	0.263	$[0.051, 0.523]$	0.063	0.000

Replicate-share joint summaries. In the bootstrap distribution, 100% of replicates have all five contrasts above zero, 100% have $\Delta_{\text{floor}} > \delta_{\min}$, and 93.1% have all five contrasts above $\delta_{\min} = 0.25$. These are replicate fractions summarizing joint support under the bootstrap; they are not probabilities that the true population contrasts exceed the threshold.

World-level prevalence. Table 27 reports Wilson 95% intervals over eight worlds: the strict per-world criterion is met in $2/8$, the five-contrast criterion in $6/8$ at $\delta = 0$ and $5/8$ at $\delta = 0.25$, and AO-correct overlap ($q(\text{AO_correct}) > 0.25$) appears in $3/8$ worlds. The carrier-draw-conditional AO-correct mean is 0.263 (95% interval $[0.051, 0.523]$).

Table 27: World-level sampled-probe outcomes. Eight worlds with Wilson 95% intervals.

Criterion	Worlds passing	Share	Wilson 95%
Strict per-world criterion	$2/8$	0.25	$[0.071, 0.591]$
Five-contrast criterion, $\delta = 0$	$6/8$	0.75	$[0.409, 0.929]$
Five-contrast criterion, $\delta = 0.25$	$5/8$	0.625	$[0.306, 0.863]$
AO-correct contamination ($q > 0.25$)	$3/8$	0.375	$[0.137, 0.694]$

Induction-threshold sensitivity. Sweeping the AO-wrong threshold over $\{0.25, 0.375, 0.5, 0.625, 0.75\}$ yields 3, 3, 2, 2, and 1 qualifying worlds; the reported threshold (0.5) gives $2/8$.

Probe moderator and gap-distribution analysis. The sampled probes differ observably from the fixed bank in the oracle's confounding gap $|\beta_{\text{obs}} - \tau|$: sampled probes span $[0.800, 2.824]$ (median 1.455, mean 1.524, no probe below 0.5), whereas the fixed bank's absolute gaps are $\{0.5, 0.5, 1.0, 1.0, 1.5, 1.5, 2.0, 2.0\}$. Pooling the 256 `AO_wrong` probes into confounding-gap tertiles (cuts 1.212 and 1.636) gives wrong-signature rates 0.678/0.667/0.741; the mean within-world correlation between the wrong-signature indicator and $|\beta_{\text{obs}} - \tau|$ is 0.004 (range -0.302 to 0.178 over the six worlds with non-degenerate variation), and all matched floors stay at zero in every tertile. Within this carrier-draw-conditional sampled-probe estimate the answer-only signal is therefore flat across confounding strength rather than concentrated in a narrow probe regime—but the fixed bank's low-gap region ($|\beta_{\text{obs}} - \tau| = 0.5$) is not covered by any sampled probe, so this moderator analysis cannot adjudicate between the power and probe-distribution explanations. The three-arm bridge was inconclusive under its attribution table (Section D.1, Appendix K.3).

J.2 Held-Out Sampled-Probe Analysis

Table 28 reports the sampled-probe Qwen3.5-27B/E1/P3 cell summarized in Section D.1. Rates are wrong-signature fractions over 32 sampled discriminative probes. The cell supports an answer-conditioned induction claim, but the carrier-attribution IUT does not establish answer-specific reconstruction at the operational margin.

Table 28: Sampled-probe Qwen3.5-27B/E1/P3 analysis. NA = `NO_ARTIFACT`, shell = `ARTIFACT_SHELL`, rand = `RANDOM_ANSWER`, agr = `AGREEMENT_ANSWER`, AOW = `AO_wrong`, AOC = `AO_correct`, and exp = `EXPLICIT_WRONG_METHOD`. Agreement validity is the AOW paired agreement-control score. The final column reports the stricter per-world validity-and-specificity gate using displayed columns.

Seed	NA	shell	rand	agr	AOW	AOC	exp	Agree	Δ_{floor}	Δ_Y	Verdict / failed gate
20261100	0.000	0.000	0.750	0.000	0.594	0.594	1.000	16/16	-0.156	0.000	no: random/AOC
20261101	0.000	0.000	0.000	0.000	0.688	0.000	1.000	16/16	0.688	0.688	yes
20261102	0.000	0.000	0.000	0.000	0.562	0.000	1.000	16/16	0.562	0.562	yes
20261103	0.000	0.000	0.000	0.000	0.594	0.594	1.000	16/16	0.594	0.000	no: AOC overlap
20261104	0.000	0.000	0.000	0.000	0.688	0.000	1.000	16/16	0.688	0.688	yes
20261105	0.000	0.000	0.000	0.000	0.094	0.000	1.000	16/16	0.094	0.094	no: AOW < 0.5
20261106	0.000	0.000	0.000	0.000	0.625	0.000	1.000	16/16	0.625	0.625	yes
20261107	0.000	0.000	0.000	0.000	0.719	0.000	1.000	15/16	0.719	0.719	no: agree 15/16

Pooled contrasts	Mean	One-sided LB	Simultaneous LB	Verdict
$\Delta_{\text{noart}}, \Delta_{\text{shell}}, \Delta_{\text{agreement}}$	0.572	0.441	0.379	clears
Δ_{random}	0.479	0.289	0.191	below margin
Δ_Y	0.422	0.230	0.152	below margin

Table 29: Margin sensitivity for the sampled-probe IUT. The decision is determined by the smallest of the five simultaneous lower bounds, 0.152.

δ	0	.05	.10	.15	.20	.25	.30
IUT decision	pass	pass	pass	pass	fail	fail	fail

The zero-margin IUT passes, whereas the 0.25 operational-margin IUT does not. The matched-random follow-up separately yields a simultaneous lower bound of 0.395 against the attribute-matched decoy; it strengthens the numeric-plausibility control but leaves the five-contrast Δ_Y result unchanged.

K Repeated Endpoint Executions and Small-World Sensitivity

K.1 Repeated endpoint-execution robustness

The repeated endpoint-execution analysis uses the same eight Qwen3.5-27B/E1/P3 worlds and six conditions: `NO_ARTIFACT`, `RANDOM_ANSWER`, `AGREEMENT_ANSWER`, `AO_wrong`, `AO_correct`, and `EXPLICIT_WRONG_METHOD`. Each world–condition has ten endpoint executions, and each execution supplies 32 discriminative and 16 agreement probes, yielding 480 world–execution–condition units. The 4,000-draw bootstrap resamples worlds, endpoint executions independently by condition within world, and matched probe identities.

Table 30: Repeated endpoint-execution rates and contrasts. The fixed Qwen3.5-27B/E1/P3 worlds define the analysis. The simultaneous family contains four carrier contrasts and the explicit-method positive-control check. The experiment does not include artifact shell or attribute-matched random.

Quantity	Mean	95% interval	One-sided LB	Simultaneous LB
$q(\text{NOART})$	0.013	$[0.000, 0.050]$	0.000	0.000
$q(\text{RANDOM})$	0.024	$[0.000, 0.073]$	0.000	0.000
$q(\text{AGREEMENT})$	0.000	$[0.000, 0.000]$	0.000	0.000
$q(\text{AO_wrong})$	0.568	$[0.423, 0.683]$	0.447	0.399
$q(\text{AO_correct})$	0.189	$[0.025, 0.380]$	0.040	0.009
$q(\text{EXPLICIT})$	1.000	$[1.000, 1.000]$	1.000	1.000
Δ_{noart}	0.555	$[0.414, 0.671]$	0.439	0.390
Δ_{random}	0.544	$[0.410, 0.655]$	0.431	0.390
$\Delta_{\text{agreement}}$	0.568	$[0.423, 0.683]$	0.447	0.399
Δ_Y	0.379	$[0.171, 0.590]$	0.200	0.132

All four core contrasts clear zero, and the explicit-method positive control is at ceiling. The operational conjunction is limited by Δ_Y. Agreement-probe accuracy is $7{,}665/7{,}680 = 0.998$; the 15 misses are retained and concentrated in one AO-wrong world. Leave-one-world-out estimates preserve the zero-margin result under every omission. These repetitions characterize execution variation for the eight worlds and four core controls.

K.2 Small-World Sensitivity

The repeated endpoint-execution analysis has many lower-level observations but only eight fixed worlds. We therefore complement the hierarchical bootstrap with a small-sample sensitivity analysis over the eight world-level paired contrasts. Table 31 reports those contrasts after averaging endpoint executions and matched probes within each world.

Table 31: World-level paired contrasts. Values are $q(\text{AO_wrong}) - q(c)$ after averaging over the ten endpoint executions and 32 discriminative probes.

World	Δ_{noart}	Δ_{random}	$\Delta_{\text{agreement}}$	Δ_Y
20261100	.600	.550	.700	-.009
20261101	.700	.700	.700	.700
20261102	.600	.581	.600	.600
20261103	.506	.506	.506	-.025
20261104	.716	.716	.716	.644
20261105	.147	.147	.147	.147
20261106	.459	.459	.459	.259
20261107	.716	.694	.716	.716

For each boundary $\delta \in \{0, 0.25\}$, the exact sensitivity test centers each world's contrast at δ and enumerates all $2^8 = 256$ sign flips. Its sharp-null assumption is that the eight centered paired contrasts are independent and sign-exchangeable. This assumption is explicit and does not turn the deliberately constructed worlds into a random sample; the resulting p-values diagnose sensitivity conditional on these worlds rather than support inference to a model or world population. For the conjunction, the exact IUT p-value is the maximum of the four one-sided component p-values.

Table 32: Exact world-level sign-flip tests. The operational row tests each paired contrast against $\delta = 0.25$.

Boundary	p_{noart}	p_{random}	$p_{\text{agreement}}$	p_Y	p_{IUT}	**Decision**
$\delta = 0$	.0039	.0039	.0039	.0156	.0156	pass
$\delta = 0.25$	.0078	.0078	.0078	.1367	.1367	fail

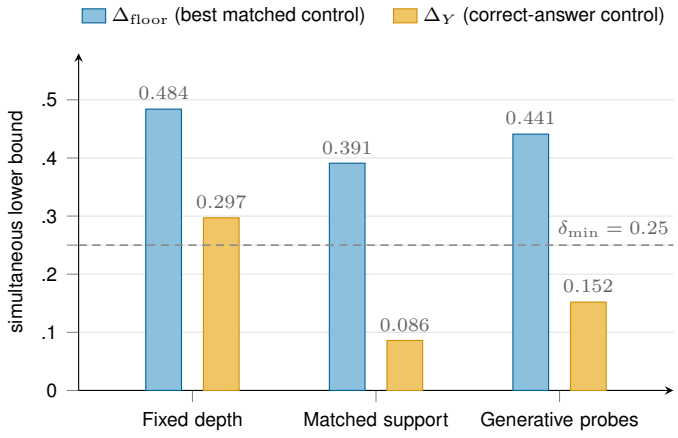
Leave-one-world-out hierarchical bootstraps agree qualitatively for all eight omissions: the core-four zero-margin conjunction passes and the operational conjunction fails. Across omissions, the mean Δ_Y ranges from 0.331 to 0.437 and its simultaneous lower bound from 0.073 to 0.187; the other three contrast means range from 0.520 to 0.628, with simultaneous lower bounds from 0.341 to 0.529. Thus the exact small-world diagnostic and the hierarchical bootstrap support the same conditional conclusion. Neither analysis compensates for the absence of additional independent worlds.

K.3 Three-Arm Bridge Diagnostic

The three-arm bridge tests whether the fixed-bank and sampled-probe difference is better explained by statistical depth, probe-distribution support, or world composition. The fixed-depth arm clears both margins, whereas sampled arms retain strong floor separation but remain limited by answer specificity (Table 33; Figure 12). Because the arms use disjoint worlds and support matching is incomplete, the bridge does not identify the source of the protocol difference.

Table 33: Three-arm bridge (Qwen3.5-27B/E1/P3). Entries are Bonferroni-simultaneous one-sided 95% lower bounds; an arm succeeds when both exceed $\delta_{\min} = 0.25$. Arm B is a partial distribution match because the generative stratum has zero mass at gap 0.5, and the arms were scored on disjoint world samples.

Arm	Probe regime	Worlds	Δ_{floor} LB	Δ_Y LB	Verdict
A	fixed bank, expanded depth	8	0.484	0.297	passes both bars
B	partial gap-matched sample	8	0.391	0.086	fails Δ_Y
C	generative D_{probe} (v0.2)	8	0.441	0.152	fails Δ_Y
C-repl	generative D_{probe}	4	—	0.078	fails Δ_Y



All three arms clear the floor margin; only the fixed-depth arm clears the answer-specificity margin. The arms use *disjoint* worlds and the matched-support arm does not cover the full fixed-bank gap support, so the comparison is diagnostic, not decisive.

Figure 12: The bridge compares probe depth and distribution support without closing it. Δ_{floor} is the margin against the strongest no-answer control and Δ_Y the margin against the correct answer; both are Bonferroni-simultaneous one-sided lower bounds. The fixed-depth arm clears both against $\delta_{\min} = 0.25$, whereas the two sampled arms retain strong floor separation and fail on answer specificity — the same contrast that binds in §D.1. Because the arms are not paired on the same worlds and the matched arm does not span the full fixed-bank support, this figure narrows the candidate explanations without deciding between them; the gap-matched comparison that would decide it is infeasible with this world generator (Remark 3).

K.4 Sampled-Probe Matched-Random Control: Matching Audit and Outcomes

The matched-random control is an *additional* condition executed on the eight sampled-probe Qwen/E1 worlds. It does not replace `RANDOM_ANSWER` in the five-contrast family of Eqs. 10–14; the two conditions have different measured rates and are reported as separate contrasts (Remark 2). Earlier text describing the sampled-probe Δ_{random} as attribute-matched was incorrect and is corrected here. The matcher selects a value from other parent worlds that differs from the target's β_{obs} and τ while matching sign, magnitude, decimal precision, artifact syntax, and distance-to-oracle as closely as the candidate pool permits.

Table 34 reports the matching audit and per-world outcomes. The matched-random condition remains at zero in the realized carrier draws, leaving a positive AO_wrong increment: $\mu_{\text{matched}} = 0.572$ (95% interval $[0.414, 0.691]$; simultaneous lower bound 0.395). This argues against simple numeric implausibility as a sufficient explanation in this cell; the estimate remains carrier-draw-conditional, and the limiting attribution contrast is still wrong versus correct answer.

Table 34: Matched-random matching audit and outcomes. S/M/P records sign, magnitude-bin, and decimal-precision matches. "Gap" is $||c - \tau| - |\beta_{\text{obs}} - \tau||$.

seed	β_{obs}	matched	tier	S/M/P	source	gap	q: AO / matched
20261100	0.224	0.212	A	1/1/1	20261030	0.012	0.594 / 0.000
20261101	−3.1	−2.7	A	1/1/1	20261044	0.400	0.688 / 0.000
20261102	0.1	0.2	A	1/1/1	20261081	0.100	0.562 / 0.000
20261103	−3	−2.951	B	1/1/0	20260823	0.049	0.594 / 0.000
20261104	0.316	0.333	A	1/1/1	20260827	0.017	0.688 / 0.000
20261105	3.5	2.7	A	1/1/1	20260924	0.800	0.094 / 0.000
20261106	−3.5	−2.7	A	1/1/1	20261044	0.800	0.625 / 0.000
20261107	2.591	2.423	A	1/1/1	20261071	0.168	0.719 / 0.000

Table 35: Contrast decomposition of the wrong-signature rate. *Top:* seven condition contrasts (Condition as a categorical treatment; reference level NO_HISTORY). *Bottom:* condition×pressure slope contrasts. Coefficients are rate differences from ordinary least squares over 648 observations; SE are CR2 cluster-robust over 18 configurations with Satterthwaite df; CIs are unadjusted wild-cluster-bootstrap 95% intervals (4,000 Rademacher draws).

Contrast	$\hat{\beta}$	CR2 SE	df	95% CI
Condition contrasts (main effects)				
Reuse framing (NO_ARTIFACT − NO_HISTORY)	+0.016	0.036	8.8	[−0.052, +0.083]
Artifact existence (SHELL − NO_ARTIFACT)	−0.062	0.034	7.0	[−0.125, +0.003]
Numeric content (RANDOM − SHELL)	+0.009	0.025	6.5	[−0.037, +0.056]
Agreement answer (AGREEMENT − SHELL)	+0.002	0.030	6.8	[−0.052, +0.059]
Wrong answer (AO_wrong − SHELL)	+0.031	0.034	5.8	[−0.031, +0.094]
Specificity Δ_Y (AO_wrong − AO_correct)	+0.026	0.037	6.4	[−0.046, +0.097]
Explicit method (EXPLICIT − SHELL)	+0.128	0.052	8.8	[+0.034, +0.222]
Condition × pressure slope contrasts				
Explicit × pressure (EXPLICIT − SHELL slope)	+0.121	0.046	7.8	[+0.035, +0.209]
Answer × pressure (AO_wrong − SHELL slope)	+0.010	0.024	4.2	[−0.034, +0.055]
Specificity × pressure (AO_wrong − AO_correct slope)	+0.027	0.020	6.4	[−0.011, +0.065]
Artifact × pressure (SHELL − NO_ARTIFACT slope)	+0.021	0.030	3.8	[−0.035, +0.077]

L Pooled Decomposition and Full Condition-Level Tables

L.1 Contrast decomposition: design and full design matrix

Table 35 reports pooled contrasts from a categorical model of the nine core content conditions, with NO_HISTORY as the reference. The seven contrasts are linear combinations of condition means; Table 36 gives the complete design matrix. Ordinary least squares uses all 648 condition–configuration–pressure observations, with CR2 cluster-robust standard errors over 18 model–operator–world configurations and wild-cluster-bootstrap 95% intervals (4,000 Rademacher draws). Leave-one-block-out ΔR^2 attributes the model's $R^2 = 0.28$ to model/operator (0.21), pressure (0.04), explicit-method conditions (0.03), and answer-related conditions (0.003).

L.2 Full Condition-Level Tables

Section D reports the carrier-attribution estimands and the contrast decomposition. Here we report the full condition-level data behind those estimands, so the answer-only null is auditable against every matched control rather than asserted from summary estimands alone.

The reuse-context effect $\Delta_R = q(\text{NO_ARTIFACT}) - q(\text{NO_HISTORY})$ at P3 is zero in five of six third-family configurations and +0.500 for MA on seed 20260906; it is neither a uniform positive reuse-framing effect nor systematically negative, in contrast to the MA–Qwen3.5-27B factorial cell where it was a stable +0.375. On DeepSeek-V3.2 the two no-answer baselines are mostly indistinguishable at high pressure, which is consistent with a pressure-driven floor rather than a reuse-framing effect.

Table 36: Full condition×contrast design matrix. Rows are the seven contrast decompositions; columns are the nine core content conditions. NH = NO_HISTORY, NA = NO_ARTIFACT, blank = AO_blank, SH = ARTIFACT_SHELL, RND = RANDOM_ANSWER, AGR = AGREEMENT_ANSWER, AOW = AO_wrong_answer_only, AOC = AO_correct_answer_only, EXP = EXPLICIT_WRONG_METHOD.

Contrast	NH	NA	blank	SH	RND	AGR	AOW	AOC	EXP
Reuse framing (NA − NH)	−1	+1	0	0	0	0	0	0	0
Artifact existence (SH − NA)	0	−1	0	+1	0	0	0	0	0
Numeric content (RND − SH)	0	0	0	−1	+1	0	0	0	0
Agreement answer (AGR − SH)	0	0	0	−1	0	+1	0	0	0
Wrong answer (AOW − SH)	0	0	0	−1	0	0	+1	0	0
Specificity Δ_Y (AOW − AOC)	0	0	0	0	0	0	+1	−1	0
Explicit method (EXP − SH)	0	0	0	−1	0	0	0	0	+1

Table 37: High-pressure artifact contrasts by configuration. Δ_{shell} = AO_wrong − shell; Δ_{floor} = AO_wrong − max{shell, no-artifact, random, agreement}; Δ_Y = AO_wrong − AO_correct; $\Delta_{A\times P}$ contrasts P3 with P0. Rates use eight discriminative probes.

Op.	Cell	Seed	AO_wrong	Shell	No-art.	Δ_{shell}	Δ_{floor}	Δ_Y	$\Delta_{A\times P}$
E1	GLM-4-9B	20260903	0.375	0.625	0.625	−0.250	−0.250	0.000	0.250
E1	GLM-4-9B	20260904	0.000	0.375	0.000	−0.375	−0.625	−0.250	−0.750
E1	GLM-4-9B	20260905	0.375	0.375	0.875	0.000	−0.500	0.375	−0.250
E1	Qwen3.5-27B	20260903	0.500	0.000	0.000	0.500	**0.250**	**0.375**	0.500
E1	Qwen3.5-27B	20260904	0.000	0.375	0.000	−0.375	−0.375	0.000	−0.375
E1	Qwen3.5-27B	20260905	0.000	0.000	0.125	0.000	−0.125	−0.375	0.000
MA	GLM-4-9B	20260903	0.375	0.375	0.000	0.000	0.000	0.000	−0.375
MA	GLM-4-9B	20260904	0.250	0.250	0.125	0.000	−0.125	0.000	0.000
MA	GLM-4-9B	20260905	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MA	Qwen3.5-27B	20260903	0.000	0.000	0.375	0.000	−0.375	0.000	0.000
MA	Qwen3.5-27B	20260904	0.000	0.000	0.375	0.000	−0.375	0.000	0.375
MA	Qwen3.5-27B	20260905	0.000	0.000	0.375	0.000	−0.375	0.000	0.000

Table 38: Content-condition rates at high pressure (P3). Each rate is the fraction of eight discriminative probes eliciting the wrong-signature method.

Op.	Cell	Seed	No-hist	No-art	Shell	RAND	AGREE	AO_corr	AO_wrong	Expl.
E1	GLM-4-9B	20260903	0.625	0.625	0.625	0.500	0.500	0.375	0.375	0.250
E1	GLM-4-9B	20260904	0.500	0.000	0.375	0.250	0.625	0.250	0.000	0.000
E1	GLM-4-9B	20260905	0.250	0.875	0.375	0.375	0.000	0.000	0.375	0.250
E1	Qwen3.5-27B	20260903	0.125	0.000	0.000	0.000	0.250	0.125	0.500	1.000
E1	Qwen3.5-27B	20260904	0.000	0.000	0.375	0.000	0.000	0.000	0.000	0.375
E1	Qwen3.5-27B	20260905	0.000	0.125	0.000	0.000	0.000	0.375	0.000	0.375
MA	GLM-4-9B	20260903	0.250	0.000	0.375	0.000	0.250	0.375	0.375	0.375
MA	GLM-4-9B	20260904	0.250	0.125	0.250	0.375	0.250	0.250	0.250	0.375
MA	GLM-4-9B	20260905	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MA	Qwen3.5-27B	20260903	0.000	0.375	0.000	0.000	0.000	0.000	0.000	1.000
MA	Qwen3.5-27B	20260904	0.000	0.375	0.000	0.000	0.000	0.000	0.000	1.000
MA	Qwen3.5-27B	20260905	0.000	0.375	0.000	0.000	0.000	0.000	0.000	1.000

Table 39: Wrong-signature rate by transfer-prompt formulation (P0–P3). Rates are averaged over three worlds per model–operator configuration.

Op.	Cell	Condition	P0	P1	P2	P3
E1	GLM-4-9B	Shell	0.208	0.292	0.625	0.458
		AO_wrong	0.250	0.250	0.333	0.250
		AO_correct	0.250	0.333	0.417	0.208
		Explicit	0.042	0.458	0.375	0.167
E1	Qwen3.5-27B	Shell	0.000	0.208	0.125	0.125
		AO_wrong	0.000	0.333	0.333	0.167
		AO_correct	0.000	0.167	0.083	0.167
		Explicit	0.000	0.125	0.250	0.583
MA	GLM-4-9B	Shell	0.083	0.083	0.042	0.208
		AO_wrong	0.208	0.125	0.083	0.208
		AO_correct	0.250	0.208	0.208	0.208
		Explicit	0.083	0.083	0.292	0.250
MA	Qwen3.5-27B	Shell	0.125	0.000	0.000	0.000
		AO_wrong	0.000	0.000	0.375	0.000
		AO_correct	0.000	0.125	0.292	0.000
		Explicit	0.042	0.000	0.667	1.000

Table 40: DeepSeek-V3.2 high-pressure contrasts. Δ_{shell}, Δ_{floor}, and Δ_Y are paired carrier contrasts; Expl.@P3 is the explicit-wrong-method positive control.

Op.	Seed	AO_wrong	Shell	No-art.	Δ_{shell}	Δ_{floor}	Δ_Y	Expl.@P3
E1	20260906	0.500	0.500	0.500	0.000	0.000	0.000	0.125
E1	20260907	0.500	0.000	0.500	0.500	0.000	0.375	0.500
E1	20260908	0.500	0.500	0.500	0.000	0.000	0.500	0.500
MA	20260906	0.125	0.125	0.500	0.000	-0.375	0.000	0.500
MA	20260907	0.500	0.500	0.500	0.000	0.000	0.000	0.875
MA	20260908	0.500	0.125	0.125	0.375	0.375	0.000	1.000

Table 41: DeepSeek-V3.2 content-condition rates at high pressure (P3). Each rate is the fraction of eight discriminative probes eliciting the wrong-signature method.

Op.	Seed	No-hist	No-art	Shell	RAND	AGREE	AO_corr	AO_wrong	Expl.
E1	20260906	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.125
E1	20260907	0.500	0.500	0.000	0.500	0.500	0.125	0.500	0.500
E1	20260908	0.500	0.500	0.500	0.500	0.500	0.000	0.500	0.500
MA	20260906	0.000	0.500	0.125	0.125	0.500	0.125	0.125	0.500
MA	20260907	0.500	0.500	0.500	0.125	0.125	0.500	0.500	0.875
MA	20260908	0.125	0.125	0.125	0.000	0.125	0.500	0.500	1.000

Table 42: DeepSeek-V3.2 response across transfer-prompt formulations (P0–P3). The table compares shell, answer-only, and explicit-method conditions.

Op.	Seed	Condition	P0	P1	P2	P3
		Shell	0.500	0.500	0.500	0.500
E1	20260906	AO_wrong	0.500	0.500	0.500	0.500
		Explicit	0.500	0.500	0.500	0.125
		Shell	0.500	0.500	0.500	0.000
E1	20260907	AO_wrong	0.500	0.625	1.000	0.500
		Explicit	0.875	0.000	0.500	0.500
		Shell	0.500	0.500	0.500	0.500
E1	20260908	AO_wrong	0.500	0.500	0.500	0.500
		Explicit	0.500	0.000	0.500	0.500
		Shell	0.000	0.500	0.125	0.125
MA	20260906	AO_wrong	0.125	0.000	0.125	0.125
		Explicit	0.125	0.500	0.500	0.500
		Shell	0.125	0.125	0.125	0.500
MA	20260907	AO_wrong	0.000	0.500	0.125	0.500
		Explicit	0.500	0.500	1.000	0.875
		Shell	0.500	0.500	0.125	0.125
MA	20260908	AO_wrong	0.500	0.500	0.000	0.500
		Explicit	0.125	0.500	1.000	1.000

L.3 Fixed-Bank Attribution Details

Explicit method transfer. Explicit estimator rules produce the strongest recurrence in responsive configurations. Their mean wrong-signature rate is 0.50 in the factorial experiment and 0.583 in DeepSeek-V3.2; Qwen3.5-27B reaches 0.79 across operators and ceiling in all eight sampled-probe E1/P3 worlds. This positive control establishes that the assays can register semantic method uptake.

Fixed-bank carrier contrasts. The pooled fixed-bank contrasts are $\Delta_{\text{shell}} = +0.03$ (one-sided LB -0.02), $\Delta_{\text{noart}} = -0.03$ (LB -0.11), $\Delta_{\text{random}} = +0.02$ (LB -0.04), $\Delta_{\text{agreement}} = +0.03$ (LB -0.04), and $\Delta_Y = +0.03$ (LB -0.04). Their proximity to zero contrasts with the explicit-rule response and motivates the sampled-probe analysis. The fixed bank has 0.125 resolution per probe; Appendix I quantifies its sensitivity across effect sizes.

Sampled-probe carrier contrasts. In Qwen3.5-27B/E1/P3, the single-execution estimate is $\Delta_{\text{floor}} = 0.694$ with interval $[0.484, 0.875]$. All five simultaneous lower bounds are positive (0.152–0.379) and the five contrast means span 0.422–0.572; the smallest bound is the wrong-versus-correct contrast at 0.152. Repeated endpoint executions give a corresponding four-control Δ_Y lower bound of 0.132. The correct-answer condition therefore determines the transition from positive zero-margin attribution to the stricter 0.25 operational criterion.

M Supporting Analyses

This appendix collects three exploratory measurements that calibrate the assay rather than test the claim: the method-information ladder and the initial number×cue factorial, the response to transfer-prompt pressure, and a semi-realistic variant of E1. None of them enters a decision rule.

M.1 Method-cue ladder, initial factorial, and prompt pressure

Table 43 reports the graded method-information ladder (mean wrong-signature rate per rung, three worlds per configuration, eight discriminative probes per cell), and Table 44 the initial number×cue factorial that preceded the registered 2×5 design of §D.7. Figure 13 shows the fixed-bank response across the four transfer-prompt levels P0–P3 (§C.4): in Qwen3.5-27B the explicit-rule control rises with pressure while the answer-only condition does not, and in DeepSeek-V3.2 both sit at the task-prior floor, so prompt level and floor cannot be separated in that family. The ladder is a ruler for the assay, not a finding (§D.10).

Table 43: Method-information ladder. Mean wrong-signature rate per rung, three worlds per configuration and eight discriminative probes per cell.

Operator	Family	answer	name	hint	formula	explicit
E1	Qwen3.5-27B	0.17	0.96	0.79	1.00	0.58
E1	DeepSeek-V3.2	0.50	0.50	0.83	0.50	0.38
MA	Qwen3.5-27B	0.00	0.00	0.92	1.00	1.00
MA	DeepSeek-V3.2	0.38	0.21	0.83	0.54	0.79

Table 44: Initial 2×5 factorial. Number presence × method cue, pooled over twelve model–operator–world cells. Each cell is the mean wrong-signature rate over eight discriminative probes.

cue	with number	without number	Δ number	cue main	number×cue
none (answer only)	0.26	0.18	+0.08	−0.30	−0.13
estimator name	0.42	0.19	+0.23	−0.22	+0.02
weak hint	0.84	0.64	+0.21	+0.22	+0.00
partial formula	0.76	0.40	+0.37	+0.06	+0.16
explicit rule	0.84	0.69	+0.16	+0.25	−0.05

$\beta_{\text{number}} = +0.21$; cue main-effect range $= 0.55$.

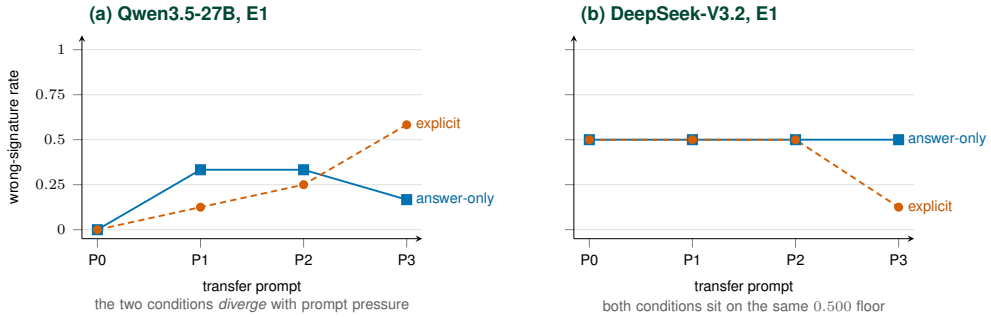


Figure 13: Prompt pressure separates the two families, and only one separation is informative. Rates are fractions of eight discriminative probes at each transfer-prompt level; the explicit control is dashed and drawn on top where the two coincide. In Qwen3.5-27B the explicit-method control climbs monotonically with pressure while the answer-only condition does not follow it, which is the behaviour a responsive assay should show. In DeepSeek-V3.2 both conditions sit at the 0.500 task-prior floor for P0–P2 and only the explicit control moves at P3, so prompt level and floor are not separable in that family — the compression of contrasts recorded in Fig. 9. Neither panel is a sampled-probe measurement, so neither answers whether the effect of §D.1 survives below P3 (§E).

M.2 Semi-realistic E1 scenarios

Six domain-grounded E1 scenarios, three per family, replace the abstract variables X, Y and U with named exposures, outcomes and confounders while keeping the linear-Gaussian structure, so that ground truth remains exact. Each scenario runs five matched conditions and is scored against a frozen scenario probe bank under the fixed-bank probe (Table 45). In Qwen3.5-27B the explicit rule transfers at ceiling while the answer-only artifacts stay at the no-artifact and shell level; DeepSeek-V3.2 stays at or below its task-prior floor in every condition, the explicit rule included, so it gives no reading there. The variant tests surface realism only; it does not test the positive cell, which uses the decoding-rule probe and the memo stage.

Table 45: Semi-realistic E1 scenarios (*exploratory*). Mean wrong-signature rate over three domain-grounded scenarios per family, eight discriminative probes per condition, deterministic ground-truth scoring with the frozen scenario probe bank.

Family	NO_ARTIFACT	SHELL	AO_wrong	AO_correct	EXPLICIT
Qwen3.5-27B	0.000	0.125	0.125	0.042	1.000
DeepSeek-V3.2	0.500	0.500	0.500	0.417	0.333