STATESWAP: SEPARATING PROPAGATION FROM DECISION EFFECTS IN PERSISTENT-STATE LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluations of persistent-state agents can conflate three questions: whether an intervention was executed, whether its consequences propagated through retrieval to rendered evidence, and whether the downstream decision responded. StateSwap verifies intervention execution separately and reports, on paired latent worlds, a *state-intervention profile* $\Pi_S = (p_A, p_R, p_D, \tau)$: allocation disagreement, rendered-evidence disagreement, observed decision disagreement, and the paired correctness contrast. Its *Intervention–Propagation–Decision Separation* principle states that rendered-evidence disagreement requires allocation disagreement, whereas complete upstream disagreement does not identify a decision-side effect.

We instantiate this protocol with real, externally served readers. Within each pair, the task, plans, repository realisation and paired gold target are fixed; only the stored allocation policy differs, inducing a $5/3$ versus $3/5$ split of a fixed retrieval budget, and the reader is masked from state, arm and gold. In a pre-specified held-out block of 2000 paired synthetic worlds, allocation and rendered evidence changed in every pair while $\tau = +0.0100$ (95% interval -0.003 to +0.023), compatible with a small positive, near-zero or slightly negative effect.

An annotation-assisted calibration on 1000 HotpotQA distractor-validation questions under two reader families shows why this distinction matters: two one-paragraph replacements with identical upstream change rates have sharply different decision-side consequences. The mean contrast for replacing an unannotated paragraph relative to baseline met a pre-specified ± 5 pp equivalence criterion under both readers, whereas replacing an annotated supporting paragraph reduced accuracy by 34.5 pp (DeepSeek) and 32.2 pp (Qwen).

A separate fresh-sample validation on 3200 new, R9-disjoint synthetic worlds tested an imposed stored policy. For the pinned Qwen primary, the reliability-dependent decision-side interaction met the prespecified conjunctive sign-reversal criterion; the prespecified DeepSeek secondary did not reach the reversal. The design was fixed after the exploratory R9 outcome was known, so it is not a preregistration or an independent replication, and S35's held-out identity is unchanged. StateSwap therefore treats propagation and decision response as distinct estimands; it neither measures model-internal state nor establishes a general decision benefit from persistent state.

1 INTRODUCTION

A state change can alter what an agent retrieves without altering what it decides, and two calls with identical inputs can disagree. Evaluations should therefore distinguish intervention execution, changes in retrieved and rendered evidence, and downstream decisions. An outcome difference alone does not locate these changes.

StateSwap models the chain $H \to S \to A \to R \to D$: interaction history, application-level persistent state, retrieval allocation, rendered evidence, and the served reader's decision. We intervene at S, not on model internals, and do not identify $H \to S$. Paired worlds share their task, repository and

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.

gold target; only the stored allocation policy changes, from $3/5$ to $5/3$ of a fixed retrieval budget. The reader is masked from state, treatment labels and gold.

Intervention execution is verified separately as a design and provenance fact. The *state-intervention profile* $\Pi_S = (p_A, p_R, p_D, \tau)$ reports allocation disagreement, rendered-evidence disagreement, decision disagreement and the paired correctness contrast (Section 2). Its central implication is that complete upstream propagation does not identify a decision-side effect (Section 3). This is a diagnostic principle, not a claim of formal mediation or an internal state of the reader.

Evidence and contributions. The pre-specified held-out block of 2000 paired worlds gave $\tau = +0.0100$ (95% interval -0.003 to +0.023), despite changes in selected and rendered evidence in every pair. Identical-input controls disagreed in 2.5% of pairs, motivating a mean contrast rather than deterministic attribution of each changed answer. Our contributions are:

- A paired measurement protocol with fixed latent truth, recorded allocation and exact rendered bytes, a masked reader, and a four-coordinate profile; simple formal statements delimit what these measurements imply.
- Controlled tests, a held-out block, path controls and external stress tests that distinguish propagation from decision response while preserving each block's evidence status.
- Annotation-assisted calibration on 1000 fixed questions across two reader families: equal upstream change rates accompany sharply different decision responses under two paragraph-replacement conditions, assessed with prospectively specified contrasts and an equivalence criterion.

The calibration tests a measurement configuration, not state learning. Neither it nor the held-out result establishes that a persistent state improves decisions.

Beyond the held-out block we also report, in Section 5.2, a prospectively frozen *fresh-sample validation* of a persisted application-level allocation policy. Its design was fixed after the corresponding exploratory outcome had been observed, so it is not a preregistration and not an independent replication; on 3200 new worlds, the pinned Qwen primary showed a reliability-dependent decision-side interaction meeting the prespecified conjunctive sign-reversal criterion, while the prespecified DeepSeek secondary did not reach the reversal, and S35's held-out identity is unchanged.

2 Framework and the state-intervention profile

We represent the system as a chain of five objects (Figure 3): H the upstream interaction history; S an application-level persistent state — an allocation policy stored between episodes, not a model-internal hidden state and not claimed to correspond to any internal representation of the reader; A the retrieval allocation the state induces; R the evidence actually rendered; and D the decision of a real, externally served language model. The experiments intervene at S and measure $S \to A \to R \to D$.

Definition 1 (state-intervention profile). Index paired units by w and arms by $a \in \{0, 1\}$ (a=1 the active state $s+$, a=0 the reference state $s-$), and write $A_w(a)$, $R_w(a)$, $D_w(a)$ and $Y_w(a) = \mathbf{1}[D_w(a) = d^*_W]$ for the selected record set, the exact rendered bytes, the served decision and correctness against the paired gold decision/target d^*_W, fixed within the pair. With $p_A = \mathbb{P}(A_1 \neq A_0)$, $p_R = \mathbb{P}(R_1 \neq R_0)$, $p_D = \mathbb{P}(D_1 \neq D_0)$ and $\tau = \mathbb{E}[Y_1 - Y_0]$, the *state-intervention profile* is $\Pi_S = (p_A, p_R, p_D, \tau)$: p_A the *allocation-response / allocation-disagreement* coordinate, p_R the *rendered-evidence propagation / disagreement* coordinate, p_D *observed paired decision disagreement*, and τ the *paired correctness contrast*. Π_S is a vector of four distinct estimands reported side by side on the same paired units; it is *not* a mediation decomposition and *not* a joint identified causal object. State-intervention *execution* is a separate verified design and provenance fact — the harness records that S was applied in both arms — and it is *not* p_A: p_A measures the downstream allocation *response*, so a successfully executed intervention can still leave $p_A = 0$.

Intervention–Propagation–Decision Separation (IPDS). *Allocation response* (p_A), *rendered-evidence propagation* (p_R) and *decision response* (p_D, τ) must be measured separately. Because the rendering map is frozen and deterministic, $R = \mathcal{R}(A)$ and therefore $\{R_1 \neq R_0\} \subseteq \{A_1 \neq A_0\}$,

so $p_R \leq p_A$. Second, and this is the separation result of the paper, $p_A > 0$ and $p_R > 0$ do *not* imply $p_D > 0$ and do *not* imply $\tau \neq 0$; conversely $p_D > 0$ does *not* imply $\tau \neq 0$. Separation means that *conflation is unlicensed*, not that the layers are mutually independent.

Estimand, claim boundary and reader contract. D is observed from the served reader; repeated-call controls show a reader and service variability component (2.5% of pairs under byte-identical reader input), so we report a paired mean contrast and never treat an individual disagreement as deterministically produced by S. The paired gold target d^*_W is fixed within the pair, so the outcome does not move with the evidence; an earlier qualification branch used a different, reader-visible oracle, which Appendix B keeps separate. We analyse τ as the mean paired difference with a paired-world bootstrap whose resampling unit is the paired world (Appendix A); the reader contract, parser and deterministic decoding configuration are documented there, and the direct decision task replaces an earlier universal-completion protocol that failed at qualification (Appendix B).

3 Formal separation results

Full proofs, the assumption table, the counterexample construction and the empirical profile table are in Appendix F. The three statements below are deliberately simple: Proposition 1 is an identity of the harness rather than a behavioural finding, and Proposition 2 is a logical non-implication witnessed by an explicit counterexample reader.

Proposition 1 (pipeline attribution for A and R). Assume (S1) the paired latent object is fixed — d^*_W, task, plans, constraints and the pre-state repository realisation identical across the arms; (S2) S is the sole manipulated application-level variable; (S3) treatment-independent allocator randomness ξ_w is fixed and shared within the pair; and (S4) the allocator and renderer are frozen and deterministic, with no arm-conditional editing, injection, reordering or dropping. Then $A_w(a) = \mathcal{A}(w, S(a), \xi_w)$ and $R_w(a) = \mathcal{R}(A_w(a))$, so $\{A_1 \neq A_0\}$ is definitionally the event that changing S changes the allocator output under the frozen pipeline and $\{R_1 \neq R_0\} \subseteq \{A_1 \neq A_0\}$. This is an identity about the harness and must not be cited as evidence of a decision effect; reader masking is a decision-side isolation property and is not an assumption of this attribution.

Proposition 2 (propagation non-implication). Within the system class, $p_A > 0$ and $p_R > 0$ do not imply $p_D > 0$ or $\tau \neq 0$, and $p_D > 0$ does not imply $\tau \neq 0$: **upstream propagation does not identify a decision-side effect**. An evidence-blind reader $g_\varnothing(R, u, \zeta) = h(u)$ gives $D_w(1) = D_w(0)$, hence $p_D = \tau = 0$, while $p_A = p_R = 1$ by construction; conversely an evidence-using reader $g_\star$ can be constructed with $\tau = +1$ at the same upstream profile. Two systems may therefore agree on the upstream coordinates and need not agree on the decision coordinates.

Proposition 3(a) (clamped path: null in expectation). Assume (C1) the reader-visible evidence is byte-identical in the pair, $R_w(1) = R_w(0)$; (C2) the same reader contract applies to both arms; and (C3a) the marginal serving mechanism is arm-invariant. Then the two arm-specific decisions share the same conditional law, so with a fixed pair gold $\mathbb{E}[Y_1] = \mathbb{E}[Y_0]$ and $\tau = 0$. Finite paired decision disagreement may still occur: under C1–C3a it is compatible with serving variability and is not evidence of an arm-visible treatment response, and we do not claim it is the unique cause. Appendix F adds a collapsed-path lemma (Proposition 3(b)) under the strictly stronger condition that the two served draws are conditionally independent, marked explicitly as a theoretical condition that is not empirically verified.

4 Paired-world design

A world fixes six latent fields (three durations, three capacities, integers in 1..9). Two candidate plans each draw on a subset of the three work items and carry a fixed overhead, a deadline and a minimum item capacity; exactly one plan is feasible in every generated world and is the gold d^*_W (external blocks retain their own externally defined golds). The five domains are semantic renderings of one latent task distribution. Each world also fixes a 16-record repository of two reporters with accuracy 0.60 and 0.50 whose source labels are neutralised before rendering, and the population is

stratified over 40 strata (five domains × two depths × two display orders × two gold plans). The full construction, allocator, renderer and a record-level worked example are in Appendix A.

S is an application-level allocation policy — a stored pair of per-source parameters — and the registered allocator is a pure function of the world and the stored state: it converts the two parameters into a quota split of a fixed budget of eight record slots, clipping each quota to $[2, 6]$, so $s+$ induces $5/3$ and $s-$ induces $3/5$. Which records fill the quota follows a uniform permutation of the six field names drawn from a treatment-independent stream, so both arms share it exactly, and any overlap between sources is recorded rather than suppressed. No evidence is edited, injected or reordered by hand. We record A changed and R changed for every pair, and in all reported blocks both links changed in essentially every pair (Table F2).

The state, arm, allocation, reliabilities, world identifier, record identifiers and gold plan are withheld from the reader. The exploratory block (S34) was extended after its first worlds had been observed, so its estimate is outcome-informed. The held-out block (S35) draws 2000 fresh paired worlds from a public randomness source committed to before its output existed, verifies the beacon, derives the population seed through the frozen key-derivation path, and fixes the sample size, estimator, bootstrap resamples (10000, analysis seed 20261001) and analysis order in advance (Appendix A). The mechanism controls (S36) and the cross-model block (S37) are exploratory by construction. We did not tune the reader, prompt, parser, allocation, strata or analysis on the held-out result, remove unfavourable worlds or blocks, or pool exploratory and held-out blocks.

5 Experiments

This section reports the main evidence blocks. Each block states its evidence class once, is never pooled with another, and is never upgraded by narrative prominence. Table 1 lists the classes; the remaining blocks are reported in the appendices as indicated.

Table 1: Evidence classes used in this paper. Only the pre-specified held-out block (S35) is confirmatory. No block is pooled with another, and no class is upgraded by narrative prominence.

Block	Evidence class
S35	pre-specified *held-out* (only confirmatory)
S34, S37, S38, S39R1, R7, R9	*exploratory*
R5, R6	prospectively specified *calibration*
R10	*fresh-sample validation* (not preregistration, not replication)
S40-R3C	*secondary*

Table 2: Primary state-intervention results. τ is the mean paired difference in decision correctness between the $s+$ (5/3) and $s-$ (3/5) arms; intervals are paired bootstrap intervals over worlds.

Block	N pairs	Calls	Acc $s+$	Acc $s-$	τ	95% CI	Δ $(+1/0/-1)$	D chg.
S34 first 400	400	800	0.9125	0.8850	+0.0275	[+0.000, +0.058]	23/365/12	35
S34 combined	1000	2000	0.9180	0.8990	+0.0190	[+0.003, +0.035]	44/931/25	69
S35 held-out	2000	4000	0.9025	0.8925	+0.0100	[-0.003, +0.023]	97/1826/77	174

The S34 block is outcome-informed exploratory evidence and is never pooled with the S35 held-out estimate. In every block the allocation and the rendered evidence differ in all pairs, so the intervention provably changes the reader input.

5.1 Pre-specified held-out paired block (S35)

Evidence class: pre-specified held-out; this is the only confirmatory block in the paper.

The held-out block draws 2000 new paired worlds over the same 40 strata from a fresh, publicly committed randomness source and repeats the identical intervention, reader, prompt, parser and analysis. All 4000 calls completed with no technical or parse failures, and the intervention again changed A and R in all 2000 and 2000 pairs. Arm accuracies were 0.9025 ($s+$) and 0.8925 ($s-$).

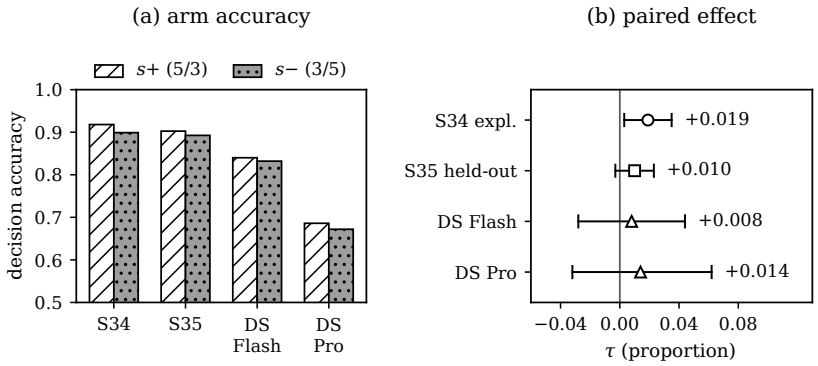

Figure 1: Primary results. (a) decision accuracy in the $s+$ (5/3) and $s-$ (3/5) arms; (b) paired effect τ with paired bootstrap intervals. The held-out and cross-model intervals include zero; the exploratory interval does not.

The decision changed in 174 pairs (8.7%), with 97 favouring $s+$, 77 favouring $s-$ and 1826 ties. The estimate is $\tau = +0.0100$ (+1.00 pp, 95% paired interval -0.003 to +0.023). This is a positive point estimate whose uncertainty includes zero, compatible with a small positive, near-zero or slightly negative effect. We do not describe it as a confirmed or proven improvement. The exact paired (McNemar) statistic on the 97 versus 77 discordant pairs is descriptive ($p = 0.1495$) and not the primary summary.

Per-stratum estimates are reported in Table A2. The stratum values are small and heterogeneous, with both positive and negative values, and no stratum was removed from any analysis.

5.2 Fresh-sample validation of the persisted allocation policy (R10)

Evidence class: fresh-sample validation, fixed after the exploratory R9 outcome was seen. It is not a preregistration and not an independent replication, and it does not change the identity of S35.

Before any R10 reader call we froze the material, schedule, code, reader contract, run policy and analysis seed. No pilot, interim effect check, optional extension, model substitution or response-conditioned stopping was allowed. The block draws 3200 new worlds over the same 40 strata, with the same task, environment, repository, state, allocator and rendering algorithms and the same policies, prompts, parser and reader parameters as R9. Worlds are excluded by canonical latent identity against R9 and every earlier comparison set; the recorded overlap is exactly zero. 13200 calls per reader were scheduled, with Qwen (`qwen-plus-2025-12-01`) primary and DeepSeek (`deepseek-flash`) prespecified secondary.

Table 3: Fresh-sample validation of the persisted allocation policy (R10). 3200 new latent worlds, 13200 calls per reader. $d1 = Y(\mathrm{E1}, S1) - Y(\mathrm{E1}, S2)$, $d2 = Y(\mathrm{E2}, S1) - Y(\mathrm{E2}, S2)$ and $\mathrm{mean}(z) = \frac{1}{2}(d1 - d2)$; all values are percentage points with 95% percentile bootstrap intervals over base worlds within the 40 fixed strata. The designated primary is Qwen; its unscaled interaction $2\,\mathrm{mean}(z)$ is +5.75 pp. "Reversal" means the $d1$ interval lies entirely above zero *and* the $d2$ interval entirely below zero.

Reader	Contrast	$d1$ (pp)	$d2$ (pp)	$\mathrm{mean}(z)$ (pp)
Qwen (primary)	point [95%]	+3.25 [+1.22, +5.28]	-2.50 [-4.56, -0.47]	+2.88 [+1.70, +4.08]
DeepSeek (secondary)	point [95%]	+2.19 [-0.06, +4.41]	+0.16 [-2.09, +2.37]	+1.02 [+0.02, +2.02]

Qwen meets the primary criterion and the conjunctive reversal criterion; DeepSeek's two component intervals both include zero, so the secondary leg does *not* reach the reversal. Its $\mathrm{mean}(z)$ lower endpoint is $+0.02$ pp (exact $+0.015625$ pp); the descriptive reader-minus-reader difference in $\mathrm{mean}(z)$ is +1.86 pp [+0.42, +3.31]. The two readers are fixed serving setups on the same materials, not independent samples, and the secondary is never a fallback primary.

With the primary Qwen reader $d1 = +3.25$ pp $[+1.22, +5.28]$, $d2 = -2.50$ pp $[-4.56, -0.47]$ and $\mathrm{mean}(z) = +2.88$ pp $[+1.70, +4.08]$. Both the primary criterion and the conjunctive reversal criterion are met. The prespecified DeepSeek secondary does not reach the reversal: $+1.02$ pp $[+0.02, +2.02]$, and both of its component intervals include zero. It is never promoted, and no rule, threshold or seed changed after outcomes.

Disclosures travel with these numbers (Appendix L). Qwen has 2 strict parse failures, counted zero under the full-N endpoint, and no world was dropped. The identical-input control's structural gate passes; identical inputs yield the same canonical answer in 89/87 (Qwen) and 89/84 (DeepSeek) of 100 pairs per environment; these rates are never subtracted from any estimate. Unlike R9, the full raw response bodies and outgoing wire-byte hashes were archived for every attempt. The persisted policy is an imposed application-level object. The bounded claim is that a reliability-dependent decision-side interaction recurs for the pinned Qwen reader on new worlds; it establishes no learned appraisal, no model-internal state and no broad state benefit.

5.3 Other blocks (appendices)

The exploratory paired block S34 (Appendix A), the external-validity stress tests S38 and S39R1 (Appendix E), the exploratory evidence-swap probe R7 (Appendix J), the exploratory synthetic policy test R9 (Appendix K) and the secondary S40-R3C block (Appendix H) are reported in the appendices. None is pooled with the held-out block, and none changes its identity.

5.4 Mechanism controls

Evidence class: exploratory controls on the mechanism path.

The controls modify one link of the reader-visible path at a time. They do not identify a formal mediation effect, and the paired-arm intervals for C1, C2, C3 and C5 include zero. The two negative controls instantiate Proposition 3(a). Under C1 the allocation is forced to $4/4$ in both arms, so the reader is shown identical inputs by construction, yet the decision still differs in 5 of 200 pairs (2.5%), the identical-input disagreement baseline. Under C2 the upstream allocation still differs by arm in 200 of 200 pairs while the reader-visible evidence changes in 0, and decision changes occur at 5 of 200 pairs. The direction perturbations are reported in Appendix C: with equal average source accuracy (C3) arm-dependent differences persist, and reversing which source is better (C5) reverses the sign of the point estimate without excluding zero.

5.5 External calibration across two reader families

Evidence class: prospectively specified calibration; not confirmatory and not an independent replication.

This calibration asks whether replacing an annotated supporting paragraph and an unannotated paragraph produces different decision responses under two reader families. It directly manipulates paragraph selection under a frozen rule, with no intervention on S or evaluation of a learned allocation policy. Upstream change rates are construction properties, not evidence of state learning.

Design. The population is 1000 HotpotQA distractor-validation questions. Each world fixes one question and one pool under three four-paragraph reader-visible conditions: B keeps both officially annotated supporting paragraphs and two distractors; I replaces one paragraph with no supporting-fact annotation; K replaces one annotated supporting paragraph. Exactly one of four slots is exchanged and lengths are matched by the frozen reference-length rule, so the interventions are equal-edit in that restricted sense, not in semantic edit distance or served token counts. The reader is not given the arm, the annotations, the allocation or the gold; paragraph indices and source titles are visible by design. Official annotations participate in compiling the pool, making this an *annotation-assisted* calibration with an external gold (provenance in Appendix G), and both interventions change the allocation and the rendered evidence in every pair by construction ($p_A = p_R = 1$ is a construction property, not an estimate).

Two readers are served under one frozen contract each, with identical messages: a DeepSeek reader (3400 calls, no truncation) and a Qwen reader (3400 HTTP responses, 3393 usable; 7 cap hits —

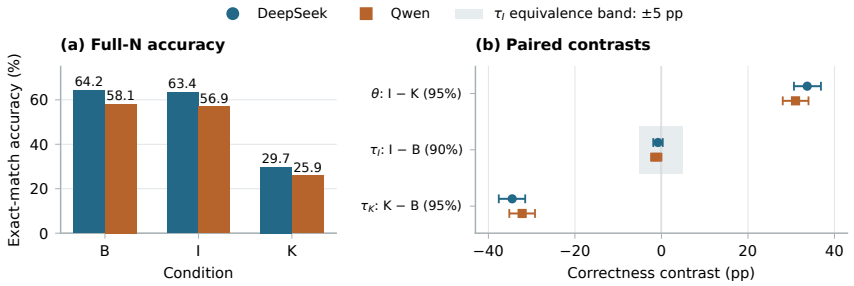

Figure 2: External calibration on 1000 shared questions. (a) Full-N exact-match accuracy: B retains four paragraphs, I replaces one unannotated paragraph and K replaces one annotated supporting paragraph. Truncations count as incorrect; no per-arm intervals were pre-specified. (b) Paired contrasts: τ_I uses 90% intervals and its shaded ± 5 pp equivalence band; θ and τ_K use 95% intervals.

B 1, I 2, K 3, closed-book 1). Cap-hit and invalid cells count as incorrect under the frozen full-N operational endpoint; no item is dropped and no cap was raised.

Results. Both readers show the same pattern. Mean exact-match accuracy over the full N was 0.642/0.634/0.297 ($B/I/K$) for the DeepSeek reader and 0.581/0.569/0.259 for the Qwen reader (full table in Appendix G; Figure 2). The prospectively specified primary contrast of each block is $\theta = \mathrm{EM}_I - \mathrm{EM}_K$: $+33.70$ pp (95% cluster interval $[+30.63, +36.87]$) for DeepSeek and $+31.00$ pp ($[+28.04, +34.00]$) for Qwen. Removing an *annotated* supporting paragraph is therefore accompanied by a much larger accuracy drop than replacing an unannotated one, and the same pattern appears under both readers, which are not independent samples.

The secondary I-versus-B contrast is -0.80 pp (90% interval $[-1.89, +0.30]$) and -1.20 pp ($[-2.40, +0.00]$). Both intervals lie inside the pre-specified ± 5 pp equivalence band, so both readers *met the prespecified equivalence criterion* for the *mean* contrast; this is not a statement that individual answers are unchanged, since canonical-answer disagreement between I and B was still 0.092 for DeepSeek ($n = 1000$) and 0.136 for Qwen over its usable pairs ($n = 997$). The accuracy drops for K versus B were -34.50 pp for DeepSeek and -32.20 pp for Qwen. The two-reader difference in θ is -2.70 pp (95% interval $[-5.79, +0.40]$); because the interval includes zero we report no detected difference and explicitly do *not* read it as model equivalence. The practical point is that two interventions with identical upstream change rates produced clearly different decision-side responses, which is why the profile is reported jointly.

Abstention is reported separately, with its own status. Under K both readers abstain more often. For the Qwen reader, whose three-way decomposition was specified before its outcomes existed, 451 of 1000 answers became the prompt-allowed UNKNOWN in K versus 36 in B while 0.259 remained correct; the DeepSeek decomposition was post-hoc and is descriptive only. We do not merge the two statuses, and neither UNKNOWN nor the residual non-exact-match class is described as hallucination (Figure 7, Appendix G).

Evidence status and disclosures. The DeepSeek and Qwen blocks are prospectively specified calibrations (**not confirmatory**, **not independent replications**); both reuse one external task population and one annotation-assisted pool-compilation rule, both retain the four flagged items, and neither replaces the pre-specified held-out block of Section 5.1, which keeps its original identity. For the Qwen block the frozen analyser passed an inherited default bootstrap seed to the resampler, an implementation deviation found by an independent audit; the intervals reported here use the prespecified seed passed explicitly, the earlier values are disclosed alongside, and the joint decision is unchanged under either seed (Appendix G).

A separate exploratory probe applies the same exchange rule to 1000 questions with a single reader and shows the paired mean change can be zero while 58 questions change in opposite directions (Appendix J); it is exploratory and replaces nothing.

6 Related work

The distinction between observed differences and identified effects is classical (Rubin, 1974; Angrist et al., 1996; Imbens & Rubin, 2015; Pearl, 2009). Mediation (Imai et al., 2010; VanderWeele, 2009) and construct validity (Cronbach & Meehl, 1955; Borsboom et al., 2004) inform our claim boundaries; our profile is not a mediation estimand. Memory architectures (Packer et al., 2023; Park et al., 2023; Luo et al., 2026), context manipulation (Greshake et al., 2023; Qian, 2026; Huang et al., 2026), and internal attribution (Vig et al., 2020; Huang et al., 2023; Zhang & Nanda, 2024; Syed et al., 2023) study related but different intervention surfaces.

Memory intervention and selection. CMI (Srivastava, 2026) compares no-memory, with-memory and perturbed-memory responses for the same query and candidate, then uses answer-based utility and stability scores to select memories. It reports both answer quality and memory-selection metrics; its implementation uses annotated memory roles and a gold-dependent deterministic scorer in the selection loop (§§3.2–3.4, 5.2–5.4 of that paper). Memory intervention and intermediate selection measurements are therefore established ingredients. StateSwap instead intervenes on a stored allocation policy under fixed paired-world conditions and reports allocation, exact rendered evidence, decisions and correctness separately; it offers no competing selector.

Tracing and fault injection. Provenance and workflow observability (Alam & Wang, 2021; Souza et al., 2025a;b) provide inspectable intermediate records; llmmas-otel (Seyedghorban et al., 2026) aligns baseline and faulty runs of the same task, records message and input hashes, and repeats tasks to measure runtime amplification (§§II–III); chaos engineering for LLM multi-agent systems (Owotogbe, 2025) proposes a broader robustness programme. Our focus is the separation of evidence propagation from answer disagreement and correctness, with identical-input decision controls; we claim neither priority for tracing, pairing or repetition nor superiority over these systems, and Appendix I records the source-specific comparison and its limits.

7 Limitations and scope

Identification and serving variability. Proposition 1 is definitional conditional on the paired world and shared treatment-independent randomness; Proposition 2 is a logical counterexample, not an empirical discovery. Proposition 3(a) requires an arm-invariant marginal serving mechanism, which is not verified; Proposition 3(b) additionally requires conditional independence. Identical-input disagreement is descriptive, not a lower bound or a quantity to subtract from an effect. The controls isolate the reader-visible path, not all possible consequences of state.

Generalisation and evidence status. The S35 held-out and S37 exploratory cross-model intervals include zero; the exploratory and secondary alignment contrasts have different evidence status and are not pooled. Five synthetic domains are renderings of one latent task distribution. The operational-text blocks broaden the materials but do not establish a general state advantage; their incoming-only baselines are descriptive. S40-R3C remains secondary.

R5/R6 share one HotpotQA population and construction rule; they are prospectively specified calibrations, not confirmatory tests or independent-sample replications. Support annotations guide construction but do not establish that a replaced paragraph is necessary or the sole answer source. The 30-item human review flagged four retained source issues; it does not certify the other questions. Mean equivalence does not imply unchanged individual answers, and the cross-reader interval containing zero does not establish reader equivalence.

Scoring and implementation. Prompt-allowed abstention and residual non-exact-match answers are not identified hallucinations; the latter can include correct aliases. The Qwen abstention analysis is prospective and the DeepSeek analysis post-hoc. All 7 Qwen truncations count as incorrect under the full-N endpoint, with no item dropped or cap raised. An inherited bootstrap seed caused an implementation deviation that changed the interval computation. Reported intervals use the pre-specified seed explicitly; both versions and the unchanged joint decision are disclosed in Appendix G.

Exploratory synthetic policy test. The R9 persisted-policy test (Appendix K) is exploratory, and imposes an application-level allocation policy rather than testing learned appraisal; its primary interaction is positive but its component intervals touch or cross zero, so its sign-reversal criterion is not met, and its secondary interval includes zero. The R10 fresh-sample validation (Section 5.2; Appendix L) re-tests the same policy on a new frozen block; its design was chosen *after* the R9 outcome was seen, so it is not a preregistration, an independent replication or a new task distribution. Only the Qwen primary reaches the conjunctive reversal, the DeepSeek secondary does not, S35's identity and interval are unchanged, and identical inputs still changed canonical answers in a minority of control pairs in both blocks.

8 CONCLUSION

StateSwap reports intervention execution, propagation and decision response separately. The held-out state intervention changed allocation and evidence throughout, without establishing a correctness gain. External calibration on the same questions under two reader families shows why the distinction matters: two paragraph replacements produce identical upstream change rates but different correctness and abstention responses. The contribution is a measurement protocol with explicit controls and claim boundaries, not evidence of model-internal state or a general benefit from persistent state. In an outcome-informed exploratory synthetic test (Appendix K), a persisted application-level allocation policy gave a positive reliability-aligned interaction for the designated Qwen reader but not for the DeepSeek secondary, and its sign-reversal criterion was not met. A later fresh-sample validation on 3200 new synthetic worlds (Section 5.2) found a reliability-dependent decision-side interaction meeting the prespecified conjunctive sign-reversal criterion for the pinned Qwen primary, while the prespecified DeepSeek secondary did not reach the reversal. That design was fixed after the exploratory R9 outcome was known, is not a preregistration or an independent replication, and leaves S35's held-out identity unchanged.

REPRODUCIBILITY STATEMENT

Freeze scope is stated per block rather than uniformly. For each experiment block this paper reports the artifact that defines it; the seeds, estimator, reader contract, parser and analysis order are documented in Appendix A, with the mechanism controls in Appendix C, the formal statements and proofs in Appendix F, and the external-calibration provenance and denominators in Appendix G. The R5/R6 calibrations and the probe controls were sealed under the reader contracts recorded in Appendix G. The exploratory R7 probe is an explicit exception: its local seal covers only its original 40-question, 130-call design, the enlargement to 1,000 questions happened after the smaller probe's outcome had been observed and is therefore outside that seal, and one sentence of the frozen R7 scale report was corrected in an appended statement rather than by rewriting the report (Appendix J). Key reported statistics are emitted from archived artifacts by the generators described in Appendix A and are re-derived in the numerical audit, which compares the generated macros against the frozen source records and asserts that they agree. The anonymous supplementary archive contains the paper source and the generated numerical macros; frozen analysis objects and compact archived inputs for the packaged R5, R6, R7, R9 and R10 blocks; per-attempt flat scoring records for R9 and R10; and self-contained offline scripts, listed in the supplement `README.md`, that recompute the packaged R5/R6, R7, R9 and R10 summaries without network access, an API key or a GPU.

REFERENCES

Md Morshed Alam and Weichao Wang. A comprehensive survey on data provenance: State-of-the-art approaches and their deployments for IoT security enforcement. *Journal of Computer Security*, 29(4):423–446, 2021. doi: 10.3233/JCS-200108.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. doi: 10.1080/01621459.1996.10476902.

Denny Borsboom, Gideon J. Mellenbergh, and Jaap van Heerden. The concept of validity. *Psychological Review*, 111(4):1061–1071, 2004. doi: 10.1037/0033-295X.111.4.1061.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955. doi: 10.1037/h0040957.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2023. doi: 10.1145/3605764.3623985.

Charoes Huang, Xin Huang, Ngoc Phu Tran, and Amin Milani Fard. Model context protocol threat modeling and analyzing vulnerabilities to prompt injection with tool poisoning. *arXiv preprint arXiv:2603.22489*, 2026.

Jing Huang, Zhengxuan Wu, Kyle Mahowald, and Christopher Potts. Inducing character-level structure in subword-based language models with type-level interchange intervention training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12163–12180. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.770. URL https://aclanthology.org/2023.findings-acl.770/.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334, 2010. doi: 10.1037/a0020761.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Jinghao Luo, Yuchen Tian, Chuxue Cao, Ziyang Luo, Hongzhan Lin, Kaixin Li, Chuyi Kong, Ruichao Yang, and Jing Ma. From storage to experience: A survey on the evolution of LLM agent memory mechanisms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 41622–41652. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.2069. URL https://aclanthology.org/2026.findings-acl.2069/.

Joshua Owotogbe. Assessing and enhancing the robustness of LLM-based multi-agent systems through chaos engineering. In *2025 IEEE/ACM 4th International Conference on AI Engineering - Software Engineering for AI (CAIN)*, 2025. doi: 10.1109/CAIN66642.2025.00039.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009. doi: 10.1017/CBO9780511803161.

Jiachen Qian. SilentRetrieval: Hijacking retrieval-augmented generation via semantically-preserving adversarial data poisoning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026. doi: 10.1145/3770855.3818186.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Zahra Seyedghorban, Egor Klimov, Arie van Deursen, Annibale Panichella, and Burcu Kulahcioglu Ozkan. Observability and fault injection for LLM-based multi-agent systems in software engineering. In *2026 IEEE International Conference on Software Testing, Verification and Validation (ICST)*, 2026. doi: 10.1109/ICST69053.2026.00037. URL https://arxiv.org/abs/2608.24271.

Renan Souza, Amal Gueroudji, Stephen DeWitt, Daniel Rosendo, Tirthankar Ghosal, Robert Ross, Prasanna Balaprakash, and Rafael Ferreira da Silva. PROV-AGENT: Unified provenance for tracking AI agent interactions in agentic workflows. In *Proceedings of the 2025 IEEE 21st International Conference on e-Science*, 2025a.

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 on Supercomputing (SC)*, 2025b. doi: 10.1145/3731599.3767582.

Saksham Sahai Srivastava. Causal intervention-based memory selection for long-horizon LLM agents. *arXiv preprint arXiv:2605.17641*, 2026.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *NeurIPS 2023 Workshop on Attributing Model Behavior at Scale (ATTRIB)*, 2023. URL https://arxiv.org/abs/2310.10348.

Tyler J. VanderWeele. Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):880–883, 2009. doi: 10.1097/EDE.0b013e3181bd5638.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

AI USE STATEMENT

Generative-AI tools were used in this work for tasks that require disclosure under the ICLR AI policy: assistance with conceptual framing and study design, statistical planning, implementation of the experimental and analysis code, data preparation, interpretation of results, literature comparison, figure preparation, and drafting and editing of the manuscript. All experimental-design decisions, scientific claims and reported numbers were reviewed and approved by the authors, who take responsibility for the final content of this work. Served reader models are experimental subjects: their outputs were scored under the specified scoring procedures and were not used as manuscript prose. All reported statistics are computed from archived experimental outputs; the numerical-audit procedure is documented in Appendix A, while calibration-specific implementation deviations and corrections are documented in Appendix G. Generative-AI tools are not authors of this work, and the authors have reviewed AI-assisted material; these reviews do not replace the authors' responsibility for verification.

A REPRODUCIBILITY PACKAGE

A.1 WHAT IS FROZEN, AND HOW

Every block is defined by a frozen artifact that is hashed before execution and never rewritten afterwards. Each block has a seed commitment recorded before any world existed: the exploratory block uses a 32-byte OS-CSPRNG seed with digest `669EE9E9D7C9C76981F296516837B70A7388723602D266060710696E0C431B4C` (commitment file `7B158FAEABC6A617451E7E615BF0949B01A3487A7C9EA017FE2D10A6FD01F2A5`), while the held-out block commits in advance to a future public randomness round, verifies the beacon against a pinned root of trust, and derives the population seed from the frozen commit digest and the verified randomness through a named key-derivation path. Worlds are generated as paired arms from one repository realisation, and cross-experiment overlap is checked by latent-world key rather than identifier alone.

The allocator and renderer are the frozen production implementations; the state changes only the allocation policy, and no code path edits, injects, reorders or drops records for one arm only. The prompt text is frozen with digest `13793AF36A14A296C9D722C8AF388271E71913D72716C011D33657D678D9004E` and the reader identity and request configuration with `F445D432323B6296285EC7AC4CB48C0BE15C723C42D69FF7291663784518D068` (temperature 0, top-p 1, single sample, no streaming, minimal output budget, direct HTTP POST to the provider endpoint with no SDK layer). Where a provider's default reasoning mode consumed the output budget before any decision was emitted, the adapter that disables it was recorded, validated on a probe call and then used for every subsequent scientific call; probe calls are archived separately and excluded from every estimate. For the original S34–S37 programme, the complete logical call order is frozen before the first call, calls are checkpointed, an interruption may only pause and resume from the exact next unresolved logical call, and a finalised response is not re-issued. The external-validity extension contains one separately disclosed deterministic replay incident described in Appendix E.4. Except for the single C3 technical missing call described in Appendix C, the reported scientific calls completed without technical missingness or parse failure.

A.2 INTEGRITY SCOPE OF EACH ARCHIVED OBJECT

We keep four integrity layers separate, because they license different claims. *(1) Protocol/freeze integrity:* the freeze artifacts (reader, prompt, parser, allocator, renderer, oracle, sample size, estimator) are hashed before execution and never rewritten afterwards. *(2) Request integrity:* for every stored scientific call the request digest was re-derived from the frozen call order and the frozen population and matches the recorded value, so the archived request agrees with the frozen protocol call by call. *(3) Response-archive integrity:* for each archived completed provider response, the experiment stores the response bytes and a response digest, and the stored response digest was re-computed from the stored payload and matched; technical-missing attempts are recorded as such and carry no response payload. A request digest does not by itself prove that a provider's response bytes could not have changed, and we do not claim that; the response layer is an archive-integrity claim about the bytes we hold, not a proof about the provider's servers. *(4) Result/analysis integrity:* the analysis is derived from the archived per-world records, and the headline paired estimates are recomputed

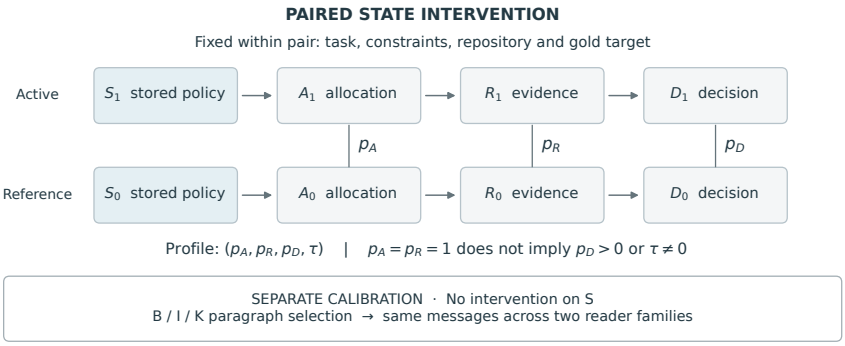

Figure 3: Paired state interventions and separate external calibration (moved here to keep the main text within the nine-page limit). In the upper module, the task, repository and gold remain fixed while S changes; the profile compares allocation, rendered evidence and decisions across arms. The lower module directly replaces paragraphs under a fixed rule, with no intervention on S (§5.5), using the same questions and messages across two readers, each with its own frozen contract.

directly from raw records and asserted to match. The single forensic repair in the cross-model block is scoped separately in Appendix B.

A.3 Protocol detail needed to read the results

What S is. S is an application-level persistent allocation policy stored between episodes: a pair of per-source parameters. It is not a model-internal hidden state. $s+$ and $s-$ are the two stored policy values used here; the frozen allocator maps them to the quota pairs $5/3$ and $3/5$ over a budget of eight record slots, with each source's quota clipped to $[2, 6]$. *How records are selected.* The allocator is a pure function of (world, domain, depth, stored state, field permutation); the permuted order of the six field names is drawn from a treatment-independent stream, so both arms of a pair share it exactly. Source 1 receives the first q_1 fields of the permutation and source 2 the last q_2; because $q_1 + q_2 = 8 > 6$, a field can be selected from both sources, and the overlap is recorded rather than suppressed. *Depth.* The registered SCD-5 rule is used: at $D2$ each plan's item subset is drawn uniformly from the three 2-element subsets of $\{1, 2, 3\}$, and at $D3$ both plans use all three items. *Worked record.* In world `S35-shipment_routing-D2-a_first-A-000002` the latent fields, the two candidate plans, the 16-record repository and the gold plan A are fixed; under $s+$ the allocator selects five fields from reporter 1 and three from reporter 2, under $s-$ three and five, and both arms render the same format with neutral source labels. *Uncertainty.* The primary interval resamples paired worlds with replacement from the pooled set of worlds (the resampling unit is the paired world; no stratification is applied), 10000 draws with analysis seed 20261001. For the alignment contrast, each of the two input blocks is resampled in the same way and independently (20000 draws, seed 20261003), and the two resampled estimates are differenced.

Execution order. For the held-out block the call order was frozen before the first model call (the freeze artifact records zero calls at freeze time) and the executed sequence matches it call by call (4000/4000, with request digests re-derived from the frozen order and the frozen material). The two arms of each pair are executed adjacently (logical indices $2k - 1$ and $2k$); which arm runs first is derived from the material stream, so 979 pairs ran $s+$ first and 1021 ran $s-$ first. Because the two calls of a pair are adjacent in wall-clock time, this design does not fully eliminate temporal or service drift between arms. We therefore report a post-hoc, no-new-call execution-order sensitivity analysis (POST_HOC_EXECUTION_ORDER_SENSITIVITY): the paired estimate is $+0.0075$, 0.0000 and $+0.0225$ across chronological thirds of execution order and ranges from -0.030 to $+0.035$ across tenths, the within-pair gap between the first response and the second request has median 0.40 s and 95th percentile 0.40 s (maximum 117 s, one interruption), the served model identity is constant (4000/4000), and no single arm occupies a long execution block. The primary estimate is unchanged; details are archived in `audit/S35_ARM_EXECUTION_ORDER_AUDIT.json`.

Table A1: Design of the experimental blocks. All blocks share the latent-world construction, the repository generator, the allocator, the renderer, the outcome definition and the reader-visible masking contract; they differ in population, sample size, source-quality regime and allocation condition.

Design element	S34 exploratory	S35 held-out	S36 controls	S37 cross-model
Role	proof of concept	held-out formal	mechanism controls	reader replication
Paired worlds	1000	2000	200 / 200 / 999 / 500 / 1000	500
Strata	40	40	40	40
Arms	$s+$ 5/3 vs $s-$ 3/5	$s+$ 5/3 vs $s-$ 3/5	varied per control	$s+$ 5/3 vs $s-$ 3/5
Source quality	0.60 / 0.50	0.60 / 0.50	0.60/0.50, equal, reversed	0.60 / 0.50
Readers	one served LLM	one served LLM	one served LLM	two served LLMs
Reader calls	2000	4000	400 / 400 / 2000 / 2000 / 2000	2000
Classification	exploratory, outcome-informed	held-out formal	exploratory controls	exploratory cross-model

Every block hides the state, the arm, the allocation, the source reliability and the gold plan from the reader. A: allocation; R: rendered evidence.

A.4 Worked record (one real world)

World `S35-shipment_routing-D2-a_first-A-000002`: domain `shipment_routing`, depth $D2$, stratum `shipment_routing|D2|reporter_a_first|goldA`, latent fields $t = (5, 2, 8)$ and $c = (6, 5, 1)$, gold $d^*_W = A$. Under $s+$ (quota $5/3$) reporter 1 contributes c_1, t_2, t_1, t_3, c_3 and reporter 2 contributes t_3, c_3, c_2; under $s-$ (quota $3/5$) reporter 1 contributes c_1, t_2, t_1 and reporter 2 contributes t_2, t_1, t_3, c_3, c_2. The rendered-evidence digests differ between arms (first 12 hex `9ea9faf8c037` versus `f5942c98e27f`), as do the reader-visible user-text digests (`c75544f82b57` versus `cc7685ca07c7`, 1169 and 1172 characters); both arms receive the same task, the same plans and neutralised source labels.

The block design is summarised in Table A1.

A.5 Readers, providers and wire configuration

The primary reader is `qwen3.8-max-0902` (Alibaba Cloud Model Studio), reported as a provider-published dated snapshot identifier in the cross-model freeze and used with the same model string in S34, S35 and S36. The cross-model readers are `deepseek-flash` and `deepseek-v4-pro` (DeepSeek Official API), whose published names are rolling aliases without an immutable snapshot identifier. All requests used temperature 0, top-p 1, one sample, streaming off, an output budget of 16 tokens, and a direct HTTP POST to the provider's chat-completions endpoint with no SDK layer; Qwen calls carried `enable_thinking=false` and the DeepSeek calls carried the provider's explicit non-thinking configuration (thinking disabled, reasoning effort none), which was validated on a probe call before scientific use. Decoding was configured deterministically (temperature 0, top-p 1, one sample, streaming off), but served inference is not perfectly repeatable, as the identical-input controls show. The frozen parser accepts exactly `{"decision":"A"}` or `{"decision":"B"}`; anything else is a non-correct outcome, and a transport failure is a technical missing excluded from paired estimation. Because the DeepSeek names are rolling aliases, the cross-model block is exploratory and is never described as snapshot-exact replication.

A.6 Freeze, commitment, execution and seal

For the held-out block the order of operations is: (1) protocol freeze; (2) commitment to a future public randomness round before its output existed; (3) beacon verification against the pinned root of trust and seed derivation through the named key-derivation path; (4) world generation and call execution in the frozen logical order with checkpointing and resume-from-next-unresolved-call; (5) result seal. Hash values identify the frozen objects; what makes the ordering meaningful is the commitment plus the future beacon and the recorded execution order, not a hash timestamp.

A.7 Generated numbers, tables, figures and citations

The reported statistics are not typed into the source. A generator reads the archived artifacts, emits the macros used by the text, the tables and the figures, and writes a provenance map recording, for every macro, the artifact and JSON path it came from; the same pipeline re-derives

the headline paired estimates directly from raw per-world records and asserts that they match, and the numerical audit re-checks the emitted macros against the frozen records before the PDF is produced. Figures are produced at the paper's true text width and pass programmatic layout checks that archive the drawn-text overlap result for the calibration figures and canvas containment for the external figure (`audit/calibration/figure7_layout_check.json`, `audit/figure_qa.json`). References were checked programmatically against Crossref and arXiv — and, for documentation entries, against the reachability of the cited URL — with every cited entry verified before it appears in the text; entries that could not be verified were excluded rather than cited (`audit/reference_factual_audit.json`).

A.8 S40-R3C: POST-OUTCOME REPAIR VALIDATION PROVENANCE

S40-R3C validates the frozen pipeline on a cross-fitted LongMemEval-V2 evaluation population that reuses the earlier LongMemEval-V2 evaluation block's questions and substrate (this is not the S35 synthetic 2000-world held-out block); it was specified and analysed after that block's outcomes had been observed, so it is *post-outcome repair validation* and is classified as secondary evidence throughout. Its records were sealed before analysis. The reader issued 1353 of 1353 planned calls (missing 0, duplicates 0) under frozen reader contract C1; the deterministic evaluator completed 885 of 885 items with no errors; and the frozen judge completed 468 of 468 rows with 0 parse failures (J1 association audit PASS). Reading stopped at the configured generation cap on 54 of 1353 calls (0.0399); cap hits are taken from the recorded per-call token accounting rather than from the finish reason alone, and no item was selectively re-run, in this block or in any other. The block carries a recorded run-level protocol deviation (`R3C_RUNTIME_PROTOCOL_DEVIATION` = true) with `SCIENTIFIC_CONTRACT_CHANGED` = false: the deviation concerns operational handling of the run, while the estimand, the estimator, the bootstrap procedure and its seed were unchanged. The per-condition counts $Y(\text{matched}) \neq Y(\text{mismatched}) = 47$, $Y(\text{matched}) \neq Y(\text{reference}) = 81$ and $Y(\text{mismatched}) \neq Y(\text{reference}) = 86$ are correctness-discordance (Y-disagreement) counts; they compare correctness, not decisions, and nothing is derived from them.

A.9 PER-STRATUM BEHAVIOUR

Table A2: S35 held-out τ by stratum (50 paired worlds each). The Δ column is $50\,\tau$, i.e. discordant pairs favouring $s+$ minus those favouring $s-$.

Domain	Depth	Reporter order	Gold	Δ	τ
crew_dispatch	D2	reporter_a_first	A	1	+0.020
crew_dispatch	D2	reporter_a_first	B	0	+0.000
crew_dispatch	D2	reporter_b_first	A	4	+0.080
crew_dispatch	D2	reporter_b_first	B	-2	-0.040
crew_dispatch	D3	reporter_a_first	A	-1	-0.020
crew_dispatch	D3	reporter_a_first	B	1	+0.020
crew_dispatch	D3	reporter_b_first	A	-1	-0.020
crew_dispatch	D3	reporter_b_first	B	-1	-0.020
pipeline_window	D2	reporter_a_first	A	1	+0.020
pipeline_window	D2	reporter_a_first	B	3	+0.060
pipeline_window	D2	reporter_b_first	A	0	+0.000
pipeline_window	D2	reporter_b_first	B	3	+0.060
pipeline_window	D3	reporter_a_first	A	0	+0.000
pipeline_window	D3	reporter_a_first	B	2	+0.040
pipeline_window	D3	reporter_b_first	A	1	+0.020
pipeline_window	D3	reporter_b_first	B	0	+0.000
production_batch	D2	reporter_a_first	A	-1	-0.020
production_batch	D2	reporter_a_first	B	1	+0.020
production_batch	D2	reporter_b_first	A	0	+0.000
production_batch	D2	reporter_b_first	B	2	+0.040
production_batch	D3	reporter_a_first	A	0	+0.000
production_batch	D3	reporter_a_first	B	-2	-0.040
production_batch	D3	reporter_b_first	A	3	+0.060
production_batch	D3	reporter_b_first	B	3	+0.060
sample_courier	D2	reporter_a_first	A	-3	-0.060
sample_courier	D2	reporter_a_first	B	3	+0.060
sample_courier	D2	reporter_b_first	A	1	+0.020
sample_courier	D2	reporter_b_first	B	1	+0.020
sample_courier	D3	reporter_a_first	A	1	+0.020
sample_courier	D3	reporter_a_first	B	-3	-0.060
sample_courier	D3	reporter_b_first	A	0	+0.000
sample_courier	D3	reporter_b_first	B	1	+0.020
shipment_routing	D2	reporter_a_first	A	-1	-0.020
shipment_routing	D2	reporter_a_first	B	5	+0.100
shipment_routing	D2	reporter_b_first	A	3	+0.060
shipment_routing	D2	reporter_b_first	B	-1	-0.020
shipment_routing	D3	reporter_a_first	A	-2	-0.040
shipment_routing	D3	reporter_a_first	B	-2	-0.040
shipment_routing	D3	reporter_b_first	A	0	+0.000
shipment_routing	D3	reporter_b_first	B	0	+0.000

Per-stratum estimates are descriptive; the primary analysis is the pooled paired estimate.

A.10 ACCOUNTING

Across the programme 17297 reader calls were issued: 14800 in the five analysed blocks (S34, S35, the S36 controls, and the S37 DeepSeek readers), 1500 in the S33F development branch, 750 in the pre-TQ qualification pilot, 245 in the budget-truncated S37 arm, and 2 DeepSeek wire-compatibility probes. The probes and the truncated arm are excluded from every estimate, and the pilot and development calls never enter an effect estimate. Every issued call has its request/attempt record archived; every completed provider response archived by the experiment carries a response payload and a response digest; technical-missing attempts are recorded as such and carry no response payload. No archived record was deleted, rewritten or regenerated for any analysis in this paper, with the single scoped exception described in Appendix B. The frozen source artefacts behind every reported number are listed in Table A3.

Table A3: Frozen source artifacts behind every number, table and figure (first 12 hexadecimal characters of SHA-256; full digests in `audit/artifact_integrity_audit.json`).

Artifact (basename)	SHA-256 (12 hex)
S33E_PILOT_BRANCH_CLOSURE.json	2a65d15878e5
S33E_PILOT_QUALIFICATION_SUMMARY.json	d2de36f8b490
S33E_PILOT_SUFFICIENT_FAILURE_RCA_SUMMARY.json	bc6e53995bfe
S33E_PILOT_SUFFICIENT_FAILURE_RCA_VERIFY.json	06ed7d53af39
S33F_READER_REMEDIATION_RESULT.json	ab12416fadfc
S34_FINAL_EXPLORATORY_SEAL.json	d7bd4371c4a8
S34_POC_SEED_COMMITMENT.json	7b158faeabc6
S35_FINAL_RESULTS.json	9c1562c22e6e
S35_PROTOCOL_FREEZE.json	a77e39f47175
S35_QUICKNET_BLS_VERIFICATION.json	aa0712fe64e6
S35_QUICKNET_ROUND_COMMITMENT.json	682415eefd15
S35_SEED_DERIVATION.json	14482179e397
S36_C1_C2_IDENTICAL_INPUT_AUDIT.json	92ead88f4f33
S36_C3_EQUAL_QUALITY_RESULTS.json	6009819d92e9
S36_C4_ALLOCATION_DOSE_RESPONSE_RESULTS.json	879fe84b6897
S36_INTEGRATED_MECHANISM_TABLE.json	615f14b4c26a
S36_QUALITY_ALIGNMENT_INTERACTION.json	6698f30a7215
S37_BUDGET_AMENDMENT_AND_FINAL_PI_ADJUDICATION.json	a5e046d343a7
S37_CROSS_MODEL_FREEZE.json	c74de46461f6
S37_DEEPSEEK_WIRE_PROBE_EXCLUSION_INDEX.json	8f3154395ebb
S37_PROVISIONAL_DEEPSEEK_RESULTS.json	0b8179fcb821
S37_SHARED_POPULATION.json	ad81109fcf66

B Development history and disclosure

B.1 The qualification pilot that failed, and why we changed the task

An earlier protocol in this project required the reader to answer under *universal completion* semantics: it had to consider every legal completion of every field left open by the visible evidence and answer only when all completions forced the same unique feasible plan, abstaining otherwise. That protocol ran 750 calls and failed at qualification. On the 600 packets in the pre-registered "at least one single-source field" class — the class the SUFFICIENT gate counted — the reader answered correctly 146, abstained 333 and answered incorrectly 121 times. Of those abstentions, 211 occurred on the 416 packets with zero unknown fields, and among the 444 packets whose relevant fields were all known the abstention rate was 50.9%, i.e. 226 of 444. In this appendix "fully specified" always refers to the material/world specification; the reader-visible evidence can still leave fields open, which is exactly why the two classes above are separated. Against the pre-registered requirement of 570 correct responses the observed value was 146, and the branch closed with its confirmatory stage withheld (0 confirmatory calls). No response filtering, rewriting or regeneration was performed.

A read-only root-cause audit then rebuilt the *reader-visible* oracle d_R^* — the answer implied by the packet the reader would see, which is a different object from the latent-world gold d_W^* used as the outcome in this paper (Section 4) — for all 600 packets of that class, by exhaustive completion enumeration and by an independent reimplementation, and reproduced every persisted qualification gold label exactly. The audit concluded that the failure was a reader-behavioural limitation with secondary prompt-contract incompleteness — not an oracle, rendering, material or accounting defect — and that a new random seed would not reasonably have repaired it. Since the universal-completion requirement was a property of that reader protocol rather than of the causal question, we kept the world, the allocator, the renderer, the state definitions and the oracle, and replaced the reader-facing task with the direct evidence-based decision used throughout this paper. A further remediation branch froze two candidate readers and one prompt that restated the completion contract explicitly and ran 33 development gates over the same packet bytes; candidate 1 (750 calls) passed 11 gates with 540 abstentions and candidate 2 (750 calls) passed 2 gates with 286 abstentions. Both failed, the branch returned FAIL-STOP after 1500 development calls, and no new protocol was created from it.

Two consequences matter for how this paper should be read. First, the reader was changed after those failures were observed, so no block here is pre-pilot: blocks are labelled exploratory or held-out, and the held-out block is the only pre-registered estimate. Second, the exploratory sample was extended after its first worlds had been observed, and it is reported as outcome-informed.

B.2 BUDGET AMENDMENT AND WRITE RESIDUE

The cross-model block was originally specified with three readers. The third arm was terminated by provider budget exhaustion after 245 of 1000 calls; its completed responses are archived verbatim, it is classified as a budget-truncated partial arm, no partial effect estimate was computed or promoted from it, and it is excluded from every number in this paper. This is an operational decision, not a scientific failure of that reader. Separately, one archived call file in the cross-model block was left with two stray bytes by a concurrent writer; the original bytes are preserved, the file was restored to the byte content the frozen writer produces (verified by re-deriving the record's request digest from the frozen order), no model call was re-issued and no response was regenerated. The forensic evidence binds the stored record to the frozen request and shows that the repaired file is byte-identical to the writer's canonical serialisation; it does not establish that the provider's original response bytes are immutable, and we do not claim that. The incident is recorded as a non-scientific write-atomicity residue.

C MECHANISM CONTROLS, PER CONTROL

C1 (allocation fixed at $4/4$). Forcing both arms to $4/4$ means the state cannot change what is retrieved: upstream allocation change is 0 and reader-visible evidence change is 0 of 200 pairs, so the reader sees identical inputs within each pair and the only decision changes are repeated-call disagreement (5 of 200, 2.5%). Arm accuracies are 0.8950 and 0.8900.

C2 (rendering clamped at the reader boundary). C2 preserves the ordinary paired intervention upstream and clamps only what the reader sees. The upstream allocation therefore still differs by arm in 200 of 200 pairs, while the reader-visible rendering is identical within every pair (0 changed). Decision changes again occur at the baseline rate (5 of 200), with arm accuracies 0.9050 and 0.8900. Neither the text nor the tables may describe C2 as an identical allocation: it is an identical *reader input*.

C3 (equal average source quality). With both sources at 0.55 and the allocation unchanged, arm-dependent decision differences remain observed and the point estimate remains positive: $\tau = +0.0130$ (+1.30 pp, interval -0.40 to +3.10, which includes zero), arm accuracies 0.9099 and 0.8969, with 999 and 999 pairs showing allocation and rendering change and 81 decisions changing. Preferential allocation towards a better source is therefore not necessary for arm-dependent differences to appear; this does not show that source quality is irrelevant, because the realised quality of the retrieved evidence can still differ between arms in a given world. *Bookkeeping.* C3 generated 1000 worlds (25 per stratum across the 40 strata) and issued 2,000 calls; for one world a single call was recorded as a technical missing (no recoverable provider response), so 1 pair was excluded from the paired analysis, leaving 999 analysable pairs. No pair was dropped because of its outcome, and nothing was resampled.

C4 (allocation sweep). On a common set of 500 worlds and 2000 calls, accuracy is 0.9040 ($4/4$), 0.9100 ($5/3$), 0.9080 ($6/2$) and 0.9400 ($7/1$). Relative to $5/3$ the $4/4$ and $6/2$ differences are small and uninformative, while $7/1$ is higher by +3.00 pp (+0.60 to +5.40). The observed point estimates across the four tested allocation levels are non-monotonic; we label the pattern NON-MONOTONIC_ALLOCATION_RESPONSE, no condition was dropped, and monotonicity was neither required nor observed.

C5 (quality alignment reversed) and the alignment contrast. Placing the better source on the $s-$ side ($0.50/0.60$, allocation unchanged) reverses the point estimate: $\tau = -0.0150$ (-1.50 pp, interval -3.40 to +0.40), with 1000 and 1000 pairs changing allocation and rendering and 93 of 1000 decisions changing. The interval includes zero, so this is a point-estimate reversal rather than established negative causation. Contrasting the held-out block with C5 gives an alignment contrast of +2.50 pp (+0.20 to +4.80; 20000 paired bootstrap draws, seed 20261003, analytic check +0.21 to +4.79). This contrast was computed after both input blocks had been run and is therefore secondary and non-pre-registered.

Observed between-arm decision disagreement. The observed fraction of paired worlds with different decisions between arms is 8.7% for Qwen S35, 16.4% for DeepSeek Flash and 29.0% for DeepSeek Pro. Because the Qwen and DeepSeek rows come from different populations, and reader-

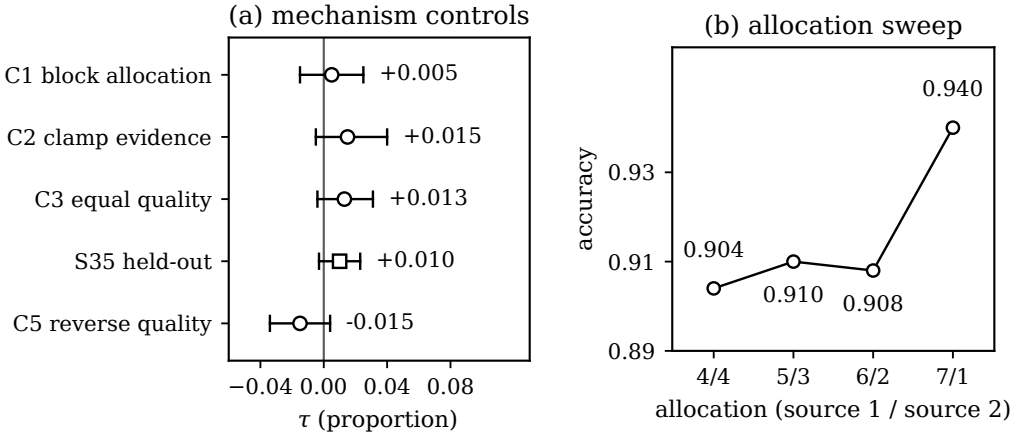

Figure 4: Mechanism controls. (a) paired effect τ for C1, C2, C3 and C5, with the held-out block as a reference; the two blocking controls leave only identical-input disagreement, whereas C3 retains a positive point estimate and C5 reverses it. (b) observed point estimates across the four tested allocation levels, which are non-monotonic.

specific identical-input repeat baselines are unavailable for the DeepSeek readers, these values are descriptive: they are not an intrinsic reader-sensitivity measure, not a capability comparison, and not a benchmark.

D EXPLORATORY SYNTHETIC CROSS-MODEL REPLICATION (S37)

One shared population. On one shared population of 500 paired worlds (40 strata) we issued the identical decision task — the same worlds, repository realisations, selected evidence and instructions — to two further, independently served readers, 1000 calls each (Table D1, Figure 5). Reader 1 gives $\tau = +0.0080$ (+0.80 pp, interval -0.028 to +0.044; 82 of 500 decisions differ between arms). Reader 2 gives $\tau = +0.0140$ (+1.40 pp, interval -0.032 to +0.062; 145 of 500 differ). Both intervals include zero. Across the held-out Qwen population and this shared DeepSeek replication population, all three served readers produced positive but uncertain point estimates under the same state intervention; because none of the readers exposes an immutable snapshot identifier, we label this an exploratory cross-model replication rather than snapshot-exact replication.

Observed between-arm decision disagreement across readers. The observed fraction of paired worlds with different decisions between arms is 8.7% for Qwen S35, 16.4% for DeepSeek Flash and 29.0% for DeepSeek Pro (Figure 5b). Because the Qwen and DeepSeek rows come from different populations, and reader-specific identical-input repeat baselines are unavailable for the DeepSeek readers, these values are descriptive and are not interpreted as an intrinsic reader-sensitivity ranking or as a comparison of reader capability.

Disclosure. The cross-model block was specified with three readers[1]; the two reported here are the arms that completed.

[1]The third arm was stopped by provider budget exhaustion after 245 of 1000 calls. Its completed calls are archived verbatim, no partial effect estimate was computed from them, and it is excluded from every number above (Appendix B).

Table D1: S37 cross-model exploratory replication on a shared 500-world paired population. Both readers receive identical world-specific evidence and must answer A or B; abstention is not permitted.

Reader	N pairs	Calls	Acc $s+$	Acc $s-$	τ	95% CI	Δ $(+1/0/-1)$	D chg.
deepseek-flash	500	1000	0.8400	0.8320	+0.0080	[-0.028, +0.044]	43/418/39	82 (16.4%)
deepseek-v4-pro	500	1000	0.6860	0.6720	+0.0140	[-0.032, +0.062]	76/355/69	145 (29.0%)

The pre-specified Qwen replication arm was budget-truncated after 245/1000 calls and is excluded from effect estimation; complete primary Qwen S34/S35 results appear in Table 2, and the truncated S37 Qwen arm has no effect estimate. The DeepSeek Official aliases are rolling, so this block is exploratory.

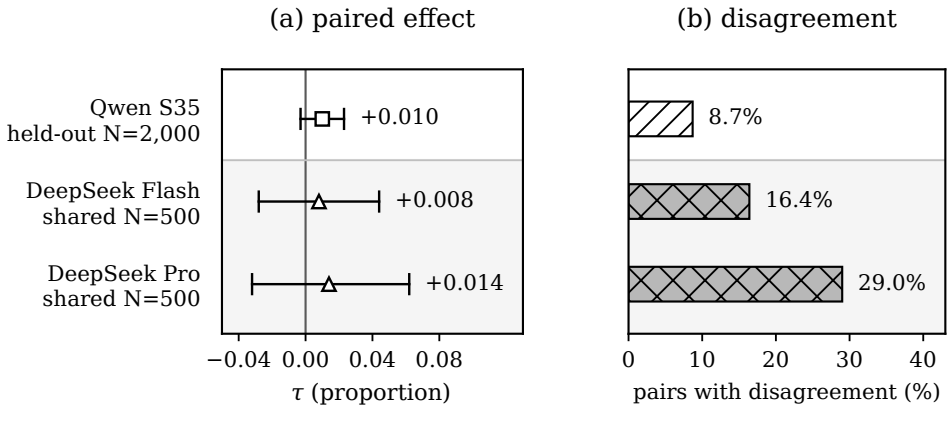

Figure 5: (a) paired effects with 95% bootstrap intervals for the held-out Qwen reference (square) and the two exploratory DeepSeek readers (triangles). The divider separates the Qwen population ($N = 2000$) from the shared DeepSeek population ($N = 500$). (b) observed between-arm decision disagreement; descriptive, not an intrinsic reader-sensitivity ranking.

E EXTERNAL-VALIDITY METHODS, CONTROLS AND DISCLOSURES

E.1 FROZEN DESIGN OF THE EXTERNAL BLOCKS

Both external blocks reuse the StateSwap structure end to end: a paired case, one application-level persistent retrieval policy state, an exactly recorded allocation A, the exact reader-visible evidence R (with a digest), a masked served reader, and a decision D scored against a case-level gold that is identical across arms and never shown to the reader.

S38 (NVD severity transfer). $N = 180$ paired cases, 60 per gold class (AUTO_HANDLE < 4.0, REVIEW 4.0–7.0, ESCALATE ≥ 7.0 on the incoming CVE's published CVSS v3.1 base score). Candidate memory is drawn only from records published strictly before the incoming CVE; twelve candidates are sampled deterministically from that temporally valid pool and each arm selects five. Gold never enters retrieval scoring.

S39R1 (CISA KEV prioritization). $N = 120$ paired cases, 60 of which are in the frozen CISA KEV catalogue (decision time frozen at the catalogue timestamp). Non-KEV cases were matched 1:1 on CVSS band and publication year (± 1), with a vendor match preferred where derivable. Retrieval uses the same two policies; historical KEV status is deliberately not shown to the reader.

E.2 READER CONFIGURATION AND PROVIDER ACCOUNTING

Both readers were served through frozen single-turn, non-thinking, JSON-only configurations with a fixed 24-token output budget and no tools or conversation memory. The DeepSeek reader is the rolling alias `deepseek-flash`; the Qwen reader is the rolling alias `qwen-flash` (the provider's dated snapshot identifier is not served by its own model list, so the alias is frozen instead and this is disclosed).

Call accounting is deliberately reported as separate quantities rather than one aggregate "N" (Table E1). The total number of provider requests is much larger than the number of valid analysed calls; the two are never combined into a single scientific sample size.

Category	DeepSeek	Qwen	Note
Provider requests issued	1880	1660	all external blocks plus the invalid predecessor
Provider responses received	1880	1040	a received response is not automatically a valid analysed call
Archived responses surviving	1460	1040	420 DeepSeek responses were overwritten before archival recovery
Deterministic replay requests	420	–	byte-identical frozen requests; NOT independent scientific samples
Valid analysed calls	1040	1040	S38 620 + S39R1 420 per reader
Invalid predecessor calls	420	–	original S39 execution; scientifically invalid, excluded from every estimate
Blocked attempts with no response	–	620	DashScope arrears; zero reader responses obtained
Ad-hoc provider probes	2	5	metadata/connectivity checks; no scientific content

Table E1: Provider call accounting for the external-validity programme. Quantities are deliberately kept separate: issued requests, received responses, valid analysed calls, invalid predecessor calls, replay requests and blocked attempts are not combined into a single scientific sample size.

E.3 DISCLOSURE: AN INVALID EXTERNAL PREDECESSOR (ORIGINAL S39)

A first execution of the S39 design was run before the external-validity freeze was finally reviewed. That execution inherited the S38 three-class decision vocabulary through a renderer mismatch, so the served reader was never asked the S39 question. Its responses are therefore *scientifically invalid for the S39 estimand*: no paired accuracy, no τ, no interval, no D-change interpretation and no effect estimate from that execution enters this paper, and no post-hoc label conversion (for example

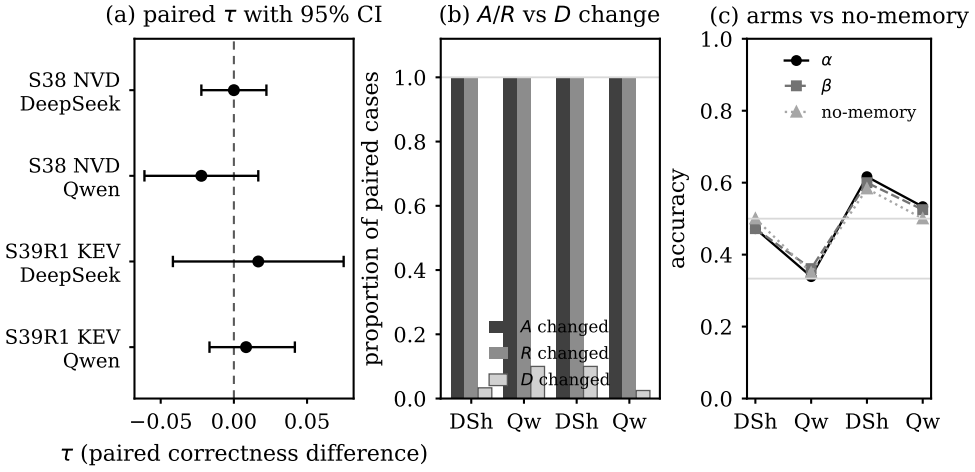

Figure 6: External-validity stress tests over two semi-naturalistic tasks and two independently served reader families (four blocks); the corresponding numeric summary is Table H2 in the main text. (a) paired correctness contrast τ with 95% percentile paired-bootstrap intervals; the zero reference line is dashed and every interval crosses it. (b) proportion of paired cases in which the selected records (A), the rendered evidence (R) and the decision (D) changed: A and R change in every pair while D rarely does. (c) arm accuracy against the incoming-only (no-memory) baseline. Readers are not ranked and raw disagreement counts are not interpreted as reader sensitivity.

mapping ESCALATE to URGENT_PATCH) was performed. The raw responses are preserved verbatim as an invalid predecessor. The corrected experiment (S39R1) was separately frozen and separately executed after a machine audit that asserts zero occurrences of the wrong vocabulary in every reader-visible request.

E.4 DISCLOSURE: ONE RESPONSE-REPLAY INCIDENT

During the external-validity execution, one engineering incident overwrote 420 archived DeepSeek S38 responses because two tasks shared a single logical-index filename namespace. The uniquely missing logical calls were determined mechanically, and those requests were deterministically re-issued using byte-identical frozen request bodies. No population, prompt, state, retrieval rule, gold label, analysis or selection was changed, and no outcome-based choice determined which calls were replayed. Replay calls are *not* independent scientific samples and are counted separately (Table E1). For this reason the archive guarantee claimed in this work is stated with an explicit scope: archived completed responses are re-derivable from the frozen request bytes, and the single namespace-collision incident above is the recorded exception.

E.5 STATESWAP-LITE: AN ILLUSTRATIVE LOW-COST DEPLOYMENT PROTOCOL

The full programme above is expensive. For a reader who wants a lightweight path evaluation rather than a precision effect estimate, we package an illustrative low-cost deployment protocol: a paired path-sanity block ($\approx$ 100 calls), an identical-input repeat block ($\approx$ 60 calls) and one clamped-evidence control ($\approx$ 100 calls) — in total $\approx$ 260 calls. The profile targets *pathway observability and serving variability*; it is not intended to detect a $\sim$ 1 pp effect reliably, and a retrospective subsampling analysis of the held-out block shows why (at $N = 200$ the mean bootstrap interval width is roughly 0.11 in paired-correctness units). A dependency-free reference implementation is provided in which a user supplies only three application-specific components (state, retrieve, render) and the harness logs S, A, R, the request digest, the archived response, D, Y, the paired difference and the control summaries.

E.6 BOOTSTRAP EXACTNESS

Every interval reported for the primary and mechanism blocks is a 95% percentile paired-bootstrap interval, resampling paired worlds with replacement (10,000 draws; analysis seed 20261001). The alignment contrast is an independent two-block percentile bootstrap (20,000 draws; seed 20261003). The external-validity blocks use their own frozen analysis seeds; no seed is reused across blocks, and no BCa, basic (reverse-percentile) or studentized bootstrap is used anywhere.

F FORMAL STATEMENTS, ASSUMPTIONS AND PROOFS

F.1 SYSTEM CLASS

A paired system in the class $\mathcal{S}$ realises each unit w twice, once per arm $a \in \{0, 1\}$, differing only in an application-level persistent state $S(a)$; the allocation is a frozen deterministic map $A_w(a) = \mathcal{A}(w, S(a), \xi_w)$ of the world, the state and treatment-independent randomness ξ_w; the rendering is a frozen deterministic map $R_w(a) = \mathcal{R}(A_w(a))$; and the reader is a possibly stochastic map $D_w(a) = g(R_w(a), u_w, \zeta_w(a))$ of the reader-visible evidence, the arm-invariant request content u_w and serving randomness $\zeta_w(a)$. Nothing outside $\mathcal{S}$ is claimed.

F.2 ASSUMPTIONS AND THEIR STATUS

Table F1: Assumptions used by the propositions, and where each one comes from.

Id	Assumption	Status and source
S1	paired latent object fixed: d^*_W, task, plans, constraints and the pre-state repository realisation identical across the arms of a pair	frozen design property (paired-world construction)
S2	S is the sole manipulated application-level variable	frozen design property (only the stored allocation policy differs)
S3	treatment-independent allocator randomness fixed and shared within the pair	frozen design property (field permutation drawn from a treatment-independent stream)
S4	allocator and renderer frozen and deterministic, with no arm-conditional editing, injection, reordering or dropping	frozen design property (frozen production implementations)
O1	the exact $A_w(a)$ and $R_w(a)$ are observed for both arms	frozen design property (per-pair A changed / R changed and rendered-bytes digests)
M1	state, arm, allocation, reliabilities, world identifier and gold withheld from the reader	frozen design property (reader contract); used by the clamped path, *not* by the attribution result
C1	reader-visible evidence byte-identical in the pair	frozen design property of the clamp controls
C2	same reader contract in both arms	frozen design property (single prompt, arm-invariant content, single parser)
C3a	arm-invariant *marginal* serving mechanism	**theoretical condition, not verified**; supported only indirectly by balanced execution order and the post-hoc execution-order sensitivity analysis
C3b	the two served draws are conditionally independent with a common discrete law	**theoretical condition, not empirically verified**; strictly stronger than C3a

F.3 EMPIRICAL STATE-INTERVENTION PROFILE

Table F2: State-intervention profile $\Pi_S = (p_A, p_R, p_D, \tau)$ for the analysed blocks, from frozen counts only (p_A and p_R are A changed and R changed over pairs; p_D is observed paired decision disagreement). Intervals are the frozen paired intervals and have not been recomputed. The C1/C2 rows are the clamped-path reference: for C2 the reader-visible rendering/evidence is unchanged while the upstream allocation still differed by arm in 200 of 200 pairs. The final four rows are exploratory external stress tests.

Block	N	p_A	p_R	p_D	τ	95% interval
S34 exploratory	1000	1.000	1.000	0.069	+0.0190	[+0.003, +0.035]
S35 held-out	2000	1.000	1.000	0.087	+0.0100	[-0.003, +0.023]
C1 fixed allocation	200	0.000	0.000	0.025	+0.0050	[-0.015, +0.025]
C2 clamped rendering	200	1.000	0.000	0.025	+0.0150	[-0.005, +0.040]
S38 × DeepSeek	180	1.000	1.000	0.033	+0.0000	[-0.022, +0.022]
S38 × Qwen	180	1.000	1.000	0.100	-0.0222	[-0.061, +0.017]
S39R1 × DeepSeek	120	1.000	1.000	0.100	+0.0167	[-0.042, +0.075]
S39R1 × Qwen	120	1.000	1.000	0.025	+0.0083	[-0.017, +0.042]

in the ordinary unclamped intervention blocks (S34, S35, S38 and S39R1), $p_A = p_R = 1$, while p_D is small. Zero coverage differs by block and is read off the interval column: the exploratory S34 interval excludes zero, whereas the pre-specified held-out S35 interval, the two path controls and the four exploratory external rows all include zero. This is the empirical face of the separation, not a test of it, and an interval that excludes zero is not proof of a non-zero effect any more than an interval that includes zero is proof of an exact zero. The external rows are exploratory external validity and are never pooled with the held-out block.

F.4 PROOF OF PROPOSITION 1 (PIPELINE ATTRIBUTION FOR A AND R)

By S2 and S4, the allocation of arm a is the evaluation $A_w(a) = \mathcal{A}(w, S(a), \xi_w)$ of one frozen map whose free arguments are the state, the world and the randomness. By S1 the world argument agrees across the arms, and by S3 the randomness argument agrees; hence the two arms evaluate the same frozen function at the same point with the state as the only differing argument, and

$$\{A_w(1) \neq A_w(0)\} = \big\{\mathcal{A}(w, S(1), \xi_w) \neq \mathcal{A}(w, S(0), \xi_w)\big\},$$

an identity rather than a set inclusion to be established. By S4, $R_w(a) = \mathcal{R}(A_w(a))$ is a deterministic function of $A_w(a)$, so $A_w(1) = A_w(0)$ forces $R_w(1) = R_w(0)$ and therefore $\{R_w(1) \neq R_w(0)\} \subseteq \{A_w(1) \neq A_w(0)\}$. O1 makes both events observable; M1 is not used. □

Proposition 1 is an identity about how the harness computes A and R and is not evidence of any behavioural effect. It does not extend to D, because the serving term $\zeta_w(a)$ is not constrained by S1–S4; the case $D_w(1) \neq D_w(0)$ with $R_w(1) = R_w(0)$ is exactly the clamped path of Proposition 3(a).

F.5 COUNTEREXAMPLE FOR PROPOSITION 2

Evidence-blind reader. Let $g_\varnothing(R, u, \zeta) = h(u)$ for a fixed deterministic h that depends only on the arm-invariant request content. Then for every unit $D_w(1) = h(u_w) = D_w(0)$, so $D_1 = D_0$, $p_D = \mathbb{P}(D_1 \neq D_0) = 0$, and because the arm decisions coincide, $Y_w(1) = Y_w(0)$ and $\tau = \mathbb{E}[Y_1 - Y_0] = 0$, while $p_A = p_R = 1$ by the construction of the intervention.

Evidence-using reader. Let $g_\star$ follow the arm-dependent part of R. On every constructed unit let the active arm expose the record that implies the paired gold target and let the reference arm not expose it, and let $g_\star$ return the implied plan. Then $D_w(1) = d^*_W$ and $D_w(0) \neq d^*_W$, so $Y_w(1) = 1$, $Y_w(0) = 0$, hence $\tau = \mathbb{E}[Y_1 - Y_0] = +1$, while the upstream coordinates are unchanged at $p_A = p_R = 1$. The two readers therefore share an upstream profile and differ on the decision side. In addition, a decision may change between arms while staying on the same side of the oracle (both correct or both incorrect), so $p_D > 0$ does not imply $\tau \neq 0$. □

F.6 Proof of Proposition 3(a) (clamped path)

By C1 and C2 the reader-visible input is a deterministic function of quantities that agree across the two arms, so the input is identical in both arms. By C3a the serving randomness is arm-invariant in law, so the marginal laws of $D_w(1)$ and $D_w(0)$ coincide: $\mathcal{L}(D_1 \mid R, u) = \mathcal{L}(D_0 \mid R, u)$. Equal laws have equal expectations, so $\mathbb{E}[Y_w(1)] = \mathbb{E}[Y_w(0)]$ because d^*_W is fixed within the pair, and hence $\tau = 0$. Nothing in the argument forces $D_w(1) = D_w(0)$ on a realised pair; the residual finite disagreement is compatible with serving variability and is not evidence of an arm-visible treatment response. □

Proposition 3(b) (collapsed path; conditional independence). *This lemma carries the theoretical condition C3b, marked in Table F1 as a theoretical condition that is not empirically verified. It is an optional addition and may be dropped without affecting Proposition 3(a).* Assume C1, C2, C3a and C3b, so that the two served decisions are independent draws from one common discrete law π on the decision alphabet. Then

$$p_D = \mathbb{P}(D_1 \neq D_0) = 1 - \sum_d \pi_d^2 = 2 \sum_{d<d'} \pi_d \pi_{d'},$$

the collision (Gini-impurity) functional of π. *Proof.* Independence and a common law give $\mathbb{P}(D_1 = D_0) = \sum_d \pi_d \pi_d = \sum_d \pi_d^2$, and $p_D = 1 - \mathbb{P}(D_1 = D_0)$; the second expression follows from $1 - \sum_d \pi_d^2 = \sum_d \pi_d(1 - \pi_d) = 2 \sum_{d<d'} \pi_d \pi_{d'}$. □ Because p_D is a dispersion functional of the reader's law while τ is a difference of means, no algebraic identity relates the two, so the clamp disagreement rate cannot be subtracted from τ; the condition C3b is required even for this collapsed form, and it is not verified by the recorded data.

G External calibration: provenance, disclosures and denominators

What the calibration measures and what it cannot measure. Both calibration blocks reuse *one* fixed question population and *one* annotation-assisted pool-compilation rule. Official supporting-fact annotations participate in compiling the pool, so the calibration studies the observable request path on a frozen population; it contains no intervention on an application-level state S and does not evaluate a learned allocation policy, but the conditions do differ in their selected paragraphs and rendered reader-visible evidence, so it cannot be used as evidence that a persistent state improves decisions or that a reader acquired an internal state. The pre-specified held-out block of the main experiments keeps its original identity; the calibration is an additional, separately specified block, not a retroactive promotion of this paper's primary experiment.

Population, masking and pool compilation. The population is 1000 HotpotQA distractor-validation questions with two officially annotated supporting paragraphs and two distractor paragraphs per question. Condition B renders all four; I replaces one paragraph that carries no supporting-fact annotation; K replaces one annotated supporting paragraph. Exactly one of four paragraph slots is exchanged, so the two interventions are equal-edit in that restricted sense — the number of exchanged paragraphs under the fixed reference-length rule — which is not a claim about semantic edit distance or about served token counts; p_A and p_R are 1 for every pair as a property of the construction, not an empirical finding. Lengths are matched by the frozen reference length rule, and the two arms of a pair share the reader-visible ordering rule. The reader is not given the treatment arm, the supporting-fact annotations, the allocation or the gold answer; the rendered evidence itself (paragraph indices and source titles) is visible by design. The question, the gold and the frozen messages are identical in the two blocks. Replacing an annotated paragraph does not guarantee that all answer-bearing evidence is removed, and replacing an unannotated paragraph does not guarantee that the paragraph was irrelevant; the human review recorded both cases, and we do not claim otherwise.

Retained source-question defects. The material review passed 26 of 30 items and recorded four *source-question* issues rather than compiler or transcription defects. These four items (`5a8284b5...`, `5ae00a27...`, `5ab7f746...`, `5a773220...`; questionnaire items 2, 13,

Table G1: External annotation-assisted calibration on one fixed population of 1000 questions, two independently served reader families, and three four-paragraph conditions (B all four; I one unannotated paragraph replaced; K one annotated supporting paragraph replaced). Accuracy is the full-N operational exact match, so truncated cells count as incorrect. $\theta = \mathrm{EM}_I - \mathrm{EM}_K$ carries the 95% cluster interval; τ_I carries the 90% interval compared with the pre-specified ± 5 pp equivalence band; τ_K carries its 95% interval. p_D is canonical-answer disagreement over usable pairs. The two readers are not independent samples and are not ranked.

Quantity	DeepSeek reader	Qwen reader
Questions N	1000	1000
Cap-hit cells (of 3400 HTTP)	0	7
EM B / I / K	0.642 / 0.634 / 0.297	0.581 / 0.569 / 0.259
$\theta = \mathrm{EM}_I - \mathrm{EM}_K$	$+33.70$ pp	$+31.00$ pp
95% interval	$[+30.63, +36.87]$	$[+28.04, +34.00]$
$\tau_I = \mathrm{EM}_I - \mathrm{EM}_B$	-0.80 pp	-1.20 pp
90% interval	$[-1.89, +0.30]$	$[-2.40, +0.00]$
$\tau_K = \mathrm{EM}_K - \mathrm{EM}_B$	-34.50 pp	-32.20 pp
95% interval	$[-37.60, -31.49]$	$[-35.17, -29.23]$
$p_D(I)$ / $p_D(K)$	0.092 / 0.545	0.136 / 0.588
usable pairs for $p_D(I)$ / $p_D(K)$	1000/ 1000	997/ 996
θ difference (Qwen − DeepSeek)	-2.70 pp, 95% $[-5.79, +0.40]$	

19 and 25) are *retained*, disclosed and not repaired, and the same four items are used by both reader blocks. Support redundancy was additionally noted on review items 8, 15, 19, 20, 21 and 24. No item was dropped or replaced after any reader outcome existed.

DeepSeek block: evidence class and freeze. The reader contract was frozen at `2026-09-24T17:15:25Z`; the first request of the block followed at `17:15:55Z` and the last at `17:50:34Z`, so every scientific request postdates the freeze (`BUILD_SEAL.json 50B B81FBFEAF6CA386DF3AFD083259D5AF73D2FA3B05096A837F60187B662E93`). All 3400 logical calls completed with no cap hit and no invalid cell. The sealed analyser hard-codes the conservative label NOT_FROZEN; the evidence class is recorded separately as PROSPECTIVELY_SPECIFIED_EXTERNAL_CALIBRATION_AFTER_R3C, determined after verifying the seal, the frozen contract, the completed human review and the attempt-to-request binding. A disclosed post-freeze edit to the analyser was detected by the seal, restored bit-exactly before the final analysis, and the sealed analyser reproduced identical statistics before and after the restoration; the additional reporting lives outside the seal. The block is **not confirmatory** and **not an independent replication**.

Qwen block: freeze, run facts and the implementation deviation. The reader contract was frozen at `2026-09-25T01:42:47Z`, the first request followed at `01:43:10Z` and the last completed at `02:17:18Z`. All 3400 HTTP responses returned status 200 with no retry; 3393 responses were usable and 7 were truncated at the completion cap (B 1, I 2, K 3 inside the main conditions, and 1 in the closed-book probe). Truncated and invalid cells count as incorrect under the frozen full-N operational endpoint; no item was dropped, no cap was raised and no truncated call was retried for a longer answer.

The frozen Qwen analyser re-exported the resampling routine from the DeepSeek block and then rebound the module-level bootstrap seed. Re-export does not rebind a function's already-evaluated default arguments, so the analyser passed the inherited DeepSeek seed while labelling the intervals with the replication seed. An independent audit found the mismatch. The corrected intervals reported in the main text pass the pre-specified replication seed explicitly, using the same components, the same 20000 resamples, the same full-N operational endpoint and the same percentile method; the correction lives outside the seal, the original sealed results and figure are preserved unchanged, and no model call was made to produce it. Both interval sets are reproduced in Table G2.

Table G2: Qwen block: intervals under the pre-specified replication seed (used in the main text) and under the inherited default seed that the frozen analyser actually passed (as originally reported). The point estimates are identical; the joint decision is unchanged.

Quantity	pre-specified seed	inherited default seed
θ 95% interval	$[+28.04, +34.00]$	$[+28.07, +34.03]$
τ_I 90% interval	$[-2.40, +0.00]$	$[-2.40, +0.00]$
τ_K 95% interval	$[-35.17, -29.23]$	$[-35.26, -29.26]$

Under either interval set, the θ lower bound stays above zero and the τ_I 90% interval stays inside the ± 5 pp equivalence band, so the pre-specified joint pattern is met; the deviation is disclosed, not silently repaired, and the seal was not re-created. Source artefacts: `r6_corrected_results.json` `AD2E2D62DAAB357E32B5C4FF4BD9342087A841DC5657C8B8A8A78FC41825DE3C`, `IMPLEMENTATION_DEVIATION_DISCLOSURE.json` `079BBA7527BEB5A106080FB57F82F1EBD4534E286AA04079567621FA7C533835`, `R6_RUN_EVIDENCE_CORRECTED.json` `2B4FC217D72D107B90EE4457A98E5936A7856A2BB65EC9FA277CBABE906F36F8`. The Qwen block is a prospectively specified cross-family reader replication on the same fixed sample: **not confirmatory**, **not an independent replication**, and not an independent sample of questions.

Denominators. Full-N operational accuracies use all 1000 questions per condition with truncated cells scored as incorrect. Quantities that require a usable answer on both sides carry their own denominators: the Qwen canonical-disagreement and Y-disagreement quantities use 997 pairs for I versus B and 996 for K versus B, and the K abstention transitions are reported over 996 usable pairs. The DeepSeek block had no truncated cells, so its corresponding denominators are the full 1000. The full-N and usable-only quantities are never mixed in one comparison.

Abstention status differs between the two blocks. For the Qwen block the three-way answer decomposition was specified before its outcomes existed and is a secondary descriptive quantity. For the DeepSeek block the equivalent decomposition was computed after its outcomes were known and is post-hoc. We report them separately and do not merge the two statuses; neither decomposition changes a primary endpoint or the joint decision. The UNKNOWN class is the abstention the prompt explicitly allows and does not establish that the model lacked the parameter knowledge; the remaining non-exact-match class can contain correct aliases that the pinned scorer does not accept, and is not called hallucination.

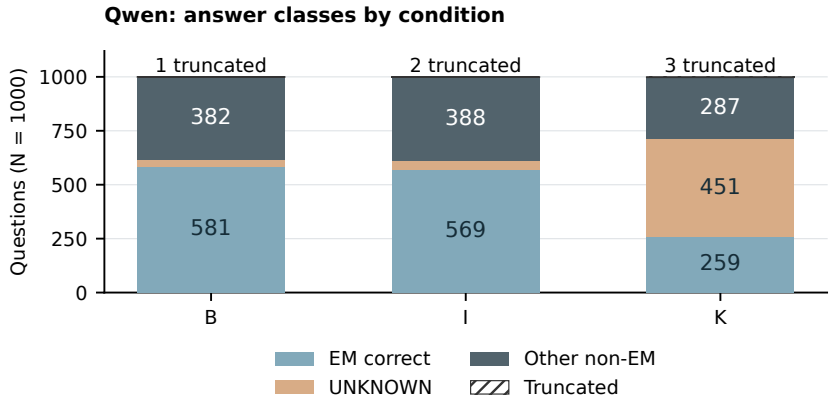

Figure 7: Qwen reader: correct, prompt-allowed UNKNOWN, and remaining non-exact-match answers per condition, with the truncated cells annotated above each bar. Category boundaries are scorer- and prompt-defined; the decomposition is descriptive and does not change the primary endpoints.

What the calibration does not establish. It does not show that either reader is generally competent, that persistent-state interventions improve decisions, that the removed annotation was necessary on every question, that the two readers are equivalent (the difference interval includes zero, which is the absence of a detected difference, not equivalence), or that retrieved memory is useless. It is one external task population with two readers and one annotation-assisted compilation rule, and magnitudes should not be extrapolated without in-situ replication.

H SUPPORTING BLOCKS: MECHANISM CONTROLS, STRESS TESTS AND THE POST-OUTCOME REPAIR VALIDATION

This appendix carries the per-control and per-block detail for the supporting experiments that the main text summarises. The evidence classes stated in the main text apply unchanged: the mechanism controls and the cross-model block are exploratory, the held-out block is the paper's pre-specified primary block, and S40-R3C is post-outcome repair validation. The mechanism figure and the external-validity figure are Figure 4 and Figure 6.

H.1 MECHANISM CONTROLS, PER CONTROL

With equal average source accuracy (C3, $0.55/0.55$) arm-dependent decision differences remain observed and the point estimate remains positive, $\tau = +0.0130$ (+1.30 pp, interval -0.40 to +3.10), with arm accuracies 0.9099 and 0.8969: preferential allocation towards a *better* source is therefore not necessary for arm-dependent differences to appear, although realised evidence quality can still differ between arms within a world. Sweeping allocation intensity (C4, 500 worlds $\times$ four allocations) gives the non-monotonic observed point estimates listed in Table H1 across the four tested levels; no condition was dropped and monotonicity was not required. Reversing which source is better (C5, $0.50/0.60$) reverses the point estimate, $\tau = -0.0150$ (-1.50 pp, interval -3.40 to +0.40), with 93 of 1000 pairs changing decision; because the interval includes zero this is a reversal of the point estimate rather than established negative causation. The secondary quality-alignment contrast, TAU_NORMAL − TAU_REVERSED, is +2.50 pp (+0.20 to +4.80), computed after both input blocks had been run.

H.2 EXTERNAL-VALIDITY STRESS TESTS

S38 maps an NVD vulnerability description to AUTO_HANDLE, REVIEW or ESCALATE, with the unseen gold label taken from its published CVSS v3.1 band. **S39R1** asks for URGENT_PATCH or STANDARD_QUEUE at a frozen CISA KEV catalogue timestamp, with KEV membership as gold; that label reflects real-world exploitation and is not fully inferable from the description alone. Each task compares a recency-first policy (α) with a lexical analogy-first policy (β, BM25 over the incoming description). The policies rank the same frozen candidate pool; gold never enters retrieval scoring, every memory predates the decision, and the reader never sees the state, arm, score, severity or gold.

Table H2 reports the four task×reader blocks; Table F2 and Figure 6 give the profile and visual summary. S38 shows complete propagation through A and R but little decision response (DeepSeek $\tau = +0.000$ $[-0.022, +0.022]$; Qwen $\tau = -0.022$ $[-0.061, +0.017]$), and S39R1 gives small, uncertain point estimates (DeepSeek $\tau = +0.017$ $[-0.042, +0.075]$; Qwen $\tau = +0.008$ $[-0.017, +0.042]$). In all four blocks $p_A = p_R = 1$, D changes rarely, and every interval includes zero, so propagation through A and R does not identify a decision-side effect. The incoming-only baseline reproduces S38's severe-side bias, consistent with a reader/task-side bias already present without retrieved history rather than one uniquely attributable to retrieval; S39R1 tests the framework on an operational label, not a KEV-prediction claim.

Table H1: Mechanism controls. Each control modifies one link of the $S \to A \to R \to D$ chain while the other links are held fixed; intervals are paired bootstrap intervals over worlds.

Control	Modification	N	τ	95% CI	Upstream A	Visible R	D chg.
C1 fixed allocation	A forced to 4/4 in both arms	200	+0.0050	[-0.015, +0.025]	0	0	5
C2 clamped evidence	R clamped at the reader boundary	200	+0.0150	[-0.005, +0.040]	200	0	5
C3 equal quality	source accuracies 0.55/0.55	999	+0.0130	[-0.004, +0.031]	999	999	81
C5 quality reversal	source accuracies 0.50/0.60	1000	-0.0150	[-0.034, +0.004]	1000	1000	93

C1 and C2 remove the reader-visible difference: for C2 the upstream allocation still differed by arm in 200/200 pairs while the rendered evidence and therefore the reader input were identical within every pair. C4 (dose response: 500 worlds $\times$ four allocations $=$ 2000 calls) appears in Figure 4: accuracy is 0.9040 at 4/4, 0.9100 at 5/3, 0.9080 at 6/2 and 0.9400 at 7/1, a non-monotone response in which 7/1 exceeds 5/3 by +3.00 pp [+0.60, +5.40]. The quality-alignment contrast is +2.50 pp [+0.20, +4.80] (secondary, not preregistered).

	S38 $\times$ DS	S38 $\times$ Qwen	S39R1 $\times$ DS	S39R1 $\times$ Qwen
N pairs	180	180	120	120
Acc α	0.472	0.339	0.617	0.533
Acc β	0.472	0.361	0.600	0.525
τ	+0.000	-0.022	+0.017	+0.008
95% CI	$[-0.022, +0.022]$	$[-0.061, +0.017]$	$[-0.042, +0.075]$	$[-0.017, +0.042]$
D changed	6/180	18/180	12/120	3/120
A changed	180/180	180/180	120/120	120/120
R changed	180/180	180/180	120/120	120/120
No-memory acc.	0.500	0.350	0.583	0.500
Repeat disagree.	0/40	5/40	1/20	0/20
Clamp disagree.	2/60	4/60	1/40	2/40

Table H2: External-validity stress tests; one column per task$\times$reader block. α is the recency-first policy state and β the analogy-first state. A and R change in every paired case, D changes rarely, and every interval *in this table* includes zero (the exploratory S34 interval in Table F2 does not). Repeat and Clamp are bounded-input disagreement counts (controls), reported descriptively and never subtracted from τ. Readers are not ranked, and raw disagreement is not reader sensitivity.

H.3 S40-R3C, SECONDARY POST-OUTCOME REPAIR VALIDATION

S40-R3C reuses the held-out block's 451-question cross-fitted population and postdates its outcomes, so it is post-outcome repair validation (secondary), not confirmatory and not an independent replication. Three conditions (*matched*, *mismatched*, preregistered reference) are question-crossfitted and trajectory-shared across three folds at a fixed budget (K_{FINAL}=16 units, TOK_{FINAL}=4096 tokens; reader contract C1). Over $N = 451$ questions, $\Delta_{\mathrm{CF}} = +0.0200$ (95% paired question-bootstrap interval $[-0.0089, +0.0488]$): the estimate is *positive* and the interval *includes zero*, so we do *not* read it as a matched-state decision advantage. Fold estimates $+0.0000$, $+0.0658$ and -0.0068 are not directionally consistent while both domains are positive ($+0.0250$, $+0.0142$); slice intervals are not recomputed, and the reference-policy differences $+0.0599$ and $+0.0399$ are descriptive only. The stored per-condition correctness yields Y-discordance counts 47, 81 and 86, which compare correctness, not decisions. The accuracies are low (0.2439, 0.2239, 0.1840) and the primary interval includes zero; low accuracy is a property of the task and reader, not evidence of an invalid experiment. The condition-level results are in Table H3, and the full run record is in Appendix A.

Table H3: S40-R3C, secondary post-outcome repair validation ($N = 451$ frozen LongMemEval-V2 questions): correctness by condition and Δ_{CF} with its frozen 95% question-bootstrap interval, which includes zero.

Quantity	Count	Value
Matched accuracy	110/451	0.2439
Mismatched accuracy	101/451	0.2239
Reference accuracy	83/451	0.1840
$\Delta_{\mathrm{CF}} =$ matched $-$ mismatched	—	+0.0200
95% paired interval	—	$[-0.0089,\ +0.0488]$

I Comparison with directly related work

The comparison below is grounded in the inspected arXiv versions rather than inferred from titles. It distinguishes reported objectives and measurements; it is not a head-to-head evaluation or an exhaustive priority claim.

Work and source	Intervention and reported measurements	Relation to StateSwap
CMI / Causal-LoCoMo (Srivastava, 2026) §§3–5, 8	Same-query no-memory, with-memory and perturbed-memory calls; utility/stability used for selection; answer scores and selected-memory metrics.	Shared interest in external-memory interventions. CMI selects memories; our controlled blocks diagnose a stored-policy intervention. R5/R6 directly calibrate evidence replacement and do not evaluate a selector.
llmmas-otel (Seyedghorban et al., 2026) §§II–III	Baseline and faulty executions of the same task; aligned traces, optional input/message hashes; five repetitions per condition on 30 tasks; runtime amplification.	Pairing, intermediate traces and repeated runs overlap in purpose. StateSwap measures answer disagreement and correctness alongside allocation/evidence changes, rather than runtime amplification.
LLM-MAS chaos engineering (Owotogbe, 2025) §§III–V	Research programme for modelling faults, running controlled robustness experiments and auditing systems; first results concern literature and tools, followed by an evaluation plan.	Relevant motivation for fault-based diagnosis. It should not be presented as a completed empirical comparator for our answer-level estimands.

Table I1: Source-grounded comparison of objectives and measurements. Statements are limited to the inspected versions, not to all capabilities of the associated code or later publications.

CMI's annotation and scoring scope. CMI explicitly uses annotated roles as a risk signal (§3.2) and a deterministic scorer involving the expected answer within its intervention loop (§§3.3, 5.2). Its final metric combines deterministic and judge scores; it is not exact-match accuracy. Its §8 acknowledges that the study does not establish label-free causal-role inference. StateSwap likewise does not infer autonomous selection ability from its annotation-assisted calibration. These differences in estimands, data and scoring preclude comparing the reported numerical performance as if it were measured on a common benchmark.

What the comparison does not establish. Table I1 documents positive statements about the inspected works. A repeated task run is not automatically a byte-identical reader-input control, but the reviewed descriptions do not justify claiming that other systems cannot implement one. Likewise, masking details not established by the inspected sections are left unspecified. The contribution claimed here is a concrete measurement protocol and its empirical diagnostics, not proof that no prior work combines similar ideas.

Source versions. CMI: arXiv:2605.17641v1; llmmas-otel: arXiv:2608.24271v1; chaos engineering: arXiv:2505.03096, inspected HTML version. Sources were retrieved on 25 September 2026.

J EXPLORATORY EVIDENCE-SWAP PROBE (SCALED QUESTION SET)

Status. This appendix reports an *exploratory* probe of the same construction family as the calibration in Section 5.5. Its evidence class is *exploratory* only. It is not a confirmatory test, not a pre-registered study, and not an independent replication, and none of its estimates are pooled with S34, S35, S38, S39R1, R5 or R6. It is included because it isolates one property that the calibration's aggregated estimates cannot show: a paired mean change of exactly zero can coexist with substantial and exactly opposite-signed observed paired correctness changes.

Design. The population is 1000 HotpotQA distractor-validation questions, selected by a fixed hash order from the same pinned adapter used by R5/R6 (171 of the 1000 questions also appear in the R5/R6 population, so the populations are not disjoint). Each question is realised under three four-paragraph reader-visible conditions that exchange exactly one paragraph: B keeps both officially annotated supporting paragraphs and two distractors; I exchanges one unannotated distractor; K exchanges one annotated supporting paragraph. The exchange partner is matched to the replaced paragraph in frozen reference length and is rendered at the same position, so I and K are equal-edit in the same restricted sense as Section 5.5. The reader contract is the frozen R5 contract and is identical across the three conditions apart from the rendered evidence; the reader is not given the arm, the annotations, the allocation or the gold answer. The question text and system message are byte-identical across conditions. The message template and four-block format are fixed; paragraph content changes according to the intervention.

Records and calls. The probe archives 3000 answered cells, i.e. one adopted answer per (question, condition). Of these, 120 are byte-identical reused responses from an earlier 40-question probe whose questions form the exact prefix of this set, and 2880 correspond to HTTP calls issued for this run. The archive holds 3010 HTTP attempts (the 3000 cells plus 10 identical-input repeats carried over from the earlier probe), with 0 retries, no truncation and no parse failure. The statistical unit is the question.

Condition	EM	ΔEM (pp)	95% interval (pp)	disagree
B: support intact	0.620	—	—	—
I: unannotated exchange	0.620	+0.00	[-1.50, +1.50]	0.095
K: support exchange	0.262	-35.80	[-38.90, -32.60]	0.587

Table J1: Exploratory probe at 1000 questions. Intervals are paired percentile bootstraps over questions (20000 draws, pre-specified seed), as defined for this probe; they are not cluster-aware. "Disagreement" is the canonical-answer disagreement rate against B. The annotated-support exchange decreases accuracy by 35.8 pp relative to the unannotated exchange.

Results. Both intervention contrasts, I versus B and K versus B, change the allocation and rendered evidence in every question by construction, so $p_A = p_R = 1$ here is a property of the construction and not an empirical finding. Under the unannotated exchange the paired mean change is exactly zero (+0.00 pp), but this is not inactivity: 58 of 1000 questions show observed paired correctness changes, 29 from incorrect to correct and 29 from correct to incorrect, and 0.095 of the canonical answers differ. These opposite-signed observed paired changes cancel in the mean while the canonical answer moves in roughly one question in ten. Under the annotated-support exchange the same construction decreases accuracy by 35.8 pp with 382 observed paired correctness changes, again with the disagreement concentrated in abstention rather than in wrong content answers.

Sensitivity. Questions in this family can share paragraphs. Grouping questions into connected components that share any rendered paragraph text or title yields 971 components, the largest containing three questions. That grouping is one specific graph over shared material and does not exhaust all possible forms of inter-question dependence. As a *post-hoc* sensitivity only, resampling those components instead of individual questions gives $[-1.50, +1.50]$ pp for the unannotated exchange and $[-38.99, -32.63]$ pp for the annotated-support exchange; the intervals reported in Table J1 remain the primary ones for this probe.

Limits. (i) The probe is exploratory and locally sealed only; the seal covers a 40-question, 130-call design, and the sample was enlarged to 1000 questions after the smaller probe's outcome had been observed, so the enlargement is outcome-informed and the enlarged analysis is not covered by that seal. (ii) 171 questions are shared with the R5/R6 population and the reader family is the same DeepSeek leg, so this is neither an independent sample nor an independent replication. (iii) "Irrelevant" means only that the exchanged paragraph carries no official supporting-fact annotation; it does not establish that the paragraph is irrelevant, and removing an annotated paragraph does not establish that all answer-bearing evidence was removed. (iv) Two further constraint deviations are recorded, not corrected: the frozen report for this probe stated "0/1000 discordant" in one place, which the transitions above contradict, and the generator that wrote that sentence was not updated; the audit record and the correction statement accompany the supplementary archive. (v) Neither condition is a state intervention: this probe manipulates paragraph selection directly and carries no claim about memory appraisal. Reported numbers are re-derived from the raw archived responses by the supplementary scoring script.

K Outcome-informed exploratory synthetic test of a persisted allocation policy

Status and provenance. This appendix reports an *outcome-informed, exploratory, synthetic* test. Its evidence class is exploratory; it is not confirmatory, not an independent replication, and none of its estimates are pooled with S34, S35, S38, S39R1, R5, R6 or the R7 probe. The design was informed by previously observed outcomes, and it imposes an *application-level* persisted allocation policy; it does not test learned appraisal, model-internal state, or any naturally occurring memory benefit, and it does not measure consciousness, feelings or personhood.

Design. 1600 latent worlds (40 strata: five domains $\times$ two depths $\times$ two display orders $\times$ two gold positions) are each realised in two source-reliability environments and under two persisted policies, giving four cells per world and 6800 logical calls per reader (13600 for both readers). In E1 internal source 1 is reliable (0.85) and source 2 is not (0.50); in E2 the reliabilities are reversed (0.50 / 0.85). Each environment pairs a persisted policy that allocates the eight-slot budget as $S1 = (6, 2)$ with one that allocates $S2 = (2, 6)$; the two policies are exact mirrors and are derived from the frozen allocator formula with stored parameter objects. Per (world, source, field) a single uniform draw is shared across environments, so the environments are two materialisations of one draw object and are not independent samples. Within a world the task, gold, repository, field permutation and rendering program are identical across policies: only the loaded policy object changes. Reader payloads hide state, arm, environment, reliability parameters, the oracle and execution identifiers. The two readers are fixed before the first scientific response: *Qwen primary* (`qwen-plus-2025-12-01`) and *DeepSeek prespecified secondary* (`deepseek-flash`); neither leg's outcome determines whether the other runs.

Endpoint and analysis. Y is exact correctness under the frozen strict parser, with capped, invalid and technical cells scored zero under the full-N operational endpoint. For each base world w we use $d1_w = Y(\mathrm{E1}, S1) - Y(\mathrm{E1}, S2)$, $d2_w = Y(\mathrm{E2}, S1) - Y(\mathrm{E2}, S2)$ and the scaled interaction $z_w = (d1_w - d2_w)/2$. Inference resamples *base worlds* jointly across all four cells within the 40 fixed strata (20,000 draws, percentile intervals, seed `R9-bootstrap-20260926-v1`); the unscaled interaction $2\,\mathrm{mean}(z)$ is reported separately.

Results as measured. Cell accuracies (E1S1, E1S2, E2S1, E2S2) are 0.5900, 0.5650, 0.5594 and 0.5844 for Qwen and 0.6406, 0.6200, 0.6344 and 0.6219 for DeepSeek; Table K1 reports the contrasts. The Qwen primary interaction is +2.5000 pp with 95% interval [+0.8750, +4.156] pp, i.e. positive with an interval excluding zero: the reader's operational correctness contrast is larger when the persisted policy favours the reliable source. The DeepSeek secondary interaction is +0.4062 pp [-0.906, +1.7500] pp, an interval that includes zero. The two component contrasts for Qwen are $d1 = +2.5000$ pp and $d2 = -2.5000$ pp, and *both* component intervals touch or cross zero (-0.3750 to +5.3750 and -5.2500 to +0.1875). The prospectively fixed **conjunctive sign-reversal** criterion therefore **fails**: the result is a symmetric one-sided pattern for the primary reader, not a demonstrated reversal, and it is reported as observed rather than reinterpreted. Canonical-answer

Quantity	Qwen (primary)	DeepSeek (secondary)
$d1$ (pp)	+2.5000 [-0.3750, +5.3750]	+2.0625 [-1.000, +5.188]
$d2$ (pp)	-2.5000 [-5.2500, +0.1875]	+1.2500 [-1.875, +4.312]
primary $\mathrm{mean}(z)$ (pp)	+2.5000 [+0.8750, +4.156]	+0.4062 [-0.906, +1.7500]
unscaled $2\,\mathrm{mean}(z)$ (pp)	+5.0000 [+1.7500, +8.312]	—

Table K1: Outcome-informed exploratory synthetic test. Intervals are paired percentile bootstraps over base worlds within the 40 fixed strata; the designated primary reader is Qwen. The reader-minus-reader difference in $\mathrm{mean}(z)$ is +2.0938 pp [+0.125, +4.062] on the same resampled worlds and is descriptive only: the two readers are fixed serving setups on the same materials, not independent samples.

disagreement among usable pairs is 0.438 (E1) and 0.415 (E2) for Qwen and 0.448 / 0.463 for DeepSeek, so answer disagreement is by itself not evidence of a decision-side effect.

Denominators, technical status and control variation. All 6800 calls per reader returned HTTP 200 on a single attempt with an explicit terminal, no retry and no transport failure; both readers cover all four cells for all 1600 worlds, and the full-N denominators are used throughout. Qwen has 2 strict parse failures (markdown-fenced JSON, counted zero, no world dropped) and DeepSeek 0. The identical-input control submits one arm-independent four-four render twice per environment with local state tags; its structural gate passes for all 200 pairs per reader, which is a statement about tags, byte-identical reader payloads and the pre-clamp allocation audit — *not* about deterministic answers. From the archived completions, identical inputs still yield different canonical answers in 31 (Qwen) and 28 (DeepSeek) of 200 pairs, and differ in correctness in 26 and 24 pairs respectively. These rates are descriptive, are never subtracted from any estimate, and they bound how individual changes may be interpreted.

Disclosures that must travel with these numbers. (i) The independent reanalysis used the archived *completions*; the full raw provider HTTP response bodies were not retained, so recomputation is conditional on the executor's archived extraction. (ii) A metadata correction is appended to the run record: the archived `outgoing_request_sha256` is the SHA256 of the canonical key-sorted request object, not of the transmitted wire bytes; the source-payload and schedule bindings are unaffected. (iii) The persisted policy is an imposed application-level object; this test does not establish learned appraisal, a general benefit of persistent state, an explanation of S35's uncertainty, or any model-family difference. (iv) Post-hoc per-domain breakdowns are sensitivity only and are not part of the frozen claim. (v) S35's held-out primary result and its interval are unchanged by this appendix.

L Fresh-sample validation of the persisted allocation policy (R10)

Status and lineage. This appendix reports a *prospectively frozen fresh-sample validation* of the persisted-policy effect for which Appendix K supplies the outcome-informed exploratory result. The design was fixed *after* the exploratory R9 outcome had been observed, so it is **not a public preregistration**, **not an independent replication**, **not a new task distribution**, and **not** evidence of learned appraisal, model-internal state or any naturally occurring memory benefit. It is not pooled with S34, S35, R5, R6, the R7 probe or R9, and it does not change S35's held-out identity or its unconfirmed outcome. The procedure used to select the sample size is recorded explicitly: it was chosen from the independent R9 audit standard deviations ($sd_z = .3383$, $sd_{d1} = .5902$, $sd_{d2} = .5586$), which roughly halves the variance of the R9 means at $N = 3200$; this is a planning approximation and not a promise of replication.

Frozen contract. 3200 new latent worlds (40 strata: five domains $\times$ two depths $\times$ two display orders $\times$ two gold positions, 80 worlds per stratum) were generated by the *same* task, environment, repository, state, allocator and rendering algorithms as R9, under a fresh deterministic R10 seed fixed in the freeze manifest before any R10 reader call. Excluded by canonical latent identity against every

earlier comparison set, the overlap is exactly zero (S34 PoC 400, S34 extension 600, S35 2,000, R9 1,600). Each world is realised in two source-reliability environments and under the two persisted mirror policies $S1 = (6, 2)$ and $S2 = (2, 6)$, giving four cells per world and 13200 logical calls per reader (26400 for both readers). Prompts, task rules, reader-visible masking, gold scoring, the strict parser and the reader parameters are byte-identical to R9 except for the new worlds' content. The readers are fixed before the first scientific response: *Qwen primary* (`qwen-plus-2025-12-01`) and *DeepSeek prespecified secondary* (`deepseek-flash`); each leg was run independently of the other leg's outcome, there was no pilot, interim effect check, optional extension, model substitution or response-conditioned stopping, and exactly one fixed N was used. The local reservation ceiling was CNY 160 per reader at CNY 0.01 per attempt; the funded account was checked at start and at least every 250 attempts, with a pause before any dispatch below CNY 8.

Endpoint and analysis. Y is exact correctness under the frozen strict parser, with capped, invalid and technical cells scored zero under the full-N operational endpoint. For each base world w we use $d1_w = Y(\mathrm{E1}, S1) - Y(\mathrm{E1}, S2)$, $d2_w = Y(\mathrm{E2}, S1) - Y(\mathrm{E2}, S2)$ and the scaled interaction $z_w = (d1_w - d2_w)/2$. Inference resamples *base worlds* jointly across all four cells (and, for the descriptive cross-reader contrast, both readers) within the 40 fixed strata, with 20,000 draws, percentile intervals and the explicitly passed seed `R10-bootstrap-20260926-v1`; the unscaled interaction $2\,\mathrm{mean}(z)$ is reported separately. The primary criterion is a two-sided 95% interval for $\mathrm{mean}(z)$ whose lower endpoint is above zero; the conjunctive reversal criterion additionally requires the $d1$ interval to lie entirely above zero *and* the $d2$ interval entirely below zero.

Results. Table 3 (main text) reports the contrasts. Cell accuracies (E1S1, E1S2, E2S1, E2S2) are 0.6062, 0.5737, 0.5806 and 0.6056 for Qwen and 0.6334, 0.6116, 0.6266 and 0.6250 for DeepSeek. The Qwen primary interaction is +2.88 pp with 95% interval [+1.70, +4.08] pp, and the component contrasts are $d1 = +3.25$ pp [+1.22, +5.28] and $d2 = -2.50$ pp [-4.56, -0.47]; both intervals exclude zero in the prespecified directions, so the primary criterion and the conjunctive sign-reversal criterion are met on new worlds. The DeepSeek secondary interaction is +1.02 pp [+0.02, +2.02] pp (lower endpoint +0.015625 pp) and *both* of its component intervals include zero, so the secondary leg does **not** reach the reversal; it is reported as observed and never promoted. The descriptive reader-minus-reader difference in $\mathrm{mean}(z)$ is +1.86 pp [+0.42, +3.31] on the same resampled worlds; the two readers are fixed serving setups, not independent samples.

Denominators, technical status and control variation. All 13200 calls per reader returned HTTP 200 on a single attempt with an explicit terminal, no retry and no transport failure, and both readers cover all four cells for all 3200 worlds; the scientific usable denominators are 13198/13200 (Qwen) and 13200/13200 (DeepSeek). Qwen has 2 strict parse failures (a markdown-fenced `json` block instead of a bare JSON object, one in `shipment_routing-D3-a-B-000331` cell E2S2 and one in `pipeline_window-D3-a-B-000393` cell E1S2); both are counted zero under the full-N endpoint, no world was dropped, the parser was not changed and neither response was retried. DeepSeek has 0. Canonical-answer disagreement among usable pairs is 0.435/0.428 for Qwen and 0.466/0.456 for DeepSeek. The identical-input control submits one arm-independent render twice per environment with local state tags; its structural gate passes, and the archived completions have identical canonical answers in 89/87 (Qwen) and 89/84 (DeepSeek) of 100 pairs per environment, with correctness differing in 9/12 and 11/15 pairs. These rates are descriptive, are never subtracted from any estimate, and caution against attributing individual answer changes solely to the intervention.

Provenance of the archived objects. For every attempt the run archived the full raw provider response body together with the frozen source-payload hash, the canonical request-object hash and the hash and byte count of the actual outgoing wire bytes; the served model and `finish_reason` were re-checked against the pinned identifiers, and the run's budget ledger, per-call reservations and stop-file/funding controls are recorded in the run artifacts. Unlike R9, the R10 recomputation is therefore not conditional on an executor extraction: the independent audit re-parsed the archived raw bodies with an independent strict parser and reproduced every point estimate and interval endpoint.

Disclosed implementation deviations (offline only). The material generator reuses the audited R9 path with only: a fresh namespace/seed; 80 rather than 40 worlds per stratum; the addition of

the R9 latent population to the historical exclusion set; raw-body and wire-byte-hash archiving; and a CNY 160 per-reader reservation ceiling. At freeze time the engineering probe of the same-day R9 qualification was inherited (same endpoints, pinned served models and byte-identical request parameters) rather than re-run; the first R10 live call was the staged funding balance check. In the offline acceptance fixtures, the contract-test subset was made to contain at least one control world and the executor-acceptance expectations were scaled to 13200 calls; neither change touches a scientific condition, threshold, statistical rule, model parameter or seed. No R10 outcome informed a design change; the initial R10 design was informed by R9.

Bounded claim. Within this synthetic task family and the pinned Qwen serving setup, an imposed application-level stored policy produced a reproducible reliability-dependent decision-side interaction, with the prespecified sign reversal, on a fresh block of worlds. This does not establish a broad state benefit, learned appraisal, natural-task efficacy, model-internal state, or a general impossibility theorem, and the prespecified DeepSeek secondary did not reach the reversal.