CREDIT WITHOUT GROUND TRUTH: AUDITING STEP-LEVEL CREDIT ASSIGNMENT IN LLM AGENTS AGAINST EXECUTED REPLAY

Heady Zhang
University of Southern California
Los Angeles, CA, USA
haiyuez@usc.edu

ABSTRACT

Audited against causal ground truth from executed replay in a single-agent tool environment (ALFWorld), none of the step-level credit signals used to train LLM agents — LLM-judge scores, outcome-conditioned logprob ratios, or the policy's own confidence — identifies which steps causally matter better than chance. Existing evaluations grade these signals against annotated step *correctness*; we audit them against step *contribution* — what re-sampling the policy's own alternatives at each decision point and rolling forward actually changes about the outcome — and the two come apart. The ground truth itself is structured: causal contribution is sparse (30.5% of decision points where ground truth is defined carry measurable effect), and measurability is model-dependent — the fraction of points with no policy-supported counterfactual differs by a factor of two (13.1% vs. 26.8%) between two similar-scale policies. The failure mode is identifiable: implicit credit echoes the policy's fluency (median rank correlation $+0.75$, replicating at $+0.70$ in a second family under a corrected instrument), while conditioning on the outcome adds no causal information (partial correlation -0.004, Qwen). A confidence-only router recovers pivotal steps at chance level, but cuts judge cost by 13.1% per turn (14.0% per trajectory). In a seven-arm pre-registered training experiment, no arm reliably outperforms the untrained policy, and the checkpoints' apparent

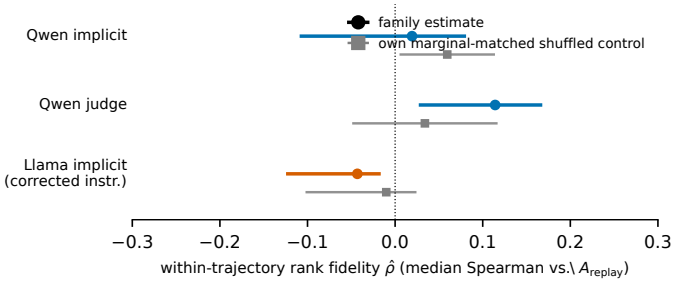

Figure 1: **The audit verdict: in both model families, credit is indistinguishable from its own shuffled control.** Each family's rank-fidelity estimate against executed-replay ground truth (median within-trajectory Spearman vs. A_{replay}, Section 2; filled markers) beside its *own* marginal-matched shuffled control (open markers, paired within family; the figure plots rank fidelity only — the precision-at-pivotal lift discussed in Section 4 is a different statistic and is not plotted): Qwen2.5-7B implicit 0.0193 $[-0.109, 0.081]$ (n_{eff}=37 of 50; own control $[0.005, 0.114]$) and judge 0.1142 $[0.027, 0.168]$ (n=37, 32 computable; own control $[-0.049, 0.117]$), data `runs/collect_v2/analysis_summary.json`; Llama-3.1-8B implicit under the corrected instrument -0.043 $[-0.125, -0.016]$ (n=21 trajectories retained of 28; own control $[-0.102, +0.024]$), data `runs/xfam_ext/redo/redo_family_analysis.json`, coverage 88.3%[1]. Under the frozen controls-first verdict order, every family–control pair overlaps under the controls-first order: verdict H3 (placebo-level, frozen verdict order, Section 4) throughout; the binding comparison is each family against its own control.

instrument signature is fully explained by training dose — sparser credit retains fewer examples, an order-of-magnitude spread in optimizer steps — not credit content. Comparisons of credit rules must therefore match effective sample size, or they measure dose, not credit.

1 INTRODUCTION

Agentic reinforcement learning is converging on a bet: train LLM agents through long tool-use episodes by scoring each step, not just the outcome. Step-level credit signals — LLM judges scoring turns, outcome-conditioned log-probability ratios, the policy's own confidence — are moving from evaluation harnesses into training loops (Tan et al., 2025; Li et al., 2026b; Tan et al., 2026; Chen et al., 2026; Zhang et al., 2026), and surveys catalogue dozens of methods built on them (Zhang, 2026). The bet rests on an assumption that, to our knowledge, no one has tested: that the credit these signals assign to a step tracks what the step actually contributed to the outcome. C3 audits multi-agent credit against its own replay advantages (Chen et al., 2026); grading signals already in LLM-agent training use against an external, executed causal ground truth is a different object (Appendix L).

Existing evaluations measure something else: step-level benchmarks grade credit signals against *annotated step correctness* (Zhang et al., 2025; Liu et al., 2026). But a correct step can contribute nothing (the trajectory was already determined) and an incorrect one can be pivotal (it opened the state from which recovery happened); correctness and contribution are different quantities, and only the second is what a training loop pays for. Auditing the signals the field actually trains on, at the level they operate requires ground truth for contribution itself.

We build that ground truth by executed replay. At each decision point of a collected trajectory, we re-sample alternative actions the policy itself supports, roll each forward to completion, and measure the shift in the outcome distribution — a per-step causal effect with a per-estimate noise floor and an explicit resolution bound. Nothing in the construction consults the credit signals under audit.

The ground truth itself is the first surprise: causal contribution is sparse — under a third of decision points carry any measurable effect — and *measurability is model-dependent*: the no-counterfactual fraction differs by a factor of two between two similar-scale policies. Against it, the audit's null survives every probe: neither credit family ranks steps better than its own marginal-matched shuffled control, in either model family, under either instrument generation. The mechanism is identifiable: implicit credit largely echoes the policy's own fluency (median rank correlation $+0.75$, replicated across families), and conditioning on the outcome adds no causal information. In a seven-arm pre-registered training experiment, apparent differences between credit rules are fully accounted for by training *dose* — sparser credit buys fewer optimizer steps — not by what the credit says.

Our contributions, by type:

1. **Measurement.** A measurability map of causal ground truth in a replayable single-agent tool environment: contribution is sparse, and measurability itself is model-dependent in both directions — steps are unmeasurable either because the policy supports no alternative or because the outcome is already absorbed.
2. **Audit.** In a replayable single-agent tool environment, no off-the-shelf signal — judge scores, implicit outcome-conditioned logprob ratios, outcome conditioning itself, or the policy's own confidence — identifies which steps causally matter better than chance.
3. **Mechanism.** The implicit family's scores are a fluency echo, replicated across two model families on three pre-registered supports; the outcome-conditioned increment carries no causal information, with the primary evidence on the Qwen policy.
4. **Decision rule (cost-only).** A frozen confidence-routing rule that cuts judge calls by 13.1% per turn and 14.0% per trajectory — the two granularities move oppositely and are always reported together — with chance-level pivotal recall, and prospective deployment of the router is not evaluated.
5. **Protocol.** A named toolset for credit comparisons: dose matching, measured perturbation strength for controls, an MDE ladder for reading training nulls, and a four-dimension integrity taxonomy with its incident catalogue.

Our method-level contributions are the decision rule and the protocols, not an end-to-end system.

2 MEASURING CAUSAL GROUND TRUTH BY EXECUTED REPLAY

Our instrument is executed replay: at each decision point of a collected trajectory we re-execute the environment under sampled alternative actions and measure how the outcome distribution shifts. We instantiate it in ALFWorld, a replayable single-agent tool environment, with Qwen2.5-7B-Instruct as the policy: 50 trajectories on a task list frozen before collection, sampled at temperature 0.7 under the HCAPO method's published ALFWorld agent template (Tan et al., 2026), extracted verbatim from the paper's appendix. Environment determinism was verified before collection and re-verified on every host that touched the data (one trajectory hash, four independent environments). A concurrent formalization of the same instrument validates it under planted effects (Shah, 2026); the contrast is in Appendix L.

The audited signals are the two families of step-level credit that agentic RL pipelines train on. The *implicit* family is HCAPO's outcome-conditioned token log-probability ratio ρ_t, computed exactly as published; its scorer is the policy checkpoint itself, by HCAPO's design, and the instrument constants, the hindsight-injection construction, and the matched *policy* scoring mode are detailed in Appendix G.

Ground truth at a turn is a contrast between two outcome distributions, never a single continuation. At every action turn t we re-execute the factual action at least three times and sample $K{=}4$ distinct admissible alternatives from the same policy snapshot at the collection temperature (within at most 300 seeded draws), rolling each alternative to terminal at least three times; the replay advantage is $A_{\text{replay}}(t) = \text{mean}(\text{outcome} \mid \text{factual replays at } t) - \text{mean}(\text{outcome} \mid \text{alternative rollouts at } t)$. The realized continuation of the original trajectory is never used as the factual estimate. When four distinct admissible alternatives cannot be sampled from the policy, the policy-supported counterfactual is *undefined* at that turn; such turns are excluded and counted rather than imputed. Under Qwen2.5-7B this leaves 1,768 complete turns of 2,034 intervened; the excluded fraction is itself a finding (Section 3). Prefix-restore determinism held across the full sweep: zero divergences over its 20,538 replay rollouts.

The same factual replicas give the estimator its noise floor and its resolution. Per turn, $\sigma_{\text{floor}}(t)$ is the standard deviation of the outcome across factual replays, and a turn is *pivotal* exactly when $A_{\text{replay}}(t) \neq 0$ under the literal stored value — no significance filter, no floor threshold. Because outcomes are discrete and replicas few, exact zeros are common: 69.5% of complete turns carry $A_{\text{replay}} = 0$, and for the 1,184 turns where both arms are all-zero the data exclude only $|\Delta p| > 0.632$ at one-sided 95% confidence. Every zero in this paper is therefore a resolution-bounded statement — "indistinguishable from the factual action at the achieved sampling resolution" — and never a claim of no causal effect.

Fidelity is scored where credit is consumed: within trajectories. For each family we compute the within-trajectory Spearman correlation between the credit values and A_{replay}, aggregated as the median across trajectories with a 10,000-resample bootstrap interval; trajectories on which the statistic is degenerate (fewer than four complete turns, or constant A_{replay}) are excluded from this computation only, symmetrically across families, and counted. Random, uniform, and within-trajectory shuffled credit — the shuffle preserves each trajectory's marginal exactly — run through the identical pipeline in the same batch. The verdict order is frozen, and its controls gate comes first: a family whose interval overlaps that of its own shuffled control is placebo-level regardless of its point estimate. Three structural bias tests accompany the verdicts under a Holm correction over all six family-level hypotheses; one of them, T3, carries a registered directional prediction — positive for the implicit family — that fluent, high-probability actions receive inflated credit regardless of causal effect.

Thresholds, exclusion rules, and the verdict order were frozen and signed before collection under immutable tags; every number here is regenerated by script from raw artifacts, never transcribed; checkpoints are content-verified shard-by-shard before any stage loads them (Section 8). The audit is repeated in a second model family, Llama-3.1-8B-Instruct. For the implicit family the swap changes the measured object, the scoring instrument, and the ruler at once — HCAPO's scorer *is* the policy, and replay truth is defined over the policy's own action distribution — so the cross-family run tests the system-level claim, that a policy's own implicit credit does not track its own causal ground truth, under a second system; it cannot and does not evaluate the scoring instrument apart from the policy. The Llama arm collected 28 trajectories, of which 27 enter the analysis set after one exclusion. Its replay layer was re-executed in full under a corrected instrument after a chat-template defect was

found and quantified (Section 8); the re-execution's coverage is stated once here[1] and inherited by every Llama-side quantity in the paper.

3 THE STRUCTURE OF CAUSAL GROUND TRUTH

3.1 THE MEASURABILITY MAP

Before asking whether credit tracks causal contribution, we ask where causal contribution can be measured at all — and the answer is that the ground truth is sparse and mostly silent. Under Qwen2.5-7B, 30.5% of complete turns are pivotal (n=1,768): fewer than a third of decision points have any measurable causal effect on the outcome at all. The dynamics are strongly absorbing — $\sigma_{\text{floor}} = 0$ at 86.0% of complete turns — and every turn with A_{replay} exactly zero carries the resolution bound of Section 2: the data exclude only $|\Delta p| > 0.632$ there. Most of the trajectory is causally inert; a credit signal earns its keep only by finding the minority of turns that are not.

3.2 MEASURABILITY IS MODEL-DEPENDENT — IN BOTH DIRECTIONS

Under the same environment, the same K=4 alternatives, and the same 15-rollout budget, the policy-supported counterfactual is undefined at 13.1% of intervened turns for Qwen2.5-7B (n=2,034) but at 26.8% for Llama-3.1-8B (n=1,082; corrected instrument[1]) — a factor of 2.05, with non-overlapping intervals. And where ground truth *is* defined, the same comparison reverses: the family that is harder to measure carries more causal signal, 38.3% of Llama's complete turns pivotal against Qwen's 30.5%. The full three-rate map is Figure 5 in Appendix J; the two divergences run in opposite directions. The Wilson 95% intervals are $[11.7, 14.6]\%$ (Qwen) and $[24.2, 29.5]\%$ (Llama). The mechanism is concentration, not capability: at some decision points Llama's probability mass is too tight for four distinct admissible alternatives to be sampled within budget. This is a scope condition on replay methodology itself.

The pivotal reversal's precision: 38.3% of Llama's complete turns (n=792; corrected instrument[1]) against Qwen's 30.5% (n=1,768). The absorbing structure, by contrast, transfers approximately — $\sigma_{\text{floor}} = 0$ at 80.6% versus 86.0% of turns. Measurability structure therefore varies with the model in both directions: the family that is harder to define counterfactuals for is the one with more causal signal where counterfactuals exist. It is a property of the (policy, environment) pair — not an environment constant, and not a statement that either family is uniformly harder to audit. Concurrent work reports a similar sparsity pattern (Shen et al., 2025); ours differs in kind: zeros are noise-floored and resolution-bounded, and measurability itself is bidirectionally model-dependent (Appendix L). Figure 4 in Appendix J shows both divergences.

The exclusions are not random, and the direction of their bias is measured rather than assumed. The turns without a policy-supported counterfactual are systematically the low-entropy ones, so exclusion correlates with the implicit family's own predictor: for Qwen, the included-minus-excluded difference in mean policy log-probability is -0.70 (n=1,768), the interval excluding zero. We report this selection check for Qwen only. For Llama, what the corrected instrument re-establishes is the exclusion *rate* just stated; the selection-check *correlation* was computed on pre-correction inputs and cannot be recomputed — a permanent state, disclosed in Appendix B. The analysis sets are therefore right-truncated on the fluency axis, with the direction of the truncation measured for Qwen, and Section 4 states what this range restriction does to the fidelity estimates.

An occupancy-style predictor built from the policy's own action statistics ranks turns correctly while missing the aggregate level — the ranking is trustworthy, the level is not (Appendix J).

4 THE FIDELITY AUDIT: IMPLICIT CREDIT AGAINST ITS OWN GROUND TRUTH

The audit question is whether the credit a trained scorer assigns to each step tracks the causal contribution that executed replay measures for that same step — and the answer, everywhere the audit

[1] Coverage 88.3% (1,082/1,225 parseable turns). Missingness is concentrated exclusively in long ($\geq$44-turn) trajectories (13/28 partially covered; short/medium buckets 100%). The 143 turns outside the analysis set — 142 never generated at the cap trip + 1 unparseable — were not re-run, by pre-registered precedent ($\sim$1% expected conversion). Length-sensitive quantities inherit this caveat (gate disclosure: Section 8).

Table 1: **Both credit families are placebo-level at within-trajectory credit ranking under the frozen verdict order (H3): each is indistinguishable from its own marginal-matched shuffled control.** The implicit estimate is centred on zero; the judge estimate is weak, possibly nonzero, yet indistinguishable from its own control — the families are reported separately and earn the same verdict class. Estimator: median within-trajectory Spearman vs. A_{replay}, 10k bootstrap (estimator note: Appendix A); controls: marginal-matched within-trajectory shuffles; instruments: original (Qwen), corrected (Llama; coverage 88.3%[1]); data: `runs/collect_v2/analysis_summary.json`, `runs/xfam_ext/redo/redo_family_analysis.json`. Sign agreement: per-step vs trajectory-median splits, Wilson 95%; implicit $267/527 = 50.7\%$ $[46.4, 54.9]$, judge $84/139 = 60.4\%$ $[52.1, 68.2]$; n_{eff} = trajectories retained by the frozen exclusion rules.

Family	$\hat{\rho}$ (median) [95% CI]	n_{eff}	Shuffle [95% CI]
Implicit (HCAPO ρ_t)	0.0193 $[-0.109, 0.081]$	37 of 50	$[0.005, 0.114]$
Judge (Qwen2.5-72B)	0.1142 $[0.027, 0.168]$	37 (32 comp.)	$[-0.049, 0.117]$

reaches, is that it does not. Coverage differs by signal, and we scope it per signal: implicit credit is audited in both model families and under both instrument generations; judge fidelity in one family (Qwen); the confidence signal in one family (Qwen).

Three measurements that share no estimator, no test statistic, and no failure mode — rank fidelity against a marginal-matched shuffled control, per-step sign agreement against chance, and partial correlation with the causal increment after conditioning out fluency — agree on the same null: the first two in both families, the third on the registered Qwen set (corrected-instrument partials: Appendix D).

For the Qwen policy under the original instrument, the family-level rank fidelity of implicit credit is read under the frozen verdict order, and the controls gate fires before any effect gate is reached: the verdict is H3 (Table 1; Figure 1). Per-step sign agreement sits at chance. The audit's negative result is not that credit anti-tracks its ground truth; it is that nothing distinguishes the trained scorer's ranking from the same scores with their step-assignment destroyed.

One face of the judge signal does clear chance, and we state it plainly before locating its limit: per-step sign agreement for the judge family is $84/139 = 60.4\%$ (Wilson 95% $[52.1, 68.2]$; `runs/collect_v2/analysis_summary.json`), an interval excluding 50%. It does not amount to identifying which steps causally matter — a rank-and-concentration claim — and the two families fail it differently (registered exploratory, per family, n=35 trajectories each, of 50 loaded; `runs/collect_v2/r12_precision_at_pivotal.json`): the implicit family's precision-at-pivotal lift is 0.940 $[0.760, 0.997]$, an interval lying entirely below the chance line of 1.0, while the judge's 1.000 $[0.935, 1.001]$ contains it. The judge's per-step sign signal buys no concentration on the turns that mattered; the implicit family's ranking is, if anything, mildly anti-concentrated. Agreement with the anchor judge is reported in Appendix J.

The Llama replication was run twice: once under the original instrument, and in full again under the corrected instrument after the chat-template defect of Section 8 was found. Under the corrected instrument the verdict class is unchanged: the family estimate and its own marginal-matched shuffled control overlap, under the frozen order the controls gate fires first, and the verdict is H3, matching Qwen's (Figure 1, which carries both intervals). Descriptively, the family point estimate is negative with a confidence interval excluding zero (-0.043, $[-0.125, -0.016]$; n=21 trajectories retained of 28[1]), and it is indistinguishable from its own marginal-matched shuffle (-0.010, $[-0.102, +0.024]$) — we note the sign and draw no claim from it, since the gate that binds is the control comparison, not the zero crossing. Sign agreement between implicit credit and replay is 0.449 ($[0.393, 0.505]$; n=301 turn pairs[1]) — chance-level, the third measure returning the same answer as the first two.

That the null itself replicates is established by the pre-registered transfer criterion. T3 replicates across model families under the corrected instrument: on the registered primary (full) set, the cross-family median ($+0.7008$ $[0.6482, 0.7683]$, n=27, coverage 88.3%[1]) satisfies the pre-registered point criterion — it lies within KS1's registered CI $[0.647, 0.793]$. The corrected interval's lower edge, 0.6482, clears the registered band's exact lower bound (0.6474) by 0.0008 — the cross-family replication holds on both the point criterion and the full interval, with essentially zero margin, which we state proactively. The second, independent criterion also holds: re-run under a script frozen before execution, the shuffled control for T3 is -0.0155 ($[-0.1290, +0.0844]$, n=27), non-overlapping with

the observed interval, so the control criterion fires; we note its shuffling procedure executes in a remote frozen script and its control value is read from the archived artifact rather than recomputed locally. And the full-set family difference, re-run by the recovered original bootstrap on the corrected inputs, spans zero for 40/40 seeds (seed-11 primary -0.0188 $[-0.1178, +0.0954]$; n_{xfam}=27 vs n_{KS1}=47)[1] — the families do not measurably differ on the statistic whose replication the criterion certifies.

The two instrument generations bracket the result. The defect that forced the re-run moved the measurability map materially (Section 8); it did not move a single verdict class. Fidelity is H3 against the shuffled control under the broken template and under the fixed one, at pilot power and at n=21; sign agreement is chance-level in both; the transfer criterion holds in both. A result that survives its own instrument being repaired is the closest an audit of this kind comes to an internal replication.

Two readings are excluded by construction. The null is not an artifact of unmeasurable ground truth — every fidelity quantity above is computed only over turns where replay truth is defined (Section 3). And it is not a statement that the scorer's outputs are uninformative: they are strongly structured, just not by causal contribution — which raises the question Section 5 answers, of what they are structured by.

5 MECHANISM: CREDIT ECHOES FLUENCY, NOT EFFECT

If the credit does not track causal effect, what does it track? We registered the answer before the data, and it was confirmed: implicit credit rises with the policy's own probability of the action it is scoring — credit, in this family, is largely an echo of fluency. The structural test against policy log-probability is the implicit family's one pre-registered positive prediction, and it survived multiplicity correction (median rank correlation $+0.752$, CI $[0.647, 0.793]$, n=47 trajectories; Holm-adjusted p=0.0002).

The decomposition makes the echo quantitative. Regressing implicit credit jointly on the action's fluency (its mean policy log-probability) and on the causal increment replay measures, fluency carries roughly two and a third times the weight of the increment for the Qwen policy (standardised 0.955 against 0.402; n=1,768), and once fluency is conditioned out, the partial correlation between credit and the causal increment is -0.004 (p=0.87). Nothing of the causal signal survives the conditioning; the increment's apparent weight in the joint model is the shadow fluency casts on it.

One trajectory shows the mechanism whole. In trajectory `seed017`, the policy issues the same action — *take cloth 1 from toilet 1* — four consecutive times, the environment answering the first attempt with *Nothing happens.*; at turns 12, 16, and 20 the scorer's confidence in these repetitions exceeds 0.9999. Replay judges the same turns pivotal, with causal increments of $+0.333$, -0.167, and $+0.250$. A policy stuck in a loop is maximally confident, because the next token is maximally predictable — and a credit signal that echoes that confidence will spend its certainty on exactly the steps where the trajectory has stopped going anywhere.

Nor does the conditioning on outcome rescue the signal. The scorer's hindsight increment — what seeing the outcome adds to the policy's own log-probability — is uncorrelated with the causal increment on the registered Qwen set, with the corrected-instrument Llama estimators disagreeing (Appendix D): knowing how the trajectory ended changes the score, but not in the direction of what each step contributed.

The cross-family run is the check on this mechanism, and its headline is stability of the dominance structure rather than any single coefficient: the fluency-dominance ratio is template- and family-robust (KS1 2.37; xfam old 2.29 $\rightarrow$ corrected 2.35). Whatever the template regime and whichever the family, the same ordinary-least-squares (OLS) regression places roughly two and a third times the weight on fluency that it places on the causal increment[1].

The corrected-instrument partial correlations — both estimators, their disagreement, and the selection caveat together — are in Appendix D. This section rests on the Qwen analysis, where the conditioning null is measured on the registered set; the Llama arm's role is the one its ratio plays — showing that fluency dominance, the structure the mechanism needs, survives a family swap and an instrument repair.

A scorer that pays out for predictability is not thereby harmless: whether the echo damages the policy it trains is a separate, empirical question, and Section 6 takes it up with the scorer placed inside a live training loop.

6 THE TRAINING LAYER: CREDIT SIGNALS INSIDE A LIVE TRAINING LOOP

The audit's natural objection is that fidelity to replay might not matter if the credit still trains a better policy. We closed the loop: seven training arms — outcome-only, implicit credit, judge credit, their shuffled and inverted controls, and a reduced-resolution replay-truth arm — trained under common random numbers (CRN) and evaluated on 128 held-out tasks; the full seven-arm table is Table 3 in the appendix. No arm reliably beats the untrained base policy (0.422; $54/128$), and every one of the six pre-registered confirmatory comparisons came back inconclusive under the frozen ± 3 pp equivalence band with Holm correction; none of the three verdict lines fired. We report this as inconclusive, not as "no difference": implicit credit trained 2.3 pp below its own shuffled control (Holm $p{=}0.43$) and 7.0 pp below outcome-only training ($p{=}0.91$), and neither gap is resolvable at this power. What the experiment rules out is any effect large enough to survive its own design; what it cannot do is certify equivalence.

The power shortfall is itself a pre-registration finding: the frozen proxy inverted, leaving the minimum detectable effect at ≈ 11.8 pp against the registered ± 3 pp band (Section 7's reading rule); the full account is a pre-registration-instrument failure story, Appendix A.

One arm-seed carries an instrument note rather than a result — a format collapse, not a competence collapse; the note, with its integrity lesson, is in Appendix G.

Yet the checkpoints are not interchangeable. On held-out states the seven arms' policies agree in their discrete choices — measured gaps $+0.41$ pp ($p{=}0.343$, Stage 1) and $+0.50$ pp ($p{=}0.185$, Stage 2) against the frozen A3B bar of ≥ 5 pp at $p < 0.05$ — while their full action *distributions* separate cleanly: between-arm Jensen–Shannon divergence exceeds within-arm by $2.6\times$ (0.0520 against 0.0198; restricted-permutation $p{=}0.0001$) — credit moved where the policies put their probability, without (yet) moving what they do.

The frozen record's own summary of this experiment is one sentence, and we quote it rather than paraphrase: *"Credit source leaves a measurable signature in the policy's output distribution — arms cluster by instrument family (rho-based vs judge-based), not by information content — yet this signature does not reach greedy action selection: checkpoints agree on actions at the same rate whether or not they share a credit rule."* Two caveats travel with the sentence wherever it appears. The masked arm moved far less than any other (adapter norm 1.37 against ≈ 4.7 for the dense arms), so its position in any clustering is confounded with how little it moved — the divergence result survives removing it entirely (0.0212 within-arm against 0.0421 between-arm; $n{=}16$ and $n{=}80$ checkpoint pairs). And the Stage-2 canonicalizer initially normalized raw generations without first extracting the action, producing spuriously unique strings; the error is retained on the record, and every number here is from the corrected pass. Two further checkpoints with elevated canonicalization-failure counts were flagged per the pre-registration rather than silently averaged in, and the format-collapsed arm-seed (Appendix G) was excluded from all primary analysis by pre-registration.

The phrase *not by information content* is licensed by a pre-registered partition comparison, fixed before the checkpoint-pair matrix was computed: the instrument partition separates the checkpoints in both stages (500-state bank: $+0.0346$, $p{=}0.0002$; $n_W{=}30/n_B{=}141$ cross-arm pairs), the information partition in neither (-0.0126, $p{=}0.9777$; $36/135$); the full two-stage decomposition, with its estimator and turn-1 notes, is in Appendix H. What the weights remember, then, is which instrument scored them — and the dose analysis below shows that even this signature is fully accounted for by how many optimizer steps each instrument's credit sparsity bought.

The instrument signature has a mundane explanation, and the explanation survives a positive control. Credit rules that zero out many turns drop those turns from the training batch — one filter line in the update step — so sparser credit buys fewer optimizer steps and smaller parameter displacement. Controlling for realised parameter-change magnitude removes the instrument correspondence entirely (partial Mantel $\rho{=}+0.078$ on realised credit scale, $p{=}0.774$), against the positive control's zero-order association — credit sparsity — which the same test certifies at $+0.912$ ($p{=}0.0001$; $n{=}21$ arm pairs, non-independent, both tests). The mediation evidence carries its own strength label: the pair set is small and non-independent, the association is robust to removing the sparsest arm; the design remains a partial-correlation design — no $\|\Delta W\|$-matched comparison exists — and the mediation claim is stated at that strength. The matched comparison was permanently cancelled by ruling, for three recorded reasons in the ruled order — science first, resources last: the partial

correlation already answers the question; the design has an inherent flaw ($\|\Delta W\|$ overlap across arms exists only across rounds, so caliper matching necessarily introduces a round confound, and round covaries with training volume); and cost against a calendar in which writing had not started. The ruling ships verbatim in Appendix K.5. What the weights remember about their credit instrument is, on this evidence, how much training its sparsity allowed — dose, not doctrine.

The natural rescue for the null — that the substrate, not the credit, was at fault — was given its own experiment and returned a negative of its own. With the three diagnosed substrate defects repaired, outcome-only training diverged in every configuration tried: at the original hyperparameters, at reduced learning rate ($5\text{e}{-5} \rightarrow 2\text{e}{-5}$), and at five-fold tighter gradient clipping ($1.0 \rightarrow 0.2$), each run's pre-clip gradient norm escaped one to several batches before KL breached its guardrail, under per-step norm clipping throughout. Divergence insensitive to two orthogonal stabiliser knobs, originating below the clip, indicates accumulated directional drift rather than oversized single steps: the stabiliser family and the corrected advantage signal are structurally mismatched. The method line was closed by the frozen criterion — no configuration reached evaluation — and the trainability of the corrected substrate is not established. The null of the training study (tag KS2) therefore cannot be laundered into a substrate complaint; the complaint was tested and failed to train at all.

Finally, the practitioner's shortcut fails first as a detector and only then earns its keep as a cost rule. A confidence-only router — frozen in advance as a replay-free decision rule, its threshold an order statistic on the policy's own confidence, its inputs free at inference time — routes low-confidence turns to the judge and recovers 11.9% of pivotal turns (Wilson 95% $[9.4, 14.9]$; n=540), against the 13.1% of all turns it routes: chance-level recall, the mechanism being exactly Section 5's — the stuck, pivotal steps are the high-confidence ones, so a confidence threshold is aimed away from them (exploratory, pre-registered as such). The two 13.1% figures coincide by definition: the routing threshold was set so the routed fraction equals the no-counterfactual rate by construction (`ROUTE_NUMBERS.md:8--9`). The rule's safety has two layers, only one by construction: on high-confidence steps the policy-gradient signal is inert ($\nabla \log \pi \approx 0$); the absorbing-step half is an audit finding — σ_{floor} is replay-derived, invisible to the router ex ante. As a cost mechanism the same threshold is well-behaved: it cuts judge calls by 13.1% at per-turn and 14.0% at per-trajectory granularity, and neither may be read alone. We release the rule with its negative result attached — a rule that travels without its evaluation will be used for the thing it cannot do (sweep and cost account: Appendix J, Figure 3). Entropy separates critical from non-critical states at the distribution level (Shen et al., 2025: Cliff's $\delta = 0.42$); that separation and chance-level recall at a cost-matched threshold are not in conflict (Appendix L).

7 What to do differently tomorrow

The audit's negative results convert into named, practitioner-facing artifacts, each of which exists because one of our own comparisons would have misled us without it: **dose matching before any credit comparison** (any update step that drops zero-weight examples lets the credit rule silently set the training dose — our arms' optimizer steps ranged from 112 to 8 per round under identical budgets); **measured perturbation strength for every control** (a control named "shuffle" is not automatically an equal-strength manipulation, and unmeasured it invalidated our own first family dissociation); **a routing rule that knows what it is for** (a cost mechanism, not a detector — full treatment in Section 6); and **reading a training-loop null** (every equivalence claim travels with the registered margin, ± 3 pp, *and* the design's minimum detectable effect, ≈ 11.8 pp at 128 held-out tasks, single seed — a null reported with its margin but not its detection floor invites "no effect" where the design can only say "no effect this large"). The reporting tables, the one-sentence reading rules, the frozen rule text, and the full MDE ladder are in Appendices C and F.

8 Evidence decay and an integrity taxonomy

Auditing credit signals produced a second record we did not plan for: a catalogue of integrity failures in our own pipeline, written down as they happened. Sorting them yields four distinct questions an integrity check can answer — whether the object checked is the object the system loads, whether the bytes were correct when they were written, whether every item is still readable now, and whether the

check preserves the evidence needed to act on it — and one recurring mistake. *No single check covers all four dimensions; the failure mode throughout is assuming one check answers another's question.*

The catalogue then made a prediction, and its own next failure confirmed it: we recorded, before the next audit ran, that the next incident should land on one of the two never-tested defences — and it did (the pack-vs-repository check; the full arc, with the second defence still an open prediction, is in Appendix A, beside the closure rule that caught its sibling failure).

The most consequential failure ran the whole arc: a metadata inconsistency, noticed by a person rather than by a check, led to a static trace, an A/B quantification, and a full re-replay confirming the move in the predicted direction (magnitudes: 12.0 pp predicted, `REDO_REPORT.md:19`; 12.3 pp measured, `REDO_REPORT.md:19`). All verdict classes were unchanged under the corrected instrument; the map rates moved materially. The lesson is narrower than *verify your inputs*. Every replica of the affected artefact was byte-identical and every hash matched, because the defect was semantic: two instruments that are the same file are not thereby the same instrument.

The same re-replay met a pre-registered gate and failed it. Our protocol required at least 20 of 27 trajectories to be complete before the analysis could run; the run finished with 15 of 28 complete, and the analysis was withheld. It was released afterwards at the adjudication layer rather than by the process that had stopped — in two separately recorded steps, on grounds quantified in Appendix A, and on the condition that every quantity derived from the run carry the partial-coverage label it carries throughout this paper. The stop and the two releases are separately recorded, in that order, in the project's decision record, whose entry also records that the binding threshold was a task-brief invention whose trajectory-level wording conflicted with the frozen pipeline's turn-level inclusion rule — the gate that fired was itself an unvalidated instrument. We report the miss because a threshold that is reinterpreted when it binds is not a threshold, and because the label alone does not tell a reader that a gate was crossed to earn it.

9 Limitations, and what would change our mind

Every scope condition below is one we measured rather than assumed: the audit runs in a single environment with a binary outcome, and every claim is scoped accordingly. The training-loop experiments are single-family and single-scale (Qwen2.5-7B, LoRA (low-rank adaptation), offline); the cross-family arm formally replicates the structural result under the corrected instrument, not the training loop. KS2 lacks a validated positive control, which is why its null is reported as inconclusive; three convergent checks — an expert-cloning arm at -10.9 pp on held-out evaluation (exact McNemar p=0.0488, n=128 tasks, 29/15 discordant), measured behavioural displacement, and off-policy fit — bound how badly the harness could be lying, without substituting for the control. The cross-family analysis set excludes 26.8% of intervened turns (n=1,082; corrected instrument[1]). The truncation's direction is measured for Qwen only: excluded turns are high-probability (included-minus-excluded mean policy log-probability -0.70, n=1,768), which attenuates the correlation there; the Llama counterpart cannot be recomputed (Appendix B), so cross-family conservativity is qualitative — and the families truncate unequally (Qwen excludes 13.1%, n=2,034). The fidelity estimates run at small effective n, with wide intervals stated wherever they appear. Every zero-ground-truth statement carries its sampling resolution bound. One mid-run batch-composition correction (skip undefined turns, keep trajectories) is disclosed with its byte-level regression verification. And the family-by-length interaction our design intended to test is unmeasurable in this data — the realized length distribution put 42 of 50 trajectories in the long bucket (Qwen; the two shorter buckets kept 2 and 3 after amendments) — so that motivating axis is dead here and inherited by future work.

Our audit is confined to a single environment. We note, however, that ALFWorld is not an arbitrary choice: it is a primary evaluation venue for the credit methods we audit — the implicit-credit instrument family we test reports its state-of-the-art results on this very benchmark.[2] Our audit therefore meets these methods on their own evaluative ground; the burden of showing that step-level credit tracks causal contribution *elsewhere* falls on settings where such credit has not yet demonstrated success either.

[2]The implicit-credit instrument family we adapt (Tan et al., 2026) and the closest step-credit method (Zhang et al., 2026) report their headline results on ALFWorld. StepOPSD evaluates at 1.7B/3B model scale against our 7B/8B policies; any comparison carries that scale gap. It is cited as an evaluation venue only.

REPRODUCIBILITY STATEMENT

Every number in this paper is generated, not transcribed: a single script regenerates the complete ledger from raw artifacts, two independent regenerations must agree byte-for-byte before any release, and each ledger row carries its value, its n, and its source path. Where a row's provenance is composite, the chain is demonstrated rather than asserted — perturbing each source artifact separately and showing each column follows its own source. All thresholds, exclusion rules, verdict orders, and analysis plans were frozen under signed, hash-pinned pre-registration tags before the data they govern existed, and the frozen files ship verbatim in the supplement, under the verified environment pins (alfworld 0.4.2, textworld 1.7.0, numpy 2.4.6, scipy 1.17.1, pyyaml 6.0.3). We release, through an anonymized repository: the replay archive (29,402 files), per-turn and per-trajectory summaries, credit and log-probability dumps, trained adapters, evaluation outputs, the generator, and the ledger itself. Two quantities are released as-is with their deaths on record rather than recomputed: the common-support medians, whose generating procedure was never archived and whose inputs predate the instrument correction — re-deriving them would require new code we did not authorise. One recomputation is released with its recovery story: the full-set difference bootstrap, whose original script was recovered, committed as received, shown to reproduce the original run byte-identically, and only then run on corrected inputs. The integrity incidents of Section 8, including the ones our own checks missed, are catalogued in the appendix with their detection channel and their fix; we regard that catalogue as part of the reproducibility surface, since a number that cannot survive its own pipeline's history is not reproducible in any sense that matters.

AI USE STATEMENT

AI systems contributed substantively to this project: operating experimental pipelines under pre-registered protocols, drafting prose, auditing drafts for consistency against frozen claim wording, and executing the generation of the number ledger from archived artifacts. Authority remained with the human authors under pre-registered governance: every experimental adjudication, threshold freeze, stopping criterion, and claim-strength decision was made by human ruling, and every number in the main text traces to archived, hash-verified artifacts; the central ledger regenerates byte-identically from a single script. Appendix B retains superseded values precisely because their artifacts are dead — they are disclosed, not recomputed. The LLM judge and the policy models audited in this paper are its research subjects, and are distinct from any AI assistance used in its preparation; no experimental data, measurement, or verdict was produced by writing assistance. Errors arising from AI assistance during this project are themselves documented and taxonomized in Section 8.

ETHICS STATEMENT

This work audits the validity of machine-generated training signals against executed environment replay in a synthetic household-task environment (ALFWorld). It involves no human subjects, no personal data, and no sensitive content; the audited models are publicly released open-weight checkpoints used under their licenses. We report negative results about widely used credit signals so that training practice is not misdirected by unvalidated per-step scores; we see no ethical risks specific to this study beyond those general to research on agent training.

REFERENCES

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative LLM agents. *arXiv preprint arXiv:2603.06859*, 2026.

Mukai Li, Qingcheng Zeng, Tianqing Fang, Zhenwen Liang, Linfeng Song, Qi Liu, Haitao Mi, and Dong Yu. Verified critical step optimization for LLM agents. *arXiv preprint arXiv:2602.03412*, 2026a.

Zhongyi Li, Wan Tian, Jinju Chen, Huiming Zhang, Yang Liu, Yikun Ban, and Fuzhen Zhuang. Counterfactual credit policy optimization for multi-agent collaboration. *arXiv preprint arXiv:2603.21563*, 2026b.

Jiale Liu, Huajun Xi, Shaokun Zhang, Yifan Zeng, Tianwei Yue, Chi Wang, Jian Kang, Qingyun Wu, and Huazheng Wang. Who&when pro: Can LLMs really attribute failures in AI agents? *arXiv preprint arXiv:2607.09996*, 2026.

Jaineet Shah. Causal agent replay: Counterfactual attribution for LLM-agent failures. *arXiv preprint arXiv:2606.08275*, 2026.

Leyang Shen, Yang Zhang, Chun Kai Ling, Xiaoyan Zhao, and Tat-Seng Chua. CARL: Criticality-aware agentic reinforcement learning. *arXiv preprint arXiv:2512.04949*, 2025. v3, 2026-05.

Hui-Ze Tan, Xiao-Wen Yang, Hao Chen, Jie-Jing Shao, Yi Wen, Yuteng Shen, Weihong Luo, Xiku Du, Lan-Zhe Guo, and Yu-Feng Li. Hindsight credit assignment for long-horizon LLM agents. *arXiv preprint arXiv:2603.08754*, 2026.

Weiting Tan, Xinghua Qu, Ming Tu, Meng Ge, Andy T. Liu, Philipp Koehn, and Lu Lu. Process-supervised reinforcement learning for interactive multimodal tool-use agents. *arXiv preprint arXiv:2509.14480*, 2025.

Zhang et al. Who&when: Multi-agent failure attribution. *arXiv preprint arXiv:2505.00212*, 2025. ICML 2025 Spotlight.

Chenchen Zhang. From reasoning to agentic: Credit assignment in reinforcement learning for large language models. *arXiv preprint arXiv:2604.09459*, 2026.

Yanfei Zhang, Xu Lin, and Chenglin Wu. StepOPSD: Step-aware online preference distillation for agent reinforcement learning. *arXiv preprint arXiv:2605.27140*, 2026.

A INTEGRITY: THE FOUR DIMENSIONS, THE INCIDENT CHAIN, AND EVIDENCE DECAY

This appendix expands Section 8. The taxonomy, verbatim from the governance record:

Table 2: The four dimensions of the integrity taxonomy, each with the incident that named it and the defence adopted. *No single check covers all four dimensions; the failure mode throughout is assuming one check answers another's question.*

Dimension	Question	Failure instance	Defence
Identity	is the object checked the object the system *uses*?	the gate verified a different physical copy than the stages loaded — hashing the wrong file passes perfectly	realpath binding: a verification claim names the physical path the stage will load
Creation-time validity	were the bytes correct when *written*?	a NUL-filled file's hash is a valid hash of a NUL-filled file; every replica verifies	structural assertions against *intent*, not against the writer's own output
Persistence	is every item still readable *now*?	a file-count check passes while one file is unreadable	per-file content hashing — you cannot hash what you cannot read
Diagnosability	does the check preserve the evidence needed to *act*?	the guard reported a verdict and discarded the errno; EIO and EDQUOT call for opposite responses	log the raw errno, never the verdict

Semantic instrument identity (extension clause). The chat-template defect of Section 8 adds an extension to the Identity row: tokenizer and template belong to the *instrument's* identity, and no manifest, hash, or closure rule can see a mismatch of this kind — two instruments that are the same file are not thereby the same instrument.

The prediction arc, in full. Two of the four defences had never met the failure they exist to catch; we recorded, before the next audit ran, that the next incident should land on one of the two. It did: the release pack had been verified repeatedly — every time against its own manifest rather than against the repository it was cut from, a check of internal consistency, not of being the thing we claim to ship. The prediction predates the confirming diff, and the second untested defence remains an open prediction, stated as one rather than retired.

The cleanest instance. A generator met a configuration failure and recorded it with our marker for a number that cannot be traced — the same marker a genuinely untraceable number receives. The verdict survived; the reason it was reached did not, and the two conditions, which call for opposite responses, were indistinguishable by the time anyone read the output.

Evidence decay (the Persistence instance, retention inverted). An inventory check passes at 29,402 files discovered against 29,402 expected while one file is unreadable [Tier 3, `ALL_EXPERIMENTS_RECORD.md:2037`]; the archive verification run records 29,401 match / 0 mismatch / 0 missing / 1 named unreadable [Tier 3, `ALL_EXPERIMENTS_RECORD.md:2206-2207`]. The derived summary survives, so the number stays traceable while its re-derivation is lost; the verified tarball is accordingly the *primary* copy of the replay evidence and the volume secondary.

The withheld-then-released analysis: coverage, cost, and the pillar corrected by its own run. The gate paragraph of Section 8 has its pillars quantified here: this run covered $1{,}082/1{,}225 = 88.3\%$ of parseable turns [Tier 3, `REDO_STATUS.md:20`]; the original XFAM-v1 analysis had itself proceeded at $1{,}129/1{,}225 = 92.1\%$ [Tier 3, `REDO_STATUS.md:41`]. The second pillar as worded at the stop — the record's claim that the frozen v3.2 rule "would retain all 28" trajectories [Tier 3, `REDO_STATUS.md:38--41`] — was corrected by the executed analysis, which retained 26 of 28 [Tier 3, `REDO_REPORT.md:27`]; the record's claim overstates by two. The release stands either way, since 26 clears the 20-trajectory threshold under both readings, and the divergence is recorded

in the decision record rather than smoothed here. There were two releases, not one, on the same day and in sequence: the salvage ruling released the *analysis*; the under-review annotations on the affected claim text were released only later, by a separate close-out item, after its own condition passed. The ruling-mandated missingness disclosure is likewise reported with its structure visible: per-bucket coverage is B1 $11/11 = 1.000$ $[0.741, 1.000]$, B2 $17/17 = 1.000$ $[0.816, 1.000]$, B3 0.881 $[0.861, 0.898]$, overall 0.883, with all 142 missing and 1 unparseable turns falling exclusively in long-bucket (B3) trajectories of length ≥ 44 [Tier 3, `REDO_REPORT.md:55--70`]. The stated flag criterion — a bucket is marked insufficient-coverage when its Wilson 95% CI excludes the overall rate — fires for no bucket; that is not evidence of balanced coverage, because B3 dominates the universe and its CI contains the overall rate by construction, so the flag is structurally incapable of firing for the only bucket with missing data. The missing turns sit at the cap-trip end of long trajectories and are plausibly harder than average; the B3-level numbers inherit that caveat. Context of the stop: generation halted at the 13 GPU-hour cap trip with 1,083 summaries written, 13.17 GPU-hours metered against a $40 cap [Tier 3, `REDO_STATUS.md:14,16`].

Two 2026-08-14 additions. (i) The packaging-boundary defect of the main text is filed to the Retention family; mechanism: *closure-rule blind spot — future citation*. (ii) A comparator defect, self-reported by the execution side: a column-diff tool split rows on every pipe, including one escaped inside a field, and reported a correct ledger as FAILED. The ledger was right; the checker was wrong. This is a live instance of this section's own meta-requirement — the checks are themselves unvalidated instruments — and the better instance precisely because the error ran in the harmless direction: a comparator that calls a correct ledger FAIL will, another day, call a wrong one PASS.

Fail-closed, worked example. A task required re-running a frozen procedure that did not exist and never had; the session stopped and escalated rather than reconstructing the specification from prose [Tier 3, `E051_RERUN_BLOCKED.md:1--6`]. It pairs with the comparator instance above: one is a check that failed safe, one is a task that refused to improvise.

A frozen power proxy, inverted. The KS2 adaptive rule computed power from the outcome-only arm's between-seed spread, treated as an upper bound on the paired spread; realised paired spreads ran 7.6 to 35 times that proxy — the anchor arm happened to be the most stable one, so the bound was inverted, not conservative, and the rule reported full power for a study underpowered at the registered band. It was executed verbatim and escalated rather than adapted mid-run: a power calculation frozen before the data is only as conservative as its proxy, and ours taught us which proxy not to freeze.

Reference closure is not need closure. One failure is about the checks rather than the data. Our closure rule verifies that every artefact cited by a governance document is in the release pack; it did not catch a directory left out entirely, because nothing cited it yet — the citation arrived a batch later, with the analysis that needed it. The rule guarantees reference closure, not need closure. We added no check, because need closure is not mechanically decidable and the defence that worked was already in place: the session that could not find the artefacts stopped and escalated instead of substituting a neighbouring one.

A length check against a float-saturated flow. A length check anchored to a document endpoint assumes the flow is incompressible: that removing content moves the anchor. The assumption fails in float-saturated regions — where a page range's area is jointly allocated to floats and text, text removals re-balance around the committed floats and release no page fraction, so the endpoint reads zero movement across full edit rounds while a word-level diff certifies the removals as real. The check's unit (endpoint position) and the edit's unit (words) live in different layers, and the discrepancy was resolved only by a downstream deletion probe, which localized the responsive region. The instance is typographic; the pattern is not: a checker whose invariant lives below the layer being edited will silently report stasis.

Defence status. Each dimension names a defence, and each defence is reported with its own validation status rather than a blanket claim (statuses as frozen at the taxonomy's 2026-08-09 ruling):

dimension	defence	known-positive	status
creation-time validity	NUL count	e38/e40/e41	validated
persistence	per-file content hashing	e49	validated
identity	realpath binding	never	holds by design, not by test
diagnosability	log the raw errno	never	holds by design, not by test

The taxonomy was derived from incidents, so its validation is likewise incident-supplied. Two of its four defences have met a known-positive only because a real failure happened to supply one; the other two have never been tested against the failure they are designed to catch, and hold by design rather than by demonstration. We report the taxonomy as a way of reasoning about integrity checks, not as four validated instruments — claiming otherwise would be the same category of error this appendix catalogues. No synthetic known-positives were constructed for the two untested rows: writing is the critical path, and the limitation stated plainly is worth more than a hurried test. We recorded that the next incident should land on identity or diagnosability, the two defences that had never met a known-positive. Recorded before the diff was run, the prediction was confirmed by e58 (the pack-vs-repository check), which supplied identity's known-positive; diagnosability remains outstanding as a live prediction. The table is shown as frozen at its ruling date, with e58's confirmation carried by the arc rather than by a silent status edit — an append-only record updates by appending.

A trim that outlived its premise. One typographic entry belongs with the rest: a negative vertical space inserted between a table caption and its body during an early compression round became a defect once the caption later grew — the table's top rule came to overlap the caption's final line. The lesson generalises beyond typography: a compensation tuned to one state of an artefact is not annotated with the state it assumed, so it silently becomes wrong when the artefact changes under it.

A note on identifiers. PC3-series incident ids carry the PC3- prefix; incident identifiers (lowercase e52–e58) and experiment identifiers (uppercase E052–E058) are different series that collide numerically and are never resolved into one another — renumbering an append-only record would itself be an integrity edit of the class this appendix catalogues. (The prefix rule was an oral ruling first written to the record on 2026-08-19, after an exhaustive search — 189 governance documents, every local project root, and all unpacked archives — returned no earlier written instance; the record gains the rule by ruling, dated, rather than by a reconstructed memory of it.)

The week the taxonomy was tested. Five incidents in three days exercised every dimension of the catalogue, and the chain is reproduced verbatim in Appendix K.5; what belongs here is what each one did to the record. e52 (persistence): three heartbeat lines NUL-zeroed in place, length unchanged — the file was sealed unrepaired, and append-type logs joined the periodic read-back that had previously covered only newly created governance files. e53 (a gate doing its job): the pre-registered MDE gate hard-stopped a training diagnostic at zero GPU when the required n exceeded the frozen cap and the 5 pp target proved structurally unreachable on the intended held-out split; a carve-out that would have let the run proceed was computed, labelled not-adopted because its motivation post-dated seeing the failing number, and escalated — the amendment that resolved it re-based the held-out table and left the threshold untouched. e54 (diagnosability): a filesystem stall under concurrent small-file writes was diagnosed at the syscall level before any restart, and an intermediate misreading — a dead process taken for a finished one because a fresh file existed — is retained with its lesson: an exit code and a census, not the presence of output, decide success. e55 (identity, the e34/e46 shape again): a dependency was declared missing after being checked in an interpreter the pipeline never uses; the false report nearly caused a substrate change, and the discipline it minted — existence checks run in the interpreter that will do the work, interpreter path recorded with the conclusion — is now standing. e56 (a scope breach, recorded rather than excused): a sealed training script was modified in place instead of copied; the compensations were an opt-in flag preserving every existing invocation, a bit-exact regression of the default path against committed records, and the explicit note that "unchanged since seal" is no longer assertable for that file. e57 (a nominal dose that was not the delivered dose): a control arm's effective optimizer-facing sample was roughly half its nominal collection because degenerate rollout groups carry zero advantage; the step target was still met, so no gate fired, but the gap is recorded as part of the arm's interpretation rather than as a post-hoc caveat. Alongside the incident series, the record keeps a channel for assertions that entered through

conversation rather than from a verified source — four relayed-assertion instances and one unsourced reasoning error as of this submission's record close, counted and classed separately because no lookup discipline could have caught the latter — each corrected by an appended entry, never by deletion.

Estimator note (Table 1). The shuffled control's own interval may exclude zero without carrying signal: under marginal matching, the median of within-trajectory rank correlations over small turn counts is not constrained to centre on zero — which is why the verdict gate is the family-versus-its-own-control comparison, not either interval's position relative to zero.

B SUPERSEDED AND DISCLOSED VALUES

E051 full-set difference, as published. The originally published value $+0.0432$ $[-0.0903, +0.2044]$ has no artifact, no script, and therefore no recorded procedure and no recorded seed; it is disclosed here and not printed in the main text. The main-text disclosed re-computation is the seeded one ($+0.0319$ $[-0.0903, +0.2000]$, script in repo, seed 11) — a re-runnable computation of the same quantity, not a reproduction of the original run — whose script was later shown to regenerate its artifact byte-identically, and whose corrected-instrument successor (-0.0188 $[-0.1178, +0.0954]$) is what Section 4 reports. That the two historical intervals share a lower bound suggests procedural similarity but does not constitute proof, and no same-procedure comparison between them is computable.

Common-support medians. The common-support medians ($+0.7673$ cross-family, $+0.6757$ KS1) are released as-is with their deaths on record: the generating procedure was never archived and cannot be rebuilt, and the cross-family side's inputs predate the instrument correction. No re-derivation was authorised; no claim rests on them.

The Llama selection check. The included-minus-excluded log-probability check of Section 3 exists for Qwen only: the Llama counterpart would need the corrected-instrument analysis set, whose membership was not separably recorded, so the check does not exist for Llama and cannot be reconstructed.

Pre-redo instrument values. All pre-redo Llama-side quantities (the pre-redo T3 values, the pre-redo measurability rates, and their derivatives) are retained in the ledger under explicit SUPERSEDED / NOT-citable labels and appear nowhere in this paper's claims, except where a pre-correction value is quoted, explicitly labeled, solely to demonstrate robustness across the instrument fix.

E057. The contaminated-era family-residual observation is void and is recorded here in one line only.

C THE PROTOCOL, IN FULL

The dose-matching table. One reporting table per credit comparison: surviving examples, optimizer steps, tokens updated, realised parameter displacement (free for LoRA adapters via the trace identity, no materialisation), and the measured strength of every control — with the behavioural measure in the last column. The reading rule is one sentence: if the behavioural column is monotone in the dose columns, the comparison has measured dose, not credit. In our own arms, optimizer steps ranged from 112 to 8 per round under identical budgets.

Measured strength, worked. Within-trajectory shuffling changed 90.4% of positions for our continuous credit but only 26.0% for the tie-heavy discrete judge scores, so comparing the two families against "their shuffles" confounded information destruction with perturbation magnitude. Every control should ship with its fraction-changed, displacement, and rank-correlation against the original; our inversion control's assumed -1 correlation held exactly, but held *because we finally checked*.

Table 3: **Seven-arm training outcome: every pre-registered confirmatory comparison is inconclusive under the frozen ± 3 pp band (C1–C4 TOST equivalence; C5–C6 one-sided superiority; Holm over the six), and no arm reliably beats the untrained base policy (0.422; 54/128).** Top: final-round win rates on the frozen 128 held-out tasks (greedy; n=128 per arm-seed; seed-paired CRN seeds {0,1,2}). † arm 5 seed 2 is the format-collapse instrument note of Appendix G, excluded from primary analysis by pre-registration; the printed arm-5 mean and SD include it and inherit the flag, and the registered-exploratory sensitivity dropping it leaves all six comparisons inconclusive. Bottom: paired comparisons; the design's minimum detectable effect at this n and seed count is ≈11.8 pp (Appendix F), so "inconclusive" is the frozen reading, never "no difference". Data: `runs/ks2/eval/final/<arm>/<seed>/eval.json` (win rates), `runs/ks2/eval/r1/base_control/s0/eval.json` (base), `runs/ks2/CONFIRMATORY_RESULT.json` (comparisons).

arm	s0	s1	s2	mean	between-seed SD
1 outcome-only	0.438	0.445	0.438	0.440	0.5 pp
2 implicit (ρ)	0.352	0.414	0.344	0.370	3.9 pp
3 judge	0.359	0.531	0.445	0.445	8.6 pp
4 shuffled-ρ	0.375	0.383	0.422	0.393	2.5 pp
5 inverted-ρ	0.406	0.391	0.094†	0.297†	17.6 pp†
6 pivotal-masked	0.469	0.344	0.320	0.378	8.0 pp
7 shuffled-judge	0.445	0.375	0.406	0.409	3.5 pp
base policy (no adapter)		—		0.422	(54/128)

id	comparison	mean diff	Holm p	conclusion
C1	implicit vs shuffled	-2.34 pp	0.4273	inconclusive
C2	judge vs shuffled-judge	$+3.65$ pp	0.5326	inconclusive
C3	implicit vs outcome	-7.03 pp	0.9116	inconclusive
C4	judge vs outcome	$+0.52$ pp	0.3265	inconclusive
C5	inverted vs implicit	-7.29 pp	0.2543	not demonstrated
C6	masked vs outcome	-6.25 pp	0.8422	not demonstrated

D CROSS-FAMILY DESCRIPTIVE PARTIALS

The partial correlations under the corrected instrument are reported descriptively, both estimators side by side: after conditioning on fluency, the Pearson partial correlation between implicit credit and the causal increment is -0.086 (p=0.015, n=792; Llama, corrected instrument), while the Spearman partial is -0.019 (p=0.591, n=792)[1]. We draw no confirmatory claim from the nominally significant Pearson value: the recomputation was not registered, the two estimators disagree, and the missing turns are concentrated non-randomly in the longest trajectories (Section 8), so a selection effect cannot be excluded.

E THE SEVEN-ARM RESULTS TABLE

Table 3 reports the full seven-arm outcome that Section 6 summarizes: final-round win rates per arm and seed, and the six pre-registered confirmatory comparisons under the frozen ± 3 pp equivalence band with Holm correction. Every comparison is inconclusive; no verdict line fires. The arm 5 seed 2 evaluation carries the format-collapse instrument note of Appendix G and was excluded from all primary analysis by pre-registration; it is printed here for completeness and flagged wherever a mean or SD includes it.

F THE MDE LADDER

Minimum detectable effects for the paired binary design of Section 6, at the registered ± 3 pp band; every MDE carries its (n, split) label, since the MDE moves with both.

n (split)	discordant pairs	MDE
96 (train, 1 seed)	$\sim$21	$\sim$13.8 pp
128 (held-out, 1 seed)	$\sim$28	$\sim$11.8 pp
384 (3 seeds pooled)	$\sim$84	$\sim$7.0 pp

G INSTRUMENT DETAILS AND NOTES

The implicit instrument, in detail. The ratio ρ_t is computed by Eqs. 6–7 of the HCAPO paper exactly, with the published constants (T_{temp}=5.0, clip $[0.8, 1.2]$). Hindsight conditioning is a single outcome-injection line appended to the prompt — our addition, not HCAPO's; the *policy* scoring mode uses the identical prompt without it, which keeps Section 5's fluency contrast clean. Deviation note: KS1 rollouts run at temperature 0.7, frozen in the pre-registration; HCAPO's published rollout temperature is 1.0, and the deviation is a KS1 design choice, logged.

The judge-family adaptation preserves the TARL judge's prompt structure (Tan et al., 2025), response contract, and $\{1, 0, -1\}$ scale. TARL's published judge receives ground-truth tool-call annotations; ALFWorld has none, so the judge runs without them — a disclosed deviation from TARL's regime, and the condition under which such judges are actually used during training.

One arm carries an instrument note rather than a result. The inverted-credit arm's third seed collapsed to near-zero evaluation not because inverted credit destroyed the policy's competence but because training degraded its output format: 87.7% of that seed's actions embedded the action tag inside the action text, making them inadmissible to the environment, where every other arm-seed sits at or near zero. The format gate passed it, because the gate tests for loops and command-likeness, not tag placement — an integrity lesson (Section 8) arriving through the training layer.

H THE PRE-REGISTERED PARTITION DECOMPOSITION (E031(B))

The phrase *not by information content* is licensed by a comparison that was fixed in advance rather than chosen afterwards. The pre-registration froze two competing partitions of the seven training arms — one grouping them by which instrument produced the credit, one grouping them by whether the credit carried task information — together with the criterion that decides between them, before the checkpoint-pair matrix was computed. The instrument partition won in both stages on both criteria: on the 500-state bank, mean between-partition JS exceeds mean within-partition JS by $+0.0346$ ($p = 0.0002$; n_W=30 within-partition and n_B=141 between-partition cross-arm checkpoint pairs), while the information partition separates the same checkpoints in neither stage (Table 4). What the weights remember is which instrument scored them — Section 6's reading; the decomposition below is its full evidential basis.

Table 4: **The pre-registered instrument partition separates the checkpoints; the information partition does not.** Each row reports mean between-partition minus mean within-partition JS divergence over cross-arm checkpoint pairs, under the partition named in column 1; positive means checkpoints sharing a group are more similar to each other than to the other group. Partitions, criterion, and the reading to record were frozen in `a3b-prereg-v1` before the matrix was seen; each partition's p is a restricted permutation test under the pre-registered within-seed scheme (10,000 draws), the scheme attribution following the pre-registration. Both partitions divide the same 171 cross-arm pairs; the split differs, so the two rows carry different n_W/n_B. Data: `a3b_stage1_structure.json`, `a3b_stage2_structure.json`.

Partition	Stage / statistic	Δ (B$-$W)	p	n_W/n_B
Instrument	Stage 1, turn-1 canonical agreement	$+0.0291$	0.0124	30/141
Information	Stage 1, turn-1 canonical agreement	-0.0077	0.7205	36/135
Instrument	Stage 2, 500-state bank JS	$+0.0346$	0.0002	30/141
Information	Stage 2, 500-state bank JS	-0.0126	0.9777	36/135

I RELEASE INVENTORY

Items released through the anonymized repository (counts from the sealed final-state record):

Item	Count	Size
KS1 `collect_v2` replay archive	29,402 files	4.7 G
per-trajectory `replay_summary.json`	1,026	—
per-turn `summary.json`	17,682	—
credit files `*.credit.json`	6,058	—
ρ files / judge files	2,275 / 1,596	—
`eval.json`	118	—
adapters	87	13.1 G
logprob dumps (`.pol`/`.hind`)	4,601	—
500-state bank (`a3b_state_bank.jsonl`)	1	0.8 M

J SUPPLEMENTARY FIGURES

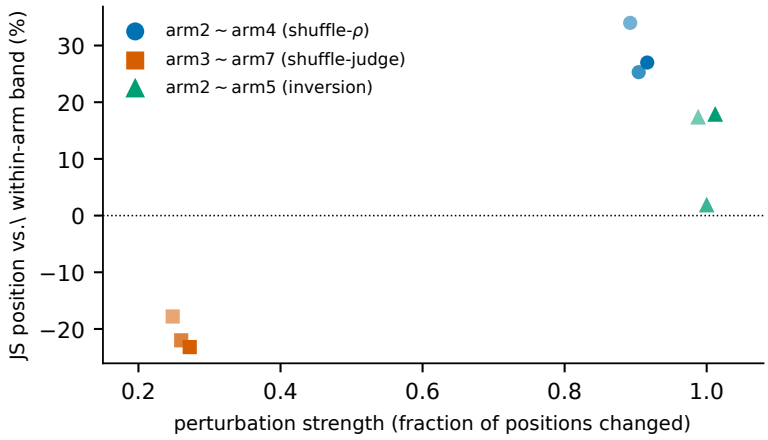

Figure 2: **Perturbation strength does not order the checkpoints' distributional positions: the position readout is not a dose meter for information destruction along the magnitude axis.** Each contrast pair's JS position relative to the within-arm band (%), across training rounds r1–r3 (marker shade), against the measured strength of its perturbation — fraction of credit positions actually changed: shuffle-ρ 0.9043, shuffle-judge 0.2602, inversion 1.0000 (n=288 trajectories each; `runs/pc2/a3b_perturbation_strength.json`). Position values are the nine round-by-contrast cells of `runs/pc2/a3b_position_monotonicity.json`; the pre-registered monotonicity test returns p=0.9413 (10k perm; n=6 adjacent comparisons; grid 3×3, randomization unit = contrast, k=3). Four qualifiers ride this figure, per its record: the strongest perturbations sit in a saturated regime of the readout; separation is monotone in information *destruction*; the non-monotonicity is confined to the magnitude axis; and the judge-shuffle pair carries no positional signal.

The cost account. At the registered threshold, 43 of 50 baseline judge calls remain; the two cost granularities move in opposite directions across the sweep (Figure 3), so neither substitutes for the other. Two disclosures travel with these numbers: the per-trajectory denominator 50 is the frozen value and the flattering one — the analysis set spans 48 trajectories (seed010 and seed020 have no complete turns), and the 48-based figure is 0.1042; and the confidence input is a per-token geometric mean, which runs high for multi-token actions. Route to save money, not to find the steps that matter.

These rates are reported as an observation, not as a predictor, and the observation is a single-family one: for Qwen, an occupancy-style model built from the policy's own action statistics ranks turns correctly — observed completion rises from 0.650 to 0.981 across its predicted deciles — while

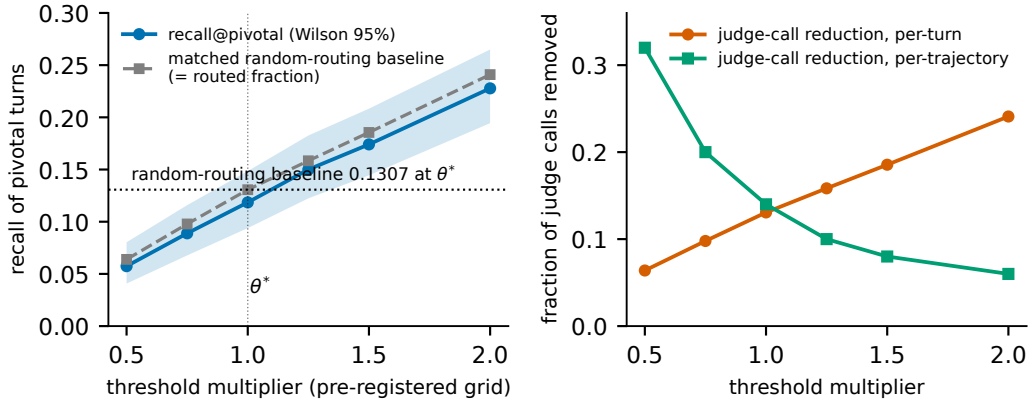

Figure 3: **[Exploratory — R_ROUTE] The confidence router recovers pivotal turns at chance level at every pre-registered threshold — its recall never exceeds the matched random-routing baseline; its two cost granularities move in opposite directions, so neither may be read alone.** Left: recall of pivotal turns (pivotal = stored $A_{\text{replay}} \neq 0$; zeros are resolution-bounded, Section 2) across the six pre-registered threshold multipliers, with Wilson 95% band (n=540 pivotal turns at θ^*); the dashed curve is the matched random-routing baseline (the routed fraction), and the dotted line marks the baseline 0.1307 at the registered threshold θ^*, against recall 0.1185 $[0.0939, 0.1485]$ — chance-level, never exceeding the matched baseline at any of the six multipliers. Right: judge-call reduction at per-turn and per-trajectory granularity; the curves cross directions across the sweep, and neither curve may be read as a single unlabelled cost figure. Sweep is sensitivity only; the single main result is multiplier 1.0. Data: `route_pack/artifacts/route_theta_sweep.json`, `route_pack/route_retrospective.json` (frozen prereg `1814166...`).

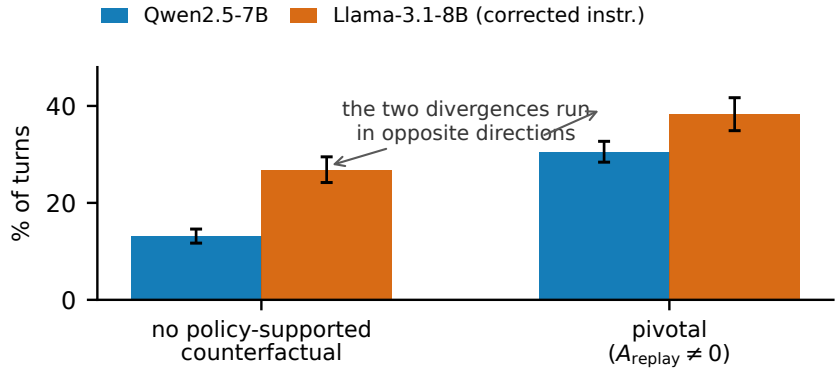

Figure 4: **Measurability of causal ground truth is model-dependent in both directions: the family with more undefined counterfactuals is the one with more causal signal where they exist.** Bars: fraction of turns with no policy-supported counterfactual (Qwen 13.1% $[11.7, 14.6]$, n=2,034 intervened; Llama 26.8% $[24.2, 29.5]$, n=1,082; corrected instrument[1]) and fraction of complete turns that are pivotal (Qwen 30.5% $[28.4, 32.7]$, n=1,768; Llama 38.3% $[34.9, 41.7]$, n=792); Wilson 95% intervals; the two divergences run in opposite directions. The absorbing axis is near-flat across families and is reported here rather than plotted: σ_{floor}=0 at 86.0% $[84.3, 87.6]$ of Qwen's and 80.6% $[77.7, 83.2]$ of Llama's complete turns; every A_{replay}=0 turn carries the resolution bound of Section 2 (unexcluded $|\Delta p| \leq 0.632$ at one-sided 95%). Data: `measurability_counts_{ks1,xfam_redo}.json`; full three-axis map: Figure 5 (appendix).

missing the aggregate level by 20 percentage points, so the ranking is trustworthy and the level is

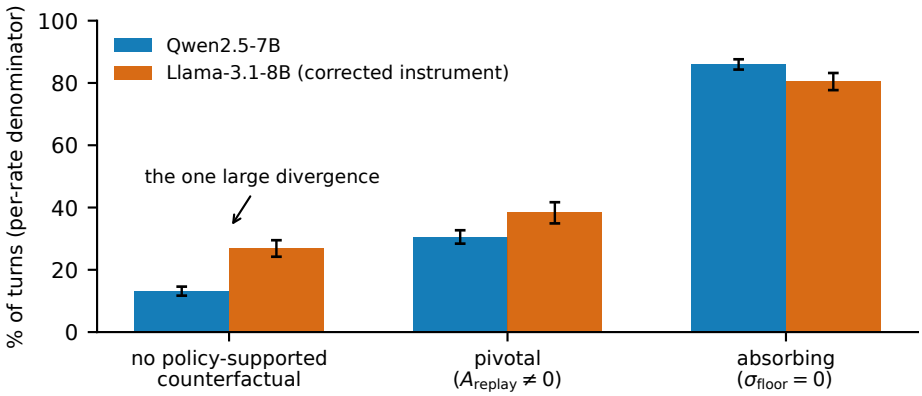

Figure 5: **Where per-step causal ground truth can be measured depends on the model — in both directions.** For each family, three fractions of the replay map: turns with no policy-supported counterfactual (Qwen2.5-7B 13.1%, n=2,034 intervened turns; Llama-3.1-8B 26.8%, n=1,082, corrected instrument[1]), pivotal turns ($A_{\text{replay}} \neq 0$: 30.5%, n=1,768; 38.3%, n=792), and absorbing turns (σ_{floor}=0: 86.0%, n=1,768; 80.6%, n=792). Error bars are Wilson 95% intervals throughout — Qwen: $[11.7, 14.6]$, $[28.4, 32.7]$, $[84.3, 87.6]$; Llama: $[24.2, 29.5]$, $[34.9, 41.7]$, $[77.7, 83.2]$ respectively. The large divergence is the first bar — Llama lacks a policy-supported counterfactual at twice Qwen's rate — yet the pivotal fraction runs the *other* way, so neither family is uniformly harder to audit. All zero-valued A_{replay} turns are resolution-bounded ($|\Delta p| \leq 0.632$, Section 2). Data: `measurability_counts_ks1.json`, `measurability_counts_xfam_redo.json`.

not; the predictor's uniform-residual assumption makes it an upper bound and therefore optimistic by construction, and no pre-GPU estimate of the measurable fraction is claimed for either family.

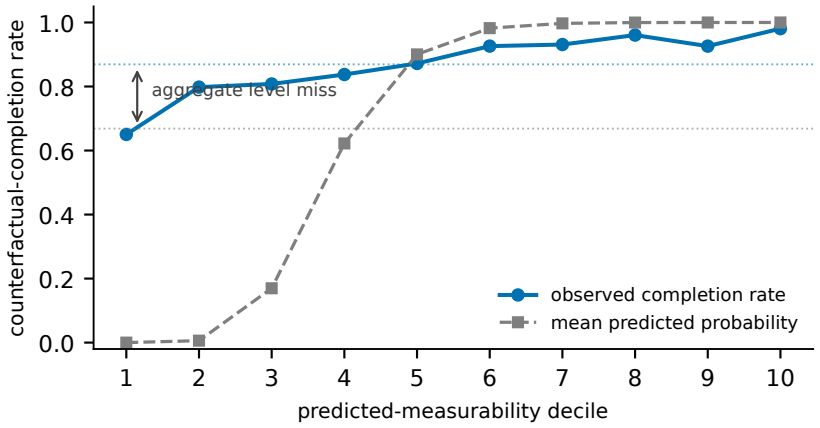

Figure 6: **An occupancy-style model of the policy's own action statistics ranks turns by measurability but does not level-calibrate.** KS1/Qwen; the cross-family replicate is reported numerically in §5. Single-family presentation is the figure's specification: the predictor is reported as a one-family observation, and the ranking-vs-level contrast (observed completion rising $0.650 \rightarrow 0.981$ across predicted deciles against a 20-percentage-point aggregate level miss) is stated in Section 3. n=2,034 intervened turns; data `runs/ks1xfam/measurability_predictor.json` (`ks1` key).

Pre-registered secondaries of the judge audit, reported. Agreement of the primary judge (Qwen2.5-72B) with the anchor judge (GPT-4.1, the audited method family's own judge, frozen 11-trajectory selection) is exact 74.1% (Wilson 95% $[68.0, 79.4]$, n=224 turns), Cohen's $\kappa = 0.496$, pooled Spearman 0.484 (`runs/collect_v2/judge_anchor_agreement.json`)

— the judge-family credit signal is substantially judge-model-sensitive. The self-preference check (self-minus-cross ρ scorer) detected none: -0.011, CI $[-0.077, 0.064]$ (`runs/collect_v2/sec7_self_preference.json`). The family $\times$ length interaction registered alongside them is not testable in this run (Section 9).

K PRE-REGISTRATIONS, PROMPTS, AND DIAGNOSTICS (VERBATIM)

Rendering policy. Every document in this appendix is included *verbatim* from its frozen source; nothing is summarized and nothing is reflowed. Two rendering tiers apply. (i) English-major sources are typeset below via a recorded, content-preserving glyph transliteration for the PDF engine (§→S, Δ→Delta, ρ→rho, → →->, ≤ / ≥ →<=/>=, ±→+/-, ×→x, ✓-class marks →[OK], crosses →[X]/[FAIL], isolated CJK tokens →[ZH]); the byte-exact file, whose sha256 is stated per item, is authoritative and ships in the supplementary pack. (ii) Chinese-major sources (the R_ROUTE pre-registration, the PC3 verdict, the DECISIONS excerpt body) ship byte-exact in the supplementary pack with sha256 and size stated here; where the paper needs their structured content readable, an English rendering that passed the element-by-element faithfulness check of the assembly record is given, and labelled as a rendering. (iii) Double-blind redaction: author names in sign-off and ruling blocks are rendered as [AUTHOR-1] in this PDF and in the supplementary files; the stated sha256 hashes refer to the unredacted originals retained by the authors, and the unredacted files ship with the camera-ready.

K.1 PRE-REGISTRATIONS, WITH AMENDMENTS AND SIGNED TAGS

```
# KILLSHOT_PREREG --- KS1, project credit-validity-paradox (v2, dual-family audit)

Status: PRESENTED FOR SIGN-OFF. Data collection blocked until [AUTHOR-1] signs (S9).
Supersedes: v1 skeleton (handoff S4). Direction per ratified Option B (2026-07-31).

## 1. Question

Two families of per-step credit are used to train LLM agents:
- **judge family** (explicit LLM-judge turn scores; known-inaccurate on attribution tasks,
accuracy length-dependent)
- **implicit family** (outcome-conditioned logprob ratios, HCAPO Eqs. 6--7; never validated
against causal ground truth)

KS1 reconciles both against environment-replay ground truth in a replayable single-agent tool
environment, per trajectory-length bucket, and issues a per-family H-branch verdict.

## 2. Environment / models (frozen)

| Item | Value |
|---|---|
| Environment | ALFWorld (TextWorld backend), determinism verified before any collection |
| Policy | Qwen2.5-7B-Instruct (HF revision pinned at harness commit), vLLM, temp 0.7 sampling
 for rollouts |
| Implicit-credit scorer | Same policy checkpoint (faithful to HCAPO); logprob eval
deterministic |
| Cross-family implicit control | Llama-3.1-8B-Instruct as rho_t scorer |
| Judge (primary) | Qwen2.5-72B-Instruct (DECISION A, ratified 2026-07-31; deviation from TARL'
s GPT-4.1 judge logged in PROVENANCE Sb) |
| Judge anchor control | GPT-4.1 on 12 trajectories --- 4 per length bucket, lowest seeds
within each bucket (frozen once buckets are known) |
| n | 50 trajectories, seeds 0--49, fixed task list committed before collection |
| Replay truth | at turn t: K=4 alternative actions resampled from same policy snapshot, each
rolled to terminal >=3x; **plus the factual action a_t re-executed >=3x at every intervened
turn**; all outcome distributions recorded |

## 3. Metric cards

### 3.1 replay_credit (ground truth)
- Definition: replay advantage A_replay(t) = mean(outcome | factual-action replays at t, >=3) -
 mean(outcome | alternative rollouts at t, K=4x>=3). Both terms are replay distributions; the
single realized continuation is never used as the factual estimate.
- Tool: harness `replay.py` (version = git SHA at run).
- Denominator: turns with >=Kx3 completed replays; incomplete turns excluded and counted.
- Failure modes: truncation/timeout (bucketed separately, S6); env non-determinism (blocked by
 smoke test).

### 3.2 implicit_credit (rho_t)
```

```
- Definition: HCAPO Eqs. 6--7 exactly --- pi_hind(a_t) = exp(mean token logprob of a_t | s_t,
s_final), self-normalized by trajectory mean pi_hind, clipped to [C_min, C_max] = HCAPO
defaults (from paper appendix; if unpublished, C=[0.5, 2.0] and logged in PROVENANCE).
- Tool: `score_implicit.py`, reads trajectory store only.
- Denominator: all action turns.
- Failure modes: prompt-template mismatch vs HCAPO appendix templates (mitigated: templates
reconstructed from paper Appendix C.2; deviation logged).

### 3.3 judge_credit
- Definition: TARL-style per-turn score from LLM judge given the complete trajectory + outcome.
 Scale: TARL's published scale (adopted verbatim from arXiv:2509.14480 prompt; adaptation for
ALFWorld action vocabulary logged). **Deviation from TARL (logged, PROVENANCE Sb): no ground-
truth tool-call annotations exist for ALFWorld, so judge runs in the w/o-gold regime.**
- Prompt: verbatim, appendix-ready, committed as `prompts/judge_credit_v1.txt` before any
scoring.
- Tool: `score_judge.py`, reads trajectory store only.
- Denominator: all action turns; unparseable judge outputs counted and reported (not silently
retried more than 2x).

### 3.4 rank_fidelity (per family)
- Definition: within-trajectory Spearman rho (primary) and Kendall tau (secondary) between {
rho_t} or {judge score_t} and {A_replay(t)}, computed per trajectory, aggregated as median
with bootstrap 95% CI (10k resamples).
- Also: sign agreement rate (credit above/below its trajectory median vs A_replay above/below
its trajectory median), Wilson CI when n<50.

**Amendment v3.2 (signed 2026-07-31, pre-analysis): trajectories with <4 complete turns are
excluded from rank_fidelity computation for both families symmetrically (completeness is
family-independent); excluded trajectories are counted and reported per length bucket.
Rationale: Spearman on <4 points is degenerate. No threshold or test spec altered.**

**Amendment v3.3 (signed 2026-08-01, day-11 REVISED ruling R5'; this wording supersedes the
draft wording carried by tag ks1-prereg-v3.3-signed --- authoritative form is tagged
ks1-prereg-v3.3r2-signed): trajectories whose A_replay is constant across all complete turns
are excluded FROM rank_fidelity COMPUTATION ONLY, for both families symmetrically, because
within-trajectory Spearman is mathematically undefined on a constant variable. This is not a
data-quality judgement and these turns are NOT discarded: they are the primary sample of the
R7 null-truth analysis and enter the R10 detectability bounds and R11 step-weighted secondary
outcome in full. Excluded trajectories are counted and reported per length bucket, alongside
the v3.2 exclusions, with effective n stated next to every rank_fidelity number. Registered
sensitivity (R5b, not primary): rank_fidelity recomputed on the subset with >=4
nonzero-A_replay turns; primary remains all Spearman-computable trajectories. No threshold
or test spec altered.**

### 3.5 structural_bias (Delta , per family) --- replaces global mean-Delta
Delta _t = normalized credit - normalized A_replay (both z-scored within trajectory). No
global mean-Delta  claim is made. Three confirmatory structural tests, specified now:
- **T1 Delta xturn-position**: per-trajectory Spearman(Delta _t, t/T); aggregate = median
across trajectories, bootstrap 95% CI (10k). Significant if CI excludes 0.
- **T2 Delta xaction-type**: ALFWorld actions binned into frozen categories {goto, take, put,
open/close, clean/heat/cool, toggle, examine/inventory/look}; Kruskal--Wallis statistic on
pooled Delta _t, but the p-value is obtained by **trajectory-level cluster permutation** (
category labels re-permuted per whole trajectory, 10k permutations) --- pooled asymptotic p is
 invalid because turns within a trajectory are not independent. If significant, Dunn post-hoc
for direction (descriptive).
- **T3 Delta xpolicy-logprob(a_t)** --- *named predicted bias for the implicit family*: per-
trajectory Spearman(Delta _t, mean token logpi(a_t|s_t) under the policy without s_final);
aggregate as T1. Directional prediction, registered now: **positive for the implicit family**
(fluent/high-probability actions receive inflated credit regardless of causal effect). No
directional prediction for the judge family.
- **Multiple-comparison correction**: the confirmatory set is 3 tests x 2 families = 6
hypotheses; Holm--Bonferroni at family-wise alpha = 0.05. Anything outside these 6 is
exploratory and labeled as such.
- **sigma_floor scope (Edit 3)**: T1--T3 run on **all turns** as primary --- censoring by |
Delta _t| would select on the dependent variable's magnitude and distort correlation structure.
 The censored (|Delta _t| > 2sigma_floor(t)) versions are reported as sensitivity analyses
only. The sigma_floor gate applies exclusively to magnitude claims of the form "Delta  at turn
 t is significantly non-zero."

## 4. Controls (same batch as intervention --- non-negotiable)

1. **Replica noise floor**: factual-action replays (>=3x at *every* intervened turn, per S2/S3
.1) double as the noise-floor instrument --- sigma_floor(t) = SD of outcome across factual
replicas at turn t; pooled sigma_floor reported per length bucket. Any |Delta _t| enters S3.5
tests only if |Delta _t| > 2sigma_floor(t); censored turns are counted and reported.

### 4.5 Compute-fallback subsampling (pre-registered; trigger + scheme frozen now)
- **Trigger**: fired if, at day 10, <60% of scheduled turn-interventions are complete, OR
projected total rollouts exceed 1.3x the budget estimated at harness sign-off (estimate
committed with the harness).
- **Degradation ladder, applied strictly in order, each step logged in DECISIONS.md**:
```

```
  1. Turn stratification: trajectories with T <= 12 keep full per-turn replay; for T > 12,
intervene on ceil(T/2) turns sampled stratified by position tercile, seed = 1000 +
trajectory_id.
  2. K: 4 -> 3 alternative actions (replicas never drop below 3; factual replays never dropped
).
  3. n: 50 -> 40 trajectories, dropping highest seeds first (seeds 49->40).
- No other degradation is permitted; if step 3 is insufficient, escalate to [AUTHOR-1] (Rule H
), do not improvise.
2. Random credit, within-trajectory shuffled credit, uniform credit --- scored through the
identical rank_fidelity pipeline.
3. Cross-family scorer controls (S2) --- self-preference check for both instruments.
4. Truncated/timeout trajectories: separate bucket, never pooled into rank correlations.

## 5. Length buckets (frozen)

Primary bucketing by turn count: **B1: <=10 turns; B2: 11--20; B3: >=21.**
Secondary (reporting only): context tokens at final turn, cutpoints 3K/6K/12K (comparability
with Who&When Pro).
All S3 metrics reported per bucket per family. Bucket claims require >=8 trajectories in
bucket, else "insufficient n" is reported, not a number.

## 6. Decision rules (frozen at sign-off; per family, primary = pooled, interaction = S7)

Let rho = median within-trajectory Spearman for the family; CI = bootstrap 95% CI.

| Verdict | Rule |
|---|---|
| **H4** (credit is fine; motivation numbers were task-mismatch) | rho >= 0.6 AND **no** S3.5
structural test significant for the family after Holm correction |
| **H1** (rank-preserved but structurally biased; mechanism paper) | rho >= 0.6 AND >=1 S3.5
structural test significant for the family after Holm correction (T3-positive for the implicit
 family = the registered prediction confirmed) |
| **H3** (placebo candidate; biggest story -> immediately add training experiment) | CI(rho)
contains 0 AND rho <= 0.2 |
| **H2-partial** (partial fidelity; analysis branch) | 0.2 < rho < 0.6, or CI straddles
0.2--0.6 band |
| **INFEASIBLE** | day-7 gate: end-to-end replay of 2 trajectories not working |

**Verdict evaluation order (fixed, Edit 2)**: Controls gate -> INFEASIBLE -> H3 -> H2 (CI-
straddle takes precedence) -> H1 -> H4. A wide CI always overrides a point estimate; a family
cannot receive H1/H4 while its CI straddles the 0.2--0.6 band.
Controls gate: if the shuffled-credit control's rho CI overlaps the family's rho CI, the
family's verdict defaults to H3 regardless of point estimate.
All percentages report n; Wilson CI when n<50 (Rule S).

## 7. Pre-registered secondary analyses (named now, not post-hoc)

- **Cross-family comparison**: paired difference in per-trajectory Spearman (implicit - judge),
 bootstrap CI; "family X tracks truth better" claimed only if CI excludes 0.
- **Family x length interaction**: per-bucket rho medians; monotone-degradation claim requires
 ordered bucket medians AND non-overlapping adjacent CIs.
- **Self-preference**: same-family vs cross-family scorer rho difference, same CI standard.

## 8. Scope, timebox, escalation

- Claims locked to "replayable single-agent tool environments" (Rule H/L). Any temptation to
widen -> STOP and flag.
- Timebox 14 days from sign-off; day-7 INFEASIBLE gate; negative result is a full deliverable.
- Check-ins: day-7 gate, Rule H flags, day-14 verdict only.
- Every experiment: ALL_EXPERIMENTS_RECORD entry ([ZH][ZH]/[ZH][ZH][ZH]/[ZH][ZH]/[ZH][ZH][ZH
]/[ZH][ZH][ZH][ZH]/[ZH][ZH][ZH][ZH] + PREREG/PROVENANCE/CONTROLS/SIDE-EFFECTS) before any
number is cited.

## 9. Sign-off block

- [x] DECISION A: judge model = **Qwen2.5-72B-Instruct** (GPT-4.1 anchor, 12 trajectories, 4/
bucket, lowest seeds)
- [x] Thresholds frozen **with edits**: (1) T2 p-value by trajectory-level cluster permutation,
 10k; (2) fixed verdict order Controls->INFEASIBLE->H3->H2->H1->H4, wide CI overrides point
estimate; (3) T1--T3 primary on all turns, sigma_floor censoring = sensitivity only, gate
scoped to magnitude claims.
- [x] [AUTHOR-1] sign-off: **2026-07-31, ratified in conversation ("[ZH][ZH][ZH][ZH][ZH][ZH][
ZH][ZH] ... with edits")**. Thresholds now IMMUTABLE. Data collection unlocked.



# KS1 CROSS-FAMILY AMENDMENT --- Llama-3.1-8B-Instruct as POLICY

Status: **SIGNED 2026-08-07, tag `ks1-xfam-v1-signed`.** Countersigned by [AUTHOR-1] in the
ideation packet of 2026-08-07 with the S0 reframing and the S0.1 implicit/judge distinction.
```

```
Execution authorised in the pipeline order of S8, under the $60 ceiling and the $40
post-collection projection stop.

## 0. WHY THIS IS AN AMENDMENT AND NOT A RE-RUN --- three simultaneous changes

KILLSHOT_PREREG.md **S2 line 19** freezes the policy:

> `| Policy | Qwen2.5-7B-Instruct (HF revision pinned at harness commit), vLLM, temp 0.7
sampling for rollouts |`

**This is a design-level finding, not bookkeeping.** Swapping the policy changes **three
things at once**, because KS1's design couples all three to the same checkpoint:

1. **The measured object** --- the rollout distribution, i.e. the trajectories themselves.
2. **The measuring instrument** --- S2 line 20 defines the implicit-credit scorer as *"Same
   policy checkpoint (faithful to HCAPO)"*. The rho_t instrument is **made of** the policy.
3. **The ruler** --- S2 line 25 builds replay ground truth by resampling alternatives *"from
   same policy snapshot"*.

So a policy swap swaps the object, the instrument and the ruler **simultaneously and
inseparably**. No design can decompose that at any n: it is a property of KS1's architecture,
not of this amendment's budget. Any result must therefore be reported as *"under
Llama-3.1-8B as policy-scorer-and-ruler"*, never as *"KS1 with a different model"*.

### S0.1 The whole-system framing applies to the IMPLICIT family ONLY

This distinction is load-bearing and must appear wherever either family is described.

**For the implicit family the whole-system framing is CORRECT, and it is a feature, not a
limitation.** HCAPO's scorer **is** the policy by definition, and replay ground truth is
defined over **the policy's own action distribution**. Policy, scorer and ruler are therefore
**one object**, and T3 is a coherent system-level claim about that object:
> *"a policy's own implicit credit does not track its own causal ground truth."*
Replicating T3 under a second policy family tests exactly that claim, with nothing held
artificially fixed that the theory says should vary.

**For the judge family it does NOT apply.** The judge is Qwen2.5-72B-Instruct and **does not
change when the policy changes**, so for that family the swap is a **clean single-variable
change**: same instrument, different measured object. This is why the (descriptive) judge
column, underpowered as it is, is interpretable in a way the implicit column is not.

**THE NARROWER CLAIM THIS DESIGN CANNOT SUPPORT, stated so it is never implied:** that the
**scoring instrument is bad independently of the policy**. For the implicit family that
statement is not even well-formed --- there is no instrument separable from the policy to
evaluate. For the judge family the design lacks the power to assert it (S4). **No document
arising from this amendment may claim that implicit credit scoring is intrinsically
unfaithful; only that a policy's own implicit credit does not track its own causal truth, in
the systems tested.**
**KS1's frozen prereg, its analysis pipeline, and VERDICT.md are UNTOUCHED and remain the
record for Qwen2.5-7B.** Nothing here may revise or reinterpret a KS1 confirmatory number.

## 1. SCOPE --- one primary endpoint

### PRIMARY (the sole registered endpoint): **T3**
`T3 = per-trajectory Spearman(Delta _t, mean token logpi(a_t|s_t) under the policy without
s_final)`,
aggregated as T1 (KILLSHOT_PREREG S3.5 line 73), with the registered KS1 directional
prediction: **positive for the implicit family**.

**Why T3 and only T3** ([AUTHOR-1]'s ruling, reasoning recorded): T3 *is* the mechanism claim
--- the
fluency echo, that fluent high-probability actions receive inflated credit regardless of
causal
effect. It is the finding most in need of cross-family support, and per S4 it is **the only
quantity this n can actually resolve** (effect +0.752 against an MDE of 0.116--0.127, roughly
six-fold headroom).

### DESCRIPTIVE, SAME BATCH, PRE-LABELLED UNDERPOWERED (not endpoints, no verdict)
Family fidelity for both instruments (rho) and the measurability map are computed from the
same
trajectories at no extra cost and reported **with their MDEs printed beside every point
estimate**, explicitly labelled *underpowered, descriptive only*. **They cannot support or
refute anything under this amendment.** Reporting them is for completeness and for whatever a
future better-powered study can pool.

**NOT replicated and not to be inferred:** the rest of the S3.5 battery, the Holm-corrected
confirmatory family, the S6 verdict order, self-preference, precision@pivotal, cross-family
paired differences. **No H1/H2/H3/H4 verdict is issued.**

## 2. THE DIFF vs KS1
```

```
| Element | KS1 (frozen) | This amendment | Changed? |
|---|---|---|---|
| Policy | Qwen2.5-7B-Instruct | **Llama-3.1-8B-Instruct** | **YES** |
| Implicit scorer | same policy checkpoint | **= Llama-3.1-8B** | **YES, by coupling (S2 L20)
** |
| Replay-truth policy | same policy snapshot | **= Llama-3.1-8B** | **YES, by coupling (S2 L25
)** |
| Judge | Qwen2.5-72B-Instruct, w/o-gold | Qwen2.5-72B-Instruct, w/o-gold | no --- **held
fixed deliberately** |
| Judge prompt | frozen TARL-derived | unchanged | no |
| Trajectories | 50 collected / 37 effective | **~=25--30 collected** | **YES (scope)** |
| Replay K | K=4 alts, >=3 rollouts each, factual re-executed >=3x | unchanged | no |
| Temperature | 0.7 rollouts | unchanged | no |
| rho_t definition | HCAPO Eqs 6--7 | unchanged | no |
| T3 statistic | per-traj Spearman(Delta _t, mean token logpi) | unchanged | no |
| **T3's status** | one of the S3.5 battery | **SOLE PRIMARY ENDPOINT** | **YES (v2 narrowing)
** |
| **Family fidelity status** | confirmatory | **descriptive, underpowered** | **YES (v2
narrowing)** |
| sigma_floor scope | T1--T3 all turns primary | unchanged | no |
| Analysis pipeline | frozen KS1 code | **run THROUGH it, unmodified** | no |

**The judge is held fixed on purpose.** Changing policy *and* judge together would confound
"survives a different policy family" with "survives a different judge". Holding the judge
constant keeps the (descriptive) judge column interpretable; the implicit column is
unavoidably confounded with the policy swap **by construction** --- see S0 --- and that
asymmetry
is restated wherever the implicit numbers appear.

## 3. CRITERION --- applies to T3 alone

**T3 replicates iff BOTH:**
1. the T3 point estimate falls **within KS1's original 95% CI [0.647, 0.793]**, AND
2. the **controls gate fires in the same direction** (shuffled control does not reproduce the
   T3 effect), as in KS1.

Reference values from VERDICT.md, untouched by this amendment:

| Quantity | KS1 point | KS1 95% CI | Status here |
|---|---|---|---|
| **implicit T3** | ***+0.752** | **[0.647, 0.793]** | **PRIMARY** |
| implicit rho | 0.0193 | [-0.109, 0.081] | descriptive, underpowered |
| judge rho | 0.1142 | [0.027, 0.168] | descriptive, underpowered |
| implicit shuffled control | --- | [0.005, 0.114] | controls gate |
| judge shuffled control | --- | [-0.049, 0.117] | controls gate |

## 4. MDE AT THIS n --- and why it forced the narrowing

Derived from KS1's own published CIs (half-width -> SE -> implied per-trajectory SD -> SE at
the
new n); 80% power, two-sided alpha = 0.05:

| Quantity | implied SD | n=25: 95% half-width / MDE | n=30: 95% half-width / MDE |
|---|---|---|---|
| **implicit T3 (PRIMARY)** | 0.227 | 0.089 / **0.127** | 0.081 / **0.116** |
| implicit rho (descriptive) | 0.295 | 0.116 / **0.165** | 0.106 / **0.151** |
| judge rho (descriptive) | 0.219 | 0.086 / **0.123** | 0.078 / **0.112** |

**The arithmetic that drove the ruling:** the judge-rho MDE (0.112--0.123) is **as large as
KS1's
entire judge point estimate (0.114)**. At this n that comparison cannot distinguish "the
effect
vanished" from "the effect is exactly what KS1 measured" --- both lie inside one CI.
Registering
it as an endpoint would manufacture a decision the data cannot support. **T3 clears its MDE by
roughly six-fold and is therefore the only defensible primary.**

**BINDING INTERPRETATION RULE (retained verbatim from v1):**
> **FAILURE TO FALL WITHIN THE ORIGINAL CIs IS INCONCLUSIVE, NOT A REFUTATION.**
> A miss is reported as *"not replicated at n~=25--30, MDE = X, below which this design cannot
> discriminate"* --- never as *"KS1 does not generalize"*. Only a **T3 SIGN REVERSAL** (median
> Spearman significantly negative) is positive evidence against the KS1 mechanism claim,
> because T3 is the only quantity this n resolves. This applies with even more force to the
> descriptive outputs, which cannot be called a failure to replicate at all.

## 5. ABORT GATE --- retained; the likeliest failure mode

**ONE-DAY PARSER-ADAPTATION ABORT GATE.** The action parser is checkpoint-sensitive --- a
known,
twice-realized landmine (e17 tag-format drift; e30, the canonicalizer that silently produced
1--3% agreement). Llama-3.1's action formatting will differ from Qwen2.5's.
```

```
- **One day** to adapt the parser and reach an acceptable parse rate on a pilot batch.
- **Pre-registered pilot gate: >=90% strict-tag parse rate on 20 pilot trajectories.** Below
  that, **ABORT** and report the parse rate as the finding --- an unparseable policy is a real
  and publishable measurability result, not a failed run.
- Parser changes are **additive only**: the Qwen path stays byte-identical, verified by
  re-running the KS1 smoke and matching its trajectory hash. **A parser edit that changes any
  KS1 number is an automatic stop.**

## 6. OPERATIONAL PRECONDITIONS

- **`HF_TOKEN` REQUIRED**: `meta-llama/Llama-3.1-8B-Instruct` is a **gated** repo. Access must
  be granted on the token's account *before* the run; a gated-repo 401 at load time is a stage
  failure, not a retry.
- **R21 content gate (iron rule 7)** unchanged: sha256 of every `*.safetensors` shard against
  the HF blob filename, on the pod that will run it, path-bound to the ambient `HF_HOME`
  (e34), report written to `runs/<run>/checkpoint_verification*.json`.
- Disk: a second 8B policy alongside the existing cache; F6 verify-then-free applies.
- Iron rules 8 (supervised-run), 9 (report=disk, **including the e40/e41 three-clause
  read-back: NUL count 0, text file type, and a content assertion against intended
  structure**) and 10 (F7 conformance gates) apply throughout.

## 7. BUDGET AND DISCIPLINE

GPU cost **not yet estimated**; no figure is asserted here. A costed plan accompanies the
countersignature request. Negative and inconclusive outcomes are **full deliverables**.
ALL_EXPERIMENTS_RECORD entry per stage boundary. Exploratory throughout; touches no KS1 or KS2
confirmatory family.

---
**SIGNED. Execution authorised in S8 pipeline order. The abort gate of S5 and the budget stop
of S7 bind absolutely.**

## 8. PIPELINE ORDER ([AUTHOR-1], 2026-08-07) --- ceiling cuts the least valuable item first

1. **Parser pilot + abort gate** --- >=90% strict label-parse on 20 pilot trajectories, one
day
   max. Three prior incidents (e17 tag drift, e30 canonicalizer, the A3B canonicalizer) say to
   expect adaptation work. **Do not fight past the gate.**
2. **Collection** (n = 25--30).
3. **Replay ground truth.**
4. **Logprob dumps** (policy + hindsight, self scorer).
5. **T3 analysis --- THE SOLE PRIMARY ENDPOINT.** The run has delivered its purpose once this
   lands.
6. **Implicit family fidelity + measurability map** --- descriptive, underpowered, MDE beside
   every estimate.
7. **72B judge scoring --- LAST and OPTIONAL.** It serves only a pre-labelled underpowered
   descriptive output, and costs ~2 GPU-h plus a 145 GB staging operation that has caused
   repeated incidents. **If the budget ceiling or the calendar is anywhere near, SKIP IT and
   record that it was skipped BY DESIGN, not by failure.**

**BUDGET STOP:** after step 2, measure the realized rate, project the total for steps 3--6,
and
post the projection before continuing. **If the projection exceeds $40, STOP and queue for
[AUTHOR-1].** Overall ceiling $60. `HF_TOKEN` required; if auth fails, **abort Task 3, log it,
 and
continue the queue --- do not stall the night.**




# KS2_PREREG --- Training-loop closure of the H3 (placebo) verdict (v1.2)

Status: PRESENTED FOR SIGN-OFF. Training blocked until [AUTHOR-1] signs (S10) and the analysis
pipeline is fixture-tested and tag-frozen (S8, CC-5 pattern). Parent: KS1 VERDICT.md
(double-H3, Controls gate; implicit:T3 confirmed). Governing kit: CLAUDE.md iron rules,
R21 content-hash gate, Rules A--S.

## 1. Question

KS1 established that both per-step credit families are placebo-level at *ranking* causal
truth. The standard rebuttal: "training may not consume rank fidelity" (group-relative
aggregation, batch effects). KS2 closes the loop: **do the credit families differ from
each other, from no step-credit at all, and from sign-inverted credit, in the training
outcome they were built to improve?**

## 2. Training regime (Plan A --- iterated offline-scoring loop; frozen)

Per [AUTHOR-1] ruling (Rule-A fix): NOT pure on-policy. Each round:
1. Current policy collects a batch of trajectories on the frozen training task list
   (temp 0.7, seeded).
2. Training pauses; all seven credit sources are scored OFFLINE over the same batch.
```

```
3. Each arm applies its own credit to the SAME base update rule (step-weighted
   group-relative advantage, LoRA on Qwen2.5-7B-Instruct; optimizer/LR/steps frozen at
   harness sign-off); next round.
- Rounds: **3**. Batch: **96** trajectories/round/arm-seed. (Reduced from the v1 draft's
  4x128 --- BUDGET-DRIVEN, frozen now via the S9 budget table; any further change is amended
  BEFORE any training, never after.) Round-1 collection is SHARED across arms (identical
  policy + CRN streams make the seven round-1 batches byte-identical by construction;
  collected once, consumed seven ways).
- **Registered deviation**: one step removed from the audited papers' on-policy setting;
  accepted because it makes the six arms' training dynamics strictly isomorphic --- the only
  varying factor is the credit source. Logged here, restated in the paper.

## 3. Arms (seven; identical everything except credit source)

| # | Arm | Credit at round r |
|---|-----|-------------------|
| 1 | outcome-only | group-relative outcome advantage, no per-step weighting |
| 2 | implicit | HCAPO rho_t (Eqs. 6--7, published constants) scored by the CURRENT policy
checkpoint on the round's own trajectories |
| 3 | judge | Qwen2.5-72B judge, frozen KS1 prompt (blob a1b59989...), w/o-gold, scored per
round --- **AWQ int4 permitted only if S7 consistency gate passes** |
| 4 | shuffled | implicit instrument (HCAPO rho_t) run BY THIS ARM on ITS OWN current-round
batch, then permuted within trajectory (seeded) |
| 5 | inverted | implicit instrument run BY THIS ARM on ITS OWN current-round batch, then sign-
flipped about the trajectory mean |
| 6 | pivotal-masked truth | cheap replay on the round's own trajectories, **stratified 50% of
 turns** (position-tercile stratified, seed = 3000 + traj_id; unsampled turns carry credit 0):
 K=2 alternatives x 2 replicas + factual x 2, sign(A_replay) only; credit = +/-1 on sampled
turns with defined nonzero sign, 0 elsewhere |
| 7 | shuffled-judge | judge instrument run BY THIS ARM on ITS OWN current-round batch, then
permuted within trajectory (seeded) --- judge's OWN marginal ({1,0,-1}, tie-heavy), so the C2
control matches the arm it controls for |

**Control-arm credit source rule (Rule-A fix, v1.2):** from round 2 the arms' policies
diverge and each arm collects its own batch, so "arm-2/arm-3 values" do not exist on a
control arm's trajectories. Each control arm therefore RUNS THE RELEVANT INSTRUMENT ITSELF
on its own batch (arms 2/4/5 -> implicit scoring; arms 3/7 -> judge scoring) and applies its
transform to those scores. Semantics: the control measures the effect of the instrument's
score DISTRIBUTION with its information removed, on that arm's own data. At round 1 the
shared CRN batch makes this exactly equivalent to the earlier phrasing.

Arm-6 replay budget is a REDUCED-RESOLUTION instrument (sign only); this is stated wherever
arm-6 is discussed and never conflated with KS1's K=4x3 estimator.
**Arm-6 degradation ladder (frozen now, applied in order, each step logged):**
1. Turn sampling 50% -> 33% (same stratification, same seed rule).
2. K=2x2+factualx2 -> K=2x1+factualx1 (3 rollouts/turn; halves arm-6 cost).
No other degradation permitted; beyond step 2 -> escalate to [AUTHOR-1].

## 4. Paired design (common random numbers; frozen)

- Seeds: S = {0,1,2} per arm; **identical across arms** --- same task order, same env seeds,
  same round-r collection RNG streams (arm enters the seed derivation only downstream of
  collection). All arm-vs-arm comparisons are seed-paired.
- SD source for the power check (stated precisely): arm-1's 3 seeds yield the BETWEEN-SEED
  SD, which is a proxy UPPER BOUND on the paired SD (CRN pairing can only shrink it), so
  the S4 power projection computed from it is conservative in the safe direction.
- Adaptive rule, FROZEN NOW: after arm-1's 3 seeds complete, compute the realized paired SD;
  if projected paired-TOST power at +/-3pp < 80%, extend ALL arms to 5 seeds (S += {3,4});
  budget pre-approved (+~$35). No other adaptation permitted.

## 5. Evaluation protocol (frozen)

- Held-out set: ALFWorld unseen split, task list of **128 tasks committed before any
  training**; disjoint from the training list and from collect_v2.
- Per seed per task: **1 episode**, greedy decoding (temp 0). (v1 had 2 episodes --- a
  determinism bug: ALFWorld + temp 0 makes the second episode byte-identical to the first,
  so n=256 would be fictitious and the Wilson CI falsely narrow. Honest n = 128; task count
  is the variance driver.)
- Metric: win rate over 128 episodes/seed, reported with n and Wilson 95% CI, per arm per
  round (learning curve) and at the final round (primary endpoint).

## 6. Confirmatory analysis (frozen; paired TOST + Holm)

Primary endpoint: final-round win rate difference, seed-paired.
- **Equivalence margin +/-3pp** ([AUTHOR-1]-ratified; = half the Spurious-Rewards random-vs-
truth
  gap). Paired TOST at alpha=0.05.
- Confirmatory comparisons (Holm over exactly these six), each with its REGISTERED test type:
  C1 implicit vs shuffled(rho) --- paired TOST (equivalence)
  C2 judge vs **shuffled-judge (arm-7)** --- paired TOST (equivalence); control matches the
     controlled arm's own marginal distribution
```

```
  C3 implicit vs outcome-only --- paired TOST (equivalence)
  C4 judge vs outcome-only --- paired TOST (equivalence)
  C5 inverted vs implicit --- one-sided paired superiority, registered direction: inverted
WORSE
  C6 pivotal-masked vs outcome-only --- one-sided paired superiority, registered direction:
masked BETTER
  The S7.4 fixture test must plant all three truth types (equivalent / directionally
  non-equivalent / inconclusive) and the pipeline must recover each under this mixed family.
- Reporting rule: "equivalent" ONLY if TOST significant; TOST non-significant AND
  difference-test non-significant -> **inconclusive** (never "no difference"). Power
  calculation with assumed SD in S9; validated against arm-1 realized SD (S4).
- Verdict lines (independent, never merged):
  V1 placebo-closure: C1--C4 all equivalent -> per-step credit content is inert in training.
  V2 information-without-alignment: C5 shows inverted significantly WORSE -> credit carries
     signal that is misaligned, not absent.
  V3 method byproduct: C6 shows pivotal-masked significantly BETTER -> sparse causal
     masking beats dense heuristic credit.

## 7. Phase-0 gates (before any training)

1. R21 content-hash verification of every checkpoint on the new pod (iron rule).
2. **AWQ consistency gate**: score the KS1 anchor set (11 frozen trajectories, 224 turns)
   with 72B-AWQ vs the stored 72B-bf16 judgments; require exact-agreement >= 90% AND
   kappa >= 0.8 (i.e., AWQ-vs-bf16 must agree far better than judge-vs-anchor's kappa=0.496,
so
   quantization noise is small relative to judge-model noise). Pass -> arm-3 uses AWQ on
   1xA100; fail -> judge scoring rounds run on 2xA100 bf16 (+~$20), logged.
3. **R29 decomposition (pre-training, exploratory, into the paper regardless of outcome)**:
   from existing collect_v2 dumps --- increment g_t = logpi_hind - logpi_policy;
   report corr(g_t, A_replay) (all complete turns + pivotal-only), partial
   corr(rho_t, A_replay | logpi_policy), and Delta  regressed on (fluency, increment).
   Claim template it feeds: "outcome-conditioning buys no causal information."
4. Analysis pipeline fixture test (CC-5): synthetic runs with planted
   equivalent / non-equivalent / inconclusive truths; pipeline must recover all three;
   tag `ks2-analysis-frozen` BEFORE first training batch lands.

## 8. Records & discipline

- ALL_EXPERIMENTS_RECORD entry per arm-seed-round, written at each round's scoring point,
  never batched at the end.
- One config change per run; configs are files; every round writes trajectories, credit
  files, update logs, eval JSON, git SHA.
- Repro package skeleton (env, prompts, scoring scripts, analysis) grows alongside runs
  (Rule R).
- Monitors auto-advance the 18+ runs; human appears only at round-boundary scoring points
  and the S10 gates.

## 9. Scope, budget, calendar

- **Scope lock (frozen wording)**: all KS2 conclusions are limited to the Qwen2.5 model
  family, ALFWorld, LoRA fine-tuning at the frozen scale, under the Plan-A iterated
  offline-scoring regime. No cross-family or cross-environment generalization is claimed.
- **Budget table (per stage, at KS1 measured rates: 1,700 continuation-rollouts/h and
  ~800 trajectories/h at the 48-worker tier on 1xA100+16vCPU; 112-worker tier measured
  4,465/h on the judge pod is upside margin, not assumed):**

  | Stage | Volume | GPU.h | $ |
  |---|---|---|---|
  | Collection (r1 shared + r2--3 x 7 arms x 3 seeds x 96) | 4,320 traj | 5.4 | $8 |
  | Arm-6 replay, default tier (50% turns x 6 rollouts x 3r x 3s) | 51,840 rollouts | 30.5 |
$45 |
  | Judge scoring, arms 3 AND 7 (2 x 17,280 turns, AWQ @~1,000 turns/h) | 34,560 turns | 34.6
| $52 |
  | LoRA updates (all arms) | --- | 4 | $6 |
  | Eval (7 arms x 3 seeds x 128 tasks x 1 ep, greedy) | 2,688 eps | 3.4 | $5 |
  | Overhead (loads, gates, R29) | --- | ~3 | $4 |
  | **Base total** | | **~81** | **~$120** |
  | Adaptive 5-seed extension (x5/3 on seed-scaled stages) | | | cap rises to $200 (rule below
) |
  | Arm-6 ladder step 2 (if invoked) | -25,920 rollouts | -15.2 | -$23 |
  | AWQ-fail contingency (judge stage on 2xA100 bf16) | | | +$20 |

  Assumed paired SD for the S6 power calc: 1.5pp (validated per S4; between-seed proxy is a
  conservative upper bound).
- **Cap-interaction rule (frozen now):** if the S4 adaptive 5-seed rule fires, the budget
  cap rises automatically to **$200** (pre-approved by [AUTHOR-1]) and arm-6 REMAINS at its
  default tier --- the degradation ladder does NOT co-fire with the seed extension.
  Instrument resolution is never traded against statistical power; the ladder is a
  triggered fallback for its own S9 conditions only.
- Arm-6 default tier CONFIRMED at 50% turns x 6 rollouts ([AUTHOR-1] ruling): C6 is a one-
sided
```

```
  superiority claim, and coarser sign estimates attenuate the effect toward zero --- the
  ladder is a fallback, not the default.
- Calendar: sign -> pod up -> S7 gates + R29 (day 1) -> training rounds (days 1--3) ->
  eval + analysis + VERDICT_KS2 (day 3--4) -> numbers by 2026-08-10. ScienceWorld checkpoint
  2026-08-20 per [AUTHOR-1] ruling (3).

## 10. Sign-off block

- [x] Plan A ratified (registered deviation from on-policy accepted)
- [x] Arms 1--7 as specified; arm-6 reduced-resolution instrument + frozen ladder; arm-7
      marginal-matched control for C2
- [x] Budget table reviewed; rounds=3 / batch=96 accepted as budget-driven freezes
- [x] Eval n=128 (1 greedy episode/task) accepted
- [x] Paired CRN design + frozen adaptive-seed rule
- [x] TOST +/-3pp, Holm over C1--C6, inconclusive-reporting rule
- [x] Phase-0 gates incl. AWQ kappa>=0.8 / exact>=90% threshold
- [x] [AUTHOR-1] signature: **[AUTHOR-1], 2026-08-02** --- SIGNED as v1.2 with the six/seven
word-level fix. Thresholds, arms, budget rules, and gates IMMUTABLE from tag ks2-prereg-v1.2-
signed.

## Amendment v1.3 (2026-08-03, [AUTHOR-1] rulings R33/R34 on the E014 budget stop; signed
BEFORE round 1 --- no training data existed at signing)

1. **Arm-6 ladder step 1 INVOKED** (its registered S9 budget condition fired --- E014 ledger):
   turn sampling 50% -> 33% (want = ceil(T/3)), same tercile stratification, same frozen seed
   rule (3000 + traj_id), T<=12 keep-all unchanged. Logged per S3 ("each step logged").
2. **Arm-6 resolution statement**: the alternative-draw cap is 20 seeded draws/turn. Turns
   where 2 distinct admissible alternatives are not found within 20 draws carry credit 0 ---
   identical in kind to arm-6's existing semantics for unsampled/undefined turns. This
   restates the reduced-resolution instrument's resolution; the estimator (sign(A_replay),
   K=2x2+factualx2 on sampled defined turns) is unchanged.
3. **Eval schedule (R34)**: primary endpoint UNCHANGED --- final round, all 7 arms x 3 seeds,
   128 tasks, greedy, Wilson CI. Per-round learning curves demoted to crn-seed-0 only
   (7 evals per round for rounds 1--2), labeled exploratory/descriptive.
4. Budget: operational cap $160 standalone / $220 if the S4 5-seed rule fires (R37). LoRA
   line rebased to measured 16.8 GPU.h (R35). Batch 96 / rounds 3 untouched (R36).

[AUTHOR-1] signature: **[AUTHOR-1], 2026-08-03** --- SIGNED as v1.3. All other v1.2 freezes
unchanged.

## Amendment v1.4 (2026-08-04, [AUTHOR-1] ruling R41 on the E016 policy collapse; signed
BEFORE any confirmatory number exists --- the r1 evals under v1.3 are diagnostics of harness
failure, quarantined per R41c)

S2 update rule REPLACED for all seven arms identically (credit source remains the only
cross-arm difference):
1. PPO-style per-token ratio clipping against the collection-time logprobs stored in
   gen_logprobs, eps = 0.2.
2. Advantage z-normalization per batch before weighting: A_z = (outcome - batch mean)/batch SD
   (sd=0 -> degenerate-batch rule unchanged). Credit w_t is applied to A_z exactly as before;
   zero-credit => zero-gradient is PRESERVED (normalizing the product would give unsampled/
   undefined turns nonzero gradient and silently break arm-6/-3/-7 semantics). Pre/post stats
   recorded per D2.
3. Optimizer budget: max 1 epoch (unchanged), lr 5e-5 (was 1e-4).
4. KL anchor to the round-start policy, coefficient 0.05, estimated on the realized tokens
   from the same stored logprobs; realized KL logged per update.
R41b: post-update generation sanity gate (4 greedy episodes; assert no >8-token repeat loops
and nonzero action-extraction) added to the fixtures AND run per-arm before every eval.
R41c: v1.3-rule adapters/evals quarantined, kept as the training-fragility evidence base.
R41e: C6 is read jointly with per-arm effective-optimizer-step counts; sparsity-mediated
survival is a named confound.
Budget: re-run inside existing ledger lines. Calendar: verdict target 8/7, hard line 8/10.

[AUTHOR-1] signature: **[AUTHOR-1], 2026-08-04** --- SIGNED as v1.4. All other freezes
unchanged.




# A3B_PREREG --- Does credit source systematically shape policy behaviour?

Status: PRE-REGISTERED BEFORE ANY AGREEMENT NUMBER WAS COMPUTED. Tag `a3b-prereg-v1`.
Exploratory throughout; touches no confirmatory family. KS1/KS2/PC/PC2 preregs and both
verdicts remain IMMUTABLE.

## The sentence under test

E026/B23 recorded: **"seven behaviourally distinct policies with no detectable performance
difference."** A3-RL established that each arm differs *from base* (~50% greedy action
agreement). It did NOT establish that the arms differ *from each other in an arm-specific
way*. If checkpoints differ from each other by roughly the same amount whether or not they
```

```
share a credit rule, then "seven behaviourally distinct policies" overstates the evidence and
the sentence must be WITHDRAWN. This prereg decides that, with thresholds fixed in advance.

## Metrics

- **Primary:** pairwise greedy action agreement over a FIXED state bank, on **canonicalized**
  actions.
- **Secondary (string):** raw-string agreement, same pairs, same states.
- **Secondary (continuous):** pairwise symmetric-KL and Jensen--Shannon divergence between
  next-action-token distributions (teacher-forced logprobs from the same forward pass).
  Rationale: with only 21 checkpoints, permutation power comes from checkpoint count; a
  continuous metric cuts per-pair noise and sidesteps parser sensitivity entirely.

## Pair classes

- **A** --- same arm, different seeds (n = 7 arms x 3 seed-pairs = 21 pairs)
- **B** --- different arms, same seed (n = 3 seeds x 21 arm-pairs = 63 pairs)
- **C** --- different arms, different seeds (n = 126 pairs)

### CRN asymmetry --- stated deliberately, in advance

Under the frozen CRN design, **round-1 collection was SHARED across arms**. Therefore:
- **B pairs share the entire round-1 batch** (and differ only in credit rule thereafter).
- **A pairs share NO collected data at all** --- different seeds means different round-1,
  round-2 and round-3 batches. They share *only the credit rule*.

Requiring `mean(A) > mean(B)` therefore asks whether **the credit rule shapes behaviour more
than a shared collection history does**. That is the strong bar, and it is chosen deliberately.

A result where B > A means shared data dominates credit rule --- i.e. collection randomness is
the larger force --- and the claim fails regardless of the A-C gap.

## Primary statistic

`gap = mean(A) - mean(C)`

## Null --- RESTRICTED PERMUTATION (replaces free permutation)

Permute **arm labels WITHIN each seed independently**, 10,000 times, recomputing `gap` each
time. Rationale recorded now: free permutation across all 21 checkpoints creates pseudo-arms
containing two checkpoints from the *same* seed, which share round-1 data and therefore
inflate
the null, wasting power. Within-seed permutation preserves the exact design structure (exactly
one checkpoint per seed per arm) and is the correct exchangeability null.

## FROZEN DECISION RULE

The sentence is **SUPPORTED** iff **all three** hold:
1. `gap = mean(A) - mean(C) >= 5 pp`, AND
2. restricted-permutation `p < 0.05`, AND
3. `mean(A) > mean(B)`.

Otherwise the sentence is **WITHDRAWN** and the null is reported plainly. `mean(A)` and
`mean(B)` are reported separately in every case. If `B >= A`, collection randomness dominates
and the claim fails regardless of the gap.

## Structural hypotheses --- FROZEN BEFORE SEEING THE 21x21 MATRIX

Two mutually exclusive partitions of the seven arms:

- **H_instrument:** {2,4,5} rho-family . {3,7} judge-family . {6} . {1}
- **H_information:** {2,3,6} informative . {4,7} shuffled . {5} inverted . {1} none

**Fixed comparison criterion (stated now):** for each partition compute
`Delta  = mean(within-group pairwise similarity) - mean(between-group pairwise similarity)`,
using arm-level similarity (the mean over that arm-pair's constituent checkpoint pairs).
Each partition gets its **own restricted permutation test** (same within-seed scheme, 10,000
draws) yielding its own p. The partition with the larger Delta  **and** the smaller p is
declared
the better explanation; if they disagree, that is reported as indeterminate rather than
resolved post hoc.

**Reading fixed in advance:** if **H_instrument** wins, the recorded interpretation is that
behaviour is shaped by the credit signal's **DISTRIBUTIONAL SHAPE, not its information
content** --- consistent with KS1's fluency-echo finding and R29's zero-information increment.
**A partition will NOT be chosen after seeing the matrix.**

## Exclusions

**arm5/seed2 is EXCLUDED from all primary analysis** (87.7% tag-in-action format collapse,
E020). It is reported separately as a degenerate case. Every A/B/C count above is reduced
```

```
accordingly, and the realized counts are reported.

## Stages

- **Stage 0 (operational):** verify all 21 final arm-seed adapters exist on the RETAINED
  NETWORK VOLUME (not container disk), with content hashes. Anything living only on the pod is
  copied to the volume and re-hashed. Full inventory reported. The pod is NOT shut down until
  Stage 2 completes.
- **Stage 1 (zero GPU):** turn 1 of each of the 128 held-out tasks is faced by every
checkpoint
  from a byte-identical initial observation. Extract each checkpoint's turn-1 action from the
  stored final-round eval trajectories, canonicalize, compute A/B/C + restricted permutation
  over these 128 unbiased states. **Ceiling risk reported explicitly:** the fraction of states
  where ALL checkpoints agree (turn-1 ALFWorld may be near-trivially constrained).
- **Stage 2 (GPU, ONLY if Stage 1 is at ceiling / underpowered / ambiguous):** ~500 states
  sampled from **BASE-model** trajectories on held-out tasks (base is the neutral common
  ancestor --- sampled from no arm's own rollouts), stratified by turn position. **State bank
and
  its hash committed BEFORE scoring.** All 21 checkpoints scored greedy (temp 0) on identical
  prompts using the same template as RL training; raw generations AND next-action-token
  logprobs stored. Everything recomputed on this bank.

## Normalization (parser is checkpoint-sensitive --- cf. the `[action>` / `<action>` / `[
action]` lesson)

Canonicalize to a bare action string: strip CoT/`<think>` spans, strip all tag variants,
lowercase, collapse whitespace, map to the ALFWorld admissible vocabulary where possible.
Report (a) canonical agreement, (b) raw-string agreement, and (c) **canonicalization-failure
count PER CHECKPOINT**. A checkpoint with a high failure rate is a format-degraded outlier and
must be FLAGGED, never silently averaged in.

## Also reported (exploratory, cheap)

- A/B/C restricted to **CONTESTED** states (those where not all checkpoints agree), as a
  sensitivity against ceiling effects.
- The 21x21 pairwise agreement matrix + clustering/MDS.
- Whether **arm1** (outcome-only, no step credit) sits apart from the step-credit arms.
- **arm6's much smaller parameter movement (||lora_B|| 1.37 vs ~4.7 for the others)** as an
  explicit caveat wherever arm6 appears in the clustering.
- Behavioural distance vs performance distance across pairs (expected underpowered; labelled
  exploratory).

## Rules

Thresholds are frozen above and were fixed before any result was viewed. **A WITHDRAWN claim
is
a full deliverable** --- if A ~= C, that is stated plainly and the sentence is dropped from
the
paper. ALL_EXPERIMENTS_RECORD entry per stage boundary. Budget: Stage 1 free; Stage 2 <= $10,
overruns -> Rule H. ntfy at each stage boundary and at the final stop.

Signed (execution pre-registration, exploratory): **2026-08-07**, tag `a3b-prereg-v1`.
# A3B-Delta  PREREG (addendum to A3B_PREREG.md) --- mechanism chain for the A3B signature

Status: PRE-REGISTERED BEFORE ANY MATRIX, STATISTIC OR CORRESPONDENCE WAS COMPUTED.
Tag `a3b-delta-v1`. Exploratory throughout (the D4 chain); touches NO confirmatory family.
A3B_PREREG.md (`a3b-prereg-v1`), KILLSHOT_PREREG.md, KS2_PREREG.md, PC/PC2 preregs and both
VERDICTs are IMMUTABLE and UNMODIFIED --- this is a separate addendum file, not an edit to any
of them (iron rule 1).

**[AUTHOR-1]'s ratified A3B wording (E031) remains the single source of truth, unchanged by
anything here: the discrete null and the continuous positive are ALWAYS stated together.**
Nothing in this addendum may be used to restate, strengthen or soften E031.

Zero GPU. Every input is already on disk: the A3B JS matrix, the per-arm credit files, the
round logs, and the trainer source. **If any step turns out to require GPU, it STOPS and is
reported as blocked rather than started.**

---

## 0. PREMISE --- the Step-0 finding this chain is built on (DECISIONS R43, E032)

The v1.4 per-batch z-normalization applies to **A_i ALONE** (`scripts/ks2_train_update.py
:107-108`);
the product `w_t . A_i` is formed afterwards at `:208` and never rescaled. `w_t` is never
normalized. The divisor `_sd` is a function of **outcomes alone** and is therefore
**credit-rule-blind**.

**SELECTED narrative (a):** credit marginal SCALE survives normalization => arms with
different
```

```
credit scales receive systematically different **effective step sizes**.
**EXCLUDED narrative (b):** the arXiv:2509.02534 S6.3 noise-amplification route. Barred from
all drafts (R43). It does not reappear in this document except as an exclusion.

This addendum TESTS the selected narrative. A null is a full deliverable and is reported as
plainly as a positive.

---

## 1. THE BY-CONSTRUCTION SPLIT

Verified from source, `scripts/ks2_credit.py`:
- `:78-82  shuffle_within` --- permutes credit values **within a trajectory** (seeded).
  arm4 = shuffle_within(arm2's rho_t); arm7 = shuffle_within(arm3's judge scores).
- `:85-87  invert_about_mean` --- `w_t = 2.mean_traj(rho) - rho_t`. arm5 = invert(arm2's rho_t
).

Consequences, stated now:
- Shuffling preserves the per-trajectory **multiset** of credit values EXACTLY => mean, SD,
  kurtosis, sum, and sum of absolute values are all identical between arm2/arm4 and arm3/arm7.
- Mirroring preserves the **mean and SD** exactly (reflection preserves even central moments;
  odd ones flip sign) but does **NOT** preserve the absolute-value marginal: |2m - v| != |v|
in
  general.

Therefore **mean/SD/kurtosis of the credit marginal cannot distinguish arms within {2,4,5} or
within {3,7}**. Clustering on them reproduces H_instrument BY CONSTRUCTION and is not evidence.


### GROUP A --- construction-invariant (completeness only; NEVER carries the mediation claim)
Credit-marginal **mean, SD, kurtosis**. Reported for completeness and labelled as
non-evidential wherever it appears.

### GROUP B --- construction-variant (the only statistics that can carry the claim)
Internal priority frozen now: **B-PRIMARY outranks B-CORROBORATIVE**, and within B-PRIMARY,
ALLOCATION is the only sub-part that can distinguish arms *within* a family.

#### B-PRIMARY (input side --- what the gradient was actually fed)
Computed from the realized applied weights: for every turn actually trained on, the quantity
`|w_t . A_i|` with `A_i` **post-normalization** (the value the trainer applied), and each turn
carrying its **response token count** as its frequency weight (`:242` averages the surrogate
over response tokens, so token count is the correct frequency).

**B-PRIMARY-SCALE --- construction-invariant WITHIN {2,4,5} and WITHIN {3,7}; Group-A-like,
and
labelled as such.** Statistics: **mean and total of |w_t . A_i|**.
The selected narrative predicts this separates **ACROSS** families (rho-family vs judge-family
 vs
arm6 vs arm1) but **NOT WITHIN** them. That is a direct, falsifiable consequence of Step 0 and
is reported as such.

> **PIPELINE SELF-CHECK, frozen now (not a finding --- a test of my own code).** Because
> `shuffle_within` permutes within a trajectory and `A_i` is constant within a trajectory,
> `Sum _t |w_t.A_i| = |A_i|.Sum _t|w_t|` is **EXACTLY** preserved by shuffling. So on any
batch shared
> between them, B-PRIMARY-SCALE totals must be **exactly equal** for arm2 vs arm4 and for arm3
> vs arm7. Round 1's collection was SHARED across all seven arms, so at round 1 this equality
is
> exact and any deviation indicates a BUG IN MY COMPUTATION, not a result. At rounds 2-3 each
> arm collected its own batch, so equality is expected only approximately.
> **arm5 is explicitly NOT predicted equal to arm2**: mirroring preserves mean and SD but not
> the absolute-value marginal. Recording this in advance so that an arm2!=arm5 scale
difference
> is not later misread as either a bug or a discovery.

**B-PRIMARY-ALLOCATION --- construction-variant; the ONLY part that can distinguish arms
within a
family.** Shuffling changes *which* turns receive which weight, and turns differ in token
count,
so this genuinely differs between arm2 and arm4. Statistics, all token-count-weighted over
`|w_t . A_i|`:
- token-count-weighted **quantiles** q10/q25/q50/q75/q90;
- **Gini coefficient** (token counts as frequency weights);
- **effective sample size** of the weight distribution, Kish form `ESS = (Sum w)^2 / Sum w^2`,
  reported both raw and normalized by the number of turns.

#### B-CORROBORATIVE (training outcomes --- may support, NEVER carry)
Per-round **gradient-norm trajectory** and **policy-entropy trajectory** (R38 logs).
Explicitly
demoted: using training outcomes to prove "input form drives output distribution" risks
circularity, since both are downstream of the same updates. Reported, never decisive.
```

```
---

## 2. POSITIVE CONTROL --- the yardstick that makes a Group-B null interpretable

Per-arm **credit non-zero coverage / sparsity**: the fraction of turns with `w_t != 0`.
Known, by construction, to span a wide range --- arm1 uniform `w_t=1.0` on every turn; arm6
sparse
(sign of A_replay on sampled turns only, 0 elsewhere); {2,4,5} continuous and near-fully
non-zero; {3,7} three-valued in {-1,0,1} with many exact zeros. **This spread is large and
known
in advance, so it is a yardstick with adequate range.**

The same correspondence procedure (S3) is run against the JS matrix for sparsity.

### FROZEN READING RULE --- exactly one branch is declared to have fired
- **sparsity corresponds AND B-PRIMARY corresponds** -> mediation **SUPPORTED**.
- **sparsity corresponds AND B-PRIMARY null** -> the Group-B null is **MECHANISTIC**: the
design
  demonstrably detects a real correspondence when one exists, and this one is absent.
- **sparsity does NOT correspond** -> n=21 pairs is **UNDERPOWERED for ANY correspondence
claim**;
  A3B stays descriptive and **no mediation conclusion is drawn in either direction**.

### FROZEN READING RULE for the B-PRIMARY split (this addendum's addition)
- **SCALE separates across families AND ALLOCATION corresponds to JS** -> the mediation chain
is
  supported at **both links**.
- **SCALE separates across families BUT ALLOCATION is null** -> the effect is carried by **
scale
  alone**; within-family JS similarity is then explained **by construction**, and the
mechanism
  claim is **limited to across-family differences**. This is stated plainly, not softened.
- **SCALE does NOT separate across families** -> the Step-0 selected narrative is **
contradicted
  by its own predicted consequence**. Report that and **RE-OPEN the narrative question rather
  than proceeding** --- do not fall back to the excluded narrative (b), which remains barred;
the
  correct response is to declare the mechanism question open.

---

## 3. CORRESPONDENCE METHOD (frozen)

Distance-matrix correspondence against the **A3B JS-divergence matrix**, computed SEPARATELY
for
Group A, B-PRIMARY-SCALE, B-PRIMARY-ALLOCATION, B-CORROBORATIVE, and the sparsity control.

- Arm-level JS distance = mean JS over that arm-pair's constituent checkpoint pairs => a 7x7
  matrix, **21 unique off-diagonal pairs**.
- Feature distance between arms: `|f_i - f_j|` for scalar features; Euclidean distance over
the
  standardized vector for multi-component features (the quantile vector).
- **Statistic:** Pearson correlation between the 21 JS pair-distances and the 21 feature
  pair-distances (Mantel form).
- **Null:** the SAME within-seed restricted permutation as A3B --- permute arm labels **within
 each
  seed independently**, 10,000 draws, recomputing the arm-level matrices and the correlation
each
  time. Report the observed statistic, the null distribution (mean, SD, quantiles) and p.
- **arm5/seed2 excluded** from primary, as in A3B_PREREG. arm2|s1 and arm7|s1 flagged (D2).

**POWER LIMITATION, stated in advance and repeated beside every correspondence result:** with
7
arms there are only 21 unique pairs, and they are not independent (each arm appears in 6 pairs
).
The permutation null respects that dependency, but the design can only detect LARGE
correspondences. A null correspondence therefore does **not** establish absence of mediation
unless the sparsity positive control fired --- which is precisely why the control is pre-
registered.

---

## 4. REGISTERED DIRECTIONAL PREDICTION --- accumulation across rounds

Between-arm separation in Group B **accumulates**: **r1 < r2 < r3**.

The monotonicity itself is tested, not merely final-round magnitude.
- **Statistic:** mean between-arm separation in the Group-B feature at each round; Spearman
  correlation with round index (1,2,3), computed per seed and averaged.
- **Primary test:** monotonicity permutation test --- permute round labels within seed, 10,000
```

```
  draws. **Also reported:** Page's L trend test.
- **Power caveat frozen now:** 3 rounds x 3 seeds is a very short series; Page's L on 3
  conditions has minimal resolution. Reported with that limitation attached.

**Why r1 is the cleanest baseline (recorded in advance):** round 1's collection was **SHARED
across all seven arms**, so every arm saw the identical batch and r1's separation arises from
**exactly one update's worth of credit difference**. A monotone r1<r2<r3 pattern is therefore
evidence **independent of absolute magnitude**.

---

## 5. STATISTIC CORRECTION (frozen)

The previously-contemplated "fraction of zero-variance (all-fail / all-succeed) groups" is
**ill-defined at G=1**: the baseline is the mean over the whole 96-trajectory batch, and at
~40%
win rate no batch is all-fail, so it is trivially 0 for every arm. **Withdrawn.** Replaced by:

- **per-batch SD of A_i**, and
- **fraction of trajectories with |A_i| below a threshold fixed here: 0.5 post-normalization**
  (half a batch SD) and **0.1 pre-normalization**.

**Two degeneracies recorded NOW, before computing, so they are not later mistaken for findings
:**
1. Outcomes are binary and the baseline is the batch mean, so within a batch `A_i` takes
exactly
   **two** values, `(1-p)` and `-p` for win rate p. Hence the "fraction below threshold" is a
   deterministic function of p and carries no information beyond the win rate.
2. **Post-normalization the per-batch SD of A_i is exactly 1.0 by construction.** The per-
batch
   SD is therefore reported **PRE-normalization** (`= sqrt(p(1-p))`), where it is informative.

---

## 6. WRITING-SIDE CONSTRAINT (travels with the data)

The **StepOPSD credit-budget-normalization implication is STRICTLY CONDITIONAL on B-PRIMARY
correspondence holding.** If B-PRIMARY is null, that implication appears in **NO document**.
Recorded here so the condition travels with the result rather than living only in
correspondence.

---

## 7. RULES

All thresholds and reading rules above are frozen BEFORE any result was viewed. D4 is
exploratory
and labelled throughout. **A null B-PRIMARY correspondence is a full deliverable and is
reported
as plainly as a positive one.** ALL_EXPERIMENTS_RECORD entry per step at the step boundary.
Every stage claim opens with a verbatim disk probe (F1/F4). Commits verified by re-reading
`git log`, never by trusting a return value (e38). ntfy at each stage boundary and at the
final
stop.

Signed (execution pre-registration, exploratory): **2026-08-07**, tag `a3b-delta-v1`.

# A3B-Delta  AMENDMENT 1 (Addenda 2 + 3) --- frozen before any matrix was computed

Amends A3B_DELTA_PREREG.md (`a3b-delta-v1`). Tag `a3b-delta-v2`. Exploratory; zero GPU.
A3B_PREREG.md, KILLSHOT_PREREG.md, KS2_PREREG.md, PC/PC2 preregs and both VERDICTs remain
IMMUTABLE and untouched. E031 remains the single source of truth for the A3B wording: the
discrete null and the continuous positive are always stated together.

Status: **nothing had been computed when this was frozen** --- no matrix, no statistic, no
correspondence. The defect in SA was caught before any run (e39).

---

## A. e39 RESOLVED --- allocation is a NEGATIVE CONTROL, not a mediator candidate

The token-count frequency weighting in `a3b-delta-v1` S1 is **WITHDRAWN**. `ks2_train_update.
py:242`
divides by `rmask.sum(-1)`, so token count normalizes OUT: every turn contributes equally
regardless of length and the effective per-token weight is `|w_t.A_i| / ntok_t`. Replaced by
four
**pairing** statistics (weight-to-CONTENT, the only thing shuffling changes):
- Spearman `corr(w_t, turn index)` within trajectory, averaged over trajectories
- Spearman `corr(w_t, ntok_t)`
- Spearman `corr(w_t, per-turn mean sampled-token NLL)`
- ratio of token-weighted to unweighted mean `|w.A|` (scalar summary of w-to-length alignment)
```

```
**ROLE CHANGE (frozen):** allocation is now a **NEGATIVE CONTROL**. Reasoning recorded: A3B
already shows the JS matrix clusters by instrument family, and *within* a family the arms
differ
ONLY in allocation (shuffle/mirror). So within-family JS similarity **IS** the finding that
allocation does not drive behaviour. The allocation statistics are reported **expecting them
to
differ sharply within family while JS does not** --- that contrast is itself the result, and
**no
correspondence test is run or needed to establish it**.

**REQUIRED NUMBER (frozen):** from the 21x21 matrix, the between-arm JS distances for
**arm2-vs-arm4** and **arm3-vs-arm7** specifically, reported beside within-arm (0.0198) and
overall between-arm (0.0520). If within-family between-arm ~= within-arm, the negative-control
argument holds.

**PAPER NOTE (frozen):** `corr(w_t, per-turn mean sampled-token NLL)` is **KS1's T3 mirrored
onto
the training input** --- arm2 predicted high, arm4 near zero. It re-quantifies the fluency-
echo
finding on the quantity the gradient actually consumed, independently of A3B.

---

## B. OPTIMIZER INVARIANCE --- the pure-scale reading is DOWNGRADED

AdamW's update `lr.m/(sqrtv+eps)` is invariant to uniform rescaling of the loss (`g->cg` => `m
->cm`,
`v->c^2v`, `m/sqrtv` unchanged). A uniform difference in credit SCALE across arms is therefore
**largely ABSORBED** and does not translate into effective step size. What survives Adam:
1. **ZERO STRUCTURE** --- `w=0` contributes exactly zero gradient (trainer `:105-106`, and the
   `:135` filter drops those examples entirely), so arms differ in **EFFECTIVE SAMPLE SIZE**;
2. **WITHIN-BATCH RELATIVE WEIGHTING** --- which examples dominate the gradient direction;
Adam
   does not normalize across examples;
3. the **early transient**, plus **gradient clipping**.

**Code facts, verified (B actions 1 and 2):**
```
scripts/ks2_train_update.py:252   gn = torch.nn.utils.clip_grad_norm_(
scripts/ks2_train_update.py:253   (p for p in model.parameters() if p.
requires_grad),
scripts/ks2_train_update.py:254   opt_cfg["grad_clip_norm"])
configs/ks2_training_config.json optimizer: {"name":"AdamW","lr":5e-05,"betas":[0.9,0.999],
                                              "grad_clip_norm":1.0,"weight_decay":0.0,
                                              "schedule":"constant"}
```
- **Clipping EXISTS at norm 1.0**, so scale invariance is **not perfect** and BOTH routes are
  reported. Empirically it **binds rarely and only late**: fraction of optimizer steps with
  grad-norm > 1.0 is **0.000 at r1** for arms 1/2/6, rising to **0.055 (arm1) and 0.096 (arm2)
  at r3**, and **0.000 for arm6 at every round**. So the scale route is available mainly to
the
  dense arms in round 3.
- **weight_decay = 0.0** => AdamW's decoupled-decay channel (which is NOT scale-invariant) is
  **INERT in this design**. Recorded as a channel that exists in principle and is switched off
  here.

**AMENDED R43 NARRATIVE (supersedes the pure-scale wording):**
> **credit zero-structure + within-batch relative weighting -> effective sample size and
gradient
> direction -> policy distribution.**
The pure-scale reading ("different credit scale => different effective step size") is
**DOWNGRADED** to a partial channel operative only where clipping binds. Reason written out:
under
AdamW a uniform rescale of the loss leaves `m/sqrtv` unchanged. The **excluded** narrative (b),
 the
arXiv:2509.02534 S6.3 noise-amplification route, **remains excluded** (R43) and is not revived
 by
this amendment.

**arm6 REFRAME (frozen):** ||lora_B|| 1.37 vs ~4.7 is ~3.4x, while a coverage-based prediction
 would
give roughly 8--14x --- a **compression consistent with Adam absorbing much of the magnitude
difference while zero-structure still costs real gradient**. Therefore: report **predicted-by-
coverage ORDERING vs observed ||Delta W|| ORDERING** (Spearman). **NO falsifiable ratio
prediction is
stated**, because under AdamW no simple analytic map exists.

**Consequent amendment to a3b-delta-v1 S2 branch 3** (Addendum-2 item 4): "SCALE does not
separate
```

```
across families" does **NOT** by itself contradict the Step-0 narrative --- the narrative
predicts
scale differences PROPAGATE, not that arms must differ in credit scale. The **upstream**
quantity
(per-arm mean `|w_t|` and non-zero coverage of the RAW credit) is reported FIRST. The
narrative is
contradicted only if credit scales **genuinely differ across arms AND** `|w_t.A_i|` / `||Delta
 W||`
**fail to track them**.

---

## C. Delta W: MAGNITUDE **AND** DIRECTION, reported separately

New mediator node, measured independently from the 21 stored adapters (zero GPU):
- **Magnitude:** `||lora_B||` and the full `||Delta W|| = ||B.A||` Frobenius, **per target
module and
  aggregate**.
- **Direction:** flatten each `Delta W = B.A`; **21x21 pairwise cosine similarity**; plus
  **principal angles / subspace overlap** between LoRA subspaces.

Rationale recorded: two adapters of equal norm can point in entirely different directions, and
"did the same credit rule push the model the same WAY" is the closest parameter-space analogue
 of
the JS behavioural measure. **Magnitude and direction are orthogonal and can dissociate; they
are
reported separately and never merged.**

- The **same A/B/C + within-seed restricted permutation** analysis is run on the **direction**
  matrix.
- **Partial Mantel** of JS on `||Delta W||` distance, **with direction as a SECOND covariate**.
 Both
  **zero-order and partial** statistics reported with their nulls.
- **FROZEN READING:** signature vanishes after controlling step size -> the effect is
  **scale-mediated and H_instrument is really H_scale --- state it plainly**; residual
signature
  survives -> **allocation/content contributes beyond scale**.

---

## D. STALE ARTIFACT --- corrected, with the two questions kept separate

`runs/ks2/r1/r38_d2_round1_stats.json` asserts `"std_normalized": false` citing
`ks2_train_update.py:96` (a **v1.3** line). Under **v1.4** the advantage **IS** std-normalized
(A_i alone, R43). Two things are corrected **separately and must stay separate**:
1. **The annotation was stale --- factual, fixed.**
2. **Whether Dark Room's mechanisms now apply is a SEPARATE question**, argued on its merits,
not
   inherited from (1). Our divisor is computed **from outcomes alone and is credit-blind**,
and at
   **G=1 there are no groups** for the all-fail-group proposition to range over.
**"The annotation was wrong" must NOT expand into "the conclusion reverses" without that
argument.** Feeds V4.

---

## E. ENTROPY --- corroborative only; availability accepted

Full-distribution policy entropy was **never logged** and **stays unrun** (would need GPU).
The
**R38 sampled-token NLL proxy** is reconstructed from `gen_logprobs` for **r2/r3** per arm; **
r1 is
shared collection and therefore arm-independent**. Labelled exactly as R38 did: *mean -logprob
 of
sampled tokens, a realized cross-entropy proxy, NOT full-distribution entropy.*

**B-CORROBORATIVE standing reminder (frozen):** grad-norm and NLL-proxy trajectories are
**OUTCOMES of training, not input-side quantities**; they may correlate with the final policy
distribution because both are produced by the same process. They remain **corroborative**. The
mediation claim rests on **`|w_t.A_i|` and `||Delta W||`**.

---

Signed (execution pre-registration amendment, exploratory): **2026-08-07**, tag `a3b-delta-v2`.


# A3B-Delta  ADDENDUM 3 (ideation packet 2026-08-07) --- dose-matched contrast

Amends A3B_DELTA_PREREG.md (`a3b-delta-v1`) and A3B_DELTA_AMENDMENT.md (`a3b-delta-v2`).
Tag `a3b-delta-v3`. Exploratory. Written **before any Task-1 computation was run**.
E031 and its five travelling conditions are unchanged and remain the single source of truth;
```

```
**clause (e) is mandatory whenever the JS clustering is cited.**

---

## 0. POST-HOC DISCLOSURE --- READ THIS BEFORE ANY NUMBER BELOW

> **THIS IS A POST-HOC EQUIVALENCE ANALYSIS, NOT A PRE-REGISTERED CONFIRMATORY TEST.**
> The three dose-matched distances (0.0272 / 0.0124 / 0.0281) and the within-arm floor
> (0.0198) were reported in E034 and appear in the ideation packet that commissioned this
> task. They were therefore **visible before any equivalence margin could be chosen**. No
> margin selected now can be called pre-registered, and this analysis must never be presented
> as one.
>
> **WHAT REMAINS VALID:** the restricted-permutation p-values. Their null is generated by
> permutation with the observed statistic held fixed, so their calibration does not depend on
> when the observed value was seen.
>
> **WHAT IS COMPROMISED:** the choice of equivalence margin, and any pass/fail verdict that
> rests on it.
>
> **MITIGATION:** the margin is fixed by a rule statable **independently of the observed
> dose-matched values** --- see S2 --- and the primary read-out is the **continuous position
> measure** of S4, which needs no margin at all.

This disclosure is reproduced verbatim in every report of this result.

---

## 1. CORRECTION 1 --- arm5 is NOT magnitude-multiset-identical

Mirroring `v -> 2m - v` preserves the **sum, mean, SD and range** (the `Sum |2m-v| = Sum |v|`
identity established in E033, valid because no mirrored value goes negative: 0/3563 turns,
min 0.7917). It does **NOT** preserve the magnitude **multiset** --- the distribution is
**REFLECTED**, so **skew flips sign**.

Registered statements, to be used verbatim:
- **arm2<->arm4 and arm3<->arm7**: dose-matched **AND magnitude-multiset-identical** (a
  permutation of the same values), **allocation destroyed**.
- **arm2<->arm5**: dose-matched, **first- and second-moment matched**, magnitude distribution
  **REFLECTED (skew inverted)**, and the credit--fluency correlation **sign-flipped**
  (-0.2890 -> +0.2911, E034).

**The phrase "magnitude-multiset-identical" must never be applied to arm5** in any document.

---

## 2. CORRECTION 2 --- the equivalence margin, and the rule that fixes it

**MARGIN RULE, statable independently of the observed dose-matched distances:**
> The equivalence margin is the **95th percentile of the within-arm across-seed JS distance
> distribution** (pair class A: same arm, different seeds; n=19 pairs, arm5|s2 excluded).

**RATIONALE:** class A is the design's own estimate of *"how far apart two checkpoints get
when the credit rule is held constant and only the seed differs"* --- i.e. seed noise. Its 95
th
percentile is the natural upper edge of that noise. The rule references only class-A pairs,
which are **not** the quantity under test, so the margin is not chosen by looking at the
dose-matched values. It was, however, chosen *after* the class-A distribution was itself
visible, which is why the post-hoc label in S0 still applies and is not lifted by this rule.

**EQUIVALENCE VERDICT:** a dose-matched pair is *equivalent within the stated margin* iff its
distance <= margin. Reported as a binary alongside --- never in place of --- S4's position
measure.

---

## 3. CORRECTION 3 --- CRN asymmetry runs the OPPOSITE way here and is controlled

Round-1 collection was **SHARED across all seven arms**. Therefore, for dose-matched pairs:
- **(a) SAME-seed cross-arm** (e.g. arm2/s0 <-> arm4/s0): **shares the round-1 batch**,
differs
  in credit. Extra data sharing can pull such a pair toward the within-arm floor for reasons
  that have nothing to do with credit content being inert.
- **(b) CROSS-seed cross-arm** (e.g. arm2/s0 <-> arm4/s1): **shares no data**, differs in
  credit. This matches the within-arm floor's own data-sharing status (class A pairs also
  share no data), so it is the **clean contrast and the PRIMARY quantity** for the
  equivalence framing.

**BOTH are computed and reported, and every reported number states whether it is (a) or (b).
If (a) and (b) diverge, that divergence is itself a result and is reported as one.**
```

```
Note the direction of the bias this controls: in A3B's original framing the CRN asymmetry made
`mean(A) > mean(B)` a *strong* bar; here it runs the other way, making same-seed pairs look
*artificially close*. Same asymmetry, opposite sign, because the comparison target changed.

---

## 4. CORRECTION 4 --- continuous position measure (the primary read-out)

For each dose-matched pair:

> **position = (d_pair - d_within) / (d_between - d_within)**

where `d_within` = mean class-A (within-arm, across-seed) JS and `d_between` = mean overall
between-arm JS. 0% = indistinguishable from seed noise; 100% = a typical between-arm pair.
**Reported with a CI** (bootstrap over the constituent checkpoint pairs, 10,000 resamples).

This is **more informative than a pass/fail** and **does not depend on the compromised
margin**, so it is the primary read-out. Computed values are reported; no anticipated value is
recorded here, precisely because the inputs were already visible.

---

## 5. CORRECTION OF AN EARLIER CLAIM OF MINE --- "at or below within-arm" was WRONG

E034 and PROGRESS B27 state that the within-family cross-arm distances are *"at or BELOW the
within-arm level (0.0198)"*. **That is false and the error is mine.** Two of the three are
**ABOVE** it:

| dose-matched pair | JS | vs within-arm floor 0.0198 |
|---|---|---|
| arm3<->arm7 | 0.0124 | **below** |
| arm2<->arm4 | 0.0272 | **ABOVE** |
| arm2<->arm5 | 0.0281 | **ABOVE** |

The supportable claim is: **"above the within-arm floor, but far below the overall
between-arm level (0.0515)"**. The original wording overstated the negative control. It
originated in my own report and was passed through unchecked; both sites are corrected in
place with the correction noted, and the corrected form is what every later document uses.

---

## 6. FROZEN READING (with the post-hoc label permanently attached)

- **Equivalent within the stated margin** -> the training-layer sentence is supported, and is
  written **only** with the S0 post-hoc disclosure attached.
- **Not equivalent** -> report the distances plainly as **"above seed noise but far below
  between-arm"**, give the continuous position measure of S4, and **DO NOT write the
  sentence.**

**POWER:** there are only **3 dose-matched contrasts** available in the entire design
(arm2<->arm4, arm3<->arm7, arm2<->arm5). This is reported beside every verdict. With n=3
contrasts
no equivalence claim can be strong, and the analysis cannot distinguish small real differences
from noise.

---

Signed (exploratory addendum, written before any Task-1 computation): **2026-08-07**,
tag `a3b-delta-v3`.

# A3B-Delta  ADDENDUM 4 --- the r1 dose-matched contrast

Amends `a3b-delta-v1` / `-v2` / `-v3`. Tag `a3b-delta-v4`. Exploratory throughout.
**Written before any r1 agreement, JS or position number was computed.**
E031 and its five travelling conditions unchanged; clause (e) mandatory wherever the JS
clustering is cited.

## 0. POST-HOC DISCLOSURE --- stronger here than in Task 1, and stated first

> **THIS IS AN EXPLORATORY ANALYSIS WHOSE THRESHOLDS WERE SET KNOWING THE r3 RESULT.** I have
> already seen the r3 dose-matched numbers (+27.0% / +17.8% / -23.2%, E036) and the r3
> perturbation strengths (E040). Nothing computed under this addendum can be called
> pre-registered in the confirmatory sense, and it must never be presented as such.
>
> **WHAT REMAINS VALID:** the restricted-permutation p-values --- their null is generated by
> permutation with the observed statistic held fixed, so calibration does not depend on when
> anything was seen. The **margin-free position measure** likewise needs no threshold.
>
> **WHAT IS COMPROMISED:** any equivalence margin, and the choice of "materially differ" in
> the frozen reading of S4. S4 therefore fixes that phrase numerically **now**, before the
> computation, rather than after.
```

```
## 1. WHY r1 IS THE CLEANER CONTRAST

Task 1's three dose-matched contrasts were computed on **r3** adapters. By r3 each arm had
collected its own r2 and r3 batches, so the arms differ in **training data** as well as credit
rule --- a confound Task 1 could not remove.

**Round 1 collection was SHARED across all seven arms**: the same 288 trajectories (96 tasks x
3 seeds), the same dose, exactly **one** update, and **credit is the only difference**.
Confirmed on disk: every r1 update log records
`store = /workspace/ks1/runs/ks2/r1/shared/s<seed>/trajectories`. The 21 r1 adapters exist and
are sha256-recorded (Task 5, `runs/pc2/a3b_intermediate_adapter_check.json`, 42/42 present).

## 2. PROTOCOL --- identical to A3B Stage 2 in every respect except the checkpoint round

- **Same state bank**, verified by hash before scoring:
  `sha256 = 8ec0924492f276e55fc8412e0c0ac048c7e4082f75144516b0ac4009afe22a36`
  (`runs/pc2/a3b_state_bank.jsonl`, 500 states, base-model greedy over the frozen 128 held-out
  tasks). The hash is re-verified at run time and pasted in the record.
- Greedy, temperature 0, `TOPK = 20` (vLLM's cap; JS on the top-20 support), same prompt
  template, same `parse_action` canonicalization via the harness's e17 parser.
- **arm5|s2 excluded** from primary, as in `a3b-prereg-v1`.
- Metrics: canonical action agreement, JS divergence, class A/B/C, restricted within-seed
  permutation null (10,000 draws), and the **margin-free position measure**
  `(d_pair - d_within) / (d_between - d_within)` with bootstrap CI.
- **CROSS-SEED dose-matched pairs are PRIMARY** (they share no data, matching the within-arm
  floor's status); same-seed reported beside them, per `a3b-delta-v3` S3.

## 3. COST GATE, BEFORE LAUNCH

Measure the per-checkpoint scoring rate on the **first 3 adapters**, project the total for 21,
and **post the projection before the full run**. Expected 1--2 GPU-h. **If the projection
exceeds 3 GPU-h, STOP and report instead of proceeding.** r2 is added only if it is trivial
once the server is already up, and is subject to the same combined ceiling.

## 4. FROZEN READING --- with "materially differ" fixed numerically now

- **If the r1 contrasts differ materially from r3** -> the r3 numbers are **confounded by
  collection divergence**, and **r1 becomes the reported quantity**.
- **If they agree** -> the r3 result is **strengthened**.
- **Either way, BOTH are reported.**

**"Materially differ" is fixed here, before computing, as: any dose-matched contrast whose
position measure moves by more than 15 percentage points between r1 and r3, OR any contrast
that changes sign.** 15 pp is chosen as roughly half the observed r3 spread between the
extreme contrasts (+27.0% to -23.2% ~= 50 pp), so it is a difference large enough to change
which contrast is the largest --- not a threshold tuned to the r1 answer, which is unseen.

## 4b. CONTINUATION GATE FOR THE CROSS-FAMILY RUN --- FROZEN 2026-08-08, BEFORE THE NUMBER
EXISTS

**Frozen by [AUTHOR-1] while Item 3 was still running and before any virgin-turn completion
rate or
MDE had been computed. Written into this addendum at that moment; the timestamp is the commit
.**

> **The cross-family run continues IF AND ONLY IF the projected MDE for T3's
> per-trajectory-Spearman median --- at the measured virgin-turn completion rate, after
applying
> KS1's v3.2 exclusion (trajectories retaining >= 4 complete turns) --- is <= 0.50.**

**RATIONALE, recorded with the threshold:** KS1's observed T3 is median **+0.752, CI
[0.647, 0.793]**. An MDE above 0.50 means we could not detect an effect meaningfully smaller
than KS1's own **lower CI bound**, so a replication attempt could not distinguish *"does not
replicate"* from *"underpowered"*. In that case option (a) buys nothing and option (d) is the
honest call.

- **MDE <= 0.50 -> RESUME** the replay to completion (within the signed cross-family run; the
$40
  gate is satisfied by the re-projection of $11.38--$22.84), then pipeline steps 4--5: logprob
  dumps, then T3 analysis. **T3 is the sole registered primary endpoint --- once it lands the
run
  has delivered its purpose.** The 72B judge stage is **NOT** run: it serves only a
  pre-labelled underpowered descriptive output and costs a 145 GB staging operation with a
  history of incidents. Skipped **by design**, not by failure.
- **MDE > 0.50 -> STOP** Task 3, release the GPU, record the projection and its arithmetic,
and
  queue option (d) with the numbers. **Do not resume.**

**Additional hard stop, frozen with the gate:** if the virgin-turn completion rate turns out
materially worse than the rate this gate was evaluated on, **STOP and report** --- do not
```

```
re-project mid-run to keep going.

## 5. WHAT THIS CANNOT SETTLE

The **strength-vs-structure hypothesis** (DOSE_CONFOUND.md) gains points here --- three
contrasts
per round, with realized perturbation strength computed **per round per contrast**. Going from
3 to 6 or 9 points weakens the "any two-factor model fits three points exactly" objection but
does **not** remove it: the transforms are the same three at every round, so the added points
are **not independent** draws over transform types. The hypothesis stays a hypothesis.

Signed (exploratory addendum, written before any r1 number was computed): **2026-08-08**,
tag `a3b-delta-v4`.
```

R_ROUTE (the confidence router). Chinese-major source; byte-exact in the supplementary pack. Structure, for the reader: frozen under tag `1814166...`; registered as EXPLORATORY; the single main result is threshold multiplier 1.0, with the six-multiplier grid registered as sensitivity only; recall and both cost granularities are required to be reported together.

PC3 (the substrate-repair method line). No standalone PC3 pre-registration file exists in the writing pack — roots searched before this absence claim, per the STOP-for-absent rule: `/mnt/project/*.md` (all 52 knowledge files), `/home/claude/paper/**`. The registered criterion (S6: same-type divergence ⇒ no further configurations) lives inside the PC3 verdict document, which is Chinese-major and ships byte-exact: Its boundary table is rendered in English in Section K.4 below.

K.2 THE FOUR FROZEN PROMPT FILES

The HCAPO template below is extracted verbatim from that paper's own appendix (Appendix C.2.1), not reconstructed, and is reproduced here so that the scoring instrument is reproducible from this paper alone. Two byte-differing copies of it were previously carried — a working copy and the original-from-bundle copy — and that redundancy is now merged to the single copy shown.

```
Role: Task Execution Evaluation Judge
Your core responsibility is to thoroughly and precisely evaluate a multi-turn episode in which
 an agent
interacts with a household environment (ALFWorld) through text actions. You must carefully
read the episode
to pinpoint where the agent's decisions lead to deviations from successful task completion.

Information Provided for Your Evaluation
You will be given three key pieces of information to guide your assessment:
1. Task Instruction: The household task the agent must accomplish (e.g., "put a clean mug in
the coffeemachine").
2. Episode Trajectory: The detailed record of the multi-turn episode. Each turn shows the
agent's action and the
environment's observation in response. Each turn is preceded by a label like [Turn N]. Actions
 are text commands
(e.g., "go to countertop 1", "take mug 1 from countertop 1", "clean mug 1 with sinkbasin 1", "
open fridge 1",
"heat egg 1 with microwave 1", "put mug 1 in/on coffeemachine 1").
3. Final Outcome: Whether the episode ended in task success or failure.

Evaluation Process
Deviations from successful task completion typically arise due to:
- The agent failing to gather sufficient information about the environment (e.g., not finding
the required object).
- Incorrect reasoning or understanding by the agent based on the observations it received.
- The agent executing actions that are invalid, redundant, or that move the task away from
completion.
Pay exceptionally close attention to state-changing operations (take, put, open/close, clean,
heat, cool, toggle):
these directly determine task success and are critical evaluation points and frequent sources
of error. Three kinds
of error are possible with state-changing operations:
(1) The agent might perform the operation on the wrong object or at the wrong location.
(2) The agent performs an unnecessary state-changing operation that should never be executed.
(3) The agent did not perform a state-changing operation that the task requires.
If any of the three cases above occurs, you need to carefully read the episode and identify
the turns where the
agent deviates from the path to task completion.
```

```
For each turn in the episode (identified by the [Turn N] tag), you should evaluate whether the
 agent's action in
that turn is the primary cause of a deviation from task completion. Assign a score for each
turn. You have three
kinds of score to assign:
- If the turn is correct, assign a score of 1.
- If the turn is the primary reason for a deviation, assign a score of -1. This can only be
assigned to at most one of
the episode turns if deviation is found.
- If the turn has issue (e.g., an inadmissible action or a redundant action that wastes steps),
 assign a score of 0.

Your Response Format
You must first conduct your evaluation process within a <think></think> block.
After completing your thinking process, you must output only a single JSON object. No other
text, commentary, or
explanation should be included outside of the JSON block. The JSON object must adhere strictly
 to the following format,
including all turn's scores from score_0 up to score_n (where n is the total number of turns).
{
"score_0": <turn 0's score>,
"score_1": <turn 1's score>,
// ... (include all turns up to 'n')
"score_n": <turn n's score>
}
```

Judge rubric v1. Role note: the 5-point rubric variant is the R39 exploratory rubric-format prompt (PC_PREREG Step 3); it was never used in any reported number and ships for provenance completeness only.

```
Role: Task Execution Evaluation Judge
Your core responsibility is to thoroughly and precisely evaluate a multi-turn episode in which
 an agent
interacts with a household environment (ALFWorld) through text actions. You must carefully
read the episode
to pinpoint where the agent's decisions lead to deviations from successful task completion.

Information Provided for Your Evaluation
You will be given three key pieces of information to guide your assessment:
1. Task Instruction: The household task the agent must accomplish (e.g., "put a clean mug in
the coffeemachine").
2. Episode Trajectory: The detailed record of the multi-turn episode. Each turn shows the
agent's action and the
environment's observation in response. Each turn is preceded by a label like [Turn N]. Actions
 are text commands
(e.g., "go to countertop 1", "take mug 1 from countertop 1", "clean mug 1 with sinkbasin 1", "
open fridge 1",
"heat egg 1 with microwave 1", "put mug 1 in/on coffeemachine 1").
3. Final Outcome: Whether the episode ended in task success or failure.

Evaluation Process
Deviations from successful task completion typically arise due to:
- The agent failing to gather sufficient information about the environment (e.g., not finding
the required object).
- Incorrect reasoning or understanding by the agent based on the observations it received.
- The agent executing actions that are invalid, redundant, or that move the task away from
completion.
Pay exceptionally close attention to state-changing operations (take, put, open/close, clean,
heat, cool, toggle):
these directly determine task success and are critical evaluation points and frequent sources
of error. Three kinds
of error are possible with state-changing operations:
(1) The agent might perform the operation on the wrong object or at the wrong location.
(2) The agent performs an unnecessary state-changing operation that should never be executed.
(3) The agent did not perform a state-changing operation that the task requires.
If any of the three cases above occurs, you need to carefully read the episode and identify
the turns where the
agent deviates from the path to task completion.
For each turn in the episode (identified by the [Turn N] tag), you should evaluate whether the
 agent's action in
that turn is the primary cause of a deviation from task completion. Assign a score for each
turn. You have a five-point rubric to assign:
- 5 = clearly advances the task (a necessary or efficient step toward completion).
- 4 = useful but not necessary (reasonable progress, mildly suboptimal ordering or a minor
detour).
- 3 = neutral (information-gathering with no progress and no cost, or an unavoidable step).
```

```
- 2 = wasteful (a redundant action, or one that costs a step without gathering useful
information).
- 1 = harmful (an inadmissible action, or one that actively moves away from task completion).
Assign exactly one integer from 1 to 5 to every turn. Unlike a coarse three-level scale, you
SHOULD use the
full range: reserve 5 and 1 for clear cases and use 4/3/2 to express the ordinary gradations
between them.
There is no limit on how many turns may receive any particular score.

Your Response Format
You must first conduct your evaluation process within a <think></think> block.
After completing your thinking process, you must output only a single JSON object. No other
text, commentary, or
explanation should be included outside of the JSON block. The JSON object must adhere strictly
 to the following format,
including all turn's scores from score_0 up to score_n (where n is the total number of turns;
every score is an integer from 1 to 5).
{
"score_0": <turn 0's score, an integer 1-5>,
"score_1": <turn 1's score>,
// ... (include all turns up to 'n')
"score_n": <turn n's score>
}


# Verbatim extraction: TARL (arXiv:2509.14480), Appendix C "LLM Judge Setup for
# Turn-Level Evaluation". Extracted 2026-07-31 via pdftotext from
# https://arxiv.org/pdf/2509.14480 (see PROVENANCE Sb). Lines below are the
# published judge prompt, unmodified (pdftotext ligature artifacts preserved).

Role: Task Execution Evaluation Judge
Your core responsibility is to thoroughly and precisely evaluate multi-turn conversations
between a user and an
agent. You must carefully read each conversation to pinpoint where the agent's decisions lead
to deviations from the
ground-truth function-call trajectories.

Information Provided for Your Evaluation
You will be given four key pieces of information to guide your assessment:
1. Policy: This document outlines the strict rules the agent must adhere to when making tool
calls. If an agent's
action violates this policy, you must immediately halt its current action and instruct it to
reconsider and correct
its approach.
2. Task Instruction: This is the specific instruction provided to the user. The user's
requests and responses should
always align with this instruction. The agent does not have access to this instruction.
3. Ground-Truth Function Call Trajectories: This serves as the definitive standard for
assessing the accuracy of the
agent's tool calls.
- The agent doesn't need to follow the exact order of this trajectory.
- It's acceptable for the agent to call information-gathering functions (e.g.,
get_order_details) multiple times,
but the agent's write operation (modifying, exchanging, returning, or canceling orders) needs
to match exactly
with the ground-truth function calls.
4. Conversation Trajectories: This provides the detailed record of the multi-turn conversation
 between the user and
the agent. You will use these conversations to identify executed tools and evaluate the
correctness of the agent's
processing of results. Each agent's reasoning and action within a turn will be preceded by a
label like [Turn N].

Evaluation Process
Deviations from the ground truth typically arise due to:
- The agent failing to gather sufficient or correct information, either through function calls
 or by asking the user.
- Incorrect reasoning or understanding by the agent based on the results of tool execution.
- The agent not following policy, resulting in wrong execution of tools.
Pay exceptionally close attention to operations involving modifying, exchanging, returning, or
 canceling orders. The
agent's calling for these function should match exactly with the ground-truth. These are
critical evaluation points and
frequent sources of error. Three kinds of error are possible with write operation:
(1) The agent might call the function with wrong arguments that do not match with ground-truth.

(2) The agent calls unnecessary write operation that should never be called.
(3) The agent did not call the write operation which is listed in the ground-truth.
If any of the three cases above occurs, you need to carefully read the conversation and
identify the turns where the
agent deviates from the ground-truth.
```

```
For each turn in the conversation (identified by the [Turn N] tag), you should evaluate
whether agent's reasoning and
action in that turn is the primary cause of a deviation from the ground-truth function call.
Assign a score for each
turn. You have three kinds of score to assign:
- If the turn is correct, assign a score of 1.
- If the turn is the primary reason for a deviation, assign a score of -1. This can only be
assigned to at most one of
the conversation turns if deviation is found.
- If the turn has issue (e.g., not following the policy or function call formats), assign a
score of 0.

Your Response Format
You must first conduct your evaluation process within a <think></think> block.
After completing your thinking process, you must output only a single JSON object. No other
text, commentary, or
explanation should be included outside of the JSON block. The JSON object must adhere strictly
 to the following format,
including all turn's scores from score_0 up to score_n (where n is the total number of turns).
{
"score_0": <turn 0's score>,
"score_1": <turn 1's score>,
// ... (include all turns up to 'n')
"score_n": <turn n's score>
}
```

```
# Verbatim extraction: HCAPO (arXiv:2603.08754), Appendix C.2.1 "ALFWorld Agent
# Training Template". Extracted 2026-07-31 via pdftotext (see PROVENANCE Se).
# Placeholders are Python .format() slots. History length = 2 for ALFWorld
# (Appendix C.2 text). Line-continuation arrows from the paper's listing removed.

You are an expert agent operating in the ALFRED embodied Environment.
Your task is to: {task_description}.
Prior to this step, you have already taken {step_count} step(s). Below are the most recent {
history_length} observations and the corresponding actions you took:
{action_history}
You are now at step {current_step} and your current observation is: {current_observation}.
Your admissible actions of the current situation are: [{admissible_actions}].
Now it's your turn to take an action. You should first reason step-by-step about the current
situation. This reasoning process MUST be enclosed within <think> </think> tags. Once you've
finished your reasoning, you should choose an admissible action for current step and present
it within <action> </action> tags.
```

K.3 PC/PC2 DIAGNOSTICS

The PC2 diagnostics' registered layer is carried by the Δ addenda above (perturbation strength, position monotonicity, the 500-state bank); their measured outputs enter the paper as Tier-1 ledger rows (Section 7; Figure 2); the raw record entries are part of the append-only experiment record, cited under the Tier-3 rule only inside Section 8 and Appendix A, and released in full in the replay archive.

K.4 THE PC3 BOUNDARY TABLE (ENGLISH RENDERING, FAITHFULNESS-CHECKED)

Rendered from the verdict's own "what this ruling supports / does not support" section; the element-by-element faithfulness matrix is in the assembly record.

Supports. (1) On the G=4-corrected substrate, the v1.4 stabiliser family (including its learning-rate and grad-clip variants) cannot stably train an outcome-only relative-advantage signal; the divergence lineage is "pre-clip gradient escapes first, KL breaches after," insensitive to both orthogonal knobs (step size $5\times10^{-5}\rightarrow2\times10^{-5}$; clip threshold $1.0\rightarrow0.2$). (2) The batch-level precursor — a gradient spike one to two batches before the KL breach — is reproducible and has warning value (the PC3-e60 reporting line fired ahead of the breach all three times).

Does not support (misreading guard). `[X]` "outcome-only is ineffective" — none of the three configurations reached evaluation; no held-out effect value exists. `[X]` "the G=4 correction is

useless or harmful" — G=4 made the advantage signal genuinely stronger (C0 verdict); exposing the stabiliser is not causing the pathology, and under KS2's weak G=1 signal the same stabiliser ran 63 updates without divergence, corroborating that the mismatch is the combination *strong signal* × *v1.4*. `[X]` "substrate necrosis" — each configuration trained and improved during its healthy phase (C2 win rate touched 0.604; loss declined stably; KL smooth); the collapse is a tail event, not whole-run failure. `[X]` "missing or failed gradient clipping" — clipping was active at every step from harness construction (C2 tightened to 0.2 with clip-fraction 1.00) and the divergence persisted; the lesion is directional, not magnitude.

Namespace discipline. PC3-prefixed numbers stay prefixed; incident ids (lowercase e52–e58) and experiment ids (uppercase E052–E058) are different series that collide numerically and are never resolved into one another.

Supersession note (ruled 2026-08-18). The §8 placement of the e52 instance mandated in the split text above was superseded by the page-ladder compression; the mandate is retained verbatim, the supersession recorded here. The split mandate is an editorial arrangement of the writing stage, not a registered criterion; §8's omission is therefore recorded as a supersession, not a breach. The instance's factual content is carried in full by Appendix A.

K.5 THE INCIDENT CHAIN

The ruled main-text/appendix split, verbatim.

> MAIN TEXT: e58 (the pack was only ever checked against its own manifest, never the repository — an IDENTITY failure) together with the Finding-6 prediction arc as the central worked example: the taxonomy predicted that the next failure would land on IDENTITY or DIAGNOSABILITY because those two defences had never met a known-positive; it landed on IDENTITY, and the prediction was recorded before the diff was run. Plus e52 (a generator encoding a configuration failure as `[UNVERIFIED]`, the project's marker for a genuinely untraceable number — the DIAGNOSABILITY instance). APPENDIX: everything else: the four-dimension table and its worked examples, the meta-requirement, e53–e57, the chat-layer channel (4 relayed-assertion instances + 1 unsourced reasoning error), the PROVENANCE v1 swallowed-appends sub-case, and the incident-supplied-validation limitation. Why this split: the main text carries the one thing a reviewer cannot get from a list of incidents — a taxonomy that made a falsifiable prediction and was confirmed by its own next failure — plus the single cleanest instance (e52). DIAGNOSABILITY remains an outstanding, live prediction and should be stated as such in the main text, not quietly dropped.

Identifier note (e52). The id e52 in this ruling text refers to the generator incident that encoded a configuration failure as `[UNVERIFIED]` — the DIAGNOSABILITY instance; in the frozen incident chain (Appendix A), e52 denotes the heartbeat NUL-corruption incident (persistence); the numeric collision is historical and retained under the verbatim policy.

Worked example (e57): an unvalidated verifier is worse than none — it manufactures confidence. `grep -c $'\0' FILE` as a NUL counter: a NUL cannot be passed to `grep` as a pattern at all — `execve` argv entries are NUL-terminated — so the argument collapses to a zero-length pattern matching every line. On a known-negative three-line file with no NULs, the true count is 0 and the check reports 3. A missing check leaves a known hole; a broken check reports a clean result over the same hole, and that clean result is then cited.

Identifier note. The id e57 in this worked example refers to the unvalidated NUL-counter verifier instance; in the frozen incident chain (Appendix A), e57 denotes the nominal-versus-delivered-dose incident; the numeric collision is historical and retained under the verbatim policy.

e53–e57, full narratives. Retrieved from the governance decision record (`DECISIONS.md:1519--1721`, held on the collection host; the errata log's E5 names it as their source) and shipped verbatim. The source is Chinese-major, so it takes rendering tier (ii): the byte-exact file ships in the supplementary pack as `K5_verbatim/01_e53-e57_incident_narratives.md`, sha256

3d08f01d3b83884fda64824db1a8c3faa2a66783386a82a3c26a3d904842bffe, 18,333 B, and is authoritative; the structured account of what each incident did to the record is given in English in Appendix A ("The week the taxonomy was tested"), labelled as a rendering.

R-3: why no $\|\Delta W\|$-matched comparison was run (verbatim). The ruling behind Section 6's cancellation sentence, English-major, rendering tier (i); sha256 c79df07b090869e7a8719316f9a0433d4bdab46a5353926413d238e7fa1e82e2, 3,269 B.

```
# RETRIEVAL 2 -- R-3 ruling: the three real reasons for NOT running the matched-||dW||
comparison (VERBATIM)
# SOURCE: pod /workspace/ks1/DECISIONS.md, entry (R46) item (1)
# SOURCE sha256: 57ed00135455073ffeafb28bbdda02b609df64e27d18999dfb7153625697034e
# LINES: 1369-1394
# Retrieved 2026-08-17. Verbatim -- no edits.
-----BEGIN VERBATIM-----
            (R46) [2026-08-08, [AUTHOR-1] FINAL PACKET -- ratified corrections]
                (1) ** ||dW||-MATCHED CONFIRMATION: PERMANENTLY CANCELLED. ** Added to the
                    permanently-barred list beside the 5-seed extension, CoT-preserving
                    distillation and R40. The three reasons, IN THE RULED ORDER -- science
                    first, resources last, so the answer holds if a reviewer asks:
                      (i)  THE PARTIAL CORRELATION ALREADY ANSWERS THE QUESTION. The
                           instrument-family signature vanishes entirely when ||dW|| is
                           controlled (+0.078, p=0.774) while the positive control fires
                           (+0.912, p=0.0001). A matched design would re-answer a question
                           already answered.
                      (ii) THE DESIGN HAS AN INHERENT FLAW. ||dW|| overlap across arms exists
                           only ACROSS ROUNDS -- a dense arm's r1 against a sparse arm's r3 --
                           so caliper matching NECESSARILY introduces a round confound, and
                           round covaries with training volume. The matched comparison could
                           not be clean even if run.
                      (iii) COST ($25-30, ~3 GPU-h) against a calendar in which writing has
                           not started.
                    CORRECTION TO THE INSTRUCTION, stated rather than silently complied with:
                    [AUTHOR-1] asked that "the E040-based reason" be removed from the record.
                    ** NO SUCH REASON EXISTS. ** B28's recorded reason has always been the
                    partial-correlation adequacy plus the calendar (PROGRESS B28, "WHY NOT
                    NOW"), and FINAL_STATE section 3 cites E033's partial correlation. E040
                    (shuffle-control strength is format-dependent) is nowhere given as a
                    reason for or against the matched design -- and [AUTHOR-1] is right that
it
                    would be orthogonal if it were. Nothing was removed because nothing was
                    there; the three ruled reasons above now replace the two-reason version.
-----END VERBATIM-----

# CROSS-REFERENCE in ALL_EXPERIMENTS_RECORD.md (sha256
eea4fcb1eabbe1f596dd97c85fb9df41583b7fd74951d32c1fbbdb6ce11b1bc8), lines 1176-1178.
# NOTE: AER carries only the SUPERSEDED pointer ('specified and UNRUN, PROGRESS B28, [AUTHOR
-1] ruling 2026-08-07').
# The three ruled reasons exist ONLY in DECISIONS R46(1) above. No competing three-reason text
 found in AER.
-----BEGIN VERBATIM-----
>>> robust to dropping arm6 (r=+0.9250, p=0.0004) and arm6+arm1 (r=+0.9274, p=0.0013). The
>>> partial-correlation design cannot upgrade this to a matched comparison; the ||dW||-matched
>>> confirmation that would is specified and UNRUN (PROGRESS B28, [AUTHOR-1] ruling
2026-08-07).
-----END VERBATIM-----
```

The S3-gate withheld-then-released ruling, with its silent diff (verbatim). The decision-record entry behind Section 8's gate paragraph and its two releases; English-major, rendering tier (i); sha256 f537e7d1f098af48c60857bd3d3602ab9645fedf9bf7d6f14be53d3cc30377ed, 11,522 B.

```
# RETRIEVAL 3 -- DECISIONS main entry: S3-gate withheld -> released ruling (VERBATIM) + silent
 diff
# SOURCE: pod /workspace/ks1/DECISIONS.md, entry "(REDO cap-trip salvage ruling) [2026-08-11,
[AUTHOR-1]]"
# SOURCE sha256: 57ed00135455073ffeafb28bbdda02b609df64e27d18999dfb7153625697034e
# LINES: 2205-2232
# Retrieved 2026-08-17. Archival copy per G17. Verbatim -- no edits.
-----BEGIN VERBATIM (DECISIONS main entry)-----
    (REDO cap-trip salvage ruling) [2026-08-11, [AUTHOR-1]]
        (1) Branch (b) AUTHORIZED: PARTIAL-COVERAGE analysis of the redo replay set,
            zero GPU, frozen callers only. Every new value carries the
            PARTIAL-COVERAGE label and its n (stated as n=x/27 per the task brief,
            with the 28-file store count co-stated as before).
        (2) ADDITIONAL DISCLOSURE OBLIGATION: REDO_REPORT.md gains a
```

```
            missingness-nonrandomness section -- distribution of the 142 missing
            turns over trajectory length; per-bucket (B1/B2/B3, frozen
            analysis/fidelity.py BUCKETS) coverage rates; any bucket with
            significantly imbalanced coverage has its bucket-level numbers marked
            "insufficient coverage".
        (3) TAIL 142 TURNS: NOT re-run, per the E043 precedent -- the deterministic
            bad-alternative backlog converts at ~1% (E042: retry-backlog 101 turns
            -> 1 complete, 1.0%), so re-attempting spends GPU without adding n.
            Recorded here as the reason of record.
        (4) tmp+rename FOR summary.json WRITES: entered as a TO-DO requiring
            signature; implemented only if a REDO-class replay is ever restarted.
            (The non-atomic write cost one summary this run: task024/t049,
            REDO_MANIFEST.json.)
        (5) 20/27 THRESHOLD SEMANTIC MISMATCH, recorded honestly: the overnight
            task brief's ">=20/27 complete trajectories" salvage criterion is
            TRAJECTORY-level while the frozen pipeline's inclusion rule
            (amendment v3.2, >=4 complete turns) is TURN-level; the two readings
            flip the branch at 88.3% turn coverage / 15-of-28 full coverage. The
            threshold was a task-brief invention, not prereg content; precedent
            follows E043 (registered analysis at 1,129/1,225 = 92.1% coverage).
            The unattended session withheld analysis per the strict letter and
            escalated; this ruling resolves it per E043.
-----END VERBATIM-----

======================================================================
SILENT DIFF vs REDO_STATUS.md / REDO_REPORT.md:4-6 / ALL_EXPERIMENTS_RECORD.md:2399-2400
  REDO_STATUS.md sha256 cf98e58e07bcc128df42af97cb784f0edbcac9a09a1912b288f7528290c4b3ec
  REDO_REPORT.md sha256 e3a71bc1a50a95a6165cdc564072536042bfcdfbf8e901d8140418c72af53715
  AER            sha256 eea4fcb1eabbe1f596dd97c85fb9df41583b7fd74951d32c1fbbdb6ce11b1bc8
======================================================================

## A. RELEASE CONDITIONS -- one divergence to report, one condition satisfied only in letter

A1. PARTIAL-COVERAGE labelling + n + 28-store co-statement (ruling item 1): CONSISTENT.
    REDO_REPORT:3-8 carries the blanket PARTIAL-COVERAGE header, per-row n, and both the
    task-brief "27" and the 28-file store count. No divergence.
    (Form note, non-substantive: the ruling specifies n as "n=x/27"; the report states
    turn-level n as x/1082 or x/792 for the map rates, where a /27 denominator is not
    defined. Trajectory-level rows do use n=27 / "of 28".)

A2. Missingness-nonrandomness disclosure (ruling item 2): SATISFIED IN LETTER, BUT THE
    MANDATED FLAG CANNOT BIND. The ruling required "any bucket with significantly
    imbalanced coverage" to be marked "insufficient coverage" but did not define the test.
    REDO_REPORT:62-65 defines it at report time (bucket marked when its 95% Wilson CI
    excludes the overall rate), applies it, and reports NO BUCKET TRIPS -- while itself
    stating that B3's CI "contains the overall rate by construction" because B3 dominates
    the universe. So the ruling-mandated flag is structurally incapable of firing for the
    only bucket with missing data. The report discloses this; the paper should not cite
    "no bucket flagged" as evidence of balanced coverage.
    Coverage as executed: B1 11/11 = 1.000, B2 17/17 = 1.000, B3 0.881, overall 0.883;
    all 142 missing + 1 unparseable turns fall exclusively in B3 (13 trajectories, each
    length >= 44 turns), i.e. missingness IS non-random by the report's own account.

A3. Tail 142 turns not re-run (ruling item 3): CONSISTENT. REDO_REPORT:94-95.
A4. tmp+rename to-do not implemented (ruling item 4): CONSISTENT. REDO_REPORT:95-96.

## B. THE TWO PILLARS -- ONE SUBSTANTIVE DIVERGENCE

The two pillars supporting release (REDO_STATUS:38-42, DECISIONS item 5):
  Pillar 1 -- E043 PRECEDENT: the registered original analysis ran at 1,129/1,225 = 92.1%
              coverage, so 88.3% is "materially the same regime".
  Pillar 2 -- FROZEN v3.2 INCLUSION RULE: amendment v3.2 (>=4 complete turns per
              trajectory) "would retain ALL 28 trajectories at this coverage".

  Pillar 1: CONSISTENT across DECISIONS:2230-2232, REDO_STATUS:40-41, REDO_REPORT:94.

  Pillar 2: ** DIVERGENT -- REPORT TO IDEATION. **
    REDO_STATUS:38-41 (verbatim): "the FROZEN pipeline's own inclusion rule (amendment
    v3.2: >=4 complete turns per trajectory) would retain **all 28** trajectories at this
    coverage".
    REDO_REPORT:27 (verbatim, executed run): "n_eff=21 (v3.2 excl 2, v3.3 excl 5, of 28)".
    When the frozen rule was actually run, v3.2 EXCLUDED 2 OF 28 -- it retained 26, not 28.
    The pillar as worded in the withholding report is contradicted by the executed
    analysis it was used to authorize.
    MAGNITUDE / DOES IT FLIP THE BRANCH: no. 26 >= 20 under either the trajectory-level or
    turn-level reading, so the release decision stands on the stated facts. But the
    sentence "would retain all 28" must NOT be quoted in the paper as written.

## C. DATE ORDER -- CONSISTENT, but the release is TWO-STAGE and must be described as such

  2026-08-11, unattended window : REDO_STATUS written; S3 gate read as 15/20 -> ANALYSIS
```

```
                                WITHHELD; escalated. (REDO_STATUS:1, :30-51)
  2026-08-11, [AUTHOR-1]          : "(REDO cap-trip salvage ruling)" -- branch (b)
                                AUTHORIZED; PARTIAL-COVERAGE analysis released.
                                (DECISIONS:2205-2232)
  2026-08-11, post-ruling       : REDO_REPORT written, citing that ruling.
                                (REDO_REPORT:4-6) At this point REDO_REPORT:92-94 still
                                records "the UNDER REVIEW annotations remain in place".
  2026-08-11, F1-arc closeout   : item (4) "RELEASE of the UNDER REVIEW cluster per R8's
                                sequence -- executed in this session AFTER (5) passed".
                                (DECISIONS:2251-2254)
  AER:2399-2400 records the same order ("after the cap trip, by the 2026-08-11 [AUTHOR-1]
  salvage ruling (branch (b), PARTIAL-COVERAGE)") and adds the prior authorization of the
  re-replay itself by STAGE2_VERDICT.json -- a different scope, not a competing account.

  NO DATE-ORDER DIVERGENCE. The one thing to state precisely in the paper: "withheld ->
  released" is TWO releases on the same day -- (i) the analysis is released by the salvage
  ruling, while (ii) the UNDER REVIEW annotations on the claim text are released only
  later by the F1-arc closeout item (4). A single "withheld then released" sentence
  conflates them.

  Consequence worth carrying: DECISIONS "(F1 block addendum) [2026-08-10]" -- "gen_numbers.py
  runs PROHIBITED while the UNDER REVIEW annotations hold ...; prohibition lifts with the
  review" (DECISIONS:2182-2184). The prohibition lifts at stage (ii), not at stage (i).

## SUPPORTING VERBATIM
-----BEGIN VERBATIM (REDO_STATUS.md:30-51)-----
## Why analysis is withheld (the S3 branch decision, stated honestly)

S3: "if >=20/27 trajectories are complete, run the frozen analysis on the complete
subset [...] If <20/27, no analysis." Under the reading in which a trajectory is
"complete" when every expected turn has a summary -- the only reading under which the
20-of-27 threshold and the n=<x>/27 label cohere -- the count is **15 < 20 -> analysis
withheld**.

For the morning ruling, the other reading is stated rather than buried: the FROZEN
pipeline's own inclusion rule (amendment v3.2: >=4 complete turns per trajectory) would
retain **all 28** trajectories at this coverage, and the ORIGINAL XFAM-v1 analysis ran
at 1,129/1,225 (92.1%) coverage -- 88.3% here is materially the same regime. Under that
reading the analysis is permissible and one command away:

```
/venv_gpu/bin/python runs/xfam_ext/redo/analyze_redo.py     # zero GPU, frozen callers only
```

The choice between readings flips the branch, the window was unattended, and "stopping
is always the correct unattended default" -- so it stops here. **[RECOMMENDATION,
labeled as such: run it under PARTIAL-COVERAGE labeling; the E043 precedent and the
frozen v3.2 rule both support analyzability at this coverage. [AUTHOR-1] decides.]**
-----END VERBATIM-----
-----BEGIN VERBATIM (REDO_REPORT.md:4-6)-----
1,082/1,225 parseable = 88.3%; gate: `REDO_MANIFEST.json`). Authorized by the
2026-08-11 [AUTHOR-1] ruling (DECISIONS "(REDO cap-trip salvage ruling)"), executed with
frozen callers only (`analyze_redo.py`; one glue quarantine-rename committed with its
-----END VERBATIM-----
-----BEGIN VERBATIM (REDO_REPORT.md:27, pillar-2 divergence)-----
| fidelity within-traj Spearman (P1 pilot quantity) | -0.0800 [-0.2086, +0.1424] n_eff=18 |
-0.0431 [-0.1245, -0.0164] n_eff=21 (v3.2 excl 2, v3.3 excl 5, of 28) | CI now excludes 0
on the negative side -- but see verdict row | `runs/xfam_ext/mde/xfam_ext_mde.json` (pilot
block) | `redo_family_analysis.json` -> `analysis.aggregate.spearman` |
-----END VERBATIM-----
-----BEGIN VERBATIM (REDO_REPORT.md:55-70, missingness)-----
Missingness non-randomness (ruling-mandated disclosure) -- `redo_missingness.json`

- The 142 missing + 1 unparseable turns are concentrated **exclusively in B3 (long)
  trajectories**: 13 partially-covered trajectories, every one of length >= 44 turns
  (list in the JSON); B1 (11/11) and B2 (17/17) are 100% covered.
- Per-bucket coverage (frozen `analysis/fidelity.py` BUCKETS): B1 **1.000**
  [0.741, 1.000]; B2 **1.000** [0.816, 1.000]; B3 **0.881** [0.861, 0.898]; overall
  0.883. Flag criterion (stated, then applied): a bucket is marked *insufficient
  coverage* when its 95% Wilson CI excludes the overall rate -- **no bucket trips it**
  (B3 dominates the universe, so its CI contains the overall rate by construction;
  that structural fact is itself part of this disclosure).
- Mechanism: the cap trip cut off the e27-class tail backlog, which lives in long
  trajectories' hard turns. Missing turns are therefore plausibly *harder-than-average*
  B3 turns; the per-bucket B3 numbers inherit that caveat even though the flag
  criterion does not fire. Stated for the record; no correction applied (none is
  registered).
-----END VERBATIM-----
-----BEGIN VERBATIM (ALL_EXPERIMENTS_RECORD.md:2399-2400)-----
  [10.6,14.7] outside the frozen 5pp band) and, after the cap trip, by the 2026-08-11
  [AUTHOR-1] salvage ruling (branch (b), PARTIAL-COVERAGE).
```

-----END VERBATIM-----

New incident (i): the STOP-for-absent rule's recurrence, one batch after it was written. The retrieval record of the G17 excerpt documents it in the retriever's own words: two reports in the same session asserted the ruling text was "not on this machine" while it sat in an unenumerated root; the rule is tightened — the root list must be written out before the absence claim, not reconstructed afterwards. Source: (retrieval-record header, English; ruling body Chinese-major, byte-exact in the supplementary pack).

New incident (ii): the relock comparator's own v1 failure. The mandatory lock-target re-verification after the CLAIMS change FAILED on its first run — on storage-format artifacts (blockquote prefixes, original-linebreak markers, curly quotes), the checker's defect and not the file's; kept on the record per the comparator lesson. The full log ships in this delivery (`logs/claims_relock_log.md`) and its v1-failure paragraph is retained inside it verbatim.

New incident (iii): the float-saturation law. Recorded as a taxonomy entry in Appendix A ("A length check against a float-saturated flow"), approved verbatim; the three-stage forensic record (word-diff, ragged-bottom exclusion, downstream deletion probe) is in the compression report of this delivery.

L CONCURRENT WORK, RECONCILED

Three concurrent papers touch this paper's objects. The main text carries one-sentence positions (§§1–3, 6); the full reconciliations are recorded here.

CARL (multi-hop search). CARL reports that entropy separates critical from non-critical states at the distribution level (Cliff's $\delta = 0.42$, $n = 294$, Brunner–Munzel), in multi-hop search, with criticality estimated from resampled continuations. Their criticality signal does not enter training through a classification threshold. It enters as an action-density term — expansion count per unit entropy at a state — with the update set restricted to states that have more than one child; the operating point at which their method works is therefore an emergent property of the rollout budget, not a stated rule whose recall was measured. We record the absence explicitly, because it is the substance of the difference: the paper reports no precision, recall, F1, or AUC for criticality identification, and no threshold, quantile, or top-k selection rule. Two independent passes over the v3 text were made, over the sections named in the provenance record; the absence is a finding of those passes, not an inference from silence in a summary. Nothing in our data speaks to multi-hop search, and nothing in theirs measures a router's recall; the designs answer different questions. We reconcile against v3 (2026-05); the bibliography entry carries the v1 date. The revision, not the first posting, is what "concurrent" refers to here.

CAR (formalization). CAR formalizes per-step causal effect measurement as do-operations with re-execution and validates the estimator on synthetic SCMs with planted effects — ground truth known by construction. That validation and this audit are complementary ends of one chain: it establishes that the estimator recovers planted effects where truth is known; we deploy the instrument on a live environment where truth is not planted, to grade signals already in training use. The premise their construction states — that executed re-intervention is the right ground truth for step credit — is the premise this audit tests signals against. Their own scope statement is explicit: real tools with side effects are, in their words, "out of scope" (§7). The delta is therefore theirs to state and ours to inherit: validation under planted effects on the one side, deployment in a live tool environment on the other.

CSO. The phrase "policy reachability" appears in CSO's abstract (Li et al., 2026a); the method section describes the requirement differently, as branches that remain within the policy's capability. We quote the abstract and say so, rather than attributing to their method a term their method does not use. What neither section reports is the verification rate. The rate is not a detail: it is the fraction of nominated steps whose counterfactual the policy actually supports, the quantity our measurability map measures at scale (undefined at 13.1% of intervened turns for Qwen, 26.8% for Llama; §3). Our audit is, among other things, that unreported number.

C3. C3 describes a "first method-agnostic auditing tool" for multi-agent credit. Its fidelity criterion is Spearman correlation against its own replay advantages — the tool grades credit against a quantity the same tool computes — audited on MAPPO/MAGRPO policies. The criterion is internally coherent but self-referential: it certifies agreement with the tool's own ground-truth estimate. The object here is different: signals already in LLM-agent training use, graded against an external, executed causal ground truth that consults none of them (§2). The two priority claims do not collide because the objects differ.

L.1 SOURCE INTEGRITY

All sixteen source files above ship byte-exact in `supplementary/` with `SHA256SUMS.txt`; the per-item sha256 values in the comments of this appendix's LaTeX source are the same values. Verbatim means verbatim: the transliteration applies to the PDF rendering only, and the release pack is the authority.