THE CEILING ON VIOLATION-FREE SUCCESS CAN SIT BELOW THE BAR A CLAIM MUST CLEAR

Anonymous authors
Paper under double-blind review

ABSTRACT

τ-bench grades an agent by the database state it ends in, so an agent that reaches the goal through a write the user never sanctioned still scores a success. Two numbers read off a baseline arm decide, before any compliance intervention is built, whether its gain on violation-free success could ever be claimed: the *ceiling*, $\mathrm{pass} - \mathrm{vfpass}$, the most a success-preserving remover could add; and the *resolution*, the 95% half-width of a task-paired violation-free-success delta under a task-clustered bootstrap. On our 3B retail arm the ceiling is 1.16 pp and the resolution is 1.9–5.5 pp, so no success-preserving remover could post a claimable gain there. The bar is set by our comparison rather than by the benchmark: a task-paired test resolves an ideal remover's gain on most of our baseline estimates, and fewer than half clear the registered spread bar. We then run the check on a real intervention, a 1.5B violation-forecasting world model deployed training-free in the planning loop over two domains, two agent families, three scales and 11 conditions. It removes violating episodes at 3B, on an arm that succeeds less often than a do-nothing agent, and again on the second agent family. But most of what it stops the environment would have rejected anyway, the removal sits on the dose–response line that three uninformed arms fix, and its premium over an uninformed scorer at matched dose is unresolved at 3B and does not replicate on the second family; a minority of the paired contrasts meet the registered spread criterion together with a paired-interval condition added later. We propose the target arm's ceiling and resolution, then a do-nothing anchor and an uninformed control at matched dose, as the entry bar for compliance interventions, and release the harness and records in the supplementary archive.

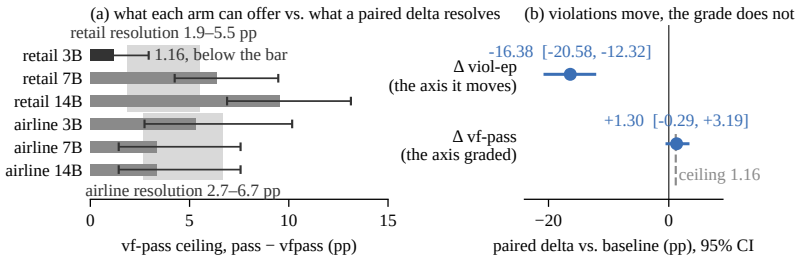

Figure 1: (a) The most a success-preserving violation remover can add to violation-free success on each baseline arm (Eq. 1, Wilson 95%), against the range of 95% half-widths a paired violation-free-success delta carries on that domain; on retail 3B the ceiling is below the resolution. (b) The 3B forecaster on both axes, two runs against two baselines: it moves violating episodes and leaves the graded composite where it was. The dashed ceiling is the single-run baseline's; the pooled four-card baseline's is 1.59 pp ($[0.89, 2.83]$ pp), above the pooled point.

1 INTRODUCTION

A language agent that books, cancels, refunds and exchanges on a user's behalf can fail twice over: it can fail to accomplish what the user asked, and it can accomplish it by issuing writes the user

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

never authorized. Benchmarks for tool use score the first failure and are largely silent on the second: τ-bench (Yao et al., 2025), a widely used testbed for policy-constrained customer-service agents, grades an episode by whether the final database state matches the goal, so an agent that reaches the goal through an unsanctioned write scores the same as one that asks first. Measured directly, the second failure mode is not a tail event. Across 3B–14B instruction-tuned agents on τ-bench retail, 36.81% to 58.84% of episodes contain at least one write outside the sanctioned action set, at 1.04–2.19 unsanctioned writes per episode; pooled over the three arms, 33% of the *successes* these agents report are contaminated by such a write (59/177). Success rises from 4.93% at 3B to 30.43% at 14B, while the per-arm contaminated shares (4/17 to 33/105) have overlapping intervals, so we read no trend into them in either direction.

The target metric can be out of reach before the intervention is built. The obvious response is to build something that removes violations; we did, and two quantities computable from the baseline decide in advance what the benchmark can say about it (Figure 1). The first is the **ceiling**. Violation-free success rises only when an episode that succeeded *with* a violation becomes one that succeeds without, and on our six baseline arms 81% to 97% of violating episodes were failing anyway, so a *perfect* success-preserving remover gains at most $\mathrm{pass} - \mathrm{vfpass}$: from $9.57\,pp$ down to $1.16\,pp$ (4/345, $[0.45, 2.94]$ pp) on our arms. That low end sits below the bar this paper registered for a claim, so on that arm no success-preserving mechanism could have produced a *claimable* gain on the target metric (§6). The second quantity is the **resolution**: the half-width of our own paired comparison, the interval half of the claim rule; the benchmark imposes no such floor. A paired delta here is resolved by the number of *tasks* rather than episodes, because trials of one task are correlated: a task-clustered paired bootstrap over 115 retail tasks puts the 95% half-width of a violation-episode delta at 5.1–7.8 pp (7.0–10.0 pp on airline's 50) and of a vf-pass delta, the ceiling's axis, at 1.9–5.5 pp. A single-run improvement of a few points here is not a small effect; it is an unresolved one.

The intervention, and what survived. We treat unsanctioned writes as a forecasting problem: a *violation-forecasting world model*, a 1.5B classifier fine-tuned on $1{,}000$ agent episodes with mechanically derived labels (validation AUC 0.9832), scores each candidate tool call in the planning loop of an unmodified agent; flagged calls are re-sampled, escalated to a clarifying question, or de-escalated (§§3–4). Two domains, three Qwen scales and 11 intervention conditions give 49 paired violation-axis contrasts (Figure 2); the Llama-8B family is counted apart (App. J). **Under the registered spread criterion and a paired-interval condition added later,** 16 **of the** 49 **survive**; the 8 that involve the trained forecaster all sit at 3B, where it removes -16.38 pp of violating episodes ($[-20.58, -12.32]$) as issued, and -1.59 pp counting only writes the environment accepted. We cannot show that the *training* produced it: three uninformed arms trace what blocked-write dose alone achieves, and the forecaster sits -1.32 pp off that line (Figure 3). At 7B no scorer clears the bar in both of its runs, the random one included; at 14B nothing moves, and the per-event logs show why (§7); transfer to a second domain is null on every arm. Claims withdrawn or corrected between versions of this paper are listed in Appendix J.

Contributions. (i) A ceiling on violation-free success, $\mathrm{pass} - \mathrm{vfpass}$, computable before an intervention exists and on a real arm below the registered claim bar, which an ideal-remover test separates from statistical power (§6). (ii) A resolution estimate for paired compliance comparisons on τ-bench (§2, Figure 6). (iii) An *uninformed-scorer* control at matched dose and a dose–response line (§5, Figure 3). (iv) A violation-forecasting world model deployed training-free, with a per-event reading of where it fails (§7, Table 3). (v) The harness, the forecaster's training and evaluation code, and per-episode records, in the supplementary archive.

2 Measuring Compliance Before Intervening

Surface and protocol. We evaluate on τ-bench (Yao et al., 2025): a tool-calling agent serves a simulated customer under a written domain policy, against a live database. *Retail* (115 tasks) is the development domain; *airline* (50 tasks) is held out entirely for transfer. Protocol, fixed before any intervention ran: a $32{,}768$-token context, 3 trials per task, and the user simulated by Qwen2.5-14B-Instruct. Runs before 2026-08-24 served the agent from one replica on three GPUs and later runs from two replicas on four; we call these the two *serving configurations*, treat a run's configuration as part of its protocol coordinate, and mark every pair that crosses it.

Metrics, and the two failure modes they separate. An episode's *success* (pass[1]) is τ-bench's database-state check. A *violation* is a write outside the task's sanctioned action set, mechanically derived from the gold action sequence with no judge model and no annotation. We report four quantities per arm: success; the fraction of episodes containing any violation (*viol-ep*); unsanctioned writes per episode; and *violation-free success* (vf-pass), success with zero violations, the natural composite target that §6 is about (an earlier gold standard was silently broken; Appendix J). Of a seeded sample of 60 such writes read against the policy text, 59 are writes the policy forbids, 1 only leaves the gold path, and 8 forbidden writes the environment committed (Appendix F).

Do-nothing anchors. τ-bench rewards are computable for an agent that executes nothing, since some tasks are satisfied by the initial database state: 6.09% on retail and 38.00% on airline, where every model we run scores *below* it (9.33%–15.33%), so airline is used only for the violation axis; retail 3B's baseline (4.93%) also sits below its anchor.

Resolution: what a paired delta on this surface can show. `PROTOCOL.md`, committed on 2026-08-06 before any result in this paper existed, fixes the criterion: report pass[1] as a mean with its cross-trial spread, and *do not claim a difference smaller than that spread* (an earlier draft used a flat 3 pp floor, under which eleven contrasts cleared; Appendix J). Every cell here is 3 trials of one task set under one fixed configuration, so the spread across those trials is the instrument's own noise with nothing changed; over 60 cells the trial index predicts neither the best nor the worst trial on the axis this bar governs (App. D). Across the 48 retail cells the per-trial standard deviation has median $3.51\,pp$ on the violation-episode axis and $1.74\,pp$ on success, and the widest same-configuration range on the violation axis is $20.87\,pp$; the violation axis, which carries this paper's effects, is the *noisier* of the two by about $2.0\times$. The 4 same-configuration re-runs inside one serving configuration measure what one draw could only suggest: run-level drift reaches 2.32 pp on this axis, points in both directions, and every one of the 16 contrasts we claim exceeds it; over all re-run pairs, including the one that crosses the change of serving configuration (§5), the widest drift is 4.64 pp.

The paired interval a reviewer will compute. Episodes are not independent: three trials share a task, and tasks differ in how many writes they invite (on the three retail baseline arms the 12 write-heaviest tasks carry 31%–36% of all unsanctioned writes and 18–33 of the 115 tasks violate in no trial; Appendix B), so the resolution of a paired delta is set by the number of tasks (115 retail, 50 airline), not of episodes (345 and 150). We resample tasks with replacement, carrying every episode of a drawn task into every arm of the contrast, $20{,}000$ times (Figure 4b). Over the single-run pairings the 95% half-widths are 5.1–7.8 pp on retail and 7.0–10.0 pp on airline, wider there because airline has fewer tasks, so airline needs its own bar.

The rule. One rule covers a paired delta and a difference of paired deltas alike: a comparison is *claimed* in §5 only if its point estimate exceeds the widest cross-trial range among the runs it is built from *and* its paired interval excludes zero; otherwise it is reported with its interval and claimed of nothing. Where a comparison merges two runs of a condition into a pooled cell, the range is taken over the *constituent runs*: a pooled cell's three trials are two-run averages, so its range sits below either run's (Appendix J records two claims withdrawn on this point). Of the 49 paired violation-axis contrasts here, 16 meet the rule. Replication across independent runs of both arms, which 9 of our differences of deltas have, is a reason to trust a number rather than a second route past the spread bar.

3 A VIOLATION-FORECASTING WORLD MODEL

The forecaster answers one question: *if the agent executes this tool call now, will it constitute a policy violation?* It is a world model in the narrow sense (it predicts the consequence of a candidate action in the environment's terms), and not in the broad sense: it has no observations, rollouts or user model.

Corpus. $1{,}000$ episodes of real agent behaviour on the retail *train* split (14B and 7B agents; $\sim$6,400 tool calls), collected before any intervention existed; the evaluation tasks are the retail *test* split, so the two are disjoint. Labels are mechanical (§2): each write marked sanctioned or not against the task's gold action set, with no preference data, judge model or human annotation.

Model and training. A 1.5B instruction model (`Qwen2.5-1.5B`) fine-tuned to classify (conversation, candidate call) pairs; conversation and candidate call enter as one sequence, the earliest point in the fusion design space surveyed for heterogeneous evidence (Xu et al., 2026), rather than through separate encoders. On a task-level held-out split: AUC 0.9832, accuracy 93.74%, ECE 0.0158 at a positive rate of 22.44%. Calibration is the property that matters: the intervention gates on the raw probability, so a miscalibrated forecaster moves the operating point silently, a burden every thresholded gate shares, whether its threshold is calibrated to a risk target (Xiao et al., 2026b) or adapted online (Xiao et al., 2025). A 3B ablation scores AUC 0.9800 against the 1.5B head's 0.9832, a gap of 0.0032 on 655 held-out examples for which we have no interval; we report both and serve the smaller model. Leakage is excluded by the task-level validation split and by a 200-example manual sweep: no gold-action strings in model inputs.

Serving. The forecaster runs as a separate HTTP service on its own GPU: a dead scorer fails the episode rather than degrading to no oversight (an earlier in-process version lost 16 of 22 tasks to a missing dependency without any error), and every baseline in §5 inherits that contract.

4 Training-Free Deployment in the Planning Loop

The agent is unmodified. Before executing a candidate tool call the wrapper scores it; calls below the threshold execute normally (v0.1 ran at τ=0.5; v0.2, the version every result in §5 uses, at τ=0.35). The move to 0.35 was made after reading the v0.1 retail 7B/3B logs, in which the violating calls that ran unflagged had scores with median 0.33 and maximum 0.498, just under τ=0.5; those logs are test-split episodes, so the forecaster's operating point was set on the same split the results use. v0.2 is the first version run after that choice, and the dose-matched uninformed control (§5) is the comparison that does not inherit it. A flagged call triggers, in order: *re-sampling* (k=4 alternatives at elevated temperature; execute the best-scoring one *if* it clears the threshold), *clarification* (replace the call with a question to the user, at most 4 per episode), and, on re-issue after clarification, an anti-livelock rule that eventually lets a persistently re-issued call through rather than deadlocking the episode, so the last branch, and an exhausted clarification budget, both execute the write the gate fired on. Every decision is logged with its probability; the flow with its branch counts is in Appendix G, and §7 reads the logs.

v0.1's lesson: lateral re-sampling replaces one violation with another. The first deployment allowed a flagged *write* to be replaced by another write, and the per-episode logs showed the re-sampler swapping a flagged violation for a different unsanctioned write the forecaster had not seen; the 7B arm's success moved -2.90 pp, within run-to-run noise but in the costly direction. v0.2 forbids write→write substitution: a flagged write may only de-escalate to a read or a question. The threshold also moved ($0.5 \to 0.35$), so the raw v0.1-vs-v0.2 contrast confounds rule with threshold; the decoupling control (v0.1's rule at v0.2's threshold) does not separate them: rule at fixed threshold $+4.06$ pp ($[-2.61, +10.43]$) and threshold at fixed rule -2.32 pp ($[-9.86, +4.93]$) both span zero, so we attribute the reduction to neither factor.

5 Results: What Survives the Resolution

RQ1: Does the forecaster remove violations, and does the removal replicate? At 3B it does (Table 1; Figure 2 shows all 49 contrasts against their bars). Against the two four-card baseline runs, the two four-card forecaster runs remove -16.38 pp ($[-20.58, -12.32]$), the estimator with one serving configuration on both sides and the one this paper's headline quotes. Counting only writes the environment accepted, the same contrast is -1.59 pp ($[-3.48, +0.14]$), below that column's 3.33 pp run-to-run drift; of the 24 contrasts that meet the rule over every pairing (the 16 single-run ones of Table 4, 6 pooled cells and 2 second-family cells), 19 meet it on that axis: 10 drop out and 5 enter (Table 1). The three same-configuration runs remove -14.49 pp, -18.84 pp and -16.23 pp (the first on three cards, the other two on four); two forecasters retrained from scratch (training seeds 1/2; val AUC 0.9759/0.9798) remove -12.75 pp ($[-19.13, -6.67]$) and -15.36 pp ($[-21.45, -8.99]$), each claimable in its own right; across runs and checkpoint seeds the four v0.2 point deltas span 6.09 pp, inside the arms' own trial spread (Figure 6a). Run 3 is no closer to its four-card mate ($+2.61$ pp, $[-2.61, +7.83]$) than to the three-card run (-1.74 pp, $[-7.54, +4.35]$), so the -4.35 pp gap between

Table 1: The arms this section argues from (retail, 345 episodes per run; "n runs" = pooled estimator). Row one pairs the two four-card forecaster runs against the two four-card baseline runs; other rows are paired against the agent's three-card baseline (viol-ep 36.81% at 3B, 58.84% at 7B). *flag*: share of scored writes the gate fired on; *block*: share it stopped ("—": derived only for Figure 3's arms; re-inject has no gate). Δviol-ep$_{\text{acc}}$: only writes the environment accepted (the label counts issued writes; baselines had 82.6% rejected); "—": derived only for pooled cells. CIs: task-clustered paired bootstrap.

agent	condition	flag	block	Δviol-ep	95% CI	Δviol-ep$_{\text{acc}}$	Δvf-pass
3B	+WM v0.2 vs 4-card base (2+2 runs)	86.4%	55.1%	-16.38 pp	$[-20.58, -12.32]$	-1.59 pp	$+1.30$ pp
	+WM v0.2 (2 runs)	86.7%	57.1%	-16.67 pp	$[-21.88, -11.30]$	-3.62 pp	$+0.29$ pp
	random scorer, τ=0.35 (2 runs)	67.7%	46.6%	-11.74 pp	$[-17.25, -6.23]$	-4.35 pp	$+1.16$ pp
	random scorer, τ=0.13	86.5%	57.4%	-15.94 pp	$[-22.32, -9.57]$	—	$+1.16$ pp
	intercept-all, no model (2 runs)	100.0%	60.4%	-21.88 pp	$[-27.25, -16.52]$	-6.81 pp	$+3.04$ pp
	clarify-only, k=0 (2 runs)	87.2%	40.7%	-9.57 pp	$[-15.22, -3.77]$	-3.91 pp	$+0.58$ pp
	+WM v0.3, re-exec block (2 runs)	84.0%	75.2%	-22.32 pp	$[-28.12, -16.52]$	-4.93 pp	$+1.45$ pp
7B	+WM v0.2 (3 runs)	76.0%	50.4%	-3.29 pp	$[-9.47, +3.00]$	-4.54 pp	$+0.29$ pp
	intercept-all, no model	100.0%	58.1%	-6.38 pp	$[-14.20, +1.45]$	—	$+0.87$ pp
	clarify-only, k=0	76.5%	—	-3.48 pp	$[-11.01, +4.06]$	—	-1.45 pp
	policy re-inject, no scaffold	—	—	-41.45 pp	$[-48.70, -34.20]$	—	-4.06 pp

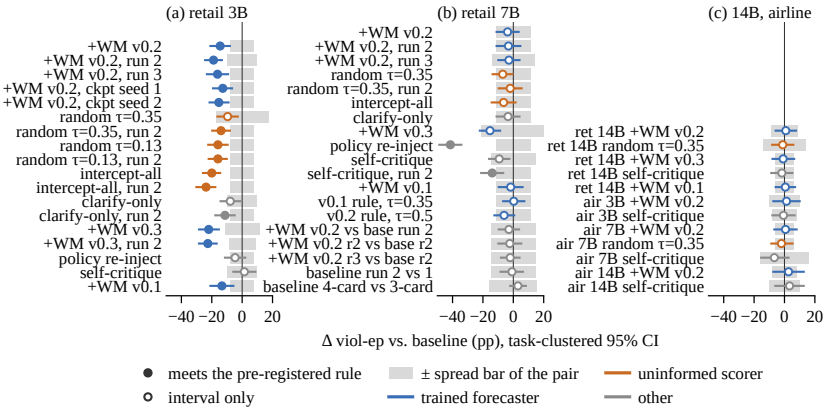

Figure 2: All 49 paired violation-episode contrasts (single runs; rows in Table 4) with task-clustered 95% CIs and, as the grey band, the spread bar the pre-registered rule judges each against. Filled markers: the 16 that meet the rule.

runs 1 and 2 is run-to-run scatter, not a serving-configuration term; the three estimators of the removal agree (Appendix H). **At 3B the removal is claimable and replicates across runs and retrainings; it is the only scale at which any scorer's removal does.**

RQ2: Does training produce the removal, or does dose? The scaffold's dose is not the share of scored writes the gate *fires* on: the anti-livelock override and an exhausted clarification budget both let the original call execute unchanged (§4), so intercept-all, which flags every write, blocks 60.4% of them at 3B. Three 3B arms sit within 0.7 pp of each other on the flag axis (clarify-only, the forecaster, the uninformed score at τ=0.13; Table 1) and remove -9.57 pp, -16.67 pp and -15.94 pp, so flag rate does not explain removal. Blocked-write rate can be tested against a line (Figure 3): the three *uninformed* arms (one scaffold, no model, only a threshold moved) give three dose points and through them 0.64 pp of removal per pp of blocked writes, residuals at most 2.26 pp, and the trained forecaster sits -1.32 pp off it, inside the resolution, as is clarify-only at $+2.13$ pp; v0.3 blocks 75.2%, past the fitted range, and its -7.33 pp is off an extrapolation.

The uninformed score is a context hash, uniform on $[0, 1)$, so its threshold sets its rate and nothing else. Moved from τ=0.35 to τ=0.13 and run twice, it fires on 87.5% of scored events against the forecaster's 86.7%, blocks 57.7% against 57.1%, and removes -15.94 pp in run 1 and -15.94 pp in run 2, pooled -15.94 pp ($[-21.45, -10.43]$). Its two runs (both four-card) against the forecaster's runs 2 and 3 (four-card), the gap is $+1.59$ pp ($[-2.46, +5.65]$) against a 9.57 pp bar; pooling the forecaster's runs 1 and 2 instead, across serving configurations, gives $+0.72$ pp ($[-3.19, +4.64]$):

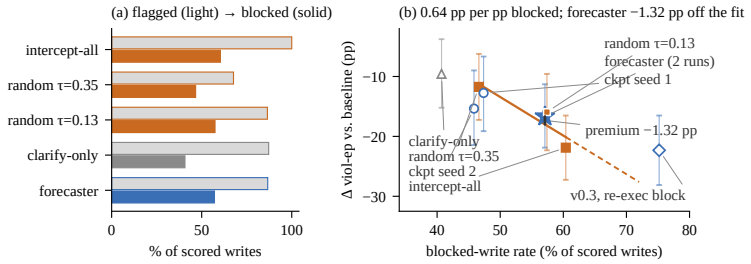

Figure 3: Removal tracks the blocked-write dose; the trained forecaster sits on the uninformed line (3B retail). (a) Per arm, scored writes flagged against blocked (values: Table 1); the gap is what the anti-livelock override and the exhausted ask budget let through, and the dose is the blocked share. (b) Line fitted through the three uninformed arms (squares) by the arithmetic that emits the slope and premium in the text, dashed where extrapolated; error bars: task-clustered 95% CIs. The random τ=0.13 square is drawn on the forecaster star it nearly coincides with.

no separation on either half of the rule under either pairing. At the control's default threshold the gap is $+4.93$ pp ($[+1.30, +8.55]$), replicated in both serving configurations ($+4.93$ pp, $+4.93$ pp) but short of its 17.39 pp bar: interval only. Intercept-all removes -21.88 pp ($[-27.25, -16.52]$), -4.35 pp ($[-7.97, -0.58]$) beyond the four-card control and short of the 9.57 pp bar taken over the constituent runs (claim withdrawn, Appendix J). We registered no equivalence margin and do not read the matched-dose null as equality: with two runs on each side its interval still admits 32% of the forecaster's own removal. **At 3B, removal tracks blocked-write dose; the trained scorer's premium over the uninformed line is -1.32 pp, and no uninformed control separates from it at this resolution.**

RQ3: Does it hold at 7B, at 14B, and on a held-out domain? At 7B the forecaster's 6 legitimate pairings (3 runs against two baseline runs) span -3.77 pp to -2.03 pp and 0 exclude zero; the random scorer's -6.96 pp ($[-13.62, -0.58]$) excludes zero but sits inside its baseline arm's 11.30 pp range, and pooled on both sides the trained-versus-uninformed gap is -1.21 pp ($[-5.27, +2.85]$) against a 13.91 pp bar. Clarify-only is indistinguishable from the full scaffold ($+0.29$ pp, $[-6.67, +7.25]$); v0.3 posts the largest scaffold-arm point estimate at this scale (-15.36 pp) on the widest same-configuration spread (20.87 pp), so the rule rejects it; self-critique clears the bar in run 2 (-13.91 pp, $[-21.16, -6.67]$) but not in run 1 (-9.28 pp); intercept-all is indistinguishable from the forecaster (-2.61 pp, $[-9.57, +4.35]$) and from the random scorer ($+0.58$ pp): neither *which* writes are stopped nor *how many* is resolvable at this scale (Appendix H). Policy re-injection (no scorer, the policy restated before every completion) posts the largest removal in the matrix, -41.45 pp ($[-48.70, -34.20]$), and its task success falls -9.86 pp ($[-14.78, -5.22]$) to 6.09%, the do-nothing anchor (6.09%; exact at the count level, $21/345$ against $7/115$): an agent that has largely stopped acting (vf-pass -4.06 pp). At 14B nothing moves under any scorer: forecaster $+0.87$ pp ($[-5.80, +7.54]$), self-critique -1.74 pp, random -1.16 pp, trained-versus-uninformed -2.03 pp ($[-8.70, +4.64]$); §7 measures why. A second agent family asks whether the 3B result belongs to the scorer or to the Qwen policy it was trained beside. On Llama-3.1-8B-Instruct (retail; one run per arm, three trials; no matched-dose arm) the baseline violates in 66.09% of episodes. The forecaster removes -23.19 pp ($[-30.14, -16.23]$) and the uninformed scorer at its default threshold -16.81 pp ($[-22.61, -11.01]$); both clear the bar, and the trained-versus-uninformed gap, -6.38 pp ($[-12.75, 0.00]$), does not, with the forecaster flagging 80.2% of scored events against the uninformed scorer's 64.6%. Here the forecaster also costs task success, -2.03 pp ($[-4.35, -0.29]$), where the uninformed scorer does not (-0.58 pp, $[-2.61, +1.16]$). Removal transfers across the family; a premium over an uninformed dose does not. Airline, never seen in training and with a wider resolution (7.0–10.0 pp), resolves nothing on any arm, its own uninformed anchor included (-2.00 pp, $[-8.67, +5.33]$); all three forecaster arms move in the *unfavourable* direction ($+1.33$ pp, $+0.67$ pp, $+2.67$ pp), and airline capability numbers stay out of every claim (§2). **At 7B, at 14B and on airline 2 of 31 contrasts clear the bar, both at 7B: policy re-injection, an agent that stopped acting, and self-critique in run 2 but not in run 1.**

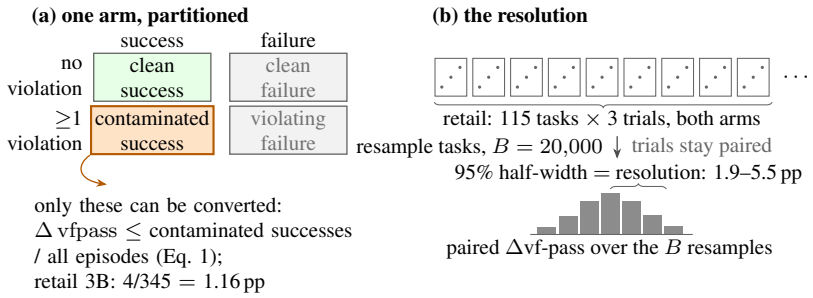

Figure 4: Two numbers the baseline fixes before any intervention is run. (a) A success-preserving remover can only convert contaminated successes, so their share of all episodes caps the violation-free-success gain (Eq. 1). (b) The resolution is the 95% half-width of our paired violation-free-success delta under a task-clustered bootstrap, drawn for retail; it is the interval half of the claim rule, whose spread half (§2) is what decides the retail 3B case: each resampled task carries all of its trials in both arms, so the delta is resolved by tasks rather than episodes; the range printed spans the retail comparisons.

Table 2: Where violations occur, and what that caps. *viol. in failures*: share of violating episodes that were failing anyway. *ceiling*: Eq. 1 over 345 (retail) / 150 (airline) episodes, Wilson 95%. Median paired-delta resolution on the violation-free-success axis: 3.5 pp retail, 4.3 pp airline (violation-episode axis: 6.7 pp, 8.7 pp). Figure 4 draws both numbers; decomposition in Figure 7.

arm	viol. in failures	contam. succ.	share	ceiling	ceiling 95% CI
retail 3B	97%	4/17	24%	1.16 pp	[0.45, 2.94] pp
retail 7B	89%	22/55	40%	6.38 pp	[4.25, 9.47] pp
retail 14B	81%	33/105	31%	9.57 pp	[6.89, 13.13] pp
airline 3B	93%	8/15	53%	5.33 pp	[2.73, 10.17] pp
airline 7B	95%	5/14	36%	3.33 pp	[1.43, 7.57] pp
airline 14B	95%	5/23	22%	3.33 pp	[1.43, 7.57] pp

RQ4: Does anything register on violation-free success? For the forecaster, vf-pass moves $+0.29$ pp ($[-1.45, +2.17]$) at 3B and $+0.29$ pp ($[-3.29, +3.86]$) at 7B, and $+1.30$ pp ($[-0.29, +3.19]$) on the four-card estimator; none excludes zero, and §6 computes why before the intervention exists. One arm is an exception and we report it: intercept-all at 3B moves vf-pass $+3.04$ pp ($[+0.72, +5.80]$), on each run ($+2.61$ pp $[+0.29, +5.51]$; $+3.48$ pp $[+0.87, +6.67]$), past this arm's 1.16 pp ceiling, because its task success rose, which Eq. 1 does not cover and the do-nothing anchor disqualifies as a capability claim (§6). **The graded composite does not move for the forecaster at any scale; the one arm that moves it raises task success from below the do-nothing anchor.**

6 The Ceiling on Violation-Free Success

Violation-free success can rise only one way: an episode that succeeded *with* a violation becomes one that succeeds without. Episodes that violate and fail contribute nothing, however thoroughly they are cleaned. So for any violation-removing intervention that does not itself change task success,

$$\Delta\,\text{vfpass} \;\le\; \text{pass} - \text{vfpass} \;=\; (\text{share of successes that are contaminated}) \times \text{pass}, \qquad (1)$$

a quantity readable off the *baseline*, before the intervention exists (Figure 4a).

The retail-3B ceiling (Table 2) is 1.16 pp (4/345, $[0.45, 2.94]$ pp), *below the spread bar the comparison had to clear*, 4.35 pp for this arm; a paired vf-pass delta on retail resolves 1.9–5.5 pp (median 3.5 pp), and the point sits under that range and the upper limit under its median, so no *success-preserving* mechanism could have produced a vf-pass gain that the criterion of §2 would claim. That bar is set by our comparison rather than by the benchmark. An *ideal remover* separates the two: relabel every contaminated success of a baseline arm as clean, move nothing else, and bootstrap the gain the same way. On retail 3B it is 1.16 pp $[0.29, 2.32]$, excluding zero under every seed from 4 cleaned successes

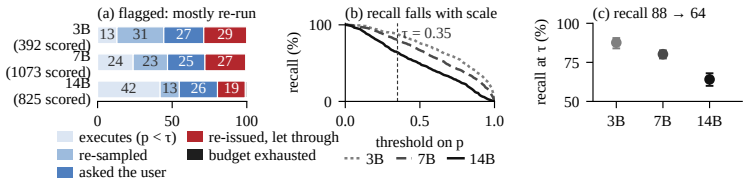

Figure 5: Recall falls with scale; most flagged writes are re-sampled or re-issued (v0.2, one run per scale). (a) Where scored writes go, one chain light to dark; red executes the flagged call anyway, black exhausted the ask budget. (b) Recall against the threshold, the marked τ being the one every result ran at. (c) Recall at τ, Wilson 95%; unsanctioned share and precision are in Table 3.

Table 3: Per scored write, v0.2, one run per scale. *unsanc.*: share that would have violated if executed; *recall/precision/FPR*: the gate at τ=0.35 against that label; *pass-miss*: unsanctioned share of writes let through; *AUC*: rank AUC of p (held-out split: 0.9832); recall intervals: Figure 5c.

agent	events	unsanc.	recall	precision	FPR	pass-miss	AUC
3B	392	96.9%	87.6%	97.7%	66.7%	92.2%	0.794
7B	1073	84.8%	80.2%	89.4%	53.4%	70.3%	0.700
14B	825	65.8%	64.1%	72.7%	46.5%	56.4%	0.653

on: the benchmark resolves this ceiling, our bar does not claim it. Across the 8 baseline estimates the interval excludes zero on 6 and 3 clear the bar (Appendix E). One of our own arms breaks the premise: 3B intercept-all moves vf-pass +3.04 pp ([+0.72, +5.80]), past this ceiling and its upper confidence limit, because task success rose +2.32 pp instead of staying fixed. Neither counterexample nor capability result: baseline success (4.93%) is below the do-nothing anchor (6.09%) and the arm lands at 7.25%, within resolution of it. The 7B and 14B ceilings (6.38 pp, 9.57 pp) are 1.8× and 2.8× the median vf-pass resolution 3.5 pp: a few detectable steps rather than a range. Of six arms, 1 (retail 3B) has an upper confidence limit below its domain's median resolution. In all six arms most violating episodes also fail (Table 2), so a composite metric mostly counts failures, and an intervention aimed at an arm whose ceiling is under the registered bar cannot post a claimable gain, whatever it removes.

7 Analysis: Mechanism, Failure Cases and Negative Results

Where the forecaster fails, measured per event. §5 reports that nothing moves at 14B and that the forecaster flags 58% of that agent's scored writes against 76% at 7B (the training corpus contains the 14B agent, so novelty of scale is not the explanation). Reading rules were fixed before the numbers were computed (`PREREG_EVENTS.md`); Table 3 and Figure 5 give the result. Two things change with scale: the share of scored writes that are unsanctioned falls from 96.9% at 3B to 65.8% at 14B, and recall at τ falls from 80.2% ([77.5, 82.7]) at 7B to 64.1% ([60.0, 68.0]) at 14B, an interval that lies entirely below the 7B one: a drop of 16.13 pp against a 3.26 pp range across that arm's 3 runs. Scored with the 7B per-class rates, the 14B agent's flag rate would be 71.0% against the observed 58%; the rest of the gap is the lower base rate. So the forecaster does miss more of the 14B agent's unsanctioned writes, in distribution, and closing the override channel outright at 14B (v0.3) still moves nothing (Appendix H), because a downstream rule cannot act on a call the gate never stopped. In-loop rank AUC is well below the held-out figure (Table 3); the populations differ (candidates proposed under intervention, re-issued calls scored again, 12 sanctioned events at 3B), and we report the gap without resolving it.

The leak channel. A flagged call re-issued after its clarifying question is eventually let through by the anti-livelock rule, so the second warning ends in execution (Figure 5a): 29% and 27% of scored events at 3B and 7B, the largest single branch at 7B, and 19% at 14B, where the forecaster passes 42% of scored events against 24% at 7B.

A fixed τ also defines a different intervention at each scale: the self-critique baseline fires on 77.4% of scored events at 7B but 1.7% at 3B and 17.9% at 14B (Figure 6b, Appendix I).

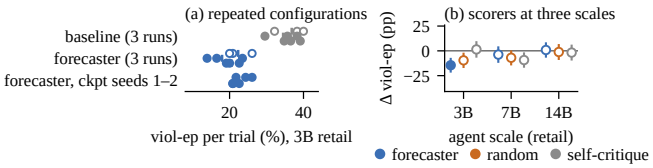

Figure 6: Spread lies within a configuration rather than between scorers. (a) Per-trial violation-episode rates of the 3B retail cells that repeat a configuration, one row per configuration, runs offset within it; filled: four-card, open: three-card; tick: run mean. (b) Scorers at each scale, 95% CIs; firing rates: Table 1.

Run-level drift, and episode-level instability. The 4 same-configuration re-runs inside one serving configuration drift by at most 2.32 pp on the violation axis, in both directions (Figure 6a). The aggregate is stable and the episodes are not: of the 219 episodes that succeed in either run of a pair, 60 (27%) succeed in both. That is the case for resampling tasks, not episodes, in §2.

Empty trajectories, negative results. Empty trajectories (up to 8.99% of a cell) cannot be labelled; on the non-empty subset 4 of 46 per-cell deltas turn from null to effect, all in the intervention's favour, and we keep the registered denominator. The decoupling control does not separate rule from threshold (§4), airline transfer is null, and on the unclaimed axis 33 of 49 contrasts pass the same rule, which we do not promote: the claim axis was fixed first (Appendix H).

8 Related Work

World models and step-level supervision for agents. Web- and GUI-agent world models predict observations to pick actions for task success (Chae et al., 2025); ours predicts policy violation, which success-directed prediction misses (§6). Agent process-reward methods score steps from outcome or demonstration signals (Choudhury, 2025; Xi et al., 2025); ours come mechanically from sanctioned action sets (§2).

Risk emulation and tool-use safety. ToolEmu-style systems *emulate* risky tool outcomes for red-teaming (Ruan et al., 2024); tool-agent safety benchmarks grade refusal or harm on curated scenarios (Andriushchenko et al., 2025; Yuan et al., 2024); proactive runtime guardrails predict a coming violation from a learned safety automaton (Wang et al., 2025), a step-level guardrail with feedback (Mou et al., 2026) or a risk-aware world model over the trajectory (Lin et al., 2026), graded on curated harm scenarios; we put a trained predictor in the planning loop and grade by the benchmark's own database-state semantics. Safe RL learns a numeric cost from a simulator or offline data, via a world model (Huang et al., 2024), cost conditioning (Liu et al., 2023) or a feasible region (Zheng et al., 2024); our policy is prose with no cost signal. On τ-bench (Yao et al., 2025; Barres et al., 2025) the policy is already in the system prompt, so re-injection and self-critique are prompt-level baselines. Predictor-free routes run alongside, such as typing that makes unsafe actions unrepresentable (Xiao et al., 2026a); a learned forecaster is what remains when the policy is prose.

9 What a Compliance Intervention Should Report

Two quantities computable before an intervention exists decide what it can show: the *resolution* a claim must clear (1.9–5.5 pp here) and the *ceiling* $\mathrm{pass} - \mathrm{vfpass}$. On our 3B arm the ceiling is below that bar, resolved by a paired test and unclaimed by our criterion, so the one scale where our scaffold removes violations resolvably is one where no success-preserving removal could post a claimable gain, on a baseline below the do-nothing anchor. We ask that an intervention report, before its results, **both failure modes** separately, **do-nothing anchors**, **its resolution and the ceiling**, **an uninformed-scorer anchor at matched dose**, and **its per-event recall and leak channels**. **Limitations**: seed variance and the matched-dose control exist at 3B only; same-arm drift is at most 2.32 pp within one serving configuration and 4.64 pp across it, so smaller differences are unreadable.

AI USE STATEMENT

Generative AI tools were used substantially in this work, and we disclose the required categories. LLMs helped develop the framing and hypotheses, gave feedback on methodology and experiment design, implemented the harness (the serving stack, the intervention wrapper, the metric collectors, the macro generator) and assisted in interpreting results; additionally code creation, literature search, brainstorming, and editing this text.

We have *not* used generative AI to produce, select or filter any experimental result. Every number is emitted by `collect_metrics.py` and `gen_macros.py` from per-episode records with content-hashed audit trails, and enters the text as a macro naming its source; a quantity without a landed artifact renders as a visible unfilled marker rather than a value. Two episodes in this paper's history make the discipline concrete, and we disclose both: a broken gold standard produced a plausible headline that the corrected labels reversed (§2), and the target metric's ceiling (§6) was computed only after two rounds of mechanism work; the paper reports the metric critique instead of a third mechanism. We reviewed all AI-assisted work and take responsibility for every claim, number and artifact.

REPRODUCIBILITY STATEMENT

The artifact contains the per-episode records for all 60 run directories behind Tables 1, 2 and 4 (task id, trial index, reward, full tool trace, and for every intervention arm the scorer's decision log), together with the four single-writer scripts that turn those records into every number in this paper on CPU: the violation analyzer, the episode-label emitter, the resolution bootstrap, and the macro generator. It also contains the forecaster's training and dataset-construction scripts, its held-out evaluation for both the 1.5B head and the 3B ablation, the intervention wrapper, the run scripts with serving details pinned (vLLM, context lengths, user-simulator identity), and `PROTOCOL.md`, the pre-registration §2 judges this paper against. Per-arm content hashes are in `metrics.json` and a per-file manifest in `MANIFEST.json`. One command, `recompute.sh`, stages the records beside the scripts, re-runs the chain and byte-compares what it produced against what shipped; the two steps that need τ-bench's own task tables for the gold action set report that they did not run rather than being skipped in silence. Because the records are shipped with cluster paths inside their tracebacks neutralised, their hashes differ from the pre-scrub ones recorded in `metrics.json`; both are published and the analysis accepts either. The forecaster *weights* are not included: the checkpoints are 21 GB, out of scale for a supplementary archive. The training script reproduces the head from the public base model and the evaluation artifact is shipped, so we state the omission here rather than promise an artifact we cannot attach.

ETHICS STATEMENT

This work aims to reduce unauthorized state-changing actions by tool-using agents, evaluated in simulated customer-service environments with no human subjects and no personal data. One misreading to guard against: the forecaster gates *mechanically decidable* policy violations; it neither detects nor bounds harms outside its label definition, and §6's ceiling argument shows how easily a compliance metric can overstate what an intervention achieves. Deployment beyond benchmarks would require a violation definition audited for the target domain.

REFERENCES

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.09024.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment. arXiv:2506.07982, 2025.

Hyungjoo Chae, Namyoung Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation. In *International Conference on Learning Representations*, 2025. arXiv:2410.13232.

Sanjiban Choudhury. Process reward models for LLM agents: Practical framework and directions. arXiv:2502.10325, 2025.

Weidong Huang, Jiaming Ji, Chunhe Xia, Borong Zhang, and Yaodong Yang. SafeDreamer: Safe reinforcement learning with world models. In *International Conference on Learning Representations (ICLR)*, 2024.

Wenhao Lin, Chenyu Yu, Xingwei Lin, Sicong Cao, Xiang Chen, Lei Xue, Le Yu, and Letian Sha. Dreamguard: Efficient runtime guardrail for llm agents via risk-aware world model. arXiv:2608.05695, 2026.

Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2023.

Yutao Mou, Zhangchi Xue, Lijun Li, Peiyang Liu, Shikun Zhang, Wei Ye, and Jing Shao. Toolsafe: Enhancing tool invocation safety of llm-based agents via proactive step-level guardrail and feedback. arXiv:2601.10156, 2026.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations*, 2024. Spotlight; arXiv:2309.15817.

Haoyu Wang, Christopher M. Poskitt, Jiali Wei, and Jun Sun. Probguard: Proactive runtime monitoring for llm agent safety via probabilistic prediction. arXiv:2508.00500, 2025.

Zhiheng Xi et al. AgentPRM: Process reward models for LLM agents via step-wise promise and progress. arXiv:2511.08325, 2025.

Yifan Xiao, Zhongjun Wang, Tiantian Yan, Weiwei Li, Dazhong Wu, and Jie Ma. Time series anomaly detection: A hybrid deep learning framework with adaptive thresholding for electricity industry data. In *IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)*, 2025. DOI 10.1109/CSCLOUD66326.2025.00046.

Yifan Xiao, Shijie Li, and Yuhao Ge. RunbookFX: Type- and effect-safe LLM synthesis for executable incident diagnosis and mitigation. *Proceedings of the ACM on Programming Languages*, 10(ICFP), 2026a. Article 277; DOI 10.1145/3828675.

Yifan Xiao, Shijie Li, and Yu Huang. CALM-TS: Risk-controlled LLM labeling for time-series via calibrated selective gating. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026b. DOI 10.1145/3770855.3817859.

Dexuan Xu, Yanyuan Chen, Zhongyan Chai, Yifan Xiao, Yandong Yan, Weiping Ding, Hanpin Wang, Zhi Jin, Wenpin Jiao, Weihua Yue, Hang Li, and Yu Huang. Knowledge fusion in deep learning-based medical vision-language models: A review. *Information Fusion*, 125:103455, 2026. doi: 10.1016/j.inffus.2025.103455.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.12045.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. R-Judge: Benchmarking safety risk awareness for LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. arXiv:2401.10019.

Yinan Zheng, Jianxiong Li, Dongjie Yu, Yujie Yang, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. Safe offline reinforcement learning with feasibility-guided diffusion model. In *International Conference on Learning Representations (ICLR)*, 2024.

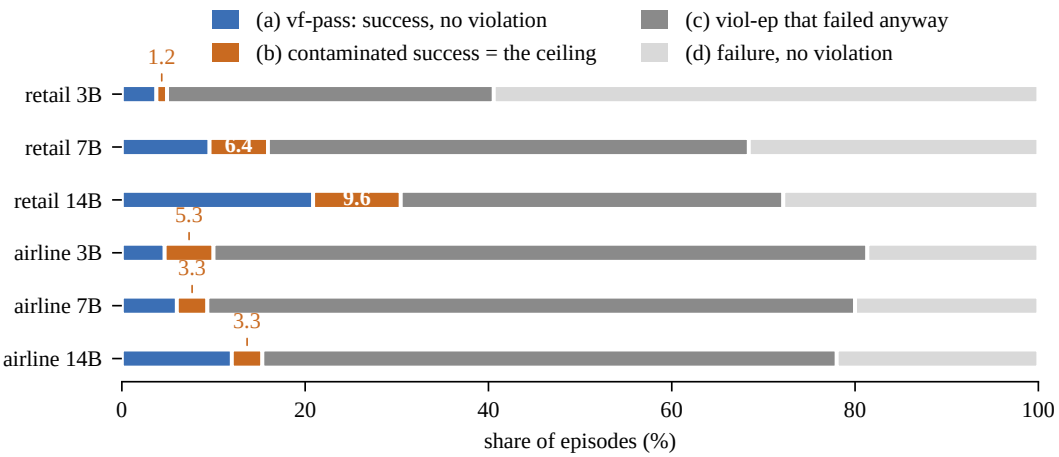

Figure 7: Where episodes go, per baseline arm; bands lettered in the legend so the figure reads without colour. Band (b), successes containing an unsanctioned write, *is* Eq. 1's ceiling; band (c) is why it is small. On retail 3B band (b) is 1.16 pp (4/345), narrower than the measured resolution 3.5 pp.

A WHERE EPISODES GO

Figure 7 is the decomposition behind Table 2 and Figure 1a: per baseline arm, the share of episodes that succeed cleanly, that succeed with an unsanctioned write, and that fail.

B TASKS ARE NOT INTERCHANGEABLE

§2 resamples tasks rather than episodes. Figure 8 shows why: on the three retail baseline arms the 12 tasks that invite the most unsanctioned writes carry 31%–36% of all such writes over three trials, while 18–33 of the 115 tasks produce no violation in any trial. Two episodes of one task are closer to each other than two episodes of different tasks, so the unit whose count sets the resolution is the task.

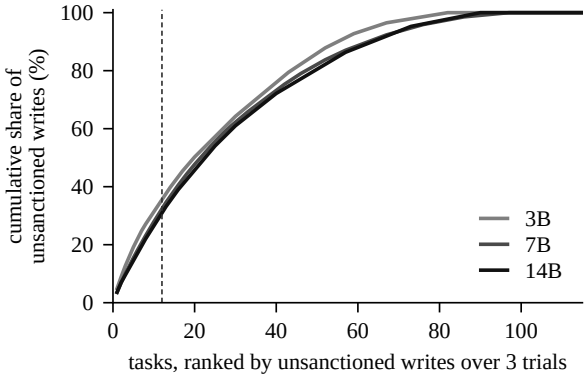

Figure 8: Cumulative share of unsanctioned writes against task rank, three retail baseline arms, three trials each; the dashed line marks rank 12.

C THE FULL MATRIX

Table 1 in §5 carries the arms the text argues from. Table 4 is the whole matrix: both domains, all three agent scales, every scorer condition, the 7B probe block and the 3B decomposition, one row per run. Nothing here is selected: every cell we ran on the violation axis appears, including the ones that resolve nothing.

Table 4: The full matrix: two domains × three agent scales × scorer condition, plus the 7B probe block and the 3B decomposition, whose verdict-carrying arms have a second run. Retail cells are 345 episodes (115 tasks × 3 trials), airline cells 150 (50 tasks); every intervention shares one scaffold (re-sample, clarify, de-escalate; §4) and differs only in what supplies the score — except intercept-all (no scorer) and policy re-inject (no scaffold). Δviol-ep is paired against the same arm's baseline; **95% CI** is its task-clustered paired bootstrap (§2), not a significance test: an interval containing zero means the benchmark cannot resolve that cell. Rows are single runs; the pooled estimators the text quotes merge every run of a condition and are in Table 1. Per-trial spreads: Table 5. Do-nothing anchors: retail 6.09%, airline 38.00% (every airline arm scores below its anchor; that domain carries violation-axis evidence only). The 16 contrasts that meet §2's rule are: retail 3B intercept-all; retail 3B intercept-all (run 2); retail 3B clarify-only (run 2); retail 3B random scorer (run 2); retail 3B random scorer, τ=0.13; retail 3B random scorer, τ=0.13 (run 2); retail 3B +WM v0.1; retail 3B +WM v0.2; retail 3B +WM v0.2 (ckpt seed 1); retail 3B +WM v0.2 (ckpt seed 2); retail 3B +WM v0.2 (run 2); retail 3B +WM v0.2 (run 3); retail 3B +WM v0.3; retail 3B +WM v0.3 (run 2); retail 7B policy re-injection; retail 7B self-critique (run 2).

agent	condition	pass[1]	viol-ep	unsanc/ep	vf-pass	Δviol-ep	95% CI
Retail (trained domain)							
3B	baseline	4.93%	36.81%	1.24	3.77%	—	—
	+WM v0.1	4.93%	23.48%	0.56	4.35%	−13.33 pp	[−20.87, −5.80]
	+WM v0.2	4.64%	22.32%	0.46	4.06%	−14.49 pp	[−20.87, −8.12]
	self-critique	6.67%	38.26%	1.11	4.93%	+1.45 pp	[−5.80, +8.70]
	random scorer	4.93%	27.25%	0.61	4.64%	−9.57 pp	[−16.23, −2.90]
	+WM v0.2, run 2	4.64%	17.97%	0.40	4.06%	−18.84 pp	[−24.35, −13.33]
	random, run 2	6.09%	22.90%	0.53	5.22%	−13.91 pp	[−19.71, −8.12]
7B	baseline	15.94%	58.84%	2.19	9.57%	—	—
	+WM v0.1	13.04%	57.10%	1.23	7.83%	−1.74 pp	[−9.57, +6.09]
	+WM v0.2	16.23%	55.07%	1.30	9.28%	−3.77 pp	[−11.01, +3.48]
	self-critique	13.62%	49.57%	1.03	8.99%	−9.28 pp	[−15.94, −2.61]
14B	baseline	30.43%	51.30%	1.04	20.87%	—	—
	+WM v0.1	29.57%	51.88%	0.95	23.19%	+0.58 pp	[−5.80, +6.96]
	+WM v0.2	27.25%	52.17%	0.93	20.00%	+0.87 pp	[−5.80, +7.54]
	self-critique	28.12%	49.57%	0.97	20.29%	−1.74 pp	[−8.70, +5.22]
	random scorer	24.93%	50.14%	0.89	19.71%	−1.16 pp	[−8.12, +5.80]
	+WM v0.3 (re-exec block)	27.83%	50.43%	0.85	22.90%	−0.87 pp	[−7.83, +6.38]
Airline (zero-shot transfer domain; violation axis only)							
3B	baseline	10.00%	76.67%	3.28	4.67%	—	—
	+WM v0.2	7.33%	78.00%	2.31	2.00%	+1.33 pp	[−7.33, +10.00]
	self-critique	10.00%	76.00%	3.06	4.67%	−0.67 pp	[−8.00, +6.67]
7B	baseline	9.33%	74.00%	2.92	6.00%	—	—
	+WM v0.2	16.00%	74.67%	2.17	8.67%	+0.67 pp	[−6.67, +8.00]
	self-critique	12.67%	67.33%	1.82	8.67%	−6.67 pp	[−15.33, +2.67]
	random scorer	16.67%	72.00%	2.09	10.00%	−2.00 pp	[−8.67, +5.33]
14B	baseline	15.33%	66.00%	1.81	12.00%	—	—
	+WM v0.2	16.67%	68.67%	1.83	12.67%	+2.67 pp	[−7.33, +12.67]
	self-critique	18.00%	69.33%	1.61	12.00%	+3.33 pp	[−6.00, +12.67]
7B probe block (retail): what owns the effect, and how far runs drift							
7B	random scorer	15.07%	51.88%	1.42	10.14%	−6.96 pp	[−13.62, −0.58]
	random scorer, run 2	12.17%	56.81%	1.53	7.83%	−2.03 pp	[−9.57, +5.51]
	v0.2 rule, τ=0.5	16.81%	52.75%	1.36	11.30%	−6.09 pp	[−12.46, +0.58]
	v0.1 rule, τ=0.35	13.91%	59.13%	1.40	10.14%	+0.29 pp	[−6.67, +7.25]
	baseline, run 2	12.75%	57.97%	2.22	7.25%	−0.87 pp	[−8.41, +6.38]
	+WM v0.2, run 2	14.78%	55.65%	1.33	10.14%	−3.19 pp	[−11.01, +4.64]
	+WM v0.2, run 3	15.36%	55.94%	1.31	10.14%	−2.90 pp	[−9.86, +4.06]
	self-critique, run 2	15.94%	44.93%	1.07	12.46%	−13.91 pp	[−21.16, −6.67]
	intercept-all (no model)	16.52%	52.46%	1.32	10.43%	−6.38 pp	[−14.20, +1.45]
	clarify-only (no re-rank)	12.75%	55.36%	1.72	8.12%	−3.48 pp	[−11.01, +4.06]
	+WM v0.3 (re-exec block)	13.04%	43.48%	0.90	10.72%	−15.36 pp	[−22.03, −8.70]
	policy re-inject (no scaffold)	6.09%	17.39%	0.35	5.51%	−41.45 pp	[−48.70, −34.20]
3B decomposition block (retail): the same probes, at the one scale with a resolvable effect							
3B	intercept-all (no model)	6.96%	16.81%	0.43	6.38%	−20.00 pp	[−25.80, −14.49]
	intercept-all, run 2	7.54%	13.04%	0.29	7.25%	−23.77 pp	[−30.14, −17.68]
	policy re-inject (no scaffold)	3.48%	32.17%	0.86	2.32%	−4.64 pp	[−11.30, +2.03]
	clarify-only (no re-rank)	6.09%	28.99%	0.78	5.22%	−7.83 pp	[−14.20, −1.45]
	clarify-only, run 2	4.06%	25.51%	0.68	3.48%	−11.30 pp	[−17.68, −4.93]
	random scorer, τ=0.13	5.80%	20.87%	0.45	4.93%	−15.94 pp	[−22.32, −9.57]
	+WM v0.3 (re-exec block)	6.96%	14.78%	0.20	6.09%	−22.03 pp	[−28.41, −15.36]
	+WM v0.3, run 2	5.22%	14.20%	0.23	4.35%	−22.61 pp	[−28.41, −16.81]
	+WM v0.2, ckpt seed 1	4.64%	24.06%	0.56	4.64%	−12.75 pp	[−19.13, −6.67]
	+WM v0.2, ckpt seed 2	5.22%	21.45%	0.57	4.64%	−15.36 pp	[−21.45, −8.99]

D PER-TRIAL SPREADS FOR EVERY CELL

`PROTOCOL.md`, committed before any result here existed, asks for pass[1] as a mean with its cross-trial spread. Table 5 supplies that on the violation-episode axis for all 44 intervention cells and their 16 baseline cells: the three per-trial values, their mean (which is the rate quoted in Table 4, since the trials are equal-sized), and their standard deviation. Nothing changes between the trials of a cell except sampling, so the spread within a row is this instrument's noise under a fixed configuration, the comparand §2 judges deltas against.

Is that spread order-free? "Nothing changes" is an assumption about three trials run in sequence against one live user simulator, so we test it (`trial_order.py`, registered in `PREREG_TRIALORDER.md` before the analysis). Over the 60 cells with three equal trials, rank the trials on each axis: if the index carried no information each would be best, and worst, a third of the time. On the violation-episode axis it is neither ($p = 0.08$ and $p = 0.20$, exact multinomial), and the same holds for success (0.71/0.22) and violation-free success (0.90/0.64); no mean trial-to-trial difference on those three axes excludes zero, by a bootstrap over cells that keeps the cells ranking discards to ties. One axis does carry the index (`unsanc_per_ep`): on unsanctioned writes per episode, trial one is the lowest in 30 of 58 untied cells ($p = 0.001$, which survives Bonferroni over the 8 tests in this family), by -0.068 (-0.194 to $+0.048$) writes per episode. The direction is what decides the consequence, and it was fixed before the test ran: the singular trial is the *cleanest* one, so a three-trial range on that axis spans a real systematic gap and therefore *over*-states pure sampling noise. Read the bar as an upper bound there, not an estimate; it costs claims nothing, and had the singular trial been the dirtiest we would owe the opposite correction.

Two features are worth reading directly off it. The violation axis is the noisier of the two axes this paper reports (median per-trial sd $3.51\,pp$ against $1.74\,pp$ on success, a factor of $2.0\times$), which is the opposite of what an earlier draft of this paper asserted from a single baseline re-run. And the widest same-configuration range here is $20.87\,pp$, which is larger than every violation delta this paper reports except 6: retail 7B policy re-injection (-41.45 pp); retail 3B intercept-all (run 2) (-23.77 pp); retail 3B +WM v0.3 (run 2) (-22.61 pp); retail 3B +WM v0.3 (pooled) (-22.32 pp); retail 3B +WM v0.3 (-22.03 pp); retail 3B intercept-all (pooled) (-21.88 pp).

E THE IDEAL REMOVER, ARM BY ARM

For each reported baseline arm (Table 6), every contaminated success is relabelled a clean success and nothing else changes; the gain in violation-free success is then bootstrapped exactly as in §2 (tasks resampled with all their trials, $20{,}000$ draws, three seeds; the interval printed is the first seed's, and *excludes zero* is required under all three). The *bar* column is the pre-registered claim criterion on the vf-pass axis for that arm. *MDE* is the smallest number of cleaned successes at which the interval excludes zero in at least 95% of random subsets of that size. A synthetic construction of 115 tasks with three trials each, one contaminated success in four tasks, reproduces the retail-3B interval $[0.29, 2.32]$.

F WRITES OUTSIDE THE GOLD SET, READ AGAINST THE POLICY

A write outside the gold action sequence is what the harness counts as a violation. To check that this mechanical rule tracks the policy rather than one path to the goal, a seeded stratified sample of 60 such writes (38 retail, 22 airline) was labelled by hand against the domain's policy document. Retail: 37 forbidden by the policy, 1 departing only from the gold path, 3 forbidden writes the environment accepted. Airline: 22 forbidden, 0 gold-only, 5 accepted. The labels ship with the records (`results/policy_label_audit.jsonl`).

G ARCHITECTURE, AND THE WRAPPER'S DECISION FLOW WITH BRANCH COUNTS

Figure 9 separates the four components the paper measures. Figure 10 draws the wrapper of §4 with the branch counts from the 7B v0.2 arm's own decision log, so the diagram and the text cannot drift

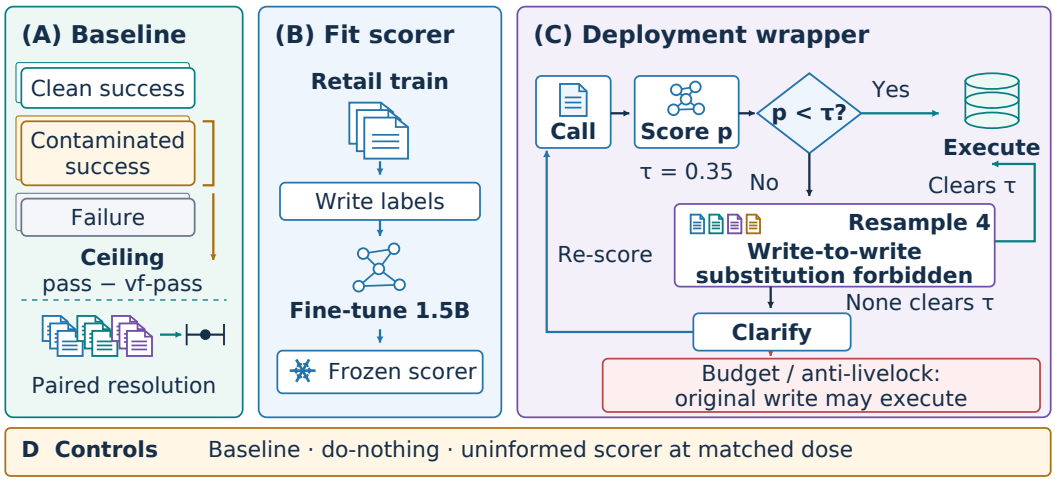

Figure 9: The audit, the forecaster and the v0.2 wrapper are distinct components, not a training chain. (A) The ceiling pass − vfpass applies to success-preserving violation removal; its resolution is the 95% half-width of the task-clustered paired delta. (B) A 1.5B classifier is fine-tuned on retail-train episodes and then frozen. (C) The frozen agent's tool call is scored; a call at or above τ is re-sampled, a flagged write can never be replaced by another write, and if none of the four alternatives clears τ the wrapper asks the user; a re-issued call is scored again, and the anti-livelock rule or an exhausted question budget lets the original write execute. The 0.35 operating point was chosen after reading v0.1 retail test-split logs (§4). (D) Every arm is read against a do-nothing anchor and an uninformed scorer at matched intervention dose. Schematic; the branch counts are in Figure 10.

apart. The shape it makes visible is the one §4 argues from: the anti-livelock override at the bottom is not a rare escape hatch but the single largest destination of a flagged call at 3B and 7B. A reader wanting the corresponding shape at 14B should note that the top branch changes instead: 42% of scored events are passed unflagged there against 24% here, which is why the branches below it carry less traffic.

H THE TWO NULL BLOCKS IN FULL

§5 summarises the arms that resolve nothing; the full reading is here, and every cell is in Table 4.

7B, inside the scaffold. Clarify-only (k=0) keeps the clarification question but removes re-sampling; it fires on 76.5% of scored events against the full scaffold's pooled 76.0%, and moves -3.48 pp ($[-11.01, +4.06]$). v0.3 forbids re-executing a clarified key; against v0.2 it is -11.59 pp ($[-18.55, -4.93]$), one run per side, interval only. Intercept-all costs no measurable capability ($+0.58$ pp, $[-4.64, +5.80]$), and its gap to the random scorer is $+0.58$ pp ($[-6.67, +7.83]$). The one 7B arm that clears the rule is the prompt-only self-critique baseline, and only in one of its two runs: run 2 moves -13.91 pp ($[-21.16, -6.67]$) against a bar of 11.30 pp, run 1 moves -9.28 pp ($[-15.94, -2.61]$) against 14.78 pp. The two runs' own violation rates are 4.64 pp apart, the widest run-level drift in this paper, so we report the verdict and do not build on it. It is unfavourable to this paper's method: a scorer with no trained model clears the rule at 7B where the trained forecaster does not.

3B, against the matched control. Clarify-only (k=0) is short of the full scaffold by $+7.97$ pp ($[+3.48, +12.61]$) against a 9.57 pp bar: dropping re-sampling costs 40.7% against 57.1% blocked at an unchanged flag rate. v0.3 is -4.78 pp ($[-8.41, -1.16]$) beyond the matched control against a 11.30 pp bar, interval only. Policy re-injection does nothing resolvable at 3B (-4.64 pp, $[-11.30, +2.03]$).

14B. Closing the anti-livelock override channel outright (the v0.3 scaffold, which forbids re-executing a clarified key) still moves nothing (-0.87 pp, $[-7.83, +6.38]$; against v0.2, -1.74 pp,

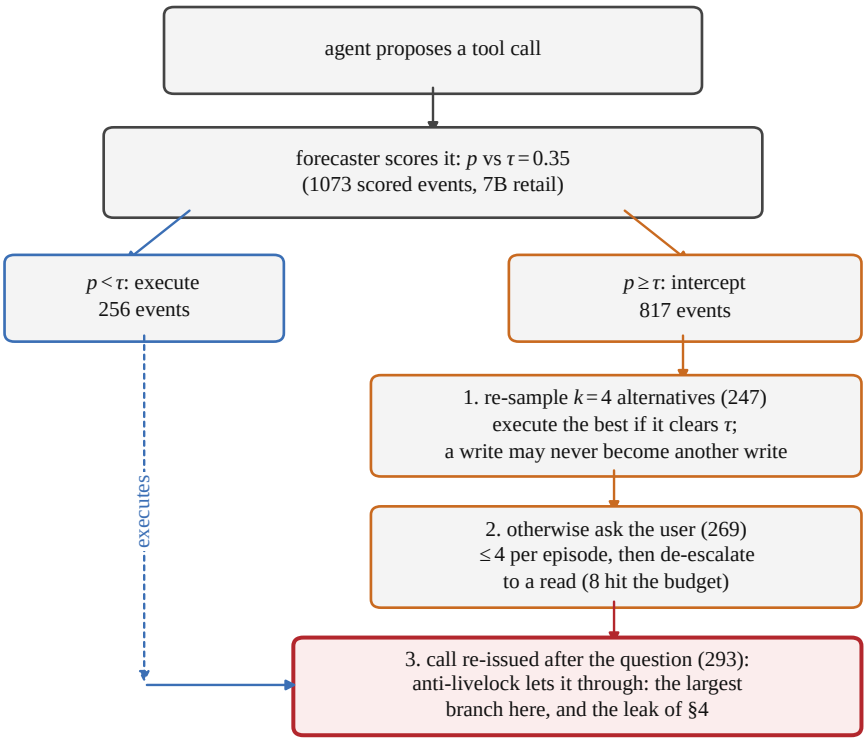

Figure 10: One scored tool call through the wrapper, with the 7B v0.2 arm's decision counts on each branch (1073 scored events over 345 episodes; each count is a scored event's terminal decision, and the five branches partition the events). A call re-issued after a question is scored again and lands in its own branch, so re-issues can exceed questions (1.09 per question). Blue: the call executes. Orange: it is intercepted. Red: it executes anyway, after the clarifying question failed to change it.

$[-8.41, +4.93]$), which is the direct test of the under-flagging diagnosis: blocking what the forecaster does flag cannot touch what it never flags. On the capability side, a same-configuration re-run of the 7B baseline moves success -3.19 pp ($[-7.83, +1.45]$) with nothing changed, which is the scale of drift the withdrawn 14B "capability cost" sat inside; the 14B random arm's success delta (-5.51 pp, $[-10.14, -0.87]$) excludes zero but sits inside the baseline's own success trial range: same verdict.

Per-event recall across runs. Table 3 reads one run per scale, which is what `PREREG_EVENTS.md` registered, and the 14B agent has only one, so its pooled row is the table's row. The other two scales have 3 runs each. Pooled over them, recall is 87.9% ($[85.9, 89.8]$, 1129 events) at 3B and 79.2% ($[77.6, 80.6]$, 3353 events) at 7B, with rank AUC 0.847 and 0.679. Pooling moves the 7B recall by less than the range across its own runs (3.26 pp), and the 14B interval stays entirely below it, so the reading in §7 does not turn on which estimator is used.

Airline. Unsanctioned writes per episode fall at 3B and 7B while violating episodes do not, so violations were redistributed across more episodes rather than removed; the release carries that axis's paired intervals and its verdict under our own rule (the 3B one both excludes zero and exceeds its bar), and we still read it as description rather than transfer; the next paragraph says why. The one airline success interval excluding zero, $+6.67$ pp at 7B, lands at 16.00%, still far below the 38.00% anchor, a change between two numbers worse than doing nothing.

Does the harness matter? Three of the four pairs are four-card and one is three-card, which makes two comparisons possible that a single harness could not support. Neither resolves.

Scale. The 3B and 7B pairs, both four-card, drift $+2.32$ pp and -1.74 pp on the violation axis, a difference far inside either pair's own width, so drift is not detectably scale-dependent here. This is

the first version of that comparison worth making: before the four-card 7B pair existed, differencing a four-card pair against a three-card one would have measured scale and harness at once, which is the confound §5 was already corrected for once.

Harness. At 7B the same arm drifts -1.74 pp on four cards and -0.87 pp on three; and pooling both runs of each side, the four-card level sits $+3.04$ pp ($[-2.03, +8.26]$) above the three-card level on that axis and $+1.88$ pp ($[-1.30, +5.22]$) on success. Judged against the same bar as every other contrast here, neither clears. We report that as a difference this instrument cannot resolve, not as evidence that the harness is irrelevant: a gap below the bar is a gap we cannot see, and the distinction is the one this paper insists on everywhere else.

The same configuration, twice. A cell's three trials share one run; two runs of one cell share nothing but the configuration, and we now have 4 such pairs. They agree on the aggregate and barely agree on which episodes carry it: of the 219 episodes that succeed in either run of a pair, 60 (27%) succeed in both. That is the empirical case for the resampling unit of §2: the paired bootstrap draws *tasks* and carries every episode of a drawn task into every arm, which is the right unit only if an episode's outcome is not a property of the episode. One run cannot show this and the first pair could only suggest it.

Three versions of one number. The 3B forecaster's violation reduction has been measured against three controls, and all three are reported because the choice among them is a methodological claim rather than a preference. -16.67 pp ($[-21.88, -11.30]$) is against the pooled arm that merges a three-card run with a four-card one, the composition MX5 was registered to test. -17.54 pp ($[-23.33, -11.74]$) restricts that arm to its two four-card runs, which removes the confound from inside the arm and leaves it standing across the contrast: it is still differenced against a three-card baseline. -16.38 pp ($[-20.58, -12.32]$) is the first with one serving topology on both sides, which the four-card baselines of `PREREG_MX6.md` made computable, and it is the one §5 now quotes.

The three agree, and that is the result. All clear the same 9.57 pp bar, the widest cross-trial range among the runs each is built from, which the topology restriction does not move because the binding constituent does not change; so the confound was worth removing and removing it does not change what the paper concludes. Had they disagreed, `PREREG_MX6_TOPOMATCHED.md` fixed in advance which one would have been reported and what would have happened to the abstract. The capability axis behaves the same way and is claimed under none of the three: violation-free success moves $+0.29$ pp, $+0.72$ pp and $+1.30$ pp respectively, and every one of those intervals contains zero.

Episodes that never started, and why we do not divide by them. In 58 of the 60 cells some episodes carry an empty trajectory. None of them is an agent that chose to do nothing: all 736 carry a harness error: 413 context-window overflows, 152 malformed action arguments, 143 API timeouts, 27 scorer timeouts (all in one cell, retail 7B v0.3) and one environment-response error. τ-bench drops the whole trajectory when the exception is raised, so a write issued before it is not recorded. They stay in the 3-trial denominator and score zero on both axes, and where present their share of a cell runs from 0.29% to 8.99%. An empty trajectory carries no recorded write, so it *cannot* be labelled as containing a violation (we checked, and 0 of them are), so an arm producing more of them gets a mechanically lower violation-episode rate, in the direction this intervention claims to move. Recomputing every one of the 46 arm-minus-baseline violation deltas (one per intervention cell; App. J separates this count from the 49 paired contrasts) on the non-empty subset moves 21 of them by a tenth or more of their own size, at most 4.30 pp.

We report that and change nothing, for a reason worth stating plainly. The shift runs *our* way: in 34 of 46 deltas the conditional delta is the larger reduction, and 4 nulls become effects. Adopting a denominator after seeing that it converts nulls into effects is the selection this paper's apparatus exists to prevent. It is also barred on its own terms: the rule of §2 judges a delta against cross-trial spreads computed on the full denominator, so a conditioned delta would be measured with an unconditioned ruler, a different and looser rule than this one. The numbers in Tables 4 and after are therefore the unconditional ones, and for most contrasts they are the conservative side; the 4 sign flips are evidence that those particular contrasts sit below what this instrument resolves, not evidence that they are real.

The rule on the two axes we do not claim. `resolution.py` computes the paired delta and interval on four metrics and, until we noticed, applied the claim rule to two of them. It now applies

to all four, which changes no existing verdict and adds a verdict for every contrast on each of the remaining axes. The result is not convenient to leave out: on unsanctioned writes per episode 33 of 49 contrasts pass (every one of them a reduction), against 16 of 49 on the violation-episode axis this paper does claim, and 1 of 49 on violation-free success. We do not promote any of them. The claim set was fixed on the episode axis before the results existed, and adopting a fourth axis because it came out well is the selection this paper's whole apparatus exists to prevent; *declining* a passing result is the only honest form that refusal can take, and it costs nothing to say so here.

Two things stop this from being a headline we are hiding. First, it is the same finding rather than a rival one: writes per episode falling while the rate of episodes *containing* a violation does not is the redistribution the airline paragraph above names, and Table 4 carries both columns for every retail arm, where the pattern is the same: the forecaster removes bad calls without emptying the episodes that carry them, and an agent that commits one violation instead of three is still a violating agent under the metric this paper argues is the right one. Second, this axis is the one whose cross-trial spread is not order-free (App. D), and its bar is conservative for the reason given there, so these verdicts are if anything easy to pass. The honest summary is that the intervention does what it was trained to do, on the axis it was trained on, and that this is not the same as making the agent safe by the measure that matters.

I SCORER FIRING RATES ACROSS SCALES

The main text summarises this in one sentence; the full reading follows. The self-critique baseline (same scaffold, same threshold, probability supplied by asking the agent to audit its own candidate call) looks from Table 4 like a scorer whose ranking against the forecaster reverses with scale. Its decision logs say something less interesting and more useful. At 7B it intervenes on 77.4% of scored events against the forecaster's 76.1%, a comparably-sized intervention; at 3B and 14B it intervenes on 1.7% and 17.9%, with a median score of 0.00 at both. The small and large agents score almost every candidate write as risk-free, so a fixed τ makes this a near-no-op there and the 3B ranking above compares an active intervention against one that barely fires. The transferable point is methodological: a self-critique scorer's output distribution is not stable across model scale, so a fixed threshold silently defines a different intervention at each scale, and any paper that gates on one owes the reader its firing rate.

J CHANGELOG OF WITHDRAWN AND CORRECTED CLAIMS

Every claim this paper has withdrawn or corrected between versions, with what replaced it. The current text carries only the replacement.

Four contrast counts appear in the text and in the shipped console output. The 49 paired contrasts of Figure 2 are the reported set: no pooled estimator, no duplicate pairing, the second family counted apart. Keeping pooled cells and duplicate pairings gives 61; adding the second family gives every registered contrast, 64, of which 24 meet the rule (§5). The 46 deltas of the empty-trajectory check (App. H) are arm-minus-baseline deltas, one per intervention cell with re-runs excluded, not paired contrasts; the two counts enumerate different objects.

Table 5: Violation-episode rate per trial, by cell. Three trials of the same task set under one fixed configuration; the mean is the value in Table 4.

arm	condition	viol-ep per trial	mean	sd
Retail				
14B	baseline	48.70, 50.43, 54.78	51.30	3.14
14B	random scorer	51.30, 42.61, 56.52	50.14	7.03
14B	+WM v0.1	49.57, 53.04, 53.04	51.88	2.01
14B	+WM v0.2	57.39, 50.43, 48.70	52.17	4.60
14B	self-critique	51.30, 46.09, 51.30	49.57	3.01
14B	+WM v0.3 (re-exec block)	53.04, 46.96, 51.30	50.43	3.14
3B	baseline	38.26, 32.17, 40.00	36.81	4.11
3B	baseline, four-card pool	36.52, 33.04, 37.39	35.65	2.30
3B	baseline, run 2	38.26, 29.57, 35.65	34.49	4.46
3B	baseline, run 3	34.78, 36.52, 39.13	36.81	2.19
3B	intercept-all (no model)	13.04, 17.39, 20.00	16.81	3.51
3B	intercept-all, run 2	13.91, 14.78, 10.43	13.04	2.30
3B	clarify-only (no rerank)	32.17, 22.61, 32.17	28.99	5.52
3B	clarify-only, run 2	25.22, 24.35, 26.96	25.51	1.33
3B	random scorer	24.35, 20.00, 37.39	27.25	9.05
3B	random scorer, run 2	26.09, 24.35, 18.26	22.90	4.11
3B	random scorer, τ=0.13	22.61, 20.87, 19.13	20.87	1.74
3B	random scorer, τ=0.13, run 2	18.26, 26.09, 18.26	20.87	4.52
3B	policy re-inject (no scaffold)	33.91, 32.17, 30.43	32.17	1.74
3B	+WM v0.1	23.48, 21.74, 25.22	23.48	1.74
3B	+WM v0.2	20.87, 20.00, 26.09	22.32	3.29
3B	+WM v0.2, ckpt seed 1	26.09, 24.35, 21.74	24.06	2.19
3B	+WM v0.2, ckpt seed 2	22.61, 20.87, 20.87	21.45	1.00
3B	+WM v0.2, run 2	23.48, 16.52, 13.91	17.97	4.94
3B	+WM v0.2, run 3	22.61, 20.00, 19.13	20.58	1.81
3B	self-critique	42.61, 33.04, 39.13	38.26	4.84
3B	+WM v0.3 (re-exec block)	9.57, 13.91, 20.87	14.78	5.70
3B	+WM v0.3, run 2	13.04, 10.43, 19.13	14.20	4.46
7B	baseline	54.78, 55.65, 66.09	58.84	6.29
7B	baseline, three-card pool	52.61, 56.96, 65.65	58.41	6.64
7B	baseline, four-card pool	57.83, 64.35, 62.17	61.45	3.32
7B	baseline, run 2	50.43, 58.26, 65.22	57.97	7.40
7B	baseline, run 3	54.78, 70.43, 61.74	62.32	7.84
7B	baseline, run 4	60.87, 58.26, 62.61	60.58	2.19
7B	intercept-all (no model)	51.30, 56.52, 49.57	52.46	3.62
7B	clarify-only (no rerank)	58.26, 52.17, 55.65	55.36	3.05
7B	random scorer	50.43, 54.78, 50.43	51.88	2.51
7B	random scorer, run 2	55.65, 56.52, 58.26	56.81	1.33
7B	policy re-inject (no scaffold)	16.52, 20.00, 15.65	17.39	2.30
7B	+WM v0.1	58.26, 57.39, 55.65	57.10	1.33
7B	v0.1 rule, τ=0.35	62.61, 61.74, 53.04	59.13	5.29
7B	+WM v0.2	53.04, 55.65, 56.52	55.07	1.81
7B	+WM v0.2, run 2	58.26, 51.30, 57.39	55.65	3.79
7B	+WM v0.2, run 3	54.78, 49.57, 63.48	55.94	7.03
7B	self-critique	50.43, 56.52, 41.74	49.57	7.43
7B	self-critique, run 2	48.70, 41.74, 44.35	44.93	3.51
7B	v0.2 rule, τ=0.5	57.39, 52.17, 48.70	52.75	4.38
7B	+WM v0.3 (re-exec block)	34.78, 40.00, 55.65	43.48	10.86
Airline				
14B	baseline	68.00, 64.00, 66.00	66.00	2.00
14B	+WM v0.2	64.00, 72.00, 70.00	68.67	4.16
14B	self-critique	74.00, 64.00, 70.00	69.33	5.03
3B	baseline	72.00, 80.00, 78.00	76.67	4.16
3B	+WM v0.2	80.00, 72.00, 82.00	78.00	5.29
3B	self-critique	78.00, 72.00, 78.00	76.00	3.46
7B	baseline	74.00, 74.00, 74.00	74.00	0.00
7B	baseline, run 2	80.00, 80.00, 74.00	78.00	3.46
7B	baseline, run 3	70.00, 76.00, 86.00	77.33	8.08
7B	random scorer	70.00, 76.00, 70.00	72.00	3.46
7B	+WM v0.2	78.00, 72.00, 74.00	74.67	3.06
7B	self-critique	76.00, 60.00, 66.00	67.33	8.08

arm	contam. succ.	gain	95% CI	$\neq 0$	clears bar	MDE
retail 3B	4	1.16 pp	$[0.29, 2.32]$	yes	no	4
retail 3B (four-card, pooled)	11	1.59 pp	$[0.58, 2.90]$	yes	no	5
retail 7B	22	6.38 pp	$[3.19, 10.14]$	yes	yes	5
retail 14B	33	9.57 pp	$[5.80, 13.91]$	yes	yes	5
retail Llama 8B	14	4.06 pp	$[1.44, 7.54]$	yes	yes	6
airline 3B	8	5.33 pp	$[2.00, 9.33]$	yes	no	5
airline 7B	5	3.33 pp	$[0.00, 8.00]$	no	no	>5
airline 14B	5	3.33 pp	$[0.00, 7.33]$	no	no	>5

Table 6: An ideal success-preserving remover applied to each of the 8 baseline estimates: the six arms of Table 2, Llama 8B, and retail 3B again, pooled over its two four-card runs. Gains and intervals in percentage points of violation-free success.

claim as first written	why it was withdrawn	what stands now
A flat 3 pp floor from a sibling harness, under which eleven contrasts cleared the bar	the pre-registered criterion is the measured cross-trial spread of both arms; the constant was substituted after results existed	§2: the measured spread, 16 of 49 contrasts
The gold action set (a missing task field made every write count as unsanctioned)	recomputation from raw records with a content-hashed audit trail reversed a headline	§2: corrected labels; every number recomputed from records
The trained-versus-uninformed gap at the default threshold ($+4.93$ pp), carried into the abstract because it replicated	it does not clear its 17.39 pp bar; replication is not a second route past the bar	§5 RQ2: interval only
Intercept-all removes *more* violations than the forecaster	the bar was read off pooled cells whose spread pooling deflates; over the constituent runs the difference is short of 9.57 pp	§5 RQ2: not distinguishable at this resolution
The withdrawal above was caused by the control mixing two serving configurations	a third forecaster run inside one configuration is no closer to its configuration-mate than to the other run	§5 RQ1: run-to-run scatter, not a configuration term
Clarify-only (k=0) claimable in its own right	same deflated-bar defect	§5 RQ2: interval only
Flag rate as the scaffold's dose	three arms flagging within 0.7 pp of each other remove very different amounts; overridden and budget-exhausted calls execute unchanged	§5 RQ2: blocked-write rate, and the dose line
14B is inert because of the anti-livelock override channel	the decision histogram shows the override is the smallest flagged branch at 14B	§7: far fewer calls are flagged there
The forecaster under-flags at 14B because it never saw that scale in training	the corpus contains the 14B agent (§3)	§7: per-event recall measured, 64.1% against 80.2%
A 14B capability cost (-3.19 pp)	its interval spans zero	§5 RQ3: withdrawn