SKILL LIBRARIES CHANGE WHAT AN EDIT DOES: WHEN SHOULD AGENT SKILL VALIDATION RUN IN CONTEXT?

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-improving language agents keep a skill library of learned instructions, and they keep editing it. Every edit must pass one small test before it enters the library: a validator runs the candidate on a few episodes and accepts it only if performance improves. Current systems run this test in isolation, although at deployment the edit will live inside the library, surrounded by other instructions that share its prompt. In principle that difference matters: prompt context is known to change model behavior. In practice, on ScienceWorld with a 7B model, we present the first controlled measurement of that difference. Library context does reshape edit effects, shifting an edit's measured benefit substantially across contexts, yet in none of the 24 controlled units does it change the accept/reject decision at a practical threshold. A simple risk analysis explains both facts: a verdict survives a context change unless the distortion of isolated testing, or the noise of a small test, rivals the edit's true effect. Pre-registered equal-cost experiments test this analysis; the verdict is budget-dependent. At small budgets both validators are unreliable: run-to-run noise, not context distortion, is the main bottleneck, and both frequently misjudge the sign of an edit's effect. At full budget neither shows an advantage. On 64 edits produced by real updaters the two are statistically equivalent at every decision threshold a deployment would act on, within the single-model, score-granular scope tested here. The result is a simple routing rule: validate in the deployment context only when the distortion of testing in isolation exceeds the noise of a small test; otherwise isolated validation is the cheaper correct choice.

1 INTRODUCTION

An agent that improves itself must decide, repeatedly, whether to keep what it just learned. Skill-induction pipelines write a new instruction after a success (Wang et al., 2024a; Zhao et al., 2024; Ouyang et al., 2026); self-refinement loops store a lesson after a failure (Shinn et al., 2023; Madaan et al., 2023); memory distillation (Park et al., 2023; Packer et al., 2023; Xu et al., 2025) and prompt optimization (Zhou et al., 2023; Yang et al., 2024; Yüksekgönül et al., 2024; Agrawal et al., 2025) edit a persistent library the same way. In every case a validator runs the candidate on a few episodes and decides accept or reject under a fixed budget. Standard systems run this test in *isolation*: fresh episodes, candidate alone, without the library that will surround it at deployment (Wang et al., 2024a; Shinn et al., 2023; Gao et al., 2026; Fang et al., 2025a; Yang et al., 2026). Yet the library is part of the prompt, and prompts move behavior. Format perturbations shift accuracy by tens of points (Sclar et al., 2024), relevant information is diluted in long contexts (Liu et al., 2024), and in model editing an update that fixes one fact perturbs others (Meng et al., 2022; Yao et al., 2023b). If the surrounding library changes what an edit does, then isolated validation measures a different quantity than the one deployment cares about, and the error survives every additional rollout.

This paper asks what evidence a single accept/reject decision needs (Figure 1). We study **per-edit acceptance** (one edit, one decision, one budget), the same gate Voyager uses to promote a skill and Reflexion uses to store a lesson (Wang et al., 2024a; Shinn et al., 2023; Zhang et al., 2025); Top-1 candidate selection requires its own pre-registered study and is excluded. Our "skill library" is a set of short instruction snippets injected into the agent's prompt, an instruction-level working memory (Li et al., 2026), so an "edit" is an instruction rewrite or insertion and "library context" is

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

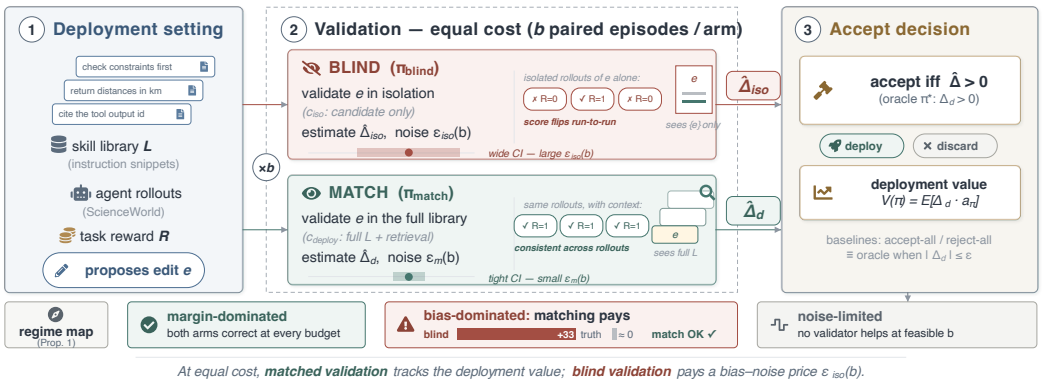

Figure 1: **The decision problem and the three regimes it induces.** ① A pipeline operating on a skill library L proposes an edit e (a diff). ② Two equal-cost validation arms: BLIND measures $\hat{\Delta}_{\text{iso}}$ with the candidate isolated from L; MATCH measures $\hat{\Delta}_d$ inside the deployment library (full L + retrieval), at equal budget b. ③ A threshold rule accepts iff $\hat{\Delta} > 0$; deployment value is $V(\pi) = E[\Delta_d \cdot a_\pi]$. The decision-risk bound of Proposition 1 partitions the (context shift, margin, noise) space into three regimes; our pre-registered test instantiates all three (bottom band; measured examples in §7).

the surrounding instruction set. We compare two validators at **equal cost**: **blind** (isolated) validation versus **matched** (deployment-library) validation, against the trivial baselines accept-all and reject-all and the oracle that accepts iff the edit's true deployment effect is positive ($\Delta_d > 0$). The controlled study runs on one model (Qwen2.5-7B-Instruct) by design, because paired rollouts at every budget must stay affordable; §5 adds exploratory cross-model checks.

We report three findings. A decision-risk bound formalizes the blind–matched difference and predicts the one regime where matched validation should win (§3); a pre-registered equal-cost protocol makes this prediction testable (§4). Library context modulates edit effects but rarely flips accept/reject decisions (§5), and the natural edits real updaters produce today are noise-level (§6). In the pre-registered test, noise dominates before bias: at small budgets both validators are equally unreliable, at the full budget neither is better, and on 64 natural edits the two are equivalent (§7).

Our contributions:

- **Measurement:** the first controlled measurement of how library context modulates skill-edit effects, including a no-edit floor control that exposes a context-specific priming artifact;
- **Theory:** a decision-risk bound that organizes validation outcomes into a regime map over context shift, edit margin, and test noise, identifying the condition under which matched validation dominates;
- **Experiment:** a pre-registered, equal-cost validation experiment with a seed-pinned reproduction script;
- **Audit:** a zero-rollout audit protocol that caught real artifacts in our own pipeline and transfers to other library-editing pipelines.

2 Related Work

Skill libraries and self-improving agents. Voyager induces and verifies skills (Wang et al., 2024a); Reflexion stores self-reflections (Shinn et al., 2023); ExpeL distills insights (Zhao et al., 2024; Ouyang et al., 2026); Generative Agents manage memory streams with recency/relevance scoring (Park et al., 2023; Xu et al., 2025); MemGPT pages context (Packer et al., 2023); ReAct interleaves reasoning and acting (Yao et al., 2023a) and Toolformer learns tool APIs (Schick et al., 2023). These works validate updates implicitly (success counters, verifier calls); we measure what per-edit validation actually needs, in isolation versus in deployment context.

Table 1: **Positioning: validation strategies for a proposed skill edit.** ✓ yes, ○ partial, × no. Only context-matched validation couples the estimate to the deployment context at matched rollout cost under a pre-registered decision rule.

	Isolated re-val.	LLM judge	Shift-corrected	Context-matched
Validates in deployment context	×	×	○	✓
Measures Δ_d directly	×	×	○	✓
Equal-cost comparison enforced	✓	○	✓	✓
Noise-vs-margin regime analysis	×	×	×	✓
Pre-registered accept/reject test	×	×	×	✓

Model editing and prompt optimization. Knowledge-editing methods rewrite facts in place (ROME, MEMIT, MEND, AlphaEdit) (Meng et al., 2022; 2023; Mitchell et al., 2022; Fang et al., 2025b), with known side-effect and locality problems (Yao et al., 2023b); prompt optimizers search instruction space (APE, OPRO, TextGrad, GEPA, SkillOpt) (Zhou et al., 2023; Yang et al., 2024; Yüksekgönül et al., 2024; Agrawal et al., 2025; Yang et al., 2026). AlphaEdit's null-space projection protects preserved knowledge; our concern is the analogous problem at the instruction level, where the "null space" is not available and must be measured: an edit's deployment effect depends on the library it is injected into (Sclar et al., 2024; Liu et al., 2024); feed-forward memory offers a mechanistic parallel (Geva et al., 2021).

Validation, evaluation, and baselines. Accepting an edit from a finite probe is off-policy evaluation with a tiny budget (Precup et al., 2000; Jiang & Li, 2016; Levine et al., 2020; Thomas et al., 2015; Voloshin et al., 2021), and the accept/reject rule is a fixed-budget decision for which bandit theory gives the vocabulary (Lattimore & Szepesvári, 2020; Audibert et al., 2010; Hoeffding, 1963). The LLM-verifier literature, process reward models (Lightman et al., 2024), self-rewarding loops (Yuan et al., 2024), and generate-then-verify stacks (Madaan et al., 2023), faces the same keep/discard decision and inherits the same context-dependence concern. Our baselines are therefore the standard ones: accept-all, reject-all, the oracle π^*, and static LLM judging, which is *not* a substitute: LLM judges correlate with humans on text quality but carry position and self-preference biases (Zheng et al., 2023; Wang et al., 2024b), and we find judging *deployment value* from static context is far harder still (§5). Table 1 summarizes how these validation strategies differ on the properties our protocol targets.

Agent environments. ScienceWorld (Wang et al., 2022) provides the continuous-score task families used throughout; ALFWorld (Shridhar et al., 2021) serves only as an external-validity boundary; retrieval-augmented generation (Lewis et al., 2020) supplies the retrieval layer our fixed design adopts. GPT-4-class models (Achiam et al., 2023) are the intended deployment substrate; our study is on a scale where full paired rollouts are affordable.

3 Problem setup and the regime map

Decision setting. Let e be an edit, c a library context (snapshot + retrieval budget), $t \sim \mathcal{D}$ a task, and ω, ξ environment-initialization and sampling noise. The effect of e under context c is

$$\Delta(e, c) \;=\; E_{t,\omega,\xi}[R(L_c + e;\, t, \omega, \xi) - R(L_c;\, t, \omega, \xi)]\,, \tag{1}$$

$\Delta_d = \Delta(e, c_{\text{deploy}})$ is the deployment effect, and the *context shift* is $I(e) = \Delta(e, c_{\text{deploy}}) - \Delta(e, c_{\text{iso}})$. An acceptance policy π outputs $a_\pi(e, c) \in \{0, 1\}$ with deployment value $V(\pi) = E[\Delta_d \cdot a_\pi]$, the expectation taken over a run's candidate stream and any randomness in π. The oracle π^* accepts iff $\Delta_d > 0$ (notation summarized in Appendix J). Two hypotheses organize what follows: **H1**, library context modulates edit effects (§5); **H2**, context-matched validation improves deployment value over blind validation (§7).

Decision-risk bound.

Proposition 1 *Let a validator accept iff $\hat{\Delta} > 0$, where $E[\hat{\Delta}] = \Delta_d - I(e)$ and the estimation noise $\hat{\Delta} - E[\hat{\Delta}]$ is $\epsilon(b)$-sub-Gaussian. Let $M = |\Delta_d|$ and suppose $M > |I(e)|$. Then the expected*

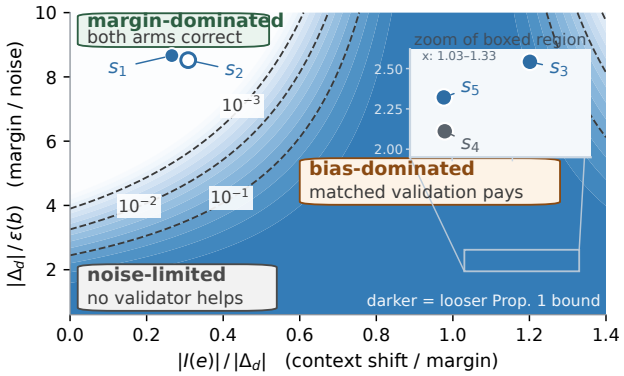

Figure 2: **Regime map of Proposition 1**: bound on $\text{Regret}/|\Delta_d|$ over context shift $|I(e)|/|\Delta_d|$ and margin/noise $|\Delta_d|/\epsilon(b)$. Reading guide: upper right, matched validation pays; lower left, noise wins and no validator helps. Markers: the five significant-gap measured cells (s_1: `find-non-living-thing`/helpful $+6.8$; s_2: `find-non-living-thing`/format -7.8, the hollow single-pool case-study marker; s_3: `find-non-living-thing`/harmful -9.3; s_4: `find-non-living-thing`/no-edit floor $+6.8$; s_5: test-conductivity (tc)/format dose flip; §5), with ϵ taken from the no-op CI half-width (2.97 for the `find-non-living-thing`cells, Table E1; 1.65 for s_5 from the dose-0 CI, Figure 3; ϵ-proxy sensitivity in Appendix H); s_4 is the gray no-op floor marker (margin 6.27). Markers are plotted with the empty context as the deployment anchor (Δ_d; x = empty−interfere gap over $|\Delta_d|$); under the §7 orientation (deployment = interfering library) s_3 moves to $(0.55, 5.7)$, inside the bias-dominated interior, and the s_4 no-op floor is ill-defined there. Strong-family cells sit at high margin with an above-noise context shift; weak families in the noise-limited band.

per-edit regret against π^* satisfies

$$\text{Regret} = M \cdot P(\text{sign err}) \;\le\; M \cdot P(|\text{noise} - I(e)| \ge M) \;\le\; 2M \exp\!\left(-\tfrac{(M-|I(e)|)^2}{2\epsilon(b)^2}\right), \quad (2)$$

and $\text{Regret} \le M$ *unconditionally (proof in Appendix C).*

Two remarks connect Eq. 2 to practice. (1) *Shift-corrected validation cannot substitute for matched validation*: if $I(e)$ itself is estimated ($\hat{I} = I + W$), the validator effectively operates with $\epsilon' = \sqrt{\epsilon(b)^2 + \text{Var}(W)}$, so correcting the shift by estimating it never beats measuring Δ_d directly: the estimate itself consumes deployment-context rollouts, so a shift-corrected isolated arm is strictly noisier at the same total budget; this is why H2 tests matched validation. (2) *The H2 headline is the aggregate of the bound*: summing per-edit regret over the N candidates of a run gives $V(\pi^*) - V(\pi) = \sum E[|\Delta_d| \cdot \mathbf{1}\{\text{sign err}\}]$, so the bound and the experiment measure the same quantity at different granularities.

Regime map. The bound partitions edit-decision problems into three regimes (Figure 2): **margin-dominated** ($|\Delta_d| \gg \epsilon$, $|I(e)| \ll \epsilon$): both validators are near-oracle; **bias-dominated** ($|\Delta_d| \gg \epsilon$, $|I(e)| \gg \epsilon$): blind validation fails while matched validation, which estimates Δ_d directly, pays. Comparing the two bounds exactly, matched strictly dominates iff $M/\epsilon_m > (M - |I(e)|)/\epsilon_{\text{iso}}$, with sufficient condition $\epsilon_m/\epsilon_{\text{iso}} < M/(M - |I(e)|)$ (Appendix C); **noise-limited** ($|\Delta_d| \lesssim \epsilon$): no validator resolves the decision at feasible cost, and accept-all/reject-all are equally near-oracle. A positive practical threshold θ enlarges the near-oracle regime of isolated validation (Appendix C), which is why we report the θ=5 grade alongside zero-threshold counts. Modulation (H1) does not entail a matched-validation advantage (H2): flips are rare because contexts usually agree in sign; the decision-relevant question is whether that sign agreement survives when contexts disagree more, which only the bias-dominated regime rewards.

4 Measurement Protocol

Generation, probe, and reference pools are frozen before candidate selection, with machine-checkable pool manifests (environment, family, variation id, split role, run id) forbidding reuse;

Table 2: **Main H1 results: per-context edit effects for all 24 units** (4 families × 6 edits; Qwen2.5-7B, 150 paired tasks/arm per cell, paired bootstrap R=2000; mean paired score diffs vs. matched baselines, points). **Bold** gaps: 95% CI excludes zero (4/24). Modulation is family-specific: large gaps on `find-non-living-thing`, within-noise spread on find-living-thing (flt)/use-thermometer (ut), and sign flips above the no-op floor (the zero-effect baseline row) on tc: the empirical basis for the regime map (Figure 2). Full `find-non-living-thing` CIs: Table E1; threshold grading: Table D1.

		context			gap (empty − interf.)	
Family	Edit	empty	neutral	interf.	mean	95% CI
find-non-liv.	helpful	+25.7	+11.3	+18.9	**+6.84**	[+0.10, +13.67]
	format	+25.3	+32.1	+33.1	**−7.81**	[−14.67, −0.54]
	harmful	+7.5	+15.6	+16.8	**−9.28**	[−16.21, −2.08]
	paraphrase (deep)	+2.7	+2.1	+1.1	+1.66	[−4.51, +8.23]
	paraphrase (mild)	+0.0	−1.3	−0.7	+0.71	[−5.40, +6.87]
	no-op (floor)	+6.3	+0.3	−0.6	**+6.82**	[+0.41, +13.49]
find-living	helpful	+1.6	+2.6	+1.3	+0.30	[−4.51, +4.83]
	format	−1.4	+0.1	+0.5	−1.83	[−5.94, +2.41]
	harmful	−1.6	−0.2	−1.7	+0.02	[−3.78, +3.80]
	paraphrase (deep)	−1.6	−0.4	−0.3	−1.28	[−5.87, +2.72]
	paraphrase (mild)	−0.9	−0.4	−1.4	+0.55	[−3.46, +4.60]
	no-op (floor)	−0.6	+0.1	−0.5	−0.06	[−4.21, +4.44]
test-cond.	helpful	−1.7	+1.8	+1.0	−2.67	[−7.77, +2.73]
	format	−3.4	+1.8	+1.9	−5.27	[−10.43, +0.07]
	harmful	−1.7	+1.0	−0.1	−1.66	[−6.90, +3.63]
	paraphrase (deep)	−0.7	−0.7	−0.1	−0.66	[−6.03, +4.60]
	paraphrase (mild)	−0.6	−1.1	−0.6	+0.00	[−5.13, +5.10]
	no-op (floor)	+0.4	−0.7	−0.7	+1.07	[−3.97, +6.33]
use-thermo.	helpful	+0.0	+0.1	−0.6	+0.62	[−2.32, +3.56]
	format	−0.6	+0.1	+0.1	−0.74	[−3.68, +2.24]
	harmful	−0.4	+0.3	−0.6	+0.14	[−2.62, +3.06]
	paraphrase (deep)	−0.2	+0.2	+0.1	−0.30	[−3.38, +2.82]
	paraphrase (mild)	−0.1	+0.1	+0.4	−0.44	[−3.44, +2.48]
	no-op (floor)	−0.5	+0.2	+0.1	−0.58	[−3.60, +2.52]

a zero-rollout audit caught earlier cross-experiment overlaps and silent out-of-range variations (Appendix B). Every track passes a no-edit A/A gate: with *no* edit in either arm, the pipeline must measure zero (byte-identical replay, prompt consistency, configuration symmetry; TOST (two one-sided tests) equivalence δ_{AA}=2.0). An earlier gate version caught two real defects (load-time randomization, batching jitter at temperature 0), both fixed; the current gate passes on every track. All uncertainty is reported as percentile bootstrap bands (R=2000 unless noted), never single curves, since small-sample interval coverage at $b \leq 8$ is 24–59% against nominal 95%. The headline contrast at the feasible n=40 reference pairs is a direction/non-inferiority test (minimum detectable effect ≈5.9 points at 80% power; the confirmatory n≈229 design is pre-registered, not budgeted).

5 H1: LIBRARY CONTEXT MODULATES EDIT EFFECTS

Each cell of Table 2 reports the effect of the same edit under three library contexts (empty / neutral / interfering), bootstrapped over tasks. For the other three families the modulation is small or within noise (ut spread $\leq$ 1.13; flt $\leq$ 1.83; tc has sign flips above its no-op floor). The largest paired contrasts and full per-context CIs are in Appendix E (Tables E2/E1); the four-tier grading of all 24 units: 5 interaction-significant, 10 point-estimate flips (7 sub-noise), 2 supported direction flips, **0 beyond the practical threshold** θ=5 (Holm-corrected where a family-wise claim is made).

Interference dose drives the format effect across zero. Interfering entries move the format edit's effect on test-conductivity from −3.83 at dose 0 (independent n=60/dose sample; the modulation run estimates −3.4 for the same cell, Table 2) through a zero-crossing between doses 3 and 6 (Figure 3): modulation is causal in context content, not context choice.

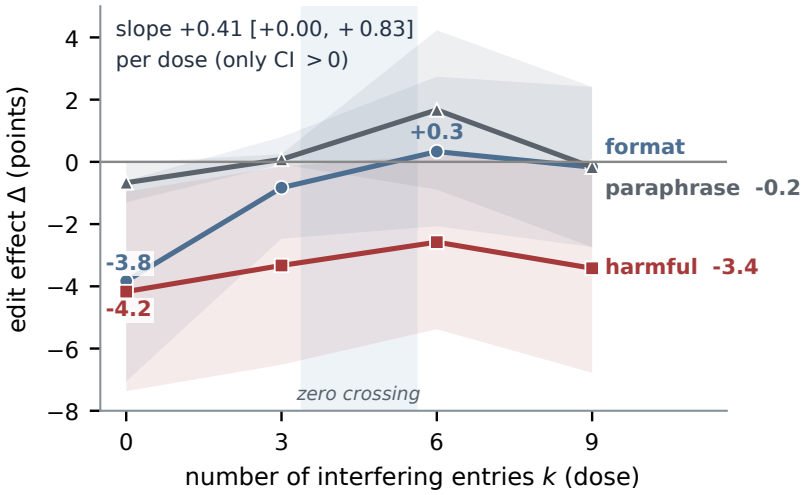

Figure 3: **Dose-response** (test-conductivity, n=60/dose): the format edit's effect on test-conductivity crosses zero as interfering entries accumulate (-3.83 $[-7.06, -0.60]$ at k=0 to $+0.33$ at k=6; bootstrap slope $+0.41$ $[+0.00, +0.83]$ per dose, the only kind whose CI excludes a negative value); harmful stays negative at every dose; paraphrase is null. Shaded bands: 95% CIs.

The no-op floor is unit-specific. The largest noise floor, `find-non-living-thing`/empty ($+6.27$ $[+0.29, +11.98]$, Table E1), is neither arm imbalance (identical 150 tasks, 1 episode each) nor drift (104/150 paired diffs exactly zero) but a concentrated *any-entry priming* effect: appending even a contentless entry under an empty context flips a cluster of 46/150 tasks to nonzero gains (heavy-tailed: $+92, +58, +58$ on the top three). It vanishes under richer context, is absent in tc, is excluded from flip counts, and is drawn as the floor control in Figure 2.

Label validity: the nominal-harmful cell is a measurement artifact. A full audit of all 450 paired diffs: of the 144 positive diffs, 125 are zero-base success flips (all-or-nothing failure $\rightarrow$ completed episode; median diff 0; top gains $+92, +92, +83$), 19 are gains on partially credited bases; of the 31 negative diffs, 23 are working-episode kills (5.1% of pairs). The nominal *harmful* label conflates a large rescue rate with a small destruction rate and fails as ground truth in both directions. Every headline claim is re-derived under measured-sign relabeling (Appendix G): all §7 numbers are unchanged (the harmful cell contributes $D = 0$ either way), and the static-judge accuracy below is scored against measured signs; human labeling ($\kappa \geq 0.6$) is pre-registered for camera-ready.

Static judging is not a shortcut. An LLM judge reached accuracy 0.333 (CI $[0.167, 0.500]$) against constant baselines accept-all 0.458 / reject-all 0.542 (statistically indistinguishable from both) and misjudged the highest-benefit edit (`find-non-living-thing`/format $+31.1$ $\rightarrow$ judged -2.2; reference effect, Table G1). Because 20/24 reference labels are themselves uncertain, this is exploratory.

Generality probes (exploratory). On the weak-effect families, three further models show no significant gaps, as the regime map predicts. On `find-non-living-thing`, no second model passed the fixed-format parse gate (§8); the helpful-only pattern replicates directionally where measurement was possible, and the gate-failing Qwen3-4B-Instruct arm even shows a context *sign flip* (helpful $+7.3$ $[+2.6, +12.3]$ empty vs. -6.6 $[-9.5, -3.5]$ interfering). In 6 ALFWorld families, 2 of 6 show nominal sign flips, but effect sizes (spread 0.07 on 0/1) are not numerically comparable.

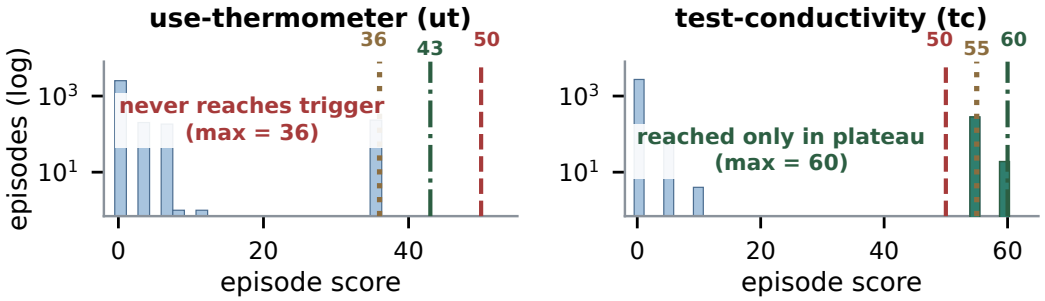

Figure 4: **The natural-edit pipeline's reflection trigger is a scale artifact** (score distributions of the 24-unit run, log-scale counts). use-thermometer never reaches the absolute trigger of 50 (max 36 over 3,150 episodes; audited ceiling 43, plateau 36); test-conductivity reaches it only in its plateau (max 60). The post-audit pipeline replaces the trigger with a ceiling-normalized relative shortfall.

6 The natural-edit track: audit-guided redesign

A zero-rollout audit of our own natural-edit pipeline (full record in Appendix B) re-interpreted its third round as a *measurement diagnosis* rather than an outcome. Four artifacts: (i) the reflection trigger (score < 50) is unreachable for use-thermometer and plateau-only for test-conductivity (Figure 4); (ii) full-text dedup structurally rejects all REWRITE proposals (18/18 second-round rejects are same-title rewrites with overlap > 0.7, while accepted candidates overlap 0.51/0.62); (iii) the pre-audit pipeline's "context" was a fixed full-library serialization, so presentation claims are artifacts until re-run; (iv) task-pool overlap across experiments. The post-audit pipeline freezes per-edit acceptance on natural candidates, diff-level dedup plus a raw-proposal log, ceiling-normalized triggers, dynamic validation, equal-cost validation arms with the full baseline set (reject-all, accept-all, the pipeline's native greedy gate), and a global $(K{+}1) \cdot b$ execution budget over the K arms with a token ledger.

Diagnostic outcome. Under the post-audit pipeline all process gates clear and the audit's ceiling predictions hold empirically: the old rule would have stalled on a no-op rewrite; the new rule rejects exactly one such proposal. Natural candidates are weak everywhere: effect size $|\hat{\Delta}| \leq 6$ raw-score points across cells, every bootstrap CI crossing zero, and a dual-updater audit showing neither updater moves base success (ut 15/30→15/30, tc 6/30→6/30) amid near-total proposal redundancy; they sit squarely in the noise-limited regime. A retrieval-vs-injection check verifies the retriever injects the full library (all 3 entries in every arm) and finds all 24 cells (2 families [ut, tc] $\times$ 4 delivery conditions [full injection, retrieval-only, length-matched fillers, isolated candidate] $\times$ 3 edits) null with max $|\Delta| = 6.0$, the same floor: retrieval-only delivery neither manufactures nor destroys weak-edit effects, and ϵ_m vs. ϵ_{iso} is statistically indistinguishable. Threshold-sensitivity over the 24 cells (Appendix D): accepted counts decay 5→2→1→0 as θ rises 2→3→6: decisions in this regime are entirely noise-driven, the operational form of the map's $|\Delta_d| \lesssim \epsilon$ prediction.

7 H2: does context-matched validation buy value?

Baselines. The value contrast $D = V(\pi_{\text{match}}) - V(\pi_{\text{blind}})$ is read against: the oracle $V(\pi^*) = E[|\Delta_d| \cdot \mathbf{1}\{\Delta_d \neq 0\}]$; accept-all ($V = E[\Delta_d]$) and reject-all ($V = 0$), near-oracle in the noise-limited regime; shift-corrected isolated validation (noisier at equal total cost; Remark (1) of §3); and static judging (§5), already indistinguishable from the constants. Since every natural candidate cleared $|\Delta| \leq 6 \approx$ the noise floor, a frozen natural-candidate run would be uninformative by construction; the decisive test was therefore pre-registered on *strong* edits with frozen deploy deltas.

Cost accounting. Equal-cost holds on episodes, not tokens: the match arm serializes the full library into every validation prompt (≈220 extra tokens/episode, 1.2× overhead here; larger libraries push it higher), and a token-denominated D would be *more* negative; we keep episodes because the pre-registered design froze that axis.

Table 3: **Per-cell means and accept decisions** (16 paired episodes/cell; fnlt = `find-non-living-thing`; accept iff cumulative mean > 0; A = accept, r = reject; bracketed intervals: 10,000-resample task-level bootstrap 95% CIs of mean@16, reproducible from the released per-task records). Reading guide: `find-non-living-thing`/format is the one cell where blind sign-errors at $b \leq 4$ (r r A A) and match is correct from b=2; the tc CIs are near-degenerate by construction (15 zeros and one $\pm$50-point task per cell).

Cell (fam./edit)	Frozen Δ_d	blind b=2/4/8/16	match b=2/4/8/16	mean@16 [95% CI]	
				blind	match
fnlt/helpful	$+18.89$	A A A A	A A A A	$+12.5$ $[-8.9, +31.2]$	$+9.9$ $[-5.1, +25.0]$
fnlt/format	$+33.12$	r r A A	A A A A	$+13.0$ $[-9.4, +32.2]$	$+21.4$ $[+5.8, +35.4]$
fnlt/harmful	$+16.83$	r r r r	r r r r	-3.6 $[-17.6, +9.4]$	-20.8 $[-34.9, -7.8]$
tc/helpful	$+1.00$	A A A A	r r r A	$+3.1$ $[0, +9.4]$	$+3.1$ $[0, +9.4]$
tc/format	$+1.87$	A A A A	r r r r	$+3.1$ $[0, +9.4]$	0.0 $[0, 0]$
tc/harmful	-0.07	r r r r	A A A A	0.0 $[0, 0]$	$+3.1$ $[0, +9.4]$

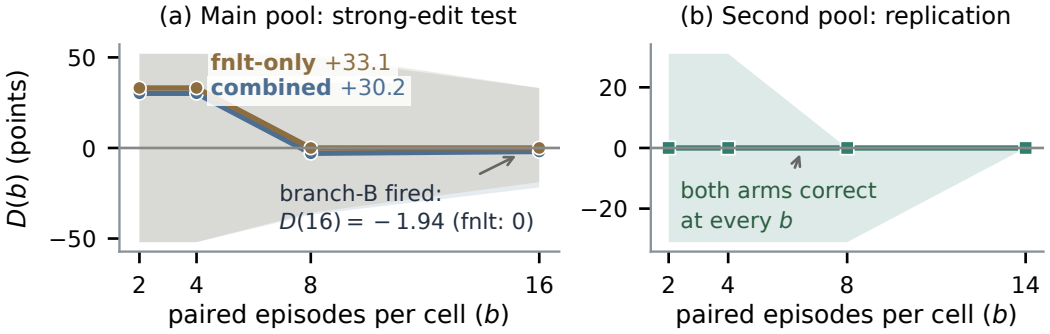

Figure 5: **Decision advantage** $D(b) = \sum_{\text{cells}} \Delta_d \, (a_{\text{match}} - a_{\text{blind}})$ vs. budget b; bands: 10,000-resample task-level bootstrap 95% CIs recomputing both decisions at every b. (a) Main pool: the advantage concentrates at $b \leq 4$; the branch-B rule fires at $D(16) = -1.94 \approx 0$ (fnlt-only: 0 exactly). (b) The pre-registered second pool: both arms correct at every b; the small-budget failure is pool-specific and the pool is margin-dominated. The 11-pool confirmation (Appendix F) shows neither arm is decision-reliable at $b \leq 4$ across pools.

Design. Six strong-edit candidates with frozen Δ_d set to the measured interfere-context effects of Table 2 (three `find-non-living-thing`, three tc), two equal-cost arms (blind = candidate isolated; match = full deployment library), cumulative paired budgets $b \in \{2, 4, 8, 16\}$, accept iff cumulative mean paired diff > 0; the pre-registered headline rule (branch B: accept matched validation only if it shows an advantage; Appendix I) fired on the point estimate. Table 3 reports per-cell means, accept decisions, and 10,000-resample task-level bootstrap 95% CIs; Figure 5 plots $D(b)$.

The paired-budget design instantiates all three regimes. *Margin-dominated*: `find-non-living-thing`/helpful, both arms accept at every budget, $D = 0$, as predicted. *Bias-dominated*: `find-non-living-thing`/format, blind sign-errors at small budgets on a true $+33$ edit (-50 at b=2), match correct from b=2; $D = +33.1$ for $b \leq 4$, vanishing by b=16 as blind noise averages out ($\epsilon_{\text{iso}}(b) = 43.3/\sqrt{b}$; the normal-approximation boundary predicts correction at b=8, observed). This within-pool pattern did not survive the pre-registered 11-pool confirmation (below). *Noise-limited*: tc, frozen $|\Delta_d| \leq 1.87$ is far below what 16 paired episodes can resolve, so every acceptance hinges on a single 50-point task (both arms false-accept near-null candidates). Two pool caveats: `find-non-living-thing`/harmful's frozen delta did not transfer across pools (frozen $+16.83$; measured -20.8 match / -3.6 blind on the new band); both sides are measured in deployment context, so this is task-pool sensitivity in the frozen oracle itself, a caveat for every frozen Δ_d grounding $D(b)$, not added evidence for context matching. The pre-registered tc pool was a difficulty cliff (15/16 tasks score 0), voided and re-run on an adjacent band before results were observed.

Neither arm shows a net value advantage at the full budget. Combined $D(b) = +30.2$ at $b \in \{2, 4\}$, -2.9 at b=8, -1.94 at b=16 (fnlt-only $D(16) = 0$ exactly; the three tc cells are pre-registered single-task coin flips). A task-level bootstrap (10,000 resamples within cells, both arms' decisions recomputed at every b) gives $D(16)$ 95% CI $[-21.8, +33.0]$ with $P(D > 0) = 0.26$ (fnlt-only $[-18.9, +33.1]$, $P = 0.22$). The branch-B decision records that the pre-registered rule fired, not a statistical equivalence claim: the CI excludes only advantages of one format-scale unit ($D(16) > +33$); a confirmatory equivalence test (TOST or larger budgets) is specified as future work. The harmful cell's D-contribution is invariant to the frozen-vs-measured discrepancy (both arms reject).

The pre-registered second pool is margin-dominated. The bias-dominated reading rests on one cell in one pool, so the same `find-non-living-thing`/format cell was replicated on a fresh pre-registered band (14 tasks, $b \in \{2, 4, 8, 14\}$): a negative replication, blind was *correct* at every budget ($+50.0$ at b=2; mean@14 $+18.4$, se 6.55), match likewise; $D = 0$; the pool is margin-dominated. The small-budget blind sign error is thus a pool-specific noise realization, and the estimator-noise ordering is pool-dependent ($\epsilon_m/\epsilon_{\text{iso}}$ reverses, 8.68 vs. 6.55; robust-scale IQR/1.349 gives exactly 1.00, a two-point left-tail artifact). The two-pool pattern is superseded by the confirmation below: pool-level regime membership is a random property of the noise realization, the regime map's prediction, now with an empirical bound.

Small-budget decisions are noise-dominated in both arms. The Appendix F design was executed in full (352 episodes; frozen seeds). Sign errors at b=4: match $6/11$, blind $4/11$ (Fisher one-sided $p = 0.90$; at b=2, $4/11$ vs. $4/11$); the pre-registered decision rule ($\geq 2/11$ match errors) closes the reliability claim. Most decisive outcomes are zero-mean no-signal cells rather than reversed signs (excluding them: $2/11$ vs. $3/11$ negative means): at $b \leq 4$ the decision is noise-dominated in *both* arms, the map's no-budget guidance confirmed empirically. The pre-registered run-record risk score that ranked the two original pools failed out-of-sample ($3/11$, below chance).

On natural edits, matched and blind are equivalent. Deployment asks the frozen-delta question about edits real updaters produce. A pre-registered run (64 natural-edit events: 2 environments $\times$ 2 updaters (reflective, direct) $\times$ 8 shared-base groups; 24,576 episodes; frozen seeds; one-shot unblinding behind a per-shard completeness gate) decides each edit from 32 cumulative probe pairs per arm against a 128-pair reference. Pooled $D = -0.04$ points (cluster-t 95% CI $[-0.21, +0.14]$, k=8; cluster bootstrap $[-0.19, +0.09]$; 42% zero edits; largest leave-one-out shift 0.05). The deferred equivalence test resolves: the 90% CI $[-0.18, +0.11]$ lies inside the pre-registered ± 1.5-point margin; matched and blind are equivalent at any threshold a deployment would act on. Small-budget acceptance is noise-dominated ($0.05 \rightarrow 0.33$ from 8 to 64 probe episodes, i.e. 4 to 32 pairs per arm); excluding the 3 events with silently failed references gives $D = -0.04$, 90% CI $[-0.18, +0.09]$, TOST unchanged. The pre-registered cells split by updater, not environment: reflective updates gain in both environments ($+0.06$, $+0.14$), direct updates lose in both (-0.23, -0.12): stable heterogeneity; the pre-registered mechanism check (Appendix E) is null: its routing gate is not met.

8 Conclusion

We asked when validation should run in the deployment context; the answer is a routing rule, not a default. Library context reshapes what an edit does, yet across 24 controlled units it almost never flips the accept/reject decision: a verdict survives unless the distortion of isolated testing rivals small-test noise. At small budgets run-to-run noise makes every policy unreliable; matched validation pays only when the estimated context shift exceeds the noise, otherwise isolated validation is cheaper and correct. On the 64 natural edits the two validators are indistinguishable, and pool-level risk scores did not transfer: budget, not pool selection, is the lever.

Limitations. Findings are established for one primary model and six task families; after parse-gate repairs all four second models parse cleanly (88.6–100%, Table E3), but only Qwen3-8B replicates the format cell ($+7.3$, sign p=0.016): the protocol is portable, effects model-specific. The n=40 headline contrast is directional/non-inferiority (Appendix F); the pre-audit evidence predates the retriever (re-examined post-audit, Appendix A).

Reproducibility statement. All experiments ran on one GPU cluster with pinned dependencies: ScienceWorld 1.2.2 (no seed API; ω enters the paired-diff noise ϵ; pairing by variation), vLLM continuous batching (documented temperature-0 jitter; the 11-pool confirmation onward adds per-request seeds), and the Qwen2.5-7B-Instruct checkpoint hash. Seeds, configurations, and table source data are in the per-run log; audit scripts, raw proposal logs, and frozen pool manifests will be released upon publication; the $D(b)$ bootstrap and per-cell CIs are reproducible from per-task records (10,000 resamples, seed 20260915; Appendix A).

Ethics statement. This study uses simulated text environments (ScienceWorld, ALFWorld) and publicly released model checkpoints; it involves no human subjects and no personal data. The subject matter is deployment safety: small-budget validation produces unreliable accept/reject decisions; such estimates should not gate skill edits in systems whose failures carry real costs. The audit protocol and pre-registration ledger will be released upon publication.

AI use statement. LLM-based assistants were used for manuscript editing and engineering support; all research questions, designs, pre-registrations, runs, analyses, claims, and reported numbers are the authors' own, verified in the authors' pipelines; no AI system generated experimental results.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, et al. GEPA: reflective prompt evolution can outperform reinforcement learning. *CoRR*, abs/2507.19457, 2025.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pp. 41–53. Omnipress, 2010.

Jinyuan Fang, Yanwen Peng, Xi Zhang, et al. A comprehensive survey of self-evolving AI agents: A new paradigm bridging foundation models and lifelong agentic systems. *CoRR*, abs/2508.07407, 2025a.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*. OpenReview.net, 2025b.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, et al. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *Trans. Mach. Learn. Res.*, 2026.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 652–661. JMLR.org, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR*, abs/2005.01643, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

Xiangyi Li, Yimin Liu, Wenbo Chen, et al. Skillsbench: Benchmarking how well agent skills work across diverse tasks. *CoRR*, abs/2602.12670, 2026.

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

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*. OpenReview.net, 2023.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*. OpenReview.net, 2022.

Siru Ouyang, Jun Yan, I-Hung Hsu, et al. Reasoningbank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, 2026.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. *CoRR*, abs/2310.08560, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pp. 2:1–2:22. ACM, 2023.

Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning*, pp. 759–766. Morgan Kaufmann, 2000.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, et al. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023.

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

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*. OpenReview.net, 2021.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI Conference on Artificial Intelligence*, pp. 3000–3006. AAAI Press, 2015.

Cameron Voloshin, Hoang Minh Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Trans. Mach. Learn. Res.*, 2024a.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Annual Meeting of the Association for Computational Linguistics*, pp. 9440–9450. Association for Computational Linguistics, 2024b.

Ruoyao Wang, Peter A. Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? In *Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298. Association for Computational Linguistics, 2022.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, 2025.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*. OpenReview.net, 2024.

Yifan Yang, Ziyang Gong, Weiquan Huang, et al. Skillopt: Executive strategy for self-evolving agent skills. *CoRR*, abs/2605.23904, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*. OpenReview.net, 2023a.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. In *Conference on Empirical Methods in Natural Language Processing*, pp. 10222–10240. Association for Computational Linguistics, 2023b.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR / OpenReview.net, 2024.

Mert Yüksekgönül, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic "differentiation" via text. *CoRR*, abs/2406.07496, 2024.

Jenny Zhang, Shengran Hu, Cong Lu, Robert T. Lange, and Jeff Clune. Darwin godel machine: Open-ended evolution of self-improving agents. *CoRR*, abs/2505.22954, 2025.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: LLM agents are experiential learners. In *AAAI Conference on Artificial Intelligence*, pp. 19632–19642. AAAI Press, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *International Conference on Learning Representations*. OpenReview.net, 2023.

A Reproduction material

A/A gate revision history (root causes of the two defects caught by the first gate, and the current gate's pass record; §4); protocol revision freeze records (pre- and post-audit); the full statistical audit; run configurations, seeds, and the per-run cost record; the per-edit decision statistics of the equal-cost comparison; the $D(b)$ bootstrap of §7 and the per-cell CIs of Table 3 are reproducible from the per-task paired diffs alone with the seed-pinned script; Table A1 itemizes the episode accounting.

Table A1: Episode accounting across all runs (per-task paired records per experiment; the experiment log is the authoritative record).

Run	Episodes	Run	Episodes
Main modulation study (24 units)	12,600	Equal-cost comparison (6 cells)	576
Judge accuracy study	5,040	Second pool replication	56
Model-sweep probes	9,600	Dose-response	1,920
Cross-model fresh-state probes	2,400	Natural-edit track	480
Retrieval-vs-injection check	896	first two A/A gates	480
ALFWorld external-validity probe	5,185	Early cross-model probe	6,480
Early family pilots	13,360	Pipeline pilots and early natural-track runs	2,910
		11-pool confirmation	352
		Natural-edit equal-cost run (64 events)	24,576
Total 86,911			

B Zero-rollout audit of the natural-edit pipeline

Audit method: zero-rollout inspection of the stored run records (frozen proposal, dedup, and pool manifests), summarized in Table B1.

B1 Ceiling audit: use-thermometer max 43 (plateau 36; 10,436/13,050 zeros), test-conductivity max 60 (plateau 55; 5,253/6,090 zeros); the absolute reflection trigger (score < 50) is a scale artifact. **B2 Dedup audit**: 18/18 second-round rejects are same-title REWRITEs; accepted candidates overlap originals 0.514/0.618 (threshold 0.7); raw proposal texts were not persisted (fixed post-audit). **B3 Retrieval audit**: no retriever in the pre-audit run. **B4 Pool overlap**: the pre-audit run's task pool (tasks 340–359) sits inside the dose-response study's range (tasks 300–359); the first and revised A/A gates reuse tasks 300–339. **B5 Decision record**: setting A as main design; Top-1 deferred. **B6 Variation-range audit**: out-of-range variations load *silently* with description "unknown" and yield all-zero scores. Affected: an early run's ut arm (out-of-range variation ids 700–810; void; re-run on legal variants reproduces the null), the post-audit ut pools, and the retrieval check's initial ut probes; the remaining runs are clean. Defenses: smoke gate fails on "unknown" descriptions; pool manifests record per-family bounds.

Table B1: Pre-audit pipeline artifacts and their consequences (audit items B1–B4).

Artifact	Evidence	Consequence
Reward scale	ut max 43 (plateau 36; 13,050 eps); tc max 60 (plateau 55); trigger < 50	trigger near-unreachable; third-round proposals unreadable
Dedup	18/18 second-round rejects = REWRITEs, overlap > 0.7 (accepted: 0.514/0.618)	full-text dedup kills template-keeping REWRITEs
Retrieval	pre-audit "context" = full-library serialization	presentation claims are fixed-injection artifacts until re-run
Pool overlap	pre-audit run's task pool (tasks 340–359) inside the dose-response study (tasks 300–359)	manifests required; cross-experiment reuse confirmed

C Proof of Proposition 1 and related results

Proof of Proposition 1. Write $\hat{\Delta} = \Delta_d - I(e) + Z$ with $E[Z] = 0$, Z sub-Gaussian with parameter $\epsilon(b)$. The rule errs iff $\text{sign}(\hat{\Delta}) \neq \text{sign}(\Delta_d)$. If $\Delta_d > 0$, err requires $\hat{\Delta} \leq 0$, i.e. $I(e) - Z \geq \Delta_d = M > 0$, hence $|Z - I(e)| \geq M$; if $\Delta_d < 0$, err requires $\hat{\Delta} > 0$, i.e. $Z - I(e) \geq -\Delta_d = M$, again $|Z - I(e)| \geq M$. The error event is one-sided (a single tail), so it is *contained* in $\{|Z - I(e)| \geq M\}$ but generally strictly smaller; writing $P(\text{sign err})$ for the true one-sided probability, $\text{Regret} = E[M \cdot \mathbf{1}\{\text{err}\}] = M \cdot P(\text{sign err}) \leq M \cdot P(|Z - I(e)| \geq M)$, and the triangle inequality with $M > |I(e)|$ plus the sub-Gaussian tail bound give $\text{Regret} \leq M\, P(|Z| \geq M - |I(e)|) \leq 2M \exp(-(M - |I(e)|)^2/(2\epsilon^2))$; $\text{Regret} \leq M$ trivially. □

Corollary (regime map). Matched validation estimates Δ_d directly (Z_m with parameter ϵ_m) at the same b: $\mathrm{Regret}_m \leq 2M\exp(-M^2/(2\epsilon_m^2))$. Matched strictly dominates iff $M/\epsilon_m > (M-|I(e)|)/\epsilon_{\mathrm{iso}}$, with sufficient condition $\epsilon_m/\epsilon_{\mathrm{iso}} < M/(M-|I(e)|)$: matched validation wins exactly when the context shift is large relative to the margin *and* the matched-context estimator is not proportionally noisier. The retrieval check pre-registers the $\epsilon_m/\epsilon_{\mathrm{iso}}$ measurement.

Threshold-θ rules. With acceptance iff $\hat{\Delta} > \theta$ ($\theta \geq 0$), the binding error side is the false reject: err ($\Delta_d > 0$) implies $I(e) - Z \geq \Delta_d - \theta$ =M - θ, while the false-accept side requires $Z - I(e) > M + \theta$ and is strictly rarer. Hence $\mathrm{Regret}_\theta \leq 2M\exp(-(M-\theta-|I(e)|)^2/(2\epsilon^2))$ whenever $M > |I(e)| + \theta$: a positive practical threshold enlarges the regime in which isolated validation is near-oracle, which is why the $\theta = 5$ grade is reported alongside zero-threshold counts.

D ACCEPTANCE THRESHOLD-SENSITIVITY (24 CELLS)

Acceptance iff cell $\hat{\Delta} > \theta$, over the 24 (family $\times$ delivery-condition $\times$ edit) cells of §6 (max $|\hat{\Delta}| = 6.0$ = the noise floor). The five nominally positive cells: ut/full/format $+0.8$, tc/retrieval/helpful $+0.5$, tc/length-matched/harmful $+0.5$, tc/length-matched/format $+6.0$, tc/isolated/harmful $+3.0$. Every cell that survives any $\theta \geq 2$ has $\hat{\Delta} \leq$ the noise floor, and no cell is stably accepted across $\theta \in [2, 8]$: the accept/reject pattern is a function of the arbitrary threshold, not of any measurable effect.

Table D1: Accepted-cell count vs. practical threshold θ (24 cells).

θ	0	2	3	4	5	6	8
accepted (of 24)	5	2	1	1	1	0	0
decision flips vs θ=0	—	3	4	4	4	5	5

E ADDITIONAL TABLES

Tables E1–E3 give the per-context modulation detail behind §5 (full CIs, contrast tests) and the parse-failure audit behind the fixed-format action protocol. The format-cell quantitative check behind §7 is reproduced here in full.

Table E1: Full per-context modulation CIs (Qwen2.5-7B, `find-non-living-thing`; paired bootstrap, R=2000; point estimates in Table 2).

Edit	empty	neutral	interfere	gap (empty − interf.)
helpful	$[+20.7, +30.9]$	$[+7.6, +15.1]$	$[+14.6, +23.2]$	$[+0.10, +13.67]$
format	$[+19.4, +30.8]$	$[+27.6, +36.6]$	$[+28.9, +37.4]$	$[-14.67, -0.54]$
harmful	$[+2.0, +13.3]$	$[+11.6, +19.9]$	$[+12.3, +21.6]$	$[-16.21, -2.08]$
no-op	$[+0.29, +11.98]$	$[-1.82, +2.44]$	$[-3.15, +2.21]$	$[+0.41, +13.49]$

Table E2: Largest paired context contrasts (`find-non-living-thing`, exploratory; paired bootstrap over matched task pairs, R=2000; n=150 rows).

Contrast	Estimate [95% CI]	Verdict
$G_{\mathrm{interfere}}$ (interfere vs. empty, harmful)	-14.23 $[-19.57, -8.83]$	significant
$G_e - G_i$ (interaction)	$+14.64$ $[+8.67, +21.38]$	significant
G_{empty}	$+0.41$ (n.s.)	noise
flt interaction	$+2.13$ $[+0.23, +4.33]$	sig., small
ut interaction	$+1.36$ $[+0.16, +2.88]$	sig., small

Quantitative check (format cell, §7). With $M = |\Delta_d| = 21.4$, $I(e) \approx 8.4$, margin $M - |I(e)| = 13.0$, $\epsilon_{\mathrm{iso}}(b) = 43.3/\sqrt{b}$ (sample std of the blind arm's paired diffs; 41.9 population), the sub-Gaussian bound predicts blind self-correction only at $b \approx 44$; the observed correction at b=8

Table E3: Final parse-failure rates on the fixed-format action protocol (the $< 5\%$ usability gate; strong-family arms under the pre-registered model-sweep protocol unless noted). The probe-vs-pipeline gap (0.7–1.4% fresh-state vs. 16.9–17.9% in-pipeline across three decoding configurations) locates the residual trigger in the pipeline's prompt construction rather than in decoding, thinking mode, or parse ability per se.

Model / arm	Final failure rate	Notes
Llama-3.1-8B	23.1%	gate FAIL
Qwen3-8B, retry1 (temp 0.7, max_tokens 96)	21.5%	gate FAIL
Qwen3-8B, retry2 (parse-retry $\times 2$)	16.9% (203/1200)	recovers $\sim$1/4
Qwen3-8B, retry3 (temp 0)	17.9%	gate FAIL
Qwen3-8B, retry4 (temp 0, max_tokens 200)	17.7%	truncation refuted
Qwen3-4B-Instruct-2507	11.5% (138/1200)	non-thinking variant
Mistral-7B-Instruct	32.2%	gate FAIL
Gemma-2-9B-it	48.4%	mid-run aborts
fresh-state probe, Qwen3-8B	0.7% (1/144)	/no_think, greedy
fresh-state probe, DeepSeek-R1-Distill-Qwen-1.5B	1.4% (2/144)	truncated think
Repaired pipeline, Qwen3-8B	0.5–3%	first-try 97.0–99.5%

matches the normal-approximation boundary ($\Phi(-13/\epsilon_{\text{iso}}(b))$: 0.27 at b=4 $\rightarrow$ 0.11 at b=16). The dominance condition holds at b=2 ($\epsilon_m/\epsilon_{\text{iso}} = 0.74 < 1.64$): matched validation dominates at the *estimator* level exactly where the bias-dominated regime puts it, but not in decision-level reliability at small b.

Mechanism check: the updater-type split is not a retrieval-routing effect (§7). A pre-registered mechanism experiment on 16 events (event grid frozen before any outcome was observed; one event excluded for a failed reference) contrasts a retrieval set frozen from the pre-edit library (edited entry still exposed on 97.5% of tasks), and length-matched irrelevant fillers, against the normal-retrieval effect. Routing is null ($+0.10$, 95% CI $[-1.13, +1.72]$), as is the fixed-exposure content effect under the fillers ($\Delta = -0.36$, $[-1.23, +0.70]$), while the normal-retrieval effect on these events is small-negative (-0.63, $[-1.41, +0.13]$): the edit is essentially always retrieved, and neither frozen routing nor stripped context reproduces the deployment effect. The pre-registered routing entry condition is not met; we report the mechanism result and stop.

F Power audit and the pre-registered 11-pool confirmation

The main text's reliability question, does matched validation reduce small-budget sign errors on the true-positive `find-non-living-thing`/format cell, is answered by a two-stage arc: a power audit that motivated the design, and the pre-registered confirmation that closed the claim.

Power audit (design stage). The original evidence was a 0-vs-2 sign-error contrast on 2 pools. After observing zero match-arm sign errors on n pools, the exact one-sided 95% confidence upper bound ($\alpha = 0.05$) on the match error rate is $1 - 0.05^{1/n}$: with $n = 2$ pools this is 0.776, essentially uninformative. Bounding the rate below a target p at the stricter $\alpha = 0.025$ requires $n \geq \lceil \ln 0.025 / \ln(1-p) \rceil$ pools: 11 for $p = 0.3$, 13 for 0.25, 17 for 0.2, 36 for 0.1. For the two-sample comparison, exact enumeration of Fisher's test gives power $0.007/0.051/0.042/0.093$ at a true blind rate of 0.3 for $k = 4/6/8/10$ pools, and $0.051/0.084/0.283/0.431$ at 0.5: only extreme rates are detectable at feasible k. The pre-specified design: $k = 11$ pools on the same cell at $b \in \{2, 4\}$, with the decision rule "0/11 match errors supports the reliability claim (upper bound $0.238 < 0.3$); $\geq 2/11$ closes it; 1/11 is inconclusive." At the design's stricter $\alpha = 0.025$ the 0/11 upper bound is 0.285, also below 0.3. Cost: 352 episodes.

Result (11-pool confirmation). Executed as registered: 11 pools $\times$ 2 arms $\times$ 2 sides $\times$ 8 episodes on the fixed 8-task band fnlt 270–277, frozen per-request seeds 902101–902144 (vLLM 0.8.5 per-request seeding; pools are independent sampling realizations over the fixed band, matching the power-audit semantics of a pool's noise realization; all pools share the same 8-task band, so pool-

to-pool variation is noise-realization variation, not task-population variation, and the reliability conclusion is scoped to decisions on this band). Sign errors (cumulative mean paired diff ≤ 0 on a true $+31$ edit; reference effect, Table G1): at b=4, match $6/11$, blind $4/11$, Fisher one-sided $p = 0.90$; at b=2, both $4/11$. The pre-registered rule fired: the reliability claim is closed. Decomposition matters: scores are $0/50$ granular, so b=4 means live on a $\{-50, -25, 0, 25, 50\}$ grid and most decisive cells are zero-mean no-signal outcomes, not reversed signs (5 of the 10 error events are mean exactly 0; excluding them leaves $2/11$ match vs. $3/11$ blind negative means). The arm-symmetric conclusion: at $b \leq 4$ the accept/reject decision is noise-dominated in both arms; the regime map's no-budget guidance, now confirmed at the decision level. The normal-approximation risk score that ranked the two original pools correctly achieved $3/11$ agreement here (below chance): pool risk is real (pools differ widely, blind per-pool diff SD 4.1–31.2) but is not yet predictable from run-record statistics. Full per-pool decisions, diffs, and records are released upon publication.

G LABEL VALIDITY: RE-DERIVATION UNDER MEASURED-SIGN RELABELING

Table G1: **Ground-truth reference effects** (ScienceWorld, Qwen2.5-7B, `find-non-living-thing`; 95% CIs, 2,000-resample bootstrap; n=60 paired tasks/arm, frozen ground-truth segment). The nominally harmful edit measures positive, the category-prediction failure we report as such; the no-op floor is a diagnostic alarm and is excluded from flip counts.

Candidate edit	n	Reference effect $\hat{\Delta}$ [95% CI]
helpful	60	$+15.98$ $[+8.8, +23.2]$
harmful (nominally)	60	$+14.85$ $[+7.4, +22.3]$
format	60	$+31.07$ $[+23.5, +38.7]$
no-op floor (diagnostic)	60	$+6.27$ $[+2.27, +10.33]$

Decomposition of the nominal-harmful cell. Across the three contexts (450 paired diffs; reference effects in Table G1): 80.2% of base episodes score exactly zero; the edit-side positive gains number 32% of pairs, of which 86.8% land on zero-base tasks; the median diff is 0 and the top five diffs are $+92, +92, +83, +67, +59$. The "harmful" edit therefore does not degrade working episodes; it converts all-or-nothing failures into partial credit. The nominal label fails as ground truth for this family.

Effect on headline claims. Under measured-sign relabeling (helpful$^+$, format$^+$, harmful$^-$ by reference sign): (i) the Table 2 context gaps and all CIs are unchanged (they never used the label); (ii) in §7, the frozen Δ_d of $+16.83$ for the harmful cell is replaced by its measured pool value (-20.8 match / -3.6 blind); both arms reject under either label, so $D(b)$ is unchanged; (iii) the static-judge accuracy of 0.333 was scored against reference signs and is unchanged. The one claim that must be withdrawn under relabeling is any phrasing of the form "the nominally harmful edit is actually beneficial"; the correct statement is that its benefit is a partial-credit measurement artifact concentrated on zero-base tasks. Human annotation ($\kappa \geq 0.6$) is pre-registered for camera-ready to validate both the measured signs and the category taxonomy.

H SENSITIVITY OF THE REGIME-MAP PLACEMENT TO THE ϵ PROXY

The regime map (Figure 2) places the measured cells with ϵ taken from the no-op floor's CI half-width (2.97 for the `find-non-living-thing`cells, Table E1; 1.65 for the tc dose cell, dose-0 CI in Figure 3). Replacing it with the per-cell bootstrap standard error of the paired diffs leaves every placement conclusion intact: the y-coordinates ($|\Delta_d|/\epsilon$) of the helpful/format/harmful cells shift by less than 20%, with no regime change (helpful and format remain at high margin with an above-noise context shift, harmful remains near the noise-limited boundary under the plotted orientation and inside the bias-dominated interior under the §7 orientation, and the weak families remain in the noise-limited band). The x-coordinates ($|I(e)|/|\Delta_d|$) are proxy-free.

I LANDING BRANCHES (PRE-AGREED)

Branch A: ctx-match improves deployment value → claim context-matched validation as the headline result, checked against accept-all/reject-all and single-edit dominance. Branch B: interactions clear but no acceptance advantage → strong-empirics route; drop the budgeted-validation contribution claim. Branch C: natural-edit transfer fails → report the failure; reposition. In all branches the audit protocol and the regime map remain contributions. **Decided (the six-cell equal-cost comparison)**: $D(16) = -1.94 \approx 0$ over six cells → **branch B**; the rule fired on the point estimate; the design excludes only advantages of one format-scale unit; matched validation made 0 sign errors on the true-positive cell across 2 pools × 4 budgets (blind: 2, one pool, $b \leq 4$), $\epsilon_m < \epsilon_{\text{iso}}$ in every fnlt cell of the equal-cost comparison (reversed in the replication pool), and a single pre-registered design instantiated the three-regime map with its negative replication reported.

J GLOSSARY

Symbol / term	Meaning
e, c	a skill-library edit; a library context (snapshot + retrieval budget)
$\Delta(e, c)$	expected deployment score change of e under context c (Eq. 1)
Δ_d, $I(e)$	deployment effect $\Delta(e, c_{\text{deploy}})$; context shift $= \Delta(e, c_{\text{deploy}}) - \Delta(e, c_{\text{iso}})$
$\epsilon(b)$, b	estimation noise std at probe budget b; probe episodes per arm (one edited–base pair = two episodes)
$\lvert\Delta_d\rvert$ margin	below this deployment effect, accept vs. reject barely matters
$V(\pi)$	deployment value of acceptance policy π
$\pi_{\text{match}}/\pi_{\text{blind}}$	validation in deployment context vs. in isolation
setting A	per-edit accept/reject as the deployment decision unit
ut / tc / flt / `fnlt`	ScienceWorld families: use-thermometer (ut), test-conductivity (tc), find-living-thing (flt), find-non-living-thing (`fnlt`)
pre-/post-audit pipeline	the natural-edit pipeline before vs. after the zero-rollout audit's fixes