

Visible but Not Avoided: State-Dependent Re-Proposal of Rejected Candidates in Catalog-Based LLM Optimization Loops

Jianlin Chen

Wenhui Chen

Chi Man VONG

Abstract

LLM optimization loops often show the proposer the candidates it has already tried and rejected, but leave it to the proposer to avoid them. We ask whether avoidance measured in a static probe predicts avoidance in the states an optimizer’s own trajectory produces. In catalog-based loops where three open LLMs improve a GUI agent by naming one of 40 edits, a re-proposal is identified exactly by the edit’s name. Clean base-state probes underpredict re-proposal in the proposers’ own loops (macro-averaged loop-minus-probe rate 0.32 [0.27, 0.37]), and proposals of already evaluated harness–edit pairs consume 28–63% of valid proposer iterations. The two settings reject different edits: a probe’s rejected edits are mostly ones the proposer did not choose, whereas a loop’s are its own earlier choices, which remain preferred when the history is removed. Ordinary history lowers re-proposal for two proposers but raises it for the third, and the effects of failure evidence are model-specific. Stating the rejection status in one sentence lowers re-proposal in every model (−0.16 [−0.22, −0.09] relative to a neutral list of the same IDs). For catalog-based loops with exact edit identities, rejection avoidance should be measured on the rejected sets the loop produces.

1 Introduction

LLM optimizers improve prompts, programs and agent harnesses by proposing a change, evaluating it and recording the outcome for the next proposal (Yang et al., 2024; Khattab et al., 2024; Agrawal et al., 2026; Novikov et al., 2025; Hu et al., 2025). Many show the proposer its failed attempts and leave the avoidance to it: OPRO asks for solutions different from the shown ones (Yang et al., 2024), self-evolving agents keep rejected edits “to discourage repetition” (Lu et al., 2026) or in a rejected-edit buffer (Yang et al., 2026), and Meta-Harness credits its gains to richer access to prior

candidates (Lee et al., 2026). Showing a rejection is not enforcing it, as a tabu list does (Glover, 1986). Under cached, deterministic evaluation, proposing an already evaluated candidate spends an iteration on a known outcome.

Avoidance of visible rejections is naturally tested on a fixed prompt, as instruction following is (Zhou et al., 2023a; Jiang et al., 2024). Such a probe can show rejected candidates chosen without regard to the proposer’s preferences. In a loop, the rejected candidates are the proposer’s own earlier choices: it proposed them because they looked promising, and a failed evaluation need not change that judgement. A proposer can therefore pass over arbitrary rejected candidates in a probe and still return to its own rejected favourites in a loop. We call the loop behaviour *visible-rejection avoidance*, reserving *compliance* for an explicit prohibition.

We study this question where re-proposal is exact. A proposer improves a screenshot-only GUI agent by naming one of 40 catalog edits, so a re-proposal is identified exactly by the edit’s name. We separate *visible re-proposal*, choosing an edit shown as rejected with the current harness, from *duplicate-pair proposal*, choosing any harness–edit pair already evaluated in the run. Three open proposers (Gemma-4, Mistral-Small-3.2, Qwen3-VL-32B) run the loop with two surface realizations of the catalog, which vary identifiers and presentation while holding the candidates and task fixed.

In a clean base-state probe, two proposers never chose a shown rejected edit and the third rarely did. In their own loops all three re-propose: the macro-averaged loop-minus-probe rate is 0.32 [0.27, 0.37], positive in all six model–catalog cells, and duplicate-pair proposals consume 28–63% of valid proposer iterations. Low re-proposal in a static probe does not guarantee low re-proposal in the states an optimizer reaches (§5.1).

To locate the gap, we rebuild sampled loop states. Keeping only the reached harness and the rejected

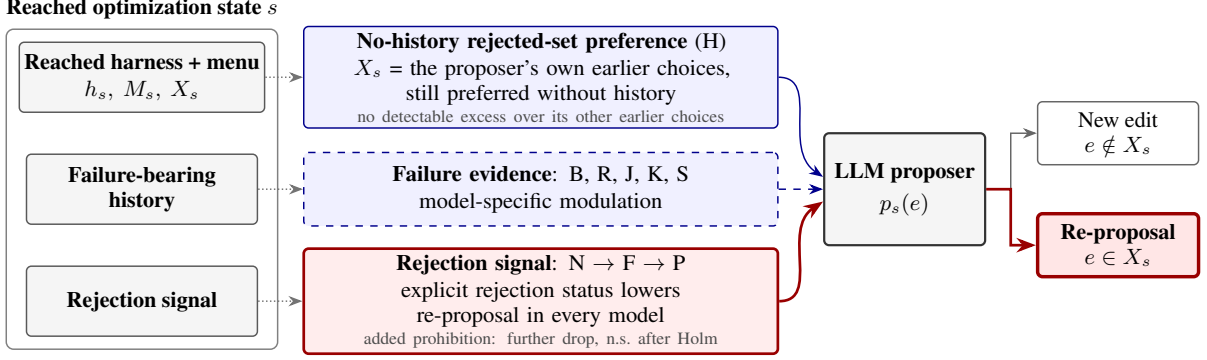


Figure 1: State-dependent re-proposal and the interventions used to localize it. H: no history; B, R, J, K, S: failure-evidence edits; N, F, P: neutral ID list, status sentence, status plus prohibition. Without history, X_s is strongly enriched, with no detectable excess over the proposer’s other earlier choices (post hoc; §5.1). Failure-evidence edits have model-specific effects (§5.2). Stating rejection status lowers re-proposal in every model; a prohibition gives a further negative rate estimate, not significant after Holm correction (p 0.054; §5.3). $\alpha = a - \lambda$ is a conceptual accounting; channel-specific readings require Assumption 2.

entries comes close to the original state, passing the preregistered ± 0.10 equivalence margin but not ± 0.05 . Removing the whole history block (scores, failures and rejection marks) shows that the rejected edits remain preferred: every proposer puts more mass on them than a uniform choice would. A post hoc comparison finds its other earlier choices enriched about as much, which points to self-selection without identifying a complete causal mechanism. Ordinary history lowers re-proposal for Gemma-4 and Qwen3-VL-32B and raises it for Mistral.

The history carries evidence about failures and the bare fact of rejection, so we edit the two separately on fresh states. Failure-evidence effects differ by model: consistent with failure content for Qwen3-VL-32B, mostly non-semantic for Mistral and small for Gemma-4. Rejection-status wording acts consistently: a sentence stating that the listed edits are in the harness or were tried with it and rejected lowers re-proposal relative to a neutral list of the same IDs in every model (-0.16 [$-0.22, -0.09$]). Adding “Do not propose any of them” lowers the rate further (-0.05), but this step did not pass the registered multiplicity-corrected test (Holm p 0.054; §5.2–5.3). Reading these wording effects as changes in an internal exclusion process requires the assumptions of §3.

Contributions.

1. A measurement framework that separates visible re-proposal from duplicate-pair proposal, scores re-proposal against a uniform choice over the menu, and states what probe-to-loop comparisons and matched-state edits identify (§3).
2. Evidence across three proposers that clean-

probe avoidance does not transport to the rejected sets their own loops produce, and that these rejected edits remain preferred without history (§5.1).

3. Matched-state interventions showing model-specific failure-evidence effects and a consistent effect of rejection-status wording (§5.2–5.3).

2 Related Work

LLM optimizers and the shown-history premise. Prompt optimizers propose instructions with an LLM and select them by score (Zhou et al., 2023b; Pryzant et al., 2023; Guo et al., 2024; Fernando et al., 2024; Khattab et al., 2024; Opsahl-Ong et al., 2024; Yuksekogonul et al., 2024; Yang et al., 2024; Agrawal et al., 2026). Program and agent optimizers keep archives of evaluated candidates (Romera-Paredes et al., 2024; Novikov et al., 2025; Lange et al., 2025; Lee et al., 2025; Hu et al., 2025; Zhang et al., 2025a; Lee et al., 2026; Xia et al., 2026). Whether these systems work, and why, has been examined through their scores and trajectories (Ma et al., 2024; Zhang et al., 2024, 2026a). Classical search treats exclusion as part of the algorithm, through tabu lists and novelty archives (Glover, 1986; Lehman and Stanley, 2011). Some LLM systems enforce it as well, for example by rejection sampling on code novelty (Lange et al., 2025), but most show the history and expect the proposer to act on it. How a system should define and enforce novelty is studied in a related manuscript (Anonymous, 2026); here the proposer is only shown its history. We test that expectation at the states these loops reach. Zhang et al. (2026a) find that zero-shot ability explains only part of an optimizer’s

trajectory. We isolate one transport failure, rejection avoidance, with exact candidate identity and matched-state counterfactuals.

Instruction following and negative constraints.

Instruction-following benchmarks score one response against constraints written into one prompt (Zhou et al., 2023a; Jiang et al., 2024; Qin et al., 2024; Zhang et al., 2026b). Negation is a known weakness (Jang et al., 2022; García-Ferrero et al., 2023), telling a model to avoid an entity can elicit it (Castricato et al., 2024), prohibitions decay over long conversations (Gamage, 2026), and context compaction can erase constraints (Chen, 2026). Rana (2026) explain violations of a word prohibition by the model’s prior pressure towards the word. Our no-history baseline measures such pressure at the states a loop reaches, and shows that it is directed at the proposer’s own earlier choices.

Failure records and repetition in context. Language models repeat what is in their context: induction heads copy earlier tokens (Olsson et al., 2022), decoding that maximises likelihood produces repetitive text (Holtzman et al., 2020), and a repeated sentence becomes more likely the more often it appears (Xu et al., 2022). For agents, a failed attempt left in context raises the error rate of a retry (Yang, 2026). The closest work is that of Gumaan (2026). In 135M–1.7B tool-using agents, a failure record raises the probability of repeating the failed call over a fixed candidate set, mostly through the surface form of the call, and a “do not repeat” instruction leaves the effect unchanged. Our primary distinction is the object of study: an evolving, trajectory-generated rejected set in an optimization loop, measured by exact identity at the states the loop actually reaches. Unlike their “do not repeat” instruction, our DO-NOT sentence lowers re-proposal substantially, and most of the reduction is already achieved by stating the rejection status. Reflection and stored experience help agents when feedback is informative (Shinn et al., 2023; Madaan et al., 2023; Zhao et al., 2024; Alazraki et al., 2025; Huang et al., 2024; Kamoi et al., 2024), while irrelevant context distracts (Shi et al., 2023).

Knowing versus acting, reasoning, and long context.

Models recall the constraints they violate (Kruthof, 2026), show a gap between accessible memory and action (Dong et al., 2026), and act unfaithfully to their own execution history (Zhang et al., 2025b). Our proposers show a related dis-

sociation: they justify re-proposals by the failures they see, yet for two of three models the content of those failures has little measurable effect on the choice (§5.2). Stronger reasoning lowers compliance with static instructions (Fu et al., 2025; Li et al., 2025), and reasoning traces rarely follow instructions about themselves (Kwon et al., 2026). Long histories also suffer from weak use of the middle of the context, short effective context, position effects and failed in-context exploration (Liu et al., 2024; Hsieh et al., 2024; Guo and Vosoughi, 2024; Krishnamurthy et al., 2024). App. H lists each line of work with the assumption we test.

3 Measurement and Behavioural Accounting Framework

This framework was developed after the first experiments, whose registered endpoints were rate differences (App. B); the fresh-state round (§4.3–4.4) was registered afterwards.

3.1 States and repeated proposals

A catalog loop keeps a harness $h_t \subseteq E$ and asks the proposer for one edit from the menu $M_t = E \setminus h_t$. The edit is kept if development accuracy rises by more than δ ; either way the attempt, its outcome and some failures are appended to the history the proposer sees, so a rejected edit can be proposed again.

Definition 1 (State and visible rejection). The state is $s_t = (h_t, H_t, q_t)$, with history H_t and rendered prompt q_t . The visible rejected set $X_t \subseteq M_t$ contains the edits shown in q_t as tried with harness h_t and not kept. The proposer’s choice follows the serving distribution $p_s(e) = P_{\text{serve}}(e_t = e \mid s, \text{endpoint, decoding, valid reply})$ over M_s , which includes provider and serving nondeterminism. A re-proposal is $B_t = 1[e_t \in X_t]$, with rate $\beta(s) = p_s(X_s)$; a state is eligible if $\emptyset \neq X_t \neq M_t$. We write s for a generic state and X_s, M_s, h_s for its components.

Replies naming no catalog edit are invalid and excluded (harness edits: App. C). A visible re-proposal ($B_t = 1$) chooses an edit in the visible rejected set. A duplicate-pair proposal, $D_t = 1[(h_t, e_t) \text{ evaluated earlier in the run}]$, chooses an evaluated pair whether or not its rejection is still visible: $B_t = 1$ implies $D_t = 1$, but under the truncated W8 view $D_t = 1$ can occur with $B_t = 0$. Under cached evaluation, D_t counts iterations spent without exploring a new pair.

3.2 A reference for re-proposal

States differ in how much of the menu is rejected, so a higher raw rate need not mean a stronger relative preference. As an illustration, not data: if 8 of 40 available edits are rejected, a uniform choice re-proposes with probability 0.20, and a proposer that puts 0.35 on the rejected set exceeds this reference.

Definition 2 (Rejection-enrichment log-odds). For a state s and a target set $T \subseteq M_s$ with $0 < |T| < |M_s|$, the enrichment log-odds of T are $\alpha_T(s) = \text{logit } p_s(T) - \text{logit}(|T|/|M_s|)$, where $|T|/|M_s|$ is the probability that a uniform choice over the menu lands in T . For the visible rejected set, the uniform-menu reference is $\beta^U(s) = |X_s|/|M_s|$, and the rejection-enrichment log-odds are $\alpha(s) \equiv \alpha_{X_s}(s) = \text{logit } \beta(s) - \text{logit } \beta^U(s)$.

Eligibility gives $0 < |X_s| < |M_s|$, so α is always defined. $\alpha > 0$ means more mass on rejected edits than the uniform reference, $\alpha < 0$ less and $\alpha = 0$ the same (the illustration has $\alpha \approx 0.77$); $\alpha = \pm\infty$ when $\beta(s) \in \{0, 1\}$. The reference removes the rejected fraction of the menu, not other differences between states.

3.3 Preference without history

Do a reached state's rejected edits remain attractive without the history? Let $s^\emptyset$ have the same harness, menu and question but an empty history. The whole history block (scores, failures, rejection marks, accuracy trajectory) is removed, so $s^\emptyset$ does not isolate rejection information. Its own rejected set is empty; we evaluate the source state's X_s as a target set, $\alpha_{X_s}(s^\emptyset) = \text{logit } p_{s^\emptyset}(X_s) - \text{logit } \beta^U(s)$. Because the rejected set contains the proposer's earlier choices at the current harness, stable preferences can yield positive enrichment even when history is removed; we compare it with the proposer's other earlier choices (§5.1). It measures preference over a self-selected set, not a causal effect or a specific mechanism.

3.4 Accounting for changes between states

Assumption 1 (Exclusion-penalised choice). For every state considered in the accounting, including matched counterfactual states, $p_s(e) \propto w_s(e) \exp(-\lambda_s \mathbf{1}[e \in X_s])$ for $e \in M_s$, with attraction weights $w_s(e) \geq 0$ and an exclusion strength $\lambda_s \geq 0$.

Attraction is shorthand for residual relative choice weight, not a cognitive mechanism: $w_s(e)$ absorbs every factor other than the explicit penalty.

Single weights may be zero, but an evaluated target set T and $M_s \setminus T$ need positive total weight. Any distribution with $0 < \beta(s) < 1$ has this form with $\lambda_s = 0$ and $w_s = p_s$, so the assumption is a parametrisation (weights up to a common factor): a and λ are not identified from choices, and Identities 1–2 follow algebraically from Definition 2.

Identity 1 (Decomposition). Under Assumption 1, $\alpha(s) = a(s) - \lambda_s$, where $a(s) \equiv a_{X_s}(s)$ and $a_T(s) = \log \bar{w}_s(T) - \log \bar{w}_s(M_s \setminus T)$ is the log ratio of the mean attraction inside and outside a set T .

Identity 2 (Probe transport). For eligible states s_0 and s , write $\Delta f = f(s) - f(s_0)$. Under Assumption 1,

$$\Delta \text{logit } \beta = \Delta a - \Delta \lambda + \Delta \text{logit } \beta^U,$$

and the corresponding enrichment identity is $\Delta \alpha = \Delta a - \Delta \lambda$.

Re-proposal thus differs between states through the rejected fraction of the menu ($\Delta \text{logit } \beta^U$), which the reference removes, and through enrichment, written as a change in attraction (Δa) minus a change in exclusion ($\Delta \lambda$). For a fixed reference, $\beta^U(s) = \beta^U(s_0)$, a probe at s_0 bounds re-proposal at s exactly when attraction has not risen by more than exclusion. In $s^\emptyset$ no edit is marked as rejected, so $\alpha_{X_s}(s^\emptyset) = a_{X_s}(s^\emptyset)$, and the history contrast

$$\alpha_{X_s}(s) - \alpha_{X_s}(s^\emptyset) = [a_{X_s}(s) - a_{X_s}(s^\emptyset)] - \lambda_s$$

is the net effect of the ordinary history on the mass assigned to X_s ; it does not separate attraction from exclusion.

3.5 What matched-state edits identify

A matched-state edit's effect on re-proposal is observed directly; attributing it to attraction or exclusion needs a further assumption.

Assumption 2 (Channel separation). A state edit ϕ leaves h_s , X_s and M_s unchanged. An evidence edit (changing the failures shown with history entries) leaves λ_s unchanged. Two sentences listing the same IDs at the same position change $a(s)$ equally.

Proposition 1 (Identification under channel separation). Under Assumptions 1–2, the log odds ratio of re-proposal between $\phi(s)$ and s equals $\alpha(\phi(s)) - \alpha(s)$; for an evidence edit it equals $a(\phi(s)) - a(s)$, and between two sentence edits ϕ_1, ϕ_2 listing the same IDs at the same position (for example F and N , or P and F) it equals $-(\lambda_{\phi_1(s)} - \lambda_{\phi_2(s)})$.

Proofs are in App. A. Assumption 2 is not identified from choices, so the contrasts identify effects on observed re-proposal and their attribution to a channel is conditional on it. Matched controls test its implications: if deleting failure spans acts through their content, an approximately length-matched neutral replacement should behave like their absence (§5.2).

Estimation. α is estimated by logistic regression of B_t with offset logit $\beta^U(s_t)$, or by an exact one-sided 95% upper bound when no re-proposal occurs. Matched-state contrasts are rate differences, also reported as Mantel–Haenszel log odds ratios stratified by state (the sample analogue of Proposition 1), with state-level and hierarchical bootstrap CIs (App. A).

4 Experimental Design

4.1 Optimization setting

The loop improves a screenshot-only GUI agent, Qwen3-VL-8B-Instruct (Bai et al., 2025), on phone screens from nine synthetic apps with numbered set-of-mark tags (Yang et al., 2023); the agent replies TAP n or ANSWER *text*. A harness is a set of edits from a 40-edit catalog (29 prompt, 9 rendering and 2 decoding edits), scored on 60 development items; agent calls are cached, so each harness has one accuracy (base: 0.75). For 30 iterations the proposer sees the task, the catalog, the current harness with its accuracy and the history, and names one edit, which is kept if accuracy rises by more than $\delta = 0.02$. Each history entry gives the edit, the new accuracy, *kept* or *not kept*, and two failed items (instruction, agent reply, correct tag). The view is the last 8 entries (W8) or the full history (FULL); decoding uses temperature 0 and at most 200 tokens. The catalog has two independently written surface realizations of the same 40 edits: catalog A with IDs E01–E40 and terse lines, catalog B with codes such as XR4, reworded and grouped descriptions and different task, question and history wording. The two versions hold candidate semantics and domain fixed while varying identifiers and presentation (Sclar et al., 2024) (App. F). Gemma-4-26B-A4B-it, Mistral-Small-3.2-24B-Instruct and Qwen3-VL-32B-Instruct (text-only prompts), accessed through OpenRouter, run 36 loops: 3 models $\times$ 2 catalogs $\times$ 2 views $\times$ 3 seeds.

4.2 Does static avoidance transport to reached states?

Clean probe. At the base harness, the probe finds the model’s zero-history favourite (the first rejected edit of its removal chain) and shows a window of $L \in \{8, 16\}$ real rejected base evaluations with their failures, the favourite at recency position $d \in \{1, L/2, L\}$ and the other entries drawn at random; five catalog orders give 30 cells per model and catalog. The clean probe and the reached loop states differ in harness, history and rejected-set composition. *Own loops.* We record B_t at every eligible step of the 36 loops. *Reached-state reconstruction.* From the eligible loop steps we sample states uniformly, regardless of the reply: 120 states in a first round and 360 fresh states in a second (§4.6). Condition M keeps the harness and the rejected entries with their failures and drops every other entry. Condition H keeps the harness, catalog and question and empties the history (§3.3); its replies are scored against the original X_s .

4.3 Edits of the failure evidence

On the same sampled states (overview: Fig. 1), with the unchanged state (A) as reference: *presence*, B deletes the failures from every entry; *text bulk*, R replaces each failure segment with approximately length-matched neutral text (a rural passage with no digits or task words) inserted at the same position with the same entry layout, within 13 characters in the realized prompts; *type*, C replaces failures with unshown failures of the same evaluation and type, D with failures without an ordinal or positional word (“eighth”, “row”), and T, on the 13 states where one exists, with a non-ordinal failure of the same evaluation, app and length (± 10 characters); *attachment*, J deletes failures only from the entries of X_s , K only from the other entries, and S permutes the failure pairs among the entries of X_s .

4.4 Edits of the rejection signal

Each condition adds one sentence before the question, listing the IDs that are in the harness or rejected with it: N a neutral list (“For reference, the following edit IDs appear in this prompt: ...”); F a status sentence (“The following edits are already in the current harness or were already tried with it and rejected: ... This note records their status.”); P the same status sentence ending “Do not propose any of them.”; and E the original sentence (“DO NOT PROPOSE any of the following edits. They

are already in the current harness or were already tried with it and rejected: ...”). The contrasts $F - N$ and $P - F$ separately test the incremental effects of rejection-status wording and of an explicit prohibition, with the listed IDs and prompt position held fixed; mapping these effects to the exclusion channel remains conditional on Assumption 2.

4.5 Boundary and supporting analyses

The regime comparison answers 40 fresh states with Qwen3-30B-A3B-Instruct-2507 and Qwen3-30B-A3B-Thinking-2507 (Yang et al., 2025) under the same 4000-token cap, first with unpinned routing and then with both endpoints pinned to one provider and three seeds. For stated reasons, two blinded LLM annotators coded 60 replies with a registered eight-code taxonomy (App. G).

4.6 Registration and inference

Every experiment was preregistered in a locally timestamped JSON file before its first paid call (App. B). The loops, probes and the first 120 states (A–E) were registered before any analysis, while M and N on those states are follow-up tests registered after A–E were observed. The 360 fresh states (A, H, M, B, R, S, J, K, T, N, F, P, E on 120; A and M on 240 more) were registered after the framework of §3, with one-sided hypotheses for B, R, J – K, S, P – F and E (Holm over six), two-sided ones for A – H, F – N and R – B (Holm over three), and an equivalence test of $M - A$ at ± 0.10 (secondary ± 0.05). Per-model contrasts were registered as descriptive, without multiplicity correction, and the self-selection comparison of §5.1 is post hoc. Each state and condition is queried three times at temperature 0 with provider seeds 1–3, and the unit is the state. Loops are resampled by run, probes by order, and states by state and by run, and tests are sign-flip tests. All paid calls cost \$4.88.

5 Results

5.1 Static avoidance does not transport to the rejected sets loops produce

In the clean probe all three proposers avoid the window: Gemma-4 and Qwen3-VL-32B named no window edit in any of their 60 cells ($\alpha \leq -2.19$), and Mistral named one in 5 of 60. In their own loops they re-propose: the registered loop-minus-probe difference is 0.32 [0.27, 0.37], positive in 6 of 6 cells (Table 1), and robust to how invalid replies are counted (0.31 to 0.36). Against the

uniform-menu reference, Gemma-4 and Qwen3-VL-32B choose rejected edits about as often as a uniform choice would (+0.35, −0.03) and Mistral far more often (+3.11; Fig. 2). Proposals of already evaluated harness–edit pairs ($D_t = 1$) consume 32% (103/324), 28% (98/350) and 63% (226/357) of valid proposer iterations (App. D).

The probe and the loop differ in harness, rejected set and history. The same-harness reconstruction (M) keeps the first two and drops other history. On 360 fresh states its rate is 0.30, against 0.34 for the unchanged state (A). $M - A$ is −0.04 (90% CI [−0.07, −0.01]; hierarchical [−0.08, 0.00]), meeting the preregistered ± 0.10 equivalence margin but not ± 0.05 . The A and M prompts are identical in 117 states. In the 243 states where M removes history, a post hoc analysis gives $M - A$ −0.057 (90% CI [−0.097, −0.016]; hierarchical [−0.113, 0.003]; App. E). The hierarchical interval crosses the ± 0.10 boundary, so equivalence is established for the registered average over all states, not for every state where the prompt changes. The base probe remains 0.32 below the loops.

With history removed (H), every proposer assigns more probability mass than uniform to the source-state rejected set X_s : +2.72 (Gemma-4), +1.28 (Mistral) and +2.11 (Qwen3-VL-32B). This is consistent with self-selection, since X_s consists of the proposer’s own earlier choices at h_s . In a post hoc comparison on the states where it had earlier chosen edits at other harnesses, those edits are enriched about as much without history (+2.13, +1.79, +1.75), and X_s shows no detectable excess over them (+0.07 [−1.31, 1.61], −1.62 [−3.74, 0.07], +0.12 [−1.16, 1.67]; App. D). The rejected set thus consists of the proposer’s earlier choices, which remain preferred relative to the uniform-menu reference when history is removed. Ordinary history lowers re-proposal relative to no history for Gemma-4 (−0.57 in rate) and Qwen3-VL-32B (−0.36), though not to probe levels, and raises it for Mistral (+0.31 [0.12, 0.49]).

The probe’s window holds one preferred edit, which the proposer avoids by choosing another, and 24 of 56, 147 of 215 and 27 of 58 loop re-proposals name one of a model’s three most frequent probe choices. In the loops, avoidance fails when a model’s two most frequent probe choices are themselves in X_t (post hoc; $\hat{\alpha}$ without and with them: Gemma-4 −0.45 and +0.61, Qwen3-VL-32B −0.42 and +0.39; Mistral is attracted either way). These observations suggest that rejected-set

Proposer	Cat.	Own loops		Clean probe		Δ [95% CI]
		n_B/n	$\hat{\alpha}$ [95% CI]	n_B/n	$\hat{\alpha}$	
Gemma-4-26B-A4B	A	22/130	+0.43 [−0.56, 1.21]	0/30	≤ -1.46	0.17 [0.08, 0.26]
Gemma-4-26B-A4B	B	34/156	+0.30 [−0.27, 0.84]	0/30	≤ -1.46	0.22 [0.13, 0.31]
Mistral-Small-3.2-24B	A	121/167	+3.52 [3.27, 3.69]	2/30	−1.84	0.66 [0.52, 0.75]
Mistral-Small-3.2-24B	B	94/158	+2.69 [1.98, 3.38]	3/30	−1.40	0.49 [0.31, 0.66]
Qwen3-VL-32B	A	16/161	−0.89 [−1.74, −0.30]	0/30	≤ -1.46	0.10 [0.05, 0.16]
Qwen3-VL-32B	B	42/152	+0.61 [−0.26, 1.23]	0/30	≤ -1.46	0.28 [0.13, 0.42]
Macro (6 cells)		329/924		5/180		0.32 [0.27, 0.37]

Table 1: Probe and loop. n_B/n : re-proposals over eligible valid loop steps (6 runs per cell) or valid probe cells. $\hat{\alpha}$: rejection-enrichment log-odds with run-cluster CI; $\leq$: exact one-sided 95% upper bound for a probe with no re-proposal. Δ : registered primary endpoint, loop rate minus probe rate, two-level cluster bootstrap CI.

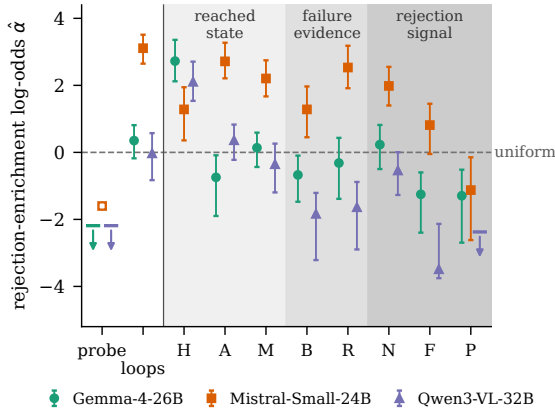


Figure 2: Rejection-enrichment log-odds against a uniform choice (dashed); arrows: exact 95% upper bounds. H–P: conditions on the 120 fresh states (Table 2); state-cluster CIs.

composition contributes to the probe–loop gap, although the post hoc comparisons do not quantify its causal contribution.

5.2 Failure evidence has model-specific text-bulk and content effects

Deleting the failures from every entry lowers re-proposal, on the first states (−0.19 [−0.25, −0.13]) and on the fresh ones (−0.17 [−0.23, −0.10]; Table 2). The approximately length-matched control weakens a common semantic-content explanation of the deletion effect. Replacing the deleted spans with neutral text restores much of the pooled re-proposal rate, but the pattern differs sharply across models (R − A −0.07 [−0.13, −0.01], Holm p 0.071; R − B +0.10 [0.05, 0.14]). Qwen3-VL-32B retains a substantial effect under neutral replacement, consistent with a content-sensitive component (R − A −0.21). For Mistral, most of the deletion effect is recovered by restoring a neutral span (−0.31 for deletion, −0.04 for neutral text), suggesting a largely non-semantic text-bulk or span-presence effect. Gemma-4 shows little effect from

either (+0.03, +0.05). The neutral replacement thus shows that deletion does not isolate semantic failure content. Consequently, B − A should not be interpreted as a pure attraction contrast.

We find no clear evidence that re-proposals systematically follow the history entries to which failure examples are attached. Deleting failures only from the rejected entries (J) or only from the others (K) has small effects (−0.05 and −0.00), with the registered contrast J − K in the wrong direction (+0.07 [−0.02, 0.16]), and after a shuffle of failures across rejected entries, choices follow the moved failures no more often than before (0.03 vs. 0.05; Holm p 0.425). Other-type replacements lowered re-proposal in the first round (D − A −0.11), but type is nearly confounded with app, and a strict within-app swap, possible on only 13 fresh states, is inconclusive (−0.10 [−0.33, 0.10]). Failure-bearing history matters, but the evidence that its content drives re-proposal is limited to one model.

5.3 Restating rejection status lowers re-proposal across all three main proposers

The no-history results show that previously selected edits can remain preferred. We therefore test whether stating their status changes proposals, with the listed IDs and their position held fixed. Repeating the IDs without saying what they are (N) changes little (−0.06 [−0.13, 0.00]). Saying that they are in the harness or were tried with it and rejected (F) lowers re-proposal from 0.30 to 0.09. Against the neutral list, F − N is −0.16 [−0.22, −0.09] (log odds ratio −1.44), and it is negative in every model (−0.14, −0.22, −0.11). Adding “Do not propose any of them” (P) lowers it further to 0.03. In rate this step is small and the registered test is not supported (P − F −0.05 [−0.10, −0.01], Holm p 0.054), because little re-proposal remains

Role	Condition	Original 120 states			Fresh R4 intervention subset (120 states)		
		Rate	$X - A$ [95% CI]	log OR	Rate	$X - A$ [95% CI]	log OR
Reference	A: unchanged state	0.34	–	–	0.30	–	–
Baseline preference	H: no history (matched)				0.51	+0.21 [0.11, 0.31]	+0.68
Localization	M: same-harness probe	0.30	−0.05 [−0.11, 0.01]	−0.35	0.27	Δ_{360} −0.04 [−0.07, −0.01] [†]	−0.36 [†]
Failure evidence	B: failure examples removed	0.15	−0.19 [−0.25, −0.13]	−1.92	0.14	−0.17 [−0.23, −0.10]	−1.49
	R: neutral text, approx. length-matched				0.24	−0.07 [−0.13, −0.01]	−0.56
	J: failures removed from X_t entries only				0.25	−0.05 [−0.12, 0.01]	−0.50
	K: failures removed from other entries only				0.30	−0.00 [−0.04, 0.04]	−0.09
	S: failures shuffled across X_t entries				0.32	follow 0.03→0.05	
	C: same-type failures	0.37	+0.04 [−0.03, 0.10]	+0.29			
	D: other-type failures	0.22	−0.11 [−0.18, −0.05]	−1.00			
Rejection signal	T: strict other-type swap ($n = 13$)				0.54	−0.10 [−0.33, 0.10]	−0.47
	N: neutral ID list	0.27	−0.06 [−0.13, 0.01]	−0.45	0.24	−0.06 [−0.13, 0.00]	−0.50
	F: status sentence				0.09	−0.22 [−0.28, −0.16]	−2.38
	P: status sentence + prohibition				0.03	−0.27 [−0.33, −0.21]	−2.88
	E: original DO-NOT sentence	0.04	−0.30 [−0.36, −0.24]	−3.20	0.03	−0.27 [−0.32, −0.22]	−3.61

Table 2: State edits, by role. Rate: mean of state-level rates (3 calls per state). $X - A$: state-level difference with stratified bootstrap CI ([†]: 90% CI over 360 states for the equivalence test). log OR: Mantel–Haenszel, stratified by state. S: share of choices that follow a shuffled failure pair. For M, the displayed rate is from the 120-state subset; the daggered equivalence contrast and log OR (Δ_{360}) use all 360 fresh states, on which the unchanged and reconstructed rates are 0.34 and 0.30. Subset-defined contrasts (T, on the n states where a strict swap exists; J – K in App. B) are paired against A on the same subset; the overall A rate cannot be used to reconstruct them. Registered tests and hierarchical CIs: App. B.

after F, and it is carried by Mistral (−0.15; Gemma-4 +0.00, Qwen3-VL-32B −0.01). In log odds it is substantial (−1.15 [−2.40, −0.31]), comparable to the status step (−1.44). The original DO-NOT sentence behaves like P (−0.27; −0.30 on the first states).

F brings rejected edits below the uniform reference for Gemma-4 and Qwen3-VL-32B ($\hat{\alpha} - 1.25$, −3.49), not for Mistral (+0.81). Reading F – N as stronger exclusion requires Assumption 2.

5.4 Boundary conditions and supporting evidence

Proposer regime. On 40 fresh states with the same 4000-token cap, the thinking endpoint re-proposed in 2 of 40 states and the instruct endpoint in 18 of 40 (−0.40 [−0.55, −0.25]; App. E); a 200-token instruct cap changed nothing detectable (+0.06 [−0.06, 0.17]). Pinned to Alibaba with three seeds, the rates were 0.06 and 0.38 (−0.33 [−0.43, −0.22]), so the difference is not a routing artefact; the endpoints also differ in post-training and completion tokens (3896 vs. 107).

Stated reasons. Among sampled re-proposals, the most frequent code is the preregistered rationale code *failure-target attraction*, a response that cites a failed item and proposes an edit to address it (16 and 21 of 30 for the blinded pair, $\kappa = 0.62$), whereas explicit retry is rare (2 and 3 of 30). The label describes the stated rationale, not the attachment-level causal mechanism tested and not supported in §5.2.

6 Discussion

A static probe asks whether a proposer avoids rejected candidates it mostly had little reason to choose; a loop asks it to avoid its own rejected favourites. The same-harness reconstruction meets the registered equivalence margin on average, while the post hoc analysis of changed prompts shows larger effects for Mistral and Qwen3-VL-32B (App. E). The comparison does not separate the effects of the harness and rejected-set composition. The self-selection comparison is post hoc and rests on few states, and history moves the three proposers in different directions, so no single mechanism describes them all. For catalog-based loops that delegate avoidance to the proposer, the results support evaluating it on rejected sets of the composition the loop produces and stating rejection status explicitly; systems needing hard non-repetition may prefer deterministic filtering, which we do not evaluate.

7 Conclusion

A rejection that an LLM proposer can see is not necessarily one it avoids. In our catalog-based loops, proposers without history put more probability on a loop’s rejected set than a uniform choice would; our clean probes only partially represent such self-selected sets, and stating rejection status lowers re-proposal in every model. For loops with exact candidate identities, avoidance should be measured on the rejected sets those loops produce.

Limitations

Scope. All loops improve one synthetic GUI agent with one 40-edit candidate set. Catalog B is a second surface realization of the same candidates, so the results show robustness to presentation, not generalisation across candidate sets or domains. Free-text optimizers, where re-proposals cannot be matched exactly by name, are not studied.

Identification. The framework was developed after the first experiments, and the fresh round tests it but uses the same loops. The no-history baseline assumes that an empty history leaves the proposer’s preference over the menu as it would be at that harness, and it also removes the accuracy trajectory. Because X_s consists of the proposer’s own earlier choices, the no-history condition measures preference over a self-selected set. The comparison with the proposer’s other earlier choices is post hoc and covers only the states where such choices exist (26, 24 and 26 for Gemma-4, Mistral and Qwen3-VL-32B, respectively), so it cannot exclude a modest harness-specific component. Likewise, $F - N$ directly identifies an effect of rejection-status wording, and interpreting that effect specifically as exclusion remains conditional on the equal-attraction clause of Assumption 2. The neutral replacement shows that deletion does not isolate semantic failure content, so $B - A$ is not a pure attraction contrast. Failure type is nearly confounded with app on this benchmark. The same-harness reconstruction passes the preregistered ± 0.10 equivalence margin, but not ± 0.05 , and the registration states this margin without a separate substantive justification.

Models and serving. Three open 24–32B models and one 30B instruct/thinking pair, and no proprietary model. Measured rates include serving non-determinism, and the models differ markedly from one another, so per-model results matter more than pooled ones. Provider routing was not pinned for the main state interventions. Conditions received similar provider mixtures; key earlier contrasts and the post hoc within-provider $F - N$ check have the same direction in their larger groups (App. C), but smaller groups are unstable and other round-4 contrasts lack provider-specific replication. Only the regime comparison was pinned.

Annotation. All annotators are LLMs and the codebook was refined on the coded replies. Human annotation of a held-out sample is pending.

Ethical Considerations

The study measures optimizer behaviour on synthetic phone screens, and no personal data and no human subjects are involved. Models were accessed through a commercial API within their terms of use. Lower re-proposal makes automated self-improvement cheaper, which inherits the oversight concerns of such loops. An LLM (the authoring assistant) served as one annotator and assisted with code and writing. All numbers are generated from logged data by scripts released with the paper.

References

- Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziemis, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, et al. 2026. [GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning](#). In *Proceedings of ICLR 2026*.
- Lisa Alazraki, Maximilian Mozes, Jon Ander Campos, Tan Yi-Chern, Marek Rei, and Max Bartolo. 2025. [No Need for Explanations: LLMs can implicitly learn from mistakes in-context](#). In *Proceedings of EMNLP 2025*.
- Anonymous. 2026. What Counts as Tried? Candidate Identity and Duplicate Control in Structured LLM Optimization. Under review.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, et al. 2025. [Qwen3-VL Technical Report](#). arXiv preprint arXiv:2511.21631. *Preprint*, arXiv:2511.21631.
- Louis Castricato, Nathan Lile, Suraj Anand, Hailey Schoelkopf, Siddharth Verma, and Stella Biderman. 2024. [Suppressing Pink Elephants with Direct Principle Feedback](#). arXiv preprint arXiv:2402.07896. *Preprint*, arXiv:2402.07896.
- Shiyang Chen. 2026. [Governance Decay: How Context Compaction Silently Erases Safety Constraints in Long-Horizon LLM Agents](#). arXiv preprint arXiv:2606.22528. *Preprint*, arXiv:2606.22528.
- Jiajun Dong, Yutao Hu, Fengrui Fan, Shihan Dou, Yueming Wu, and Deqing Zou. 2026. [MemArbiter: Decision-Time Memory Arbitration for Long-Horizon LLM Agents](#). arXiv preprint arXiv:2608.02113. *Preprint*, arXiv:2608.02113.
- Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2024. [Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution](#). In *Proceedings of ICML 2024*.

- Tingchen Fu, Jiawei Gu, Yafu Li, Xiaoye Qu, and Yu Cheng. 2025. [Scaling Reasoning, Losing Control: Evaluating Instruction Following in Large Reasoning Models](#). arXiv preprint arXiv:2505.14810. *Preprint*, arXiv:2505.14810.
- Yeran Gamage. 2026. [Omission Constraints Decay While Commission Constraints Persist in Long-Context LLM Agents](#). arXiv preprint arXiv:2604.20911. *Preprint*, arXiv:2604.20911.
- Iker García-Ferrero, Begoña Altuna, Javier Álvarez, Itziar González-Dios, and German Rigau. 2023. [This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models](#). In *Proceedings of EMNLP 2023*.
- Fred Glover. 1986. [Future paths for integer programming and links to artificial intelligence](#). *Computers & Operations Research*, 13(5):533–549.
- Esmail Gumaan. 2026. [Feedback That Backfires: Why Small Language Model Agents Repeat the Call They Just Watched Fail](#). arXiv preprint arXiv:2608.23651. *Preprint*, arXiv:2608.23651.
- Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. [EvoPrompt: Connecting LLMs with Evolutionary Algorithms Yields Powerful Prompt Optimizers](#). In *Proceedings of ICLR 2024*.
- Xiaobo Guo and Soroush Vosoughi. 2024. [Serial Position Effects of Large Language Models](#). arXiv preprint arXiv:2406.15981. *Preprint*, arXiv:2406.15981.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The Curious Case of Neural Text Degeneration](#). In *Proceedings of ICLR 2020*.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesht, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. [RULER: What’s the Real Context Size of Your Long-Context Language Models?](#) In *Proceedings of COLM 2024*.
- Shengran Hu, Cong Lu, and Jeff Clune. 2025. [Automated Design of Agentic Systems](#). In *Proceedings of ICLR 2025*.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. [Large Language Models Cannot Self-Correct Reasoning Yet](#). In *Proceedings of ICLR 2024*.
- Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2022. [Can Large Language Models Truly Understand Prompts? A Case Study with Negated Prompts](#). arXiv preprint arXiv:2209.12711. *Preprint*, arXiv:2209.12711.
- Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjuan Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. [FollowBench: A Multi-level Fine-grained Constraints Following Benchmark for Large Language Models](#). In *Proceedings of ACL 2024*.
- Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. [When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs](#). *Transactions of the Association for Computational Linguistics* (2024).
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, et al. 2024. [DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines](#). In *Proceedings of ICLR 2024*.
- Akshay Krishnamurthy, Keegan Harris, Dylan J. Foster, Cyril Zhang, and Aleksandrs Slivkins. 2024. [Can large language models explore in-context?](#) In *Advances in NeurIPS 2024*.
- Garvin Kruthof. 2026. [Models Recall What They Violate: Constraint Adherence in Multi-Turn LLM Ideation](#). arXiv preprint arXiv:2604.28031. *Preprint*, arXiv:2604.28031.
- Yongchan Kwon, Shang Zhu, Federico Bianchi, Kaitlyn Zhou, and James Zou. 2026. [ReasonIF: Large Reasoning Models Fail to Follow Instructions During Reasoning](#). In *Findings of ACL 2026*.
- Robert Tjarko Lange, Yuki Imajuku, and Edoardo Cetin. 2025. [ShinkaEvolve: Towards Open-Ended And Sample-Efficient Program Evolution](#). arXiv preprint arXiv:2509.19349. *Preprint*, arXiv:2509.19349.
- Kuang-Huei Lee, Ian Fischer, Yueh-Hua Wu, Dave Marwood, Shumeet Baluja, Dale Schuurmans, and Xinyun Chen. 2025. [Evolving Deeper LLM Thinking](#). arXiv preprint arXiv:2501.09891. *Preprint*, arXiv:2501.09891.
- Yoonho Lee, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, and Chelsea Finn. 2026. [MetaHarness: End-to-End Optimization of Model Harnesses](#). arXiv preprint arXiv:2603.28052. *Preprint*, arXiv:2603.28052.
- Joel Lehman and Kenneth O. Stanley. 2011. [Abandoning Objectives: Evolution Through the Search for Novelty Alone](#). *Evolutionary Computation*, 19(2):189–223.
- Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. 2025. [When Thinking Fails: The Pitfalls of Reasoning for Instruction-Following in LLMs](#). arXiv preprint arXiv:2505.11423. *Preprint*, arXiv:2505.11423.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the Middle: How Language Models Use Long Contexts](#). *Transactions of the Association for Computational Linguistics* (2024).

- Yuxing Lu, Yicheng Chen, Shanchan Wu, and Sercan Ö. Arik. 2026. [Procedural Graphs: Self-Evolving Execution Structures for LLM Agents](#). arXiv preprint arXiv:2609.09153. *Preprint*, arXiv:2609.09153.
- Ruotian Ma, Xiaolei Wang, Xin Zhou, Jian Li, Nan Du, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. [Are Large Language Models Good Prompt Optimizers?](#) arXiv preprint arXiv:2402.02101. *Preprint*, arXiv:2402.02101.
- 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, et al. 2023. [Self-Refine: Iterative Refinement with Self-Feedback](#). arXiv preprint arXiv:2303.17651. *Preprint*, arXiv:2303.17651.
- Alexander Novikov, Ngán Vŭ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, et al. 2025. [AlphaEvolve: A coding agent for scientific and algorithmic discovery](#). arXiv preprint arXiv:2506.13131. *Preprint*, arXiv:2506.13131.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, et al. 2022. [In-context Learning and Induction Heads](#). arXiv preprint arXiv:2209.11895. *Preprint*, arXiv:2209.11895.
- OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. [gpt-oss-120b & gpt-oss-20b Model Card](#). arXiv preprint arXiv:2508.10925. *Preprint*, arXiv:2508.10925.
- Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. [Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs](#). In *Proceedings of EMNLP 2024*.
- Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chengguang Zhu, and Michael Zeng. 2023. [Automatic Prompt Optimization with "Gradient Descent" and Beam Search](#). In *Proceedings of EMNLP 2023*.
- Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. [InFoBench: Evaluating Instruction Following Ability in Large Language Models](#). arXiv preprint arXiv:2401.03601. *Preprint*, arXiv:2401.03601.
- Shailesh Rana. 2026. [Semantic Gravity Wells: Why Negative Constraints Backfire](#). arXiv preprint arXiv:2601.08070. *Preprint*, arXiv:2601.08070.
- Bernardino Romera-Paredes, Mohammadamin Barekatin, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. 2024. [Mathematical discoveries from program search with large language models](#). *Nature*, 625(7995):468–475.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting](#). In *Proceedings of ICLR 2024*.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. [Large Language Models Can Be Easily Distracted by Irrelevant Context](#). In *Proceedings of ICML 2023*.
- Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: Language Agents with Verbal Reinforcement Learning](#). arXiv preprint arXiv:2303.11366. *Preprint*, arXiv:2303.11366.
- Peng Xia, Rujun Han, Zifeng Wang, Yanfei Chen, Yufan Zhuang, Yoonho Lee, Chengsong Huang, Han Yu, Zhongying CuiZhu, Yifei Ming, Huaxiu Yao, Burak Gokturk, et al. 2026. [RRSI: Regularized Recursive Self-Improvement of Agent Har- nesses](#). arXiv preprint arXiv:2609.24972. *Preprint*, arXiv:2609.24972.
- Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. [Learning to Break the Loop: Analyzing and Mitigating Repetitions for Neural Text Generation](#). In *Advances in NeurIPS 2022*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, et al. 2025. [Qwen3 Technical Report](#). arXiv preprint arXiv:2505.09388. *Preprint*, arXiv:2505.09388.
- Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. [Large Language Models as Optimizers](#). In *Proceedings of ICLR 2024*.
- Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chun- yuan Li, and Jianfeng Gao. 2023. [Set-of-Mark Prompting Unleashes Extraordinary Visual Ground- ing in GPT-4V](#). arXiv preprint arXiv:2310.11441. *Preprint*, arXiv:2310.11441.
- Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, Kai Qiu, Yuqing Yang, et al. 2026. [SkillOpt: Executive Strategy for Self-Evolving Agent Skills](#). arXiv preprint arXiv:2605.23904. *Preprint*, arXiv:2605.23904.
- Zhanfu Yang. 2026. [Why Retrying Fails: Context Con- tamination in LLM Agent Pipelines](#). arXiv preprint arXiv:2605.08563. *Preprint*, arXiv:2605.08563.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. [TextGrad: Automatic "Differentiation" via Text](#). arXiv preprint arXiv:2406.07496. *Preprint*, arXiv:2406.07496.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. 2025a. [Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents](#). arXiv preprint arXiv:2505.22954. *Preprint*, arXiv:2505.22954.

Tuo Zhang, Jinyue Yuan, and Salman Avestimehr. 2024. [Revisiting OPRO: The Limitations of Small-Scale LLMs as Optimizers](#). In *Findings of ACL 2024*.

Weichen Zhang, Yiyue Sun, Pohao Huang, Jiayue Pu, Heyue Lin, and Dawn Song. 2025b. [MIRAGE-Bench: LLM Agent is Hallucinating and Where to Find Them](#). arXiv preprint arXiv:2507.21017. *Preprint*, arXiv:2507.21017.

Xinhao Zhang, Xi Chen, François Portet, and Maxime Peyrard. 2026a. [What Makes an LLM a Good Optimizer? A Trajectory Analysis of LLM-Guided Evolutionary Search](#). In *Findings of ACL 2026*.

Yunfan Zhang, Yijie Bei, Jetashree Ravi, and Pawel Garbacki. 2026b. [FireBench: Evaluating Instruction Following in Enterprise and API-Driven LLM Applications](#). arXiv preprint arXiv:2603.04857. *Preprint*, arXiv:2603.04857.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. [ExpeL: LLM Agents Are Experiential Learners](#). In *Proceedings of AAAI 2024*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023a. [Instruction-Following Evaluation for Large Language Models](#). arXiv preprint arXiv:2311.07911. *Preprint*, arXiv:2311.07911.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023b. [Large Language Models Are Human-Level Prompt Engineers](#). In *Proceedings of ICLR 2023*.

A Proofs and Estimators

Identity 1. Write $W_{\text{in}} = \sum_{e \in X_s} w_s(e)$ and $W_{\text{out}} = \sum_{e \in M_s \setminus X_s} w_s(e)$. Both sums are positive under Assumption 1, and

$$\beta(s) = \frac{W_{\text{in}} e^{-\lambda_s}}{W_{\text{in}} e^{-\lambda_s} + W_{\text{out}}},$$

so $\text{logit } \beta(s) = \log W_{\text{in}} - \log W_{\text{out}} - \lambda_s$. The reference has $\text{logit } \beta^U(s) = \log |X_s| - \log |M_s \setminus X_s|$. Subtracting,

$$\alpha(s) = \log \frac{W_{\text{in}}}{|X_s|} - \log \frac{W_{\text{out}}}{|M_s \setminus X_s|} - \lambda_s,$$

which is $a(s) - \lambda_s$ because $W_{\text{in}}/|X_s| = \bar{w}_s(X_s)$ and $W_{\text{out}}/|M_s \setminus X_s| = \bar{w}_s(M_s \setminus X_s)$. A uniform proposer has w_s constant and $\lambda_s = 0$, hence $\alpha = 0$. Any p_s with $0 < \beta(s) < 1$ admits the representation with $\lambda_s = 0$ and $w_s = p_s$; more generally, for any chosen $\lambda_s \geq 0$, set $w_s(e) = p_s(e) e^{\lambda_s \mathbf{1}[e \in X_s]}$. Thus the split between a and λ is not identified from p_s alone, which is why Proposition 1 needs Assumption 2.

Identity 2. By Definition 2, $\text{logit } \beta(s) = \alpha(s) + \text{logit } \beta^U(s)$ at every eligible state. Apply Identity 1 at s and s_0 and subtract to obtain the enrichment identity; adding $\text{logit } \beta^U(s) - \text{logit } \beta^U(s_0)$ to both sides gives the raw transport identity. When $\beta^U(s) = \beta^U(s_0)$ the last bracket vanishes, and since logit is increasing, $\beta(s) \leq \beta(s_0)$ holds exactly when $a(s) - a(s_0) \leq \lambda_s - \lambda_{s_0}$.

Proposition 1. An edit ϕ leaves X_s and M_s unchanged, so $\beta^U(\phi(s)) = \beta^U(s)$ and the log odds ratio $\text{logit } \beta(\phi(s)) - \text{logit } \beta(s)$ equals $\alpha(\phi(s)) - \alpha(s)$, which by Identity 1 is $[a(\phi(s)) - a(s)] - [\lambda_{\phi(s)} - \lambda_s]$. For an evidence edit the second bracket is zero. For two sentence edits ϕ_1, ϕ_2 listing the same IDs at the same position, both leave X_s and M_s unchanged, so the same argument with $\phi_2(s)$ in place of s gives $[a(\phi_1(s)) - a(\phi_2(s))] - [\lambda_{\phi_1(s)} - \lambda_{\phi_2(s)}]$, and the first bracket is zero by Assumption 2.

No-history baseline. In $s^\emptyset$ no edit is displayed as rejected, so under Assumption 1 the exclusion term multiplies nothing and $p_{s^\emptyset}(e) \propto w_{s^\emptyset}(e)$. The menu is still M_s , and the source-state set X_s satisfies $0 < |X_s| < |M_s|$, so the computation of Identity 1 for the target set X_s , with no exclusion factor, gives $\alpha_{X_s}(s^\emptyset) = a_{X_s}(s^\emptyset)$. For the original state, $\alpha_{X_s}(s) = a_{X_s}(s) - \lambda_s$. Hence $\alpha_{X_s}(s) - \alpha_{X_s}(s^\emptyset) = [a_{X_s}(s) - a_{X_s}(s^\emptyset)] - \lambda_s$: the net effect of adding the ordinary history on the mass assigned to X_s . The contrast identifies λ_s only if the history leaves a_{X_s} unchanged; we do not assume that, and report the net effect.

Estimators. *Rejection-enrichment log-odds.* For steps t with outcomes B_t and offsets $o_t = \text{logit } \beta^U(s_t)$, $\hat{\alpha}$ maximises $\sum_t [B_t(\alpha + o_t) - \log(1 + e^{\alpha + o_t})]$, a concave function; we use damped Newton steps. It is the maximum-likelihood estimator if α is constant over the steps and the replies are independent given their states. Steps of one run are not independent, so we treat it as a working likelihood, solving $\sum_t [B_t - \sigma(\hat{\alpha} + o_t)] = 0$, and take CIs from

the cluster bootstrap. When α varies across steps, $\hat{\alpha}$ is a weighted summary. *Probe bound.* With no re-proposal in cells with offsets o_i , the bound is the largest α with $\prod_i (1 - \sigma(\alpha + o_i)) \geq 0.05$, exact if the cells are independent with a common α ; with $k > 0$ re-proposals, $P(\sum_i B_i \leq k)$ under the Poisson-binomial distribution replaces the product. *State edits.* With a_s, b_s the re-proposals and non-re-proposals of state s under the edit and c_s, d_s under the comparison, the Mantel–Haenszel estimate is $\log(\sum_s a_s d_s / n_s) - \log(\sum_s b_s c_s / n_s)$, consistent for a common odds ratio as the number of states grows with few calls per state, assuming independent calls within a state and condition; when the state-level log odds ratios differ, it estimates a weighted average of them.

B Registration, Claims and All Registered Tests

Each confirmatory experiment has a JSON registration file (`exp/prereg/`). The anonymous Data supplement contains the registration files, per-call ledger and raw replies; the Software supplement contains the analysis scripts. The files carry their own UTC timestamps; they were not deposited with an external registry, and the order of registration and data collection can be checked against the per-call ledger (Table 3). Each file states the question, hypothesis, falsifier, primary endpoint, models, seeds, sampling, unit of analysis, exclusions, test, CI, multiple-testing family, stopping rule and budget. Ten pilot calls preceded registration; they were used only to check the parser and are excluded from every analysis. The registered analyses are differences in rates; the log-odds analyses of §3 were added afterwards, at no cost, from the same logged data (`analysis/model.py`).

Registration file	Registered (UTC)	First paid call (UTC)	USD
VRR_CONFIRM.json	2026-10-04 06:49:05	2026-10-04 06:49:43	0.393
VRR_INTERVENE.json	2026-10-04 06:49:05	2026-10-04 07:15:52	0.375
VRR_REASONING.json	2026-10-04 06:49:05	2026-10-04 07:17:11	1.208
VRR_REASONING_AMEND1.json	2026-10-04 08:34:01	2026-10-04 08:34:25	0.144
VRR_QUAL.json	2026-10-04 06:49:05	2026-10-04 07:20:41	0.014
VRR_R2_REASONING.json	2026-10-04 10:10:24	2026-10-04 10:12:04	0.282
VRR_R2_STATEPROBE.json	2026-10-04 10:10:24	2026-10-04 10:12:05	0.131
VRR_R2_SEEDS.json	2026-10-04 10:10:24	2026-10-04 10:12:04	0.087
VRR_R2_QUAL.json	2026-10-04 10:10:24	2026-10-04 10:12:04	0.029
VRR_R3.json	2026-10-04 14:06:25	2026-10-04 14:07:34	0.029
VRR_R4.json (VRR-R4)	2026-10-07 15:01:24	2026-10-07 15:01:38	0.973
VRR_R4.json (VRR-R4-REGIME)	2026-10-07 15:01:24	2026-10-07 15:46:36	1.215

Table 3: Registration vs. first paid call (UTC) and spend per experiment.

Amendment to VRR-P1. The reasoning experiment’s registered stop rule (\$1.00) was checked before each 30-iteration run, not before each

call. Thinking calls cost about \$0.013 each, roughly ten times the estimate, so the third thinking loop ended at \$1.20. At that point the loop-level counts had been printed, and no cross-evaluation call had been made. The amendment (VRR_REASONING_AMEND1.json) reduced the matched-state cross-evaluation from 30+30 to 15+15 states, dropped the thinking-model probe, and moved the stop rule (\$1.75) to every call. This test (H3: $-0.54 [-0.73, -0.35]$ on 26 paired states, unequal budgets) is superseded by the fresh test with one token cap in §5.4.

Round 2. A second set of registrations (VRR_R2_*.json) was written in response to a review and before any of its calls: the fresh three-arm regime test (H4a, H4b), the same-harness probe and the neutral-list control (H5, H6a, H6b), five more seeds per condition on 30 states, and the blinded re-annotation. H5 and H6a/b use the 120 sampled states. They were registered after the results of conditions A–E on those states had been seen, and only conditions M and N were new; we therefore call them preregistered follow-up tests, not independent confirmations. In the blinded re-annotation, gpt-oss-120b’s first reply hit its 600-token limit in 11 of 60 cases and was re-queried with 2000 tokens.

Round 3 (stopped). VRR_R3.json registered a re-test of H2a, H2b and H6a on 120 fresh states, an equivalence test of the same-harness probe (TOST, margin ± 0.10), an approximately length-matched neutral replacement of the failures and a shuffle of failures across entries. The run stopped after 179 of about 2,484 planned calls (11 Gemma-4-26B-A4B states; \$0.029), when the API key, which is shared with other projects, reached its spending limit. These calls cover too few states for any registered test, so we do not analyse them; the raw replies are released and the spend is included in all totals. Its design was carried into round 4.

Round 4 (fresh states). VRR_R4.json was registered after the framework of §3 and before any call of the round. It sampled 360 fresh eligible states (60 per model and catalog, none used before), ran conditions A, H, B, R, S, J, K, T, N, F, P, E and M on 120 of them and A and M on the other 240, and pinned the regime comparison to one provider (VRR-R4-REGIME). The first two states were a pilot, read reply by reply; no rendering fault was found, so they are part of the data. Table 5 lists

Claim	Where	Evidence	Falsifier	Status
Loop re-proposal exceeds clean-probe re-proposal	§5.1	H1: $\bar{\Delta} = 0.32$ [0.27, 0.37]; H1b: 6/6 cells	CI includes 0; fewer than 4 cells	registered, supported
M and A satisfy the preregistered ± 0.10 equivalence criterion	§5.1	H15: $M - A = -0.04$ [$-0.07, -0.01$] (90%), 360 fresh states	CI outside ± 0.10	registered, passes ± 0.10 , not ± 0.05
Ordinary history changes re-proposal relative to matched no-history	§5.1	H14a: $A - H = -0.21$ [$-0.30, -0.11$]	CI includes 0	registered, supported pooled (decrease); reversed for Mistral
Failure deletion lowers re-proposal	§5.2	H2a: -0.19 ; H13a: -0.17 [$-0.23, -0.10$]	CI includes 0	registered, supported (both rounds)
Failure content beyond approximate text-bulk control	§5.2	H13b: $R - A = -0.07$ [$-0.13, -0.01$]	Holm $p \geq 0.05$	registered, not supported (Holm p 0.071)
Choice follows shuffled failure attachment	§5.2	H13c: $J - K = 0.07$; H13d: follow $0.03 \rightarrow 0.05$	no difference	registered, not supported
Status wording beyond neutral same-ID list	§5.3	H14b: $F - N = -0.16$ [$-0.22, -0.09$]	CI includes 0	registered, supported in every model
Prohibition beyond status wording	§5.3	H13e: $P - F = -0.05$ [$-0.10, -0.01$]; log OR -1.15 [$-2.40, -0.31$]	Holm $p \geq 0.05$	registered rate test not supported (Holm p 0.054)
Instruct vs. thinking, same token cap	§5.4	H4a: -0.40 ; pinned H16: -0.33 [$-0.43, -0.22$]	CI includes 0	registered, supported; endpoint difference

Table 4: Claims, evidence and falsifiers. Registered tests are differences in rates (sign-flip tests, Holm within their family). H13–H16 were registered after the framework of §3.

every registered test. Of 5,754 calls, a share were served from the cache because the prompt was identical to another condition’s (for example K equals A in states whose shown entries are all rejected).

Outcomes of all registered hypotheses. H1 (macro gap > 0): supported. H1b (positive in ≥ 4 of 6 cells): supported, 6/6. H2a ($B - A < 0$): supported. H2b ($E - A < 0$): supported, -0.30 [$-0.36, -0.24$]. H2c ($C - A$, two-sided): CI includes 0. H2d ($D - A$, two-sided): negative, $p_{\text{Holm}} 0.002$. H3: supported under the amendment; superseded by H4a. H4a: supported. H4b: CI includes 0. H5: CI includes 0. H6a: supported. H6b: CI includes 0. QUAL: κ reported as registered (first round $\kappa = 0.50$; blinded $\kappa = 0.62$). VRR-R3: not tested (run stopped). Round 4 (Table 5): H13a and H13f supported; H13b, H13c, H13d and H13e not supported after Holm; H14a, H14b and H14c supported; H15 passes the ± 0.10 margin but not ± 0.05 ; H16 supported.

Heterogeneity and sensitivity of the registered endpoint. The six cells share three models, one candidate set and one task, so we do not summarise them as a random sample of cells. Per model, averaged over catalogs, Δ is 0.19 [0.13, 0.26] (Gemma-4), 0.58 [0.46, 0.67] (Mistral) and 0.19 [0.11, 0.27] (Qwen3-VL-32B). The difference between catalogs (catalog B minus catalog A) includes 0 for Gemma-4 and Mistral and is positive for Qwen3-VL-32B ($+0.18$ [0.02, 0.34]; [0.06, 0.30] with the run bootstrap stratified by view; $+0.18$ [0.05, 0.32] with view and $|X_t|$ fixed). Subtracting the exact 95% upper bound of each probe rate from the lower

bound of the loop CI leaves the gap positive in all cells except Gemma-4 and Qwen3-VL-32B on catalog A (2 of 6), where loop rates are lowest.

C Protocol, Models and Costs

Models. Proposers via OpenRouter: google/gemma-4-26b-a4b-it, mistralai/mistral-small-3.2-24b-instruct, qwen/qwen3-v1-32b-instruct, qwen/qwen3-30b-a3b-instruct-2507, qwen/qwen3-30b-a3b-thinking-2507. Annotators: deepseek/deepseek-chat-v3.1, openai/gpt-oss-120b. Agent: qwen/qwen3-v1-8b-instruct. Prices were fetched from the provider API before the first call (exp/configs/pricing.json). Temperature is 0 throughout; the state edits also set provider seeds. The loop prompt asks for at most two sentences of reasoning and a final line CHOICE: $\langle id \rangle$.

Serving providers. OpenRouter routed every call to one of the providers that serve the model; we did not pin providers. The ledger records the serving provider of each call (Table 7). Within a model, re-proposal rates differ between providers; in condition A, for example, Mistral re-proposed in 26 of 33 valid DeepInfra replies and 47 of 85 valid Venice replies. The provider mix was similar across conditions (Mistral: Venice served 71–87 of 120 calls per condition), and $B < A$ and $E < A$ hold within each of the 5 model-provider combinations with at least 25 valid replies per condition. These checks reduce concern about routing but cannot establish its absence as a confounder. Round 4 also routed freely, except for the regime rerun (VRR-

Test	Contrast	n	Rate difference [95% CI]	Hierarchical CI	log OR [95% CI]	p_{Holm}
H13a	B – A < 0	120	−0.17 [−0.23, −0.10]	[−0.25, −0.08]	−1.49 [−2.21, −0.94]	<0.001
H13b	R – A < 0	120	−0.07 [−0.13, −0.01]	[−0.15, 0.02]	−0.56 [−1.16, −0.04]	0.071
H13c	J – K < 0	65	+0.07 [−0.02, 0.16]	[−0.03, 0.18]	+0.53 [−0.10, 1.24]	0.928
H13d	follow S > A	102	+0.02 [−0.01, 0.05]	–	–	0.425
H13e	P – F < 0	120	−0.05 [−0.10, −0.01]	[−0.11, 0.00]	−1.15 [−2.40, −0.31]	0.054
H13f	E – A < 0	120	−0.27 [−0.32, −0.22]	[−0.34, −0.20]	−3.61 [−5.05, −2.89]	<0.001
H14a	A – H	120	−0.21 [−0.30, −0.11]	[−0.34, −0.07]	−0.68 [−1.06, −0.34]	0.002
H14b	F – N	120	−0.16 [−0.22, −0.09]	[−0.25, −0.06]	−1.44 [−2.30, −0.81]	<0.001
H14c	R – B	120	+0.10 [0.05, 0.14]	[0.04, 0.16]	+1.25 [0.64, 2.01]	0.002
H15	M – A, equivalence ± 0.10	360	−0.04 [−0.07, −0.01] (90%)	[−0.08, 0.00]	−0.36	equivalent
H16	thinking – instruct, pinned	40	−0.33 [−0.43, −0.22]	–	–	<0.001

Table 5: Round-4 registered tests on fresh states. One-sided tests H13a–f (Holm over six), two-sided H14a–c (Holm over three); stratified and hierarchical (run, then state) bootstrap CIs; Mantel–Haenszel log odds ratios. H15: equivalence of M – A within ± 0.10 (90% CI over 360 states). H16: provider-pinned regime comparison.

Test	Contrast	n	Effect [95% CI]	p_{Holm}
H4a	thinking – instruct (both 4000 tok.)	40 states	−0.40 [−0.55, −0.25]	<0.001
H4b	instruct 4000 – 200 tok.	36 states	+0.06 [−0.06, 0.17]	0.689
H5	same-harness probe – loop state	117 states	−0.05 [−0.11, 0.01]	0.177
H6a	DO-NOT – neutral list	120 states	−0.23 [−0.29, −0.18]	<0.001
H6b	neutral list – unchanged	117 states	−0.06 [−0.13, 0.01]	0.177

Table 6: Round-2 registered tests. Paired differences in re-proposal rate over states; sign-flip tests, Holm within each registration (H4a/b; H5, H6a/b).

R4-REGIME), whose 240 calls were all served by the pinned provider.

Model	Calls	Providers	Largest shares
Gemma-4-26B-A4B	3295	13	NextBit 34%, Cloudflare 22%, DekaLLM 11%, ...
Mistral-Small-3.2-24B	2962	3	Venice 57%, DeepInfra 34%, Parasail 9%
Qwen3-VL-32B	3021	1	Alibaba 100%
Qwen3-30B-A3B-Instruct	347	6	Alibaba 39%, StreamLake 32%, SiliconFlow 10%, ...
Qwen3-30B-A3B-Thinking	256	1	Alibaba 100%
Qwen3-VL-8B (agent)	2640	3	PrimeIntellect 44%, Alibaba 43%, Parasail 13%
DeepSeek-V3.1 (annotator)	120	4	DeepInfra 52%, AtlasCloud 43%, SiliconFlow 4%, ...
gpt-oss-120b (annotator)	71	6	DigitalOcean 92%, CoreWeave 3%, Amazon Bedrock 1%, ...

Table 7: Serving providers per model over all paid calls except the pilot (from the per-call ledger).

Post hoc provider check for F – N. Table 8 compares valid F and N replies only where both conditions reached the same state through the same provider; multiple calls within a state–condition–provider are averaged before the state contrast. All four model–provider groups with at least 25 valid calls under each condition have negative estimates. The Gemma-4/IO Net group has a positive estimate from five paired states; several other groups have wide intervals. This check is descriptive because provider assignment was not controlled, and one additional group with fewer than five paired states is omitted from the table.

Harness edits. A reply naming an edit already in h_s is invalid in the loops and counted as $B = 0$ in the state edits, where it makes up 1.3% and 0.5% of valid replies in the two rounds.

Caching. Every call is stored under the hash of (model, messages, max tokens, temperature, provider, seed), and a repeated identical request

returns the stored reply. Agent evaluations are cached per harness, so a harness has one development accuracy. The response cache of an earlier exploratory study on the same benchmark was read (never written) as a fallback; it supplied cached agent evaluations and the base-harness rejections used in the probe.

Budget. A hard project cap of \$10 applied; the client refuses any call whose estimated cost would exceed the project budget, which was \$3.5 until round 4 and was then raised to the cap with the authors’ approval. Each registration also set its own stop rule. Table 9 gives the spend per experiment.

D Verification Log

Every boundary case and unexpected result was traced through the code path and the logged data before it was interpreted. Each entry gives the observation, the check and the outcome.

- *Probe rates of exactly zero.* Checked that the probe’s window edits are rejected base evaluations, that the parser accepts both catalogs’ IDs, and that the favourite is in the window. Outcome: genuine zeros; reported as exact upper bounds, not as $-\infty$.
- *Raw rates mislead when $|X_t|$ varies.* Full views show more rejections (mean $|X_t|$ 6.5 vs. 4.6). Outcome: the uniform-menu reference (Definition 2); the FULL – W8 loop-rate difference shrinks from +0.13 to +0.09 [0.00, 0.17] with cell and $|X_t|$ fixed in a linear probability model. Qwen3-VL-32B’s catalog effect survives the same adjustment (+0.18 [0.05, 0.32]).
- *Harness or rejected-set size?* Base-harness loop steps have $|X_t| \leq 3$ for Gemma-4 and Qwen3-VL-32B. Outcome: the harness contrast holds at matched $|X_t| \leq 3$ (§5.1).

Model	Provider	F/N calls	Paired states	F – N [95% CI]
Gemma-4	Cloudflare	26/26	9	−0.15 [−0.33, 0.00]
Gemma-4	Io Net	14/13	5	+0.07 [0.00, 0.20]
Gemma-4	Makora	18/18	7	−0.33 [−0.67, −0.05]
Gemma-4	NextBit	47/45	16	−0.11 [−0.28, 0.01]
Mistral	DeepInfra	18/11	9	−0.11 [−0.39, 0.11]
Mistral	Parasail	15/14	7	−0.21 [−0.79, 0.43]
Mistral	Venice	78/81	33	−0.26 [−0.43, −0.09]
Qwen3-VL	Alibaba	117/115	40	−0.11 [−0.18, −0.05]

Table 8: Provider sensitivity of F – N (post hoc). Valid F/N calls are counted within paired states, which have at least one valid call under both conditions from the same provider. Differences average state-level rates within each provider, with catalog-stratified 95% bootstrap CIs over paired states. Only groups with at least five paired states are shown.

Experiment	USD
VRR-P0-1	0.3381
VRR-P0-1-probe	0.0545
VRR-P0-2	0.3748
VRR-P1	1.1997
VRR-P1-cross	0.1440
VRR-P1-probe	0.0085
VRR-QUAL	0.0141
VRR-R2-QUAL	0.0290
VRR-R2-REASON	0.2816
VRR-R2-SEEDS	0.0875
VRR-R2-SPROBE	0.1310
VRR-R3	0.0288
VRR-R4	0.9731
VRR-R4-REGIME	1.2154
VRR-pilot	0.0013
Total	4.8815

Table 9: Spend per experiment (USD, from the per-call ledger).

- *Depletion of the menu.* Hypothesis: avoidance falls as edits are kept. Test: slope of α in $|h_t|$ with cell intercepts. Outcome: not consistent across models (Gemma-4 +0.09 [−0.46, 1.12], Mistral +0.33 [0.02, 0.89], Qwen3-VL-32B −0.78 [−1.28, 0.26]); not used as an explanation.
- *Self-selection of the rejected set.* Every edit of X_s was chosen by the proposer at h_s , so H’s enrichment of X_s is expected under a stable preference. Null set O_s : edits the proposer chose earlier in the run at other harnesses, minus X_s and h_s , on the states where $0 < |O_s| < |M_s|$ (26, 24 and 26 states). Under H, $\hat{\alpha}$ for X_s and O_s is +2.20 and +2.13 (Gemma-4), +0.16 and +1.79 (Mistral), +1.87 and +1.75 (Qwen3-VL-32B); the excess of X_s is +0.07 [−1.31, 1.61], −1.62 [−3.74, 0.07] and +0.12 [−1.16, 1.67]. With O_s taken from the whole run (35, 29 and 30 states) the excess is +1.18 [−0.02, 2.56], −0.38 [−1.91, 1.20] and +0.98 [−0.49, 2.62]. Out-

come: no detectable harness-specific excess; §5.1 describes the enrichment as consistent with self-selection (post hoc, paired bootstrap over states, `analysis/r4.py`).

- *Base-harness zero-history favourites.* Hypothesis: loops fill X_t with the proposer’s zero-history favourites at the base harness. Outcome: rejected; the zero-history favourites mostly remain available outside $X_t \cup h_t$, and the edit Mistral re-proposes most often (E33) is not among them.
- *Failure-target availability (exploratory; used in §5.1).* Hypothesis: avoidance fails when the edits a model chooses in its probe (its two most frequent probe choices per catalog) are in X_t . Outcome: for Gemma-4, $\hat{\alpha} = -0.45$ [−1.02, −0.02] without and +0.61 [0.00, 1.13] with a target in X_t ; for Qwen3-VL-32B, −0.42 [−1.16, 0.00] and +0.39 [−1.50, 1.26]; for Mistral, +2.92 [0.91, 3.54] and +3.15 [2.71, 3.55]. Supportive for two models, not for Mistral, which is attracted either way; the targets were chosen after seeing the probe data.
- *Low agreement among seeded temperature-0 calls* (0.25–0.38 of state-condition triples agree). Check: provider routing. Outcome: pairs served by the same provider agree no more often than pairs served by different providers (0.44 vs. 0.54); the state rate is treated as a Monte Carlo estimate. With eight calls on 30 states, B – A is −0.14 [−0.24, −0.05] (three calls: −0.13) and E – A −0.28 [−0.38, −0.18] (−0.29).
- *Dependence among states from one run.* The 120 states come from 36 source runs. Outcome: a hierarchical bootstrap that resamples runs within

- model and catalog and then states within runs gives log odds ratios with CIs $[-3.10, -1.17]$ ($B - A$), $[-0.28, 0.90]$ ($C - A$), $[-1.98, -0.28]$ ($D - A$), $[-5.07, -2.19]$ ($E - A$), $[-1.12, 0.15]$ ($N - A$), $[-0.94, 0.28]$ ($M - A$) and $[-4.64, -1.83]$ ($E - N$), and rate differences $[-0.27, -0.11]$, $[-0.04, 0.12]$, $[-0.21, -0.02]$, $[-0.39, -0.22]$, $[-0.14, 0.01]$, $[-0.12, 0.02]$ and $[-0.31, -0.16]$. Every conclusion drawn from the state-level CIs holds.
- *The neutral replacement (R) restores most of the rate that deletion (B) removes.* This was the most consequential unexpected result. We rebuilt R prompts and checked that each failure segment was replaced in place, with the same entry layout and a length within 13 characters, and that no failure text survived. Outcome: R is constructed as registered. The neutral replacement restores most of the deletion effect for Mistral; Gemma-4 shows little effect from either deletion or neutral replacement; Qwen3-VL-32B retains a substantial effect under neutral replacement, consistent with a content-sensitive component.
 - *A strict type swap is rarely possible.* Only 73 of 2,294 shown ordinal failures have an unshown non-ordinal failure from the same evaluation and app within 10 characters, because failure type and app nearly coincide on this benchmark (mail-list and file-grid failures are ordinal). Outcome: T is descriptive (13 states).
 - *The history raises re-proposal for Mistral.* $A - H$ is negative for Gemma-4 and Qwen3-VL-32B but positive for Mistral ($+0.31$ $[0.12, 0.49]$). Checked that H prompts keep the harness line, catalog and question and differ from A only in the history block. Outcome: reported as a model difference, not pooled away.
 - *Cached replies in round 4.* 900 of the round's calls were served from the cache because their prompt equalled another condition's prompt for the same state and seed (for example K equals A when every shown entry is rejected, and M equals A when the history holds only rejected entries). These are genuine replies to the identical request and are counted once per condition.
 - *Iteration cost versus visible re-proposal.* The 28–63% figure counts duplicate pairs (D_t), which Table 1 does not show. Recomputed from the loop logs: every step with $B_t = 1$ has $D_t = 1$, and steps with $D_t = 1$, $B_t = 0$ (47 Gemma-4, 11 Mistral, 40 Qwen3-VL-32B) occur only under the W8 view, where the earlier attempt has left the window. Denominators are all valid loop steps, not only eligible ones (analysis/model.py, COST).
 - *Condition-D fallback.* Recomputed from the stored states: 47% of replaced entries used base-harness failures, in 95 of 120 states; on the 25 states without fallback, $D - A$ is -0.20 $[-0.32, -0.09]$. Only 1 state has replacements that come from the same evaluation and app and match the original failures in length within 10 characters, so a strict type swap needs newly constructed prompts.
 - *Cached replies in the regime test.* Matching replies against the ledger: 17 of 120 replies were served from the cache because the request equalled an earlier one; 16 are the original loop replies (I-200 on instruct-loop states, T-4000 on thinking-loop states). States were sampled without regard to these replies.
 - *Invalid replies.* 49 of 1080 loop steps were invalid: 35 Gemma-4 and 10 Qwen3-VL-32B replies hit the 200-token limit before CHOICE:, Mistral named an edit already in the harness 3 times, and one catalog-B reply named a code that does not exist. Counting them as compliant or as re-proposals gives $\bar{\Delta} = 0.31$ $[0.26, 0.36]$ and 0.36 $[0.31, 0.40]$.
 - *Formulas and estimators.* Identity 1 was checked numerically on 10,000 random weightings (agreement to numerical precision). Every reported $\hat{\alpha}$ satisfies its score equation, and the Mantel–Haenszel estimator recovered a known common log odds ratio in simulated sparse strata of three calls. No eligible state has $X_s = M_s$, so $\beta^U \in (0, 1)$ throughout. Replies naming an edit already in the harness are invalid in the loops and counted as $B = 0$ in the state edits (1.3% and 0.5% of valid replies).
 - *Divergent log-odds fits.* An undamped Newton iteration diverged for Mistral, whose rate is far above the reference. Outcome: damped steps with a log-likelihood check; all fits converge and the intercept-only estimates match a grid search of the likelihood.
 - *Earlier replay disagreed for Mistral.* On violation-selected states, removing failures did

not lower Mistral’s rate; on random states it does (-2.02 [$-3.63, -1.09$]). Outcome: attributed to selection on a prior violation (App. I).

- *Pipeline faults.* A non-reentrant lock in the budget check hung the first pilot (fixed before registration; no data affected). The first-round second annotator used the state summary against the codebook; we report that round rather than re-prompting, and blinded the second round.

E Supplementary Results

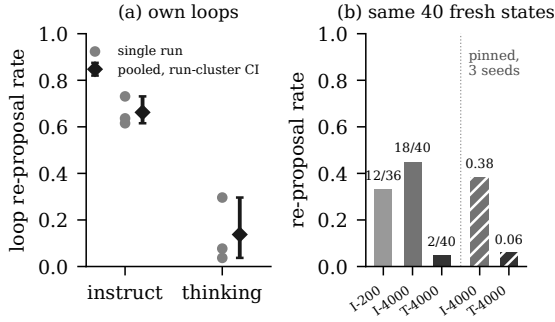


Figure 3: Instruct vs. thinking (Qwen3-30B-A3B-2507). (a) Re-proposal rate in each model’s own loops. (b) The same 40 fresh states answered by the instruct model with a 200- or 4000-token cap and by the thinking model with a 4000-token cap: unpinned routing, one call per state (left), and both endpoints pinned to one provider with three seeds (hatched).

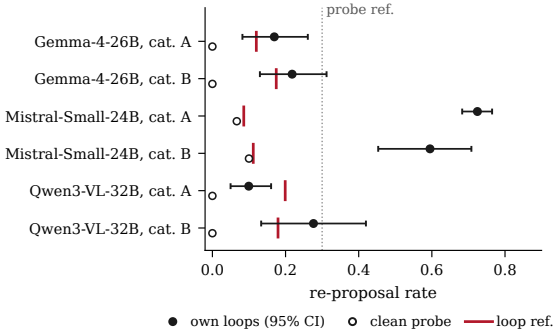


Figure 4: Raw re-proposal rates against the uniform-menu reference. Filled: own loops, run-cluster 95% CIs; open: clean probe. Red ticks: each loop’s reference rate $|X_t|/(40 - |h_t|)$; dotted line: the probe’s reference, $L/40$.

Table 11 gives the round-4 contrasts per model, Fig. 3 the regime comparison and Fig. 6 the code frequencies of §5.4. Fig. 4 shows the raw rates behind Table 1, and Fig. 5 the registered rate differences behind Table 2. Table 12 gives the per-model registered contrasts and Table 13 every loop. In the regime study, on states sampled from the instruct model’s loops the thinking model’s rate is lower by -0.62 , and on states from its own loops by -0.46 ;

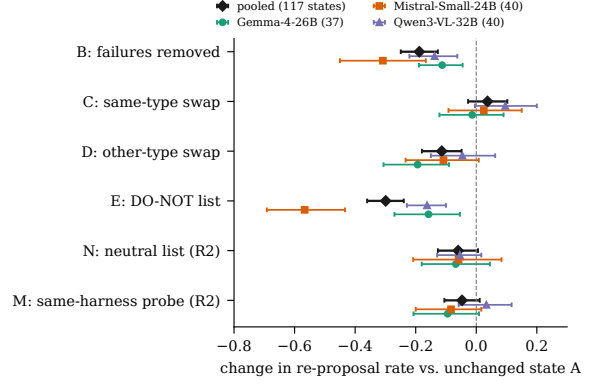


Figure 5: State edits as differences in rate, pooled and per model (state counts in the legend), 95% bootstrap CIs over states.

the instruct model gave 10 invalid replies on eligible loop steps and the thinking model 0.

Reconstruction sensitivity. Table 10 separates states whose A and M prompts are identical from those whose prompts change. The first 117 states contribute zero by construction to $M - A$. On the remaining 243 states, $M - A$ is -0.057 (90% CI [$-0.097, -0.016$]; run-hierarchical [$-0.113, 0.003$]). This split and its per-model estimates are post hoc; the preregistered H15 result uses all 360 states.

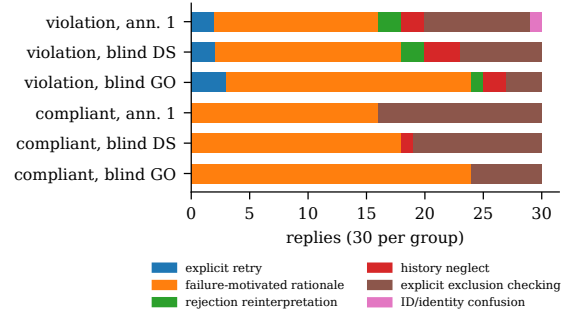


Figure 6: Stated reasons for re-proposals and avoiding choices: per annotator 1 and the two blinded LLM annotators (30 replies per group; blinded pair $\kappa = 0.62$). “Failure-motivated rationale” abbreviates the preregistered code *failure-target attraction* (App. G), which describes the stated rationale, not a causal mechanism.

F The Two Catalogs

Table 14 lists the 40 edits with their catalog-A and catalog-B IDs. Catalog-B descriptions are reworded; for example, E03 “Accuracy is very important for this task.” becomes “Add to the agent’s instructions: getting this right matters a lot.” Catalog B groups the options under headings (prompt instructions, rendering, decoding), and its task text,

State subset	n	M - A [90% CI]	Run-hierarchical [90% CI]
A and M identical	117	+0.000 [0.000, 0.000]	–
A and M differ	243	–0.057 [–0.097, –0.016]	[–0.113, 0.003]
Gemma-4-26B-A4B	98	+0.005 [–0.060, 0.068]	–
Mistral-Small-3.2-24B	71	–0.099 [–0.178, –0.019]	–
Qwen3-VL-32B	74	–0.099 [–0.164, –0.036]	–

Table 10: Same-harness reconstruction M – A by whether the prompt changes (post hoc). The A/M equality was checked by rerendering all 360 prompts. Differences average state-level rates from valid replies; 90% catalog-by-model stratified bootstrap CIs, with a run-then-state hierarchical sensitivity for the changed subset. Per-model rows subdivide that subset. The registered H15 equivalence test remains the all-state contrast in Table 5.

Contrast	Gemma-4-26B-A4B	Mistral-Small-3.2-24B	Qwen3-VL-32B
B – A	+0.03 [–0.05, 0.12]	–0.31 [–0.45, –0.17]	–0.22 [–0.30, –0.13]
R – A	+0.05 [–0.03, 0.14]	–0.04 [–0.18, 0.10]	–0.21 [–0.30, –0.12]
J – K	+0.17 [0.05, 0.30]	+0.04 [–0.17, 0.24]	–0.02 [–0.16, 0.12]
A – H	–0.57 [–0.71, –0.42]	+0.31 [0.12, 0.49]	–0.36 [–0.53, –0.19]
F – N	–0.14 [–0.25, –0.04]	–0.22 [–0.37, –0.05]	–0.11 [–0.18, –0.05]
P – F	+0.00 [–0.07, 0.08]	–0.15 [–0.26, –0.05]	–0.01 [–0.04, 0.00]
E – A	–0.05 [–0.09, 0.00]	–0.53 [–0.64, –0.41]	–0.24 [–0.33, –0.16]

Table 11: Round-4 state-level rate differences per model (fresh states; registered as descriptive, no multiplicity correction), with bootstrap CIs over states. The models differ in kind: deletion and content effects are small for Gemma-4, deletion acts mostly through non-semantic text bulk for Mistral, and the history raises re-proposal for Mistral only; the status sentence (F – N) is negative for all three.

question and history line format differ from catalog A.

G Codebook

The codebook (notes/CODEBOOK.md) defines eight registered codes: explicit retry, failure-target attraction, rejection reinterpretation, history neglect, salience copying, explicit exclusion checking, ID/identity confusion and other. Its decision rules were added by annotator 1 after reading the 60 sampled replies, before recording any code: (i) codes 1 and 3 concern the chosen edit; (ii) code 6 applies when the reply names an earlier attempt that was not kept and uses it as a reason to choose something else, or says that it must pick something untried; (iii) code 2 applies if the rationale cites the shown failures, and code 4 if it argues only from the catalog or from general principles; (iv) code 7 includes treating an edit not in the harness as if it were. Precedence is 1, 3, 6, 7, then the dominant of 2, 4 and 5. Blinded DeepSeek agrees with annotator 1 at $\kappa = 0.80$, and all three annotators agree on 43 of 60 replies, 29 of them failure-target attraction.

H Positioning Matrix

Table 15 summarises how each line of related work (§2) relates to the question of this paper.

I Earlier Evidence on the Same Benchmark

An earlier study on this benchmark and catalog A (unpublished; its data are read-only inputs here) motivated the present design. Its results are not part of this paper’s registration, and we list them with their original status.

Probe and loop (exploratory). Four open proposers declined their own rejected favourite at 81–100% of probe positions, but declined visible rejections in only 28–82% of loop steps.

Near-miss control (exploratory). Over (visible rejection, step) pairs, clearly failed edits (development change ≤ -2 items) were re-proposed in 254 of 4,626 pairs (0.055), and near-misses (within one item of the threshold) in 467 of 4,921 (0.095). Within runs neither was consistently higher (near-miss higher in 23, failed higher in 24 of 57 runs), so a rational retry of near-misses does not explain re-proposal.

Instructed loops (earlier registered). A true sentence stating that evaluation is deterministic and that rejected edits must not be proposed again changed the pooled rate from 0.33 (330/988) to 0.29 (292/997), in runs compared with uninstructed runs on the same seeds.

Replay on violation-selected states (earlier registered; biased). In 119 stored states where a re-proposal had occurred, a single replay call re-produced a visible re-proposal in 67. Removing failure examples gave 44 and a DO-NOT list gave 8 (Table 16). Selecting states on a prior violation biases these rates upward and makes removal effects partly regression to the mean; one call per state cannot separate an intervention from temperature-0 variation. The random-state design of §4.2 replaces this analysis.

Reasoning-mode proposer (earlier registered). gpt-oss-120b (OpenAI et al., 2025) honoured

Proposer (states)	$B - A$	$C - A$	$D - A$	$E - A$
Gemma-4-26B-A4B (37)	-0.11 [-0.19, -0.05] (0.054)	-0.01 [-0.12, 0.09] (1.000)	-0.19 [-0.31, -0.09] (0.018)	-0.16 [-0.27, -0.05] (0.063)
Mistral-Small-3.2-24B (40)	-0.31 [-0.45, -0.17] (0.002)	+0.03 [-0.09, 0.15] (1.000)	-0.11 [-0.23, 0.01] (0.546)	-0.57 [-0.69, -0.43] (<0.001)
Qwen3-VL-32B (40)	-0.14 [-0.22, -0.06] (0.021)	+0.10 [0.00, 0.20] (0.546)	-0.05 [-0.15, 0.06] (1.000)	-0.16 [-0.23, -0.10] (<0.001)

Table 12: Per-model state-level rate differences on the original 120-state intervention round (registered secondary family; Holm correction over 12 contrasts; p_{Holm} in parentheses). States with no valid reply under A are dropped from that model’s contrasts.

Proposer	Cat.	View	Seed	Elig.	Valid	B	Kept	Final dev
Gemma-4-26B-A4B	A	FULL	20	26	21	5	3	0.850
Gemma-4-26B-A4B	A	FULL	21	27	17	1	3	0.867
Gemma-4-26B-A4B	A	FULL	22	25	23	8	3	0.850
Gemma-4-26B-A4B	A	W8	20	25	22	2	3	0.850
Gemma-4-26B-A4B	A	W8	21	26	24	1	3	0.850
Gemma-4-26B-A4B	A	W8	22	25	23	5	3	0.867
Gemma-4-26B-A4B	B	FULL	20	28	27	8	1	0.800
Gemma-4-26B-A4B	B	FULL	21	28	26	5	1	0.800
Gemma-4-26B-A4B	B	FULL	22	27	26	11	2	0.817
Gemma-4-26B-A4B	B	W8	20	28	26	5	1	0.800
Gemma-4-26B-A4B	B	W8	21	27	26	3	2	0.817
Gemma-4-26B-A4B	B	W8	22	27	25	2	2	0.817
Mistral-Small-3.2-24B	A	FULL	20	28	28	21	1	0.800
Mistral-Small-3.2-24B	A	FULL	21	27	27	19	2	0.817
Mistral-Small-3.2-24B	A	FULL	22	29	29	23	0	0.750
Mistral-Small-3.2-24B	A	W8	20	27	27	19	2	0.850
Mistral-Small-3.2-24B	A	W8	21	28	28	18	1	0.783
Mistral-Small-3.2-24B	A	W8	22	28	28	21	1	0.800
Mistral-Small-3.2-24B	B	FULL	20	27	27	20	2	0.817
Mistral-Small-3.2-24B	B	FULL	21	27	27	20	2	0.833
Mistral-Small-3.2-24B	B	FULL	22	27	27	13	2	0.817
Mistral-Small-3.2-24B	B	W8	20	28	28	20	1	0.783
Mistral-Small-3.2-24B	B	W8	21	26	25	14	3	0.850
Mistral-Small-3.2-24B	B	W8	22	26	24	7	3	0.850
Qwen3-VL-32B	A	FULL	20	27	27	1	2	0.817
Qwen3-VL-32B	A	FULL	21	27	27	6	2	0.817
Qwen3-VL-32B	A	FULL	22	26	26	3	3	0.867
Qwen3-VL-32B	A	W8	20	27	27	4	2	0.817
Qwen3-VL-32B	A	W8	21	27	27	1	2	0.817
Qwen3-VL-32B	A	W8	22	27	27	1	2	0.817
Qwen3-VL-32B	B	FULL	20	27	25	8	2	0.817
Qwen3-VL-32B	B	FULL	21	28	28	16	1	0.800
Qwen3-VL-32B	B	FULL	22	27	22	6	2	0.817
Qwen3-VL-32B	B	W8	20	28	28	9	1	0.800
Qwen3-VL-32B	B	W8	21	26	24	0	3	0.850
Qwen3-VL-32B	B	W8	22	26	25	3	3	0.850

Table 13: Every replication loop. Elig.: steps with $X_t \neq \emptyset$; Valid: those with a valid reply; B : re-proposals of a visible rejection; Kept: edits in the final harness; Final dev: final development accuracy.

290/290 probe cells and re-proposed in 16 of 327
loop steps (rate 0.05).

A	B	Kind	Catalog-A description
E01	RF1	prompt	Briefly describe the screen before choosing an action.
E02	ZQ7	prompt	Double-check your answer before replying.
E03	NK3	prompt	Accuracy is very important for this task.
E04	WV1	decode	Lower the output token limit from 48 to 16 to save cost.
E05	VG3	prompt	If you are unsure, choose the most prominent element on the...
E06	YV2	prompt	Put the final reply on the last line, prefixed with FINAL:.
E07	AF9	render	Enlarge the tag labels to 13px bold.
E08	FG6	prompt	Consider what the user is ultimately trying to accomplish.
E09	DG4	prompt	In label-value rows the value is right-aligned on the same ...
E10	YJ8	render	Outline every tagged element with a thin box in the tag's c...
E11	BK6	prompt	Calendar events are coloured blocks whose label starts with...
E12	AN3	prompt	In the inbox, the sender's name is the bold first line of e...
E13	WN6	prompt	For questions, copy the value exactly as displayed, includi...
E14	NM3	render	Increase screenshot contrast by 30%.
E15	PZ9	prompt	Take a deep breath and work on this problem step by step.
E16	VK8	prompt	Icon legend: [icon] back, [icon] archive, [icon] delete, [i...
E17	CV5	render	Hide the tags of elements that are covered by another eleme...
E18	GZ3	prompt	Switches: green with the knob on the right means On; grey w...
E19	MV5	decode	Raise the output token limit from 48 to 256 so the agent ca...
E20	CD7	prompt	If a dialog or pop-up is open, only the elements inside the...
E21	RT8	prompt	To toggle a setting, tap its switch rather than the text la...
E22	JV4	prompt	Be careful and precise.
E23	DZ8	prompt	First list the tags you can see, then choose one.
E24	RX7	prompt	Here is an example of a correct reply to 'Open settings': T...
E25	JJ6	prompt	Do not explain your answer; output only the single reply line.
E26	HW7	prompt	Each tag number is drawn just ABOVE the top-left corner of ...
E27	CM1	render	Downscale screenshots to 0.75x to save visual tokens.
E28	XH2	render	Render screenshots at 1.5x resolution (more visual tokens).
E29	RA1	prompt	You are an expert mobile GUI agent with years of experience.
E30	XR4	render	Draw each tag inside the top-left corner of its element ins...
E31	KR8	prompt	Tasks that ask you to do something (open, turn on, confirm,...
E32	HR6	prompt	Prefer elements in the top toolbar when several elements co...
E33	JG7	prompt	Tags are numbered in reading order: top to bottom, left to ...
E34	WA9	prompt	Answer confidently; do not hedge.
E35	QD8	prompt	Greyed-out buttons are disabled and cannot be pressed.
E36	ZI8	render	Convert screenshots to grayscale.
E37	PM2	prompt	Reply as JSON: {"action": "TAP" or "ANSWER", "value": <tag ...
E38	XX3	prompt	Ordinal words in a task (first, fifth, ...) count items on ...
E39	AA3	prompt	Think step by step before giving your final answer.
E40	BW8	render	Give each tag a distinct colour instead of all red.

Table 14: The 40 edits. A: catalog-A ID; B: catalog-B code; kind; catalog-A description (truncated at 62 characters).

Line of work	What is established	Assumption we test	How we test it
LLM optimizers with shown history (Yang et al., 2024; Agrawal et al., 2026; Lee et al., 2026; Lu et al., 2026; Yang et al., 2026)	showing tried candidates and failures helps search	a shown rejection is avoided	α in reached states; no-history baseline (§5.1)
Static instruction following (Zhou et al., 2023a; Jiang et al., 2024)	compliance in one response	avoidance is a trait measured once	clean probe vs. own loop vs. reconstructed state (§5.1)
Negative and omission constraints (Castricato et al., 2024; Rana, 2026; Gamage, 2026)	prohibitions are fragile; prior pressure drives violations	pressure is a fixed prior	no-history baseline at the reached harness; status vs. prohibition (§5.3)
Failure records and repetition (Gumaan, 2026; Yang, 2026; Xu et al., 2022)	a visible failed action is repeated; “do not repeat” does not help (135M–1.7B)	the failure record drives repetition	approximately length-matched neutral text, attachment shuffle, status sentence (§5.2–5.3)
Knowing vs. acting (Kruthof, 2026; Dong et al., 2026)	models restate what they violate	stated reasons explain choices	stated reasons vs. content controls (§5.4)
Reasoning vs. compliance (Fu et al., 2025; Li et al., 2025)	reasoning lowers static-instruction compliance	this generalises to all constraints	instruct/thinking pair, provider-pinned (§5.4)

Table 15: Positioning matrix. Each line of work, what it establishes, and the assumption that a history-implied exclusion constraint puts under test.

Proposer	States	Replay (A)	No failures (B)	DO-NOT (E)
Gemma-4-26B-A4B	27	10	2	2
Mistral-Small-3.2-24B	48	39	42	6
Qwen3-VL-32B	44	18	0	0

Table 16: Earlier replay on violation-selected states (one call per state): states with a visible re-proposal.