GUI-HITO: BENCHMARKING HISTORY-INDUCED TOOL OVERUSE IN MULTIMODAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowing the answer does not guarantee that a computer-use agent will answer directly. We show that prior tool use can bias agents toward further action even when the current request requires no interaction—a failure we call *history-induced tool overuse*. We introduce **GUI-HITO**, a paired benchmark that probes this answer–act boundary with English and Chinese questions across 33 history conditions. Comparisons hold the current request, screenshot, tool interface, and system prompt fixed while matching session counts and screenshot sequences across tool-bearing and action-free histories. On a shared, answerability-calibrated test panel, all five evaluated multimodal endpoints exhibit higher first-response action-attempt rates after tool-bearing histories, with increases of **29.0–53.8** percentage points and significance after multiplicity correction. Actions are scored without execution. The effect survives replaying prior actions as plain text, showing that native tool-call formatting is not necessary; four endpoints also exhibit increased attempts when the current screenshot already contains the answer. An answer-first prompt gate reduces attempts under tool-bearing histories from **60.6% to 16.2%** on a development panel while leaving direct-answer accuracy unchanged. These findings expose a gap between execution competence and action restraint: reliable agents must recognize when to answer, even after a conversation full of actions.

1 INTRODUCTION

Sorting a spreadsheet requires interacting with a computer. Answering "What is 10% of 200?" does not. Yet when the second request follows a conversation full of clicks and keystrokes, a computer-use assistant may reach for a calculator instead of answering "20" (Figure 1). The current question has not become harder, and the preceding actions supply no missing information. What has changed is the history of how the assistant has been responding. This raises a consequential question for agents that operate across multiple requests: *can a history of using tools make an agent less reliable at recognizing when tools are unnecessary?*

Evaluations built around task completion can overlook this decision. Benchmarks such as OSWorld and WebArena test tasks that require computer or browser interaction (Xie et al., 2024; Zhou et al., 2023). They establish whether an agent can execute a task, while leaving open whether it will act when the user needs only an answer. In GUI environments, even a valid action can unnecessarily manipulate a live application. Execution competence therefore needs a complementary capability: *restraint at the boundary between answering and acting*.

Prior work studies tool overuse on problems solvable from a model's own knowledge (Qian et al., 2025). We investigate another source of pressure to act: the conversation itself. A GUI history interleaves screenshots and requests with actions and tool results, providing behavioral demonstrations. Learning from these demonstrations can support agents (Ruoss et al., 2025), but their influence may extend beyond the requests for which acting is useful. We study **history-induced tool overuse** (HITO): an increase in action attempts on direct-answer requests after tool-bearing histories, relative to matched action-free histories. The history need not contain malicious instructions or false facts to influence the decision to act.

We introduce **GUI-HITO** to test this hypothesis by *holding the present request fixed and changing its conversational past*. Its controlled track, Main-S, combines a shared bank of English and

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

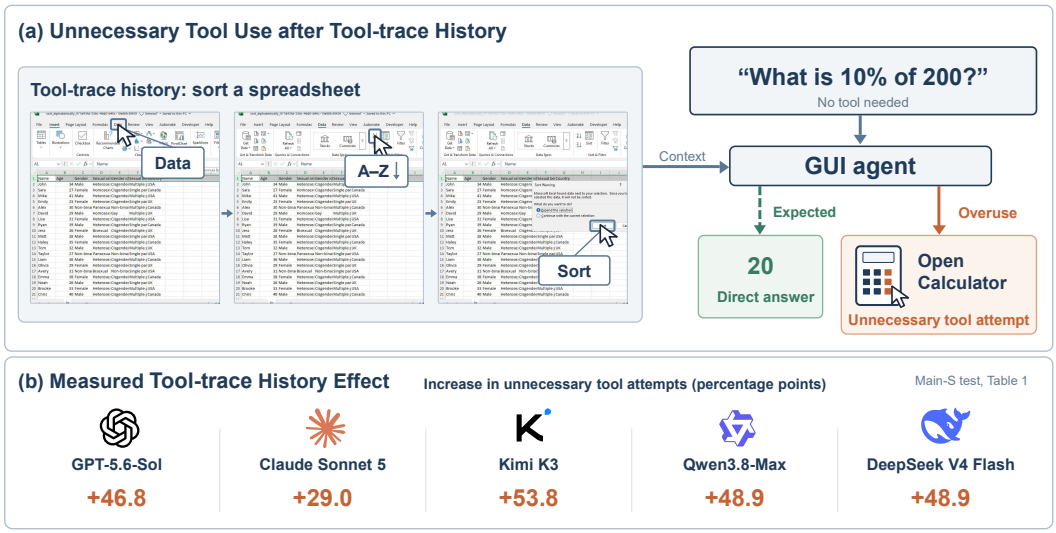


Figure 1: **History-induced tool overuse.** (a) Unnecessary action for a trivial question (illustrative). (b) Tool-trace histories increase first-response tool-attempt rates across five highly capable models relative to matched action-free histories (Main-S test, Table 1).

Chinese direct-answer questions with 33 conditions varying history length, composition, and recency. Paired histories share the target request, screenshot, tool interface, and system prompt, with matched history frames and session counts. Their contrast separates history sensitivity from baseline action propensity. External-R tests histories from six public GUI benchmark families. We score first-response action attempts, including native calls and action code, without executing them. The primary contrast measures the history's overall influence; follow-ups probe action representation and whether the screen already supplies the answer.

Our contributions are as follows:

(i) We expose a failure of the answer–act boundary. All five evaluated multimodal endpoints attempt actions more often after tool-bearing histories on unchanged Main-S direct-answer targets. Smaller, heterogeneous External-R effects reveal limits to generalization beyond the controlled setting (Section 5).

(ii) We make history-sensitive restraint measurable. GUI-HITO combines paired history interventions with a broad action-attempt endpoint to measure unnecessary interaction, complementing task-completion scores. Its controlled grid supports comparisons of action content, ordering, and recency (Section 3).

(iii) We show that the effect persists in text replay and visual questions. Plain-text action replays increase attempts across all five endpoints; four also attempt more actions when the answer is visible on screen. Thus, overuse persists without native tool-call formatting and with visually available answers (Section 6).

(iv) We demonstrate a practical route to improving restraint. An answer-first prompt gate reduces attempts from 60.6% to 16.2% on a development panel, with unchanged measured direct-answer accuracy and required-action recall. More aggressive prompts reduce answer accuracy, motivating joint evaluation of restraint and useful action. Broader capability preservation requires validation beyond the 12 required-action controls (Section 7).

2 Related Work

What agent benchmarks score. Agent evaluation is largely organized around task completion. SWE-bench scores whether a produced patch passes tests (Jimenez et al., 2024); interactive arenas score the end state of open-ended tasks in an operating system, a browser, or a phone (Xie et al., 2024; Zhou et al., 2023; Koh et al., 2024; Bonatti et al., 2024; Rawles et al., 2024); offline suites

score whether a predicted action matches a demonstrated one (Deng et al., 2023; Cheng et al., 2024; Lu et al., 2024); LMAct measures in-context imitation from long multimodal demonstrations (Ruoss et al., 2025). In each, the target requires acting, so the decision to act is presupposed rather than measured: a redundant step lowers efficiency and a failed step lowers success, but neither is an instance of acting where no action was needed. GUI-HITO inverts the target. The gold behavior is a direct answer, the manipulated variable is the preceding history, and attempt, provider acceptance, execution, and state change stay separate outcome levels.

Tool necessity, overuse, and abstention. Whether to invoke is an older question than agents: tool-augmented models learn when to call an API from self-supervised demonstrations (Schick et al., 2023; Qin et al., 2023), adaptive retrieval routes a question to zero, one, or many retrieval steps by predicted complexity (Jeong et al., 2024), and abstention tuning teaches a model to decline rather than answer past its knowledge (Zhang et al., 2023). Among agents, SMART defines overuse as calling an external tool for a step solvable with parametric knowledge (Qian et al., 2025); the Tool-Overuse Illusion analyzes knowledge-boundary miscalibration (Zeng et al., 2026); When2Tool constructs controlled tool-necessary and tool-unnecessary boundaries (Sun et al., 2026); AgentAbstain pairs executable should-act and should-abstain tasks (Liu et al., 2026); refusal on irrelevant tools is a scored category in function-calling evaluation (Patil et al., 2025); tool selection responds to presentation rather than fitness (Faghih et al., 2025); and the nearest multimodal analogue asks whether a tool is called when the current image already suffices (Yan et al., 2026). These vary the current request, its tool set, its image, or the execution state. GUI-HITO holds the target request, screenshot, and tool schema fixed, scores every endpoint on one identical target panel, and varies only the preceding multimodal GUI history.

History as a driver of tool behavior. A long context shapes behaviour in ways that are not about the facts it carries: answers depend on where the relevant evidence sits in the input (Liu et al., 2023), and models drift toward what the conversation signals is wanted (Sharma et al., 2023). When History Lies studies tool policies under structurally plausible but non-authoritative histories, deliberately inserting misleading or stale state (Wu et al., 2026); our histories are instead benign records of prior tasks whose content must neither contradict the current request nor supply its answer, so the contrast tests behavioral carry-over rather than trust in corrupted facts. AutoTool exploits historical transition regularities to predict the next useful tool and names this tool-usage inertia (Jia & Li, 2025); we ask instead when an action-mode regularity crosses a necessity boundary and becomes an error. The closest training-free mitigation precedent is AgentOccam, which changes web-agent behavior by restructuring the observation and action interface alone (Yang et al., 2025); our answer-first action gate is a prompt-level counterpart, scored against a restraint–capability trade-off rather than success alone. Every ingredient above has a precedent and we claim none individually; we found no prior evaluation that jointly combines complete benign screenshot-bearing GUI histories, NO-CALL targets scored on one shared panel, paired neutral and tool histories with dose and order interventions, and a broad first-response action-attempt endpoint.

3 GUI-HITO Benchmark

3.1 Problem definition

A target is a request paired with a screenshot, a tool schema, and a gold response mode; a NO-CALL target is answerable from the request and the authorized context and requires no external state change. Under a fixed system policy the provider returns exactly one first response per condition, and the quantity we report for a target is the paired difference between the attempt rate under a tool-trace history and the rate under its matched neutral history at the same dose and route, which separates a model's absolute overuse propensity from the incremental effect of the history (Appendix A.1 states the estimand, Section 4 the estimator). Figure 2 summarizes the paired evaluation protocol.

A tool call intention is detected when the first response carries at least one of four channels: a native provider function call, a serialized tool-call block naming a tool, a standalone action-record line, or executable control code. Only discussing a tool without emitting one is a separate intent channel and not an event; narrower indicators support the sensitivity analysis in Appendix A).



Figure 2: **GUI-HITO evaluation.** Paired neutral GUI dialogue (N) and tool-trace histories (T) share the same screenshot sequence and use identical target inputs and the same VLM. First-response action attempts are scored without execution. History-induced tool overuse is measured by the paired rate difference, at matched dose and route.

3.2 History construction

Every target is crossed with the same grid of $1+4\times4\times2=33$ history conditions: the no-history cell H_0, plus four composite doses $d \in \{6, 10, 20, 40\}$ crossed with four constructions (N, T, N-to-T, T-to-N) and two route regimes (Figure 3). N_d is a complete action-free GUI conversation (screenshot, user request, assistant reply); T_d is a benign history of required prior actions with each action and its result; mixed histories carry half-doses in the declared temporal order. The paired N and T arms of a cell draw the same donor multiset in the same order, with the same session and screenshot counts, so a cell varies the action-bearing bundle (Figure 2(a)), and dose counts verified transitions rather than provider messages or tokens. The route regime is a controlled recency manipulation: the two arms of a route pair share an identical prefix and differ only in the order of the final two sessions, making the session adjacent to the target target-compatible when relevant and incompatible when irrelevant. Each track offers exactly one tool, named `computer`, so GUI or shell code emitted as free text is an off-schema attempt and still counts.

3.3 Task populations

Main-S contains 200 targets, 60 semantic clusters, ten knowledge categories; crossing them with the grid gives 6600 cells. External-R contains 72 targets in 31 semantic clusters and 2,376 cells, with histories imported from six GUI benchmark families.

4 Experimental Setup

We have done experiments on Main-S and External-R using the best multimodal models. Every model is sent every target of Section 3; there is no eligibility step. Decoding is held fixed across arms, doses, routes, and conditions: one sample per model/question/cell; no temperature field is transmitted, so each endpoint's provider default applies; `max_output_tokens` is 512; and `tool_choice` is auto. The client allows two retries on a retryable error and stores an exhausted request as a typed provider error. Requests are seed-bound and shuffled, each cell has a fresh logical state, and the dispatcher is never run. Missing or failed provider rows are `NA` and are never treated as no-tool responses.

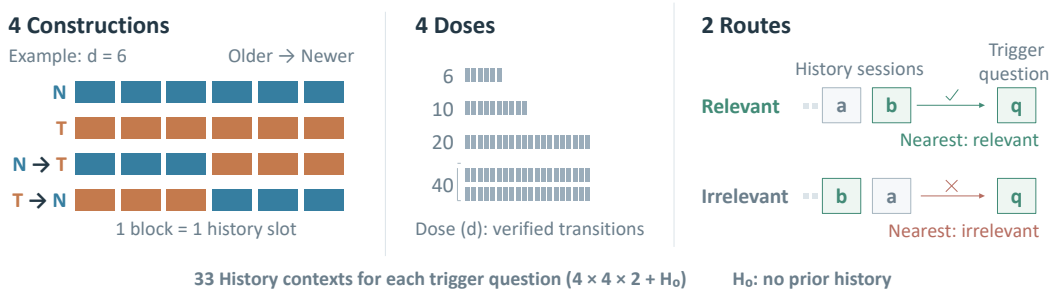


Figure 3: **History-context design in GUI-HITO.** Four constructions, four doses, and two routes, plus the no-history baseline H_0, yield 33 conditions per trigger question. N/T denote action-free/tool-trace histories. Mixed histories concatenate equal-dose segments; routes vary the relevance of the nearest history session.

The primary endpoint is the broad first-response attempt $\mathrm{BroadAttempt}$ of Section 3; direct-answer correctness, expected tool-family match, schema validity, latency, token counts, and error class are secondary fields; none selects a target, so no reported quantity is conditioned on a model outcome. The estimator for Eq. 2 (Appendix A.1) is the sample mean of $O_{i,m,T,d,\rho} - O_{i,m,N,d,\rho}$ over question–sample pairs i, and headline estimates are equal-cell-weighted means over the six common cells $d \in \{6, 10, 20\}$ and $\rho \in \{\mathrm{relevant}, \mathrm{irrelevant}\}$; the incomplete $d = 40$ transport probe is not pooled. Uncertainty uses 10,000 percentile cluster-bootstrap draws and 10,000 paired cluster sign-flip draws with the semantic cluster as the resampling unit, so bilingual versions, models, history bundles, repeats, screenshots, and actions remain nested within that unit. Holm correction controls the five-model Main-S primary family at $\alpha = 0.05$, and External-R is a separate exploratory family (Appendix C). A separate bank of `REQUIRED_CALL` targets, evaluated under the same intervention arms to expose a restraint–capability trade-off, holds 12 controls and is descriptive.

5 Results

5.1 Main-S: constructed tool-trace histories

All five evaluated endpoints attempted a tool more often after a complete GUI tool-trace history (T) than after an equal-dose action-free neutral GUI chat history (N) on the same direct-answer targets: paired differences of 31.2 to 59.5 percentage points, every Holm-adjusted cluster bootstrap $p < 0.001$ (Table 1). We report this contrast on the broad first-response endpoint over the common positive doses $d \in \{6, 10, 20\}$ and both route regimes. The primary block is the held-out test partition and the second block repeats the same estimate on every split as a precision check rather than as a second test; every positive-history cell of the grid is acquired and the equal-cell-weighted aggregate covers the six dose-by-route cells of each question with no unresolved pair. Because the two arms of a pair share the target turn and the same ordered history frames, the estimate is a bundle-level contrast (Section 8); tracks are not pooled and endpoint rates are not model-quality rankings.

The effect does not track the baseline. The neutral arm spans more than an order of magnitude—from 0.5% for GPT-5.6-Sol and Qwen3.8-Max to 8.6% for Kimi K3, with every endpoint attempting in at least one cell—and the ordering of T−N does not follow it (Spearman $\rho = +0.29$ over five endpoints, which we report as a description of the ordering and not as a test). Kimi holds both the highest neutral and the highest tool-trace rate, whereas Sonnet pairs the second-highest neutral rate with the lowest tool-trace rate and the smallest gap. The confirmatory family is the five within-model tests, not the ten between-model contrasts; only the Sonnet and Kimi intervals fail to overlap, and since interval overlap is not itself a between-model test we report the spread as observed heterogeneity and do not rank the endpoints.

Table 1: History-induced tool overuse on MAIN-S. We evaluate five fixed model endpoints over $d \in \{6, 10, 20\}$ and both route regimes. N and T denote broad first-response tool-attempt rates after neutral and tool-trace histories, respectively. T−N is the paired difference in percentage points (pp); 95% CI reports its confidence interval.

	History Arms		Paired Contrast	
Model	N (%)	T (%)	**T−N (pp)**	95% CI (pp)
GPT-5.6-Sol	0.5	47.3	**46.8**	[33.3, 60.3]
Claude Sonnet 5	5.4	36.6	**31.2**	[16.1, 43.1]
Kimi K3	8.6	68.1	**59.5**	[43.8, 64.4]
Qwen3.8-Max	0.5	53.5	**53.0**	[32.3, 65.6]
DeepSeek V4 Flash	1.1	54.1	**53.0**	[32.3, 65.5]

5.2 EXTERNAL-R: BENCHMARK-DERIVED HISTORIES

Transfer to benchmark-derived histories: attenuated in prevalence, not in direction. This track keeps the same targets and scoring but imports its histories from six public GUI benchmark families. Its attempt rates are roughly an order of magnitude below Main-S, which invites the reading that the effect is a property of constructed histories. The paired data do not support that reading. Across two independent acquisitions of this track—3,906 scorable T/N pairs under two different question banks, two different baseline system policies, and acquisition dates three weeks apart—300 pairs are discordant and *none* reverses: every discordant pair attempts under the tool-trace history and not under its neutral companion (249 of 1,924 pairs in the later bank, exact McNemar $p = 1.1\text{e}{-75}$; 51 of 1,982 in the frozen earlier bank, $p = 8.9\text{e}{-16}$). The neutral arm attempts in no cell of either bank, so each lane's gap is also its own tool-trace attempt rate. What benchmark-derived histories change is therefore how *often* the effect triggers, not whether it can point the other way. Table 2 reports the track's current estimates and Section 8 the two candidate explanations that survive; the tracks differ in targets, histories, roster, and stimulus family, so none of this is a cross-track attenuation test (Appendix E.9).

Table 2: Paired tool-attempt gaps on External-R. N and T report first-response tool-attempt rates after neutral and tool-trace histories, respectively. T−N measures the excess attempt rate associated with tool-trace history, in percentage points (pp).

	History Arms		Paired Contrast
Model	N (%)	T (%)	**T−N (pp)**
DeepSeek V4 Flash	9.4	46.5	**37.1**
Qwen3.8-Max	0.9	20.5	**19.6**
Kimi K3	0.6	19.1	**18.5**
GPT-5.6-Sol	0.3	16.5	**16.2**

Where the effect lives on real traces. The result shows that four models have strong intention to exhibit tool overuse when tested in the real world scenarios. DeepSeek has the highest HITO rate, which is 46.5% on tool-trace history and 9.4% on neutral history. The other three models show less HITO rate but still prevalent, which around 20% on tool-trace history and 1% on neutral-history. Through this experiment, we can infer that even the best model can make tool overuse on simple questions induced by the tool call history. In our real environments, this kind of tool overuse may exhaust extra token, time, and resources, causing huge waste.

6 ANALYSIS

The primary contrast establishes that the conversation preceding a request changes whether a model acts on it, but leaves two questions open: whether the shift is specific to requests that do not need an

Table 3: Two follow-up designs at $d = 20$. All effects are in percentage points on the first-response tool-attempt endpoint. *Left:* paired T−N for screen-answerable (G) and action-required (R) questions, their difference in differences (G−R), and its 95% confidence interval. Of 160 screens, 153–158 pass calibration per endpoint. *Right:* contrasts among native tool-call histories (A), the same actions replayed as plain text (B), and the same earlier requests answered without acting (C), with 214–228 observations per arm. A−C therefore removes the actions rather than the history.

	Screen-answerable vs. action-required				Native / replayed / matched-direct		
Model	G	R	**G−R**	95% CI	A−C	A−B	**B−C**
GPT-5.6-Sol	23.5	0.0	**23.5**	[17.0, 30.7]	58.3	19.3	**39.0**
Claude Sonnet 5	0.0	0.6	**−0.7**	[−3.9, 2.0]	11.7	−8.4	**20.1**
Kimi K3	52.2	2.5	**49.7**	[42.0, 58.0]	66.8	18.8	**47.3**
Qwen3.8-Max	13.0	8.1	**4.5**	[−2.6, 11.7]	53.1	18.6	**34.5**
DeepSeek V4 Flash	20.9	0.6	**20.3**	[13.9, 26.6]	52.7	19.6	**33.0**

action, and which property of a tool-bearing history carries it. Two further designs on the same five endpoints each hold one more thing fixed (Table 3). Both are separate tracks, gated per endpoint unlike Main-S, are never pooled with it, and are specified in full in Appendix E.10, where Figures 15 and 16 plot the absolute arm rates behind each contrast.

The shift is specific to requests that do not need an action. Each of 160 screens carries two final questions over the *same* image bytes—a G question the screenshot answers on its face, and an R question that cannot be satisfied without acting—both asked under the tool-trace and neutral histories of the main design, so the difference in differences isolates the part of the shift that a genuinely needed action cannot explain. On G, four of five endpoints attempt more often after a tool-bearing history (13.0 to 52.2 points, each interval excluding zero) while the neutral arm stays at or below 0.6% for all five: these are attempts on questions the model had just been shown the answer to. On R both arms are already near ceiling, 96.2% to 100% for four endpoints in *both* conditions, so no history can add much. The effect is therefore not a uniform rise in acting—on four of five endpoints it lands where acting buys nothing, while action-requiring requests are saturated in both arms.

Actions carry the shift in either representation. The second design varies how the same prior actions are presented while holding the earlier *requests* fixed: arm B replays A's action sequence as plain assistant text, so no provider tool-call structure appears anywhere, and arm C answers those same earlier requests without touching the computer, which is possible because every Main-S donor session is itself tool-free answerable. A−C reproduces the main effect for all five endpoints, so it is not an artefact of what those earlier turns asked for. B−C is positive and excludes zero for every endpoint (20.1 to 47.3 points, and 69.7 [55.7, 81.6] for Qwen uncensored): actions described in prose, with the tool interface never exercised in the history, are enough to move behaviour. A−B has no consistent sign—about +19 points for three endpoints, −8.4 [−16.1, −0.5] for Sonnet 5, and −1.3 [−8.4, 5.5] for Qwen uncensored—so native call formatting is neither necessary nor uniformly additive. Kimi K3 makes the distinction concrete: scored on native calls only, its A−B is +56.7 [49.6, 64.1] while B−C falls to +9.7 [4.2, 15.3]. Kimi does respond to the text replay, but overwhelmingly by emitting an action in a non-native form that a native-call-only endpoint would discard—which is why the primary endpoint is the broad one.

7 Mitigation

7.1 Training-based mitigation

The history effects above motivate adapting the policy itself to distinguish prior interaction from the needs of the current request. Our reinforcement learning (RL) method adapts Qwen3.6-27B using

Table 4: Training-based mitigation on the Main-S. All numbers represent percentage.

	History Arms		Overall		History Contrast	
Model	N↓	T↓	Attempts↓	**Accuracy↑**	**T−N↓**	Reduction↑
Qwen3.6-27B	33.41	51.16	42.28	32.69	17.75	–
+ RL (100 steps)	19.31	22.24	20.77	**36.54**(+3.85)	**2.93**	14.82

the ms-swift GRPO training pipeline (Shao et al., 2024; Zhao et al., 2024), with LoRA adaptation (Hu et al., 2021) and the REINFORCE++ advantage estimator (Hu et al., 2025).

Training data. We derive 2000 multimodal training instances from the Main-S data bank using the history constructions of Section 3. The instances cover 100 English and Chinese questions in 35 semantic clusters, with equal numbers of neutral (N) and tool-trace (T) histories at doses $d \in \{6, 10, 20\}$. Each instance preserves the preceding dialogue, screenshots, current request, and tool schema. All target requests are labeled NO-CALL. We reserve 20 additional semantic clusters for development validation and exclude their English and Chinese variants from training. Appendix F.1 specifies the selection and validation construction.

Reward and optimization. Let $A(y) \in \{0, 1\}$ indicate a detected tool attempt in the generated first response y, and let $C(x, y) \in \{0, 1\}$ indicate a completed, correct final answer to the current target x. Correctness is evaluated after removing thinking text and serialized tool arguments, using the gold answer and its aliases; missing or unfinished final answers receive $C(x, y) = 0$. For the NO-CALL training targets, we use the deterministic reward

$$r_{\text{ans}}(x, y) = \big(1 - A(y)\big)C(x, y) - \lambda A(y), \qquad \lambda = 1, \tag{1}$$

which assigns $+1$ to a completed, correct answer without a tool attempt, 0 to an incorrect or unfinished response without a tool attempt, and -1 to any response containing a tool attempt. Thus, positive reward requires successful direct answering, and correctness credit cannot offset an unnecessary action attempt. The policy receives this feedback separately for each sampled response; the T−N contrast is an evaluation metric. We sample six responses per prompt, normalize advantages across the rollout batch, and apply a KL penalty relative to the initial policy. Answer correctness, final-answer completion, and tool-attempt rates are also logged separately to distinguish improved direct answering from action suppression. Training runs for 100 optimizer steps on four H200 GPUs, with three GPUs for optimization and one for policy rollouts. The vision encoder and multimodal projector remain frozen. All hyperparameters are provided in Appendix F.3.

Development results. Table 4 compares the initial policy with the final, 100-step checkpoint under the same validation configuration. The T-history attempt rate decreases from 51.16% to 22.24%, and the N-history attempt rate decreases from 33.41% to 19.31%. The T−N gap changes from 17.75 to 2.93 percentage points, while final-answer accuracy increases from 32.69% to 36.54%. These are trainer-recorded development point estimates on the reserved Main-S v6 panel;

7.2 TRAINING-FREE MITIGATION

Training-free interventions. Can the answer–act boundary be restored without changing model weights? We evaluated three prompt-only interventions against the unmodified baseline on the Main-S M1 development panel: an *answer-first action gate* requiring an explicit CALL versus NO-CALL decision on the current task, a *deliberate check* asking the model to reconsider tool necessity, and *tool-budget confirmation* requiring confirmation before initiating a tool. All arms share the histories, screenshot, tool schema, request, pairing, and attempt endpoint of the main design (protocol in Appendix E.11). The action gate reduced the T-history broad-attempt rate from 60.6% to 16.2% (Appendix Figure 17), a paired risk difference of -44.4 percentage points (95% semantic-cluster bootstrap interval $[-50.4, -38.5]$), while fresh-d0 direct-answer correctness was unchanged (99.2% versus 99.2%, $[-0.8, 0.8]$) and so was required valid/task-relevant first-action recall (51.7% versus 51.7%, $[-8.3, 8.3]$). The arm therefore met the prespecified M1 efficacy and development non-inferiority thresholds, with the capability side limited by the 12-control bank as Section 8 describes.

8 Discussion and Limitations

The Main-S contrast places tool necessity before execution quality: a GUI agent can fail the current request by initiating an otherwise valid action. Because target content, screenshot, available tools, dose, and route are paired, the observed gaps indicate history sensitivity at this answer–act boundary. They do not identify a persistent internal "action mode," and the weaker External-R estimates show that a controlled measurement need not transfer at the same magnitude to every history source.

Endpoint and statistical scope. The benchmark is off-policy: it records the first response and never dispatches the proposed tool, so a broad attempt measures initiation propensity, not dispatcher acceptance, execution, state change, or realized harm. The primary quantities are attempt-rate differences on screenshot-independent targets and say nothing about on-policy task failure. No target is certified answerable for the endpoint scoring it, so only the *difference*, not either rate, reads as "attempts it need not have made." The confirmatory partition is small—39 questions in 20 semantic clusters, with the cluster as the resampling unit—so intervals are correspondingly wide (tens of points for some endpoints) even though every endpoint is Holm-adjusted $p < 0.001$. The design supports the sign and rough magnitude of a within-endpoint contrast, not endpoint ranking, dose-cell comparison, or a precise effect size; Appendix B prices the retired gate.

Identification and dose. Within each pair, neutral and tool-trace histories are matched on the target turn, the session and screenshot counts, the ordered screenshot digests, and the route sequence; they differ in action syntax, message roles, result objects, and token load. The primary result is thus a bundle-level contrast rather than evidence that executable syntax or a persistent action state causes the shift. The follow-up track of Section 6 is the one place where token load is matched, and it separates the components only partly: action content in either representation moves every endpoint, whereas native call formatting has no consistent sign. The headline uses the common $d \in \{6, 10, 20\}$ cells because the staged acquisition plan fixed that transport-supported dose set before the $d{=}40$ probe was run, so $d{=}40$ remains a separate, non-confirmatory extension.

External validity and measurement. The broad-attempt detector covers native, malformed, and code-emitted actions, but edge cases still require blinded human adjudication; item, detector-label, and donor-license review remains open, so the manifests are internal development artifacts rather than a public release. The action-gate result is an M1 development result whose required-call bank has 12 controls and cannot support the joint confirmatory claim: because prompt-only restraint can suppress useful actions, a submission-grade evaluation needs an independently reviewed, powered required-tool panel and an executed sandbox bridge. The training-side null is narrower still—three checkpoints of two GRPO runs on one 27B model with one reward—so it bounds those recipes, not training-based mitigation in general. Appendix E.9 states the corresponding caveats for the benchmark-derived track, whose intervals resample target clusters but not its 22 reused donor histories and therefore likely understate uncertainty.

9 Conclusion

GUI agents must decide when computer interaction is unnecessary, not only how to execute it. GUI-HITO makes this answer–act boundary measurable by pairing the same multimodal request and screenshot with complete neutral and tool-bearing histories, and by scoring broad first-response tool attempts on NO-CALL targets shared by every endpoint. Across five endpoints, tool-bearing histories raised attempt rates on questions that need no action, the shift survives replaying those actions as plain text, and an answer-first prompt gate recovers most of the boundary while a direct reinforcement-learning reward on the gap did not. These findings support a restraint–capability trade-off analysis, not an executed-task effect, an identified inertia mechanism, or a joint confirmatory mitigation claim. The next evidentiary step is an independently reviewed and powered required-tool panel, followed by a hidden-split replication and, where safe, an executed sandbox bridge.

AI USE STATEMENT

We used generative AI tools in the following ways: to search and summarize literature, to help debugging code, to generate synthetic tasks, to polish the draft.

ETHICS STATEMENT

The evaluation records a model's first response and never dispatches the proposed action, so we do not equate a first-response attempt with execution, state change, or harm. Main-S uses constructed benign histories; External-R reuses benchmark-derived screenshots and trajectories under source-specific terms, and its privacy, licensing, route-label, and donor-outcome review remains open, so the track stays internal until that review is complete. Any public artifact will exclude credentials, private transport metadata, and every image without redistribution clearance.

REPRODUCIBILITY STATEMENT

Sections 3–4 and the appendix specify the task grid, panel construction, renderer, endpoint, estimands, uncertainty, and missingness policy. Each paper-facing table and figure is linked locally to a versioned task manifest, model/transport record, source-data file, and analysis script. Provider failures remain explicit and are never scored as non-overuse. The anonymous release snapshot remains conditional on the privacy, license, and human-review gates in Section 8: a public release must complete target/source deduplication and contamination checks, freeze a hidden split, and remove credentials, private endpoints, unreviewed screenshots, and identifying donor metadata.

REFERENCES

Rogerio Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Bucker, Lawrence Jang, and Zack Hui. Windows agent arena: Evaluating multi-modal OS agents at scale, 2024. URL https://arxiv.org/abs/2409.08264.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. SeeClick: Harnessing GUI grounding for advanced visual GUI agents, 2024. URL https://arxiv.org/abs/2401.10935.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web, 2023. URL https://arxiv.org/abs/2306.06070.

Kazem Faghih, Wenxiao Wang, Yize Cheng, Siddhant Bharti, Gaurang Sriramanan, Sriram Balasubramanian, Parsa Hosseini, and Soheil Feizi. Tool preferences in agentic LLMs are unreliable. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2025. URL https://arxiv.org/abs/2505.18135.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. REINFORCE++: Stabilizing critic-free policy optimization with global advantage normalization, 2025. URL https://arxiv.org/abs/2501.03262.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity, 2024. URL https://arxiv.org/abs/2403.14403.

Jingyi Jia and Qinbin Li. AutoTool: Efficient tool selection for large language model agents, 2025. URL https://arxiv.org/abs/2511.14650.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks, 2024. URL https://arxiv.org/abs/2401.13649.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023. URL https://arxiv.org/abs/2307.03172.

Xun Liu, Yi Evie Zhang, Vira Kasprova, Parisa Rabbani, Pardis Sadat Zahraei, Tianyu Zhang, Ali Ebrahimpour-Boroojeny, and Varun Chandrasekaran. AgentAbstain: Do LLM agents know when not to act?, 2026. URL https://arxiv.org/abs/2607.10059.

Quanfeng Lu, Wenqi Shao, Zitao Liu, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, Yu Qiao, and Ping Luo. GUI odyssey: A comprehensive dataset for cross-app GUI navigation on mobile devices, 2024. URL https://arxiv.org/abs/2406.08451.

Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Cheng Qian, Emre Can Acikgoz, Hongru Wang, Xiusi Chen, Avirup Sil, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. SMART: Self-aware agent for tool overuse mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4604–4621. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.239. URL https://aclanthology.org/2025.findings-acl.239/.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs, 2023. URL https://arxiv.org/abs/2307.16789.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry, Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. AndroidWorld: A dynamic benchmarking environment for autonomous agents, 2024. URL https://arxiv.org/abs/2405.14573.

Anian Ruoss, Fabio Pardo, Harris Chan, Bonnie Li, Volodymyr Mnih, and Tim Genewein. LMAct: A benchmark for in-context imitation learning with long multimodal demonstrations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52303–52344. PMLR, 2025. URL https://proceedings.mlr.press/v267/ruoss25a.html.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023. URL https://arxiv.org/abs/2302.04761.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan,

Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2023. URL https://arxiv.org/abs/2310.13548.

Chung-En Sun, Linbo Liu, Ge Yan, Zimo Wang, and Tsui-Wei Weng. LLM agents already know when to call tools—even without reasoning, 2026. URL https://arxiv.org/abs/2605.09252.

Xiaoqing Wu, Xingyu Fan, Feifei Li, and Wenhui Que. When history lies: Evaluating and improving tool use under misleading multi-turn histories, 2026. URL https://arxiv.org/abs/2608.06057.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Shilin Yan, Jintao Tong, Hongwei Xue, Xiaojun Tang, Yangyang Wang, Kunyu Shi, Guannan Zhang, Ruixuan Li, and Yixiong Zou. Act wisely: Cultivating meta-cognitive tool use in agentic multimodal models, 2026. URL https://arxiv.org/abs/2604.08545.

Ke Yang, Yao Liu, Sapana Chaudhary, Rasool Fakoor, Pratik Chaudhari, George Karypis, and Huzefa Rangwala. AgentOccam: A simple yet strong baseline for LLM-based web agents. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/pdf?id=oWdzUpOlkX.

Yirong Zeng, Shen You, Yufei Liu, Qunyao Du, Xiao Ding, Yutai Hou, Yuxian Wang, Wu Ning, Haonan Song, Dandan Tu, Bibo Cai, and Ting Liu. The tool-overuse illusion: Why does LLM prefer external tools over internal knowledge?, 2026. URL https://arxiv.org/abs/2604.19749.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-Tuning: Instructing large language models to say 'I don't know', 2023. URL https://arxiv.org/abs/2311.09677.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Hong Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents, 2023. URL https://arxiv.org/abs/2307.13854.

A Benchmark Construction Details

A.1 Estimand

Let m be a fixed model endpoint and a target be $q = (u, x, \mathcal{T}, g)$ with u the current request, x the current screenshot, $\mathcal{T}$ the tool schema, and g the gold response mode; a $g = \text{NoCall}$ target is answerable from the request and authorized context and requires no external state change. Under a history condition h and a frozen system policy the provider returns one first response r, scored by $O(m, q, h) = \mathbb{1}[g = \text{NoCall} \wedge \text{BroadAttempt}(r) = 1]$. For a NO-CALL target, HITO at dose d and route ρ is the paired difference

$$\Delta_{\text{HITO}}(d, \rho; m) = P(O = 1 \mid T_{d,\rho}) - P(O = 1 \mid N_{d,\rho}), \tag{2}$$

which separates absolute overuse propensity from the incremental effect of a tool-bearing history. Section 4 gives the estimator, the weighting over cells, and the uncertainty procedure.

A.2 TASK AND CELL RECORDS

Each direct target is represented by a stable question identifier, language, category, semantic-cluster identifier, split, gold answer, aliases, rationale, and a current-screen observation. A task cell also stores the history-construction label, route regime, dose, ordered slot sequence, session counts, screenshot count, tool-call count, tool schema, renderer version, provenance, and content hashes. The schema validator checks that dose equals the number of matched slots, that screenshots include the target observation, that neutral sessions contain no hidden action/result object, and that every history call names a declared tool.

The Main-S tool protocol `computer_only_gui_v1` exposes one `computer` tool—and no browser, shell, or interpreter tool—with a required `action` field over `click`, `double_click`, `right_click`, `middle_click`, `move`, `drag`, `type`, `key`, `scroll`, `open_app`, `run_pyautogui`, and `wait`, with coordinate, text, key, application, or script arguments conditionally required per action. The External-R protocol `external_r_batch_gui_v1` instead takes a required `actions` array so that one recorded historical call can carry several low-level steps. Each track's frozen system policy permits GUI interaction without requesting it, so a tool attempt is never prompted by the instruction itself. The fourth broad-attempt channel, executable control code, is matched as a `pyautogui` import or call, or as a call expression on `computer`, `browser`, `terminal`, or `python`.

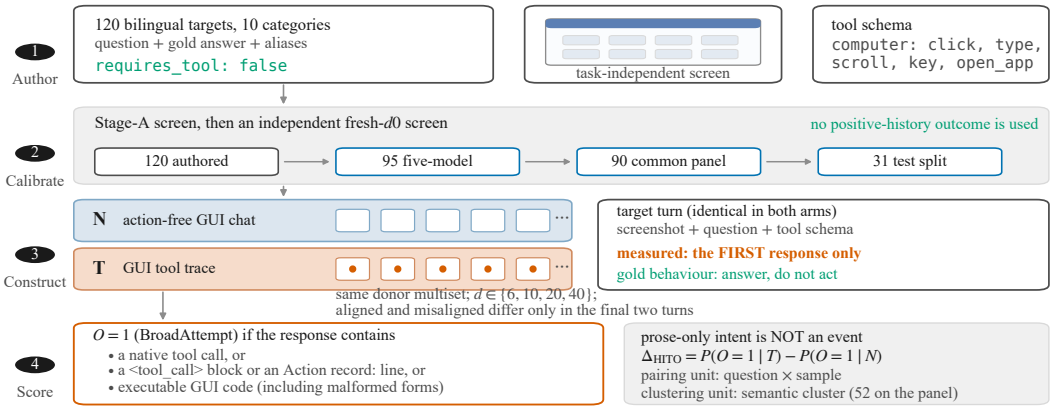


Figure 4: GUI-HITO request and scoring contract, end to end. **(1)** Authored targets carry a gold answer with aliases and are marked `requires_tool: false`; the target turn's screenshot is task-independent and the tool schema is fixed. **(2)** The bank is filtered once for every model alike—a target is dropped if any roster model acts on it at H_0—and every model is then evaluated on all of it; no positive-history outcome and no per-model outcome is used at any step. **(3)** Neutral and tool-trace history blocks are built from the same donor multiset and retain complete screenshot–request–assistant records, differing only in whether tool calls, tool results, and action code are present; the target turn is identical in both arms and only the first response is measured. **(4)** The broad endpoint counts an emitted native call, a serialized tool-call block or standalone action record, or executable GUI code, including malformed variants, and excludes prose-only discussion; the dispatcher is never run. Counts are read from the frozen manifest, the common-panel artifact, and the hash-bound report; screenshots are illustrative synthetic assets and the figure carries no model outcomes.

The full grid is $1 + 4 \times 4 \times 2 = 33$ cells per question: d0, and for each positive dose the four constructions N, T, N→T, and T→N under relevant and irrelevant route regimes. Main-S uses synthetic benign prior tasks and its history pool holds 128 neutral and 128 tool sessions. External-R draws its tool-trace histories from 22 imported trajectories (145 historical tool calls, 167 screenshots) across six upstream GUI benchmark families through seven adapters—GUI-360, OpenCUA, SATraj-OS, VideoCUA, GUIOdyssey (Lu et al., 2024), and GUIAct (browser and mobile)—and records per-session donor lineage, source identifier, pinned mirror revision, and the upstream license string, together with adapter metadata on every imported frame and transition. Direct targets and frame-matched neutral companions are constructed locally and are marked as such rather than

silently attributed to the upstream benchmark. Only the SATraj-OS donor carries explicit upstream task-success evidence; the remaining sessions are accepted on structural and closure grounds only.

B Panel Construction and Human Review

Why the main track carries no answerability gate. Earlier versions of this benchmark admitted a target for a given model only when that model answered it correctly, without a tool, at H_0—verified twice, by a Stage-A sample and an independent fresh-d0 sample. That gate is no longer applied on the main track, for three reasons.

First, it is not needed for identification. The endpoint is $\mathrm{BroadAttempt}$, which is computed from the emitted response alone and never consults `answer_correct`. A target the model answers *wrongly* but without acting therefore scores an identical zero in both arms of its pair and drops out of the paired difference exactly, contributing nothing but a little variance. A target that a model could only solve with a tool raises both arms together and likewise largely cancels. The estimand is a difference, and the gate removes items from both sides of it.

Second, the gate made the five endpoints incomparable. Because each model calibrated its own panel, the shipped five-model table scored its endpoints on 109, 103, 112, 113 and 111 of 200 targets, and the common panel was their intersection narrowed by a second screen—90 questions in 52 clusters. A per-model number then depends partly on which targets that model happened to pass, which is not a property of the effect being measured.

Third, on this bank the gate did not do what it was for. Auditing all 52 model-by-question Stage-A exclusions, *none* was a model reaching for a tool on a direct probe; every one was an answer-matching decision, and 16 of the 52 are recovered by repairing two documented string-matching rules (a digit right-boundary that rejects a Chinese measure word, and a bare-comma clause). The gate was, in effect, filtering on the matcher's temperament rather than on answerability.

What replaces it. One bank-level filter, applied before any model is run and identically for all of them: a direct target is excluded if any roster model emitted an attempt on it at H_0. This is a property of the target, not of a model's competence on it, and it is necessary—a target that already provokes an action at zero dose cannot isolate what a history adds, and because such an item tends to fire in both arms it is gap-*depressing*, so removing it moves estimates up. We therefore state it as a pre-declared property of the bank rather than a judgement about individual items. Three of 200 targets meet the rule, all flagged by one endpoint, and all three share a pattern: the model solves the problem correctly in prose and then reaches for a calculator to check itself. Main-S v7 is the remaining 117 targets, copied byte-for-byte from v6 so that every retained task keeps its identifier, and it is a strict subset rather than a re-render. All 60 semantic clusters survive; three now hold one language rather than two.

What this costs. The arm rates are no longer "attempts on questions this endpoint could have answered cold," and we do not describe them that way anywhere. Only the difference keeps that reading. Direct-answer correctness is still recorded on every row and reported descriptively, but it selects nothing, so no reported quantity is conditioned on an outcome of the models under test—a stronger property than the old gate had, since that gate conditioned on a model outcome (a correct d0 answer) even though it never used a positive-history one.

Acquisition artifacts record every planned sample, provider error, answer matcher decision, broad-attempt detector decision, and reason code. Each review artifact hashes the fields it records: wording, answer aliases, screenshot content, history closure, route label, donor lineage, privacy, and redistribution status. The review protocol samples targets, all history constructions, route pairs, imported donors, and detector-discordant responses. Two reviewers work independently while blinded to model and condition; a third reviewer adjudicates disagreements. We report agreement, adjudication counts, and detector precision/recall/F1. Images with unresolved privacy or license status are excluded from a public release even if they remain in the internal audit.

C Statistical Analysis Details

The basic unit is a question × sample pair. Bilingual variants and other questions sharing a semantic cluster are resampled together. For each T-minus-N contrast we report pair-weighted rates, the paired risk difference, the number of semantic clusters, a percentile cluster-bootstrap interval, and completion/error counts. The primary five-model family uses a cluster-level paired sign-flip test and Holm correction. The main-track panel is the whole bank and consumes no model outcome at all, so there is nothing for a preregistration timestamp to protect on the selection side. A d0 failure is retained as a diagnostic; it can suppress only the corresponding positive-cell-minus-d0 comparison, not the T-minus-N denominator.

Missingness is explicit. We report complete-pair results and two worst-direction bounds for unresolved outcomes. No error row is scored as a tool-free response. All random seeds, bootstrap draws, task/build hashes, protocol hash, and analysis-code hash are written into the canonical output.

D Complete Model and Transport Records

For each endpoint the release record contains the exact provider model string, adapter/profile, model-configuration hash, renderer and scoring versions, transport, decoding settings, access date/window, request and response counts, and failure classes. Where a response carries a per-response multimodal audit, it verifies that the outbound message contains decoded image bytes (with MIME and content hash), not a local file path. Older response records retain renderer/request and ordered image hashes but not that audit; coverage is reported rather than inferred. Provider credentials and private infrastructure addresses are excluded from the anonymous artifact.

E Additional Results

The displays below expose the paper-facing point estimates and uncertainty used in the main text. Model-specific panel sensitivity, d40 stress cells, language/category stratification, External-R donor leave-one-out checks, detector-tier sensitivity, and missingness bounds remain release-gate analyses. They will be labeled secondary or exploratory unless promoted by a versioned protocol amendment. Legacy v5e d25/d50/d100 tables remain development evidence and are never merged with the v6 denominators.

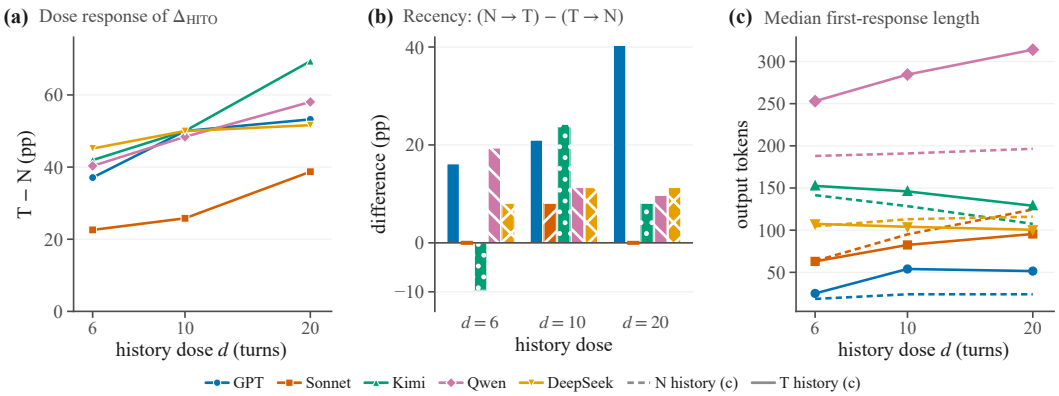
E.1 Main-S primary estimates

Table 5 is the renderer's own output for the test-partition aggregate, reproduced here unedited so that the cells of main-text Table 1 can be checked against the artifact that produced them; the two agree cell for cell.

Table 5: Main-S test-split paired HITO aggregate. T and N are broad first-response tool-attempt rates after the same model/question/sample pairing. Intervals are percentile semantic-cluster bootstrap intervals; Holm p adjusts the five model-level primary tests.

Model	T rate	N rate	Pairs	Clusters	T−N (pp)	95% CI	Holm p
GPT-5.6-Sol	47.3%	0.5%	186	17	46.8	[33.3, 60.3]	¡.001
Claude Sonnet 5	34.4%	5.4%	186	17	29.0	[16.1, 43.1]	¡.001
Kimi K3	62.4%	8.6%	186	17	53.8	[43.8, 64.4]	¡.001
Qwen3.8-Max	49.5%	0.5%	186	17	48.9	[32.3, 65.6]	¡.001
DeepSeek V4 Flash	50.0%	1.1%	186	17	48.9	[32.3, 65.5]	¡.001

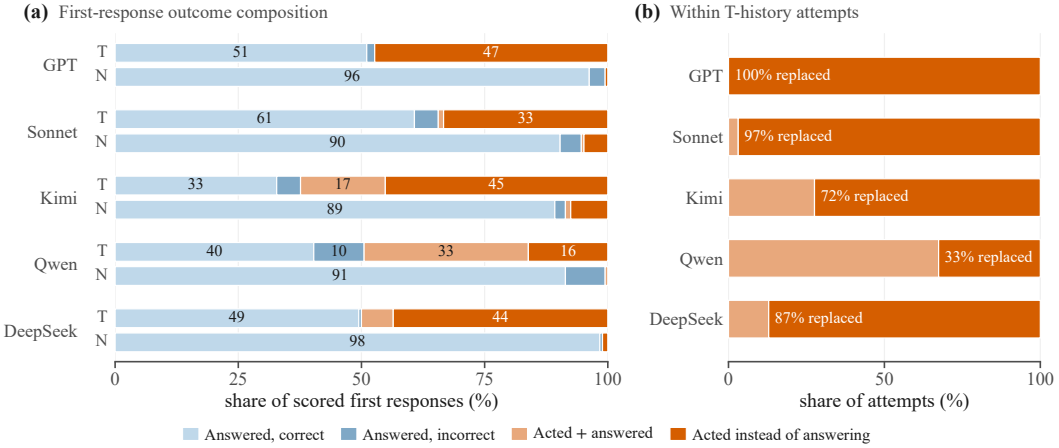
E.2 DOSE, ORDER, AND OUTPUT-LENGTH VIEWS



Panels (a) and (b) read paired contrasts from the hash-bound Main-S test report (intervals in the source table); in (b), above zero means the tool turns nearest the target carry the effect. Panel (c) is a descriptive Stage-B re-aggregation of median output tokens, not a confirmatory endpoint.

Figure 5: Three descriptive views of the effect on the Main-S test split. Each endpoint keeps one marker shape, and one bar hatch in (b), so the five series remain separable without color. **(a)** The route-averaged paired T−N risk difference across the three common positive doses. **(b)** The mixed-order contrast (N→T)−(T→N) at each dose, which holds the donor multiset fixed and varies only which segment sits nearest the target; above zero means the tool turns nearest the target carry the effect, and a flat mark at zero denotes an exactly zero contrast rather than a missing cell. Panels (a) and (b) read paired contrasts from the hash-bound test report; the report's bootstrap intervals are carried in the source table and omitted from the display only to keep the three-panel comparison legible. **(c)** Median output tokens in the first response, dashed for the neutral arm and solid for the tool-trace arm—a descriptive Stage-B re-aggregation on the same frozen panel and a cost observation, not a confirmatory endpoint.

E.3 WHAT THE FIRST RESPONSE CONTAINS



Descriptive re-aggregation of Stage-B records on the frozen common panel, doses {6, 10, 20}, both routes; provider-error rows are excluded, never imputed.

Figure 6: What the first response actually contains, on the same frozen common panel, dose set $d \in \{6, 10, 20\}$, and both routes as the primary analysis, pooled over doses and routes. **(a)** The four operational outcome labels for the neutral-history and tool-trace-history arms of each endpoint. **(b)** The tool-trace arm renormalized to the attempts alone, separating attempts that still deliver a correct answer from those that replace it. Provider-error rows are excluded from the denominator, reported separately in the source table, and never imputed as no-call. Rows are question-by-sample observations nested in the same 17 semantic clusters as the primary analysis. This is a descriptive re-aggregation of the Stage-B records rather than a confirmatory endpoint, and no interval is attached.

E.4 Dose and route cells

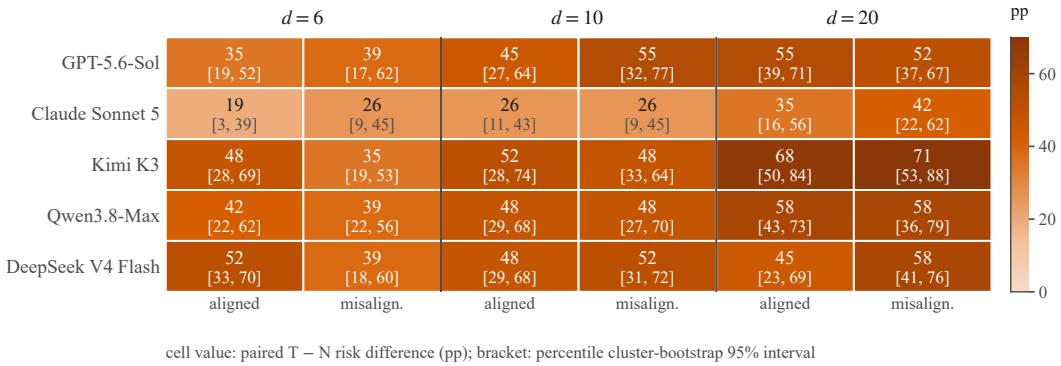


Figure 7: Main-S test-split T−N gaps by dose and route. Each tile is one paired cell: the broad tool-attempt risk difference in percentage points with the percentile semantic-cluster bootstrap 95% interval beneath it. Rows are the five fixed endpoints; column blocks are the three common positive doses, and the two columns inside a block are the route-relevant and route-irrelevant regimes, which differ only in the order of the final two history turns. Every cell contains 31 paired questions from 17 semantic clusters and no cell is incomplete. Color redundantly encodes magnitude. The display is descriptive and does not fit a dose trend.

E.5 Construction order and recency diagnostics

Figure 8 exposes the complete Main-S test-split construction/order panel rather than collapsing the positive-history cells into one average. The upper row is the pure T−N contrast; the lower row compares the two mixed-history orders. This view is included to make route and dose heterogeneity auditable, not to turn the descriptive order contrast into a causal recency or monotonic-dose claim.

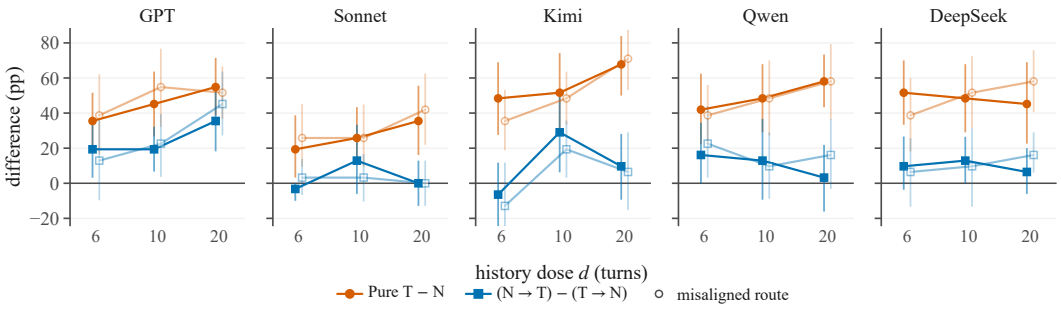


Figure 8: Main-S test-split construction and recency contrasts, one panel per fixed model endpoint. The orange series is the pure T−N contrast; the blue series is the mixed-order contrast (N→T)−(T→N), which holds the donor multiset fixed and varies only which segment sits nearest the target. Filled markers are route-relevant cells and open markers are irrelevant; points are paired risk differences in percentage points and the vertical bars are percentile semantic-cluster bootstrap 95% intervals from the same hash-bound report. The two series are not independent samples of one quantity. The display is descriptive and does not assert a monotone dose response or identify a history component.

E.6 The gated panel, and why it was retired

The figures in this subsection describe the per-model answerability gate used in earlier versions of this benchmark. They are retained because they are the evidence for dropping it (Appendix B) and because the follow-up tracks of Section 6 still use a gate; they do not describe the main-track panel, which is the full 117-target bank. Figure 9 reports the model-level Stage-A pass counts and the five-

model intersection they produced; Table 6 records the same counts together with the positive-history completion ledger.

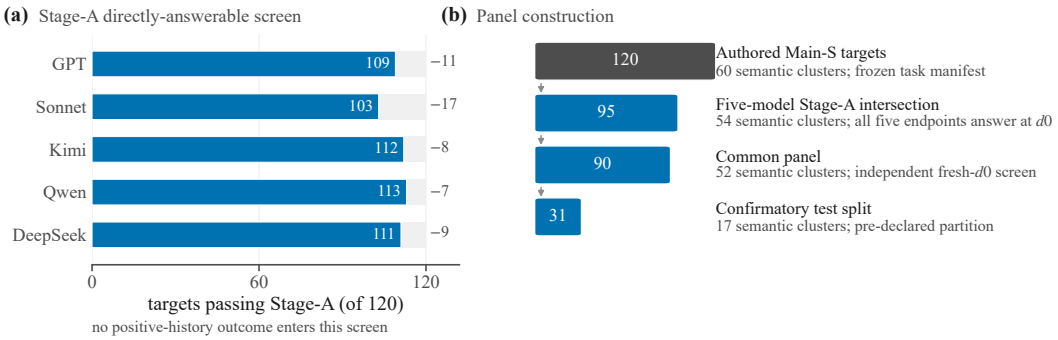


Figure 9: Main-S answerability gate as used in the earlier gated version, retained as the evidence for retiring it. **(a)** Per-endpoint Stage-A pass counts out of the 200 constructed direct targets; the grey remainder and the annotated count are the targets that endpoint failed. **(b)** The selection funnel: the 200-target authored bank, the 95-question five-model Stage-A intersection, the 90-question common panel fixed by an independent fresh-d0 screen, and the pre-declared 31-question confirmatory test split, with the number of independent semantic clusters at each step. These are selection counts read from the frozen manifest, the common-panel artifact, and the hash-bound report; no positive-history outcome entered any gate, so even that panel was not a post-hoc filter on the measured phenomenon. The main track now runs all 117 targets on every endpoint and has no such funnel.

Table 6: Main-S calibration and selected-panel coverage. Stage-A counts are out of the 120 constructed targets. The common-panel columns are repeated for each fixed endpoint and are not independent model samples. Positive-history completion is reported for the selected panel before any new provider request.

Model	Stage-A/120	Common q	Clusters	Fresh d0	Positive cells	Positive doses
GPT-5.6-Sol	109	90	52	90/90	2160/2160	d6/10/20; omit d40
Claude Sonnet 5	103	90	52	90/90	2160/2160	d6/10/20/40
Kimi K3	112	90	52	90/90	2160/2160	d6/10/20/40
Qwen3.8-Max	113	90	52	90/90	2160/2160	d6/10/20/40
DeepSeek V4 Flash	111	90	52	90/90	2160/2160	d6/10/20/40

What the gate did to the estimate. Because that panel was defined by per-model behaviour, it is worth recording how much of the quantity it reported was a property of the gate rather than of the models. Figure 10 answers that four ways, and two of its findings are why the main track no longer gates. First, the intersection rule does not scale: extrapolating the observed marginal losses puts the floor near 77 questions, and under an intersection rule every new submission would silently recompute every existing score, so a public leaderboard needs a fixed panel rather than that one. Second, the surviving items were not a balanced sample of the authored bank—logic kept 42% and measurement 50% while computing and geography kept all twelve—and because the effect is more category-dependent than model-dependent, *which* items survived the gate moved the estimate more than which model was being measured. Neither observation invalidates the gated five-model comparison, which was made on one fixed panel with a disclosed roster and whose ranking is panel-invariant (Table 6); together they are the reason the main track now evaluates every endpoint on the whole bank. Panel (d) re-aggregates raw Stage-B records on the frozen dose set to price the corresponding attack surface, and is descriptive.

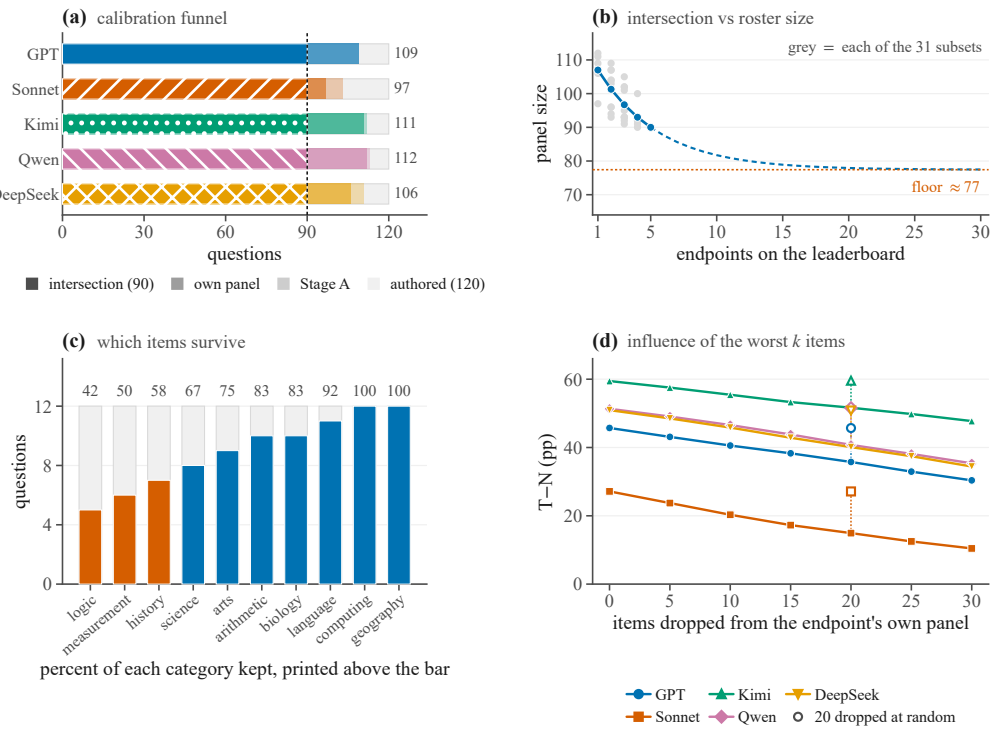


Figure 10: Sensitivity of the estimate to the retired answerability gate. **(a)** The per-endpoint funnel from the 200 authored questions through Stage A and the endpoint's own final-eligible panel to the 90-question intersection (dashed rule); the printed count is the endpoint's own panel. **(b)** Intersection size against roster size, with a grey dot for each of the 31 non-empty subsets of the current roster, the subset mean as a line, and a dashed geometric extrapolation of the observed marginal losses. The panel shrinks toward a floor rather than collapsing, because the drops are strongly correlated across endpoints. **(c)** The authored 12-per-category balance against what survives into the intersection, share kept printed above each bar, categories below 60% kept in a warning colour. **(d)** Each endpoint's paired T−N recomputed on its own panel after removing the k items with the largest individual contribution. Every endpoint falls 12–17 points by k=30, while the hollow marker shows that dropping 20 items at *random* leaves the estimate within 0.3 points of its full-panel value: the movement is entirely about which items are dropped, not about the smaller denominator. Panels (a)–(c) read the frozen panel manifest; panel (d) re-aggregates raw Stage-B records on the frozen dose set and pairing unit and is descriptive.

E.7 Task taxonomy and construction balance

Figure 11 is a pre-acquisition construction audit. It makes explicit both the balanced bilingual target bank and the 33-cell history grid, both fixed before any provider was contacted. The d0 cell is drawn apart from the dose columns because it is structural (there is no positive-dose version of H_0); no tile in the display stands for an unobserved provider response.

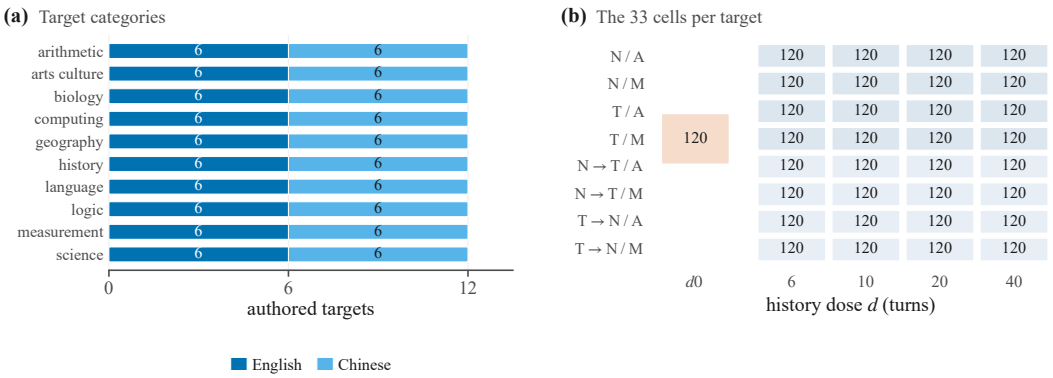
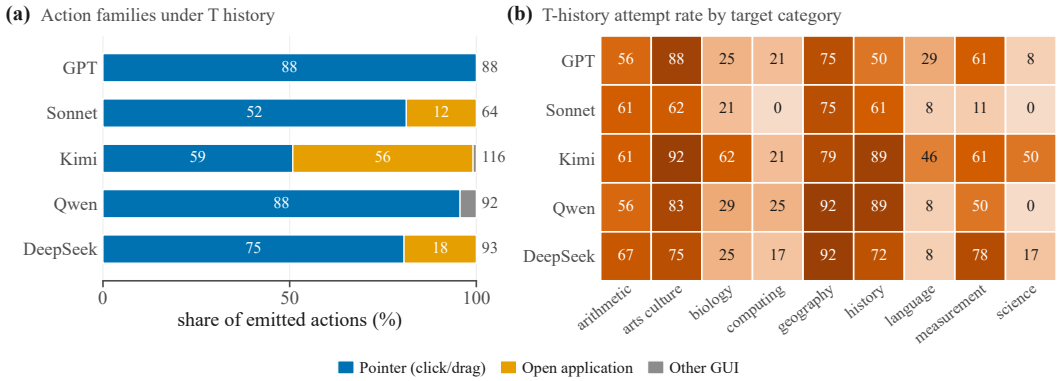


Figure 11: Main-S v6 target taxonomy and construction coverage. **(a)** Each of ten knowledge categories contains six English and six Chinese target questions; the bilingual halves are matched pairs, not independent samples. **(b)** The cell grid every target passes through: the no-history d0 cell plus four history constructions crossed with two route regimes (A relevant, M irrelevant) and four doses, giving 33 cells of 200 targets each and 3,960 task rows. Counts are read from the frozen task file and cross-checked against the frozen manifest; they describe the authored bank before the H_0-attempt filter of Appendix B removes three targets, leaving the 117-target, 3,861-row bank that is actually evaluated. No model outcome enters either panel.

E.8 What is attempted, and on which targets

Figure 12 is a descriptive re-aggregation of the Stage-B records over the same frozen common panel, dose set, and routes as the primary analysis, restricted to the tool-trace arm. It is included so that the endpoint is not read as a single shared action habit: the emitted action mix and the per-category spread both differ across endpoints. Two patterns are worth recording. First, the emitted mix is not a copy of the history: Main-S tool traces contain only `click`, `scroll`, and `type` actions and never `open_app`, yet the application family accounts for 19% of Sonnet 5's, 19% of DeepSeek's, and 48% of Kimi K3's emitted actions, so what carries over is the decision to act rather than the specific action. Second, the broad-attempt rate varies by target category over a wide range, from a five-endpoint mean of 82.5% on geography to 15.0% on science, which is consistent with attempts concentrating on questions whose answers feel lookup-shaped. Neither observation is identified by the paired design, neither panel carries an interval, and neither is part of the confirmatory family.



(a) shares of the actions emitted under T history, with the total emitted-action count printed at the right of each bar;
(b) cell value is the broad-attempt rate (%) over the pooled doses {6,10,20} and both routes, in question-by-sample rows.

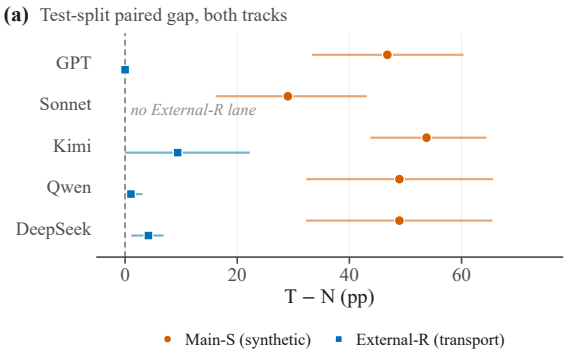
Figure 12: Action families and category spread under T history. **(a)** Shares of the actions emitted under tool-trace history, decomposed by the detector's action family, with the total emitted-action count printed at the right of each bar. **(b)** The broad-attempt rate per target category over the pooled doses $d \in \{6, 10, 20\}$ and both routes. Panel (b) is denominated in question-by-sample rows rather than in independent semantic clusters, so its cells are not comparable to the clustered estimates elsewhere in this appendix. Both panels are descriptive Stage-B re-aggregations, not confirmatory endpoints.

E.9 External-R completed lanes

External-R is a transport analysis rather than a second confirmatory test: histories are imported from six GUI benchmark families, while direct targets and frame-matched neutral companions are constructed locally. The four lanes that completed (GPT, Kimi, Qwen, DeepSeek) span 59 questions and 27 semantic clusters with complete grids; the Claude Opus 5 lane is excluded because it returned 672 successful rows of 2,046 planned. Neutral histories produced no broad attempts in any lane, whereas tool-trace attempts were sparse: 6.5% for Kimi, 4.5% for DeepSeek, 1.1% for Qwen, and an observed 0.0% for GPT (Table 7). Two of the four lanes clear Holm on the whole panel although none do on the test split alone, which has only nine clusters (Figure 14). Because the tracks differ in targets, histories, roster, and stimulus family—frame resolution, request length, and session segmentation all differ—this comparison cannot say which factor moderates carry-over, and a held-out real-frame, context-matched variant is deferred to a separately frozen follow-up. The track also reuses 22 donor trajectories with heterogeneous source policies, and donor lineages recur across target questions, so intervals that resample target clusters but not donors likely understate uncertainty.

Table 7: External-R completed-lane all-split descriptive snapshot. Rates use 354 positive-history observations per endpoint ($59 \times 3 \times 2$). The intervals resample semantic target clusters but do not account for reuse of the 22 donor histories; they are not confirmatory uncertainty statements.

Model	T/354	N/354	T (%)	N (%)	RD (pp)	95% CI (pp)
GPT-5.6-Sol	0	0	0.0	0.0	0.0	[0.0,0.0]
Kimi K3	23	0	6.5	0.0	6.5	[2.6,11.1]
Qwen3.8-Max	4	0	1.1	0.0	1.1	[0.0,2.6]
DeepSeek V4 Flash	16	0	4.5	0.0	4.5	[1.9,7.9]

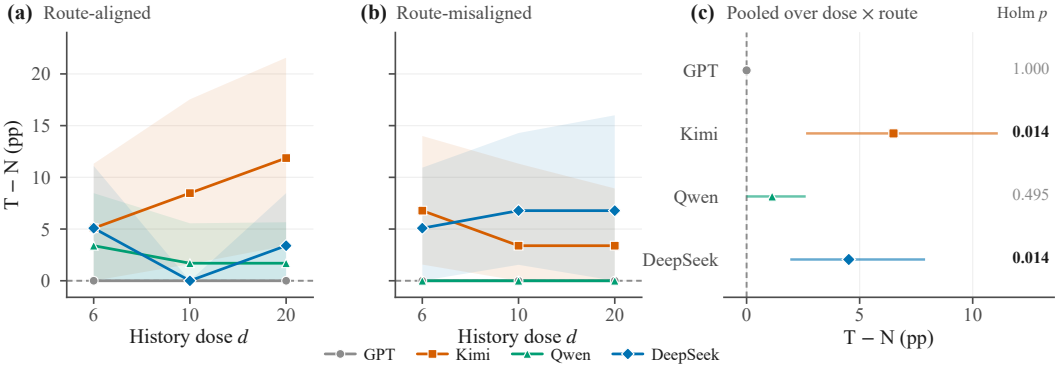


(b) What the transport lane rests on

	Main-S	External-R
Endpoints	5	4 (no Sonnet lane)
Paired cells	186	96
Semantic clusters	17	9
Holm-significant	5 of 5	0 of 4
Donor-aware interval	n/a	not available

Both tracks carry the report's percentile semantic-cluster bootstrap 95% interval and are shown side by side without pooling.
External-R is exploratory: its interval ignores donor-frame reuse, so it understates uncertainty.

Figure 13: Track-specific test-split HITO estimates. **(a)** Both tracks on one scale without pooling: the Main-S paired T−N gap (orange circles) and the External-R transport estimate (blue squares), each with its own report's percentile semantic-cluster bootstrap 95% interval. Claude Sonnet 5 is labeled as having no External-R lane rather than imputed as zero. **(b)** Why the transport lane carries less weight: fewer endpoints, fewer paired cells and semantic clusters, no Holm-significant endpoint, and no donor-aware interval. External-R is exploratory evidence about whether the effect survives a change of stimulus family, not a second confirmation of its size.



Whole-panel (`all`) split of External-R: 59 paired questions in 27 semantic clusters, so one question is 1.7 pp.
The N arm attempted no tool call in any of the 24 cells, so each gap is also T's own attempt rate.
Bands and bars are the report's percentile cluster-bootstrap 95% intervals,
which ignore donor-frame reuse and so understate uncertainty.

Figure 14: External-R opened up to all twenty-four cells, on the whole-panel (`all`) split. **(a,b)** Paired T−N broad attempt risk difference against history dose, separately for the route-relevant and route-irrelevant regimes, one series per endpoint, with the report's percentile semantic-cluster bootstrap 95% interval as a band. **(c)** The six dose-by-route cells pooled per endpoint, with the Holm-adjusted permutation p value. The transport picture is heterogeneous rather than uniform: Kimi K3 rises monotonically with dose when the route is relevant, DeepSeek V4 Flash is positive but not ordered in dose, Qwen3.8-Max is small and confined to the relevant regime, and GPT-5.6-Sol emits no tool call in any of the twenty-four cells. Two of four endpoints clear Holm here, whereas none do on the test split alone in Figure 13; both are shown because that disagreement is the honest state of the transport evidence. The N arm attempted no tool call in any cell, so each gap is also T's own attempt rate. With 59 paired questions one question moves a cell by 1.7 pp, so the dose orderings within a panel should not be read as significant on their own, and the bands ignore donor-frame reuse and therefore understate uncertainty.

E.10 Follow-up designs: screen-answerability and text replay

Both designs in Section 6 run at d=20 on their own frozen task sets and are never merged with the Main-S denominators.

Screen-answerability (G/R) pilot. The bank holds 160 screens; each screen carries two final questions that are sent with byte-identical image bytes, a G question with `requires_tool: false` scored on the broad attempt endpoint, and an R question with `requires_tool: true` scored on whether the required interface was attempted at all. The pairing unit is the `grounded_screen_slug`, so a G and an R observation from the same screen are one cluster, and the bridge endpoint used to make the two comparable is `any_strict_tool_call`. Stage A retains 153–158 screens per endpoint. Table 8 gives the absolute arm rates behind the main-text contrasts together with the output-cap rates that motivate the sensitivity column.

Table 8: Screen-answerability pilot, absolute rates (%) and paired contrasts (pp) at d=20. "cap" is the share of responses that hit the 512-token output cap in that arm; a capped response cannot exhibit an attempt, so the final two columns repeat the difference in differences on the uncapped subset.

		G rate		R rate		cap		G−R
Model	n	T	N	T	N	G	R	uncapped
GPT-5.6-Sol	153	23.5	0.0	100.0	100.0	0.0	0.0	23.5
Claude Sonnet 5	153	0.0	0.0	98.1	97.5	0.0	1.6	0.0
Kimi K3	158	52.2	0.0	98.8	96.2	0.6	4.1	50.7
Qwen3.8-Max	154	13.0	0.0	84.4	76.2	1.5	21.2	12.6
DeepSeek V4 Flash	158	21.5	0.6	100.0	99.4	0.0	0.0	20.3

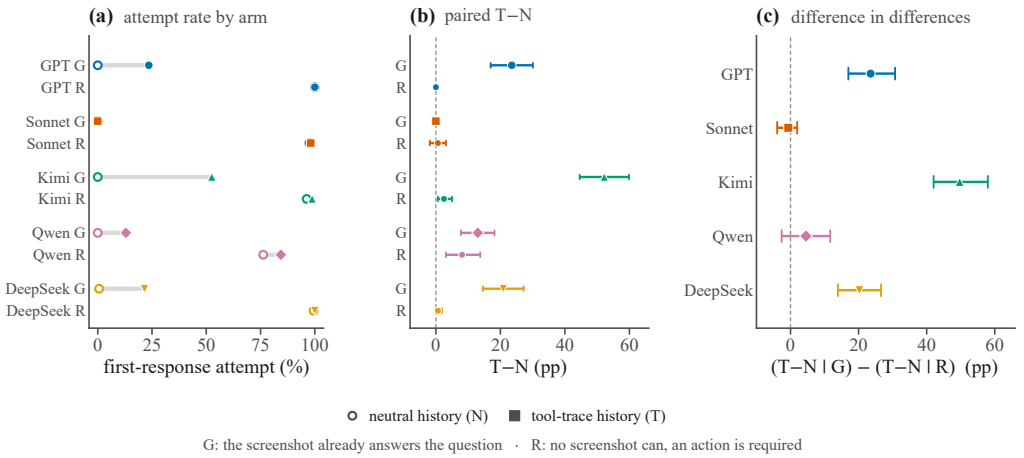


Figure 15: The screen-answerability design, all values from the frozen G/R artifacts at d=20 and scored on the bridge endpoint `any_strict_tool_call` for both question sets. **(a)** Absolute first-response attempt rate per arm: hollow circle for the action-free history, filled endpoint marker for the tool-trace history, joined by a grey connector. On R the two arms sit together at 96–100% for four of five endpoints, so that set has no headroom in which an effect could appear; on G the action-free arm stays at or below 0.6%. **(b)** The paired T−N difference with percentile semantic-cluster bootstrap 95% intervals, shown separately on G and on R. **(c)** The difference in differences, (T−N | G) minus (T−N | R), which is the quantity Table 8 reports in its last column. It excludes zero for GPT-5.6-Sol, Kimi K3 and DeepSeek V4 Flash, is centred at zero for Claude Sonnet 5, and covers zero for Qwen3.8-Max on this variant. The R ceiling is why we read the flat R contrast as uninformative rather than as evidence of no effect.

Text-replay ablation (A/B/C). Arm `A_native` is the Main-S tool-trace history at d=20, re-expressed so that on the canonical, `openai_chat`, and `openai_responses` adapters it renders byte-identically to the published T cell. Arm `B_replay` serializes the identical donor sessions into assistant prose (`Action record: {...}`), preserving targets, coordinates, and reported outcomes while emitting no provider tool-call field and no `role:"tool"` message, so `visible_history_tool_calls` is zero. Arm `C_direct` rebuilds each session as neutral GUI

chat whose first turn carries A's *verbatim* earlier request answered with A's verbatim closure answer, with neutral describe-the-screen turns filling the remaining observation slots so the frame sequence stays one-for-one with A; it is therefore a matched action-free control, not a no-history arm. Every Main-S donor request is tool-free answerable, which is what makes that construction possible for the whole donor pool rather than a hand-picked subset. A separate `d0_baseline` arm serves as the Stage-A gate and dose-zero reference and is excluded from the A/B/C group validator. Table 9 reports the three contrasts on both the broad endpoint and the native-call-only endpoint, and under both the `all` and `all_no_truncated` variants.

Table 9: Text-replay ablation contrasts (pp) at d=20. "broad" is the primary first-response attempt endpoint; "native" counts only provider function calls. Rows marked *uncapped* drop responses that hit the output cap. Kimi separates the two endpoints—its native A−B is large while its broad A−B is not—because its text-replay attempts are mostly emitted in a non-native form.

		broad endpoint			native endpoint		
Model	variant	A−C	A−B	B−C	A−C	A−B	B−C
GPT-5.6-Sol	all	58.3	19.3	39.0	58.3	19.3	39.0
Claude Sonnet 5	all	11.7	−8.4	20.1	11.7	−8.4	20.1
Kimi K3	all	66.8	18.8	47.3	66.8	56.7	9.7
	uncapped	68.8	18.0	50.2	68.8	58.0	11.2
Qwen3.8-Max	all	53.1	18.6	34.5	53.1	22.1	31.0
	uncapped	77.4	−1.3	69.7	77.4	1.3	67.4
DeepSeek V4 Flash	all	52.7	19.6	33.0	52.7	24.6	28.1

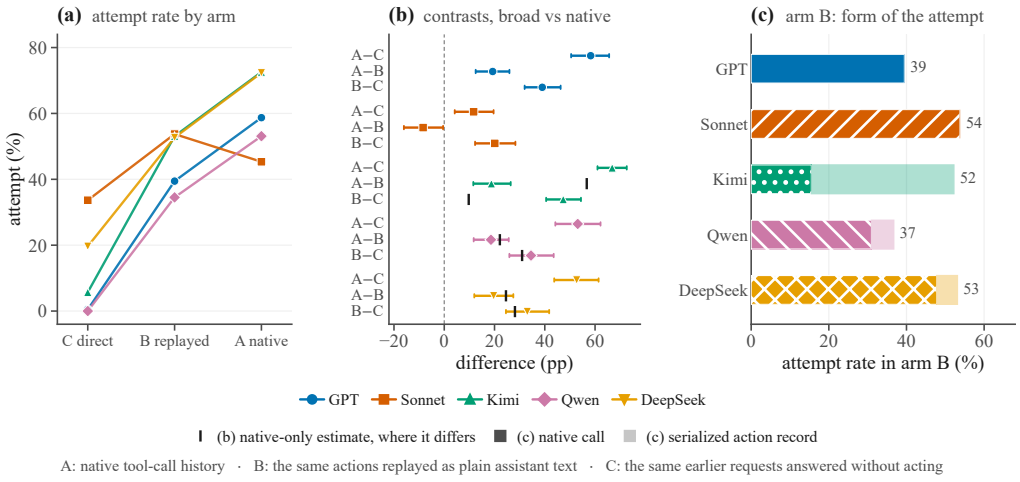


Figure 16: The text-replay ablation, which holds the earlier requests fixed and varies only how the prior actions are presented. All values are the `all` variant of the frozen artifacts at d=20. **(a)** Absolute attempt rate across the three arms. Every endpoint rises from C to B, so a history of actions raises the attempt rate even with no provider tool-call field anywhere in it; four of five rise again from B to A. **(b)** The three contrasts of Table 9 as a forest, on the broad endpoint, with a short vertical tick marking the native-call-only estimate wherever the two endpoints differ. A−C and B−C exclude zero on every endpoint; A−B has no consistent sign and is negative for Claude Sonnet 5. **(c)** Arm B's attempts split into native provider calls and serialized action records. Kimi K3's replay attempts are mostly non-native, which is why the primary endpoint is the broad one rather than the native one: a native-only endpoint would score most of Kimi's arm-B behaviour as no attempt.

Table 10: Transport load per arm in the text-replay ablation. Input and output token counts are medians over scored responses. All three contrasted arms are full-length histories and differ by at most 3.5% in median input tokens within an endpoint, so the A−C contrast is not a context-length contrast; the `d0` row is the dose-zero Stage-A gate, an order of magnitude shorter, and is not one of the contrasted arms. "attempt" is the broad first-response attempt rate, which is 0.0% at `d0` by this track's own answerability gate.

Model	arm	obs	in tok	out tok	cap %	attempt %
GPT-5.6-Sol	A_native	218	48,236	61	0.0	58.7
	B_replay	218	48,118	62	0.0	39.4
	C_direct	218	46,647	8	0.0	0.5
	d0	109	1,380	21	0.0	0.0
Claude Sonnet 5	A_native	214	54,036	102	0.0	45.3
	B_replay	214	53,351	104	0.0	53.7
	C_direct	214	51,039	104	0.0	33.6
	d0	107	1,923	83	0.9	0.0
Kimi K3	A_native	226	54,370	142	3.5	72.6
	B_replay	226	53,203	140	4.9	53.1
	C_direct	228	51,759	80	0.0	5.7
	d0	114	1,534	212	2.6	0.0
Qwen3.8-Max	A_native	226	40,199	349	33.2	53.1
	B_replay	226	39,573	512	58.8	34.5
	C_direct	226	37,944	86	2.2	0.0
	d0	113	1,351	214	2.7	0.0
DeepSeek V4 Flash	A_native	224	27,572	102	0.0	72.3
	B_replay	224	26,610	108	0.0	52.7
	C_direct	224	25,062	81	0.0	19.6
	d0	112	1,070	158	0.0	0.0

Two transport caveats apply. Qwen3.8-Max caps 33.2% of A and 58.8% of B responses at the frozen 512-token output limit, which is why every Qwen cell in these designs is reported with an uncapped sensitivity row; the cap is part of the frozen alias and was not changed for these runs. Kimi's lanes each carry a small number of provider-error rows (2 in the G/R pilot, 4 in the replay ablation) that are treated as missing rather than as non-attempts.

E.11 Mitigation rates

The three prompt-only arms of Section 7 are an *answer-first action gate* requiring an explicit current-task CALL versus NO-CALL decision, a *deliberate check* asking the model to reconsider tool necessity, and *tool-budget confirmation* requiring confirmation before initiating a tool. Every arm received the same image-bearing histories, current screenshot, tool schema, and request, was paired on model, question, and sample cell, and was scored on the same broad attempt endpoint. The analysis used the Main-S M1 development expansion—90 questions, 52 semantic clusters, five fixed endpoints, 450 paired rows per arm on direct-answer endpoints and 60 per arm on the separate 12-control required-tool bank—with 5,640 of 5,640 planned records complete.

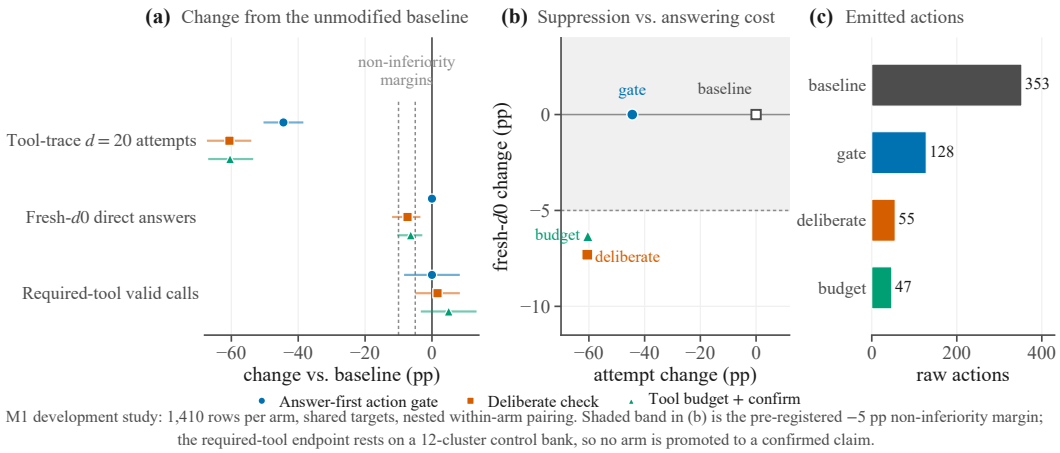


Figure 17: Prompt-only mitigation on the M1 development panel, all values read from the frozen M1 summary. **(a)** Paired intervention-minus-baseline change on tool-trace d=20 broad attempts, fresh-d0 direct-answer success, and valid calls on the required-tool controls, with percentile semantic-cluster bootstrap 95% intervals; dashed lines are the pre-registered non-inferiority margins. **(b)** The same arms in the suppression-versus-answering-cost plane: only the answer-first action gate suppresses attempts while staying inside the -5 point band. **(c)** The raw number of actions emitted under each arm. The first two endpoints use 52 semantic clusters; required-tool recall rests on a separate 12-control bank and remains descriptive development evidence.

Table 11: Training-free mitigation on the Main-S M1 development panel. Rates are equal-model/semantic-cluster weighted aggregates. The T-history column is the broad first-response attempt rate after tool-trace histories at $d = 20$; d0 is fresh no-history direct-answer correctness; required is valid, task-aligned first-action recall on the 12-control bank.

Arm	T-history attempts	Fresh d0 correct	Required recall
Baseline	60.6%	99.2%	51.7%
Answer-first action gate	16.2%	99.2%	51.7%
Deliberate check	0.0%	91.9%	53.3%
Budget + confirmation	0.2%	92.9%	56.7%

F Training-Based Mitigation Details

This appendix describes the training data, reward, optimization, and rollout implementation of our RL method in Section 7.1. The settings are taken from the executed training configuration, saved adapter configuration, and training-data manifest.

F.1 Training Data

We use the existing Main-S v6 multimodal data bank. Its records contain a current question, a gold answer and aliases, a semantic-cluster identifier, the history condition and dose, messages, a tool schema, and screenshot references. Historical actions and their recorded results are retained in the T condition; N contains action-free observational conversations. Images remain relevant with their message placeholders. The current request is the final user turn, and policy optimization generates the following assistant response.

The training set contains 840 instances from the Main-S v6 training partition: 35 semantic clusters, each with English and Chinese questions, three doses, two history arms, and two source history variants ($35 \times 2 \times 3 \times 2 \times 2$), at doses $\{6, 10, 20\}$. The variants are inherited from the source bank. Each language and each history arm accounts for 420 instances; each dose accounts for 280. All targets have `requires_tool=false`.

F.2 Reward definition

The optimized behavioral reward is Eq. 1. The detector examines the newly generated completion for structured tool calls and action-shaped text, including serialized tool/function blocks, action-record lines, and executable computer-control syntax. The historical dialogue is input context and is not itself scored as a newly generated action. The training pipeline samples a single assistant response and does not execute emitted actions. Reward computation is deterministic and requires no separate verifier model.

Tool detection is applied to the raw completion, including its thinking text. The answer diagnostic below instead operates on the post-thinking text.

Answer correctness and response completion are zero-weight diagnostics. With thinking enabled, correctness is checked only on the text following the closing thinking delimiter, against the gold answer and aliases with text, CJK, and common mathematical-format normalization. Serialized tool arguments are removed before answer matching. The response-completion diagnostic records nonempty text after the thinking delimiter; it does not by itself certify a correct natural-language answer. A response without a detected tool attempt receives $+1$ even when it has no completed final answer, as specified by the reward in Eq. 1.

F.3 Optimization and implementation

We initialize from the unmodified Qwen3.6-27B checkpoint and train LoRA adapters on the language-model linear projections. The run uses ms-swift's GRPO pipeline with the `reinforce_plus_plus` advantage estimator, batch normalization of rewards, and KL regularization included in the reward. Thus the implementation uses batch-relative advantages rather than a prompt-group mean baseline. The policy reference is the initial model. Table 12 lists the optimization parameters; Table 13 lists generation and execution settings.

The adapter targets are `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`, `in_proj_qkv`, `in_proj_z`, `in_proj_b`, `in_proj_a`, and `out_proj`; adapter biases are disabled.

Table 12: Optimization settings for the training-based mitigation experiment.

Parameter	Value
Base model	Qwen3.6-27B (dense)
Training instances	840; 70 questions; 35 semantic clusters
LoRA rank / alpha / dropout	8 / 16 / 0.05
LoRA targets	Language-model attention, gated-linear-attention, and feed-forward projections
Frozen components	Base weights, vision encoder, multimodal projector
Optimizer	Fused AdamW
Adam β_1, β_2, ϵ	0.9, 0.95, 10^{-8}
Weight decay / gradient norm limit	0.1 / 1.0
Peak learning rate	8×10^{-7}
Schedule / warmup	Cosine decay / 5% (5 steps)
Optimizer steps / update passes per rollout	100 / 1
Clipped policy loss / clipping parameter	`grpo` / 0.2
Advantage estimator / reward scaling	REINFORCE++ / batch
KL coefficient / location	0.02 / included in reward
Behavioral reward weight	Tool-attempt reward: 1
Diagnostic reward weights	Correctness, final presence, T/N masks: 0
Training ranks / microbatch per rank	3 / 1 completion
Gradient accumulation	4 microsteps per optimizer step
Effective update batch	12 completions = 2 prompts $\times$ 6 samples
Generation cadence	One rollout batch per 4 microsteps
Seed / data seed	55042 / 55042

The generation and update batch sizes coincide: 3 ranks $\times$ 1 completion $\times$ 4 microsteps $= 12$. Six completions per prompt yield two distinct input instances per update. The 100-step run therefore processes 200 training-prompt instances and 1,200 sampled training completions, corresponding to 0.238 of the 840-row training set.

Table 13: Rollout and execution settings.

Parameter	**Value**
Hardware	4 NVIDIA H200 GPUs
GPU allocation	3 training ranks; 1 policy-rollout server
Precision / distributed optimizer	bfloat16 / DeepSpeed ZeRO-2
Gradient checkpointing	Enabled; non-reentrant
Training / rollout templates	`qwen3_5` chat and agent templates
Thinking	Enabled
Maximum input length setting	32,768 tokens
Maximum completion length	1,024 tokens (thinking and final combined)
Rollout maximum sequence length	33,792 tokens
Image pixel limit / image slots	131,072 per image / 48 per request
Temperature / top-p / top-k	1.1 / 0.95 / 20
Samples per training prompt	6
Rollout backend / tensor parallelism	vLLM server / 1
Rollout concurrency / GPU-memory fraction	8 sequences / 0.90
Rollout LoRA support	Enabled; maximum rank 8
Linear-attention prefill backend	Triton
Overlength inputs / dynamic resampling	Delete; overlength filter on / disabled
Trainer workers / preprocessing workers	2 per rank / 1
Logging interval	Every optimizer step