PLACEBO AUDITS QUANTIFY AUDIT-SPECIFIC ACCEPTANCE UNDER ADAPTIVE REUSE OF VERIFIER TESTS

Anonymous authors
Paper under double-blind review

ABSTRACT

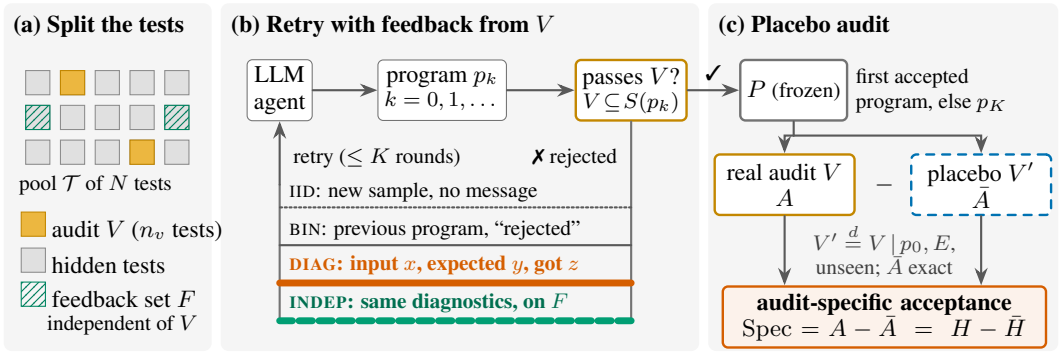
Coding agents are often given diagnostic feedback (failing inputs, expected and actual outputs) by the same small test suite that later decides whether their code is accepted. Holdout-reuse theory warns that this lets an agent fit the audit instead of the task. We make this measurable and show how to avoid it. A *placebo audit* re-evaluates each frozen output on a fresh audit drawn from the same conditional law as the real one. The gap between the two acceptance rates, *audit-specific acceptance*, equals the audit-specific part of false acceptance, is zero for any policy whose output is independent of the audit given the first attempt, is bounded by the information the feedback carries about the audit, and has a closed-form unbiased estimator. The same argument gives a remedy: diagnostics computed on a feedback set drawn independently of the acceptance audit provably add no audit-specific acceptance after one round. On 540 EvalPlus tasks, with four models from three families (8B to 32B), four audit sizes, up to 8 feedback rounds and eleven full runs, ten of them preregistered, the placebo meets its calibration prediction (i.i.d. resampling and binary feedback: $+0.2$ and $+0.1$ points after one round). One round of diagnostic feedback yields $+2.9$ $[+1.8, +4.1]$ points of audit-specific acceptance for Qwen3-8B (four runs) and a positive estimate for every other model, including $+2.6$ $[+0.8, +4.5]$ in a preregistered test at 32B; the point estimates assign roughly half of diagnostics' excess false acceptance to it. After eight retries, i.i.d. selection over nine candidates reaches a similar level. Independent feedback audits kept a fraction 0.74 [0.54, 0.95] of the diagnostic gain in useful acceptance, with audit-specific acceptance $+0.1$ $[-1.0, +1.3]$. On 444 LiveCodeBench problems, one-round audit-specific acceptance is $+2.6$ $[+1.0, +4.3]$, and the preregistered precision contrast is inconclusive. We also report a preregistered replication failure of our own exploratory headline.

1 INTRODUCTION

LLM coding agents improve programs by iterating against tests (Chen et al., 2024; Shinn et al., 2023; Madaan et al., 2023). In deployment the tests that produce feedback are often the same tests that decide acceptance: a change is merged when CI passes, and a submission is scored on the examples it was debugged against. This is *adaptive reuse of a holdout*. Adaptive data analysis shows that such reuse can overfit a holdout far faster than a fixed hypothesis would (Dwork et al., 2015b;a; Blum & Hardt, 2015), and agents are known to special-case tests (Skalse et al., 2022; Gao et al., 2023; Baker et al., 2025; Zhong et al., 2025). Diagnostic feedback ("input x: expected y, got z") looks like exactly the high-bandwidth channel that would let an agent fit a small audit.

Measuring this is harder than it looks. With diagnostic feedback, more programs pass the audit, and more of them wrongly: they fail hidden tests. This rise in false acceptances mixes three things: *selection* (any retry-until-pass loop accepts some programs that are wrong elsewhere), *transferable change* (feedback alters which programs are produced in ways that would also show up on any other audit) and *audit dependence* (programs that pass *this* audit because it was this audit). Selection and feedback are both routes to audit dependence, which is the overfitting that holdout-reuse theory is about; our estimand measures it operationally, whatever the route. Comparing with i.i.d. resampling

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.



Decomposition: $\Delta H = \Delta\text{Spec} + \Delta\bar{H}$ (audit-specific + transferable false acceptance). **Spec $= 0$ by design:** IID and BIN at $K{=}1$; INDEP at $K{=}1$; a second audit stage W drawn from the hidden tests and never shown, at any K.

Figure 1: Placebo audits. (a) Each task's test pool is split at random into an audit V of n_v tests and hidden tests. (b) After a rejected first attempt p_0, the agent retries for up to K rounds; the policies differ only in what they learn from V (line width indicates how much the message can reveal about V). INDEP gives the same diagnostics as DIAG, but on a feedback set F drawn independently of V (§4). (c) The frozen output P is scored on the real audit and, in closed form, on a placebo audit V' with the same conditional law that the agent never saw. The gap Spec is the audit-specific part of false acceptance.

controls for selection only at matched horizons, and aggregate precision cannot separate audit fitting from transferable change (cf. Roelofs et al., 2019).

Contributions.

- **Placebo audits (§3).** For each frozen output we compute, in closed form, its acceptance probability on a fresh audit with the same conditional law as the real one. The difference, *audit-specific acceptance* Spec, equals the audit-specific part of false acceptance; is exactly zero for policies whose output is independent of the audit given the first attempt; is bounded by $\sqrt{I/2}$, where I is the information the feedback transcript carries about the audit; and yields an exact decomposition of any policy's excess false acceptance (Proposition 1). The estimator is unbiased and has smaller variance than a physical placebo draw by an explicit amount (Proposition 2).
- **A remedy with a guarantee (§4).** Diagnostics computed on a feedback set drawn independently of the acceptance audit have zero audit-specific acceptance after one round, for any agent (Corollary 1); a preregistered run shows that they keep most of the utility of diagnostics. A second audit stage that the agent never saw has the same guarantee at every horizon, and we compare both at equal test budgets.
- **A preregistered, multi-model measurement (§6–7).** 540 EvalPlus tasks (Liu et al., 2023), Qwen3-8B/32B, MiniCPM4.1-8B and Llama-3.1-8B, audits of $n_v \in \{1, 2, 4, 8\}$ visible tests, and up to 8 rounds of diagnostic feedback, binary feedback or i.i.d. resampling with selection, in eleven full runs with independent audit draws and seeds, plus one run on 444 LiveCodeBench problems. We report every preregistered outcome, including failures.
- **Findings.** (i) The placebo is calibrated: policies that theory says cannot fit the audit show Spec ≈ 0 in every run. (ii) One round of diagnostic feedback yields positive audit-specific acceptance for every model, about half of the excess false acceptance over i.i.d. resampling in the point estimates. (iii) Audit-specific acceptance grows with the number of retries for every policy; by four retries, i.i.d. selection matches diagnostics. (iv) Diagnostic feedback's main effect is useful acceptance (+19 points at $K{=}1$ for Qwen3-8B); the false-acceptance share of accepted programs depends on audit size more visibly than on the feedback type. (v) Our exploratory single-run headline (+19.5 points at $n_v{=}1$, $K{=}8$) did not replicate at its preregistered criterion.

2 RELATED WORK

Adaptive data analysis and test-set reuse. Worst-case analyses show that adaptive holdout queries can overfit quickly, motivating noise, thresholding and information-based controls (Dwork et al., 2015b;a; Blum & Hardt, 2015; Russo & Zou, 2016; Xu & Raginsky, 2017). Empirically, reuse of ML test sets causes little overfitting (Recht et al., 2019; Roelofs et al., 2019), attributed to model similarity and many-class structure (Mania et al., 2019; Feldman et al., 2019); selective disclosure on LLM leaderboards is a newer form of reuse (Singh et al., 2025). Placebo audits measure adaptive fitting per output, and independent feedback audits are a protocol-level control in the spirit of holdout reuse mechanisms.

Feedback-driven code generation and weak tests. Self-debugging and reflection repair code from execution feedback (Chen et al., 2024; Shinn et al., 2023; Madaan et al., 2023), and sample-and-filter methods select by visible or generated tests (Li et al., 2022; Chen et al., 2023). Small test suites overestimate correctness (Liu et al., 2023), and patches accepted on SWE-bench often fail stronger tests (Yu et al., 2025; Wang et al., 2026; Yu et al., 2026). These works strengthen the tests; we separate how much of the overestimate a feedback loop adds through audit fitting.

Reward hacking. Optimising a proxy degrades the target (Manheim & Garrabrant, 2018; Gao et al., 2023; Skalse et al., 2022). Frontier agents special-case or tamper with tests (Denison et al., 2024; Baker et al., 2025; Zhong et al., 2025), reward hacking in coding RL can generalise to misalignment (MacDiarmid et al., 2025), and capped randomized tests can flag it (Lodkaew et al., 2026). Placebo audits quantify the *statistical* channel on ordinary tasks, without judging intent.

3 FRAMEWORK

Setting. A task has a pool $\mathcal{T}$ of N tests. The audit $V \subset \mathcal{T}$ is a uniformly random n_v-subset, and the remaining tests are hidden. Program p has pass set $S(p) \subseteq \mathcal{T}$. The audit accepts p iff $V \subseteq S(p)$, and p is *correct* iff $S(p) = \mathcal{T}$ (correct with respect to the finite pool). All policies start from the same first attempt p_0. We analyse the *activation event* $E = \{V \not\subseteq S(p_0)\}$, i.e. that p_0 was rejected. K counts the retries after p_0, so a policy with horizon K has at most $K{+}1$ candidates. It outputs P, the first accepted candidate or else its last one; on E, at $K{=}1$ this is the single retry p_1 for every policy. IID draws the retries afresh, independently of everything else. BIN shows the model its previous program and "rejected". DIAG also shows, for each failing visible test, its input, the expected output and the program's actual output. The policy's private randomness (sampling seeds) is a variable $Z \perp\!\!\!\perp (p_0, V)$. Logarithms are natural.

Definition 1 (Acceptance quantities)**.** Conditional on E, the acceptance rate, false-acceptance rate, useful-acceptance rate and false-acceptance share of accepted programs are

$$A = \Pr[V \subseteq S(P)], \qquad H = \Pr[V \subseteq S(P),\ S(P) \neq \mathcal{T}], \qquad U = A - H, \qquad R = H/A.$$

Definition 2 (Placebo audit)**.** Given (p_0, E), let V' be drawn independently of (P, V, Z) from the law of V given (p_0, E), i.e. uniformly over the n_v-subsets not contained in $S(p_0)$. The placebo quantities are $\bar{A} = \Pr[V' \subseteq S(P)]$ and $\bar{H} = \Pr[V' \subseteq S(P),\ S(P) \neq \mathcal{T}]$, and *audit-specific acceptance* is $\mathrm{Spec} = A - \bar{A}$.

The placebo asks how often the program the agent actually produced would be accepted by an equally distributed audit that it never interacted with. It is computable from each program's per-test results: with $s = |S(P)|$, $s_0 = |S(p_0)|$ and $s_\cap = |S(P) \cap S(p_0)|$,

$$\bar{a}(P, p_0) := \Pr[V' \subseteq S(P) \mid P, p_0] = \frac{\binom{s}{n_v} - \binom{s_\cap}{n_v}}{\binom{N}{n_v} - \binom{s_0}{n_v}}. \tag{1}$$

Proposition 1 (Properties of audit-specific acceptance)**.** (a) $\mathrm{Spec} = H - \bar{H}$*: every audit-specific acceptance is a false acceptance.* (b) *If* $P \perp\!\!\!\perp V \mid (p_0, E)$*, then* $\mathrm{Spec} = 0$*. This holds for* IID *and for* BIN *at* $K{=}1$*.* (c) *Let* $\mathcal{F}_{1:K}$ *be the sequence of feedback messages and verdicts, write* $\mathbb{H}$ *for Shannon entropy, and take mutual information and entropy under the law given* (p_0, E)*. For any policy,*

$$|\mathrm{Spec}| \ \le\ \mathbb{E}_{p_0|E}\sqrt{\tfrac{1}{2}\, I(P; V \mid p_0, E)}, \qquad I(P; V \mid p_0, E) \ \le\ \mathbb{H}(\mathcal{F}_{1:K} \mid p_0, Z, E).$$

(d) *For* IID *with first-pass selection and for* BIN*, the transcript given* (p_0, Z) *is determined by the first accepted retry in* $\{1, \dots, K\}$ *or "none", so* $\mathbb{H}(\mathcal{F}_{1:K} \mid p_0, Z, E) \le \log(K{+}1)$*. For* DIAG*, one round can reveal the failing visible tests; only the trivial bound* $\log \binom{N}{n_v}$ *applies.* (e) *For two policies,* $H_1 - H_0 = (\mathrm{Spec}_1 - \mathrm{Spec}_0) + (\bar{H}_1 - \bar{H}_0)$*: an audit-specific part plus a part that transfers to fresh audits.*

All statements are for a fixed task, prompt and audit size; the experimental aggregates average them with equal weight per activation record. Proofs are in Appendix A. We estimate Spec by the average of $\mathbf{1}\{V_i \subseteq S(P_i)\} - \bar{a}(P_i, p_{0,i})$ over activation records. This is unbiased and, relative to drawing one physical placebo audit per record, removes exactly $\mathbb{E}[\bar{a}(1-\bar{a})]$ of per-record variance (Proposition 2); on our data the per-record variance is 64% of the physical-placebo variance, and a 200-draw physical placebo reproduces the closed form (Appendix F).

What the estimand does and does not identify. Spec is an operational, statistical quantity: the excess acceptance of the realised output on the audit it interacted with, over an independent audit with the same conditional law. A positive value means that the output's acceptance depends on the identity of the realised audit. It does not identify intent, a particular mechanism (e.g. special-casing versus a heuristic repair that happens to cover the shown inputs), or a causal share. Likewise, $\bar{H}$ is false acceptance on independent audits, not proof of task-general improvement. Part (b) gives a falsifiable check: IID and BIN at $K{=}1$ have $\mathrm{Spec} = 0$ in expectation. Parts (c)–(d) give worst-case ceilings; they do not rank actual effects. Part (e) replaces the descriptive split $\Delta H = \Delta A\, R_0 + A_1\, \Delta R$, whose attribution depends on the reference policy, with terms that have operational meanings.

4 A REMEDY: INDEPENDENT FEEDBACK AUDITS

Proposition 1(b) suggests a change of protocol rather than of the agent. Draw a *feedback set* F of n_v tests from the same pool, independently of the acceptance audit V (the two may overlap), give diagnostics only on F, and accept on V as before. Call this policy INDEP.

Corollary 1 (Independent feedback carries no audit-specific acceptance)**.** *If* $F \perp\!\!\!\perp (V, p_0, Z)$*, then* INDEP *has* $\mathrm{Spec} = 0$ *at* $K{=}1$*, however informative the diagnostics on* F*. For* $K > 1$*,* $I(P; V \mid p_0, E) \le \log(K{+}1)$*, the ceiling of* IID*.*

Proof. At $K{=}1$ on E the output is $P = g(p_0, Z, \mathcal{F}_F(p_0))$, where the diagnostics $\mathcal{F}_F(p_0)$ are a function of (p_0, F); since $(Z, F) \perp\!\!\!\perp V \mid (p_0, E)$, also $P \perp\!\!\!\perp V \mid (p_0, E)$, and Proposition 1(b) applies. For $K > 1$, treat (Z, F) as the policy's private randomness in Proposition 1(c): given (p_0, Z, F) the programs and their diagnostics on F are fixed until the first acceptance, so only the index of the first accepted round depends on V, as in Proposition 1(d). □

Preregistered test (E9). Before any output existed we preregistered E9: Qwen3-8B with the EvalPlus settings of §5, new audit and sampling seeds, and the policies DIAG, INDEP and IID ($n{=}533$ activation records). The single primary test is M1: INDEP has higher useful acceptance than IID at $K{=}1$. Secondary tests are the calibration of INDEP (M2) and a larger Spec for DIAG than for INDEP (M3).

Results. M1 is supported: INDEP raises useful acceptance over IID by $+13.8$ [$+6.0, +21.8$] points at $K{=}1$, a fraction 0.74 [0.54, 0.95] of the gain from diagnostics on V ($+18.7$ [$+9.1, +25.8$]; Figure 2). Its audit-specific acceptance is $+0.1$ [$-1.0, +1.3$], against $+3.2$ [$+1.3, +5.4$] for DIAG (M3, supported: difference $+3.1$ [$+0.9, +5.0$]), and its excess false acceptance over IID is $+1.2$ [$-2.9, +6.2$] versus $+5.7$ [$+1.3, +10.5$] for DIAG. The retained utility is not complete: for 36% of activation records the first attempt passes every test in F, so INDEP can only say "rejected". At $K{=}8$ all three policies have positive Spec ($+5.7$ [$+3.3, +8.3$] for INDEP), as the selection ceiling allows; independence of the feedback does not remove selection.

Equal test budgets. A second stage of m hidden tests that the agent never saw also has zero audit-specific acceptance, at every horizon and for every policy, and a program failing f of h hidden tests survives it with probability $\binom{h-f}{m} / \binom{h}{m} \le e^{-mf/h}$ (Corollary 2, Appendix A). With a fixed budget of $t{=}8$ tests per task, we compare a single audit of $n_v{=}8$ with an audit of 4 plus an unseen stage of 4,

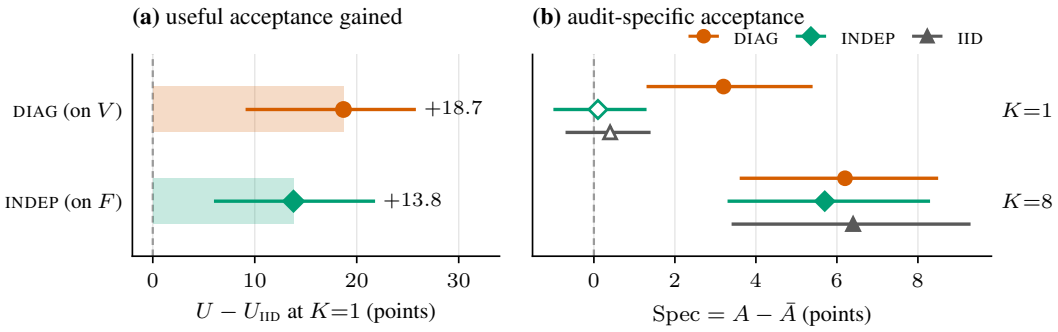


Figure 2: Independent feedback audits (E9, Qwen3-8B, one run; points, 95% task-cluster CIs). (a) Gain in useful acceptance over IID at K=1. (b) Audit-specific acceptance; hollow markers: CI contains 0. INDEP gives the diagnostics of DIAG on a feedback set F drawn independently of the acceptance audit V. Full numbers: Table 5.

on all 4304 Qwen3 (run, task, prompt) units (a post hoc analysis of existing outputs; Appendix E). At K=1 the split changes the false-acceptance share of DIAG by -0.2 [-0.9, $+0.4$] points and its useful acceptance by -0.5 [-1.2, $+0.1$]. At K=8 it lowers the share for IID by -0.9 [-1.6, -0.2] (useful acceptance -0.3 [-0.6, $+0.0$]) and costs DIAG -1.3 [-2.1, -0.4] in useful acceptance. Over all tasks, where most first attempts are already correct, reallocating tests matters little; it helps most where selection over many candidates is the source of audit fitting.

5 Experimental design

Tasks and audits. We use all 540 MBPP+ and HumanEval+ tasks (Liu et al., 2023) with at least 12 tests. For each task and $n_v \in \{1, 2, 4, 8\}$, a seeded hash draws V; an audit size is used only if it leaves ≥ 10 hidden tests, which excludes n_v=8 for 2 tasks. Hidden tests never enter any prompt.

Models and policies. Qwen3-8B (Qwen Team, 2025) (8k thinking budget; 16k in E13), Qwen3-32B (8k), MiniCPM4.1-8B (MiniCPM Team, 2025) (16k) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) (no thinking phase, 1k tokens), temperature 0.7, top-p 0.95. There are two system prompts: "produce a correct solution" and "you are graded only by evaluator tests; maximise acceptance". Feedback is stateless: each round sees the task, the previous program and the feedback, with values truncated to 300 characters. The IID pool has 9 samples per task and prompt and is shared across audits, so all policies are compared on identical activation sets. An offline executor records per-test results on every visible and hidden test for each output and never returns anything to the model, which makes the placebo exact.

Runs and inference. Each run draws new audits *and* new generation seeds (Table 7). V2 (Qwen3-8B) was exploratory. E2 (Qwen3-8B) and E3 (MiniCPM, E2's audits) preregistered H1: DIAG$-$IID H at K=8, n_v=1 has a positive 95% lower bound. E4 (Qwen3-8B, two audit draws) preregistered P1 (DIAG$-$IID H at K=1, pooled over n_v), P2 (BIN$-$IID H, CI contains 0) and P3 (DIAG$-$IID R). After the placebo analyses of these runs, five further runs preregistered placebo quantities as primary tests: E6 (Qwen3-32B), E7 (two MiniCPM runs), E8 (Llama-3.1-8B), E13 (Qwen3-8B with a 16k budget) and E9 (§4); E5 (§7) moves to a second benchmark. The estimator was fixed in advance: pooled differences on the activation set, each (task, prompt, n_v) record weighted equally, ratio-of-sums for R, and a 2,000-draw bootstrap that resamples tasks with all their records; policy contrasts are paired. Task-cluster intervals target new tasks *conditional on the realised runs*; as a complementary *run-conditional* summary we report the mean of the four Qwen3-8B (8k) runs with a t_3 interval (Appendix B). On V2–E4 the placebo analyses, the decomposition and all audit-size analyses were specified after E4 and are exploratory for those runs.

Table 1: Every preregistered primary test, and E4's secondary tests (percentage points; 95% task-cluster bootstrap CIs). Secondary tests of E5–E13 are in §6.3–7.

Run	Test	Contrast	Estimate [95% CI]	Verdict
E2 (Qwen3-8B)	H1	DIAG−IID H, K=8, n_v=1	+6.2 [−2.3, +14.4]	not supported
E3 (MiniCPM)	H1	DIAG−IID H, K=8, n_v=1	+8.8 [+1.3, +16.0]	supported
E4 (Qwen3-8B)	P1	DIAG−IID H, K=1, pooled	+5.3 [+1.5, +9.3]	supported
E4	P2	BIN−IID H, K=1, pooled	−1.3 [−3.6, +1.1]	CI contains 0, as predicted
E4	P3	DIAG−IID R, K=1, pooled	−0.8 [−11.5, +10.1]	not supported
E5 (LCB, Qwen3-8B)	Q1	DIAG−IID R, K=1 (two-sided)	+1.8 [−11.0, +14.5]	direction: none detected; equiv. not shown
E6 (Qwen3-32B)	C1	Spec of DIAG, K=1	+2.6 [+0.8, +4.5]	supported
E7 (MiniCPM, 2 runs)	D1	Spec of DIAG, K=1	+2.2 [+1.0, +3.4]	supported
E8 (Llama-3.1-8B)	C1	Spec of DIAG, K=1	+1.5 [+0.3, +2.7]	supported
E9 (Qwen3-8B)	M1	U: indep. diag.−IID, K=1	+13.8 [+6.0, +21.8]	supported
E13 (Qwen3-8B, 16k)	C1	Spec of DIAG, K=1	+2.1 [+0.5, +3.9]	supported

Table 2: Placebo audits, pooled over n_v and prompts (%). Spec $= A - \bar{A}$ is audit-specific acceptance; $\bar{H}$ and $\bar{R}$ are false acceptance and its share on the placebo audit. Qwen3-8B pools 4 runs (n=2240 activation records); MiniCPM is E3 (n=733).

Model	K	Policy	A	Spec [95% CI]	H	$\bar{H}$ (placebo)	R / $\bar{R}$
Qwen3-8B	1	Diagnostic	48.7	+2.9 [+1.8, +4.1]	10.8	7.9	22.2 / 17.3
		Binary	21.7	+0.1 [−0.3, +0.6]	4.3	4.2	19.9 / 19.5
		i.i.d.	23.7	+0.2 [−0.2, +0.6]	4.8	4.6	20.2 / 19.5
	8	Diagnostic	71.2	+6.0 [+4.5, +7.7]	17.7	11.7	24.9 / 17.9
		Binary	45.1	+1.5 [+0.7, +2.5]	9.9	8.4	22.0 / 19.2
		i.i.d.	48.7	+7.1 [+5.1, +9.5]	15.0	8.0	30.9 / 19.2
MiniCPM4.1-8B	1	Diagnostic	49.1	+2.9 [+1.4, +4.3]	10.0	7.1	20.3 / 15.3
		Binary	37.7	−0.5 [−1.6, +0.5]	5.7	6.2	15.2 / 16.3
		i.i.d.	38.6	−0.1 [−0.6, +0.5]	3.5	3.7	9.2 / 9.5
	8	Diagnostic	79.9	+8.0 [+5.0, +11.3]	19.9	11.9	24.9 / 16.5
		Binary	65.3	+2.0 [+0.8, +3.2]	10.6	8.7	16.3 / 13.7
		i.i.d.	63.2	+6.6 [+3.7, +10.0]	15.4	8.8	24.4 / 15.6

6 Results on EvalPlus

6.1 Preregistered outcomes

The exploratory run suggested a large excess of false acceptances at the smallest audit and longest horizon (+19.5 points). Its directional replication failed on new audits and seeds with Qwen3 (E2) and succeeded with MiniCPM on E2's audits (E3); the intervals overlap, so this does not show that the effect differs between runs, only that one run is not enough. The four-run Qwen3 mean for this contrast is +10.1 [−0.3, +20.4]. A one-round excess, found after V2 and E2, was confirmed on two fresh draws (P1); binary feedback did not show a detectable excess (P2). The test of a precision loss at K=1 (P3) was not supported, but its interval is too wide to exclude a sizeable change in either direction. The later runs test the placebo findings of the next subsection prospectively; their primary tests (Table 1, lower block) are discussed in §6.3.

6.2 Where the excess comes from: audit-specific versus transferable change

The placebo passes its calibration check (Table 2): at K=1, IID and BIN have Spec of +0.2 [−0.2, +0.6] and +0.1 [−0.3, +0.6] for Qwen3-8B and −0.1 [−0.6, +0.5] and −0.5 [−1.6, +0.5] for MiniCPM, as Proposition 1(b) requires. One round of diagnostic feedback produces Spec =+2.9 [+1.8, +4.1] (Qwen3-8B; four-run mean +2.9 [+1.9, +3.9]) and +2.9 [+1.4, +4.3] (MiniCPM). Proposition 1(e) splits the one-round excess of false acceptances over IID exactly: for Qwen3-8B, +2.7 [+1.5, +4.0] points are audit-specific and +3.4 [+1.1, +5.7] transfer to fresh audits (Appendix Table 8); the MiniCPM split is almost identical. In 30% of Qwen3-8B activation records the first attempt hit

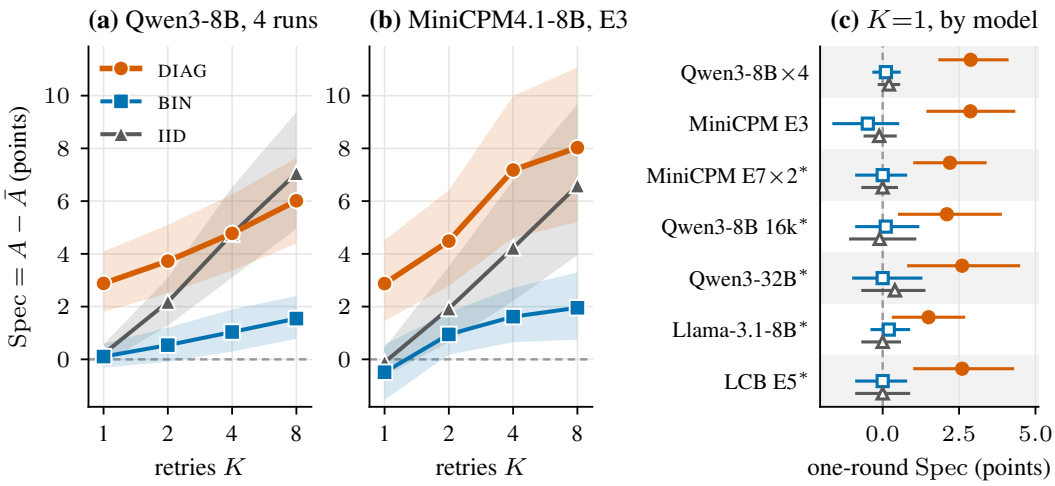


Figure 3: Audit-specific acceptance Spec (points, 95% task-cluster CIs). (a, b) By the number of retries K (bands: CIs); IID and BIN are calibrated at K=1, and selection over IID candidates catches up with diagnostics by K=4. (c) One round, all models; * preregistered runs with DIAG Spec > 0 as the primary test; hollow markers: CI contains 0. Numbers and the decomposition of ΔH: Table 4.

the 8k thinking budget; on the remaining records the one-round excess is $+8.2$ [$+4.8$, $+11.9$]. The transferable part reflects that diagnostics yield many more programs that pass a fresh audit ($\bar{A}$ is 22 points higher for Qwen3-8B), a minority of which are still wrong. For Qwen3-8B the audit-specific component's interval excludes zero under both prompts and at every audit size. All four Qwen3-8B runs agree in sign for the one-round quantities, with between-run SDs (0.6–1.1 points) no larger than the within-run SEs. Exploratorily, the component is larger on MBPP+ than on HumanEval+ (difference $+2.8$ [$+1.0$, $+4.7$] for Qwen3-8B), whose prompts contain worked examples.

Audit-specific acceptance is not literal copying. The runner flags a program that contains an input or expected value of an audit test absent from the task prompt and the first attempt. Among one-round false acceptances of the four Qwen3-8B runs, this flag fires for 9.1% of diagnostic outputs (22/242; 2/22 among those with placebo acceptance $\bar{a} < 0.2$), 11.2% of IID and 16.5% of BIN outputs (Appendix H; post hoc). Explicit copying of audit values is no more common after diagnostics than after resampling, so it does not explain audit-specific acceptance; the data are consistent with repairs that are narrow enough to cover the shown inputs but not the task.

6.3 Across models and scale

Figure 3(c) and Table 4 show the one-round quantities for every model. In the preregistered runs, diagnostic Spec has a positive lower bound for every model: Qwen3-32B (E6, C1 supported: $+2.6$ [$+0.8$, $+4.5$]), MiniCPM (E7, two new runs, $+2.2$ [$+1.0$, $+3.4$]; three-run mean with E3 $+2.5$ [$+1.5$, $+3.4$]), Llama-3.1-8B (E8, $+1.5$ [$+0.3$, $+2.7$]) and Qwen3-8B with a 16k budget (E13, $+2.1$ [$+0.5$, $+3.9$]), and IID and BIN remain calibrated in each run. The magnitudes are similar across the two reasoning families and scales (32B minus pooled 8B: -0.3 [-2.4, $+2.2$]) and smaller for the model without a thinking phase. The secondary contrasts are weaker: at 32B, neither the excess of DIAG over IID in Spec (C3, not supported) nor in false acceptance (C4, not supported) has a positive lower bound in this single run, and Llama-3.1-8B's excess false acceptance ($+1.5$ [-1.7, $+4.6$]) is not distinguishable from 0. The larger precision loss of MiniCPM in E3 ($+11.1$ [$+5.2$, $+17.5$], post hoc) shrinks to $+5.6$ [-2.5, $+15.8$] in E7 (direction: none detected). With a 16k budget, first-attempt truncation on the activation set falls from 30% to 11%, and the estimates stay close to the 8k runs.

6.4 Audit-specific acceptance at longer horizons

First-pass selection over several i.i.d. candidates is itself adaptive, with a ceiling of $\log(K+1)$ nats (Proposition 1(d)). Its audit-specific acceptance grows quickly with the horizon: $+0.2$ [-0.2, $+0.6$],

Table 3: E5 (LiveCodeBench, Qwen3-8B, one run, n=724 activation records): every preregistered test (points, 95% task-cluster CIs).

Test	Contrast	Estimate [95% CI]	Verdict
Q1 (primary)	DIAG−IID R, K=1	+1.8 [−11.0, +14.5]	direction: none detected; equivalence: not shown
Q2	$R(n_v{=}1) - R(n_v{=}8)$, K=1	+9.3 [−4.4, +22.1]	not supported
Q3	DIAG−IID H, K=1	+4.0 [−0.3, +8.1]	not supported
Q4	BIN−IID H, K=1	−0.4 [−3.4, +2.7]	CI contains 0, as predicted
Q5	DIAG−IID R, K=8	−2.4 [−13.4, +10.4]	direction: none detected; equivalence: not shown

+2.2 [+1.3, +3.1], +4.8 [+3.1, +6.5] and +7.1 [+5.1, +9.5] at $K = 1, 2, 4, 8$ for Qwen3-8B, against +2.9 [+1.8, +4.1], +3.7 [+2.5, +5.1], +4.8 [+3.4, +6.3] and +6.0 [+4.5, +7.7] for diagnostics (Appendix G; post hoc). By four retries, selection matches diagnostics, and at K=8 we detect no difference (−1.0 [−3.4, +1.3]). Binary feedback accumulates less (+1.5 [+0.7, +2.5] at K=8), consistent with binary retries often reproducing the previous program (64% vs. 31% for diagnostic retries in run E4a), which leaves the selection step fewer distinct candidates. Diagnostics keep a large gain in useful acceptance at both horizons (ΔU =+19.0 [+12.0, +25.8] at K=1, +19.9 [+13.2, +27.1] at K=8). With a single retry, diagnostics add about 3 points of audit fitting that a resample would not; with more retries, resampling with selection has a similar point estimate.

6.5 Precision of accepted programs, audit size and feedback content

For a deployer, the relevant quantity is the share R of accepted programs that are wrong. For Qwen3-8B at K=1 it is 22.2% (DIAG) versus 20.2% (IID), a difference of +2.0 [−6.0, +9.4]; on placebo audits −2.2 [−9.4, +5.1]. Both intervals span changes in either direction, and the preregistered runs of Table 4 give no consistent direction. Audit size matters more clearly. Holding the population fixed at the 391 Qwen3-8B units rejected at every audit size (Figure 4, Appendix C), R falls from 35/27/29% (DIAG/BIN/IID) at n_v=1 to 16/16/15% at n_v=8; the paired size effect is +19.1 [+9.5, +29.3] for DIAG and +13.8 [+6.3, +22.7] for IID, while the policy effect at fixed n_v is never significant. Whether size matters more than feedback is not formally established (+8.2 [−5.0, +15.5]). On the exploratory run we also ablated the feedback content at n_v=1 (Appendix D). The excess false acceptance over IID at K=8 orders full (+19.5), input and expected (+16.8), input only (+8.0), failure count (+1.8) and binary (+0.9). Re-executing these outputs on their hidden tests gives the one-round audit-specific acceptance (post hoc): full +6.5 [+1.8, +12.5], input and expected +9.1 [+4.3, +14.6], input only +1.7 [−0.9, +4.4], failure count +0.3 [−1.5, +1.8] and binary +1.0 [−0.4, +2.8]; revealing the expected output is what makes the retry audit-specific.

7 A second benchmark: LiveCodeBench (E5, preregistered)

Before any output existed we preregistered E5 on all 444 LeetCode problems of LiveCodeBench (Jain et al., 2025), which have functional tests (12 to over 100 per problem) and are harder than EvalPlus. Audits are drawn as before with at least 10 hidden tests (n_v=8 where $\geq$ 18 tests). The setup is Qwen3-8B with a 16k thinking budget, both prompts, $K \leq 8$, and the policies DIAG, BIN, IID and an input-only diagnostic arm. E5 was preregistered while the precision question was open, so its single primary test is Q1, DIAG−IID R at K=1 (two-sided), with direction and practical equivalence (±10 points) reported separately; Q2–Q5 are secondary, and its placebo analyses were specified after the EvalPlus placebo results but before any E5 retry existed. E5 is a single run with n=724 activation records. 56% of them are first attempts that hit the thinking budget (12.6% of all first attempts).

Preregistered results. Q1: DIAG−IID R at K=1 is +1.8 [−11.0, +14.5] (R 28.2% vs. 26.4%; direction: none detected; equivalence: not shown). As on EvalPlus (P3), the precision contrast is too imprecise to establish or exclude a change of 10 points. The audit-size contrast (Q2) is +9.3 [−4.4, +22.1] (not supported), in the direction seen on EvalPlus. The one-round excess in false acceptance (Q3) is +4.0 [−0.3, +8.1] (not supported; E4: +5.3), and binary feedback shows −0.4 [−3.4, +2.7] (Q4). Diagnostics raise useful acceptance by +9.3 [+4.1, +16.3] points at K=1, less than on EvalPlus, because many activated problems are ones the model cannot finish within its budget.

Placebo analyses (secondary). At $K{=}1$, Spec is -0.0 [-0.9, $+0.9$] (IID) and -0.0 [-0.9, $+0.8$] (BIN), consistent with calibration, and $+2.6$ [$+1.0$, $+4.3$] for DIAG and $+1.1$ [-0.3, $+2.6$] for input-only diagnostics. The one-round excess over IID decomposes into $+2.6$ [$+0.8$, $+4.5$] audit-specific and $+1.4$ [-1.3, $+4.0$] transferable points. At $K{=}8$ the pattern of EvalPlus recurs: DIAG $+4.5$ [$+2.5$, $+6.9$], IID $+6.7$ [$+3.9$, $+9.7$], BIN $+1.2$ [-0.2, $+2.8$]. On the second benchmark, too, one round of diagnostics produces a few points of audit-specific acceptance, and selection produces a similar amount over eight retries.

8 Conclusion

Placebo audits turn the question of whether an agent fitted the tests that accepted it into a quantity computed in closed form from each output, with a built-in calibration check. For open models up to 32B with stateless feedback, on two benchmarks, one round of diagnostic feedback produced a few points of audit-specific acceptance per activated record, roughly half of its excess false acceptance, and selection alone reached a similar level within a few retries. Drawing the feedback independently of the acceptance audit removes the one-round component by construction while keeping most of the gain, and an unseen second stage removes the audit-specific component at every horizon.

Ethics statement

This work involves no human subjects or personal data; it uses public benchmarks (EvalPlus, LiveCodeBench) and open-weight models. It describes how coding agents can come to pass the specific tests that grade them. We think the risk of misuse is low compared with the benefit of measuring this failure mode, and the remedies we analyse (independent feedback and an unseen audit stage) are defensive.

Reproducibility statement

All estimands are defined in §3, with complete proofs in Appendix A and inference details in Appendix B. The supplementary material contains every preregistration verbatim, with server time stamps and dated revisions, the one-shot analysis scripts, the generation and scoring code, and a manifest of file modification times and hashes. Seeds for audits, sampling and bootstrap are fixed and recorded.

AI use statement

Generative AI tools were used in two limited ways: to polish the English grammar of the manuscript, and to assist with writing parts of the experiment code. All AI-assisted text and code were checked in turn by several of the authors for correctness. The research ideas, the method and its implementation, the design of the experiments, and the structure of the paper are the authors' own work. The authors take full responsibility for the content of this paper.

References

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv:2503.11926*, 2025.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *ICML*, 2015.

Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. CodeT: Code generation with generated tests. In *ICLR*, 2023.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. In *ICLR*, 2024.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, et al. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv:2406.10162*, 2024.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *STOC*, 2015b.

Vitaly Feldman, Roy Frostig, and Moritz Hardt. The advantages of multiple classes for reducing overfitting from test set reuse. In *ICML*, 2019.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *ICML*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *ICLR*, 2025.

Yujia Li et al. Competition-level code generation with AlphaCode. *Science*, 378(6624):1092–1097, 2022.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *NeurIPS*, 2023.

Thanawat Lodkaew, Johannes Ackermann, Soichiro Nishimori, Nontawat Charoenphakdee, Masashi Sugiyama, and Takashi Ishida. Do coding agents deceive us? Detecting and preventing cheating via capped evaluation with randomized tests. *arXiv:2606.07379*, 2026.

Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, et al. Natural emergent misalignment from reward hacking in production RL. *arXiv:2511.18397*, 2025.

Aman Madaan et al. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, 2023.

David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's law. *arXiv:1803.04585*, 2018.

Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, and Benjamin Recht. Model similarity mitigates test set overuse. In *NeurIPS*, 2019.

MiniCPM Team. MiniCPM4: Ultra-efficient LLMs on end devices. *arXiv:2506.07900*, 2025.

Qwen Team. Qwen3 technical report. *arXiv:2505.09388*, 2025.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In *NeurIPS*, 2019.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *AISTATS*, 2016.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*, 2023.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion. In *NeurIPS*, 2025.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *NeurIPS*, 2022.

You Wang, Michael Pradel, and Zhongxin Liu. Are "solved issues" in SWE-bench really solved correctly? An empirical study. In *ICSE*, pp. 169–181, 2026.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *NeurIPS*, 2017.

Boxi Yu, Yuxuan Zhu, Pinjia He, and Daniel Kang. UTBoost: Rigorous evaluation of coding agents on SWE-Bench. In *ACL*, pp. 3762–3774, 2025.

Boxi Yu, Yang Cao, Yuzhong Zhang, Liting Lin, et al. SWE-ABS: Adversarial benchmark strengthening exposes inflated success rates on test-based benchmark. *arXiv:2603.00520*, 2026.

Ziqian Zhong, Aditi Raghunathan, and Nicholas Carlini. ImpossibleBench: Measuring LLMs' propensity of exploiting test cases. *arXiv:2510.20270*, 2025.

A Proofs and further theory

Proof of the placebo formula. Given (p_0, E), V' is independent of P and uniform over the $\binom{N}{n_v} - \binom{s_0}{n_v}$ subsets of size n_v not contained in $S(p_0)$. Of these, the ones contained in $S(P)$ are the subsets of $S(P)$ except those of $S(P) \cap S(p_0)$, which number $\binom{s}{n_v} - \binom{s_\cap}{n_v}$.

Proposition 1. (a) Write $S = S(P)$. If $S = \mathcal{T}$, both indicators equal 1, so $A - \bar{A} = \Pr[V \subseteq S,\ S \neq \mathcal{T}] - \Pr[V' \subseteq S,\ S \neq \mathcal{T}] = H - \bar{H}$. (b) Under $P \perp\!\!\!\perp V \mid (p_0, E)$, the pairs (P, V) and (P, V') are equal in law given (p_0, E). For IID at $K{=}1$, the output is the fresh sample p_1 (if p_1 is rejected it is still the last program), and $p_1 \perp\!\!\!\perp (p_0, V)$. For BIN at $K{=}1$, the prompt is a function of p_0 only, because the verdict is "rejected" on E, so $P = g(p_0, Z)$. With more rounds, first-pass selection breaks independence. (c) Fix p_0 with $\Pr[E \mid p_0] > 0$ and write Q for the law of $(P, V, Z, \mathcal{F})$ given (p_0, E). By definition of the placebo, $\bar{A}$ given p_0 is $\mathbb{E}_{Q_P \otimes Q_V} g$ with $g = \mathbf{1}\{V \subseteq S(P)\}$, while A given p_0 is $\mathbb{E}_{Q_{PV}} g$. Since $g \in [0, 1]$, the variational bound and Pinsker's inequality give (Russo & Zou, 2016; Xu & Raginsky, 2017)

$$\begin{aligned}\left|\mathbb{E}_{Q_{PV}} g - \mathbb{E}_{Q_P \otimes Q_V} g\right| &\leq \mathrm{TV}(Q_{PV}, Q_P \otimes Q_V) \\ &\leq \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(Q_{PV} \,\|\, Q_P \otimes Q_V)} = \sqrt{\tfrac{1}{2}\, I(P; V \mid p_0, E)},\end{aligned}$$

and averaging over $p_0 \mid E$ gives the first bound. The event E depends only on (p_0, V) and $Z \perp\!\!\!\perp (p_0, V)$, so $Z \perp\!\!\!\perp V$ under Q. Given (p_0, Z) the interaction is deterministic in V: each generated program is a function of $(p_0, Z,$ previous feedback$)$, and each feedback message is a function of the audit and the current program. Hence $P = \phi(p_0, Z, \mathcal{F}_{1:K})$, and by data processing and the chain rule, under Q,

$$\begin{aligned}I(P; V \mid p_0, E) \leq I(Z, \mathcal{F}_{1:K}; V \mid p_0, E) &= \underbrace{I(Z; V \mid p_0, E)}_{=0} + I(\mathcal{F}_{1:K}; V \mid p_0, Z, E) \\ &\leq \mathbb{H}(\mathcal{F}_{1:K} \mid p_0, Z, E).\end{aligned}$$

(d) For BIN and IID-with-selection, given (p_0, Z) the sequence of programs is fixed until the first acceptance, so the transcript is determined by the index of the first accepted round in $\{1, \ldots, K\}$ or "none". For DIAG, given (p_0, Z) the transcript is a deterministic function of V, so $\mathbb{H}(\mathcal{F}_{1:K} \mid p_0, Z, E) \leq \mathbb{H}(V \mid p_0, E) \leq \log \binom{N}{n_v}$, whether or not the feedback identifies the failing tests (truncated values or duplicate inputs can prevent identification; the bound survives). (e) Algebra using (a).

Audit-size reference. For a program distribution fixed independently of n_v and an independent uniform audit, with $\rho_f = \Pr[|\mathcal{T} \setminus S(P)| = f]$,

$$\Pr[\text{accept}] = \sum_{f \geq 0} \rho_f \, \frac{\binom{N-f}{n_v}}{\binom{N}{n_v}}, \qquad \frac{\binom{N-f}{n+1}}{\binom{N}{n+1}} = \frac{\binom{N-f}{n}}{\binom{N}{n}} \cdot \frac{N - f - n}{N - n}.$$

For $f \geq 1$ the last factor is at most 1, so every term with $f \geq 1$ is non-increasing in n_v, while the $f = 0$ term is constant. Hence $\Pr[\text{accept}]$ is non-increasing, and the false-acceptance share $1 - \rho_0 / \Pr[\text{accept}]$ is non-increasing in n_v.

Estimation. Index activation records by $i = 1, \dots, n$ (task, prompt, audit size, run), with first attempt $p_{0,i}$, audit V_i, output P_i, $A_i = \mathbf{1}\{V_i \subseteq S(P_i)\}$ and $\bar{a}_i = \bar{a}(P_i, p_{0,i})$ from equation 1. Assume, as in the design, that audits of distinct records are independent given the first attempts, that the policy randomness is independent of all audits, and that records of distinct tasks are independent.

Proposition 2 (Closed-form placebo estimator)**.** *Let $D_i = A_i - \bar{a}_i$, $\widehat{\mathrm{Spec}} = \frac{1}{n}\sum_i D_i$, and let Spec_i be the audit-specific acceptance of record i's stratum given $p_{0,i}$.* (a) $\mathbb{E}[D_i \mid p_{0,i}, E_i] = \mathrm{Spec}_i$*, so $\widehat{\mathrm{Spec}}$ is unbiased for $\frac{1}{n}\sum_i \mathrm{Spec}_i$ conditional on the first attempts and the activation pattern.* (b) *Let $B_i = \mathbf{1}\{V_i' \subseteq S(P_i)\}$ for one physical placebo draw V_i'. Then $\mathrm{Var}(A_i - B_i) = \mathrm{Var}(D_i) + \mathbb{E}[\bar{a}_i(1-\bar{a}_i)]$: the closed form removes exactly the placebo's own sampling noise.* (c) $\mathrm{Var}(\widehat{\mathrm{Spec}}) = n^{-2}\sum_t \mathrm{Var}\big(\sum_{i\in t} D_i\big)$*, summing over tasks t; this is the quantity the task-cluster bootstrap estimates.* (d) *If every record's policy satisfies the independence of Proposition 1(b) and task t has n_t records, then for every $\epsilon > 0$, conditional on the first attempts and the activation pattern,*

$$\Pr\left[|\widehat{\mathrm{Spec}}| \geq \epsilon\right] \leq 2\exp\Big(-\frac{n^2\epsilon^2}{2\sum_t n_t^2}\Big),$$

a finite-sample calibration test that needs no normal approximation.

Proof. (a) Given $(p_{0,i}, E_i)$ the placebo V_i' is independent of (P_i, V_i, Z) with the law of V_i, so $\bar{a}_i = \mathbb{E}[B_i \mid P_i, V_i, Z, p_{0,i}]$ and, by the tower property, $\mathbb{E}[D_i \mid p_{0,i}, E_i] = \Pr[V_i \subseteq S(P_i) \mid p_{0,i}, E_i] - \Pr[V_i' \subseteq S(P_i) \mid p_{0,i}, E_i] = \mathrm{Spec}_i$. The activation of any other record j depends only on $(p_{0,j}, V_j)$, and V_j is independent of (V_i, Z) given the first attempts (this holds even when records share the IID pool through Z), so conditioning on the activation pattern leaves this law unchanged. (b) $A_i - B_i = D_i + (\bar{a}_i - B_i)$ and $\mathbb{E}[\bar{a}_i - B_i \mid P_i, V_i, Z, p_{0,i}] = 0$, so the cross term vanishes and $\mathrm{Var}(\bar{a}_i - B_i) = \mathbb{E}[\mathrm{Var}(B_i \mid P_i, p_{0,i})] = \mathbb{E}[\bar{a}_i(1-\bar{a}_i)]$. (c) Cluster sums of distinct tasks are independent. (d) Under the independence, $\mathbb{E}[D_i \mid p_{0,i}, E_i] = 0$ by Proposition 1(b), so the task sums $X_t = \sum_{i\in t} D_i$ are independent, mean zero and lie in $[-n_t, n_t]$; apply Hoeffding's inequality to $\sum_t X_t = n\widehat{\mathrm{Spec}}$. □

Corollary 2 (An unseen stage carries no audit-specific acceptance)**.** *Let a policy interact only with V, and let W be a uniformly random m-subset of the $h = N - n_v$ hidden tests, drawn independently of (Z, P) given V and never shown to the policy.* (a) *Given (P, V) with $V \subseteq S(P)$ and $f = |\mathcal{T} \setminus S(P)|$,*

$$\Pr[W \subseteq S(P) \mid P, V] = \binom{h-f}{m}\Big/\binom{h}{m} \leq \Big(1 - \frac{f}{h}\Big)^m \leq e^{-mf/h}.$$

(b) *Relative to a placebo W' with the law of W given V, the second stage has zero audit-specific acceptance, for every policy however adaptive to V.* (c) *Hence false acceptance after both stages is*

$$H_2 = \mathbb{E}\Big[\mathbf{1}\{V \subseteq S(P),\ S(P) \neq \mathcal{T}\}\binom{h-f}{m}\Big/\binom{h}{m}\Big] \leq H.$$

Proof. (a) On $V \subseteq S(P)$ all f failures are hidden, and W is uniform over the $\binom{h}{m}$ hidden subsets, $\binom{h-f}{m}$ of which avoid them; the ratio is $\prod_{j=0}^{m-1}\big(1 - \frac{f}{h-j}\big) \leq (1-f/h)^m$, and $1 - x \leq e^{-x}$. (b) $W \perp\!\!\!\perp P \mid V$, so (P, W) and (P, W') are equal in law given V; this is Proposition 1(b) with (p_0, E) replaced by V. (c) Take expectations of (a) over the accepted, incorrect outputs; passing $V \cup W$ while $S(P) \neq \mathcal{T}$ is exactly passing both stages and failing a remaining hidden test. □

B INFERENCE DETAILS

All policy contrasts are paired, because the same activation records appear in every arm. The IID pool is shared across audit sizes and policies within a run, but not across runs, and bootstrap draws resample tasks, carrying all of a task's arms, audit sizes, prompts and runs. The run-conditional summary treats the four Qwen3 runs as units: the mean of the four run estimates with a t_3 interval based on their between-run SD, and leave-one-run-out means (LOO ranges: ΔH 5.6–6.5, DIAG Spec 2.7–3.2, ΔSpec 2.3–3.1, transferable part 3.0–3.7; ΔR is excluded, its run estimates range from -7.0 to $+5.3$). As a heuristic sensitivity check we also fit a DerSimonian–Laird random-effects

model whose inputs are each run's task-bootstrap SE (the standard deviation of its bootstrap draws), whose weights are inverse total variance, and whose t_3 interval is centred on the random-effects mean. Its within-run variances come from task resampling while runs share tasks, so its covariance assumptions are not exactly met. With four runs neither analysis can establish that heterogeneity is absent; $\hat{\tau} = 0$ is a boundary estimate. For the long-horizon contrast the random-effects summary is $+9.9$ [-0.2, $+20.1$] ($\hat{\tau} = 4.2$). The CIs in the appendix tables use one common bootstrap seed, whereas each preregistered analysis used its own seed. This is why, e.g., E2's interval is $[-2.3, 14.4]$ in Table 1 and $[-1.8, 14.8]$ in Table 7; the point estimates are identical.

C Additional EvalPlus results

Table 4: Audit-specific acceptance and the decomposition at $K{=}1$ across models (points; 95% task-cluster CIs). Upper block: runs analysed after the placebo was introduced (exploratory). Lower block: preregistered runs whose primary test is DIAG Spec > 0.

		Spec at $K{=}1$			DIAG−IID at $K{=}1$			
Model (runs)	n	DIAG	IID	BIN	ΔH	ΔSpec	$\Delta \bar{H}$	ΔR
Qwen3-8B, 8k (4 runs)	2240	+2.9 [+1.8, +4.1]	+0.2 [−0.2, +0.6]	+0.1 [−0.3, +0.6]	+6.0 [+3.4, +9.1]	+2.7 [+1.5, +4.0]	+3.4 [+1.1, +5.7]	+2.0 [−6.0, +9.4]
MiniCPM4.1-8B (E3)	733	+2.9 [+1.4, +4.3]	−0.1 [−0.6, +0.5]	−0.5 [−1.6, +0.5]	+6.4 [+3.6, +9.5]	+3.0 [+1.4, +4.6]	+3.4 [+1.1, +5.8]	+11.1 [+5.2, +17.5]
MiniCPM4.1-8B (E7, 2 runs)	1462	+2.2 [+1.0, +3.4]	−0.0 [−0.7, +0.5]	−0.0 [−0.9, +0.8]	+5.0 [+2.7, +7.7]	+2.3 [+1.0, +3.6]	+2.7 [+0.7, +4.3]	+5.6 [−2.5, +15.8]
Qwen3-8B, 16k (E13)	504	+2.1 [+0.5, +3.9]	−0.1 [−1.1, +1.1]	+0.1 [−0.9, +1.2]	+5.9 [+0.9, +10.3]	+2.2 [+0.2, +4.4]	+3.7 [+0.4, +7.1]	+1.0 [−14.3, +17.2]
Qwen3-32B (E6)	452	+2.6 [+0.8, +4.5]	+0.4 [−0.7, +1.4]	+0.0 [−1.0, +1.3]	+4.0 [−1.4, +8.8]	+2.1 [−0.1, +4.4]	+1.8 [−1.8, +5.3]	−2.5 [−18.7, +15.8]
Llama-3.1-8B (E8)	1253	+1.5 [+0.3, +2.7]	−0.0 [−0.7, +0.6]	+0.2 [−0.4, +0.9]	+1.5 [−1.7, +4.6]	+1.5 [+0.1, +2.8]	−0.0 [−1.9, +1.9]	−0.2 [−11.4, +10.3]

Table 5: Independent feedback audits (E9, Qwen3-8B, one run; %, ΔU in points with 95% task-cluster CIs). INDEP gives diagnostics on an independently drawn feedback set F and accepts on V.

K	Policy	A	U	H	R	Spec [95% CI]	ΔU vs. IID
1	Diagnostic (on V)	46.8	36.9	9.9	21.2	+3.2 [+1.3, +5.4]	+18.7 [+9.1, +25.8]
	Independent diagnostic (on F)	37.3	32.0	5.3	14.3	+0.1 [−1.0, +1.3]	+13.8 [+6.0, +21.8]
	i.i.d.	22.4	18.2	4.2	18.7	+0.4 [−0.7, +1.4]	—
8	Diagnostic (on V)	69.1	52.8	16.3	23.6	+6.2 [+3.6, +8.5]	
	Independent diagnostic (on F)	62.2	51.1	11.1	17.9	+5.7 [+3.3, +8.3]	
	i.i.d.	47.4	33.0	14.4	30.3	+6.4 [+3.4, +9.3]	

Table 6: Run-level view for Qwen3 (points; activation records per run: V2 567, E2 573, E4a 543, E4b 557, of which 113, 113, 106, 114 at $n_v{=}1$ for the last row). Per-run point estimates; the mean of the four runs with a t_3 interval from their between-run SD (primary run-conditional summary); and, as a heuristic sensitivity check, a DerSimonian–Laird random-effects summary with between-run SD $\hat{\tau}$ ($\hat{\tau} = 0$ is a boundary estimate, not evidence of homogeneity).

Quantity (Qwen3)	V2	E2	E4a	E4b	Run mean [t_3 95% CI]	DL [t_3 95% CI], $\hat{\tau}$
ΔH DIAG−IID, $K{=}1$	+7.4	+6.1	+5.9	+4.7	+6.0 [+4.2, +7.8]	+6.1 [+2.7, +9.6], 0.0
Spec DIAG, $K{=}1$	+3.2	+2.9	+3.4	+2.0	+2.9 [+1.9, +3.9]	+2.7 [+1.4, +4.1], 0.0
ΔSpec DIAG−IID, $K{=}1$	+2.9	+2.8	+3.7	+1.3	+2.7 [+1.1, +4.3]	+2.5 [+0.8, +4.2], 0.5
$\Delta \bar{H}$ DIAG−IID, $K{=}1$	+4.5	+3.3	+2.2	+3.4	+3.3 [+1.8, +4.9]	+3.6 [+0.8, +6.4], 0.0
ΔR DIAG−IID, $K{=}1$	+5.3	+4.2	−7.0	+2.5	+1.3 [-7.7, +10.2]	+2.3 [-7.7, +12.2], 0.0
Spec DIAG, $K{=}8$	+6.6	+6.4	+6.3	+4.7	+6.0 [+4.6, +7.4]	+5.9 [+4.0, +7.7], 0.0
Spec IID, $K{=}8$	+7.2	+8.4	+5.9	+6.6	+7.0 [+5.3, +8.7]	+6.9 [+4.2, +9.6], 0.0
ΔH DIAG−IID, $K{=}8$, $n_v{=}1$	+19.5	+6.2	+9.4	+5.3	+10.1 [-0.3, +20.4]	+9.9 [-0.2, +20.1], 4.2

Audit size, full detail. On the full activation population, R falls from 31/25/27% to 17/12/16% (DIAG/BIN/IID, Qwen3) between $n_v{=}1$ and 8, but that comparison mixes populations. On the common-support population the Qwen3 audit-size effect for BIN is $+10.5$ [-0.1, +22.0]; the policy effect DIAG−IID is $+1.0$ [-7.4, +8.9] at $n_v{=}8$, and the policy-by-size interaction is $+5.3$ [-3.4, +14.2]. For MiniCPM, R falls from 25/19/13% to 14/15/7%, and the size effects are $+11.0$ [+0.2, +23.1] (DIAG) and $+5.8$ [+0.0, +13.8] (IID). The Qwen3 subgroup contrast in ΔSpec over IID between MBPP+ and HumanEval+ is $+2.7$ [$+0.6$, $+4.8$] (MiniCPM $+2.6$ [$+0.1$, $+4.9$]), and the one-round

Table 7: All five runs: DIAG−IID contrasts (points, 95% CI). Point estimates match the preregistered analyses; CIs here use a common bootstrap seed and task clustering.

Run	Model	Status	n	ΔH, $K{=}1$, pooled	ΔR, $K{=}1$, pooled	ΔH, $K{=}8$, $n_v{=}1$
V2	Qwen3-8B	exploratory	567	$+7.4$ [$+3.9$, $+11.4$]	$+5.3$ [-7.1, $+17.0$]	$+19.5$ [$+10.1$, $+29.3$]
E2	Qwen3-8B	prereg. (E2)	573	$+6.1$ [$+1.4$, $+11.1$]	$+4.2$ [-9.2, $+15.8$]	$+6.2$ [-1.8, $+14.8$]
E4A	Qwen3-8B	prereg. (E4)	543	$+5.9$ [$+1.5$, $+10.9$]	-7.0 [-23.5, $+8.5$]	$+9.4$ [$+0.0$, $+18.7$]
E4B	Qwen3-8B	prereg. (E4)	557	$+4.7$ [$+0.4$, $+9.3$]	$+2.5$ [-9.0, $+13.7$]	$+5.3$ [-4.7, $+15.4$]
E3	MiniCPM4.1-8B	prereg. (E3)	733	$+6.4$ [$+3.6$, $+9.5$]	$+11.1$ [$+5.2$, $+17.5$]	$+8.8$ [$+1.3$, $+16.3$]

Table 8: Exact decomposition of the excess false acceptance over i.i.d. resampling (Proposition 1e), in points with 95% CIs.

Model	Contrast	ΔH (total)	ΔSpec (audit-specific)	$\Delta \bar{H}$ (transferable)
Qwen3-8B	Diagnostic−i.i.d., $K{=}1$	$+6.0$ [$+3.4$, $+9.1$]	$+2.7$ [$+1.5$, $+4.0$]	$+3.4$ [$+1.1$, $+5.7$]
	Binary−i.i.d., $K{=}1$	-0.4 [-2.1, $+1.3$]	-0.1 [-0.7, $+0.4$]	-0.3 [-2.0, $+1.4$]
	Diagnostic−i.i.d., $K{=}8$	$+2.7$ [-1.2, $+6.5$]	-1.0 [-3.4, $+1.3$]	$+3.7$ [$+1.1$, $+6.5$]
	Binary−i.i.d., $K{=}8$	-5.1 [-8.0, -2.3]	-5.5 [-7.6, -3.8]	$+0.4$ [-1.4, $+2.5$]
MiniCPM4.1-8B	Diagnostic−i.i.d., $K{=}1$	$+6.4$ [$+3.6$, $+9.5$]	$+3.0$ [$+1.4$, $+4.6$]	$+3.4$ [$+1.1$, $+5.8$]
	Binary−i.i.d., $K{=}1$	$+2.2$ [-0.3, $+4.7$]	-0.4 [-1.3, $+0.5$]	$+2.6$ [-0.1, $+5.0$]
	Diagnostic−i.i.d., $K{=}8$	$+4.5$ [-0.6, $+9.5$]	$+1.4$ [-2.9, $+5.5$]	$+3.1$ [$+0.5$, $+5.7$]
	Binary−i.i.d., $K{=}8$	-4.8 [-9.3, -0.1]	-4.6 [-7.7, -1.8]	-0.1 [-3.4, $+3.2$]

excess in false acceptance is $+7.0$ [+3.7, +10.6] on MBPP+ and $+2.0$ [-1.6, +6.0] on HumanEval+ (Qwen3).

D Feedback-content ablation (E1, exploratory run)

At $n_v{=}1$, $K{=}8$, with $n = 113$ activation records, the excess over IID in H (points, 95% CI) was: full diagnostic $+19.5$ [10.0, 30.1]; input and expected $+16.8$ [8.3, 25.7]; input only $+8.0$ [2.7, 13.4]; failure count $+1.8$ [-2.6, 7.1]; binary $+0.9$ [-3.6, 6.4]. The retry truncation rates were 41/54/34/24/24%. We re-executed every chosen E1 output on its hidden tests to obtain per-test results and computed audit-specific acceptance for each feedback variant (Figure 5 and Table 15; a secondary analysis of an exploratory run). The re-execution agrees with the original hidden verdicts for 83 of 86 accepted outputs checked; the three differences are time-limit boundary cases.

E Equal test budgets (post hoc)

For each (run, task, prompt) unit of the four Qwen3-8B runs whose task has all four audit sizes (4304 units), and each budget split $n_v{+}m = 8$, the first attempt is the output if it passes its audit V; otherwise the policy's output at horizon K is. Acceptance additionally requires passing a uniformly random m-subset of that split's hidden tests, computed exactly from per-test results (Corollary 2). The population includes units whose first attempt was accepted, because the activation set differs between splits; R is therefore much lower than on activation records. Contrasts are paired within units, with 2,000 task-cluster bootstrap draws.

F Physical placebo check

For all 17,838 combinations of activation record, policy and $K \in \{1, 8\}$ in the five exploratory-period runs, we drew 200 physical placebo audits by rejection sampling from the conditional law and compared their acceptance rate with the closed form equation 1. For the 83% of cases with $\bar{a} \in \{0, 1\}$ the two agree exactly. Elsewhere the standardised differences have variance 0.97 and exceed 1.96 in absolute value in 3.9% of cases (nominal 5%); the largest absolute difference is 0.125. The variance identity of Proposition 2(b) holds up to sampling error, and the closed form's per-record variance is 64% of that of a single physical draw.

Table 9: Contrasts against IID (points, 95% CI).

Model	Contrast	ΔA	ΔH	ΔU	ΔR
Qwen3-8B	Diagnostic−i.i.d., $K=1$	+25.0 [+17.6, +32.8]	+6.0 [+3.4, +9.1]	+19.0 [+12.0, +25.8]	+2.0 [−6.0, +9.4]
	Binary−i.i.d., $K=1$	−1.9 [−6.0, +2.3]	−0.4 [−2.1, +1.3]	−1.5 [−5.2, +2.5]	−0.3 [−7.2, +6.7]
	Diagnostic−i.i.d., $K=8$	+22.5 [+15.8, +30.2]	+2.7 [−1.2, +6.5]	+19.9 [+13.2, +27.1]	−6.0 [−12.9, +0.4]
	Binary−i.i.d., $K=8$	−3.6 [−9.2, +2.5]	−5.1 [−8.0, −2.3]	+1.6 [−3.8, +7.1]	−8.9 [−14.5, −3.5]
MiniCPM4.1-8B	Diagnostic−i.i.d., $K=1$	+10.5 [+3.7, +17.3]	+6.4 [+3.6, +9.5]	+4.1 [−2.0, +10.2]	+11.1 [+5.2, +17.5]
	Binary−i.i.d., $K=1$	−1.0 [−7.1, +5.0]	+2.2 [−0.3, +4.7]	−3.1 [−8.9, +2.4]	+6.0 [+0.3, +12.6]
	Diagnostic−i.i.d., $K=8$	+16.8 [+9.9, +23.8]	+4.5 [−0.6, +9.5]	+12.3 [+6.3, +18.6]	+0.5 [−6.1, +7.0]
	Binary−i.i.d., $K=8$	+2.2 [−3.1, +7.2]	−4.8 [−9.3, −0.1]	+7.0 [+1.4, +12.4]	−8.1 [−14.7, −1.4]

Table 10: Audit-specific acceptance Spec by audit size (points, 95% CI).

Model	K	n_v=1	n_v=2	n_v=4	n_v=8
Qwen3-8B	Diagnostic, K=1	+4.0 [+1.9, +6.3]	+3.1 [+1.6, +4.7]	+2.1 [+0.8, +3.6]	+2.6 [+0.6, +5.0]
	Binary, K=1	−0.3 [−1.3, +0.8]	+0.3 [−0.3, +1.1]	+0.0 [−1.0, +1.1]	+0.2 [−0.3, +0.9]
	i.i.d., K=1	+0.7 [−0.5, +2.0]	+0.2 [−1.4, +1.7]	−0.5 [−1.7, +0.7]	+0.5 [−0.5, +1.9]
	Diagnostic, K=8	+9.1 [+6.4, +11.8]	+6.2 [+4.2, +8.6]	+5.1 [+3.0, +7.6]	+4.6 [+2.3, +7.4]
	Binary, K=8	+1.9 [+0.1, +3.6]	+1.9 [+0.2, +3.8]	+2.0 [+0.6, +3.9]	+0.5 [−0.4, +1.5]
	i.i.d., K=8	+3.6 [+2.0, +5.6]	+6.9 [+4.0, +10.3]	+8.8 [+6.0, +11.8]	+7.9 [+5.1, +11.0]
MiniCPM4.1-8B	Diagnostic, K=1	+3.8 [+0.8, +7.3]	+4.8 [+1.9, +8.9]	+1.2 [−0.3, +2.9]	+1.9 [−0.8, +5.4]
	Binary, K=1	−1.0 [−2.8, +0.7]	−1.8 [−4.1, +0.2]	+0.2 [−2.0, +2.3]	+0.5 [−1.9, +2.6]
	i.i.d., K=1	−0.1 [−1.2, +0.9]	+0.5 [−0.5, +1.9]	−0.5 [−1.9, +0.6]	−0.3 [−2.1, +1.2]
	Diagnostic, K=8	+9.1 [+4.6, +13.8]	+12.5 [+7.2, +18.9]	+5.9 [+2.7, +9.8]	+5.1 [+1.8, +9.0]
	Binary, K=8	+1.3 [−1.0, +4.5]	+2.3 [−0.6, +5.1]	+1.9 [+0.1, +3.8]	+2.2 [−0.0, +4.5]
	i.i.d., K=8	+3.4 [+1.0, +6.9]	+7.7 [+3.8, +12.2]	+7.1 [+3.4, +11.1]	+7.6 [+4.2, +11.8]

G AUDIT-SPECIFIC ACCEPTANCE BY HORIZON (POST HOC)

H LITERAL-COPYING CHECK

For every one-round output of the four Qwen3-8B runs that passed its audit and failed at least one hidden test, we read the runner's literal flag: the code contains the representation (at least 3 characters) of an input or expected value of an audit test that occurs neither in the task prompt nor as a constant of the first attempt, or branches on such a value. The flag fires for 22/242 diagnostic outputs, 2/22 of those with $\bar{a} < 0.2$ (the outputs that carry most of the audit-specific acceptance), 12/107 IID outputs and 16/97 binary outputs. The detector is coarse (for IID and BIN, constants absent from the first attempt can match by chance), so these rates bound explicit copying rather than measure it; an inspection by a judge model and by humans is left for future work.

I PREREGISTRATION TEXTS

The preregistrations are supplied verbatim in the supplementary material, with time stamps, dated revisions and analysis scripts: E2 (two revisions, both written before E2 outputs existed), E3, E4, E5 (three revisions, all before E5 outputs existed), E6 (two revisions, both before E6 outputs existed; the second concerns only process launching), E7, E8, E9 and E13. Each revision records whether any outcome had been observed. The equal-budget, horizon and literal-copying analyses (Appendices E, G, H) and the physical placebo check were specified after the corresponding outputs existed and are reported as secondary.

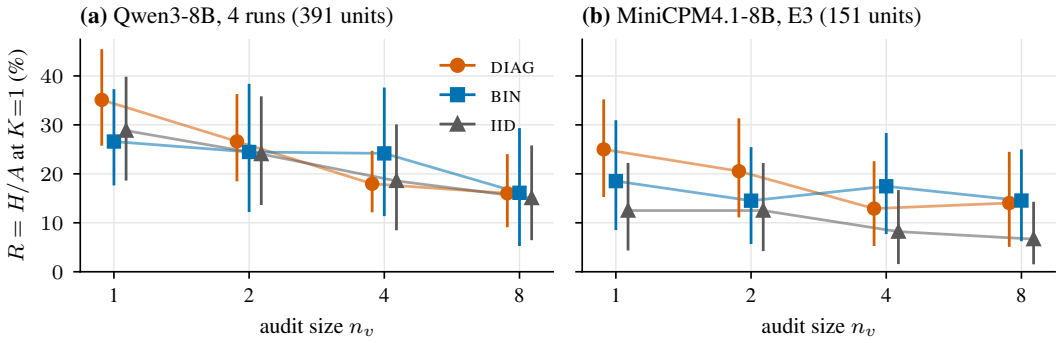


Figure 4: False-acceptance share R at $K=1$ versus audit size on a common-support population: units (same run, task and prompt) whose first attempt was rejected at all four audit sizes. These are persistently failing, harder-than-average cases. The full-population curves (population varies with n_v) are in Table 12 and show the same ordering. 95% CIs.

Table 11: R (%) at $K=1$ on the population fixed across audit sizes, with 95% CIs.

Model	Policy	$n_v=1$	$n_v=2$	$n_v=4$	$n_v=8$
Qwen3-8B (391 units)	Diagnostic	35.1 [25.7, 45.5]	26.6 [18.5, 36.3]	18.0 [12.1, 24.7]	16.0 [9.1, 24.0]
	Binary	26.6 [17.6, 37.3]	24.5 [12.2, 38.4]	24.2 [11.4, 37.6]	16.1 [5.3, 29.3]
	i.i.d.	28.8 [18.6, 39.8]	24.0 [13.6, 35.8]	18.6 [8.5, 30.1]	15.1 [6.4, 25.8]
MiniCPM4.1-8B (151 units)	Diagnostic	25.0 [15.3, 35.2]	20.5 [11.1, 31.3]	12.9 [5.3, 22.6]	14.0 [5.1, 24.5]
	Binary	18.5 [8.5, 31.0]	14.5 [5.7, 25.5]	17.5 [7.7, 28.3]	14.5 [6.2, 25.0]
	i.i.d.	12.5 [4.3, 22.2]	12.5 [4.2, 22.2]	8.2 [1.6, 16.7]	6.7 [1.5, 14.3]

Table 12: R (%) at $K=1$ by audit size on the full activation population (population varies with n_v), and DIAG−IID contrasts.

Model	n_v	R_{diag}	R_{bin}	R_{iid}	ΔH (diag−iid)	ΔR (diag−iid)
Qwen3-8B	1 ($n = 446$)	31.1	24.5	27.0	+8.3 [+3.8, +13.3]	+4.2 [−5.9, +14.5]
	2 ($n = 530$)	26.3	23.6	22.9	+6.8 [+2.9, +11.0]	+3.4 [−6.8, +14.0]
	4 ($n = 600$)	17.6	20.7	15.2	+5.3 [+2.2, +8.6]	+2.5 [−6.3, +10.6]
	8 ($n = 664$)	16.7	12.0	16.3	+4.5 [+1.7, +7.6]	+0.4 [−6.8, +7.3]
MiniCPM4.1-8B	1 ($n = 160$)	26.7	16.9	11.9	+9.4 [+3.8, +15.2]	+14.8 [+3.9, +25.1]
	2 ($n = 181$)	22.6	13.4	11.6	+7.2 [+1.7, +12.6]	+11.0 [+0.8, +21.3]
	4 ($n = 192$)	18.1	16.5	8.0	+5.7 [+2.0, +10.2]	+10.1 [+2.5, +18.4]
	8 ($n = 200$)	13.8	14.1	5.6	+4.0 [+0.5, +7.9]	+8.2 [+0.0, +16.5]

Table 13: By system prompt (points, 95% CI).

Model	System prompt	K	Spec diag	ΔSpec diag−iid	ΔH diag−iid
Qwen3-8B	correctness	1	+2.8 [+1.8, +3.9]	+3.1 [+1.9, +4.4]	+7.4 [+4.4, +10.5]
		8	+5.9 [+4.3, +7.6]	−1.2 [−3.8, +1.6]	+3.2 [−1.1, +7.6]
	maximise acceptance	1	+2.9 [+1.7, +4.4]	+2.3 [+0.8, +3.9]	+4.7 [+1.2, +8.4]
		8	+6.2 [+4.5, +7.9]	−0.9 [−3.9, +2.0]	+2.2 [−2.1, +6.4]
MiniCPM4.1-8B	correctness	1	+3.2 [+1.3, +5.4]	+3.0 [+0.7, +5.4]	+7.2 [+4.1, +10.4]
		8	+7.0 [+3.9, +10.4]	+0.9 [−4.4, +5.6]	+3.2 [−4.1, +10.0]
	maximise acceptance	1	+2.5 [+1.1, +4.1]	+3.0 [+1.5, +4.7]	+5.7 [+2.1, +9.6]
		8	+9.0 [+5.6, +12.8]	+2.0 [−3.5, +7.0]	+5.7 [−0.5, +12.3]

Table 14: False-acceptance share (%) with a second stage of m fresh hidden tests (ratios of expected counts, Corollary 2; same activation set as before).

Extra fresh tests m	R at $K = 1$ (diag / bin / iid)			R at $K = 8$ (diag / bin / iid)		
Qwen3-8B						
$m = 0$	22.2	19.9	20.2	24.9	22.0	30.9
$m = 1$	19.1	17.8	17.5	21.1	19.4	24.8
$m = 2$	17.2	16.5	16.0	19.0	17.6	21.5
$m = 4$	14.8	14.8	13.9	16.3	15.4	17.5
MiniCPM4.1-8B						
$m = 0$	20.3	15.2	9.2	24.9	16.3	24.4
$m = 1$	16.7	14.1	8.4	19.1	14.4	19.2
$m = 2$	15.1	13.3	7.9	16.8	13.2	16.7
$m = 4$	13.2	12.2	7.1	14.4	11.7	13.9

Table 15: Audit-specific acceptance by feedback content (E1, n_v=1, n=113 activation records; points, 95% CIs).

Feedback content	Spec, K=1	Spec, K=8
Full diagnostic	+6.5 [+1.8, +12.5]	+14.5 [+8.1, +22.4]
Input and expected	+9.1 [+4.3, +14.6]	+14.0 [+7.9, +21.0]
Input only	+1.7 [−0.9, +4.4]	+6.5 [+3.0, +10.8]
Failure count	+0.3 [−1.5, +1.8]	+3.5 [+1.4, +6.1]
Binary	+1.0 [−0.4, +2.8]	+3.9 [+1.2, +7.8]
i.i.d.	+0.6 [−1.1, +2.2]	+4.0 [+1.5, +7.4]

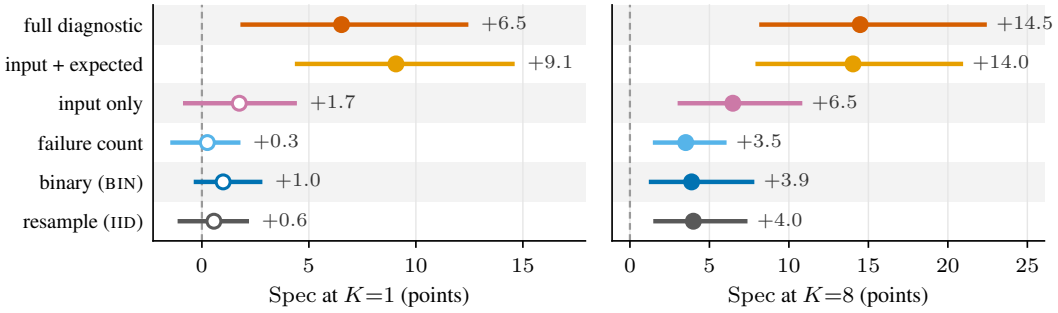


Figure 5: Audit-specific acceptance by feedback content (E1 run, n_v=1, re-executed hidden tests; points, 95% CIs; hollow: CI contains 0). Revealing the expected output is what makes a single retry audit-specific; over eight retries every variant, including IID, accumulates some.

Table 16: False-acceptance share R under a fixed budget of 8 tests per task, all Qwen3-8B units.

K	Policy	R (%) for n_v+m = 1+7 / 2+6 / 4+4 / 8+0				ΔR, 4+4 vs. 8+0
1	Diagnostic	5.5	5.6	5.8	6.0	−0.2 [−0.9, +0.4]
	Binary	5.3	5.3	5.6	5.4	+0.2 [−0.4, +0.8]
	i.i.d.	5.3	5.3	5.4	5.5	−0.1 [−0.7, +0.5]
8	Diagnostic	5.8	5.9	6.3	6.5	−0.2 [−0.9, +0.5]
	Binary	5.6	5.7	5.8	5.7	+0.1 [−0.5, +0.7]
	i.i.d.	5.5	5.6	6.0	6.9	−0.9 [−1.6, −0.2]

Table 17: Spec by number of retries K, pooled over n_v and prompts (points, 95% task-cluster CIs).

Policy	K=1	K=2	K=4	K=8
Qwen3-8B (4 runs)				
Diagnostic	+2.9 [+1.8, +4.1]	+3.7 [+2.5, +5.1]	+4.8 [+3.4, +6.3]	+6.0 [+4.4, +7.6]
Binary	+0.1 [−0.3, +0.6]	+0.5 [−0.1, +1.2]	+1.0 [+0.3, +1.9]	+1.5 [+0.8, +2.4]
i.i.d.	+0.2 [−0.2, +0.6]	+2.2 [+1.3, +3.1]	+4.8 [+3.1, +6.5]	+7.1 [+4.9, +9.4]
MiniCPM (E3)				
Diagnostic	+2.9 [+1.5, +4.5]	+4.5 [+2.8, +6.4]	+7.2 [+4.6, +10.0]	+8.0 [+5.2, +11.1]
Binary	−0.5 [−1.5, +0.6]	+0.9 [+0.2, +1.8]	+1.6 [+0.6, +2.7]	+2.0 [+0.7, +3.3]
i.i.d.	−0.1 [−0.6, +0.4]	+1.9 [+0.6, +3.5]	+4.2 [+2.2, +6.8]	+6.6 [+3.9, +9.7]