POISONING THE PROCESS, NOT THE ANSWER: RELEASE-TIME ANTI-DISTILLATION FOR REASONING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frontier reasoning models are routinely distilled by third parties who collect prompt–reasoning–answer triples through model queries and use them to fine-tune student models. We investigate whether model owners can preserve answer utility for ordinary users while making released outputs ineffective, or even harmful, as distillation data. Our approach exploits an asymmetry in how these outputs are consumed: ordinary users primarily rely on final answers, whereas distillers also learn from reasoning traces. We propose *Answer-Preserving Trace Poisoning* (APTP), a release-time transformation that rewrites only the visible reasoning trace while keeping the final answer bit-for-bit identical. We instantiate three mechanisms targeting distinct aspects of student learning: (i) *information removal*, which reduces the reasoning signal through hidden, summarized, or hollow traces; (ii) *gradient shortcut*, which introduces an early natural-language "intuition" that reveals the answer, allowing students to minimize answer loss without learning the underlying reasoning; and (iii) *gradient conflict*, which constructs a coherent decoy trace supporting a different answer while retaining the correct final answer. We also introduce *AntiDistill-Bench*, a benchmark of 16.7k problems with paired clean and poisoned reasoning traces, together with an evaluation protocol that assesses utility preservation, the imperceptibility of trace modifications to casual readers and LLM judges, and student distillation performance as a function of training-data scale.

1 INTRODUCTION

Chain-of-thought traces have become the most valuable training signal in the open model ecosystem: fine-tuning a small base model on a few hundred thousand teacher traces yields a competitive reasoner (DeepSeek-AI, 2025; Muennighoff et al., 2025; Guha et al., 2025; Hugging Face, 2025). For a provider, every API response is therefore also a free training example for a competitor. Existing responses are blunt—hide the trace behind a summary, rate-limit, or watermark—and none of them stops a determined distiller; the first degrades the product for everyone.

We start from an asymmetry between the two audiences of a response: users consume the final answer and rarely read the reasoning panel, distillers extract most of their value from it. A provider thus has a lever that costs users nothing: change what the trace *teaches* without changing what the answer *says*. We call the resulting family of release-time transforms *Answer-Preserving Trace Poisoning* (APTP). The answer is passed through unchanged, so correctness filtering—the standard first step of every distillation recipe (DeepSeek-AI, 2025; Guha et al., 2025)—cannot remove the poison; only the visible reasoning is rewritten (Figure 1).

Which rewrite? We decompose the student's fine-tuning loss into a trace term and an answer term and identify three ways to break their coupling (§4): remove the information that lets the answer be derived (*hidden*, *summary*, *hollow*); give the answer term a spurious shortcut (an *anchor*: an early "intuition" that already states the answer); or make the two terms *conflict* (a *decoy*: a fluent, full-length trace whose working supports a different value, ending in the correct answer). The decoy is the interesting case: it is the only variant that keeps a full-length trace on the screen.

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.

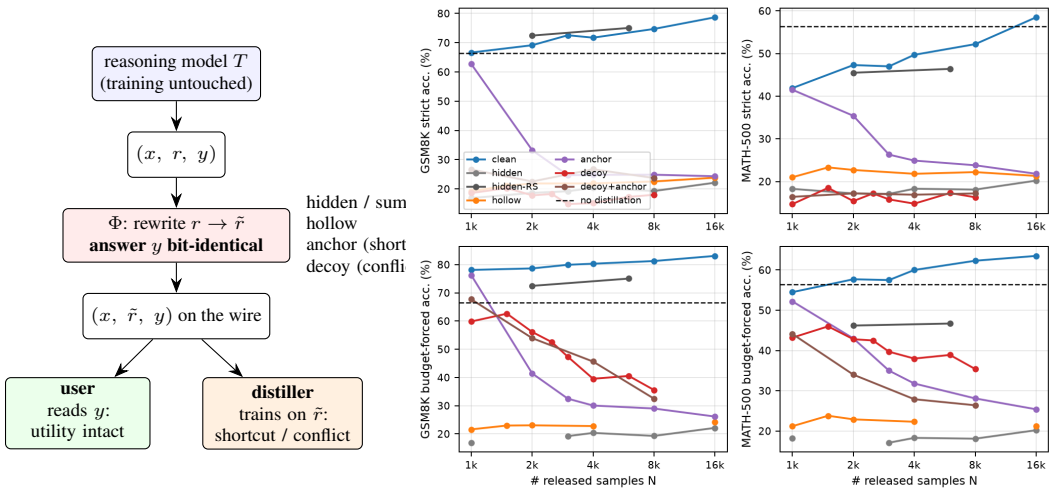

Figure 1: **Left:** release-time anti-distillation. Φ sits between an already-trained reasoning model and the wire, rewrites only the visible trace and passes the answer through unchanged, so correctness filtering cannot remove the poison. **Right:** distillation-scaling curves for a Qwen3-1.7B-Base student (strict accuracy, top; budget-forced, bottom) as a function of the number of released samples. Clean traces rise; anchor (shortcut) and decoy (conflict) traces stay at or below the answer-only floor and fall with more data. Dashed: the un-distilled base model.

Building decoys at scale is harder than it looks: *on-policy* poisoning fails because the teacher notices and repairs an injected slip 97–100% of the time, and whole-trace rewriting by a non-reasoning model lets the correct value resurface. What works is an *edit-script* protocol (§4.2) in which a thinking editor emits per-sentence replacements that we apply and verify programmatically.

We also release AntiDistill-Bench (§5): 16,691 verified teacher traces with 51 release variants, a science multiple-choice domain, and a protocol that measures distillation-scaling curves, imperceptibility and robustness to adaptive distillers. Ten students from five families (Qwen3 1.7B/4B/8B/14B, Qwen2.5-1.5B, Llama-3.2-3B, Llama-3.1-8B, Mistral-7B, GLM-4-9B) are distilled from six teachers (Qwen3-14B / 30B-A3B / 32B / 235B-A22B, gpt-oss-120b, Phi-4-reasoning-plus) with SFT, rejection-sampling fine-tuning (RFT) and GRPO.

Our contributions are: (1) A formalisation of release-time anti-distillation with the answer-preserving constraint and a threat model that includes adaptive distillers. (2) The APTP menu with a gradient-level account of three mechanisms, and a scalable decoy generator; negative results on on-policy poisoning. (3) AntiDistill-Bench and an evaluation protocol covering utility, imperceptibility and distillation scaling. (4) Extensive experiments, including mechanism analysis and adaptive-distiller robustness.

2 Related Work

Model extraction and its defences. Query-access extraction of classifiers (Tramèr et al., 2016), imitation attacks on NLP APIs (Krishna et al., 2020; Wallace et al., 2020) and partial extraction of production LLMs (Carlini et al., 2024b) established that black-box access leaks capability. Defences perturb the returned prediction: prediction poisoning (Orekondy et al., 2020), adaptive misinformation (Kariyappa & Qureshi, 2020), query-pattern detection (Juuti et al., 2019), watermarks (Kirchenbauer et al., 2023), anti-distillation sampling (Savani et al., 2025), and the *nasty teacher* (Ma et al., 2021), which retrains the provider's model to mislead students. All act on the prediction the user consumes; APTP exploits the trace/answer asymmetry instead. **Reasoning distillation.** Fine-tuning on teacher traces is the dominant recipe for open reasoning models (DeepSeek-AI, 2025; Muennighoff et al., 2025; Ye et al., 2025; NovaSky Team, 2025; Guha et al., 2025; Hugging Face, 2025), and a few thousand traces suffice (Muennighoff et al., 2025; Ye et al., 2025). **Data poisoning.** Clean-label and gradient-matching poisons (Shafahi et al., 2018; Geiping et al., 2021), unlearnable examples (Huang

Table 1: The APTP menu. All variants leave the answer unchanged.

Variant	Released trace $\tilde{r}$	Mechanism
clean	teacher trace r	none (distiller's upper bound)
hidden / summary / hollow	empty / 3–5 sentences / narrative with all intermediate values removed	information removal
anchor	r + one early sentence stating the answer as an "intuition"	gradient shortcut
decoy	rewrite of r with one confident load-bearing slip propagated coherently; neutral wrap-up + correct y	gradient conflict
decoy+anchor	both	shortcut + conflict

et al., 2021), web-scale and instruction-tuning poisoning (Carlini et al., 2024a; Wan et al., 2023), radioactive data (Sablayrolles et al., 2020), model collapse (Shumailov et al., 2024); our anchor is a clean-label shortcut and our decoy gradient-conflicting noise under the constraint that the label y must not change. **CoT faithfulness.** Chains of thought can be unfaithful (Turpin et al., 2023; Lanham et al., 2023; Chen et al., 2025) and optimisation pressure induces obfuscation (Baker et al., 2025); APTP manufactures unfaithful traces at release time, and our slip-and-continue result adds evidence on the robustness of self-verification in long CoT (Yeo et al., 2025).

3 Problem Setting

Actors A *provider* owns a reasoning model T whose pretraining, SFT and RL are finished and untouched. At deployment T answers a query x with a trace r and an answer y. *Users* consume y; *distillers* query at scale, collect (x, r, y) and fine-tune a student S.

Release transform The provider inserts $\Phi : (x, r, y) \mapsto (\tilde{r}, y)$ between the model and the wire: the answer is bit-identical and only the visible trace is replaced. Nothing about training changes, so Φ can be added to an already-trained model at release time and updated without retraining.

Constraints (C1, utility) y is unchanged, so accuracy and every downstream use of the answer are preserved by construction. (C2, imperceptibility) $\tilde{r}$ should read as a plausible trace to a casual reader and not be trivially separable from r by cheap filters. (C3, anti-distillation) for $\tilde{D} = \{(x_i, \tilde{r}_i, y_i)\}$, the student's accuracy $A(S; \tilde{D})$ should be no better than what the distiller obtains from answers alone and should not increase—ideally decrease—with $|\tilde{D}|$.

Threat model The distiller has black-box access, unlimited queries, reference answers for a verification subset, and standard tooling: correctness filtering, LLM judges, paraphrasing, mixing with other providers' data, and RL with verifiable rewards. We do not assume the distiller is unaware of the defence.

4 Answer-Preserving Trace Poisoning

Write the student's SFT loss on one triple as $\mathcal{L} = \mathcal{L}_{\text{trace}}(\tilde{r} \mid x) + \mathcal{L}_{\text{ans}}(y \mid x, \tilde{r})$. Distillation works because minimising $\mathcal{L}_{\text{trace}}$ teaches the *procedure* that makes y derivable, after which $\mathcal{L}_{\text{ans}}$ only has to read the result off the trace. Φ can break this coupling in three ways (Table 1).

Information removal If $\tilde{r}$ carries none of the procedure, the answer head must learn y from x directly—exactly what the distiller could do without the trace.

Gradient shortcut (anchor) With an anchor, y is copyable from the trace, so $\mathcal{L}_{\text{ans}}$ is satisfied by an attention shortcut and stops coupling the answer to the computation. At test time the student produces its own anchor—a guess—and has learned to trust it. The trace is otherwise verbatim, so this variant is maximally natural, but also the easiest to strip.

Gradient conflict (decoy) A decoy keeps voice, structure and length but contains one plausible, confidently stated slip in a load-bearing step, propagated so the working leads to a decoy value $y' \neq y$, while the trace ends with a neutral wrap-up and the correct y: the trace term pushes the student to reproduce erroneous computation, the answer term to output y regardless.

4.1 WHY THE TRACE GRADIENT CAN OPPOSE THE ANSWER GRADIENT

We make the three mechanisms precise with a minimal model of what a student learns from a trace (full derivations in Appendix E). Let the student's answer head read a *value* $v_\theta(x, \tilde{r}) \in \mathbb{R}$ off the trace and score answers by $p_\theta(a \mid x, \tilde{r}) \propto \exp(-\frac{1}{2\sigma^2}(a - v_\theta)^2)$, and let $a(\tilde{r})$ denote the value that the trace's working actually supports (its last computed result). A *procedure-following* student is one with $v_\theta \approx a(\tilde{r})$; distillation from clean traces makes it so, because there $a(r) = y$.

Population objective. Over a release distribution $\tilde{\mathcal{D}}$ the SFT objective is

$$\mathcal{L}(\theta) = \underbrace{\mathbb{E}\left[-\log p_\theta(\tilde{r} \mid x)\right]}_{\mathcal{L}_{\text{trace}}} + \underbrace{\mathbb{E}\left[-\log p_\theta(y \mid x, \tilde{r})\right]}_{\mathcal{L}_{\text{ans}}}, \qquad \nabla_\theta \mathcal{L}_{\text{ans}} = \mathbb{E}\Big[\tfrac{1}{\sigma^2}\,(v_\theta - y)\,\nabla_\theta v_\theta\Big]. \quad (1)$$

The answer gradient always pulls the read-off value toward the released answer y. What the trace gradient pulls toward depends on Φ.

Information removal. If $\tilde{r} \perp y \mid x$, then $\arg\min \mathcal{L}_{\text{ans}}$ over all conditionals is $p^\star(y \mid x, \tilde{r}) = p(y \mid x)$ (Appendix E, Prop. 1): the trace contributes no gradient toward answer accuracy beyond what x provides, so the student cannot beat the answer-only floor. This is the natural floor of APTP; it is also user-visible.

Gradient conflict (decoy). A decoy is generated so that $a(\tilde{r}) = y' \neq y$ while the released answer stays y. For a procedure-following student, fitting the trace means reproducing the working that leads to y', i.e. $\mathcal{L}_{\text{trace}}$ pulls v_θ toward $a(\tilde{r}) = y'$; $\mathcal{L}_{\text{ans}}$ pulls it toward y. Along the value direction the two gradients are

$$g_{\text{trace}} \propto (v_\theta - y')\,\nabla_\theta v_\theta, \qquad g_{\text{ans}} \propto (v_\theta - y)\,\nabla_\theta v_\theta, \qquad \langle g_{\text{trace}}, g_{\text{ans}} \rangle \propto (v_\theta - y')(v_\theta - y)\,\|\nabla_\theta v_\theta\|^2 < 0 \quad (2)$$

whenever v_θ lies between y' and y—which is exactly where a student that has partly followed the working sits. The population minimiser of the sum is $v^\star = \frac{1}{2}(y + y')$ (weighted by the two losses' scales), a value that is *neither* the trace's nor the answer's: the joint minimiser is a student whose answer is decoupled from its trace and whose procedure is corrupted (Prop. 2). Two consequences are testable. (i) For a student that has learned to read answers off traces, $-\log p_\theta(y \mid x, \tilde{r})$ is larger for a decoy than for a clean trace by $\frac{1}{2\sigma^2}(y - y')^2$ to first order—we measure $+0.30$ nats (§7). (ii) Because the conflict lives in the value pathway and not in the bulk of the parameters, whole-model gradient cosines stay near zero even when the value-direction gradients are anti-parallel; we observe $|\cos| \leq 0.03$. The opposition is local and directional, not a reversal of the whole gradient.

Gradient shortcut (anchor). Now let the answer head have two pathways, $v_\theta = w_c\,\hat{v}_{\text{comp}} + w_a\,\hat{v}_{\text{anchor}}$: the value computed by the working and the value stated in an early "intuition" sentence. On anchored data $\hat{v}_{\text{anchor}} = y$ exactly, so the gradient on the copy weight, $\partial \mathcal{L}_{\text{ans}}/\partial w_a \propto (v_\theta - y)\,y$, drives $w_a \to 1$ and, as $v_\theta \to y$, the residual $(v_\theta - y)$ that also drives w_c vanishes: the computation pathway is *starved* (Pezeshki et al., 2021) (Prop. 3). At test time the student writes its own anchor—a guess with accuracy q—and copies it, so its accuracy is bounded by q rather than by its computation. The shortcut exists only if the anchor is *cheaper* than computing and *uniform* enough to be learned in N samples: paraphrasing the sentence, or placing it after a passage that already carries computation (so that $\hat{v}_{\text{anchor}}$ is itself accurate), removes the effect, as §6 and §7 show.

Predictions. Hidden traces: the student converges to the answer-only floor. Decoys: below it, with a gap that grows with N because each released trace adds thousands of working tokens pulling toward y' but only one answer token pulling toward y. Anchors: to its own guessing accuracy q. All three hold empirically (Figure 1).

4.2 GENERATING DECOYS AT SCALE

Negative result 1: on-policy poisoning self-heals The most natural generator corrupts one early paragraph and lets T continue. With Qwen3-14B the teacher notices and repairs the slip in 130/134

cases (97%); anti-verification steering ("take previously computed values as given") reduces recovery only to 92–98%. Long-CoT self-verification is robust to prefix corruption, so poisoning must be *off-policy*.

Negative result 2: full rewriting by a non-reasoning model leaks Asking Qwen3-32B (non-thinking) to rewrite whole traces leaves the correct value in 50–75% of long traces and adds explicit self-corrections; thinking mode is reliable only below ∼1.5k tokens. Frontier APIs propagate well (correct-value ratio ≈0.1) but are rate-limited.

Edit-script protocol We split the trace into numbered sentences; a thinking editor (Qwen3-14B) plans the slip and emits per-sentence replacements or deletions covering every dependent sentence; we apply the edits programmatically, verify with regular expressions (correct value absent, decoy value present, no added self-correction phrases), run a repair pass on residual leaks, and finally perform *ending surgery*: any concluding sentence that states the decoy total is removed and a neutral wrap-up plus the correct answer is appended. Yield is 77% (12,867/16,691); 8,869 traces pass a strict tail-leak filter; length is 78% of the original. The editor emits 2.4× the teacher's tokens per trace (4.5 s per trace on one H200).

5 ANTIDISTILL-BENCH

Prompts come from GSM8K-train (7,130), MATH-train (5,571) and DeepScaleR (3,990), decontaminated against all test sets. The teacher Qwen3-14B (thinking) solves 19,185/26,955 prompts (71%); we keep 16,691 correct traces ≤6,144 tokens (mean 2,610). Each problem ships with clean, hidden, summary, hollow, anchor, decoy (strict/loose), decoy+anchor, external-rewriter, paraphrased, judge-passed/flagged and conceptual-slip decoys, a second teacher's traces (Qwen3-8B), answer-only self-distillation sets, mixtures, a judge-in-the-loop release, and five further teachers' (Qwen3-30B-A3B, -32B, -235B-A22B, gpt-oss-120b, Phi-4-reasoning-plus) releases on a shared 9k-prompt subset (5.1–6.4k kept traces each); counts in Appendix A. A science multiple-choice domain (MMLU-Pro physics/chemistry/biology/engineering/economics, 2,716 traces, 1,387 decoys; ARC-Challenge as OOD evaluation) tests transfer beyond arithmetic. **Protocol.** Students are fine-tuned on the first $N \in \{1k, \ldots, 16k\}$ samples (3 epochs, lr 10^{-5}, cosine, fp32 master weights); evaluation is zero-shot with strict boxed-answer matching (avg@2 on GSM8K/MATH-500, avg@8 on AMC23/AIME24/25, 12k tokens) and, because small students imitate teacher verbosity before they learn to stop (a clean 1.7B student exceeds 12k tokens on 31% of GSM8K problems yet is 93% correct when it finishes), also under s1-style budget forcing (Muennighoff et al., 2025). Strict accuracy measures usable behaviour; forced accuracy measures what the student believes; truncation rates are in the Appendix. Imperceptibility probes: judge error-detection, head-and-tail glance, pairwise naturalness, reference-LM perplexity, n-gram filters, answer-token conflict.

6 EXPERIMENTS

Setup Unless stated otherwise the student is Qwen3-1.7B-Base fine-tuned on the first N released samples; we report strict accuracy and budget-forced accuracy as strict / forced. The un-distilled base scores 66.5 / 56.4 / 29.1 (GSM8K / MATH-500 / AMC23). Seed standard deviations are 0.2–1.8 points (Appendix C).

Clean traces transfer; every APTP variant removes the transfer (Table 2, Figure 1) Clean distillation rises monotonically (GSM8K 66.6→78.7, MATH-500 41.9→58.5 from N=1k to 16k). On MATH-500 the strict score of a clean student stays *below* the base until N=16k: the student inherits the teacher's verbosity before it learns to stop (35–45% of its MATH-500 generations hit the 12k budget), whereas its forced accuracy exceeds the base from N=8k (62.3 vs 57.1). This is the reason we report both metrics. Answer-only self-distillation (hidden-RS) is a strong baseline on GSM8K (75.1) but not on MATH-500/AMC (46.4 / 21.2): the trace's marginal value grows with difficulty and, as shown below, with student size. Information-removal variants sit at the direct-answer floor (≈20%) and are flat in N. The anchor *falls* with data: 62.7→33.2→24.8→24.9→24.3 on GSM8K, 40 points below the base at $N \geq$4k. The decoy pins strict accuracy at 15–18% for all N and, under budget forcing, falls from 59.9 to 35.6; two thirds of its damage is non-termination (72–91% of generations

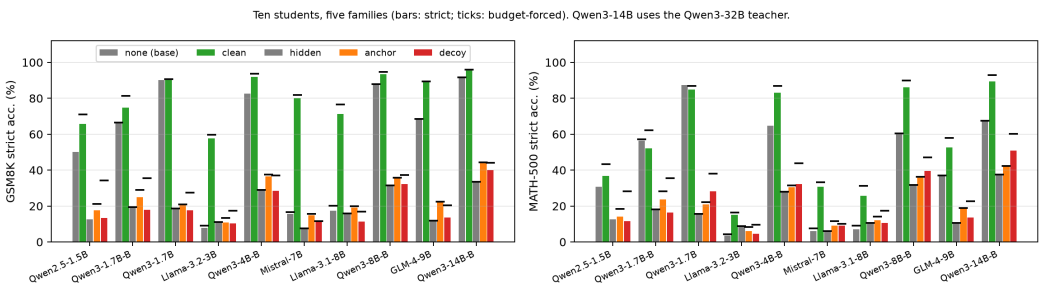

Figure 2: Ten students from five families after distillation on 8k released samples (3k for the 32B-teacher runs). Bars: strict accuracy; ticks: budget-forced. Grey = un-distilled base; clean traces help every student, hidden traces sit at the answer-only floor, anchor and decoy traces put every student at or below that floor.

hit the 12k budget) and one third is corrupted computation (accuracy *when the student finishes* drops from 34% at 2k to 20% at 8k). Releasing the full answer text instead of the boxed answer adds only $\sim$1 point for the distiller, and extending the decoy release past the strict pool with looser (tail-leaking) decoys does not help either: 12k loose decoys give 20.6 / 19.6, the same as 8k.

Mixtures are dominated by the poison Replacing 25% of a clean 8k set by decoys costs two thirds of the clean gain (74.7$\rightarrow$56.5); 50% puts the student below the base (35.6). Half the data from an honest second teacher (Qwen3-8B, 72.2 alone) does not rescue it either: decoy4k+cleanB4k reaches 39.1 against 73.0 for clean4k+cleanB4k. The 8B student loses 35 / 24 points on the same mixture (60.0 / 62.6 vs 94.6 / 86.8).

Other students (Table 3, Figure 2) The same ordering holds for every student we tried, across five model families and a 10$\times$ range of scale. The post-trained Qwen3-1.7B keeps its accuracy under clean traces (90.1 / 84.9) but collapses under decoys (38.9$\rightarrow$17.7 GSM8K from N=2k to 8k) and anchors (20.5). Qwen3-4B-Base gains 9 / 18 points from clean traces and loses 50 from decoys or anchors; Qwen3-8B-Base gains +26 MATH-500 and +32 AMC23 from clean traces (60$\rightarrow$86, 39$\rightarrow$71) and lands at 32 / 40 with decoys; its true answer-only self-distillation baseline (rejection sampling of its *own* traces against the released answers, 6k prompts) gives 86.4 / 61.7 / 33.8, i.e. nothing beyond the base, so for a capable student the entire value of the release lies in the trace; and its decoy curve also falls with data (40.1 at N=2k, 32.3 at 8k). The cross-family students are the most instructive because their un-distilled ability is low (Llama-3.2-3B 7.7, Llama-3.1-8B 17.4, Mistral-7B 15.6 GSM8K; 4-shot references in Appendix Table 9): clean traces lift them to 57.7 / 71.4 / 80.1, hidden, anchor and decoy traces leave them at 10–19—slightly *below* no trace at all—and the decoy's forced accuracy again falls with data (Llama-3.2-3B: 29.9 at 2k, 17.4 at 8k). GLM-4-9B, whose base is competent (68.5), goes to 88.9 with clean traces and to 11.8 / 13.7 with hidden / decoy traces. Qwen3-14B-Base is discussed with the teacher-size study below. Larger students retain more in absolute terms (decoy students score 32 at 8B and 40 at 14B against 18 at 1.7B), but every decoy student is far below its own base, and the gap to the clean student grows with size because the clean gain grows.

Teachers: size, architecture and family (Table 4) We repeated the release with five more teachers on the same prompts—Qwen3-30B-A3B (MoE), Qwen3-32B, Qwen3-235B-A22B (MoE, FP8), and two from other families, gpt-oss-120b (OpenAI, MoE, harmony format) and Phi-4-reasoning-plus (Microsoft, 14B)—at matched N. Within the Qwen3 family nothing changes: larger teachers are not more accurate under the 8k budget (67–73% solved), only longer; the clean gain is bounded by the *student* (14B student 85–89 MATH-500 from any Qwen teacher); decoys and anchors land in the same place with every teacher (14B student 40–48 GSM8K). Across families, trace *style* matters as much as teacher accuracy: gpt-oss-120b solves the most prompts (75%) with the tersest traces (1.6k tokens, 3 self-check phrases per 1k words vs 9–18 for Qwen; Appendix Table 11) and is the best teacher for the small student (81.4 vs 72–73 GSM8K) because the student learns to stop; verbose Phi traces teach it to ramble (73% truncation even for clean traces; forced accuracy 80.9, level with Qwen). The *decoy* carries over intact: 44.2 / 42.2 / 47.5 GSM8K for the 1.7B / 8B / 14B students with

Table 2: Qwen3-1.7B-Base student: strict / budget-forced accuracy (%) after fine-tuning on N released samples. Base = no distillation. hidden-RS = answer-only self-distillation (rejection sampling of the student's own traces against the released answers).

Release variant	N	GSM8K	MATH-500	AMC23	AIME24
none (base)	–	66.5 / 66.5	56.4 / 57.1	29.1 / 31.6	3.3 / 4.2
clean	1000	66.6 / 78.1	41.9 / 54.5	20.0 / 32.5	2.5 / 5.0
clean	8000	74.7 / 81.3	52.2 / 62.3	28.4 / 36.9	3.3 / 5.4
clean	16000	78.7 / 83.1	58.5 / 63.5	33.4 / 37.2	4.6 / 7.1
hidden-RS	6000	75.1 / 75.1	46.4 / 46.7	21.2 / 22.8	2.1 / 2.1
hidden	8000	19.3 / 19.3	18.1 / 18.1	7.2 / 7.2	0.0 / 0.0
summary	8000	22.2	22.4	8.8	1.7
hollow	8000	22.6	22.2	8.1	2.1
anchor	1000	62.7 / 76.2	41.5 / 52.2	21.6 / 31.2	2.5 / 3.8
anchor	2000	33.2 / 41.4	35.4 / 43.0	19.7 / 23.1	1.7 / 2.5
anchor	8000	24.9 / 29.0	23.8 / 28.1	10.0 / 15.0	2.5 / 3.3
anchor	16000	24.3 / 26.2	21.8 / 25.4	9.4 / 12.5	1.7 / 2.9
decoy	1000	18.6 / 59.9	14.7 / 43.2	4.7 / 24.7	0.4 / 2.9
decoy	2000	17.8 / 56.1	15.4 / 42.8	4.4 / 21.6	0.8 / 2.1
decoy	8000	18.0 / 35.6	16.3 / 35.4	7.8 / 20.9	0.8 / 2.5
decoy+anchor	8000	23.7 / 32.5	17.2 / 26.4	13.1 / 14.1	1.2 / 3.8
25% decoy in clean	8000	56.5 / 67.9	39.5 / 54.2	23.8 / 35.6	2.9 / 5.0
50% decoy in clean	8000	35.6 / 51.9	27.8 / 45.9	17.8 / 29.7	1.7 / 4.2
50% decoy + teacher-B clean	8000	39.1 / 53.6	27.5 / 43.6	15.6 / 28.1	3.8 / 4.2
50% clean + teacher-B clean	8000	73.0 / 81.5	52.6 / 62.7	31.2 / 37.8	5.0 / 7.1
judge-in-the-loop release	8000	24.1 / 26.0	22.7 / 25.9	8.1 / 10.6	1.7 / 3.8

gpt-oss traces (clean 81.4 / 94.4 / 94.7) and 45.9 / 54.6 / 62.8 with Phi traces (clean 72.6 / 87.0 / 88.7), a 26–52 point loss in every cell. The *anchor* does not: 81.0 (= clean) and 61.6 with gpt-oss traces, 67.9 and 82.4 with Phi. Appendix Table 13 shows why: anchor damage is the product of the student copying its opening guess and the guess being wrong. gpt-oss's opening paragraph already carries computation, so an anchor placed after it is copied but right 50–59% of the time (Qwen: 18–34%); Phi's long opening restatement pushes the anchor to a median offset of 500 characters and the 1.7B student never forms the habit (30%). Moving the anchor to after the first *sentence* restores the poison for the 8B student (35.6 with gpt-oss, 54.3 with Phi traces; guess right 32–33%) but not for the 1.7B student, which now emits the guess (94%) yet copies it only 12–23% of the time. The shortcut is a property of trace structure and student, to be placed after the reading but before the computation; the decoy corrupts the computation itself and needs no such care.

Distillers with RL (Table 5) Verifiable-reward post-training is the strongest recipe available to a distiller: 200 GRPO steps from the raw base reach 81.1 / 64.2, above every SFT baseline. Initialising from clean traces adds 2.6 GSM8K points; from the anchor student the base trajectory is fully recovered (82.1 / 60.7); from the decoy student it is not at this budget (66.6 / 47.3), and three rounds of rejection-sampling fine-tuning leave the decoy initialisation 19–23 points behind every other one. *The gap is a cost, not a wall.* Tripling the RL budget to 600 steps repairs most of it: the decoy initialisation reaches 82.1 / 59.8 / 32.5 against 83.7 / 66.4 / 34.4 from the base, i.e. it is still 6.6 points behind on MATH-500 but level on GSM8K, whereas the anchor initialisation is level everywhere (83.6 / 61.9). The same holds for the 8B student under RFT: from the decoy-8k student (32.3 / 39.6) three rounds recover 48.2, 60.3 and 70.5 GSM8K against 90.5 from the base, catching up on MATH-500 and AMC23 (66.7 / 46.2 vs 68.2 / 40.6) but not on GSM8K, where the decoys' arithmetic slips live; every round also has fewer correct samples to learn from (2,888 vs 3,800 prompts solved in round 1). Under an RL-capable distiller, then, our clean traces are worth a little, our poisoned traces are worth less than nothing, and the poison's residual effect is measured in the extra verifiable-reward compute needed to undo it: roughly $3\times$ the GRPO steps for the decoy, and none for the anchor.

7 Analysis

What the student learns We measure trace–answer agreement (does the boxed answer equal the last value computed in the thinking?) on the students' own generations. Clean students agree 96–99% of the time (87–92% accurate when they do); decoy students agree *100%* yet are 20% accurate—they faithfully follow a corrupted procedure; anchor students agree only 25–33% at N=8k (41–69% at

Table 3: Ten students, five families: strict / forced accuracy (%) after distillation on N=8k released samples (3k for the 32B-teacher rows, marked †). Base = un-distilled student with the same template. q17bi = post-trained Qwen3-1.7B (chat template); all other students are base models. Per-N rows and AMC/AIME columns in Appendix Table 8.

Student	Set	none (base)	clean	hidden	anchor	decoy
Qwen2.5-1.5B	GSM8K	50.1	65.8 / 71.1	12.7	17.5 / 21.2	13.4 / 34.2
	MATH-500	30.8	36.7 / 43.4	12.7	14.2 / 18.4	11.5 / 28.2
Qwen3-1.7B-Base	GSM8K	66.5 / 66.5	74.7 / 81.3	19.3 / 19.3	24.9 / 29.0	18.0 / 35.6
	MATH-500	56.4 / 57.1	52.2 / 62.3	18.1 / 18.1	23.8 / 28.1	16.3 / 35.4
Qwen3-1.7B (post-trained)	GSM8K	90.1	90.1 / 90.6	18.7 / 18.7	20.5 / 20.9	17.7 / 27.5
	MATH-500	87.5	84.9 / 86.8	15.6 / 15.6	20.9 / 22.1	28.2 / 38.0
Llama-3.2-3B	GSM8K	7.7 / 9.1	57.7 / 59.6	11.0 / 11.0	10.8 / 13.4	10.3 / 17.4
	MATH-500	3.5 / 4.4	15.2 / 16.5	8.8 / 8.8	6.0 / 8.3	4.6 / 9.5
Qwen3-4B-Base	GSM8K	82.5	91.9 / 93.6	28.9 / 28.9	36.5 / 37.6	28.6 / 37.0
	MATH-500	64.6	83.0 / 87.0	27.9 / 27.9	30.6 / 31.5	32.2 / 43.8
Mistral-7B-v0.3	GSM8K	15.6 / 16.7	80.1 / 81.8	7.7 / 7.7	14.9 / 15.5	11.4 / 11.6
	MATH-500	6.0 / 7.5	30.7 / 33.2	6.0 / 6.0	9.2 / 11.5	9.0 / 10.1
Llama-3.1-8B	GSM8K	17.4 / 20.1	71.4 / 76.4	16.0 / 16.0	19.0 / 19.9	11.4 / 17.0
	MATH-500	7.0 / 9.1	25.8 / 31.3	10.5 / 10.5	12.2 / 14.2	10.5 / 17.3
Qwen3-8B-Base	GSM8K	88.0 / 88.0	93.5 / 94.7	31.5 / 31.5	35.8 / 35.8	32.3 / 37.2
	MATH-500	60.0 / 60.4	86.2 / 90.0	31.8 / 31.8	35.7 / 36.2	39.6 / 47.0
GLM-4-9B	GSM8K	68.5 / 68.5	88.9 / 89.3	11.8 / 11.8	22.5 / 22.5	13.7 / 20.4
	MATH-500	36.9 / 37.0	52.6 / 58.0	10.6 / 10.6	18.7 / 18.8	13.5 / 22.7
Qwen3-14B-Base†	GSM8K	91.6 / 91.6	95.5 / 96.0	33.6 / 33.6	43.9 / 44.4	40.0 / 44.0
	MATH-500	67.3 / 67.6	89.3 / 93.0	37.5 / 37.5	42.0 / 42.4	50.8 / 60.2

Table 4: Teacher size, architecture and family. Students distilled from Qwen3-14B (T14, dense), Qwen3-30B-A3B (T30, MoE), Qwen3-32B (T32, dense), Qwen3-235B-A22B (T235, MoE), OpenAI gpt-oss-120b (MoE, harmony format, high reasoning effort) or Microsoft Phi-4-reasoning-plus (dense 14B) traces on the same prompts at matched N (3k, or 2.5k where a teacher's strict decoy yield is smaller). Cells: GSM8K / MATH-500 strict accuracy (%); "base" = un-distilled. Full table with forced accuracy, AMC23 and AIME24 in Appendix Table 12.

Student	Variant	T14	T30	T32	T235	gpt-oss	Phi-4
Qwen3-1.7B-Base	none (base)			66.5 / 56.4			
	clean	72.6 / 47.0	71.5 / 44.7	73.1 / 46.1	69.6 / 39.8	81.4 / 56.9	72.6 / 46.0
	hidden	19.2 / 17.1	18.1 / 17.0	17.7 / 17.7	17.1 / 16.2	16.9 / 16.5	17.0 / 17.8
	anchor	24.7 / 26.3	27.1 / 25.5	26.8 / 27.6	36.8 / 27.4	81.0 / 56.4	67.9 / 41.8
	decoy	14.8 / 15.8	20.4 / 17.2	21.5 / 19.7	10.8 / 7.5	44.2 / 41.1	45.9 / 23.8
Qwen3-8B-Base	none (base)			88.0 / 60.0			
	clean	92.5 / 83.3	93.3 / 82.4	94.2 / 84.5	91.1 / 79.4	94.4 / 85.6	87.0 / 77.0
	hidden	31.5 / 31.8	27.1 / 31.2	29.0 / 31.8	26.5 / 30.8	30.0 / 30.7	29.8 / 32.2
	anchor	37.8 / 38.4	39.1 / 36.7	37.5 / 36.7	42.1 / 42.1	61.6 / 54.2	82.4 / 69.2
	decoy	40.1 / 44.8	38.4 / 41.2	35.6 / 46.4	34.2 / 29.7	42.2 / 50.9	54.6 / 42.3
Qwen3-14B-Base	none (base)			91.6 / 67.3			
	clean	95.3 / 89.0	94.7 / 87.2	95.5 / 89.3	93.0 / 85.4	94.7 / 89.9	88.7 / 83.0
	hidden	37.1 / 37.3	35.9 / 34.8	33.6 / 37.5	34.5 / 35.5	33.9 / 34.0	34.8 / 33.2
	anchor	43.8 / 40.6	45.0 / 41.0	43.9 / 42.0	48.0 / 50.1	63.2 / 51.9	77.9 / 66.9
	decoy	40.2 / 52.8	41.5 / 48.4	40.0 / 50.8	39.7 / 40.1	47.5 / 56.9	62.8 / 48.4

2k) and are 2–3% accurate when they disagree—they copy their opening "intuition", and the habit strengthens with data. Decoy students also loop: 12–19% of their generations collapse into a repeated

Table 5: Verifiable-reward post-training by the distiller. RFT = 3 rounds of rejection-sampling fine-tuning (k=8, 4k prompts/round); GRPO = 200 or 600 steps $\times$ 32 prompts $\times$ 8 completions (1.7B only). Strict accuracy (%), GSM8K / MATH-500 / AMC23.

Student	Initialisation	RFT round 3	GRPO 200	GRPO 600
Qwen3-1.7B-Base	base	75.7 / 54.5 / 26.9	81.1 / 64.2 / 35.9	83.7 / 66.4 / 34.4
	clean-8k	71.6 / 49.5 / 29.7	83.7 / 59.3 / 35.9	–
	hidden-RS-6k	75.5 / 51.7 / 22.5	76.0 / 52.2 / 24.7	–
	anchor-8k	70.0 / 50.0 / 25.9	82.1 / 60.7 / 33.8	83.6 / 61.9 / 33.1
	decoy-8k	56.7 / 31.5 / 12.2	66.6 / 47.3 / 20.3	82.1 / 59.8 / 32.5
Qwen3-8B-Base	base	90.5 / 68.2 / 40.6	–	–
	decoy-8k	70.5 / 66.7 / 46.2	–	–

fragment at an arithmetic step (clean students $<1\%$), consistent with a next-token distribution torn between candidate results.

Answer-token conflict For a clean-distilled student the negative log-likelihood of the correct boxed value given the trace is 0.64–0.65 for clean, anchor and hollow traces but 0.96 for decoys (+0.30 nats) and 0.84 for decoy+anchor; the untrained base shows no gap (0.63–0.69). This is the first-order signature predicted by Prop. 2: the answer gradient is larger and points against the trace.

Imperceptibility (Appendix Table 6) A Qwen3-32B judge flags 1.7% of clean traces, 1.7% of anchor traces, 14% of hollow/summary traces and 75.7% of decoys; a head-and-tail glance flags 19% vs 68%; in pairwise naturalness decoys win 3% and anchors 53%. Frontier judges are sharper: GPT-5.6 detects 95% of decoys (8% false positives) and Claude Sonnet 5 93% (0%), and Sonnet's glance flags the anchor's "intuition then confirm" pattern 75% of the time. Reference-LM perplexity: clean 1.38, anchor 1.54, decoy 2.22. Decoys are therefore *not* judge-imperceptible; their value is to raise the distiller's cleaning cost and to keep poisoning whatever survives cleaning. **Cheap filters (C2).** A bag-of-n-grams logistic regression trained on 3k labelled clean/poisoned pairs separates the releases (Appendix F): AUROC 1.00 for templated anchors and v1.0 decoys, 0.99 for v1.1 decoys and for per-trace LLM paraphrases of the templates (release v1.2)—the fingerprint is semantic, not lexical—and 0.86 with the final two sentences removed; negatives from another teacher work as well (0.997–0.999), so the classifier learns the poison, not the style. Worse for the anchor, paraphrasing *removes its poison*: a student trained on 8k v1.2 anchors scores 75.5 / 52.3, the clean level, and copies its own early guess in 16% of generations against 75% for the templated anchor. The shortcut needs a uniform surface form to be learned, and the same uniformity is what a filter keys on. The decoy's poison lives in the working: it survives paraphrasing of the wrap-up (18.6 / 16.4 at N=5.5k vs 18.0 / 16.3) and a change of teacher family (§6). What filtering buys the distiller is the answer-only floor, not the trace: what passes is hollow-equivalent, and what slips through still poisons (judge-passed decoys, 30.4 vs 66–69).

Adaptive distillers (i) *Judge filtering* removes 83% of decoys; students trained on the passing 17% reach 30.4 / 68.4 (GSM8K strict/forced, N=1.5k) against 66–69 / 78–79 for clean traces—less toxic, still worse than discarding the trace (8B: 61.4, 27 points below its base). (ii) *Paraphrase* by Qwen3-14B repairs decoys (correct value present 14%$\rightarrow$74%; students recover to 53.1 / 73.7 at 1.7B and 77.9 at 8B, still 10 below base); conceptual slips are repaired just as easily because a capable rewriter re-derives the solution. (iii) *Mixing* does not dilute (§6). (iv) *Verifiable-reward RL* (Table 5) fully repairs the anchor at 200 GRPO steps (82.1 vs 81.1 from base) but not the decoy (66.6; 56.7 after three RFT rounds vs 75.7); 600 steps repair the decoy too, up to a residual MATH-500 gap (59.8 vs 66.4). The anchor is a shortcut habit that RL removes at once; the decoy corrupts the procedure and costs about three times the RL compute to undo—as Prop. 3 predicts, since RL acts on the computation pathway directly.

Judge-in-the-loop release Releasing only judge-passing decoys (17%) and hollow traces otherwise keeps the distiller at the answer-only floor (24.1 / 22.7 at N=8k; 8B student 39.4 / 44.5 vs 88.0 / 60.0 un-distilled) while exposing no decoy a strong judge would flag.

AI USE STATEMENT

Generative AI tools assisted with initial manuscript drafting, translation, and language editing. During revision, these tools also supported literature verification and consistency checks across the manuscript, implementation, and saved experimental results. References and numerical claims were verified against primary publications and project artifacts. AI tools were additionally used to monitor experimental runs, conduct scheduled checks of the server cluster, and assist with debugging the experiments in conjunction with command-line tools. The authors are responsible for the final text, claims, and research artifacts.

ETHICS STATEMENT

The released traces studied here deliberately contain misleading reasoning. Such data could harm a third party that trains on it, and presenting it as a faithful explanation could mislead users even when the final answer is correct. We do not treat unchanged answer accuracy as sufficient justification for deployment. Any distribution of the benchmark should identify the altered variants and their intended research use. We study public problem sets and do not conduct a human-subject experiment. Data and model redistribution remain subject to their original terms.

REPRODUCIBILITY STATEMENT

Section 4 specifies the release construction, and Appendix A describes training, scoring, and data selection. The supplementary tables identify sample counts, teacher choices, and alternative scoring protocols. Problem IDs link the release variants, permitting independent checks of the preserved answer field. The tables and scaling figure are generated from a frozen set of complete evaluation results; the appendix identifies the excluded evaluations and the scope of repeated-seed evidence.

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 preprint arXiv:2503.11926*, 2025.

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *IEEE Symposium on Security and Privacy*, 2024a.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model. In *International Conference on Machine Learning (ICML)*, 2024b.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jonas Geiping, Liam Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. In *International Conference on Learning Representations (ICLR)*, 2021.

Etash Guha et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *ICLR*, 2021.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1. https://github.com/huggingface/open-r1, 2025.

Mika Juuti, Sebastian Szyller, Samuel Marchal, and N. Asokan. PRADA: Protecting against DNN model stealing attacks. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2019.

Sanjay Kariyappa and Moinuddin K. Qureshi. Defending against model stealing attacks with adaptive misinformation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *ICML*, 2023.

Kalpesh Krishna, Gaurav Singh Tomar, Ankur P Parikh, Nicolas Papernot, and Mohit Iyyer. Thieves on sesame street! model extraction of bert-based apis. In *ICLR*, 2020.

Tamera Lanham et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Haoyu Ma, Tianlong Chen, Ting-Kuei Hu, Chenyu You, Xiaohui Xie, and Zhangyang Wang. Undistillable: Making a nasty teacher that CANNOT teach students. In *International Conference on Learning Representations (ICLR)*, 2021.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

NovaSky Team. Sky-T1: Train your own O1 preview model within $450. NovaSky blog, UC Berkeley Sky Computing Lab, 2025. URL https://novasky-ai.github.io/posts/sky-t1/.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Prediction poisoning: Towards defenses against dnn model stealing attacks. In *ICLR*, 2020.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *NeurIPS*, 2021.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Radioactive data: tracing through training. In *International Conference on Machine Learning (ICML)*, 2020.

Yash Savani, Asher Trockman, Zhili Feng, Yixuan Even Xu, Avi Schwarzschild, Alexander Robey, Marc Finzi, and J. Zico Kolter. Antidistillation sampling, 2025. URL https://arxiv.org/abs/2504.13146.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *NeurIPS*, 2018.

Zhihong Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

Avi Singh et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *USENIX Security*, 2016.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, 2023.

Leandro von Werra et al. TRL: Transformers reinforcement learning. https://github.com/huggingface/trl, 2020.

Eric Wallace, Mitchell Stern, and Dawn Song. Imitation attacks and defenses for black-box machine translation systems. In *EMNLP*, 2020.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *ICML*, 2023.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.

A Reproducibility and engineering notes

Training Full-parameter SFT with fp32 master weights and bf16 autocast, AdamW(0.9, 0.95), lr 10^{-5} cosine to 10%, 3 epochs, ≈128k tokens and ≤64 samples per optimizer step, gradient checkpointing; DDP for ≤4B students and FSDP (full shard, fp32 params, bf16 compute) for 8–14B students. Every rank executes the same number of forward/backward passes per step (unequal counts hang FSDP all-gathers). Checkpoints are saved as bf16 copies; converting live parameters in place corrupts DDP bucket views.

Template In Qwen3-*-Base the embeddings of `</think>` and `<|im_end|>` are essentially untrained (norm 0.37 vs 1.7 for ordinary tokens, tied LM head); small-lr SFT cannot make students emit them, so base students use a plain template (`Question:/Thinking:/Answer:` + the tokenizer's EOS) while the post-trained student keeps the official chat template.

Evaluation vLLM 0.11, temperature 0.6, top-p 0.95, 12,288 max tokens, explicit EOS stop ids (base generation configs list only `<|endoftext|>`); strict extraction requires `\boxed{}` or an explicit "final answer is"; budget forcing appends "Answer:\nThe final answer is $\boxed{" to truncated generations and lets the model emit ≤40 tokens. GLM-4-9B has an 8,192-token context, so its students are evaluated with a 7,168-token budget (all other students: 12,288); its truncation rates are therefore not comparable to the other rows.

Decoy generation Qwen3-14B thinking editor, temperature 0.6, two passes, sentence-level edit scripts; rejection rules: correct value present (regex, fraction- and comma-aware), decoy value absent, added self-correction phrases, format failures. Ending surgery removes the concluding sentences that mention the decoy value inside the final 30% of the trace.

RL attacks RFT: ReST-EM (Singh et al., 2023) with k=8 samples per prompt at 4k tokens, one correct trace kept per prompt, one SFT epoch per round. GRPO (Shao et al., 2024) via TRL 0.21 (von Werra et al., 2020): colocated vLLM, 8 generations per prompt, batch 32, lr 10^{-6}, Dr. GRPO loss, reward = correctness (+0.2 for format).

Cost of Φ Φ adds no answer latency; the decoy editor emits 2.4× the teacher's tokens per trace (1.6 passes, 4.5 s per trace on one H200), the hollow rewrite 365 tokens per trace.

Gradient alignment Whole-model cosine similarity between the gradients of $\mathcal{L}_{\text{trace}}$ and $\mathcal{L}_{\text{ans}}$ stays within ± 0.03 for all variants (clean, anchor, hollow, decoy): the conflict is local to the answer tokens (Analysis, answer-token conflict), not a global gradient direction.

B Additional results

Table 6: Imperceptibility probes (% flagged). Detect: "does the trace contain an error?"; glance: judge sees only the first and last 500 characters; pair: preference for the variant over the clean trace.

Judge	probe	clean	anchor	hollow	summary	decoy
Qwen3-32B (n=300)	detect	1.7	1.7	14.0	14.0	75.7
	glance	19.3	21.7	30.3	31.7	67.7
	pair win	–	53.0	2.3	9.7	3.3
GPT-5.6 (n=100)	detect	8.0	7.0	5.0	–	95.0
	glance	3.0	14.0	4.0	–	61.0
Claude Sonnet 5 (n=100)	detect	0.0	0.0	0.0	–	93.0
	glance	5.0	75.0	6.0	–	59.0

Table 7: Full N-grid for the Qwen3-1.7B-Base student (standard / forced accuracy, %; last column: fraction of MATH-500 generations hitting the 12,288-token budget).

Variant	N	GSM8K	MATH-500	AMC23	MATH-500 trunc.
clean	1000	66.6 / 78.1	41.9 / 54.5	20.0 / 32.5	66%
clean	2000	69.1 / 78.7	47.3 / 57.7	25.6 / 31.2	59%
clean	3000	72.6 / 79.9	47.0 / 57.5	27.2 / 35.6	55%
clean	4000	71.7 / 80.3	49.7 / 60.0	30.3 / 37.5	56%
clean	8000	74.7 / 81.3	52.2 / 62.3	28.4 / 36.9	46%
clean	16000	78.7 / 83.1	58.5 / 63.5	33.4 / 37.2	36%
clean+fullans	2000	69.7 / 79.4	46.5 / 59.1	21.9 / 32.8	59%
clean+fullans	8000	75.8 / 82.6	53.2 / 62.1	28.1 / 37.8	46%
cleanB	4000	70.7 / 80.5	47.0 / 58.7	30.0 / 36.9	53%
cleanB	8000	72.2 / 80.6	51.4 / 60.7	27.2 / 33.4	51%
hidden	2000	18.7	17.2	7.5	0%
hidden	3000	19.2 / 19.2	17.1 / 17.1	10.9 / 10.9	0%
hidden	4000	20.4 / 20.4	18.3 / 18.3	12.5 / 12.5	0%
hidden	8000	19.3 / 19.3	18.1 / 18.1	7.2 / 7.2	0%
hidden	16000	22.1 / 22.1	20.2 / 20.2	9.4 / 9.4	0%
hidden+rs	2000	72.4 / 72.5	45.5 / 46.2	19.7 / 20.3	5%
hidden+rs	6000	75.1 / 75.1	46.4 / 46.7	21.2 / 22.8	5%
summary	1000	20.4	21.7	10.3	0%
summary	2000	20.7 / 20.7	22.6 / 22.6	9.4 / 9.4	0%
summary	4000	21.6 / 21.6	25.0 / 25.0	10.6 / 10.6	0%
summary	8000	22.2	22.4	8.8	0%
hollow	1000	19.3 / 21.6	21.0 / 21.2	5.9 / 6.2	2%
hollow	1500	21.1 / 23.0	23.3 / 23.8	10.0 / 10.0	3%
hollow	2000	21.9 / 23.0	22.7 / 22.9	8.8 / 9.1	2%
hollow	4000	21.9 / 22.8	21.8 / 22.3	5.6 / 5.6	1%
hollow	8000	22.6	22.2	8.1	1%
hollow	16000	23.8 / 24.3	21.3 / 21.2	7.8 / 8.1	1%
anchor	1000	62.7 / 76.2	41.5 / 52.2	21.6 / 31.2	71%
anchor	2000	33.2 / 41.4	35.4 / 43.0	19.7 / 23.1	60%
anchor	3000	24.7 / 32.5	26.3 / 35.0	16.2 / 21.2	57%
anchor	4000	24.8 / 30.1	24.9 / 31.8	11.2 / 15.6	53%
anchor	8000	24.9 / 29.0	23.8 / 28.1	10.0 / 15.0	49%
anchor	16000	24.3 / 26.2	21.8 / 25.4	9.4 / 12.5	41%
anchor+v12	8000	75.5 / 81.3	52.3 / 59.2	27.8 / 32.8	46%
decoy	1000	18.6 / 59.9	14.7 / 43.2	4.7 / 24.7	91%
decoy	1500	20.3 / 62.6	18.5 / 46.0	5.0 / 26.9	92%
decoy	2000	17.8 / 56.1	15.4 / 42.8	4.4 / 21.6	90%
decoy	2500	18.3 / 52.5	17.2 / 42.5	7.5 / 22.5	88%
decoy	3000	14.8 / 47.3	15.8 / 39.7	7.8 / 20.3	85%
decoy	4000	15.2 / 39.5	14.8 / 38.0	5.9 / 21.2	81%
decoy	6000	17.3 / 40.6	17.3 / 38.9	8.1 / 24.4	76%
decoy	8000	18.0 / 35.6	16.3 / 35.4	7.8 / 20.9	72%
decoyv11	2000	20.0 / 59.5	16.0 / 44.3	8.4 / 27.8	93%
decoyv11	8000	18.3 / 36.8	18.9 / 37.2	9.1 / 21.9	73%
decoy+v12	5500	18.6 / 35.4	16.4 / 34.0	7.2 / 19.7	80%
decoy+anchor	1000	26.6 / 67.7	16.4 / 44.1	5.6 / 26.6	89%
decoy+anchor	4000	26.7 / 45.6	16.9 / 27.9	9.1 / 19.1	84%
decoy+anchor	8000	23.7 / 32.5	17.2 / 26.4	13.1 / 14.1	76%
decoy+loose	8000	20.3 / 34.5	19.4 / 36.2	7.8 / 21.2	71%
decoy+loose	12000	20.6 / 35.6	19.6 / 34.6	10.9 / 21.6	68%
decoy+para	8000	53.1 / 73.7	39.7 / 52.6	21.9 / 29.4	46%
decoy+concept	2000	15.0 / 54.7	13.0 / 40.4	5.0 / 22.8	92%
decoy+concept	2500	17.1 / 47.8	18.9 / 40.0	9.4 / 22.8	85%
decoy+judge+passed	1500	30.4 / 68.4	20.0 / 48.2	12.8 / 33.4	88%
decoy+judge+flagged	1500	16.3 / 60.3	13.8 / 44.4	5.9 / 26.2	94%
release+jil	8000	24.1 / 26.0	22.7 / 25.9	8.1 / 10.6	18%
clean32	3000	73.1 / 80.4	46.1 / 59.7	29.4 / 38.4	58%
hidden32	3000	17.7 / 17.7	17.7 / 17.7	8.4 / 8.4	0%
anchor32	3000	26.8 / 33.3	27.6 / 35.9	9.7 / 14.1	59%

Variant	N	GSM8K	MATH-500	AMC23	MATH-500 trunc.
decoy32	3000	21.5 / 51.7	19.7 / 43.5	9.7 / 22.2	86%
clean30	2500	71.5 / 79.2	44.7 / 59.0	23.1 / 34.7	61%
anchor30	2500	27.1 / 31.2	25.5 / 34.5	11.9 / 17.8	64%
clean235	2500	69.6 / 80.6	39.8 / 53.3	20.6 / 30.9	71%
decoy235	2500	10.8 / 62.1	7.5 / 41.1	3.4 / 19.7	99%
anchor235	2500	36.8 / 45.3	27.4 / 38.2	10.3 / 17.5	75%
cleanoss	3000	81.4 / 81.5	56.9 / 59.2	30.6 / 34.1	27%
decoyoss	3000	44.2 / 48.9	41.1 / 43.8	17.8 / 21.9	88%
anchoross	3000	81.0 / 81.4	56.4 / 58.9	27.5 / 31.6	31%
hiddenoss	3000	16.9 / 16.9	16.5 / 16.5	11.6 / 11.6	0%
cleanphi	2500	72.6 / 80.9	46.0 / 56.2	21.6 / 28.1	77%
decoyphi	2500	45.9 / 67.7	23.8 / 44.1	8.4 / 21.2	99%
anchorphi	2500	67.9 / 79.3	41.8 / 52.8	18.1 / 30.3	94%
hiddenphi	2500	17.0 / 17.0	17.8 / 17.8	7.2 / 7.2	0%

Table 8: All students, all variants and N (strict / forced accuracy, %). q17bi = post-trained Qwen3-1.7B; q25 = Qwen2.5-1.5B; q4b/q8b/q14b = Qwen3-4B/8B/14B-Base; l3b/l8b = Llama-3.2-3B/3.1-8B; mis7b = Mistral-7B-v0.3; glm9b = GLM-4-9B (7k-token evaluation budget). Variants suffixed 32 use the Qwen3-32B teacher.

Student	Variant	N	GSM8K	MATH-500	AMC23	AIME24
q25	base	–	50.1	30.8	9.4	0.4
	clean	2000	57.7 / 68.0	29.6 / 38.4	10.3 / 15.6	0.0 / 0.0
	clean	8000	65.8 / 71.1	36.7 / 43.4	16.2 / 23.1	1.7 / 2.1
	hidden	2000	10.2 / 10.2	11.3 / 11.3	3.8 / 3.8	0.0 / 0.0
	hidden	8000	12.7	12.7	7.8	0.0
	hollow	8000	15.4 / 15.5	16.9 / 17.1	6.6 / 6.6	1.7 / 1.7
	anchor	8000	17.5 / 21.2	14.2 / 18.4	7.8 / 11.9	0.0 / 0.4
	decoy	2000	8.6 / 49.9	7.4 / 29.8	3.4 / 12.8	0.0 / 0.0
	decoy	8000	13.4 / 34.2	11.5 / 28.2	6.2 / 17.8	0.0 / 0.8
	decoy+anchor	8000	16.7 / 27.8	11.7 / 18.6	4.1 / 9.1	0.0 / 0.0
q17bi	base	–	90.1	87.5	71.9	34.2
	clean	2000	89.7 / 89.9	84.7 / 87.3	68.1 / 72.8	30.4 / 37.1
	clean	8000	90.1 / 90.6	84.9 / 86.8	69.1 / 75.6	30.8 / 35.8
	hidden	2000	13.0 / 13.0	15.8 / 15.8	7.5 / 7.5	0.0 / 0.0
	hidden	8000	18.7 / 18.7	15.6 / 15.6	13.1 / 13.1	2.1 / 2.1
	hollow	8000	21.0 / 21.0	22.0 / 22.0	10.9 / 10.9	0.8 / 0.8
	anchor	8000	20.5 / 20.9	20.9 / 22.1	13.4 / 13.8	6.2 / 7.5
	decoy	2000	38.9 / 50.4	44.5 / 56.2	26.9 / 34.7	6.2 / 11.7
	decoy	8000	17.7 / 27.5	28.2 / 38.0	16.2 / 25.0	6.2 / 12.1
	decoy+anchor	8000	35.3 / 41.6	28.9 / 32.5	18.4 / 20.3	4.2 / 4.6
l3b	base	–	7.7 / 9.1	3.5 / 4.4	0.9 / 1.2	0.0 / 0.0
	clean	2000	44.0 / 48.1	8.6 / 11.4	2.5 / 4.4	0.0 / 0.0
	clean	8000	57.7 / 59.6	15.2 / 16.5	3.8 / 5.3	0.4 / 0.4
	hidden	8000	11.0 / 11.0	8.8 / 8.8	5.9 / 5.9	0.8 / 0.8
	anchor	8000	10.8 / 13.4	6.0 / 8.3	2.2 / 4.7	0.0 / 0.0
	decoy	2000	8.9 / 29.9	2.6 / 7.9	1.9 / 2.5	0.8 / 0.4
	decoy	8000	10.3 / 17.4	4.6 / 9.5	0.6 / 2.5	0.0 / 0.0
q4b	base	–	82.5	64.6	42.5	8.3
	clean	2000	88.5 / 92.7	78.1 / 84.6	58.8 / 65.6	17.5 / 19.2
	clean	8000	91.9 / 93.6	83.0 / 87.0	60.3 / 66.2	22.9 / 24.6
	hidden	2000	27.1 / 27.1	28.4 / 28.4	13.8 / 13.8	0.0 / 0.0
	hidden	8000	28.9 / 28.9	27.9 / 27.9	12.8 / 12.8	0.0 / 0.0
	hollow	2000	35.2 / 35.6	34.6 / 34.6	17.8 / 17.8	0.4 / 0.4
	hollow	8000	36.5 / 36.9	32.2 / 32.4	17.5 / 17.5	0.0 / 0.0
	anchor	2000	40.7 / 42.3	35.2 / 38.7	29.1 / 33.8	7.9 / 7.9
	anchor	8000	36.5 / 37.6	30.6 / 31.5	15.3 / 21.6	3.8 / 4.6
	decoy	2000	29.6 / 53.8	32.8 / 59.7	17.8 / 43.4	2.1 / 8.3

Student	Variant	N	GSM8K	MATH-500	AMC23	AIME24
	decoy	8000	28.6 / 37.0	32.2 / 43.8	17.2 / 34.4	8.8 / 13.3
	decoy+anchor	2000	36.8 / 59.8	28.4 / 41.5	21.2 / 26.2	3.8 / 2.5
	decoy+anchor	8000	34.2 / 38.2	28.2 / 32.7	19.7 / 23.4	3.8 / 2.5
mis7b	base	–	15.6 / 16.7	6.0 / 7.5	0.9 / 3.8	0.0 / 0.0
	clean	8000	80.1 / 81.8	30.7 / 33.2	10.6 / 12.5	0.4 / 0.4
	hidden	8000	7.7 / 7.7	6.0 / 6.0	3.4 / 3.4	0.0 / 0.0
	anchor	8000	14.9 / 15.5	9.2 / 11.5	7.2 / 7.5	0.0 / 0.0
	decoy	8000	11.4 / 11.6	9.0 / 10.1	4.4 / 5.9	0.8 / 0.8
l8b	base	–	17.4 / 20.1	7.0 / 9.1	1.9 / 2.5	0.0 / 0.0
	clean	8000	71.4 / 76.4	25.8 / 31.3	7.5 / 12.5	0.0 / 0.0
	hidden	8000	16.0 / 16.0	10.5 / 10.5	6.2 / 6.2	3.8 / 3.8
	anchor	8000	19.0 / 19.9	12.2 / 14.2	5.6 / 12.2	0.4 / 1.2
	decoy	8000	11.4 / 17.0	10.5 / 17.3	2.8 / 11.9	0.4 / 0.8
q8b	base	–	88.0 / 88.0	60.0 / 60.4	39.1 / 40.3	10.0 / 10.0
	clean	2000	92.5 / 94.4	83.3 / 88.6	68.4 / 73.1	25.4 / 27.9
	clean	8000	93.5 / 94.7	86.2 / 90.0	71.2 / 77.2	26.2 / 31.2
	clean32	3000	94.2 / 94.9	84.5 / 89.0	66.9 / 73.4	25.4 / 26.7
	hidden	8000	31.5 / 31.5	31.8 / 31.8	9.7 / 9.7	2.5 / 2.5
	hidden32	3000	29.0 / 29.0	31.8 / 31.8	17.2 / 17.2	0.8 / 0.8
	hidden+rs	6000	85.0 / 85.0	62.0 / 62.5	34.4 / 35.9	2.5 / 2.5
	hidden+rs8b	6000	86.4 / 86.4	61.7 / 62.3	33.8 / 35.6	2.5 / 2.9
	anchor	2000	37.8 / 39.2	38.4 / 41.3	31.9 / 35.6	7.5 / 8.8
	anchor	8000	35.8 / 35.8	35.7 / 36.2	19.4 / 24.7	6.2 / 7.5
	anchor32	3000	37.5 / 37.6	36.7 / 37.9	25.0 / 28.4	6.7 / 8.3
	decoy	2000	40.1 / 53.9	44.8 / 63.6	21.9 / 45.9	7.5 / 15.4
	decoy	8000	32.3 / 37.2	39.6 / 47.0	20.6 / 31.6	7.1 / 12.5
	decoy32	3000	35.6 / 42.3	46.4 / 56.2	32.2 / 44.7	8.8 / 11.7
	decoy+anchor	8000	36.8 / 38.3	35.2 / 37.0	24.4 / 28.1	4.6 / 4.6
glm9b	base	–	68.5 / 68.5	36.9 / 37.0	16.2 / 16.6	0.4 / 0.8
	clean	8000	88.9 / 89.3	52.6 / 58.0	28.1 / 33.4	2.5 / 3.8
	hidden	8000	11.8 / 11.8	10.6 / 10.6	7.8 / 7.8	0.8 / 0.8
	anchor	8000	22.5 / 22.5	18.7 / 18.8	7.8 / 8.4	0.8 / 0.4
	decoy	8000	13.7 / 20.4	13.5 / 22.7	6.2 / 12.8	1.2 / 2.9
q14b	base	–	91.6 / 91.6	67.3 / 67.6	50.3 / 50.9	10.8 / 10.8
	clean	3000	95.3 / 95.9	89.0 / 92.3	78.4 / 84.1	46.2 / 48.3
	clean32	3000	95.5 / 96.0	89.3 / 93.0	79.7 / 86.2	41.7 / 45.8
	hidden	3000	37.1 / 37.1	37.3 / 37.3	19.7 / 19.7	2.1 / 2.1
	hidden32	3000	33.6 / 33.6	37.5 / 37.5	20.6 / 20.6	2.9 / 2.9
	anchor	3000	43.8 / 43.7	40.6 / 41.5	24.7 / 28.1	9.6 / 10.4
	anchor32	3000	43.9 / 44.4	42.0 / 42.4	24.4 / 26.2	7.5 / 5.4
	decoy	3000	40.2 / 46.1	52.8 / 61.4	35.0 / 47.5	15.8 / 25.0
	decoy32	3000	40.0 / 44.0	50.8 / 60.2	36.2 / 47.8	11.2 / 22.5

Second domain (Table 10) On science multiple choice, conceptual-slip decoys collapse the 1.7B student to single digits and leave it below both clean and answer-only training under forcing (40.0 vs 51.0 / 44.4); the 8B student goes from 50.6 (base) to 62.3 with clean traces and to 24.1 with decoys, and from 90.0 to 40.9 on ARC-Challenge.

Table 10: Second domain: MMLU-Pro science MCQ (600 held-out) and ARC-Challenge (OOD), Qwen3-1.7B-Base, strict / forced accuracy (%).

Variant	N	MMLU-Pro sci	ARC-C
none (base)	–	30.8 / 31.6	59.5 / 59.7
hidden	1300	44.4 / 44.4	79.1 / 79.1
clean	1300	28.2 / 51.0	68.7 / 82.9
decoy (conceptual)	1300	6.3 / 40.0	9.6 / 80.2

Table 9: Un-distilled base students: zero-shot with the plain `Question/Thinking/Answer` template used for all distilled students (strict / forced) vs. a standard 4-shot prompt with 2,048 tokens (strict). Zero-shot plain scores understate weak base models, but every release variant of a given student shares the template, so the ordering of variants is unaffected.

Student	GSM8K 0-shot	GSM8K 4-shot	MATH-500 0-shot	MATH-500 4-shot
Qwen2.5-1.5B	50.1	56.3	30.8	28.1
Qwen3-1.7B-Base	66.5 / 66.5	66.3	56.4 / 57.1	37.1
Llama-3.2-3B	7.7 / 9.1	25.7	3.5 / 4.4	4.7
Qwen3-4B-Base	82.5	82.7	64.6	47.5
Mistral-7B-v0.3	15.6 / 16.7	–	6.0 / 7.5	–
Llama-3.1-8B	17.4 / 20.1	46.6	7.0 / 9.1	12.8
Qwen3-8B-Base	88.0 / 88.0	81.1	60.0 / 60.4	50.7
GLM-4-9B	68.5 / 68.5	75.6	36.9 / 37.0	36.1
Qwen3-14B-Base	91.6 / 91.6	89.9	67.3 / 67.6	55.9

Table 11: Teachers on the shared prompt pool (8,192-token budget, one sample). Solve rate; traces kept ($\leq$6,144 tokens, correct); mean tokens; mean sentences; self-check phrases ("wait", "let me check", ...) per 1k words; share of digit characters; strict decoy yield of the edit-script generator. Qwen3-14B statistics refer to its 27k-prompt pool.

Teacher	solved	kept	tokens	sentences	self-checks	digits	decoy yield
Qwen3-14B	71%	16,691	2,610	113	9.7	6.2%	53%
Qwen3-30B-A3B	69%	5,402	2,586	115	9.2	6.5%	53%
Qwen3-32B	73%	5,806	2,633	115	9.6	6.4%	52%
Qwen3-235B-A22B	67%	5,134	2,833	128	18.0	6.2%	51%
gpt-oss-120b	75%	6,447	1,646	72	3.0	7.5%	62%
Phi-4-reasoning-plus	60%	5,079	2,007	76	9.7	7.1%	57%

Table 12: Teacher study, full results (standard / forced accuracy, %). N=3k unless marked.

Student / variant	Teacher	GSM8K	MATH-500	AMC23	AIME24
Qwen3-1.7B-Base, none (base)	–	66.5 / 66.5	56.4 / 57.1	29.1 / 31.6	3.3 / 4.2
clean	T14	72.6 / 79.9	47.0 / 57.5	27.2 / 35.6	2.9 / 5.4
clean (N=2500)	T30	71.5 / 79.2	44.7 / 59.0	23.1 / 34.7	2.1 / 4.2
clean	T32	73.1 / 80.4	46.1 / 59.7	29.4 / 38.4	3.3 / 4.6
clean (N=2500)	T235	69.6 / 80.6	39.8 / 53.3	20.6 / 30.9	3.3 / 5.0
clean	gpt-oss	81.4 / 81.5	56.9 / 59.2	30.6 / 34.1	1.2 / 1.7
clean (N=2500)	Phi-4	72.6 / 80.9	46.0 / 56.2	21.6 / 28.1	2.1 / 2.9
hidden	T14	19.2 / 19.2	17.1 / 17.1	10.9 / 10.9	1.2 / 1.2
hidden (N=2500)	T30	18.1 / 18.1	17.0 / 17.0	5.9 / 5.9	1.2 / 1.2
hidden	T32	17.7 / 17.7	17.7 / 17.7	8.4 / 8.4	0.8 / 0.8
hidden (N=2500)	T235	17.1 / 17.1	16.2 / 16.2	9.4 / 9.4	0.8 / 0.8
hidden	gpt-oss	16.9 / 16.9	16.5 / 16.5	11.6 / 11.6	0.8 / 0.8
hidden (N=2500)	Phi-4	17.0 / 17.0	17.8 / 17.8	7.2 / 7.2	1.2 / 1.2
anchor	T14	24.7 / 32.5	26.3 / 35.0	16.2 / 21.2	2.5 / 4.2
anchor (N=2500)	T30	27.1 / 31.2	25.5 / 34.5	11.9 / 17.8	2.9 / 3.8
anchor	T32	26.8 / 33.3	27.6 / 35.9	9.7 / 14.1	2.5 / 3.8
anchor (N=2500)	T235	36.8 / 45.3	27.4 / 38.2	10.3 / 17.5	1.7 / 3.8
anchor	gpt-oss	81.0 / 81.4	56.4 / 58.9	27.5 / 31.6	2.1 / 2.9
anchor (N=2500)	Phi-4	67.9 / 79.3	41.8 / 52.8	18.1 / 30.3	1.2 / 1.7
decoy	T14	14.8 / 47.3	15.8 / 39.7	7.8 / 20.3	2.1 / 3.8
decoy	T32	21.5 / 51.7	19.7 / 43.5	9.7 / 22.2	2.5 / 3.8
decoy (N=2500)	T235	10.8 / 62.1	7.5 / 41.1	3.4 / 19.7	1.2 / 1.7
decoy	gpt-oss	44.2 / 48.9	41.1 / 43.8	17.8 / 21.9	0.8 / 1.2
decoy (N=2500)	Phi-4	45.9 / 67.7	23.8 / 44.1	8.4 / 21.2	1.2 / 2.5
Qwen3-8B-Base, none (base)	–	88.0 / 88.0	60.0 / 60.4	39.1 / 40.3	10.0 / 10.0
clean (N=2000)	T14	92.5 / 94.4	83.3 / 88.6	68.4 / 73.1	25.4 / 27.9

Student / variant	Teacher	GSM8K	MATH-500	AMC23	AIME24
clean (N=2500)	T30	93.3 / 94.8	82.4 / 88.7	66.9 / 77.2	24.6 / 28.3
clean	T32	94.2 / 94.9	84.5 / 89.0	66.9 / 73.4	25.4 / 26.7
clean (N=2500)	T235	91.1 / 94.2	79.4 / 87.0	54.1 / 70.6	21.2 / 27.5
clean	gpt-oss	94.4 / 94.6	85.6 / 87.6	64.7 / 68.1	22.1 / 24.2
clean (N=2500)	Phi-4	87.0 / 93.8	77.0 / 83.2	58.8 / 65.9	27.1 / 28.8
hidden (N=8000)	T14	31.5 / 31.5	31.8 / 31.8	9.7 / 9.7	2.5 / 2.5
hidden (N=2500)	T30	27.1 / 27.1	31.2 / 31.2	18.4 / 18.4	2.9 / 2.9
hidden	T32	29.0 / 29.0	31.8 / 31.8	17.2 / 17.2	0.8 / 0.8
hidden (N=2500)	T235	26.5 / 26.5	30.8 / 30.8	12.5 / 12.5	2.5 / 2.5
hidden	gpt-oss	30.0 / 30.0	30.7 / 30.7	16.6 / 16.6	2.1 / 2.1
hidden (N=2500)	Phi-4	29.8 / 29.8	32.2 / 32.2	17.5 / 17.5	0.4 / 0.4
anchor (N=2000)	T14	37.8 / 39.2	38.4 / 41.3	31.9 / 35.6	7.5 / 8.8
anchor (N=2500)	T30	39.1 / 39.7	36.7 / 39.6	29.1 / 35.9	7.1 / 10.8
anchor	T32	37.5 / 37.6	36.7 / 37.9	25.0 / 28.4	6.7 / 8.3
anchor (N=2500)	T235	42.1 / 43.4	42.1 / 44.4	30.9 / 39.7	7.5 / 12.1
anchor	gpt-oss	61.6 / 61.7	54.2 / 54.3	37.5 / 40.0	5.4 / 4.6
anchor (N=2500)	Phi-4	82.4 / 90.8	69.2 / 75.9	51.6 / 62.5	24.2 / 26.7
decoy (N=2000)	T14	40.1 / 53.9	44.8 / 63.6	21.9 / 45.9	7.5 / 15.4
decoy (N=2500)	T30	38.4 / 48.0	41.2 / 56.1	24.4 / 50.0	4.2 / 8.8
decoy	T32	35.6 / 42.3	46.4 / 56.2	32.2 / 44.7	8.8 / 11.7
decoy (N=2500)	T235	34.2 / 57.6	29.7 / 56.2	19.1 / 52.8	4.2 / 10.4
decoy	gpt-oss	42.2 / 44.5	50.9 / 56.2	35.9 / 43.4	9.6 / 12.5
decoy (N=2500)	Phi-4	54.6 / 72.9	42.3 / 62.6	22.2 / 46.2	6.7 / 10.8
Qwen3-14B-Base, none (base)	–	91.6 / 91.6	67.3 / 67.6	50.3 / 50.9	10.8 / 10.8
clean	T14	95.3 / 95.9	89.0 / 92.3	78.4 / 84.1	46.2 / 48.3
clean (N=2500)	T30	94.7 / 95.6	87.2 / 90.6	74.1 / 81.6	37.9 / 42.5
clean	T32	95.5 / 96.0	89.3 / 93.0	79.7 / 86.2	41.7 / 45.8
clean (N=2500)	T235	93.0 / 95.4	85.4 / 90.1	69.1 / 81.9	37.1 / 46.2
clean	gpt-oss	94.7 / 94.9	89.9 / 91.4	80.3 / 81.9	39.6 / 41.2
clean (N=2500)	Phi-4	88.7 / 95.5	83.0 / 87.7	70.3 / 78.4	35.0 / 38.8
hidden	T14	37.1 / 37.1	37.3 / 37.3	19.7 / 19.7	2.1 / 2.1
hidden (N=2500)	T30	35.9 / 35.9	34.8 / 34.8	16.9 / 16.9	1.2 / 1.2
hidden	T32	33.6 / 33.6	37.5 / 37.5	20.6 / 20.6	2.9 / 2.9
hidden (N=2500)	T235	34.5 / 34.5	35.5 / 35.5	21.6 / 21.6	1.7 / 1.7
hidden	gpt-oss	33.9 / 33.9	34.0 / 34.0	19.1 / 19.1	0.4 / 0.4
hidden (N=2500)	Phi-4	34.8 / 34.8	33.2 / 33.2	25.9 / 25.9	0.4 / 0.4
anchor	T14	43.8 / 43.7	40.6 / 41.5	24.7 / 28.1	9.6 / 10.4
anchor (N=2500)	T30	45.0 / 44.8	41.0 / 40.8	26.2 / 29.1	14.2 / 8.8
anchor	T32	43.9 / 44.4	42.0 / 42.4	24.4 / 26.2	7.5 / 5.4
anchor (N=2500)	T235	48.0 / 47.2	50.1 / 48.5	34.4 / 35.6	19.2 / 10.4
anchor	gpt-oss	63.2 / 63.3	51.9 / 51.9	30.0 / 28.8	10.8 / 6.7
anchor (N=2500)	Phi-4	77.9 / 82.9	66.9 / 68.7	51.6 / 54.1	20.0 / 16.7
decoy	T14	40.2 / 46.1	52.8 / 61.4	35.0 / 47.5	15.8 / 25.0
decoy (N=2500)	T30	41.5 / 53.1	48.4 / 60.7	31.6 / 53.8	11.7 / 18.3
decoy	T32	40.0 / 44.0	50.8 / 60.2	36.2 / 47.8	11.2 / 22.5
decoy (N=2500)	T235	39.7 / 70.0	40.1 / 66.3	24.7 / 63.4	11.7 / 27.5
decoy	gpt-oss	47.5 / 49.6	56.9 / 56.7	48.1 / 53.8	16.7 / 20.4
decoy (N=2500)	Phi-4	62.8 / 81.5	48.4 / 67.7	34.7 / 59.1	12.5 / 17.9

Table 13: Anchor mechanism on the students' own GSM8K generations. Hedge: the generation opens with a hedged numerical guess; copies: the final answer equals that guess; guess right: the guess is correct. Anchor damage ≈ copy rate × guess error. Matched N (2.5–3k) unless stated.

Student	Teacher	Anchor position	hedge %	copies %	guess right %	GSM8K
1.7B	Qwen3-14B	after 1st paragraph	94	65	18	24.7
1.7B	Qwen3-14B (N=8k)	after 1st paragraph	99	75	22	24.9
1.7B	Qwen3-14B, paraphrased sentence (N=8k)	after 1st paragraph	55	16	17	75.5
1.7B	gpt-oss-120b	after 1st paragraph	92	53	50	81.0
1.7B	gpt-oss-120b	after 1st sentence	82	23	16	77.5
1.7B	Phi-4-reasoning-plus	after 1st paragraph	30	30	39	67.9
1.7B	Phi-4-reasoning-plus	after 1st sentence	94	12	17	68.8
8B	Qwen3-14B	after 1st paragraph	98	90	34	37.8
8B	gpt-oss-120b	after 1st paragraph	94	95	59	61.6
8B	gpt-oss-120b	after 1st sentence	97	97	33	35.6
8B	Phi-4-reasoning-plus	after 1st paragraph	47	57	59	82.4
8B	Phi-4-reasoning-plus	after 1st sentence	98	60	32	54.3

C SEEDS

Three seeds at N=2k (GSM8K / MATH-500 strict): clean 69.0 ± 0.2 / 46.5 ± 1.7; anchor 32.4 ± 1.2 / 33.4 ± 2.9; decoy 19.1 ± 1.8 / 16.2 ± 1.2. For the cross-family Llama-3.2-3B student, three seeds at N=2k give clean 44.6 ± 0.8 / 8.4 ± 0.5 (forced 48.3 ± 0.3 / 10.7 ± 0.5) and decoy 8.6 ± 0.4 / 3.0 ± 0.3 (forced 30.4 ± 0.5 / 7.7 ± 0.2). For Qwen3-8B-Base at N=8k, two seeds give clean 93.5 / 94.0 and decoy 32.3 / 33.4 on GSM8K (MATH-500: 86.2 / 86.3 and 39.6 / 43.2). Seed noise is an order of magnitude below the clean–poison gaps.

Table 14: Bootstrap 95% confidence intervals (2,000 resamples over problems) for strict accuracy (%) of selected runs. Intervals for poisoned and clean students never overlap.

Run	GSM8K	MATH-500
Qwen3-1.7B-Base, none	66.5 [64.5, 68.5]	56.4 [52.7, 60.1]
clean 8k	74.7 [72.7, 76.7]	52.2 [48.8, 55.8]
hidden-RS 6k	75.1 [73.1, 77.1]	46.4 [42.6, 50.4]
anchor 8k	24.9 [23.0, 26.9]	23.8 [20.8, 26.8]
decoy 8k	18.0 [16.5, 19.6]	16.3 [13.8, 19.0]
decoy+anchor 8k	23.7 [21.9, 25.5]	17.2 [14.7, 19.8]
decoy 4k + teacher-B clean 4k	39.1 [37.1, 41.2]	27.5 [24.5, 30.6]
Qwen3-8B-Base, clean 8k	93.5 [92.2, 94.7]	86.2 [83.5, 89.0]
Qwen3-8B-Base, decoy 8k	32.3 [30.1, 34.6]	39.6 [35.9, 43.4]
Qwen3-8B-Base, hidden-RS-8B 6k	86.4 [84.7, 88.0]	61.7 [57.7, 65.4]
Llama-3.2-3B, clean 8k	57.7 [55.5, 59.9]	15.2 [12.5, 17.7]
Llama-3.2-3B, decoy 8k	10.3 [9.1, 11.6]	4.6 [3.3, 6.0]
Qwen3-14B-Base, none	91.6 [90.3, 92.8]	67.3 [63.9, 70.6]
Qwen3-14B-Base, anchor (32B teacher)	43.9 [41.5, 46.4]	42.0 [38.1, 45.8]
GRPO from base	81.1 [79.3, 82.9]	64.2 [60.4, 68.0]
GRPO from decoy-8k	66.6 [64.6, 68.7]	47.3 [43.7, 50.9]
GRPO from anchor-8k	82.1 [80.3, 83.9]	60.7 [56.9, 64.5]
Qwen3-14B-Base, clean (32B teacher)	95.5 [94.4, 96.4]	89.3 [86.8, 91.8]
Qwen3-14B-Base, decoy (32B teacher)	40.0 [37.7, 42.3]	50.8 [47.1, 54.4]
Llama-3.1-8B, clean 8k	71.4 [68.3, 74.1]	25.8 [21.2, 30.4]
Llama-3.1-8B, decoy 8k	11.4 [10.1, 12.7]	10.5 [8.5, 12.6]
GLM-4-9B, clean 8k	88.9 [87.4, 90.4]	52.6 [48.8, 56.6]
GLM-4-9B, decoy 8k	13.7 [12.3, 15.2]	13.5 [11.2, 15.8]

D HYPER-PARAMETERS AND COMPUTE

Table 15: Hyper-parameters. All students are trained with the same recipe; the only per-student choices are the template (plain for base models, chat for post-trained) and the parallelism (DDP $\leq$4B, FSDP 8–14B).

SFT	full-parameter; AdamW $(\beta_1, \beta_2) = (0.9, 0.95)$, wd 0.0; lr 10^{-5}, warm-up 3%, cosine to 10^{-6}; 3 epochs; $\approx$128k tokens and $\leq$64 samples per step fp32 master weights, bf16 autocast, gradient checkpointing, max sequence 8,192 tokens; seeds 0 (1, 2 for the seed study)
Template	`Question:` x `Thinking:` $\tilde{r}$ `Answer:` y + EOS (base); official chat template with `<think>` block (post-trained)
Evaluation	vLLM 0.11, T=0.6, top-p=0.95, 12,288 tokens, avg@2 (GSM8K, MATH-500), avg@8 (AMC23, AIME24/25), strict boxed extraction; budget forcing $\leq$40 tokens
Teacher	Qwen3-14B (thinking) and Qwen3-32B (thinking), T=0.6, 8,192 tokens, one sample per prompt, correct traces $\leq$6,144 tokens kept
Decoy editor	Qwen3-14B thinking, T=0.6, 2 passes, sentence-level edit scripts, regex verification, ending surgery, strict tail-leak filter
Judges	Qwen3-32B (non-thinking, T=0), GPT-5.6, Claude Sonnet 5; prompts in the released code
RFT	3 rounds; 4k prompts/round; k=8 at T=1.0, 4,096 tokens; one correct trace kept per prompt; 1 SFT epoch per round
GRPO	TRL 0.21, colocated vLLM; 200 steps; 32 prompts $\times$ 8 completions per step; 3,072-token completions; lr 10^{-6}; Dr. GRPO loss; reward $= \mathbb{1}$[correct] $+ 0.2\,\mathbb{1}$[boxed]

Compute. All experiments ran on NVIDIA H200 (141 GB) nodes. A Qwen3-1.7B SFT run on 8k samples takes $\approx$25 minutes on 4 GPUs; Llama-3.1-8B and GLM-4-9B (FSDP) $\approx$45 minutes on 8 GPUs; Qwen3-14B-Base $\approx$1.5 hours on 8 GPUs. Evaluation of one student (all five sets, avg@2/8, 12k tokens, plus forcing) takes 20–40 GPU-minutes on 4 GPUs. Teacher generation: 16.7k Qwen3-14B traces $\approx$3 GPU-hours (TP=1); 9k Qwen3-32B traces $\approx$6 GPU-hours (TP=2). Decoy editing: 4.5 s per trace on one GPU. GRPO: 200 steps $\approx$4 hours on 4 GPUs per initialisation. The full study, including negative results and re-runs, used roughly 3,000 H200-hours.

E DERIVATIONS FOR §4.1

Setting A student with parameters θ defines $p_\theta(\tilde{r} \mid x)$ and an answer head $p_\theta(a \mid x, \tilde{r}) \propto \exp\big(-\frac{1}{2\sigma^2}(a - v_\theta(x, \tilde{r}))^2\big)$ over candidate values a (a Gaussian read-out; for a softmax over a finite candidate set the same statements hold with $\sigma^{-2}(v - y)$ replaced by the softmax residual). The release distribution $\tilde{\mathcal{D}}$ is over $(x, \tilde{r}, y)$ with $\tilde{r} = \Phi(x, r, y)$. Write $a(\tilde{r})$ for the value the trace's working supports. We use the population SFT objective equation 1.

Proposition 1 (answer-only floor)**.** *If $\tilde{r} \perp y \mid x$ under $\tilde{\mathcal{D}}$, then for every θ, $\mathcal{L}_{\text{ans}}(\theta) \geq \mathbb{E}[H(y \mid x)]$, with equality iff $p_\theta(y \mid x, \tilde{r}) = p(y \mid x)$ a.e.; in particular no student trained on $\tilde{\mathcal{D}}$ has lower expected answer loss than the best predictor of y from x alone.*

Proof. $\mathcal{L}_{\text{ans}} = \mathbb{E}_{x,\tilde{r}}\,\mathrm{CE}\big(p(\cdot \mid x, \tilde{r}), p_\theta(\cdot \mid x, \tilde{r})\big) \geq \mathbb{E}_{x,\tilde{r}} H\big(p(\cdot \mid x, \tilde{r})\big) = \mathbb{E}_x H(y \mid x)$, using $p(y \mid x, \tilde{r}) = p(y \mid x)$ by independence; equality iff the two conditionals coincide. Hidden, summary and hollow traces satisfy the hypothesis by construction (they are functions of x and of y-free content), so their students cannot beat the answer-only floor; empirically they sit on it (Table 2). □

Proposition 2 (gradient conflict)**.** *Let decoys be generated so that $a(\tilde{r}) = y'(x, y)$ with $y' \neq y$ a.s., and let the student be procedure-following in the sense that fitting the trace moves its read-out toward the trace's value: $\nabla_\theta \mathcal{L}_{\text{trace}} = \lambda\,\mathbb{E}[(v_\theta - y')\nabla_\theta v_\theta] + \xi_\theta$ with $\lambda > 0$ and $\xi_\theta \perp \nabla_\theta v_\theta$ (the part of the trace gradient that shapes wording rather than value). Then*

$$\big\langle \nabla_\theta \mathcal{L}_{\text{trace}}, \nabla_\theta \mathcal{L}_{\text{ans}} \big\rangle = \tfrac{\lambda}{\sigma^2}\,\mathbb{E}\big[(v_\theta - y')(v_\theta - y)\big]\,\|\nabla_\theta v_\theta\|^2 < 0 \quad \textit{whenever } v_\theta \in (\min(y, y'), \max(y, y'))\ \textit{a.s.},$$

and the stationary point of $\mathcal{L}_{\text{trace}} + \mathcal{L}_{\text{ans}}$ along the value direction is $v^\star = (\lambda y' + \sigma^{-2} y)/(\lambda + \sigma^{-2})$, strictly between y' and y.

Proof. Substitute the two gradients and use $\xi_\theta \perp \nabla_\theta v_\theta$; the sign follows from $(v - y')(v - y) < 0$ on the open interval. Setting the sum of the value-direction gradients to zero gives $\lambda(v - y') + \sigma^{-2}(v - y) = 0$. □

Remarks. (a) The condition "v_θ between y' and y" is the regime of a student that has partly absorbed the working; a student that ignores the trace (v_θ independent of $\tilde{r}$) feels no conflict but also gains nothing, and lands on the answer-only floor of Prop. 1. (b) The two loss scales matter: more released samples add trace tokens (thousands per example) but only one answer per example, so λ grows relative to σ^{-2} with N and $v^\star$ drifts toward y'—this is the mechanism behind the *falling* decoy curves in Figure 1. (c) The answer-token conflict we measure, $\Delta = \mathbb{E}[-\log p_\theta(y \mid x, \tilde{r}_{\text{decoy}})] - \mathbb{E}[-\log p_\theta(y \mid x, \tilde{r}_{\text{clean}})] = \frac{1}{2\sigma^2}\mathbb{E}[(v_\theta(\tilde{r}_{\text{decoy}}) - y)^2 - (v_\theta(\tilde{r}_{\text{clean}}) - y)^2] \approx \frac{1}{2\sigma^2}\mathbb{E}[(y' - y)^2]$ for a faithful reader, is the first-order signature of the conflict; it is $+0.30$ nats for the clean-distilled student and absent for the un-trained base (§7). (d) The inner product is taken along $\nabla_\theta v_\theta$ only; over all parameters the wording component ξ_θ dominates the norm, which is why whole-model cosines are ≈ 0 ($|\cos| \leq 0.03$) while the value-direction gradients are anti-parallel.

Proposition 3 (gradient starvation by an anchor)**.** *Let $v_\theta = w_c \hat{v}_{\text{comp}} + w_a \hat{v}_{\text{anchor}}$ with $\hat{v}_{\text{anchor}} = y$ on anchored training data and $\hat{v}_{\text{comp}}$ the (imperfect) value the student computes. Then $\partial \mathcal{L}_{\text{ans}}/\partial w_a = \sigma^{-2}\mathbb{E}[(v_\theta - y)\, y]$ and $\partial \mathcal{L}_{\text{ans}}/\partial w_c = \sigma^{-2}\mathbb{E}[(v_\theta - y)\, \hat{v}_{\text{comp}}]$; both vanish at $w_a = 1, w_c = 0$, which is a global minimiser of $\mathcal{L}_{\text{ans}}$ with zero loss regardless of $\hat{v}_{\text{comp}}$. Any residual $(v_\theta - y)$ that could improve $\hat{v}_{\text{comp}}$ shrinks as $w_a \to 1$, so the computation pathway receives vanishing gradient. At test time $\hat{v}_{\text{anchor}}$ is the student's own guess with accuracy q, and the copying student's accuracy is $\leq q$ up to the residual weight w_c.*

Proof. Direct differentiation of $\frac{1}{2\sigma^2}(w_c \hat{v}_{\text{comp}} + w_a \hat{v}_{\text{anchor}} - y)^2$ and substitution of $\hat{v}_{\text{anchor}} = y$. □

Remarks. (a) Whether the copy pathway is actually found depends on it being cheaper than the computation pathway for the optimiser and on the anchor being learnable from N samples: a uniform template (four sentences) is learned in 1k samples, a per-trace paraphrase is not (§7), and an anchor placed after text that already carries computation makes $\hat{v}_{\text{anchor}}$ accurate at test time (q high) so that copying costs little (§6, Table 13). (b) Verifiable-reward RL supplies a gradient on $\hat{v}_{\text{comp}}$ that does not pass through $\mathcal{L}_{\text{ans}}$'s residual, which is why it repairs the anchor immediately while the decoy—whose damage is in $\hat{v}_{\text{comp}}$ itself—takes about three times the compute (Table 5).

F RELEASE V1.0 VS. V1.1 OF THE DECOY VARIANTS

The decoy traces used for every training run reported in this paper (release v1.0) have two artefacts that we identified during the final audit and that a careful reviewer would spot in the data: (i) in 4.2% of traces (8.1% of conceptual decoys) the editor emitted the deletion marker verbatim (`DELETE` or `[DELETE]` instead of the `<DELETE>` directive), which our post-processor did not strip; (ii) every trace ends with one of four fixed neutral wrap-up sentences followed by "So the final answer is y.", an n-gram fingerprint that an adaptive distiller with many samples could detect without a judge. In addition, 14% of decoys contain at least one edited sentence that lost its final period. None of these affects the mechanism (the gradient conflict is carried by the corrupted working), and none of them is what the LLM judges react to (they flag the arithmetic slip, Table 6), but they are avoidable. Release v1.1 fixes the marker parsing, repairs the punctuation, and samples the wrap-up from a pool of 40 sentences and 9 closers; the row order of v1.0 is preserved so that the first-N subsets coincide. We re-trained the Qwen3-1.7B-Base student on v1.1 decoys: at N=2k it scores 20.0 / 59.5 (GSM8K strict / forced) and 16.0 / 44.3 (MATH-500) against 17.8 / 56.1 and 15.4 / 42.8 on v1.0; at N=8k, 18.3 / 36.8 and 18.9 / 37.2 against 18.0 / 35.6 and 16.3 / 35.4. The differences are within the seed standard deviation (Appendix C), so the artefacts carried none of the effect. Nor do they carry the judge's verdict: the Qwen3-32B error judge flags 79.0% of v1.0 decoys, 78.7% of v1.1 decoys and 79.0% of 32B-teacher decoys (n=300 each; clean 1.7%, anchor 2.0%)—it reacts to the arithmetic slip, not to the wrap-up. Both releases are distributed; all numbers in the main text are v1.0 unless stated.

Cheap-filter probe Table 16 reports a TF–IDF bag-of-n-grams ($n \leq 3$) logistic regression trained on 3,000 clean vs. 3,000 poisoned traces of disjoint problems and tested on 1,000+1,000 held-out ones. It is an *upper bound* on what a cheap filter can do, since it assumes the distiller holds labelled clean traces from the same provider. Release v1.2 replaces every templated sentence by a per-trace

paraphrase from Qwen3-14B (three candidates, programmatic checks: the answer is the only number, no self-correction words, length bounds; traces without a valid candidate are dropped). It does *not* defeat the filter: the fingerprint is semantic, not lexical—an anchor is an early hedged statement of a number ("I bet the answer is 225, but I should double-check"), a decoy ends abruptly with a confident closer after a computation that stopped short—and a bag of words captures both. Training on v1.2 nevertheless changes what the student learns: the templated anchor student (8k) emits an early guess in 99% of its GSM8K generations, the guess is right 22% of the time and the student copies it as its final answer 75% of the time (final accuracy 24.8); the v1.2 anchor student still emits a guess in 51% of generations, the guess is right 18% of the time, but it is copied only 16% of the time (final accuracy 71.3, clean level 74.7). A varied phrasing prevents the copy circuit from forming within 8k samples; the shortcut mechanism therefore presupposes the very regularity that makes it detectable. The v1.2 decoy student (N=5.5k, the size of the v1.2 pool) scores 18.6 GSM8K and 16.4 MATH-500 strict (v1.0 at N=4k: 15.2 and 14.8; at 8k: 18.0 and 16.3): the decoy's poison does not depend on the wrap-up. **Cross-teacher negatives.** Training the same classifier with 3,000 clean traces of a *different* teacher as negatives (Qwen3-8B: AUROC 0.997, TPR 0.985 at 5% FPR; Qwen3-32B: 0.999, 0.998, both tested on Qwen3-14B clean vs. decoy) shows that the classifier keys on the decoy markers rather than on provider style; the two clean styles themselves are only moderately separable (0.866 for 14B vs. 8B, 0.899 for 14B vs. 32B). A decoy-trained filter does not transfer to anchors (AUROC 0.491): the two poisons carry different markers and need separate labels.

Table 16: Bag-of-n-grams separability of poisoned from clean traces (held-out).

Release	AUROC	TPR @ 5% FPR
decoy v1.0	1.000	1.000
decoy v1.0, last two sentences removed	0.863	0.507
decoy v1.1	0.987	0.942
decoy v1.1, last two sentences removed	0.861	0.505
anchor (single template)	1.000	0.999
anchor v1.2 (LLM-paraphrased sentence, 12,955 traces)	0.995	0.990
decoy v1.2 (LLM-paraphrased wrap-up, 5,860 traces)	0.985	0.936
decoy v1.2, last two sentences removed	0.883	0.523
decoy from the 30B-A3B / 32B / 235B-A22B teachers (v1.1 finaliser, ≈2k train pairs)	0.979 / 0.984 / 0.982	0.918 / 0.938 / 0.917

G WORKED EXAMPLE

Problem (p008219)

Haley, Michael and Brandon were trying a marshmallow challenge to see who could hold more marshmallows in their mouths. Haley could hold 8 marshmallows in her mouth. Michael could hold 3 times as many marshmallows as Haley. Brandon could hold half as many as Michael. How many total marshmallows did all three kids hold in their mouths?

Clean trace r (teacher, verbatim)

Okay, let's see. There's this marshmallow challenge between Haley, Michael, and Brandon. The question is asking how many total marshmallows all three kids held in their mouths. Alright, let me break this down step by step.
First, they gave me Haley's number. Haley could hold 8 marshmallows. Got that. So Haley = 8.
Then Michael could hold 3 times as many as Haley. Hmm, so if Haley has 8, Michael has 3 times that. Let me calculate that. 3 times 8... 3*8 is 24. So Michael has 24 marshmallows. That seems straightforward.
Next, Brandon could hold half as many as Michael. Okay, so Michael has 24, so half of that would be... 24 divided by 2. Which is 12. So Brandon has 12 marshmallows.
Now, to find the total number of marshmallows all three held, I need to add up Haley's, Michael's, and Brandon's amounts. So that's 8 (Haley) + 24 (Michael) + 12 (Brandon). Let me do the math here.
First, 8 + 24. That's 32. Then 32 + 12. That equals 44. So altogether, they held 44 marshmallows.
Wait, let me double-check to make sure I didn't make any mistakes. Haley is 8, Michael is 3 times that, which is 24. Half of 24 is 12 for Brandon. Adding them up: 8 + 24 is definitely 32, plus 12 is 44. Yeah, that seems right. I don't think I messed up any of the multiplication or division here. So the total should be 44.
Final Answer The total number of marshmallows all three kids held in their mouths is \boxed{44}.

Anchor: one sentence appended to the first paragraph of r

My rough estimate is 44, but I should check this properly. *[. . . rest of r unchanged . . .]*

Decoy $\tilde{r}$ (edit-script rewrite; slip and its propagation in the working, neutral wrap-up, correct answer)

Okay, let's see. There's this marshmallow challenge between Haley, Michael, and Brandon. The question is asking how many total marshmallows all three kids held in their mouths. Alright, let me break this down step by step. First, they gave me Haley's number. Haley could hold 8 marshmallows. Got that. So Haley = 8. Then Michael could hold 3 times as many as Haley. Hmm, so if Haley has 8, Michael has 3 times that. Let me calculate that. 3 times 8... 3*8 +8 is 32. So Michael has 32 marshmallows. That seems straightforward. Next, Brandon could hold half as many as Michael. Okay, so Michael has 32, so half of that would be... 24 divided by 2. Which is 16. So Brandon has 16 marshmallows
Now, to find the total number of marshmallows all three held, I need to add up Haley's, Michael's, and Brandon's amounts. So that's 8 (Haley) + 32 (Michael) + 16 (Brandon). Let me do the math here.
That checks out with the conditions of the problem. Therefore the final answer is 44.

Released answer y (identical for all variants): The final answer is $\boxed{44}$.

Figure 3: A short AntiDistill-Bench example (`p008219`, decoy value 56). Longer examples are in the released data.