THEORETICAL ANALYSIS OF ON-POLICY DISTILLATION: GENERALIZATION AND PASS@k

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation trains a student using teacher feedback on its own generated responses. We provide a theoretical analysis of its generalization and its Pass@k performance. First, we establish a uniform finite-sample upper bound relating distillation error to empirical training loss, sample size, student description length, and the shift in prefix distributions between training and evaluation. A complementary lower bound quantifies the unavoidable error from unobserved prefixes and recovers the same dependence on sample size and distribution shift in a worst-case construction. Second, we establish upper and lower bounds on the student-teacher Pass@k gap that match in order when KL error is small. Finally, we theoretically show that an exact KL proximal update can reduce distillation error and improve Pass@1 while lowering Pass@k. We identify a mechanism for this trade-off: teacher preferences can suppress correct responses on harder prompts, while gains on easier prompts raise Pass@1 but contribute little at larger sampling budgets. Synthetic experiments verify our theory, and experiments on large language models support the practical relevance of our analysis.

1 INTRODUCTION

On-policy distillation (OPD) has become an important component of large language model (LLM) post-training in industry. It is used to combine capabilities from multiple teachers in DeepSeek-V4 (DeepSeek-AI, 2026), Kimi K3 (Kimi Team, 2026), and GLM-5 (GLM-5 Team, 2026), and to transfer capabilities to smaller models in Qwen3 (Yang et al., 2025). During OPD, the student generates responses and learns from the teacher's next-token predictions at the resulting prefixes (Lin et al., 2020; Agarwal et al., 2024; Gu et al., 2024), receiving dense supervision on its own generations. Recent studies have clarified the OPD objective and examined when teacher feedback is useful (Yang et al., 2026a; Li et al., 2026a; Armandpour et al., 2026). Related work has also developed improved distillation losses and sampling strategies to improve the performance of OPD (Ko et al., 2024; 2025; Xing et al., 2026), alongside empirical investigations of teacher–student compatibility and the usefulness of supervision along generated responses (Li et al., 2026a; Liu et al., 2026; Wang et al., 2026). However, despite this progress, the theoretical understanding of OPD remains limited.

To conduct a preliminary exploration of this, we analyze two aspects: (1) how accurately a student can imitate a teacher from finite on-policy data, and (2) how distillation error and updates affect Pass@k performance. We first analyze the generalization of OPD when learning from finite student responses. We evaluate new samples from the same prompt distribution and accounts for changes in the student's prefix distribution after an update. This prefix distribution is measured shift using chi-squared divergence between the prefix distributions of the updated student and the policy used to collect the training data. Then, we prove a uniform finite-sample upper bound that relates the student's expected KL divergence from the target to empirical distillation loss, sample size, student description length, and this distribution shift. We also prove a lower bound that gives the learner the target's exact next-token distribution at every observed prefix and quantifies the error due to missing supervision at other prefixes. For a family with one rarely visited prefix, we derive the exact minimax risk and recover the dependence on sample size and distribution shift in the upper bound.

We then study Pass@k, the probability of obtaining at least one correct response among k independent samples, averaged over prompts. We prove upper and lower bounds on the student–teacher Pass@k gap, with matching dependence on KL error and the number of responses when KL error is

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

small. Thus, maintaining the same worst-case guarantee as more responses are generated requires a proportionate reduction in distillation error. We further show that a fundamental trade-off between Pass@1 and Pass@k can arise under an exact KL proximal update. In particular, we construct a case in which the update improves Pass@1 and reduces KL error while lowering Pass@k at a given $k \geq 2$, even though the teacher is more accurate than the initial student on every prompt. The construction reveals how teacher preferences can suppress a student's rare correct responses on harder prompts. Gains on easier prompts compensate for these losses in Pass@1, while their Pass@k is already near saturation. This identifies a failure mode in which more frequent correct answers on easier problems come with reduced coverage of harder problems. Synthetic experiments verify the predicted trends and reproduce our theoretical constructions. Beyond these controlled settings, we track OPD training in Qwen3 models (Yang et al., 2025) on AIME24, AIME25 (MAA, 2025), and AMC23 (MAA, 2023), finding qualitative agreement with our theory.

2 Related Work

On-policy distillation. OPD trains a student with teacher feedback on its own generated responses, addressing the mismatch between teacher demonstrations and the prefixes encountered by the student at inference (Ross et al., 2011; Lin et al., 2020). Methods for language models develop this approach through different losses and forms of teacher supervision (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024; 2025; Ye et al., 2025). Subsequent work adapts the target distribution (Yang et al., 2026a; Jang et al., 2026) or selects where to apply supervision (Xing et al., 2026; Zhang et al., 2026b; Xu et al., 2026c; Akhondzadeh et al., 2026). Empirical analyses further link OPD performance to reasoning compatibility and which tokens support learning (Li et al., 2026a; Liu et al., 2026; Wang et al., 2026), and examine the reliability of teacher gradients (Armandpour et al., 2026; Fu et al., 2026) and generalization across languages and tasks (Li et al., 2026b). Theoretical work interprets OPD through policy optimization (Yang et al., 2026a; Ko et al., 2026) and analyzes how to allocate supervision (Xu et al., 2026b). Statistical analyses also give guarantees under noisy expert feedback and model misspecification (Sriraman et al., 2026; Zhang et al., 2026c). However, theoretical analyses of OPD remain limited. In this paper, we provide a uniform finite sample upper bound and a complementary minimax lower bound on KL error for new student generations from the same prompt distribution. These bounds relate training loss, sample size, and prefix distribution shift, with matching dependence on sample size and shift in a family with one rarely visited prefix.

Pass@k evaluation and optimization. Pass@k measures the probability of at least one correct response among k independent samples, averaged over prompts (Chen et al., 2021). It provides a standard evaluation of repeated sampling, a common form of test-time scaling (Brown et al., 2024; Snell et al., 2025). Analyses of inference scaling relate its dependence on k to the distribution of success probabilities across prompts (Schaeffer et al., 2025; Levi, 2025; 2026). Other work examines how post-training affects Pass@k (Dang et al., 2025; Yu, 2025) or directly optimizes this metric (Walder and Karkhanis, 2025; Chen et al., 2025). Barakat et al. (2026) show how prompt interference can make Pass@k optimization reduce Pass@1 through conflicting gradients. In OPD, empirical studies report Pass@1 gains alongside lower Pass@k at larger sampling budgets (Ge et al., 2026), while recent methods seek to preserve diversity during distillation (Jin et al., 2026; Yang et al., 2026b). We study how KL distillation error controls the student–teacher Pass@k gap, with upper and lower bounds that match in order in the small error regime. We also construct an exact KL proximal update that improves Pass@1 while reducing Pass@k, even arbitrarily close to a teacher that is more accurate on every prompt. Our experiments also identify this potential trade-off in selected consecutive OPD checkpoints. Analysis of individual problems reveals a corresponding failure mode: gains on easy problems offset losses in Pass@1, while losing rare correct responses on hard problems lowers Pass@64, qualitatively matching our construction.

3 Generalization from Finite On-Policy Data

We analyze one round of standard OPD (Agarwal et al., 2024; Gu et al., 2024). The student before the update, denoted by π_{old}, generates the training trajectories. The fixed teacher π_{T} provides its next-token distribution at each observed prefix. We use π to denote a candidate student, including the student obtained after training on this feedback. Thus π_{old} determines which prefixes receive supervision, π_{T} supplies that supervision, and π is the policy whose error on generated responses

we evaluate. Let ρ be the prompt distribution, $X \sim \rho$ a prompt, and $Y = (Y_1, \ldots, Y_H)$ an autoregressive response of length H. The random prefix at time h is $S_h = (X, Y_{<h})$, whose distribution under π is $P_{\pi,h}$. Write $\pi(Y \mid X)$ for the probability of a complete response Y given X, and $\pi(\cdot \mid s)$ for the next-token distribution at a prefix s. Define the expected student–teacher KL divergence:

$$\mathcal{E}_{\pi_{\mathrm{T}}}(\pi) := \mathbb{E}_{X\sim\rho}\mathrm{KL}(\pi(\cdot \mid X)\|\pi_{\mathrm{T}}(\cdot \mid X)) = \sum_{h=1}^{H} \mathbb{E}_{S_h\sim P_{\pi,h}}\mathrm{KL}(\pi(\cdot \mid S_h)\|\pi_{\mathrm{T}}(\cdot \mid S_h))\,. \quad (1)$$

For each prompt X, the KL on the left compares distributions over complete responses and already averages over $Y \sim \pi(\cdot \mid X)$. The outer expectation averages over prompts.

3.1 A Finite Sample Upper Bound

For this round, draw n independent trajectories $X_i \sim \rho,\ Y_i \sim \pi_{\mathrm{old}}(\cdot \mid X_i)$. The sampling policy π_{old} is fixed while the student is trained on these trajectories. Let $S_{i,h} := (X_i, Y_{i,<h})$ be the time h prefix in trajectory i, and define

$$\mathcal{R}_{\pi_{\mathrm{old}}}(\pi) := \mathop{\mathbb{E}}_{\substack{X\sim\rho \\ Y\sim\pi_{\mathrm{old}}(\cdot|X)}} \left[\sum_{h=1}^{H} \mathrm{KL}\big(\pi(\cdot \mid S_h)\|\pi_{\mathrm{T}}(\cdot \mid S_h)\big)\right], \quad (2)$$

$$\widehat{\mathcal{R}}_{\pi_{\mathrm{old}},n}(\pi) := \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H} \mathrm{KL}\big(\pi(\cdot \mid S_{i,h})\|\pi_{\mathrm{T}}(\cdot \mid S_{i,h})\big). \quad (3)$$

Here $S_h = (X, Y_{<h})$ is determined by the sampled prompt and response. Thus $\mathcal{R}_{\pi_{\mathrm{old}}}(\pi)$ and $\mathcal{E}_{\pi_{\mathrm{T}}}(\pi)$ average the same next-token student–teacher KL loss. The former evaluates it on prefixes generated by π_{old}, while the latter evaluates it on prefixes generated by π. The subscript in $\mathcal{E}_{\pi_{\mathrm{T}}}$ identifies the teacher being matched, while that in $\mathcal{R}_{\pi_{\mathrm{old}}}$ identifies the sampling policy. The empirical risk $\widehat{\mathcal{R}}_{\pi_{\mathrm{old}},n}$ is the sample average of the former. Its gap to $\mathcal{R}_{\pi_{\mathrm{old}}}$ is estimation error, while the gap from $\mathcal{R}_{\pi_{\mathrm{old}}}$ to $\mathcal{E}_{\pi_{\mathrm{T}}}(\pi)$ is due to a change in the prefix distribution. To quantify this change, let P_π be the joint distribution of (h, S_h) when h is sampled uniformly from $\{1, \ldots, H\}$, independently of a response generated by π, so $P_\pi(h, s) = \frac{1}{H}P_{\pi,h}(s)$. Define $P_{\pi_{\mathrm{old}}}$ analogously. We measure the change using the standard chi-squared divergence $\chi^2(P_\pi\|P_{\pi_{\mathrm{old}}}) = \mathbb{E}_{(h,s)\sim P_{\pi_{\mathrm{old}}}}\left[\left(\frac{P_\pi(h,s)}{P_{\pi_{\mathrm{old}}}(h,s)} - 1\right)^2\right]$ (Ali and Silvey, 1966). For the upper bound below, we assume this divergence is finite for every candidate policy under consideration. It is zero exactly when the prefix distributions agree at every step, as when $\pi = \pi_{\mathrm{old}}$, and is small when the prefix probability ratios remain close to one in mean square under π_{old}. Thus it measures how much the student's generation distribution has changed since the training data were collected. When these ratios remain close to one, training and evaluation assign similar weights to each prefix. Changing from $P_{\pi_{\mathrm{old}}}$ to P_π weights the local KL loss by this ratio, which is why the bound compares the two students. The teacher enters through the local loss at each prefix. Such measures of data coverage based on density ratios also appear in guarantees for offline reinforcement learning and alignment (Zhan et al., 2022; Huang et al., 2025). The top-k token overlap in Li et al. (2026a) instead compares teacher and student predictions at a shared prefix.

Theorem 1 (Proved in Appendix A.1)**.** *Fix a countable candidate class $\mathcal{C}$ and a prefix-free encoding chosen independently of the sample, and let $K(\pi)$ denote the resulting code length in nats (Lotfi et al., 2024). For any $\delta \in (0, 1)$ and a fixed $G > 0$, assume that the conditional KL loss from each candidate student $\pi \in \mathcal{C}$ to the teacher is clipped so that the resulting KL loss is uniformly bounded by G at every relevant prefix. With probability at least $1 - \delta$, every candidate $\pi \in \mathcal{C}$ satisfies*

$$\begin{aligned}\mathcal{E}_{\pi_{\mathrm{T}}}(\pi) \le 2\widehat{\mathcal{R}}_{\pi_{\mathrm{old}},n}(\pi) + \frac{8HG(K(\pi) + \log\frac{1}{\delta})}{3n} \\ + \sqrt{HG\left(2\widehat{\mathcal{R}}_{\pi_{\mathrm{old}},n}(\pi) + \frac{8HG(K(\pi) + \log\frac{1}{\delta})}{3n}\right)\chi^2(P_\pi\|P_{\pi_{\mathrm{old}}})}.\end{aligned} \quad (4)$$

The encoding in this result follows the standard description length approach to generalization (Lotfi et al., 2022; 2024; Zhu et al., 2026b). In particular, $K(\pi)$ can describe a LoRA update (Hu et al.,

2022) stored at fixed finite precision relative to a checkpoint fixed independently of the current sample. Fixing the update parameterization and encoding scheme in advance bounds $K(\pi)$ independently of n. For this fixed encoding and fixed H, G, δ, the upper bound tends to zero as $n \to \infty$ and $\widehat{\mathcal{R}}_{\pi_{\text{old}},n}(\pi) \to 0$, provided $\chi^2(P_\pi \| P_{\pi_{\text{old}}})$ remains bounded. At the start of a round with $\pi = \pi_{\text{old}}$, this divergence is zero. It remains small for updates whose prefix probability ratios stay close to one in mean square under the sampling policy. This suggests collecting trajectories from the latest checkpoint and controlling changes in these ratios between data refreshes. A cold start on responses generated by the teacher can first improve conditional overlap between the teacher and student (Li et al., 2026a), after which the initialized student can serve as π_{old} for the next round of data collection and updates. The upper bounds also hold for any fixed target q replacing π_{T}, with the corresponding substitutions in the assumptions, losses, and policy comparisons.

3.2 A Lower Bound from Unobserved Prefixes

The upper bound shows how prefix distribution shift affects transfer for a given policy. The next result shows that this dependence reflects a fundamental information limit when supervision is restricted to trajectories generated by π_{old}. Even with the teacher's exact next-token distributions at all observed prefixes, the next-token distribution at an unobserved prefix can remain unknown.

Theorem 2 (Proved in Appendix A.2)**.** *For all integers $n, m \geq 1$ and every $0 < \eta < \frac{1}{2}$, choose positive numbers $r_1, \ldots, r_m$ and $P_1, \ldots, P_m$ satisfying $\sum_j r_j < 1$ and $\sum_j P_j < 1$. There exists a standard OPD instance with prefixes $A_1, \ldots, A_m$, a student π_{old} that generates the training trajectories, a family of teachers $\{\pi_{\text{T}}^\sigma : \sigma \in \{-1, +1\}^m\}$, and a constant $c_\eta > 0$. These policies have common support. A trajectory generated by π_{old} reaches A_j with probability r_j, whereas a trajectory generated by any π_{T}^σ reaches it with probability P_j. For teacher π_{T}^σ, let D_n contain n independent trajectories from π_{old}, together with the teacher's exact next-token distribution $\pi_{\text{T}}^\sigma(\cdot \mid s)$ at each prefix s in those trajectories. For any learner $\mathcal{A}$, set $\pi_n := \mathcal{A}(D_n)$, and let $\mathbb{E}_\sigma$ average over D_n and any learner randomization. Then*

$$\inf_{\mathcal{A}} \sup_{\sigma \in \{-1,+1\}^m} \mathbb{E}_\sigma \left[\mathcal{E}_{\pi_{\text{T}}^\sigma}(\pi_n) \right] \geq c_\eta \sum_{j=1}^{m} P_j (1 - r_j)^n. \tag{5}$$

For this single prefix construction with fixed η, when $r = \Theta(\frac{1}{n})$, its expected number of appearances among n training trajectories is $O(1)$. Given the upper bound in Theorem 1, for $P \in [r, 1)$, the exact minimax characterization above and the proof in Appendix A.2 give matching upper and lower bounds on this family's minimax expected KL risk:

$$\Theta\left(\frac{1}{n} + \sqrt{\frac{\chi^2(P_{\pi_{\text{T}}^\sigma} \| P_{\pi_{\text{old}}})}{n}} \right). \tag{6}$$

Here the chi-squared divergence compares the teacher with the student that collected the data. The upper bound in Eq. (4) compares the candidate student with that same sampling policy. These divergences coincide for students that copy the teacher's distribution of the first token in this construction with two steps. Eq. (6) exhibits the same dependence on sample size and distribution shift as the upper bound, so the distribution shift term cannot generally be removed. The difficulty is most pronounced when a prefix is common under the teacher but rarely sampled during training. If training misses this prefix, the constructed teachers agree on all observed feedback but disagree on how to continue after it. Therefore, even perfect fitting of that feedback cannot identify which continuation to imitate. Generating the prefix risks a mismatch in the continuation, while reducing its frequency to avoid this risk instead creates a mismatch in how often the student and teacher generate the prefix. Even the best balance leaves KL error of order P for some teacher consistent with the data whenever the prefix is missed, which occurs with probability $(1 - r)^n$. Thus the lower bound remains substantial when nr is small and P is large: further fitting of the same data cannot distinguish these teachers, motivating additional trajectories that reach the missing prefixes.

4 Pass@k Performance under OPD

We study how accurately an OPD student preserves the teacher's Pass@k performance and how this metric changes under distillation. Pass@k evaluates success with k independent responses per

prompt, providing a standard evaluation of repeated sampling as a form of test-time scaling (Brown et al., 2024). We first relate the student–teacher Pass@k gap to the KL error analyzed in Section 3, with upper and lower bounds that match in order when KL error is small. We then show that an exact KL proximal update can reduce this KL error and improve Pass@1 while decreasing Pass@k.

Definition 1 (Pass@k (Chen et al., 2021))**.** A binary correctness function is a map $c(X, Y) \in \{0, 1\}$. For a policy π, let $p_\pi(X) := \Pr_{Y\sim\pi(\cdot|X)}[c(X, Y) = 1]$ be the probability that one response to prompt X is correct. Its pass rate after k independent responses is

$$\mathsf{Pass}_k(\pi) := \mathbb{E}_X[1 - (1 - p_\pi(X))^k]. \tag{7}$$

For a fixed prompt, $(1 - p_\pi(X))^k$ is the probability that all k responses fail. Averaging the complement over prompts gives Pass@k, which is nondecreasing in k for a fixed policy. Increasing k makes even a small success probability useful, while prompts that are already solved with high probability have less room to improve. Analyses of inference scaling (Schaeffer et al., 2025; Levi, 2026) also emphasize that the distribution of success across prompts matters beyond Pass@1.

4.1 Bounds on the Student-Teacher Pass@k Gap

The generalization bound in Section 3 controls the student's expected KL divergence from the teacher, $\mathcal{E}_{\pi_\mathrm{T}}(\pi)$, defined in Eq. (1). We now quantify what this implies for their Pass@k difference at a given k. The following upper bound applies to any pair of policies with the same prompt distribution. The OPD analysis supplies a bound on the KL error appearing in this guarantee.

Theorem 3 (Proved in Appendix A.3)**.** *Let π_T be a teacher and π a student policy. If $\mathcal{E}_{\pi_\mathrm{T}}(\pi) \le \varepsilon$ for some $\varepsilon \ge 0$, then for every integer $k \ge 1$, $\mathsf{Pass}_k(\pi) \ge \mathsf{Pass}_k(\pi_\mathrm{T}) - \min\{1, \frac{1}{2}\sqrt{2k\varepsilon}\}$.*

The same error term bounds how far the student's Pass@k can exceed that of the teacher, so the guarantee controls the gap in both directions. A KL error bound of ε therefore guarantees similar student and teacher performance when $k\varepsilon$ is small. To retain the same worst-case guarantee as k increases, the KL error bound must decrease proportionally to $\frac{1}{k}$. The next theorem shows that the dependence on $\sqrt{k\varepsilon}$ cannot generally be improved in the stated regime of small KL error.

Theorem 4 (Proved in Appendix A.4)**.** *For every integer $k \ge 1$, there exist a prompt distribution and a teacher $\widetilde{\pi}_\mathrm{T}$ on responses $Y \in \{0, 1\}$ with correctness $c(X, Y) = Y$, such that for every $0 < \varepsilon \le \frac{1}{80k}$, a policy $\widetilde{\pi}$ with common support satisfies $\mathcal{E}_{\widetilde{\pi}_\mathrm{T}}(\widetilde{\pi}) = \varepsilon$ and $\mathsf{Pass}_k(\widetilde{\pi}) \le \mathsf{Pass}_k(\widetilde{\pi}_\mathrm{T}) - \frac{3}{40}\sqrt{k\varepsilon}$.*

Remark. At a high level, for $k \ge 2$, the construction lowers success on a harder prompt with teacher success probability of order $\frac{1}{k}$ and raises it by the same amount on an equally likely easier prompt, preserving Pass@1. The easier prompt is already likely to be solved within k attempts, so its gain contributes less to Pass@k than the loss of rare correct responses on the harder prompt.

Theorem 4 constructs a student whose Pass@k is below the teacher's by $\Omega(\sqrt{k\varepsilon})$, matching the $O(\sqrt{k\varepsilon})$ upper bound in Theorem 3 for $0 < \varepsilon \le \frac{1}{80k}$. Together, these results show that, for each $k \ge 1$ and $\varepsilon > 0$, the worst-case absolute Pass@k gap over prompt distributions, binary correctness functions, and policy pairs with common support and KL error at most ε is $\Theta(\min\{1, \sqrt{k\varepsilon}\})$, with universal implicit constants. For $\varepsilon \ge \frac{1}{80k}$, the lower bound follows by retaining the construction at KL error $\frac{1}{80k}$, which gives a gap bounded below by a universal positive constant.

4.2 Pass@1 Gains Can Accompany Pass@k Losses

Empirical OPD studies report improvements at small sampling budgets alongside declines at larger budgets (Ge et al., 2026). We examine whether lower KL error and higher Pass@1 can coexist with lower Pass@k under a distillation update. Here we compare the updated student with its initial policy at the same k. We consider an exact KL proximal update of the population objective, balancing agreement with the teacher against departure from the initial student. The construction below exhibits all three changes even when the teacher is more accurate on every prompt.

Theorem 5 (Proved in Appendix A.5)**.** *Let π_0 be the initial student and π_{T} the teacher, with common support on a finite response space, and set $\varepsilon_0 := \mathcal{E}_{\pi_{\mathrm{T}}}(\pi_0)$. For a teacher weight $\alpha \in (0,1)$, define the updated student by optimizing over all policies ν:*

$$\pi_+ \in \arg\min_{\nu} \{\alpha \mathcal{E}_{\pi_{\mathrm{T}}}(\nu) + (1-\alpha)\mathbb{E}_X \mathrm{KL}(\nu(\cdot \mid X) \| \pi_0(\cdot \mid X))\}. \tag{8}$$

Then $\mathcal{E}_{\pi_{\mathrm{T}}}(\pi_+) \le \varepsilon_0$. Furthermore, for $\alpha = \frac{1}{2}$, every integer $k \ge 2$, and every $0 < \varepsilon_0 \le \frac{1}{40k}$, there exist a prompt distribution, an initial student, and a teacher with common support and initial KL ε_0 whose update satisfies

$$\mathsf{Pass}_1(\pi_+) \ge \mathsf{Pass}_1(\pi_0) + \frac{\varepsilon_0}{48}, \qquad \mathsf{Pass}_k(\pi_+) \le \mathsf{Pass}_k(\pi_0) - \frac{k\varepsilon_0}{160}. \tag{9}$$

Proof sketch. For $\alpha = \frac{1}{2}$, the proof uses a hard prompt of probability $w \in (0, \frac{1}{4}]$ and an easy prompt of probability $1-w$, with initial success probabilities $u = \frac{1}{10k}$ and $\frac{1}{2}$. We prescribe updated success probabilities $\frac{u}{2}$ and $\frac{1}{2} + s$, where $s = \frac{3wu}{4(1-w)}$. With three responses, two correct, suitable probabilities for individual responses make $\pi_{\mathrm{T}}(y \mid x) \propto \frac{\pi_+(y|x)^2}{\pi_0(y|x)}$ a more accurate teacher on both prompts. The exact update $\pi_+ \propto \sqrt{\pi_0 \pi_{\mathrm{T}}}$ then realizes the prescribed distributions. The construction gives a Pass@1 gain of $\frac{wu}{4}$, a Pass@k loss of order wku, and initial KL of order wu. Choosing w to match the prescribed KL ε_0 yields the bounds in Eq. (9), completing the construction.

In the construction, a teacher that is more accurate on every prompt assigns little probability to the student's most likely correct response on the harder prompt, so the update suppresses this response and reduces success there. Gains on an easier prompt that is already likely to be solved within k attempts raise average Pass@1, while contributing too little to Pass@k to offset this loss. Eq. (9) quantifies this imbalance even arbitrarily close to the teacher. This suggests evaluating checkpoints at the intended k and checking teacher feedback on correct student responses (Akhondzadeh et al., 2026).

5 Experiments

5.1 Synthetic Experiments

Experimental setup. We evaluate the constructions in our proofs exactly on finite response spaces. We specify prompt and response distributions explicitly, with teacher and student policies given by probability tables. The generalization construction controls how often training data visit an informative prefix. The Pass@k constructions use two prompts with designated correct responses. We vary sample size and prefix divergence to test generalization, then vary KL error and sampling budget to test the Pass@k bounds and the trade-off under an exact proximal update. Appendix B gives the distributions, parameter grids, and evaluation procedures.

Sample size and divergence. Figure 1 shows the exact minimax risk of the single prefix family in Appendix A.2, whose matching upper and lower bounds are given in Eq. (6). The first term in that equation predominates when the divergence is much smaller than $\frac{1}{n}$, and the second when it is much larger. The curve with zero divergence follows n^{-1}. At fixed positive divergence, the second term eventually dominates, producing $n^{-\frac{1}{2}}$ decay. The curves with divergence 0.25 and 4 illustrate this regime. These trends agree with the dependence on sample size and divergence in Theorems 1-2.

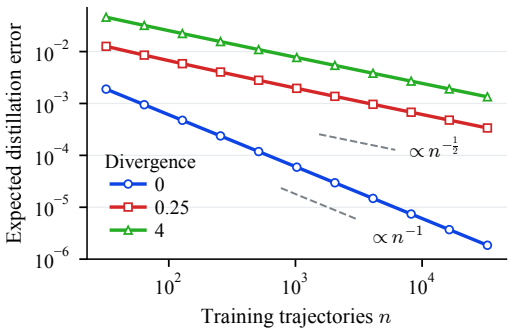


Figure 1: Exact expected error for the single prefix construction of Theorem 2. The vertical axis is $\mathbb{E}_{D_n}[\mathcal{E}_{\pi_{\mathrm{T}}}(\pi_n)]$. "Divergence" is $\chi^2(P_{\pi_{\mathrm{T}}} \| P_{\pi_{\mathrm{old}}})$, fixed as the instance varies with n. Both axes use base-10 log scales. Dashed lines show n^{-1} and $n^{-\frac{1}{2}}$ decay.

Pass@k at small distillation error. Accurate imitation motivates studying small distillation error. Section 3 gives conditions under which low training loss yields small expected error. In this regime, Theorems 3-4 give matching worst-case

bounds of order $\sqrt{k\varepsilon}$ for the student-teacher Pass@k gap. All 88 instances satisfy $\varepsilon \leq \frac{1}{80k}$. As in the lower bound, the family varies with k. Figure 2(a) exhibits approximately $\sqrt{\varepsilon}$ scaling at each fixed k, despite equal Pass@1. In Figure 2(b), ε is approximately proportional to $\frac{1}{k}$ at a fixed gap, consistent with the $\sqrt{k\varepsilon}$ dependence.

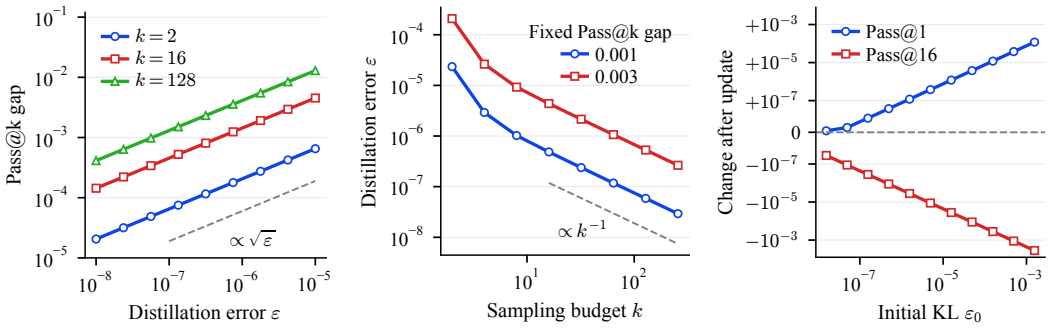
Pass@1 can improve while Pass@k declines. In all 88 instances of the proximal update, the teacher is more accurate on both prompts, yet the update reduces KL, improves Pass@1, and lowers Pass@k, as predicted by Theorem 5. Figure 2(c) shows that, for $k = 16$, the changes retain opposite signs as initial KL decreases through five orders of magnitude, while their magnitudes approach zero. The construction lowers expected KL by reducing the weight of the harder prompt while preserving the conditional policies at each prompt. Its opposing performance changes expose how gains on an easier prompt and losses on a harder one affect the two sampling budgets differently.

5.2 Experiments on LLMs

Experimental setup. We use Qwen3-0.6B-Base, Qwen3-1.7B-Base, and Qwen3-4B-Base (Yang et al., 2025). For each size, we train a GRPO teacher (Shao et al., 2024) on DAPO-Math-17k (Yu et al., 2025) for two epochs, freeze it, and apply OPD (Li et al., 2026a) to the corresponding Base student. The main figures report the first of two OPD epochs. The appendix analyzes both epochs. We evaluate every 20 iterations and at each epoch end on AIME24, AIME25 (MAA, 2025), and AMC23 (MAA, 2023), with 64 responses per problem. We write Ct and π_t for the checkpoint and student policy after t training iterations. Appendix C.1 details the hyperparameters, benchmarks, sampling and grading protocols, evaluation metrics, and prompts.

Training and held-out performance. With 64 responses on each of 100 evaluation problems, Figure 3 shows decreasing training loss at all three model sizes, while test error reaches different plateaus. Appendix C.2 estimates the prefix chi-squared divergence $\chi^2(P_\pi \| P_{\pi_{\text{old}}})$ in Eq. (4) for the transition from C120 to C139, with probability ratios truncated at 10. All nine estimates lie between 1.624 and 6.523, which are small relative to the nominal cumulative count of $n = 128 \times 139 = 17{,}792$ training responses at C139. Furthermore, at the end of training, the 0.6B and 4B students are within one percentage point of their teachers on all three benchmarks, whereas the 1.7B student exceeds its teacher's accuracy. These observations suggest that students' performance can approach and sometimes exceed their teachers' as training progresses (Li et al., 2026a). The approach to teacher-level accuracy aligns with Theorems 1–2, which relate KL error relative to the teacher to training fit, sample size, and prefix chi-squared divergence. Finally, Appendix C.3 fits the signed student–teacher error gaps over both epochs. These inverse square root fits yield Pearson $r > 0.9$ on seven of nine curves and show agreement with the $n^{-1/2}$ term in the theoretical KL bounds.

A potential trade-off between Pass@1 and Pass@k. Figure 4 shows two selected pairs of consecutive evaluation checkpoints during Qwen3-4B training in which small gains in Pass@1 accompany



(a) Gap versus error. (b) Error at a fixed gap. (c) Pass@1 gain, Pass@16 loss.

Figure 2: Pass@k gaps and proximal updates. (a) The gap $\mathsf{Pass}_k(\pi_{\mathrm{T}}) - \mathsf{Pass}_k(\pi)$ versus distillation error ε, at equal Pass@1. (b) Distillation error at fixed gaps of 0.001 and 0.003. (c) Signed changes $\mathsf{Pass}_j(\pi_+) - \mathsf{Pass}_j(\pi_0)$, for $j = 1$ (blue) and $j = 16$ (red). All horizontal axes and the vertical axes in (a)-(b) use base-10 log scales. The vertical axis in (c) is linear near zero and logarithmic in magnitude elsewhere. Dashed lines mark $\sqrt{\varepsilon}$ scaling in (a), $\frac{1}{k}$ scaling in (b), and zero change in (c).

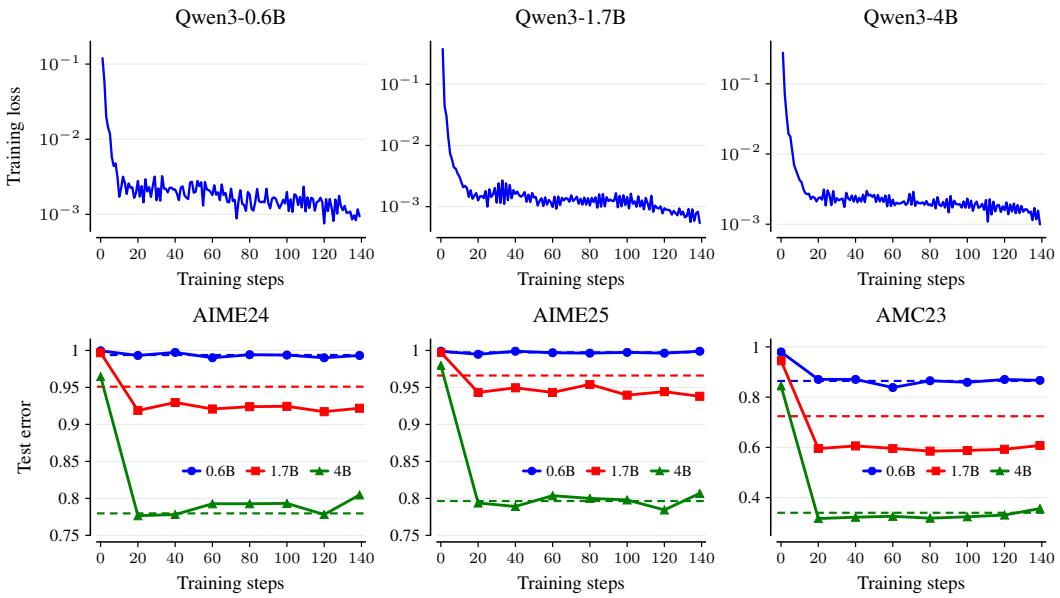


Figure 3: First epoch of OPD with a GRPO teacher of matching size. Top: training loss on a log scale for Qwen3-0.6B, 1.7B, and 4B. Bottom: test error, $1-\text{Pass@}1$, on AIME24, AIME25, and AMC23. Solid curves show student checkpoints, and dashed horizontal lines show their corresponding frozen teachers. Each error estimate uses 64 responses per problem.

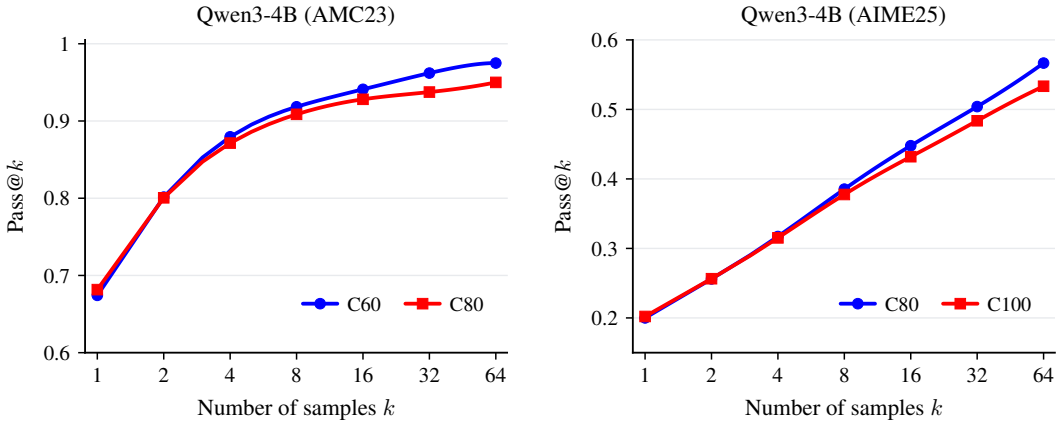


Figure 4: Pass@k for two selected pairs of consecutive Qwen3-4B evaluation checkpoints. Left: C60 and C80 on AMC23. Right: C80 and C100 on AIME25. The horizontal axes show sampling budget k. Each displayed budget uses the same 64 responses per problem. The curves cross between $k = 1$ and 2 on AMC23 and between $k = 2$ and 3 on AIME25.

substantially larger losses in Pass@64. On AMC23, C60 $\rightarrow$ C80 raises Pass@1 from 67.42% to 68.16% while Pass@64 falls from 97.50% to 95.00%. On AIME25, C80 $\rightarrow$ C100 raises Pass@1 from 20.00% to 20.21% while Pass@64 falls from 56.67% to 53.33%. Despite their higher Pass@1, the later checkpoints have lower Pass@k at every integer budget $k = 2, \ldots, 64$ on AMC23 and $k = 3, \ldots, 64$ on AIME25. This possibility is formalized in Theorem 5, which constructs an exact KL proximal update that reduces divergence to the teacher and improves Pass@1 while decreasing Pass@k. The theorem thus shows that closer imitation and higher Pass@1 can occur without improvement at every sampling budget, providing a theoretical counterpart to the observed checkpoint crossings. Appendix C.4 reports four further instances of higher Pass@1 and lower Pass@64 among 42 adjacent checkpoint comparisons for 0.6B and 1.7B students trained with an 8B teacher.

Examining the observed trade-off. To delve into the trade-off describe above, we compare each problem's correct count among 64 responses at the two checkpoints in Figure 4 and calculate its

Table 1: 4B checkpoint changes by problem group, with 64 responses per problem. Correct counts are totals before and after the update. Red denotes decreases and blue increases.

Problem group	Problems	Correct counts	ΔPass@1	ΔPass@32	ΔPass@64
AMC23 (C60 → C80)					
Easy problems	20	987 → 1058	**+2.77%**	0.00%	0.00%
Hard problems	2	4 → 1	−0.12%	**−2.52%**	**−2.50%**
Newly solved problems	0	0 → 0	0.00%	0.00%	0.00%
Other problems	18	735 → 686	−1.91%	+0.06%	0.00%
Total	40	1726 → 1745	+0.74%	−2.46%	−2.50%
AIME25 (C80 → C100)					
Easy problems	5	156 → 172	**+0.83%**	+0.01%	0.00%
Hard problems	3	9 → 3	−0.31%	**−3.71%**	**−6.67%**
Newly solved problems	1	0 → 1	+0.05%	+1.67%	+3.33%
Other problems	21	219 → 212	−0.36%	0.00%	0.00%
Total	30	384 → 388	+0.21%	−2.04%	−3.33%

contribution to changes in Pass@1, Pass@32, and Pass@64. In these evaluation samples, total correct responses increase, while the number of problems with at least one correct response falls from 39 to 38 on AMC23 and from 17 to 16 on AIME25. We discuss how *gains on easy problems offset losses in Pass@1 while losses on hard problems drive the decline in Pass@32 and Pass@64*. Table 1 summarizes the counts and sums the contributions over four groups:

1. **Easy problems.** Each problem has an estimated Pass@32 of at least 99% at the earlier checkpoint and more correct responses at the later checkpoint.
2. **Hard problems.** Each problem has 1–7 correct responses out of 64 at the earlier checkpoint and fewer correct responses at the later checkpoint.
3. **Newly solved problems.** Each problem has no correct responses at the earlier checkpoint and at least one correct response at the later checkpoint.
4. **Other problems.** These problems do not meet any of the three definitions above.

These groups use observed model performance and its change between checkpoints. On AMC23, the 20 easy problems gain 71 correct responses, raising Pass@1, while their Pass@32 and Pass@64 gains are negligible. The two hard problems lose only three correct responses, yet account for almost the entire Pass@32 loss and the full Pass@64 loss. AIME25 exhibits the same pattern, with one newly solved problem contributing 3.33 percentage points to Pass@64 and offsetting half of the loss from hard problems. The remaining problems include gains and losses, so their combined contribution can change sign across budgets. This allocation of gains and losses parallels the construction of Theorem 5. Appendix C.5 further analyzes how gains and losses on individual problems affect estimated Pass@k. The key takeaway is that, one more correct response matters most on a problem with few existing successes. This explains how losing a few successes on hard problems can outweigh more numerous gains on easy problems at larger k, even as Pass@1 improves.

6 Conclusion

In this paper, we connect finite on-policy supervision to generalization and Pass@k performance. Our bounds relate distillation loss on the training data to error on the student's own generations through differences in their prefix distributions. Maintaining the same worst-case guarantee on the student-teacher Pass@k gap as k increases requires more accurate imitation, even when their Pass@1 accuracies match. We also prove that an exact KL proximal update can bring the student closer to a teacher that is more accurate on every prompt and improve Pass@1 while reducing Pass@k at a given $k \geq 2$. The construction shows how teacher preferences can suppress correct responses on harder prompts, with gains on easier prompts insufficient to offset the Pass@k loss. Synthetic and LLM experiments support our theory and the mechanisms underlying our proofs.

AI USE STATEMENT

Generative AI tools were used to assist with the design and critical review of the experimental protocol, literature search and summarization, manuscript drafting and language editing, and LaTeX formatting. They also assisted with implementing, executing, and analyzing the synthetic simulations in Section 5.1. The authors checked the cited sources against the original papers, verified the mathematical statements and proofs, and reviewed all AI-assisted text. The authors take responsibility for all text, claims, and artifacts in this work, including those produced with generative AI.

ETHICS STATEMENT

This work develops a theoretical analysis and simulations on finite distributions of on-policy distillation. It does not involve human subjects, interaction with users, or the collection of personal or sensitive data. The simulations use explicitly specified synthetic policies and correctness functions, without pretrained models or external datasets. The main foreseeable risks are dual use of improved distillation methods, the computational cost of repeated sampling, and drawing overly strong conclusions about model capabilities from constructed examples. We report the complete parameter grids and sampling protocols, distinguish exact calculations from Monte Carlo estimates, and limit conclusions to the models studied. We have read and will adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The assumptions of our theoretical results are stated in Sections 3 and 4, with complete proofs in Appendix A. Appendix B specifies the finite distributions, parameter grids, and exact evaluation procedures for the synthetic experiments. Appendix C.1 describes the language models, training settings, benchmarks, prompts, and evaluation protocols. Appendices C.2–C.5 give the divergence estimates, problem counts, budget comparisons, teacher baselines, and descriptive error-gap fits.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *ICLR*, 2024.

Mohammad Sadegh Akhondzadeh, Vijay Lingam, Atula Tejaswi, Chanakya Ekbote, Sujay Sanghavi, and Aleksandar Bojchevski. Reward-gated on-policy distillation. *arXiv preprint arXiv:2607.04037*, 2026.

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

Mohammadreza Armandpour, Fatih Ilhan, David Harrison, Ajay Jaiswal, Duc N. M. Hoang, Fartash Faghri, Yizhe Zhang, Minsik Cho, and Mehrdad Farajtabar. Unmasking on-policy distillation: Where it helps, where it hurts, and why. *arXiv preprint arXiv:2605.10889*, 2026.

Anas Barakat, Souradip Chakraborty, Khushbu Pahwa, and Amrit Singh Bedi. Why Pass@k optimization can degrade Pass@1: Prompt interference in LLM post-training. *arXiv preprint arXiv:2602.21189*, 2026.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@k training for adaptively balancing exploration and exploitation of large reasoning models. *arXiv preprint arXiv:2508.10751*, 2025.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, second edition, 2006.

Xingyu Dang, Christina Baek, Kaiyue Wen, Zico Kolter, and Aditi Raghunathan. Weight ensembling improves reasoning in language models. In *COLM*, 2025.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Shiyuan Feng, Huan-ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Weak-to-strong generalization via direct on-policy distillation. *arXiv preprint arXiv:2607.05394*, 2026.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Xinmu Ge, Zizhuo Zhang, Yu Huang, Jianing Zhu, Lin Yuan, Wanli Gu, Weichang Wu, Weiran Huang, Xiaolu Zhang, Bo Han, Jun Zhou, and Jiangchao Yao. Towards understanding on-policy distillation through the lens of test-time scaling. *arXiv preprint arXiv:2608.11829*, 2026.

GLM-5 Team. GLM-5: From vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *ICLR*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D. Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J. Foster. Correcting the mythos of KL-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *ICLR*, 2025.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of ACL*, 2026.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *ICML*, 2026.

Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *ICML*, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *ICML*, 2025.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

Noam Itzhak Levi. A simple model of inference scaling laws. In *ICML*, 2025.

Noam Levi. Learning shrinks the hard tail: Training-dependent inference scaling in a solvable linear model. In *ICLR*, 2026.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. In *ICML Workshop on Foundations of Deep Generative Models*, 2026.

Zhaoyi Li, Deyang Kong, Yuan Wei, Evan Yang, Ranran Shen, Mahardika Krisna Ihsani, Ming Yang, Wei Zhang, Chuan Hao, Jian Yang, Ran Tao, Bryan Dai, Shikun Zhang, Wei Ye, Ying Wei, and Defu Lian. Every coin has two sides: On the dual nature of generalization in on-policy distillation of large language models. *arXiv preprint arXiv:2608.16647*, 2026.

Baohao Liao, Hanze Dong, Christof Monz, Xinxing Xu, Li Dong, and Furu Wei. Multi-turn on-policy distillation with prefix replay. *arXiv preprint arXiv:2607.04763*, 2026.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *EMNLP*, 2020.

Kaiyuan Liu, Ziyuan Zhuang, Yang Bai, Bing Wang, Rongxiang Weng, and Jieping Ye. Prefix teach, suffix fade: Local teachability collapse in strong-to-weak on-policy distillation. *arXiv preprint arXiv:2605.13643*, 2026.

Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew Gordon Wilson. PAC-Bayes compression bounds so tight that they can explain generalization. In *NeurIPS*, 2022.

Sanae Lotfi, Marc Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. In *ICML*, 2024.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. In *ICML*, 2026.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)? In *ICML*, 2025.

Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot-Desenonges, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shahriari, Sarah Perrin, Abram Friesen, Geoffrey Cideron, Sertan Girgin, Piotr Stanczyk, Andrea Michi, Danila Sinopalnikov, Sabela Ramos Garea, Amélie Héliou, Aliaksei Severyn, Matthew Hoffman, Nikola Momchev, and Olivier Bachem. BOND: Aligning LLMs with best-of-N distillation. In *ICLR*, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *ICLR*, 2025.

Ved Sriraman, Peihan Liu, Daniel Hsu, and Adam Block. Behavior cloning is not all you need: The optimality of on-policy distillation for noisy expert feedback. *arXiv preprint arXiv:2606.30923*, 2026.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Christian Walder and Deep Tejas Karkhanis. Pass@K policy optimization: Solving harder reinforcement learning problems. In *NeurIPS*, 2025.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026.

Yunbei Xu, Yuzhe Yuan, and Ruohan Zhan. Autoregressive learning in joint KL: Sharp oracle bounds and lower bounds. *arXiv preprint arXiv:2605.12316*, 2026.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, and Zhipeng Wang. PACED: Distillation and on-policy self-distillation at the frontier of student competence. *arXiv preprint arXiv:2603.11178*, 2026.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

Run Yang, Runpeng Dai, Jie Sun, Jielei Zhang, Fan Zhou, Hongtu Zhu, Peiyi Li, and Longwen Gao. Influence-directed distillation: Solving the diversity bottleneck in sampled-token on-policy distillation. *arXiv preprint arXiv:2608.29846*, 2026.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025.

Yang Yu. $\mathrm{Pass}@k$ metric for RLVR: A diagnostic tool of exploration, but not an objective. *arXiv preprint arXiv:2511.16231*, 2025.

Suqin Yuan, Jinkun Chen, Jiyang Zheng, Muyang Li, Lei Feng, Dadong Wang, Tao Xiang, Tongliang Liu, and Bo An. Understanding diversity collapse in RLVR via the lens of overtraining. *arXiv preprint arXiv:2606.15455*, 2026.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason D. Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *COLT*, 2022.

Xinsen Zhang, Zhenkai Ding, Tianjun Pan, Run Yang, Chun Kang, Xue Xiong, and Jingnan Gu. OPSDL: On-policy self-distillation for long-context language models. *arXiv preprint arXiv:2604.17535*, 2026.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D. Lyng, Sanjit Singh Batra, and Robert E. Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of ACL*, 2026.

Huaqing Zhang, Jingchu Gai, Juno Kim, Bingbin Liu, and Andrej Risteski. When does online imitation learning help in LLM post-training? The role of (non-)realizability beyond horizon. *arXiv preprint arXiv:2606.30445*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *arXiv preprint arXiv:2605.11182*, 2026.

Yuxuan Zhu, Rohan Alur, and Daniel Kang. Non-vacuous generalization bounds for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2607.14506*, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *SOSP*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *NeurIPS*, 2025.

Mathematical Association of America. American Mathematics Competitions (AMC). 2023.

Mathematical Association of America. American Invitational Mathematics Examination (AIME). 2025.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

APPENDIX CONTENTS

A Proofs 16

A.1 Proof of Theorem 1 16
A.2 Proof of Theorem 2 18
A.3 Proof of Theorem 3 21
A.4 Proof of Theorem 4 21
A.5 Proof of Theorem 5 22

B Synthetic Experimental Details 24

C Large Language Model Experimental Details 24

C.1 Detailed Experimental Setup 24
C.2 Chi-Squared Divergence between Checkpoint Prefix Distributions 26
C.3 Teacher Baselines and Error-Gap Fits 27
C.4 Additional Examples from Separate OPD Runs 29
C.5 Sampling Budgets and Individual Problems in the Selected Crossings 30

A PROOFS

A.1 PROOF OF THEOREM 1

Proof. We first prove the bound when the original conditional KL losses are bounded by G, so clipping has no effect. The same argument for clipped losses is given below. We first bound the loss on prefixes generated by π_{old}, then on those generated by the student. The class $\mathcal{C}$ and the encoding in Theorem 1 are fixed independently of the sample. Since the encoding is prefix-free and $K(\pi)$ is its code length in nats, Kraft's inequality gives $\sum_{\pi\in\mathcal{C}} e^{-K(\pi)} \le 1$ (Langford, 2005). The code may use a checkpoint fixed independently of the current sample as shared side information. All additional information needed to reconstruct the candidate, including any encoding choices made using the data, must be included in $K(\pi)$.

Loss on prefixes generated during training. The policy π_{old} generates the training responses. These responses are independent, but token losses within a response can be dependent. We apply Bernstein's inequality to the loss summed over each response. If $V_1, \dots, V_n$ are independent copies of $V \in [0, M]$, then for every $u > 0$, with probability at least $1 - e^{-u}$,

$$\mathbb{E}V - \frac{1}{n}\sum_{i=1}^{n} V_i \le \sqrt{\frac{2\mathrm{Var}(V)u}{n}} + \frac{Mu}{3n}. \tag{10}$$

If $\mathrm{Var}(V) = 0$, the sample mean equals $\mathbb{E}V$, so the bound holds. Otherwise, apply Boucheron et al. (2013, Theorem 2.10) to the centered variables $\mathbb{E}V - V_i$ and divide by n. Each variable has variance $\mathrm{Var}(V)$ and absolute value at most M.

Fix a candidate policy $\pi \in \mathcal{C}$. For training trajectory i, define

$$X_{i,\pi} = \sum_{h=1}^{H} \mathrm{KL}(\pi(\cdot \mid S_{i,h}) \| \pi_{\mathrm{T}}(\cdot \mid S_{i,h})) .$$

The variables $X_{1,\pi}, \dots, X_{n,\pi}$ are independent because each sums losses from a separate training response. By definition,

$$\mathbb{E}X_{i,\pi} = \mathcal{R}_{\pi_{\text{old}}}(\pi), \qquad \frac{1}{n}\sum_{i=1}^{n} X_{i,\pi} = \widehat{\mathcal{R}}_{\pi_{\text{old}},n}(\pi).$$

Each token loss lies in $[0, G]$, so $0 \le X_{i,\pi} \le HG$. The inequality $X_{i,\pi}^2 \le HGX_{i,\pi}$ gives

$$\mathrm{Var}(X_{i,\pi}) \le \mathbb{E}X_{i,\pi}^2 \le HG\,\mathcal{R}_{\pi_{\text{old}}}(\pi). \tag{11}$$

Apply Eq. (10) with $M = HG$ and $u = K(\pi) + \log\frac{1}{\delta}$. For this fixed policy, the bound fails with probability at most

$$e^{-K(\pi)-\log\frac{1}{\delta}} = \delta e^{-K(\pi)}.$$

A union bound over $\mathcal{C}$ gives total failure probability at most

$$\sum_{\pi\in\mathcal{C}} \delta e^{-K(\pi)} = \delta \sum_{\pi\in\mathcal{C}} e^{-K(\pi)} \le \delta,$$

where the last step uses Kraft's inequality. Substituting Eq. (11) into Bernstein therefore proves, simultaneously for every $\pi \in \mathcal{C}$,

$$\mathcal{R}_{\pi_{\text{old}}}(\pi) \le \widehat{\mathcal{R}}_{\pi_{\text{old}},n}(\pi) + \sqrt{\frac{2HG\,\mathcal{R}_{\pi_{\text{old}}}(\pi)\left(K(\pi) + \log\frac{1}{\delta}\right)}{n}} + \frac{HG\left(K(\pi) + \log\frac{1}{\delta}\right)}{3n}.$$

The bound holds simultaneously for all policies in $\mathcal{C}$, including any $\widehat{\pi} \in \mathcal{C}$ chosen using the data.

To remove the unknown loss from the right-hand side, set

$$a := \frac{HG\left(K(\pi) + \log\frac{1}{\delta}\right)}{n}.$$

The elementary inequality

$$\sqrt{2aL} \le \frac{L}{2} + a, \qquad L, a \ge 0,$$

follows because $(\frac{L}{2} + a)^2 - 2aL = (\frac{L}{2} - a)^2 \ge 0$. Apply it with $L = \mathcal{R}_{\pi_{\text{old}}}(\pi)$ and omit the argument π below. Then

$$\mathcal{R}_{\pi_{\text{old}}} \le \widehat{\mathcal{R}}_{\pi_{\text{old}},n} + \frac{\mathcal{R}_{\pi_{\text{old}}}}{2} + a + \frac{a}{3}.$$

Rearranging proves

$$\mathcal{R}_{\pi_{\text{old}}} \le 2\widehat{\mathcal{R}}_{\pi_{\text{old}},n} + \frac{8a}{3} = 2\widehat{\mathcal{R}}_{\pi_{\text{old}},n} + \frac{8HG\left(K(\pi) + \log\frac{1}{\delta}\right)}{3n}.$$

Transfer to prefixes generated by the student. For step h and prefix s, write the token loss as

$$\ell_h(s) := \mathrm{KL}(\pi(\cdot \mid s) \| \pi_{\mathrm{T}}(\cdot \mid s)).$$

The KL chain rule for autoregressive distributions (Cover and Thomas, 2006, Theorem 2.5.3) gives $\mathcal{E}_{\pi_{\mathrm{T}}}(\pi) = \sum_h \mathbb{E}_{P_{\pi,h}} \ell_h$, which is Eq. (1). The finite chi-squared divergence assumed in Section 3.1 implies $P_{\pi,h} \ll P_{\pi_{\text{old}},h}$ at every step: any set of prefixes with probability 0 under π_{old} also has probability 0 under π. Define $w_h(s) := \frac{P_{\pi,h}(s)}{P_{\pi_{\text{old}},h}(s)}$ on the support of $P_{\pi_{\text{old}},h}$. Its value outside that support is irrelevant. Then

$$\begin{aligned}\mathcal{E}_{\pi_{\mathrm{T}}}(\pi) &= \sum_{h=1}^{H} \mathbb{E}_{S_h \sim P_{\pi_{\text{old}},h}} \left[w_h(S_h)\ell_h(S_h)\right] \\ &= \mathcal{R}_{\pi_{\text{old}}}(\pi) + \sum_{h=1}^{H} \mathbb{E}_{S_h \sim P_{\pi_{\text{old}},h}} \left[(w_h(S_h) - 1)\ell_h(S_h)\right].\end{aligned}$$

The second equality uses the definition of $\mathcal{R}_{\pi_{\text{old}}}(\pi)$. Apply Cauchy–Schwarz under the joint distribution $P_{\pi_{\text{old}}}$, then multiply both sides by H, to obtain

$$\left|\sum_{h=1}^{H} \mathbb{E}_{P_{\pi_{\text{old}},h}}[(w_h - 1)\ell_h]\right| \le \sqrt{\sum_{h=1}^{H} \mathbb{E}_{P_{\pi_{\text{old}},h}}(w_h - 1)^2} \sqrt{\sum_{h=1}^{H} \mathbb{E}_{P_{\pi_{\text{old}},h}} \ell_h^2}.$$

The ratio $\frac{P_\pi(h,s)}{P_{\pi_{\text{old}}}(h,s)}$ equals $w_h(s)$, since the factors $\frac{1}{H}$ cancel. By the definition of chi-squared divergence,

$$\sum_{h=1}^{H} \mathbb{E}_{P_{\pi_{\text{old}},h}}(w_h - 1)^2 = H\chi^2(P_\pi \| P_{\pi_{\text{old}}}). \tag{12}$$

For the second factor, $0 \le \ell_h \le G$ implies $\ell_h^2 \le G\ell_h$, and therefore

$$\sum_{h=1}^{H} \mathbb{E}_{P_{\pi_{\text{old}},h}} \ell_h^2 \le G \sum_{h=1}^{H} \mathbb{E}_{P_{\pi_{\text{old}},h}} \ell_h = G\mathcal{R}_{\pi_{\text{old}}}(\pi).$$

Using these two bounds gives

$$\mathcal{E}_{\pi_{\mathrm{T}}}(\pi) \le \mathcal{R}_{\pi_{\text{old}}}(\pi) + \sqrt{HG\mathcal{R}_{\pi_{\text{old}}}(\pi)\chi^2(P_\pi \| P_{\pi_{\text{old}}})}.$$

On the event of probability at least $1 - \delta$ established for the training loss, set

$$B_\pi := 2\widehat{\mathcal{R}}_{\pi_{\text{old}},n}(\pi) + \frac{8HG(K(\pi) + \log(\frac{1}{\delta}))}{3n},$$

so that $\mathcal{R}_{\pi_{\text{old}}}(\pi) \le B_\pi$ for every candidate. For fixed $c \ge 0$, the map $z \mapsto z + \sqrt{cz}$ is nondecreasing on $[0, \infty)$. Substituting B_π for $\mathcal{R}_{\pi_{\text{old}}}(\pi)$ in the preceding inequality and expanding B_π proves Eq. (4) for every candidate on the same event. □

Bounded and clipped KL losses. The finite loss bound follows whenever the fixed teacher satisfies $\pi_{\mathrm{T}}(a \mid s) \geq \varepsilon > 0$ for every token and relevant prefix. Indeed, for every candidate policy,

$$\mathrm{KL}(\pi(\cdot \mid s) \| \pi_{\mathrm{T}}(\cdot \mid s)) \leq \sum_a \pi(a \mid s) \log \frac{1}{\pi_{\mathrm{T}}(a \mid s)} \leq \log \frac{1}{\varepsilon}.$$

A finite vocabulary and a fixed maximum input context length give a finite set of model inputs. If the fixed teacher uses a full softmax with finite logits on this set, such an ε exists. The resulting G can nevertheless be large, so existence alone does not ensure a tight finite sample bound. Alternatively, prediction smoothing as in Lotfi et al. (2024) provides an explicit probability floor. For a vocabulary $\mathcal{V}$ and fixed $\alpha \in (0, 1)$, the target $q_\alpha(\cdot \mid s) = (1 - \alpha)\pi_{\mathrm{T}}(\cdot \mid s) + \alpha \,\mathrm{Unif}(\mathcal{V})$ satisfies $\mathrm{KL}(\pi(\cdot \mid s) \| q_\alpha(\cdot \mid s)) \leq \log(|\mathcal{V}|/\alpha)$. Applying the theorem with this target bounds the KL to q_α. Transferring this bound to the original teacher requires an additional comparison.

For the clipping convention in Theorem 1, keep the sampling, encoding, and finite chi-squared divergence conditions, and allow the original tokenwise KL to be unbounded. Fix a threshold $G > 0$ independently of the sample, and define

$$\ell_\pi(s) := \mathrm{KL}(\pi(\cdot \mid s) \| \pi_{\mathrm{T}}(\cdot \mid s)), \qquad \ell_{\pi,G}(s) := \min\{\ell_\pi(s), G\}.$$

Let

$$\widehat{\mathcal{R}}^{(G)}_{\pi_{\mathrm{old}},n}(\pi) := \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{H} \ell_{\pi,G}(S_{i,h}),$$

$$B_{\pi,G} := 2\widehat{\mathcal{R}}^{(G)}_{\pi_{\mathrm{old}},n}(\pi) + \frac{8HG(K(\pi) + \log(1/\delta))}{3n}.$$

The preceding arguments apply to $\ell_{\pi,G}$, since it lies in $[0, G]$. Their concentration and change-of-measure arguments use only this boundedness and nonnegativity, so with probability at least $1 - \delta$, simultaneously for all candidates,

$$\sum_{h=1}^{H} \mathbb{E}_{S_h \sim P_{\pi,h}} \ell_{\pi,G}(S_h) \leq B_{\pi,G} + \sqrt{HGB_{\pi,G}\chi^2(P_\pi \| P_{\pi_{\mathrm{old}}})}.$$

These are the empirical and population losses used in Eq. (4) under clipping. The identity $\ell_\pi = \ell_{\pi,G} + (\ell_\pi - G)_+$, where $(u)_+ := \max\{u, 0\}$, and the KL chain rule give the following bound for the original, unclipped $\mathcal{E}_{\pi_{\mathrm{T}}}(\pi)$:

$$\begin{aligned} \mathcal{E}_{\pi_{\mathrm{T}}}(\pi) &\leq B_{\pi,G} + \sqrt{HGB_{\pi,G}\chi^2(P_\pi \| P_{\pi_{\mathrm{old}}})} + \mathcal{T}_G(\pi), \\ \mathcal{T}_G(\pi) &:= \sum_{h=1}^{H} \mathbb{E}_{S_h \sim P_{\pi,h}} \left[(\ell_\pi(S_h) - G)_+\right]. \end{aligned} \tag{13}$$

This tail term is zero when the original KL loss never exceeds G. Otherwise, controlling the original KL requires a bound on $\mathcal{T}_G(\pi)$, which requires control of the losses above the clipping threshold.

A.2 Proof of Theorem 2

Proof. Xu et al. (2026a, Theorem 4.2 and Appendix B) use Fano's inequality to prove a KL lower bound for learning from sampled responses. Here we condition on the observed prefixes and compute the remaining KL loss directly.

1. Target family and observed information. There is a single prompt, and each response has two tokens. The possible first tokens are $A_1, \ldots, A_m, B$, and the second token is 0 or 1. We use the same symbols for the prefixes after the first token. The policy π_{old} generates the training responses, choosing $A_1, \ldots, A_m$ with probabilities $r_1, \ldots, r_m$ and B with the remaining probability. Each teacher π^σ_{T} instead uses probabilities $P_1, \ldots, P_m$ and the remainder for B. After A_j, let

$$\pi^\sigma_{\mathrm{T}}(1 \mid A_j) = \frac{1}{2} + \sigma_j \eta, \qquad \sigma_j \in \{-1, +1\}.$$

After B, each teacher chooses 0 and 1 with equal probability. The training policy π_{old} does so after every first token. For standard OPD, π_{old} is the student before this round and π^σ_{T} is its teacher. The

assumed probabilities for the first token and $0 < \eta < \frac{1}{2}$ make every response possible under all policies.

Choose each σ_j independently, with equal probability for either sign. Let $U(D_n)$ contain the indices j for which no training response starts with A_j. The training responses come from the fixed policy π_{old}, so they reveal no information about the signs. Neither does the teacher's distribution for the first token, which is the same for all signs. Teacher feedback after an observed A_j reveals σ_j through the probability $\frac{1}{2}+\sigma_j\eta$. If A_j is absent, all observed feedback is independent of σ_j. Therefore, given D_n, the signs indexed by $U(D_n)$ remain independent and equally likely.

2. Conditional risk and the minimax lower bound. A learner that generates any other response has infinite KL loss, so we consider only the responses defined above. At a prefix that the learner never generates, its next-token distribution contributes 0 to the loss. Fix the data and any random choices made by the learner. Write $\pi_n := \mathcal{A}(D_n)$, let p_j and p_B be its probabilities of starting with A_j and B, and let v_j be its probability of choosing 1 after A_j. Write $\mathbf{p} := (p_1, \dots, p_m, p_B)$ and $\mathbf{P} := (P_1, \dots, P_m, P_B)$, where $P_B := 1 - \sum_j P_j$, for the distributions of the first token. Here $\mathrm{Ber}\,(v)$ assigns probability v to 1 and $1-v$ to 0. By the chain rule for KL divergence (Cover and Thomas, 2006, Theorem 2.5.3), the loss for a fixed σ is

$$\begin{aligned}\mathrm{KL}(\pi_n \| \pi_{\mathrm{T}}^{\sigma}) = \mathrm{KL}(\mathbf{p}\|\mathbf{P}) + \sum_{j=1}^{m} p_j \mathrm{KL}\left(\mathrm{Ber}\,(v_j)\middle\|\mathrm{Ber}\left(\frac{1}{2}+\sigma_j\eta\right)\right) \\ + p_B \mathrm{KL}(\pi_n(\cdot \mid B) \| \pi_{\mathrm{T}}^{\sigma}(\cdot \mid B)).\end{aligned}$$

The last term and the losses after observed prefixes are nonnegative. Dropping them and averaging over the unknown signs bounds $\mathbb{E}[\mathrm{KL}(\pi_n\|\pi_{\mathrm{T}}^{\sigma}) \mid D_n, \pi_n]$ below by

$$\mathrm{KL}(\mathbf{p}\|\mathbf{P}) + \sum_{j \in U(D_n)} p_j \frac{1}{2}\left[\mathrm{KL}\left(\mathrm{Ber}\,(v_j)\middle\|\mathrm{Ber}\left(\frac{1}{2}+\eta\right)\right) + \mathrm{KL}\left(\mathrm{Ber}\,(v_j)\middle\|\mathrm{Ber}\left(\frac{1}{2}-\eta\right)\right)\right]. \tag{14}$$

To bound the bracketed term, put $a = \frac{1}{2}+\eta$, so $1-a = \frac{1}{2}-\eta$. For every $v \in [0,1]$,

$$\begin{aligned}&\frac{1}{2}\left[\mathrm{KL}\left(\mathrm{Ber}\,(v)\middle\|\mathrm{Ber}\left(\frac{1}{2}+\eta\right)\right) + \mathrm{KL}\left(\mathrm{Ber}\,(v)\middle\|\mathrm{Ber}\left(\frac{1}{2}-\eta\right)\right)\right] \\ &= v\log v + (1-v)\log(1-v) - \frac{1}{2}\log(a(1-a)) \\ &\geq -\log 2 - \frac{1}{2}\log(a(1-a)) \\ &= -\frac{1}{2}\log(1-4\eta^2) = d_\eta,\end{aligned} \tag{15}$$

We use $0\log 0 = 0$. The inequality holds because $v\log v + (1-v)\log(1-v)$ is minimized at $v = \frac{1}{2}$. Equality is attained there, and $d_\eta > 0$ because $0 < \eta < \frac{1}{2}$.

Reducing p_j lowers the loss after the unobserved A_j. However, making p_j differ from P_j adds KL for the first token. To minimize the combined loss, define $c_j = d_\eta$ for $j \in U(D_n)$, set $c_j = 0$ for the other prefixes, and set $c_B = 0$. With z ranging over $\{1, \dots, m, B\}$, let

$$Z_U := \sum_z P_z e^{-c_z} = 1 - (1 - e^{-d_\eta}) \sum_{j \in U(D_n)} P_j, \qquad p_z^* := \frac{P_z e^{-c_z}}{Z_U}.$$

Let $\mathbf{p}^*$ have probabilities p_z^*. For any distribution $\mathbf{p}$ of the first token, expanding the KL gives

$$\mathrm{KL}(\mathbf{p}\|\mathbf{P}) + \sum_z p_z c_z = \sum_z p_z \log \frac{p_z}{P_z e^{-c_z}} = \mathrm{KL}(\mathbf{p}\|\mathbf{p}^*) - \log Z_U.$$

Nonnegativity of KL shows that the minimum is attained at $\mathbf{p} = \mathbf{p}^*$, and thus

$$\begin{aligned}\inf_{\mathbf{p}} \left\{\mathrm{KL}(\mathbf{p}\|\mathbf{P}) + d_\eta \sum_{j \in U(D_n)} p_j\right\} &= -\log\left(1 - (1-e^{-d_\eta}) \sum_{j \in U(D_n)} P_j\right) \\ &\geq (1 - e^{-d_\eta}) \sum_{j \in U(D_n)} P_j.\end{aligned} \tag{16}$$

The final inequality uses $-\log(1-z) \geq z$ for $0 \leq z < 1$. This calculation minimizes the combined loss on the two tokens. It also follows from Boyd and Vandenberghe (2004, Example 3.25).

A training response starts with a token other than A_j with probability $1 - r_j$. Independence of the responses gives $\Pr[j \in U(D_n)] = (1 - r_j)^n$. Taking expectations in Eq. (16) shows that every learner's expected loss, averaged over the random signs, is at least

$$(1 - e^{-d_\eta}) \sum_{j=1}^{m} P_j (1 - r_j)^n.$$

Set $c_\eta := 1 - e^{-d_\eta} = 1 - \sqrt{1 - 4\eta^2} > 0$. For every learner, the maximum over σ is at least the uniform average over σ. Taking the infimum over learners preserves this lower bound. This proves Eq. (5). Randomized learners are included because we first fixed their random choices and can now average over them.

3. The single prefix case and chi-squared divergence. For $m = 1$, write $r = r_1$ and $P = P_1$. Before the first token, the prefix is the fixed prompt under both policies, so this step contributes 0 to the chi-squared divergence. After the first token, the prefix probabilities are $(P, 1 - P)$ under $\pi_{\mathrm{T}}^{\sigma}$ and $(r, 1 - r)$ under π_{old}. Averaging the two steps gives

$$\chi^2(P_{\pi_{\mathrm{T}}^{\sigma}} \| P_{\pi_{\mathrm{old}}}) = \frac{1}{2}\left[\frac{(P-r)^2}{r} + \frac{(P-r)^2}{1-r}\right] = \frac{(P-r)^2}{2r(1-r)}. \tag{17}$$

Set $r = \frac{1}{2n}$ and take $P \in [r, 1)$. Bernoulli's inequality gives $(1 - r)^n \geq 1 - nr = \frac{1}{2}$. Eq. (17) and $P \geq r$ give

$$\frac{1}{n} + \sqrt{\frac{\chi^2(P_{\pi_{\mathrm{T}}^{\sigma}} \| P_{\pi_{\mathrm{old}}})}{n}} = 2r + \frac{P-r}{\sqrt{1-r}} \leq (2 + \sqrt{2})P.$$

Here $r \leq \frac{1}{2}$, so $(1 - r)^{-\frac{1}{2}} \leq \sqrt{2}$, while $2r \leq 2P$. Theorem 2 lower bounds the risk by $(1 - e^{-d_\eta})P(1 - r)^n$, which is at least $\frac{(1-e^{-d_\eta})P}{2}$. Combining the last two bounds proves the lower bound at the rate in Eq. (6), with implied constant $\frac{1-e^{-d_\eta}}{2(2+\sqrt{2})}$. At $P = r$, the distributions over prefixes coincide and the chi-squared divergence is 0. As P approaches 1, Eq. (17) is of order n. For the converse comparison, $P \geq r$ and $\sqrt{1-r} \leq 1$ give $2r + \frac{P-r}{\sqrt{1-r}} \geq P$. Together with $\frac{1}{2} \leq (1 - r)^n \leq 1$, these inequalities show that the lower bound has the order in Eq. (6), with constants depending only on η. The exact minimax calculation below supplies the matching upper bound for this family. □

Exact minimax risk for a single rare prefix. For the construction above with $m = 1$, fix $r, P \in (0, 1)$ and $0 < \eta < \frac{1}{2}$. The learner knows this family of two targets, including the shared next-token distribution after B. Only the sign σ is unknown. Under the same observation model, its minimax risk is exactly

$$\begin{aligned} R_n^* &:= \inf_{\mathcal{A}} \sup_{\sigma \in \{-1,+1\}} \mathbb{E}_\sigma\left[\mathcal{E}_{\pi_{\mathrm{T}}^{\sigma}}(\mathcal{A}(D_n))\right] \\ &= (1-r)^n \left[-\log\left(1 - \left(1 - \sqrt{1-4\eta^2}\right)P\right)\right]. \end{aligned} \tag{18}$$

To obtain the lower bound, retain the logarithmic expression in Eq. (16). The rare prefix is unvisited with probability $(1 - r)^n$, and $e^{-d_\eta} = \sqrt{1 - 4\eta^2}$. The lower bound averaged over the two signs is therefore the right-hand side of Eq. (18).

To achieve this bound, copy the target policy if the rare prefix was observed. Otherwise, use $\mathrm{Ber}\left(\frac{1}{2}\right)$ at that prefix, copy the shared next-token distribution after B, and choose probabilities for the first token

$$p_A^* = \frac{Pe^{-d_\eta}}{1 - P + Pe^{-d_\eta}}, \qquad p_B^* = \frac{1-P}{1 - P + Pe^{-d_\eta}}.$$

For either sign, the KL after the unobserved prefix equals d_η. The minimization above gives total KL $-\log(1 - P + Pe^{-d_\eta})$ when it is unvisited and 0 when it is visited. This learner attains the lower bound for each sign, proving the minimax equality. For fixed η, $-\log(1 - c_\eta P) = \Theta(P)$. Together with the comparison above, this gives the matching upper and lower bounds in Eq. (6). The equality characterizes this construction. The comparison with the general upper bound in Theorem 1 concerns the dependence on sample size and distribution shift.

A.3 Proof of Theorem 3

Proof. Fix $k \geq 1$. Let $C_i := c(X, Y_i)$, where conditional on X, the responses $Y_1, \ldots, Y_k$ are independent draws from the same policy. For $u, v \in [0, 1]$, write $\mathrm{kl}(u\|v)$ for the KL divergence between Bernoulli distributions with success probabilities u and v, using the usual boundary conventions. Write $\mathcal{L}_\pi(X, C_1, \ldots, C_k)$ for the joint distribution of $(X, C_1, \ldots, C_k)$ induced by π, and similarly for π_{T}. Because the prompt distribution is the same under both policies and the correctness indicators are conditionally independent, the KL chain rule (Cover and Thomas, 2006, Theorem 2.5.3) gives

$$\mathrm{KL}(\mathcal{L}_\pi(X, C_1, \ldots, C_k)\|\mathcal{L}_{\pi_{\mathrm{T}}}(X, C_1, \ldots, C_k)) = k\mathbb{E}_X\, \mathrm{kl}(p_\pi(X)\|p_{\pi_{\mathrm{T}}}(X))\,. \tag{19}$$

Map $(X, C_1, \ldots, C_k)$ to $Z := \mathbf{1}\{\max_{1\leq i\leq k} C_i = 1\}$. Its success probability is $\mathsf{Pass}_k(\pi)$ under π and $\mathsf{Pass}_k(\pi_{\mathrm{T}})$ under π_{T}. Grouping outcomes cannot increase KL, by the log-sum inequality (Cover and Thomas, 2006, Theorem 2.7.1). Grouping by Z in Eq. (19) gives the first inequality in Eq. (20) below. For each fixed prompt, grouping responses by $c(X, Y)$ also gives

$$\mathrm{kl}(p_\pi(X)\|p_{\pi_{\mathrm{T}}}(X)) \leq \mathrm{KL}(\pi(\cdot \mid X)\|\pi_{\mathrm{T}}(\cdot \mid X))\,.$$

Thus

$$\mathrm{kl}(\mathsf{Pass}_k(\pi)\|\mathsf{Pass}_k(\pi_{\mathrm{T}})) \leq k\mathbb{E}_X\, \mathrm{kl}(p_\pi(X)\|p_{\pi_{\mathrm{T}}}(X)) \leq k\mathcal{E}_{\pi_{\mathrm{T}}}(\pi). \tag{20}$$

Pinsker's inequality (Tsybakov, 2009, Lemma 2.5(i)) gives $\mathrm{kl}(a\|b) \geq 2(a-b)^2$ for Bernoulli parameters $a, b \in [0, 1]$, with natural logarithms. Applying it to Eq. (20) yields

$$|\mathsf{Pass}_k(\pi) - \mathsf{Pass}_k(\pi_{\mathrm{T}})| \leq \sqrt{\frac{k\mathbb{E}_X\, \mathrm{kl}(p_\pi(X)\|p_{\pi_{\mathrm{T}}}(X))}{2}} \leq \sqrt{\frac{k\mathcal{E}_{\pi_{\mathrm{T}}}(\pi)}{2}}.$$

The absolute difference between two probabilities is also at most 1. Using $\mathcal{E}_{\pi_{\mathrm{T}}}(\pi) \leq \varepsilon$ and rearranging gives the lower bound on $\mathsf{Pass}_k(\pi)$ in Theorem 3. The same argument gives the upper bound $\mathsf{Pass}_k(\pi) \leq \mathsf{Pass}_k(\pi_{\mathrm{T}}) + \min\left\{1, \sqrt{\frac{k\varepsilon}{2}}\right\}$, so the guarantee controls deviations in both directions. □

A.4 Proof of Theorem 4

Proof. We first handle $k \geq 2$, preserving Pass@1 exactly. Let the two prompts x_h, x_e each have probability $\frac{1}{2}$, let the response be $Y \in \{0, 1\}$, and set $c(X, Y) = Y$. Write $u := \frac{1}{5k}$. Choose the target success probabilities as $p_{\widetilde{\pi}_{\mathrm{T}}}(x_h) = u$ and $p_{\widetilde{\pi}_{\mathrm{T}}}(x_e) = \frac{1}{2}$. For $0 < \delta \leq \frac{u}{2}$, choose $p_{\widetilde{\pi}_\delta}(x_h) = u - \delta$ and $p_{\widetilde{\pi}_\delta}(x_e) = \frac{1}{2} + \delta$. These probabilities specify Bernoulli response policies with common support. Their average success probabilities are equal, since the changes on the two prompts cancel. Denote the response KL by

$$E(\delta) := \mathcal{E}_{\widetilde{\pi}_{\mathrm{T}}}(\widetilde{\pi}_\delta) = \frac{1}{2}\left(\mathrm{kl}(u - \delta\|u) + \mathrm{kl}\left(\frac{1}{2} + \delta\middle\|\frac{1}{2}\right)\right).$$

For $0 < p < 1$, applying $\log z \leq z - 1$ to the two terms of the Bernoulli KL gives $\mathrm{kl}(r\|p) \leq \frac{r^2}{p} + \frac{(1-r)^2}{1-p} - 1 = \frac{(r-p)^2}{p(1-p)}$. This is the standard bound of KL by chi-squared divergence (Tsybakov, 2009, Lemma 2.7). Because $u \leq \frac{1}{10}$ and $k \geq 2$, it follows that

$$E(\delta) \leq \frac{\delta^2}{2}\left(\frac{1}{u(1-u)} + 4\right) \leq \frac{\delta^2}{2}\left(\frac{50k}{9} + 2k\right) \leq 4k\delta^2. \tag{21}$$

Let $f(p) := 1 - (1-p)^k$. Its derivative $f'(p) = k(1-p)^{k-1}$ is decreasing in p. Hence $f(u) - f(u-\delta) = \int_{u-\delta}^{u} f'(p)\, dp \geq \delta f'(u)$, whereas $f(\frac{1}{2} + \delta) - f(\frac{1}{2}) \leq \delta f'(\frac{1}{2})$. Subtracting the gain on the easier prompt from the loss on the harder prompt and applying Bernoulli's inequality gives

$$\begin{aligned}
\mathsf{Pass}_k(\widetilde{\pi}_{\mathrm{T}}) - \mathsf{Pass}_k(\widetilde{\pi}_\delta) &= \frac{1}{2}\left(f(u) - f(u-\delta) - f\left(\frac{1}{2} + \delta\right) + f\left(\frac{1}{2}\right)\right) \\
&\geq \frac{k\delta}{2}\left((1-u)^{k-1} - \left(\frac{1}{2}\right)^{k-1}\right) \\
&\geq \frac{k\delta}{2}\left(\frac{4}{5} - \frac{1}{2}\right) = \frac{3k\delta}{20} \geq \frac{3}{40}\sqrt{kE(\delta)}.
\end{aligned} \tag{22}$$

Here $(1-u)^{k-1} \geq 1-(k-1)u \geq \frac{4}{5}$, and the final step uses Eq. (21). Equal changes in success probability cancel in Pass@1, but the loss on the harder prompt outweighs the gain on the easier prompt in Pass@k.

To obtain each sufficiently small KL error, observe that $E(0) = 0$ and E is continuous and strictly increasing on $[0, \frac{u}{2}]$. Indeed, $\partial_r \operatorname{kl}(r\|p) = \log \frac{r(1-p)}{p(1-r)}$ has the sign of $r-p$. Both probabilities move away from their target values as δ increases, so both KL terms increase. Set $\varepsilon_0 := E(\frac{u}{2}) > 0$. For each $0 < \varepsilon \leq \varepsilon_0$, continuity and strict monotonicity give a unique δ with $E(\delta) = \varepsilon$. Eq. (22) proves the bound throughout this interval. To check that this interval covers the range in the theorem, take $\delta = \frac{u}{2}$. The function $r \mapsto \operatorname{kl}(r\|u)$ and its first derivative vanish at $r = u$. Its second derivative is $\frac{1}{r(1-r)} \geq \frac{1}{u}$ on $[\frac{u}{2}, u]$. Taylor's integral remainder therefore gives $\operatorname{kl}(u-\delta\|u) = \int_{u-\delta}^{u} \frac{t-(u-\delta)}{t(1-t)}\, dt \geq \frac{\delta^2}{2u}$. At $\delta = \frac{u}{2}$, this lower bound equals $\frac{u}{8}$. This term has weight $\frac{1}{2}$ in E. Combining it with the upper bound in Eq. (21) gives

$$\frac{1}{80k} \leq \varepsilon_0 \leq \frac{1}{25k}, \qquad \mathsf{Pass}_k(\widetilde{\pi}_{\mathrm{T}}) - \mathsf{Pass}_k(\widetilde{\pi}_{\frac{u}{2}}) \geq \frac{3}{200}.$$

Thus a fixed positive Pass@k loss is possible with KL of order $\frac{1}{k}$.

For $k = 1$, use a single prompt with target success $u = \frac{1}{5}$ and student success $u - \delta$, where $0 < \delta \leq \frac{u}{2}$. Then $\mathcal{E}_{\widetilde{\pi}_{\mathrm{T}}}(\widetilde{\pi}_\delta) \leq \frac{25}{4}\delta^2$, whereas the Pass@1 loss is exactly δ. It is therefore at least $\frac{2}{5}\sqrt{\mathcal{E}_{\widetilde{\pi}_{\mathrm{T}}}(\widetilde{\pi}_\delta)}$, which implies the asserted constant $\frac{3}{40}$. At $\delta = \frac{u}{2}$, the curvature lower bound $\frac{1}{r(1-r)} \geq \frac{1}{u}$ on $[\frac{u}{2}, u]$ gives KL at least $\frac{u}{8} = \frac{1}{40}$. The same continuity and monotonicity argument gives every KL error in the stated interval $(0, \frac{1}{80}]$. □

A.5 Proof of Theorem 5

Proof. The KL bound is immediate when $\varepsilon_0 = \infty$, so assume finite initial KL. We derive the update, then construct the opposing changes in Pass@1 and Pass@k.

1. The update and KL decrease. First consider Eq. (8) with any $\alpha \in (0, 1)$. For a policy ν, write its expected KL to the initial student as $\mathcal{E}_{\pi_0}(\nu) := \mathbb{E}_X \mathrm{KL}(\nu(\cdot \mid X)\|\pi_0(\cdot \mid X))$, as in Eq. (1).

For almost every prompt x, write $P(y) = \pi_0(y \mid x)$ and $Q(y) = \pi_{\mathrm{T}}(y \mid x)$. All sums are over responses to which both policies assign positive probability. Let $z(y) = P(y)^{1-\alpha}Q(y)^{\alpha}$ and $Z_x = \sum_y z(y)$. Positivity on the common support gives $Z_x > 0$. Hölder's inequality with exponents $\frac{1}{1-\alpha}$ and $\frac{1}{\alpha}$ gives $Z_x \leq (\sum_y P(y))^{1-\alpha}(\sum_y Q(y))^{\alpha} = 1$. For the normalized distribution $R(y) = \frac{z(y)}{Z_x}$, the identity $\log R(y) = (1-\alpha)\log P(y) + \alpha \log Q(y) - \log Z_x$ gives, by substitution into KL,

$$\alpha \mathrm{KL}(\nu\|Q) + (1-\alpha)\mathrm{KL}(\nu\|P) = \mathrm{KL}(\nu\|R) - \log Z_x.$$

Since KL is nonnegative and equals 0 only for equal distributions, R uniquely minimizes the objective at this prompt. A policy has infinite loss if it assigns positive probability to a response with probability 0 under P or Q. We can optimize each prompt separately, so these distributions define π_+. Their expected objective is finite: the minimum $-\log Z_x$ is at most $\alpha\mathrm{KL}(P\|Q)$, whose expectation is $\alpha\varepsilon_0$. Taking expectations in the displayed identity also shows that any other minimizing policy agrees with π_+ for almost every prompt. Finally, evaluating the objective at π_0 gives $\alpha\mathcal{E}_{\pi_{\mathrm{T}}}(\pi_+) + (1-\alpha)\mathcal{E}_{\pi_0}(\pi_+) \leq \alpha\varepsilon_0$. Dividing by $\alpha > 0$ yields

$$\mathcal{E}_{\pi_{\mathrm{T}}}(\pi_+) + \frac{1-\alpha}{\alpha}\mathcal{E}_{\pi_0}(\pi_+) \leq \varepsilon_0. \tag{23}$$

Dropping the nonnegative second term proves $\mathcal{E}_{\pi_{\mathrm{T}}}(\pi_+) \leq \varepsilon_0$.

2. A construction with increasing Pass@1 and decreasing Pass@k. Set $\alpha = \frac{1}{2}$ as in Theorem 5. The prompt distribution and teacher may depend on k and ε_0.

Set $u = \frac{1}{10k}$, so $0 < u \leq \frac{1}{20}$. Let the harder prompt h have probability $w \in (0, \frac{1}{4}]$ and the easier prompt e probability $1-w$. We will choose w so that the initial KL equals ε_0. Set $s = \frac{3wu}{4(1-w)}$.

Then $0 < s \le \frac{u}{4} \le \frac{1}{80}$. Responses 1 and 2 are correct and response 3 is incorrect. Define initial distributions P_h, P_e and desired updated distributions R_h, R_e by

$$P_h = \left(\frac{u}{8}, \frac{7u}{8}, 1-u\right), \quad R_h = \left(\frac{3u}{8}, \frac{u}{8}, 1-\frac{u}{2}\right),$$
$$P_e = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right), \qquad R_e = \left(\frac{\frac{1}{2}+s}{2}, \frac{\frac{1}{2}+s}{2}, \frac{1}{2}-s\right).$$

For each prompt $x \in \{h, e\}$, define the teacher by $Q_x(i) = \frac{R_x(i)^2}{P_x(i)N_x}$, where $N_x = \sum_{i=1}^{3} \frac{R_x(i)^2}{P_x(i)}$. Set $\pi_0(\cdot \mid x) = P_x$ and $\pi_{\mathrm{T}}(\cdot \mid x) = Q_x$. All probabilities are positive. Since $\sqrt{P_x(i)Q_x(i)} = \frac{R_x(i)}{\sqrt{N_x}}$, normalizing this expression gives R_x. Thus the update in Step 1 produces exactly the chosen distributions.

The teacher is more accurate on both prompts. Write $p_{Q_x} = Q_x(1) + Q_x(2)$ for its success probability at prompt x. The two correct responses contribute $\frac{9u}{8} + \frac{u}{56} = \frac{8u}{7}$ to N_h, while the incorrect response contributes $\frac{(1-\frac{u}{2})^2}{1-u} = 1 + \frac{u^2}{4(1-u)}$. Thus $N_h = 1 + \frac{8u}{7} + \frac{u^2}{4(1-u)}$ and $p_{Q_h} = \frac{8u}{7N_h}$. The teacher assigns response 2 probability $\frac{u}{56N_h}$, compared with $\frac{7u}{8}$ under the initial student and $\frac{u}{8}$ after the update. Thus the correct response most likely under the initial student receives less probability after distillation. The inequality $p_{Q_h} > u$ is equivalent to $\frac{8}{7}(1-u)^2 > (1-\frac{u}{2})^2$. For $u \le \frac{1}{20}$, $\frac{1-\frac{u}{2}}{1-u} \le \frac{39}{38}$ and $(\frac{39}{38})^2 < \frac{8}{7}$, proving the claim. On the easier prompt, $p_{Q_e} = \frac{(\frac{1}{2}+s)^2}{(\frac{1}{2}+s)^2+(\frac{1}{2}-s)^2} > \frac{1}{2} + s > \frac{1}{2}$. Here $p_{Q_e} - (\frac{1}{2}+s) = \frac{s(1-4s^2)}{1+4s^2} > 0$ because $0 < s < \frac{1}{2}$. Since $p \mapsto 1-(1-p)^j$ is strictly increasing for every integer $j \ge 1$, higher teacher accuracy on both prompts gives $\mathsf{Pass}_j(\pi_{\mathrm{T}}) > \mathsf{Pass}_j(\pi_0)$ for every finite integer $j \ge 1$.

After the update, success probability is $\frac{u}{2}$ on the harder prompt and $\frac{1}{2} + s$ on the easier prompt. Averaging over prompts, the Pass@1 gain is $(1-w)s - \frac{wu}{2} = \frac{wu}{4}$. At the fixed k, the harder prompt contributes a Pass@k loss of at least

$$w\left[\left(1-\frac{u}{2}\right)^k - (1-u)^k\right] \ge \frac{wku}{2}(1-u)^{k-1} \ge \frac{9wku}{20}.$$

The first inequality follows from the derivative of $p \mapsto 1-(1-p)^k$ on $[\frac{u}{2}, u]$. The second uses Bernoulli's inequality and $(k-1)u < \frac{1}{10}$. On the easier prompt the derivative is at most $k(\frac{1}{2})^{k-1}$ throughout $[\frac{1}{2}, \frac{1}{2}+s]$, so its contribution to the gain is at most $(1-w)ks(\frac{1}{2})^{k-1} \le \frac{3wku}{8}$, because $k \ge 2$. Subtracting the gain from the loss gives

$$\mathsf{Pass}_k(\pi_0) - \mathsf{Pass}_k(\pi_+) \ge \frac{3wku}{40}.$$

To express these changes in terms of ε_0, write $D_h = \mathrm{KL}(P_h \| Q_h)$ and $D_e = \mathrm{KL}(P_e \| Q_e)$. Since $\log \frac{P_x(i)}{Q_x(i)} = 2\log \frac{P_x(i)}{R_x(i)} + \log N_x$, averaging over P_x gives $\mathrm{KL}(P_x \| Q_x) = 2\mathrm{KL}(P_x \| R_x) + \log N_x$. Applying $\log t \le t - 1$ gives the standard inequality $\mathrm{KL}(P\|R) \le \sum_i \frac{P(i)^2}{R(i)} - 1 = \sum_i \frac{(P(i)-R(i))^2}{R(i)}$ (Tsybakov, 2009, Lemma 2.7). On the harder prompt the differences $P_h - R_h$ are $(-\frac{u}{4}, \frac{3u}{4}, -\frac{u}{2})$, giving

$$\mathrm{KL}(P_h \| R_h) \le \frac{u}{6} + \frac{9u}{2} + \frac{u^2}{4(1-\frac{u}{2})} \le \left(\frac{14}{3} + \frac{1}{78}\right)u.$$

Also $\log N_h \le N_h - 1 \le (\frac{8}{7} + \frac{1}{76})u$. Hence $D_h \le (\frac{28}{3} + \frac{1}{39} + \frac{8}{7} + \frac{1}{76})u < 11u$. For a lower bound, $D_h \ge \log N_h \ge \log(1 + \frac{8u}{7}) \ge \frac{8u}{7+8u} \ge u$, using $\log(1+t) \ge \frac{t}{1+t}$ and $u \le \frac{1}{20}$. On the easier prompt $N_e = 1 + 4s^2$. The terms for the two correct responses in $\mathrm{KL}(P_e \| R_e)$ have total weight $\frac{1}{2}$ and ratio $\frac{1}{1+2s}$. The term for the incorrect response has weight $\frac{1}{2}$ and ratio $\frac{1}{1-2s}$. Therefore $\mathrm{KL}(P_e \| R_e) = -\frac{1}{2}\log(1-4s^2)$, and the preceding identity yields

$$D_e = \log \frac{1+4s^2}{1-4s^2} \le 16s^2.$$

Indeed, for $a = 4s^2 \le \frac{1}{4}$, $\log \frac{1+a}{1-a} \le \frac{2a}{1-a} \le 4a$. The expected KL before the update, denoted by $E(w)$, thus satisfies

$$wu \le E(w) = wD_h + (1-w)D_e \le 11wu + \frac{9w^2u^2}{1-w} \le 12wu.$$

The last inequality follows from $\frac{9wu}{1-w} \le \frac{3}{20} < 1$.

Here D_h is fixed once k is fixed, while D_e varies continuously with $s = \frac{3wu}{4(1-w)}$ and tends to 0 as $w \downarrow 0$. Thus $E(w)$ extends continuously to $[0, \frac{1}{4}]$, with $E(0) = 0$ and $E(\frac{1}{4}) \ge \frac{u}{4} = \frac{1}{40k}$. By the intermediate value theorem, there is a $w > 0$ with $E(w) = \varepsilon_0$ for every ε_0 in the stated range. Since $wu \ge \frac{\varepsilon_0}{12}$, the Pass@1 gain is at least $\frac{\varepsilon_0}{48}$, and the Pass@k loss is at least $\frac{k\varepsilon_0}{160}$. Finally, $\pi_+ \ne \pi_0$ on both prompts, each of which has positive probability, so Eq. (23) implies $\mathcal{E}_{\pi_T}(\pi_+) \le \varepsilon_0 - \mathcal{E}_{\pi_0}(\pi_+) < \varepsilon_0$. □

B Synthetic Experimental Details

Sample size and prefix divergence. We evaluate distillation error and Pass@k exactly on finite response spaces. To examine the dependence on sample size and divergence in Theorems 1-2, we use the single prefix family and the learner attaining its minimax risk in Appendix A.2. We set $\eta = 0.3$, $r = \frac{1}{2n}$, and $n \in \{32, 64, \dots, 32768\}$, and choose the teacher's prefix probability using Eq. (17) to fix the chi-squared divergence at 0, 0.25, 1, or 4. We verify the divergence from the joint prefix distributions and average the learner's error over the training data.

KL error and sampling budget. For Theorems 3-4, we use two equally likely prompts with teacher success probabilities $(u, \frac{1}{2})$ and student probabilities $(u-d, \frac{1}{2}+d)$, where $u = \frac{1}{5k}$ and $k \in \{2, 4, \dots, 256\}$. We solve for d at nine logarithmically spaced KL values in $[10^{-8}, 10^{-5}]$, then at fixed Pass@k gaps of 0.001 and 0.003.

Exact proximal updates. For Theorem 5, we use the construction with two prompts and three responses per prompt in Appendix A.5 and the exact proximal update $\pi_+ \propto \sqrt{\pi_0 \pi_T}$, with $\alpha = \frac{1}{2}$. At each budget, the weight of the harder prompt sets $\varepsilon_0 = \frac{t}{40k}$ for 11 logarithmically spaced $t \in [10^{-5}, 1]$. Figure 1 reports how expected distillation error varies with sample size at fixed divergence. Figure 2 reports the Pass@k gap as ε and k vary, together with the changes in Pass@1 and Pass@k under the proximal update.

Figure 2(c) displays the $k = 16$ slice of this grid. Its vertical scale is linear on $[-10^{-8}, 10^{-8}]$ and logarithmic in magnitude outside that interval. Changing the weight of the harder prompt changes the expected KL while retaining the same conditional response distributions at each prompt. These calculations describe the explicit constructions in the proofs. Their prevalence under neural network training is a separate empirical quantity.

C Large Language Model Experimental Details

C.1 Detailed Experimental Setup

Models and teacher checkpoints. We use Qwen3-0.6B-Base, Qwen3-1.7B-Base, and Qwen3-4B-Base (Yang et al., 2025). For each size, a GRPO run (Shao et al., 2024) on DAPO-Math-17k (Yu et al., 2025) produces the teacher checkpoint at the end of epoch 2. The teacher remains frozen throughout OPD. Each OPD student is initialized from the original Base checkpoint of the same size. The teacher and student therefore have the same parameter count and different training histories. Both stages use the `verl` framework (Sheng et al., 2024).

Training. Our hyperparameter presentation follows the on-policy training setup of Jin et al. (2026). Table 2 reports the settings of the present experiments. The OPD optimizer is AdamW (Loshchilov and Hutter, 2019), with $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.01, gradient clipping at 1.0, and zero warmup. The clipping thresholds for the policy ratio are 0.2 on either

Table 2: Training hyperparameters for GRPO and OPD.

Hyperparameter	GRPO	OPD
Training dataset	DAPO-Math-17k	DAPO-Math-17k
Initialization	Qwen3-Base	Qwen3-Base
Learning rate	3×10^{-6}	3×10^{-6}
LR scheduler	Cosine	Cosine
Optimizer	AdamW	AdamW
Training batch size	128	128
Minibatch size	32	32
Samples per prompt	8	1
Sampling temperature	1.0	1.0
Top-p	1.0	1.0
Maximum response length	7168	7168
Teacher scoring temperature	–	1.0
Student candidates per token	–	16
Training epochs	2	2
Teacher checkpoint	–	GRPO teacher

side of one (dual-clip parameter 3). Master parameters and optimizer states are FP32, forward computation uses BF16. Prompts are limited to 1024 tokens, with overlong prompts filtered, and the response budget is 7168 tokens. Decoding uses temperature 1, top-$p = 1$, and repetition penalty 1.

Benchmarks. We evaluate mathematical reasoning on AIME24, AIME25, and AMC23. AIME24 and AIME25 each contain 30 problems from the 2024 and 2025 American Invitational Mathematics Examination, respectively (MAA, 2025). AMC23 contains 40 problems drawn from the 2023 American Mathematics Competitions (MAA, 2023). These benchmarks cover competition mathematics, including algebra, geometry, number theory, and combinatorics. We evaluate all three benchmarks using free responses.

Sampling protocol. At each checkpoint, we independently sample $n = 64$ responses per problem using vLLM (Kwon et al., 2023), with temperature 1.0, top-$p = 1.0$, and no top-k filtering. The request seeds are 42–105, and the maximum response length is 15872 tokens for every checkpoint. The inference prompt is specified at the end of this subsection. Responses that reach the length limit are retained and evaluated as generated.

Answer correctness. We verify final answers using rules, as is common in mathematical reasoning evaluation (DeepSeek-AI, 2025). For a benchmark $\mathcal{B} = \{(x_i, a_i^\star)\}_{i=1}^N$, let y_{ij} denote the j-th response to problem x_i, and let $a_i^\star$ be its reference answer. We extract the boxed final answer $\hat{a}_{ij} = \mathrm{Extract}(y_{ij})$, assigning $\hat{a}_{ij} = \text{False}$ when extraction fails. The binary correctness label is

$$R_{ij} = \begin{cases} \mathbf{1}[\nu(\hat{a}_{ij}) = \nu(a_i^\star) \ \vee \ S(\hat{a}_{ij}, a_i^\star) = 1], & \hat{a}_{ij} \neq \text{False}, \\ 0, & \hat{a}_{ij} = \text{False}, \end{cases} \tag{24}$$

where ν denotes the grader's answer normalization function, and $S(a, b) \in \{0, 1\}$ indicates whether its symbolic equivalence checker accepts the two answers as equivalent. Thus, a response is correct if its extracted answer passes either normalized matching or symbolic verification, responses without an extractable boxed answer receive zero.

Evaluation metrics. Let $c_i = \sum_{j=1}^n R_{ij}$ be the number of correct responses to problem i. Following Chen et al. (2021), we estimate Pass@k as

$$\text{Pass@}k = \frac{1}{N} \sum_{i=1}^{N} \left[1 - \frac{\binom{n-c_i}{k}}{\binom{n}{k}} \right], \qquad 1 \leq k \leq n, \tag{25}$$

with $\binom{a}{k} = 0$ when $a < k$. Under independent, identically distributed sampling within each problem, this estimator is unbiased for the probability of obtaining at least one correct answer in k attempts, averaged over benchmark problems.

For $n = 64$, the two metrics reduce to

$$\text{Pass@1} = \frac{1}{64N} \sum_{i=1}^{N} \sum_{j=1}^{64} R_{ij}, \tag{26}$$

$$\text{Pass@64} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[c_i > 0]. \tag{27}$$

Pass@1 therefore measures single response accuracy, estimated using all 64 samples, whereas Pass@64 measures the fraction of problems with at least one correct response. The reported test error is

$$\text{TestError} = 1 - \text{Pass@1} = \frac{1}{64N} \sum_{i=1}^{N} \sum_{j=1}^{64} (1 - R_{ij}). \tag{28}$$

All budgets $k = 1, \ldots, 64$ share the same 64 responses per problem. The request seeds index these response samples at each checkpoint. A count $c_i = 0$ records zero observed successes in this finite batch. Appendix C.5 discusses sampling uncertainty.

Training and inference prompts. Both GRPO and OPD training use the DAPO prompt (Yu et al., 2025; Li et al., 2026a). For benchmark evaluation, all Qwen3 checkpoints use the inference prompt below, with thinking mode disabled.

Training Prompt (GRPO and OPD)

Solve the following math problem step by step. The last line of your response should be of the form Answer: `$Answer` (without quotes) where `$Answer` is the answer to the problem.

{`problem`}

Remember to put your answer on its own line after "`Answer:`".

Inference Prompt

{`problem`} Please reason step by step, and put your final answer within `\boxed{}`.

C.2 Chi-Squared Divergence between Checkpoint Prefix Distributions

For an old checkpoint Ca and a new checkpoint Cb, let π_a and π_b denote their student policies. We estimate the chi-squared divergence $\chi^2(P_{\pi_b} \| P_{\pi_a})$ appearing in Eq. (4). For each benchmark, we reuse the $M = 64N$ trajectories sampled from π_a and rescore their saved token sequences under both checkpoints. The probability ratio for the prefix before token h of trajectory m is

$$w_{m,h} = \exp\left\{ \sum_{t<h} \left[\log \pi_b(y_{m,t} \mid x_m, y_{m,<t}) - \log \pi_a(y_{m,t} \mid x_m, y_{m,<t}) \right] \right\}. \tag{29}$$

The ratio at the empty prefix is one. With $H = 15872$, the Monte Carlo estimate averages $(w_{m,h} - 1)^2$ over trajectories and prefix depths:

$$D_{\text{raw}} = \frac{1}{MH} \sum_{m=1}^{M} \sum_{h=1}^{H} (w_{m,h} - 1)^2.$$

Its population expectation is the chi-squared divergence for a uniformly sampled prefix depth when the prompt distribution, horizon, and termination convention match the definitions of P_{π_a} and P_{π_b}. We truncate probability ratios at $c = 10$ to limit the influence of large ratios and report

$$D_c = \frac{1}{MH} \sum_{m=1}^{M} \sum_{h=1}^{H} \big(\min\{w_{m,h}, c\} - 1 \big)^2. \tag{30}$$

Table 3: Estimates from Eq. (30) for C120 → C139, with ratios truncated at 10 and horizon $H = 15872$. Columns give the student parameter count.

Dataset	0.6B	1.7B	4B
AIME24	4.591	6.440	1.857
AIME25	4.296	6.523	1.624
AMC23	3.949	4.677	2.378

Distribution change across checkpoints. Table 3 gives larger values for the 1.7B model and smaller values for the 4B model on all three benchmarks during C120 → C139. Larger values reflect greater mean squared deviation of the truncated prefix probability ratios from one. This variation motivates accounting for prefix distribution change when relating training fit to KL error during generation, as in Eq. (4).

Measurement conventions. The training loss uses the top-16 approximation and optimization settings in Appendix C.1. Benchmark error uses answer correctness in Eq. (28). The theoretical risk in Eq. (1) compares the student and teacher distributions over complete responses and averages their KL over a specified prompt distribution. Truncating ratios in Eq. (30) removes part of the contribution from ratios above 10, so the reported values measure the truncated expectation. Evaluating the full shift term in Eq. (4) requires accounting for these tails and specifying a common treatment of prefixes after termination. Aligning the risk measurements also requires matching the prompt distribution, sampling policies, teacher, KL direction, and normalization by tokens or responses. The reported distribution measurements concern C120 → C139. The Pass@k comparisons below concern C60 → C80 and C80 → C100.

Model size and the observed error plateaus. Figure 3 measures answer error, whereas Theorems 1–2 control KL error relative to a teacher. Their connection follows from

$$1 - \mathsf{Pass}_1(\pi) = \big[1 - \mathsf{Pass}_1(\pi_{\mathrm{T}})\big] + \big[\mathsf{Pass}_1(\pi_{\mathrm{T}}) - \mathsf{Pass}_1(\pi)\big].$$

The first term is the teacher's answer error. The second is the signed accuracy gap, whose magnitude Theorem 3 bounds by $\sqrt{\mathcal{E}_{\pi_{\mathrm{T}}}(\pi)/2}$ at $k = 1$. As KL error vanishes, the student's answer error approaches that of its particular teacher. Thus different teachers can give different answer error levels even under accurate imitation. Here each teacher shares the student's architecture, so its policy is representable at that size. Recovering it through OPD also depends on optimization and the feedback available along student responses.

Finite data and prefix coverage. Eq. (4) separates empirical fitting loss, the description length term involving $K(\pi)/n$, and prefix shift. Eq. (5) gives a complementary mechanism: supervision at a prefix visited with probability r_j is missed with probability $(1 - r_j)^n$. This can sustain imitation error at finite sample sizes. For fixed positive r_j, that probability tends to zero. The upper bound also tends to zero under vanishing fitting loss and bounded shift, with the other quantities fixed. The plateaus in Figure 3 describe the observed training interval. Differences in teacher accuracy, achieved fit, and prefix coverage provide possible explanations for their separation. Table 3 gives the largest shift estimates for 1.7B, while the highest answer errors occur at 0.6B, so the shift estimates alone do not order model performance. Appendix C.3 supplies each teacher's benchmark accuracy and analyzes the signed answer-error gaps. Distinguishing the remaining mechanisms requires comparable student–teacher KL measurements and observations of which prefixes receive supervision.

C.3 Teacher Baselines and Error-Gap Fits

Teacher evaluation and the signed gap. We evaluate the frozen epoch-2 GRPO teachers with 64 responses per problem, request seeds 42–105, temperature 1, top-$p = 1$, and a 15872-token response limit. The teacher evaluation uses four GPUs, and the student evaluation uses eight. Matching request seeds does not ensure identical floating-point execution across these configurations. The supplied teacher report gives Pass@1 to two decimal places on the percentage scale. We use these reported values without inferring unreported correct counts. Table 4 lists the teacher baselines and the final student errors.

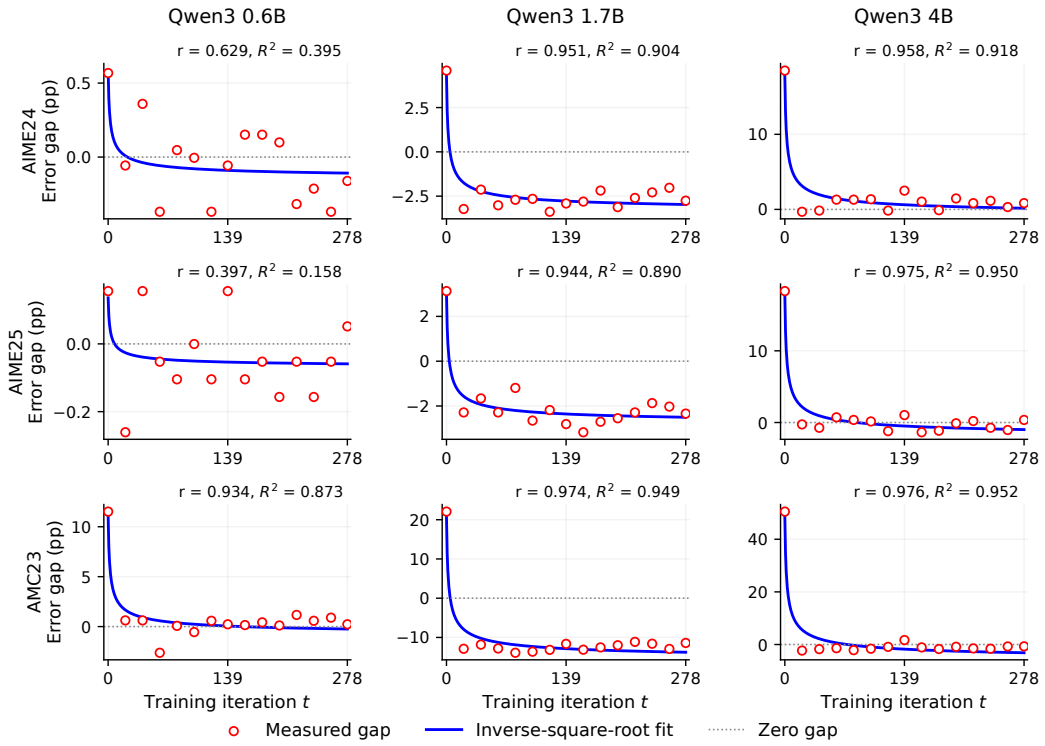


Figure 5: Signed student minus teacher test-error gaps and the descriptive fit in Eq. (32). Rows correspond to AIME24, AIME25, and AMC23. Columns correspond to model sizes. Coral circles mark all measured checkpoints. Blue curves are fitted values. The dotted zero line denotes equal student and teacher answer error. Vertical axes use percentage points and separate scales.

For each model size and benchmark, let e_{T} be the teacher's test error and e_t the student's test error at Ct. At every evaluated checkpoint we compute the signed gap

$$d_t = e_t - e_{\mathrm{T}} = \mathsf{Pass}_1(\pi_{\mathrm{T}}) - \mathsf{Pass}_1(\pi_t). \tag{31}$$

Negative values mean that the student has higher measured accuracy than its teacher. We retain these values, including the negative final gaps for all three 1.7B benchmarks. The 0.6B and 4B final gaps have magnitude below one percentage point on every benchmark. Teacher rounding introduces uncertainty of at most 0.005 percentage points into each reported gap, apart from sampling variability.

Table 4: Teacher baselines and final student errors. Pass@1 and error are percentages. d_{278} is student minus teacher error in percentage points.

Model	Dataset	Teacher Pass@1	Teacher error	C278 error	d_{278}
0.6B	AIME24	0.62	99.38	99.22	−0.16
0.6B	AIME25	0.26	99.74	99.79	+0.05
0.6B	AMC23	13.55	86.45	86.68	+0.23
1.7B	AIME24	4.90	95.10	92.34	−2.76
1.7B	AIME25	3.39	96.61	94.27	−2.34
1.7B	AMC23	27.58	72.42	61.02	−11.40
4B	AIME24	22.03	77.97	78.80	+0.83
4B	AIME25	20.36	79.64	80.00	+0.36
4B	AMC23	66.02	33.98	33.28	−0.70

Checkpoints and fitting procedure. The original metric exports contain 15 evaluation checkpoints per model–benchmark pair: C0, every 20 iterations within each epoch, and the epoch endpoints C139 and C278. Training loss is logged at every iteration, but answer error is available only

at these checkpoints. We use all 135 measured errors without interpolating unevaluated iterations or smoothing the curves. With batch size 128 and one response per prompt, write $n_t = 128t$ for the nominal cumulative number of training trajectories. To include the pre-training checkpoint C0, we use the fixed shift $t + 1$ and fit each curve by equally weighted least squares:

$$\widehat{d}_t = c + \frac{a}{\sqrt{t+1}} = c + \frac{\sqrt{128}\,a}{\sqrt{n_t + 128}}, \qquad a \geq 0, \quad c \in \mathbb{R}. \tag{32}$$

The shift fixes the origin used for fitting. The free offset accommodates a residual signed performance gap and does not impose convergence to the teacher's accuracy. We report the Pearson correlation coefficient r between the measured gaps d_t and $(t+1)^{-1/2}$, and the coefficient of determination $R^2 = 1 - \sum_t (d_t - \widehat{d}_t)^2 / \sum_t (d_t - \bar{d})^2$, where $\bar{d}$ is the mean measured gap.

Table 5: Descriptive fits of the signed error gap. Both columns use all 15 checkpoints in Eq. (32).

Model	Dataset	r	R^2
0.6B	AIME24	0.629	0.395
0.6B	AIME25	0.397	0.158
0.6B	AMC23	0.934	0.873
1.7B	AIME24	0.951	0.904
1.7B	AIME25	0.944	0.890
1.7B	AMC23	0.974	0.949
4B	AIME24	0.958	0.918
4B	AIME25	0.975	0.950
4B	AMC23	0.976	0.952

Relation to the generalization bounds. Theorems 1–2 bound teacher-relative KL risk under specified sampling and coverage conditions. At $k = 1$, Theorem 3 bounds the absolute answer-error gap by $|d_t| \leq \sqrt{\mathcal{E}_{\pi_T}(\pi_t)/2}$ for the corresponding population quantities. The signed gap in Eq. (31) summarizes the difference in answer error. In particular, an $n^{-1/2}$ KL bound would give an $n^{-1/4}$ upper bound through this inequality, rather than requiring a matching $n^{-1/2}$ answer-error law. The observed approach to teacher-level accuracy for 0.6B and 4B is qualitatively compatible with accurate imitation. The fits describe this training trajectory. Identifying a KL convergence exponent requires measurements of KL risk. The cumulative counts also combine changing sampling policies, whereas the finite-sample bounds fix the sampling policy within a round. Correlations across these sequential checkpoints are descriptive and are not significance tests across independent training runs.

C.4 Additional Examples from Separate OPD Runs

We also examine earlier runs that trained Qwen3-0.6B-Base and Qwen3-1.7B-Base students for one OPD epoch on DAPO-Math-17k with a fixed Qwen3-8B teacher. This teacher differs from the GRPO teachers of matching size used in Section 5.2. The saved benchmark summaries report 64 responses per problem and a common response cap of 15872 tokens. The earlier report records that saved responses exceeding this cap were truncated and graded again.

We compare consecutive evaluated checkpoints at $t \in \{0, 20, 40, 60, 80, 100, 120, 139\}$ on each of the three benchmarks. Among the 42 comparisons across two student sizes, seven intervals, and three benchmarks, four have higher Pass@1 and lower Pass@64 at the later checkpoint. Table 6 reports all four. These provide further qualitative examples of opposing changes during training under a different teacher configuration.

Table 6: Additional adjacent checkpoint comparisons with higher Pass@1 and lower Pass@64 in runs with a Qwen3-8B teacher. Arrows show earlier and later checkpoint values.

Student	Benchmark	Checkpoints	Pass@1	Pass@64
0.6B	AIME24	C60 → C80	1.15% → 1.46%	23.33% → 13.33%
0.6B	AMC23	C40 → C60	14.84% → 19.18%	82.50% → 77.50%
0.6B	AMC23	C100 → C120	22.19% → 23.05%	87.50% → 75.00%
1.7B	AMC23	C120 → C139	36.95% → 38.55%	92.50% → 90.00%

C.5 Sampling Budgets and Individual Problems in the Selected Crossings

Sampling budget and the checkpoint gap. We evaluate every integer budget $k = 1, \ldots, 64$, extending the powers-of-two budgets displayed in Figure 4. For AMC23, the later checkpoint is better at $k = 1$ and worse at every integer $k = 2, \ldots, 64$. Its deficit peaks at 2.83% for $k = 48$ and contracts to 2.50% at $k = 64$, so the gap need not increase monotonically with the sampling budget. For AIME25, the later checkpoint is better at $k = 1, 2$ and worse at $k = 3, \ldots, 64$, reaching a deficit of 3.33% at $k = 64$. The Pass@64 declines are approximately 3.4 and 16 times the magnitudes of the Pass@1 gains on AMC23 and AIME25, respectively.

Comparisons and group definitions. We compare C60 with C80 on AMC23 and C80 with C100 on AIME25 for the 4B student, using 64 responses per problem at each checkpoint. Old and new refer to the earlier and later checkpoints, respectively. We partition each benchmark into four groups:

- **Easy problems (A).** Each problem has an estimated Pass@32 of at least 99% at the earlier checkpoint and more correct responses at the later checkpoint.
- **Hard problems (B).** Each problem has 1–7 correct responses out of 64 at the earlier checkpoint and fewer correct responses at the later checkpoint.
- **Newly solved problems (C).** Each problem has no correct responses at the earlier checkpoint and at least one correct response at the later checkpoint.
- **Other problems (D).** These problems do not meet any of the three definitions above.

The groups are mutually exclusive and include every problem in each benchmark.

Why correct responses and coverage can move in opposite directions. Let c be the number of correct responses among the 64 samples for one problem. Its contribution before averaging over problems is the estimate

$$q_k(c) = 1 - \frac{\binom{64-c}{k}}{\binom{64}{k}}, \qquad q_1(c) = \frac{c}{64}, \qquad q_{64}(c) = \mathbf{1}[c > 0]. \tag{33}$$

For N benchmark problems with counts $c_1, \ldots, c_N$, estimated Pass@k is $N^{-1} \sum_{i=1}^{N} q_k(c_i)$. At $k = 1$, every additional correct response therefore increases the benchmark score by $\frac{1}{64N}$, regardless of which problem it belongs to. At $k = 64$, each problem contributes either $\frac{1}{N}$ or zero, according to whether its sample contains at least one correct response. For example, increasing a count from 10 to 11 raises Pass@1 but leaves Pass@64 unchanged. Reducing a count from one to zero lowers Pass@1 by $\frac{1}{64N}$ and Pass@64 by $\frac{1}{N}$. Thus, the total number of correct responses can increase while the number of problems solved at least once in the sample decreases.

The same estimator explains intermediate budgets. By the binomial identity $\binom{m}{k} - \binom{m-1}{k} = \binom{m-1}{k-1}$,

$$q_k(c+1) - q_k(c) = \frac{\binom{63-c}{k-1}}{\binom{64}{k}}, \qquad 0 \le c \le 63, \quad 1 \le k \le 64. \tag{34}$$

This increment equals $\frac{1}{64}$ for $k = 1$ and is nonincreasing in c for every $k > 1$. ***Consequently, at a fixed larger budget, adding a correct response to a problem with many existing successes contributes no more than adding one to a problem with few successes.*** In particular, $q_k(1) - q_k(0) = \frac{k}{64}$, so losing the only correct response on one problem reduces the benchmark score by $\frac{k}{64N}$. This explains how a few lost successes on hard problems can outweigh more numerous gains on easy problems, the allocation used in the constructions of Theorems 4 and 5. These identities concern the observed finite-sample estimates. Zero observed correct responses are compatible with a positive underlying success probability.

Contributions to benchmark changes. For problem i with correct counts c_i^{old} and c_i^{new}, define its contribution to the benchmark change as

$$\Delta_i(k) = \frac{1}{N} \left[\frac{\binom{64-c_i^{\text{old}}}{k} - \binom{64-c_i^{\text{new}}}{k}}{\binom{64}{k}} \right], \qquad k = 1, \ldots, 64. \tag{35}$$

The sum $\sum_i \Delta_i(k)$ is the new minus old benchmark Pass@k on the 0–1 scale. Positive and negative contributions are summed separately with the same full benchmark denominator, $N = 40$ or 30. All values are computed using exact binomial ratios before rounding. In Tables 7–12, red denotes a positive change and blue a negative change. Signs also identify the direction. Tables display $100\Delta_i(k)$ or its group sum with a percent sign. For example, $+0.74\%$ represents a probability increase of 0.0074. Values below 0.005% in magnitude round to zero.

Correct responses and problems with a sample hit. On AMC23, 21 problems gain 75 correct responses, 14 lose 56, and five are unchanged. Total correct responses increase from 1726 to 1745, while problems with a sample hit decrease from 39 to 38. On AIME25, six problems gain 17 correct responses, eight lose 13, and 16 are unchanged. The corresponding totals change from 384 to 388 correct responses and from 17 to 16 problems with a sample hit. These opposite changes arise because Pass@1 counts every correct response equally, while Pass@64 assigns equal weight to each problem with at least one hit, as expressed in Eqs. (26)–(27).

Table 7: Pass@k changes and their positive and negative contributions for the 4B comparisons in Table 1. Changes are new minus old on the percentage scale, using the full benchmark denominator. Red and blue denote increases and decreases.

Dataset	k	Positive	Negative	Change
AMC23	1	$+2.93\%$	-2.19%	$+0.74\%$
	2	$+1.99\%$	-2.08%	-0.09%
	4	$+1.16\%$	-1.97%	-0.81%
	8	$+0.66\%$	-1.64%	-0.97%
	16	$+0.41\%$	-1.70%	-1.29%
	32	$+0.06\%$	-2.52%	-2.46%
	64	0.00%	-2.50%	-2.50%
AIME25	1	$+0.89\%$	-0.68%	$+0.21\%$
	2	$+0.84\%$	-0.81%	$+0.03\%$
	4	$+1.07\%$	-1.27%	-0.20%
	8	$+1.14\%$	-1.92%	-0.77%
	16	$+1.09\%$	-2.67%	-1.58%
	32	$+1.67\%$	-3.71%	-2.04%
	64	$+3.33\%$	-6.67%	-3.33%

Contribution of gains and losses. Table 7 shows that the AMC23 gains contribute $+2.93\%$ to Pass@1 and $+0.06\%$ to Pass@32, while the losses contribute -2.19% and -2.52%, respectively. On AIME25, the corresponding gains are $+0.89\%$ and $+1.67\%$, and the losses are -0.68% and -3.71%. The sign of the change reverses because increasing the budget changes the relative contribution of improvements and declines across problems. This dependence on the distribution of success across prompts aligns with the construction of the lower bound in Theorem 4.

Table 8 reports the sizes of groups A–D, their changes in correct counts, and their contributions at seven budgets. At each budget, the four contributions sum to the full benchmark change.

Gains on problems near saturation. On AMC23, the 20 problems in group A add 71 correct responses and 2.77% to Pass@1. Their Pass@32 contribution is approximately 5.83×10^{-12} on the 0–1 scale, and their Pass@64 contribution is zero. The two problems in group B lose three correct responses, reducing Pass@1 by 0.12%, Pass@32 by 2.52%, and Pass@64 by 2.50%. At budget 32, the improving problems are already almost certain to contain a hit within the sampled batch, leaving little room for additional gains. The lost hits on the two problems with few correct responses therefore dominate the change at larger budgets. This is the imbalance used in the construction of Theorem 5: improvements on an easier prompt raise average Pass@1, and losses on a harder prompt yield the Pass@k decline in Eq. (9).

On AIME25, the five problems in group A add 16 correct responses, contributing $+0.83\%$ to Pass@1 and $+0.01\%$ to Pass@32. Group B loses six correct responses, contributing -0.31%, -3.71%, and -6.67% at budgets 1, 32, and 64. The new hit in group C offsets part of these losses, contributing $+0.05\%$, $+1.67\%$, and $+3.33\%$, respectively. The remaining problems contribute -0.36% to

Table 8: Contributions by problem group: A (easy problems), B (hard problems), C (newly solved problems), and D (other problems). The groups are defined above. Each checkpoint uses 64 responses per problem. Δc is the change in total correct responses. Pass@k changes are new minus old on the percentage scale, using the full benchmark denominator. Red and blue denote increases and decreases.

Group	Problems	Δc	$k=1$	$k=2$	$k=4$	$k=8$	$k=16$	$k=32$	$k=64$
AMC23									
A	20	+71	+2.77%	+1.71%	+0.71%	+0.10%	0.00%	0.00%	0.00%
B	2	−3	−0.12%	−0.23%	−0.45%	−0.87%	−1.58%	−2.52%	−2.50%
C	0	0	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
D	18	−49	−1.91%	−1.57%	−1.07%	−0.20%	+0.28%	+0.06%	0.00%
AIME25									
A	5	+16	+0.83%	+0.74%	+0.86%	+0.73%	+0.25%	+0.01%	0.00%
B	3	−6	−0.31%	−0.60%	−1.08%	−1.81%	−2.66%	−3.71%	−6.67%
C	1	+1	+0.05%	+0.10%	+0.21%	+0.42%	+0.83%	+1.67%	+3.33%
D	21	−7	−0.36%	−0.21%	−0.18%	−0.11%	−0.01%	0.00%	0.00%

Pass@1, approximately -1.54×10^{-7} on the 0–1 scale to Pass@32, and zero to Pass@64. Thus the new hit provides substantial gains at larger budgets, while the losses on three problems with few correct responses still dominate. This second comparison exhibits the same allocation of gains and losses underlying Theorem 5.

Grouping by the old correct count. Tables 9 and 10 group problems by their old correct count: 0, 1–2, 3–7, 8–15, 16–31, and 32–64. Membership is fixed by the old checkpoint, and each group retains both positive and negative changes. The AMC23 3–7 group improves, and AIME25 gains a hit among previously unhit problems. These contributions show how improvements on some problems with few correct responses offset losses on others. The difference between these gains and losses produces the benchmark trade-off, consistent with the averaging across prompts in Eq. (7).

Table 9: Problem counts grouped by the old correct count out of 64. "Up/down/same" counts increases, decreases, and unchanged correct counts. $\sum c_{\text{old}}$ and $\sum c_{\text{new}}$ are group totals.

Dataset	Old count	Problems	Up/down/same	$\sum c_{\text{old}}$	$\sum c_{\text{new}}$
AMC23	0	1	0/0/1	0	0
	1–2	2	0/2/0	4	1
	3–7	1	1/0/0	5	9
	8–15	1	0/1/0	13	10
	16–31	6	4/2/0	153	159
	32–64	29	16/9/4	1551	1566
AIME25	0	13	1/0/12	0	1
	1–2	4	0/2/2	5	3
	3–7	3	0/1/2	17	13
	8–15	2	2/0/0	22	29
	16–31	3	1/2/0	54	53
	32–64	5	2/3/0	286	289

Table 10: Contributions within the groups in Table 9. Changes are new minus old on the percentage scale, using the full benchmark denominator. Red and blue denote increases and decreases.

Old count	Contribution	$k=1$	$k=2$	$k=4$	$k=8$	$k=16$	$k=32$	$k=64$
AMC23								
0	Positive	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Negative	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Change	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Continued on next page

Table 10 (continued)

Old count	Contribution	$k=1$	$k=2$	$k=4$	$k=8$	$k=16$	$k=32$	$k=64$
1–2	Positive	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Negative	−0.12%	−0.23%	−0.45%	−0.87%	−1.58%	−2.52%	−2.50%
	Change	−0.12%	−0.23%	−0.45%	−0.87%	−1.58%	−2.52%	−2.50%
3–7	Positive	+0.16%	+0.28%	+0.45%	+0.56%	+0.41%	+0.06%	0.00%
	Negative	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Change	+0.16%	+0.28%	+0.45%	+0.56%	+0.41%	+0.06%	0.00%
8–15	Positive	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Negative	−0.12%	−0.19%	−0.26%	−0.23%	−0.07%	0.00%	0.00%
	Change	−0.12%	−0.19%	−0.26%	−0.23%	−0.07%	0.00%	0.00%
16–31	Positive	+1.02%	+1.07%	+0.60%	+0.09%	0.00%	0.00%	0.00%
	Negative	−0.78%	−1.10%	−1.09%	−0.53%	−0.05%	0.00%	0.00%
	Change	+0.23%	−0.03%	−0.50%	−0.43%	−0.05%	0.00%	0.00%
32–64	Positive	+1.76%	+0.64%	+0.11%	0.00%	0.00%	0.00%	0.00%
	Negative	−1.17%	−0.55%	−0.16%	−0.01%	0.00%	0.00%	0.00%
	Change	+0.59%	+0.09%	−0.05%	−0.01%	0.00%	0.00%	0.00%
AIME25								
0	Positive	+0.05%	+0.10%	+0.21%	+0.42%	+0.83%	+1.67%	+3.33%
	Negative	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Change	+0.05%	+0.10%	+0.21%	+0.42%	+0.83%	+1.67%	+3.33%
1–2	Positive	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Negative	−0.10%	−0.21%	−0.42%	−0.83%	−1.67%	−3.33%	−6.67%
	Change	−0.10%	−0.21%	−0.42%	−0.83%	−1.67%	−3.33%	−6.67%
3–7	Positive	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Negative	−0.21%	−0.39%	−0.67%	−0.97%	−0.99%	−0.38%	0.00%
	Change	−0.21%	−0.39%	−0.67%	−0.97%	−0.99%	−0.38%	0.00%
8–15	Positive	+0.36%	+0.60%	+0.80%	+0.71%	+0.25%	+0.01%	0.00%
	Negative	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Change	+0.36%	+0.60%	+0.80%	+0.71%	+0.25%	+0.01%	0.00%
16–31	Positive	+0.05%	+0.07%	+0.06%	+0.02%	0.00%	0.00%	0.00%
	Negative	−0.10%	−0.16%	−0.18%	−0.11%	−0.01%	0.00%	0.00%
	Change	−0.05%	−0.09%	−0.13%	−0.09%	−0.01%	0.00%	0.00%
32–64	Positive	+0.42%	+0.07%	0.00%	0.00%	0.00%	0.00%	0.00%
	Negative	−0.26%	−0.06%	0.00%	0.00%	0.00%	0.00%	0.00%
	Change	+0.16%	+0.01%	0.00%	0.00%	0.00%	0.00%	0.00%

Individual problems and cancellation. AMC23-11, concerning distances between vertices of an icosahedron, changes from two correct responses to zero and accounts for the full 2.50% decline in Pass@64. AMC23-20, a problem about counting dates on a display, changes from two correct responses to one and retains its Pass@64 value. Together they contribute -2.52% to Pass@32. AMC23-6 improves from five to nine correct responses and contributes $+0.06\%$ to Pass@32. Larger count gains on AMC23-15 ($35 \to 46$), AMC23-36 ($24 \to 33$), and AMC23-37 ($31 \to 40$) add almost zero Pass@32 because their old values are already near one. These problems make concrete the different effects of gains and losses on the two prompts in the construction of Theorem 5.

On AIME25, AIME25-13, about minimizing distances, and AIME25-29, about a rational function, each change from one correct response to zero. AIME25-23, concerning zeros and tangencies of a trigonometric composition, changes from zero to one and offsets one of these lost hits at budget 64. AIME25-4, which counts multiples of 22 with eight digits, changes from seven correct responses to three and contributes -0.38% to Pass@32 while retaining its Pass@64 value. The improvements in group A occur on AIME25-1 ($8 \to 13$), AIME25-2 ($56 \to 59$), AIME25-3 ($22 \to 23$), AIME25-5 ($56 \to 61$), and AIME25-18 ($14 \to 16$). Their old Pass@32 values are already close to one, which limits their combined contribution to $+0.01\%$. This contrast with the losses on problems with few correct responses follows the saturation effect in Eq. (7) used in Theorem 5.

Sampling and measurement scope. These selected comparisons describe one evaluation batch per checkpoint. For a fixed true success probability of $1/64$, the probability of zero successes in 64 independent responses is $(63/64)^{64} \approx 0.365$. Establishing a change in the underlying success

probability therefore requires repeated evaluation. The 64 budget estimates here share samples and are statistically dependent. The correspondence with Theorem 5 concerns the allocation of gains and losses across problems. Examining its mechanism at the level of individual responses would require measuring teacher probabilities on the student's correct responses and their changes under the update in Eq. (8). Comparing the KL and Pass@k changes would additionally require KL over complete responses for the relevant policies under the same evaluation protocol. Estimating the frequency of this failure mode over training would require a broader set of checkpoint pairs and training runs.

Reproducibility. Tables 11 and 12 give all 70 problems, identified by dataset `example_id` values starting at zero, and their contributions at budgets 1, 32, and 64. The correct counts determine all other budgets through Eq. (35). Anonymous supplementary CSVs give every integer budget from 1 to 64, signed contributions, and group membership. The accompanying Python script recomputes these values with exact rational arithmetic. The 4480 comparisons of individual problems and 128 benchmark comparisons reproduce the supplied evaluation summaries.

Table 11: All 40 AMC23 problems for the 4B student, C60 → C80. IDs are dataset `example_id` values starting at zero. c_{old} and c_{new} are correct counts out of 64. Changes are new minus old on the percentage scale, using the full benchmark denominator. Red and blue denote increases and decreases.

ID	c_{old}	c_{new}	Group	ΔPass@1	ΔPass@32	ΔPass@64
0	62	63	A	+0.04%	0.00%	0.00%
1	62	61	D	−0.04%	0.00%	0.00%
2	64	64	D	0.00%	0.00%	0.00%
3	62	63	A	+0.04%	0.00%	0.00%
4	45	48	A	+0.12%	0.00%	0.00%
5	44	49	A	+0.20%	0.00%	0.00%
6	5	9	D	+0.16%	+0.06%	0.00%
7	33	28	D	−0.20%	0.00%	0.00%
8	49	52	A	+0.12%	0.00%	0.00%
9	64	58	D	−0.23%	0.00%	0.00%
10	38	36	D	−0.08%	0.00%	0.00%
11	2	0	B	−0.08%	−1.88%	−2.50%
12	23	13	D	−0.39%	0.00%	0.00%
13	49	47	D	−0.08%	0.00%	0.00%
14	52	46	D	−0.23%	0.00%	0.00%
15	35	46	A	+0.43%	0.00%	0.00%
16	62	63	A	+0.04%	0.00%	0.00%
17	62	63	A	+0.04%	0.00%	0.00%
18	25	15	D	−0.39%	0.00%	0.00%
19	62	63	A	+0.04%	0.00%	0.00%
20	2	1	B	−0.04%	−0.63%	0.00%
21	62	64	A	+0.08%	0.00%	0.00%
22	61	60	D	−0.04%	0.00%	0.00%
23	62	62	D	0.00%	0.00%	0.00%
24	54	59	A	+0.20%	0.00%	0.00%
25	44	41	D	−0.12%	0.00%	0.00%
26	59	55	D	−0.16%	0.00%	0.00%
27	13	10	D	−0.12%	0.00%	0.00%
28	23	30	A	+0.27%	0.00%	0.00%
29	52	54	A	+0.08%	0.00%	0.00%
30	59	63	A	+0.16%	0.00%	0.00%
31	0	0	D	0.00%	0.00%	0.00%
32	33	33	D	0.00%	0.00%	0.00%
33	61	64	A	+0.12%	0.00%	0.00%
34	63	64	A	+0.04%	0.00%	0.00%
35	27	28	A	+0.04%	0.00%	0.00%
36	24	33	A	+0.35%	0.00%	0.00%

Continued on next page

Table 11 (continued)

ID	c_{old}	c_{new}	Group	ΔPass@1	ΔPass@32	ΔPass@64
37	31	40	A	+0.35%	0.00%	0.00%
38	48	49	A	+0.04%	0.00%	0.00%
39	48	48	D	0.00%	0.00%	0.00%

Table 12: All 30 AIME25 problems for the 4B student, C80 → C100. IDs are dataset `example_id` values starting at zero. c_{old} and c_{new} are correct counts out of 64. Changes are new minus old on the percentage scale, using the full benchmark denominator. Red and blue denote increases and decreases.

ID	c_{old}	c_{new}	Group	ΔPass@1	ΔPass@32	ΔPass@64
0	62	60	D	−0.10%	0.00%	0.00%
1	8	13	A	+0.26%	+0.01%	0.00%
2	56	59	A	+0.16%	0.00%	0.00%
3	22	23	A	+0.05%	0.00%	0.00%
4	7	3	B	−0.21%	−0.38%	0.00%
5	56	61	A	+0.26%	0.00%	0.00%
6	0	0	D	0.00%	0.00%	0.00%
7	7	7	D	0.00%	0.00%	0.00%
8	16	15	D	−0.05%	0.00%	0.00%
9	0	0	D	0.00%	0.00%	0.00%
10	0	0	D	0.00%	0.00%	0.00%
11	1	1	D	0.00%	0.00%	0.00%
12	0	0	D	0.00%	0.00%	0.00%
13	1	0	B	−0.05%	−1.67%	−3.33%
14	0	0	D	0.00%	0.00%	0.00%
15	60	59	D	−0.05%	0.00%	0.00%
16	52	50	D	−0.10%	0.00%	0.00%
17	0	0	D	0.00%	0.00%	0.00%
18	14	16	A	+0.10%	0.00%	0.00%
19	3	3	D	0.00%	0.00%	0.00%
20	16	15	D	−0.05%	0.00%	0.00%
21	2	2	D	0.00%	0.00%	0.00%
22	0	0	D	0.00%	0.00%	0.00%
23	0	1	C	+0.05%	+1.67%	+3.33%
24	0	0	D	0.00%	0.00%	0.00%
25	0	0	D	0.00%	0.00%	0.00%
26	0	0	D	0.00%	0.00%	0.00%
27	0	0	D	0.00%	0.00%	0.00%
28	0	0	D	0.00%	0.00%	0.00%
29	1	0	B	−0.05%	−1.67%	−3.33%