WHY CAN POLICY DISTILLATION EVADE DATA CONTAMINATION DETECTION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Data contamination detection is essential for reliable LLM benchmark evaluation. Conventional detectors often exploit unusually confident predictions associated with overfitting to training data. However, we find that policy distillation can expose students to evaluation data without producing this signature, as soft-target training encourages teacher agreement rather than certainty about corpus tokens. We theoretically show that distillation exposure can remain undetected by conventional detectors, whereas hard-label training can become detectable. To systematically study this problem, we construct Distillation-MIA, a benchmark for contamination detection in distillation post-training. Experiments show that conventional detectors perform strongly under matched supervised fine-tuning but fall close to chance under both online and offline distillation. We further propose two teacher–student KL-based detection methods, PD-Div and PD-Div-Gain. PD-Div-Gain improves average AUROC over the strongest conventional detector by 13.1% and 12.4% under online and offline distillation. Finally, an exploratory Qwen3 audit extends our analysis to a public-model setting. Our findings reveal the difficulty of detecting contamination introduced by policy distillation, which remains an important problem for further investigation. Code is available at `https://anonymous.4open.science/r/PD-DIv-Gain-0DB3`.

1 INTRODUCTION

Large language models (LLMs) are evaluated on benchmarks intended to measure generalization beyond their training corpora (White et al., 2025; Liu et al., 2026; Zhao et al., 2026). When evaluation data have already been used in training, the performance can no longer be interpreted as a clean measure of generalization (Zhuang et al., 2025; Wang et al., 2026). Contamination detection aims to mitigate this problem by identifying exposure and screening evaluation corpora accordingly.

Conventional contamination detectors look for unusually high token likelihood, low predictive entropy, or elevated probabilities on otherwise difficult tokens (Yeom et al., 2018; Song & Mittal, 2021; Shi et al., 2024; Zhang et al., 2024; 2025a; Ke et al., 2026a;b). However, the rapid development of post-training and the growing use of policy distillation challenge this premise (Agarwal et al., 2024; Yang et al., 2025). Distillation on evaluation corpora encourages students to match a teacher's soft predictions rather than fit hard token labels. Repeated optimization therefore need not drive token probabilities toward certainty. If the teacher does not exhibit the anomalies targeted by a detector, the student may also remain undetected despite having trained on the evaluation corpus.

Figure 1 illustrates this distinction through an exam analogy. Conventional contamination resembles seeing the exam questions and answers beforehand, which can leave a trace in the student's unusually confident responses. Distillation instead resembles studying those questions with a teacher without seeing the answers. The student's responses may look like those of the teacher taking the exam: *prior exposure has occurred, but its signature is harder to distinguish from ordinary model behavior.*

We formalize this failure by comparing hard-label and soft-target training on the same text. For a Lipschitz detection score, if the initial student and teacher both have positive non-member margins and the KL logit-flow path is sufficiently short relative to those margins, the entire distillation trajectory stays on the non-member side. In contrast, hard-label cross-entropy eventually crosses the detection boundary when the hard target lies on the detected side and the score extends continuously to it along the CE path. Empirical analysis further supports this detection limitation. Across online (on-policy)

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

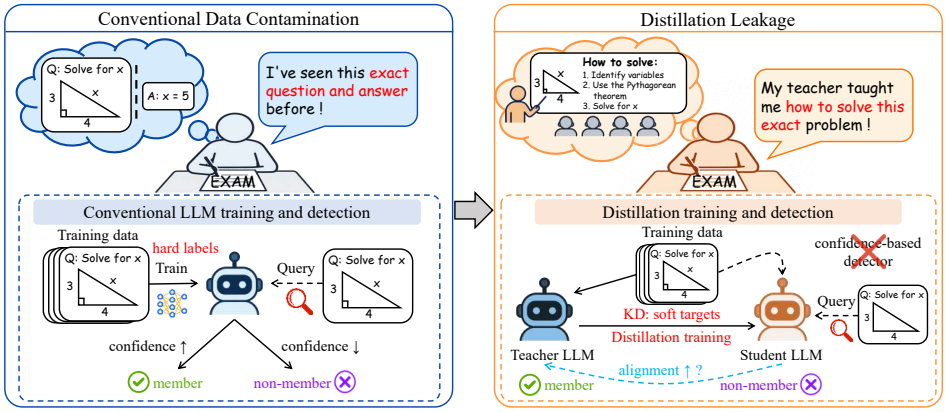

Figure 1: Direct pretraining exposure resembles seeing exam questions and answers in advance; distillation exposure resembles studying the same questions with a teacher without seeing the answers.

and offline (off-policy) distillation, representative conventional baselines achieve average AUROCs of 48.0%–54.8%, close to chance-level performance. Thus, despite known training exposure, existing detectors fail to reliably identify distillation contamination.

The analysis identifies the training target as central to this failure: the teacher itself lies on the conventional detector's non-member side. Therefore, we propose a teacher–student KL-based detection approach that places close agreement with the teacher on the detected side. We derive a lower bound on the expected non-member-minus-member KL gap; when positive, it establishes a directional membership signal. Although we have validated the effectiveness of this method in controlled experiments, this approach requires access to the relevant teacher's output distributions and, for checkpoint calibration, the student immediately before distillation. Such access is often unavailable in real-world audits. Our main contributions are therefore the identification of this detection failure and a formal characterization of conditions under which it persists, while reliable contamination auditing under realistic model access remains a major challenge.

2 Related Work

Membership inference and contamination auditing. Membership inference seeks to identify examples that participated in training (Shokri et al., 2017). For language models, likelihood and token-level statistics have been studied extensively in pretraining and supervised fine-tuning settings (Mireshghallah et al., 2022; Mattern et al., 2023; Fu et al., 2024; Shi et al., 2024; Xie et al., 2024; Dong et al., 2024; Zhang et al., 2024; 2025a; Ke et al., 2026a;b). Loss and entropy provide direct student-based signals (Yeom et al., 2018; Song & Mittal, 2021), while reference calibration and perturbation-based attacks use additional observations (Carlini et al., 2022; Choquette-Choo et al., 2021). Our negative theorem concerns a specified regular output score at a fixed threshold, not every possible membership attack. The empirical comparison is broader, but remains specific to the implementations and distillation protocols evaluated here.

Distillation and privacy. Knowledge distillation transfers a teacher's predictions to a student (Hinton et al., 2015); on-policy variants query the teacher on student-generated trajectories (Agarwal et al., 2024). Distillation has also been explored as a defense against membership leakage from a teacher's private training data (Shejwalkar & Houmansadr, 2021; Tang et al., 2022; Mazzone et al., 2022; Zheng et al., 2021). Subsequent work shows that teacher-data membership and memorization risks can persist in distilled students (Jagielski et al., 2023; Zhang et al., 2025b; Cui et al., 2026). Our focus is instead membership in the student's distillation corpus itself. Teacher-aware auditing is motivated by that training interaction, while the training–validation comparison is necessary because teacher agreement can also generalize (Stanton et al., 2021).

3 PRELIMINARIES

Student and teacher outputs. For a fixed sequence $x = (x_1, \ldots, x_m)$ over a vocabulary of size V, let $p_{\theta,j}(x), q_j(x) \in \Delta_{++}^{V-1}$ be the student and teacher next-token distributions at prefix $x_{<j}$. Here θ denotes student parameters, and $++$ means strictly positive coordinates. Write

$$P_\theta(x) = (p_{\theta,1}(x), \ldots, p_{\theta,m}(x)), \qquad Q(x) = (q_1(x), \ldots, q_m(x)). \tag{1}$$

These are tuples of conditional distributions with aligned vocabulary coordinates, not joint sequence distributions. We use natural logarithms and the flattened Euclidean norm $\|P\|_2^2 = \sum_{j,v} p_{j,v}^2$. Membership refers to use in the student's distillation corpus; a negative detector decision is not a ground-truth statement of non-exposure.

Conventional output scores. A conventional detector summarizes the student's predictions through a score s, oriented so that larger values indicate membership, and uses

$$F_\tau(P_\theta(x)) = \mathbf{1}\{s(P_\theta(x)) \geq \tau\}. \tag{2}$$

Our analysis uses Lipschitz continuity on a relevant domain $\mathcal{U}$:

$$|s(P) - s(P')| \leq L_s \|P - P'\|_2, \qquad P, P' \in \mathcal{U}. \tag{3}$$

Appendix B.1 proves this property for average log-probability, Min-K% Prob, label and maximum confidence, Shannon entropy, and fixed linear combinations of log-probability and entropy features when the relevant probabilities are bounded away from zero. It lets us translate output movement into score changes without assuming a convex non-member region.

4 WHY POLICY DISTILLATION CAN EVADE DETECTION

Training on evaluation corpora may not produce the anomalies sought by conventional detectors when the objective is teacher agreement rather than one-hot prediction. Section 4.1 formalizes this mechanism by comparing hard-label and soft-target gradient flows on the same text, establishing sufficient conditions for non-detection throughout training. Section 4.2 then provides experimental validation of the detection failure by evaluating eight conventional methods under controlled online and offline distillation.

4.1 TWO TARGETS, TWO PATHS

Hard-label cross-entropy (CE) increases the probability of observed corpus tokens, whereas knowledge distillation (KD) encourages agreement with a teacher, including its uncertainty. Repeated training can therefore reinforce different output patterns even when the same corpus supplies supervision. To isolate this difference, fix a sequence x and its prefixes, and start both objectives from the same student output P_0. Suppress the dependence on x and let $E = (e_{x_1}, \ldots, e_{x_m})$ collect the one-hot observed-token labels:

$$\mathcal{L}_{\mathrm{KD}}(P; Q) = \frac{1}{m} \sum_{j=1}^{m} \mathrm{KL}(q_j \| p_j) = \frac{1}{m} \sum_{j,v} q_{j,v} \log \frac{q_{j,v}}{p_{j,v}}, \tag{4}$$

$$\mathcal{L}_{\mathrm{CE}}(P; E) = -\frac{1}{m} \sum_{j=1}^{m} \log p_{j,x_j}. \tag{5}$$

The relevant distinction is where these targets lie relative to the detector. As illustrated in Figure 2 (left), a hard target can exhibit the extreme confidence that the detector flags, while the teacher may itself be classified as a non-member. Approaching such a teacher does not necessarily create a detectable signature. This endpoint intuition, however, is not enough: the student could cross the detection boundary during training before converging to the teacher.

We therefore analyze the entire optimization path. Write $p_j = \mathrm{softmax}(z_j)$ and $J(p) = \mathrm{diag}(p) - pp^\top$ for the softmax Jacobian (Goodfellow et al., 2016). Under ordinary Euclidean gradient flow on independent logits, the probability dynamics are

$$\dot{p}_j^{\mathrm{KD}} = J(p_j^{\mathrm{KD}})(q_j - p_j^{\mathrm{KD}}), \qquad \dot{p}_j^{\mathrm{CE}} = J(p_j^{\mathrm{CE}})(e_{x_j} - p_j^{\mathrm{CE}}). \tag{6}$$

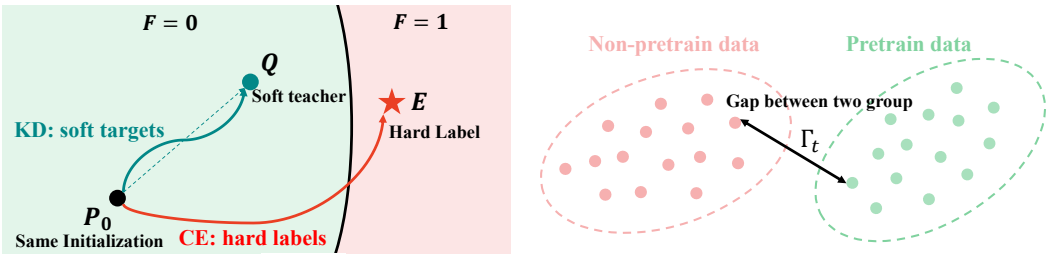

Figure 2: Overview of our theoretical framework. **Left:** Starting from the same initialization P_0, hard-label cross-entropy drives the model toward the one-hot target E, whereas knowledge distillation drives it toward the soft teacher distribution Q. **Right:** A positive lower bound in Theorem 2 establishes an expected teacher–student divergence gap between distillation members and independent validation examples.

These paths are generally curved, so a non-detection guarantee must account for intermediate outputs as well as the endpoints. For a strictly positive teacher, the independent-logit KD flow converges to Q with finite path length, as proved in Appendix A.1.

Lipschitz continuity connects movement along the path to changes in the detection score. Let $\rho_0 = \tau - s(P_0)$ and $\rho_Q = \tau - s(Q)$ denote the initial student's and teacher's margins below the threshold. If both margins are positive, any path that reaches the detected side and then converges to Q must cover enough distance for the score first to rise from $s(P_0)$ to the threshold and then to fall to $s(Q)$. A path whose length is smaller than the sum of the margins divided by the Lipschitz constant cannot make this excursion. This yields the following theorem.

Theorem 1 (Distillation can evade a detector that recognizes CE exposure)**.** *Let P_0, Q be strictly positive, and let the KD and CE outputs follow Equation* (6)*. Suppose $\rho_0, \rho_Q > 0$ and s is L_s-Lipschitz, for some $L_s > 0$, on a region containing P_0, Q and the entire KD trajectory. If*

$$L_s \int_0^\infty \left\| \frac{d}{du} P^{\mathrm{KD}}(u) \right\|_2 du < \rho_0 + \rho_Q, \tag{7}$$

then $F_\tau(P^{\mathrm{KD}}(u)) = 0$ for every $u \geq 0$. If, additionally, s has a continuous extension to E along the CE path and $s(E) > \tau$, a finite $u_\star$ exists such that

$$F_\tau(P^{\mathrm{CE}}(u)) = 1, \qquad F_\tau(P^{\mathrm{KD}}(u)) = 0 \quad \textit{for all } u \geq u_\star. \tag{8}$$

The proof is in Appendix A.1. Theorem 1 formalizes the distinction introduced above: the same text is used in both training processes, yet the detector eventually recognizes hard-label exposure while missing distillation exposure throughout the entire trajectory. Thus, the failure cannot be explained solely by stopping training too early: under the stated conditions, continued distillation toward a non-member-side target never produces the signature expected by the detector.

The theorem gives sufficient conditions for a specified detector to fail; it does not imply that all distillation is undetectable. Endpoint margins alone do not guarantee the path-length condition, and independent-logit flow does not capture the coupling of examples through shared neural parameters. Next, we test the practical implication: whether conventional detectors can reliably distinguish exposed from unexposed prompts after online or offline distillation.

4.2 Experimental Validation of Conventional Detection in Distillation

Distillation-MIA: A Benchmark for Data Contamination Detection in Distillation Post-training. To the best of our knowledge, no benchmark has been specifically designed to evaluate data contamination detection in distillation post-training for LLMs. To fill this gap, we construct Distillation-MIA, which simulates controlled data contamination by selectively including a subset of examples in the distillation training corpus. We use MATH (Hendrycks et al., 2021), MulDimIF (Ye et al., 2026), XSum (Narayan et al., 2018), and AIME 2026 (Dekoninck et al., 2026) to construct the benchmark, covering a diverse range of tasks. Prompts included in a student's distillation training set are labeled as members (i.e., label = 1), while held-out prompts are labeled as non-members (i.e., label = 0). For

each dataset, we partition the prompts into two disjoint groups, A and B, and train separate students S_A and S_B. The two groups exchange member and non-member roles across the students, thereby reducing potential dependence on sample-specific characteristics or any particular split. Details of the benchmark construction and experimental setup are provided in Appendix C.

Models and training methods. We use LoRA fine-tuning to train the student model Qwen2.5-1.5B-Instruct, with Qwen2.5-7B-Instruct serving as the frozen teacher (Yang et al., 2024). In the online setting, training uses evolving student-generated prefixes; in the offline setting, training uses teacher-generated prefixes. We also conduct supervised fine-tuning (SFT) under matched settings as a comparison. Figure 3 compares Pass@4 before and after distillation on MATH and AIME 2026. It means we actually achieved a score improvement in this setting.

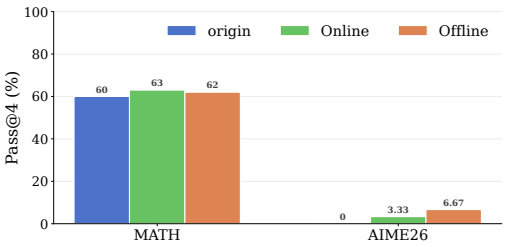

Figure 3: Pass@4 (%) before distillation (origin) and after online or offline distillation.

Conventional detectors. We evaluate eight conventional detection methods. **Loss** (Yeom et al., 2018) uses the likelihood of the generated completion; **Zlib**, **Reference**, and **Lowercase** (Carlini et al., 2021) normalize or compare likelihood using compression, a separate reference model, or a lowercased completion. **Min-K%** (Shi et al., 2024) selects the $k\%$ of tokens with the lowest log probabilities, while **Min-K%++** (Zhang et al., 2025a) additionally standardizes each token's log probability using the model's full next-token distribution over the vocabulary. **ETD** (Ke et al., 2026a) and **AECA** (Ke et al., 2026b) incorporate entropy- or token-frequency-based calibration, respectively.

Metrics. We report AUROC and true positive rate (TPR) at low false-positive rates (FPRs). For each dataset–setting combination, we average the results of S_A and S_B. For MATH, MulDimIF, and XSum, we report TPR@5%FPR. Since AIME 2026 has only 15 non-members per student, we report TPR@2FP, equivalent to TPR@13.3%FPR.

Table 1: Conventional detection AUROC (%) under SFT and offline (Off) or online (On) distillation.

	MATH			MulDimIF			XSum			AIME26			AVG		
Method	SFT	Off	On	SFT	Off	On	SFT	Off	On	SFT	Off	On	**SFT**	**Off**	**On**
Loss	62.0	52.0	52.4	90.8	49.8	52.3	98.1	54.4	56.4	78.4	48.3	53.6	**82.3**	**51.1**	**53.7**
Zlib	60.9	51.3	51.7	92.2	51.6	53.1	98.8	54.8	55.5	79.1	52.0	53.6	**82.7**	**52.4**	**53.5**
Reference	66.6	49.5	48.9	59.9	51.2	49.5	96.8	52.7	48.5	60.9	49.3	45.2	**71.0**	**50.7**	**48.0**
Lowercase	60.8	51.9	52.7	73.9	49.3	50.6	94.2	55.6	56.6	78.2	53.4	59.1	**76.8**	**52.6**	**54.8**
Min-K%	64.3	52.0	52.4	91.9	49.6	50.4	98.4	53.2	54.9	79.1	49.2	54.0	**83.4**	**51.0**	**52.9**
Min-K++	71.1	50.2	52.1	89.4	50.7	52.7	98.4	49.6	52.4	89.3	54.0	54.0	**87.0**	**51.1**	**52.8**
ETD	65.7	51.7	51.7	92.4	49.6	50.7	99.0	52.1	53.3	94.4	49.6	57.6	**87.9**	**50.8**	**53.3**
AECA	61.6	51.1	52.2	78.1	49.8	51.0	96.4	53.7	52.7	66.4	49.6	51.8	**75.6**	**51.1**	**51.9**

Detection fails under both online and offline distillation. In Table 1, the average AUC scores of all eight conventional detection methods range from 48.0–54.8% across the online and offline settings, close to chance-level performance. In contrast, under the matched SFT setting, the same methods achieve substantially higher average AUC scores of 71.0–87.9%. Some cases show stronger signals, including 56.6% on online XSum and 59.1% on online AIME 2026 for Lowercase. Thus, the results do not imply that every case is exactly at chance, but rather that *conventional detectors fail to consistently distinguish members from non-members across tasks and distillation settings.* Table 2 further shows that TPR@5%FPR under distillation is substantially lower than under SFT, indicating that conventional detectors do not reliably transfer from hard-label training to distillation.

The theoretical and empirical findings support the same conclusion at different levels. Theorem 1 gives conditions under which soft-target exposure remains undetected throughout optimization; the controlled experiments show that the evaluated conventional detectors fail to reliably identify known

Table 2: Conventional detection at low false-positive rates (%), with the same column as Table 1.

	MATH			MulDimIF			XSum			AIME26			AVG		
Method	SFT	Off	On	SFT	Off	On	SFT	Off	On	SFT	Off	On	**SFT**	**Off**	**On**
Loss	9.5	4.5	6.0	52.5	6.5	9.5	89.5	8.5	7.0	56.7	13.3	26.7	**52.0**	**8.2**	**12.3**
Zlib	8.5	6.5	7.5	59.0	7.0	11.0	95.0	10.5	6.0	53.3	23.3	23.4	**54.0**	**11.8**	**12.0**
Reference	22.5	6.5	5.0	11.5	7.5	8.0	82.5	6.5	4.0	33.3	16.7	13.4	**37.5**	**9.3**	**7.6**
Lowercase	12.5	3.5	6.0	35.5	7.5	11.0	81.0	11.0	6.5	70.0	20.0	36.7	**49.8**	**10.5**	**15.1**
Min-K%	10.0	4.5	6.0	55.0	7.0	9.0	92.0	10.5	9.5	56.7	13.3	26.7	**53.4**	**8.8**	**12.8**
Min-K++	12.5	3.5	4.0	27.0	9.0	7.0	92.0	10.5	6.5	80.0	20.0	26.7	**52.9**	**10.8**	**11.1**
ETD	13.5	4.5	6.0	51.5	8.5	6.5	93.5	10.5	8.5	90.0	23.4	33.3	**62.1**	**11.7**	**13.6**
AECA	14.5	4.0	5.5	13.0	10.0	8.0	83.5	5.5	13.5	30.0	13.4	23.4	**35.2**	**8.2**	**12.6**

exposure under both online and offline distillation. Although the practical runs use shared neural parameters and JSD rather than the theorem's independent-logit KL flow, they demonstrate the auditing limitation that motivates the analysis: **a negative conventional detection result cannot reliably rule out participation in the distillation corpus.**

5 A Teacher-Aware Direction and Its Limits

Returning to Theorem 1, we identify a key factor in the detectability of training exposure: whether the training target lies on the detector's member or non-member side. In the failure mechanism above, distillation moves the student toward a teacher that the conventional detector itself classifies as a non-member. This motivates a different decision rule that places close teacher agreement on the detected side. In this section, we use teacher–student KL to construct such a detector, derive a lower bound on its expected non-member-minus-member divergence gap, and evaluate its empirical signal and practical access limitations.

5.1 Teacher–Student KL as a Detector

A natural way to break the negative-teacher-endpoint condition is to make agreement with the teacher member-oriented, regardless of the teacher's confidence. We implement this decision rule using the forward-KL training objective as a score:

$$k_\theta(x) := d_{\mathrm{KL}}(P_\theta(x), Q(x)) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{KL}(q_j(x) \| p_{\theta,j}(x)), \tag{9}$$

and predict membership using

$$G_\gamma(P_\theta(x), Q(x)) = \mathbf{1}\{k_\theta(x) \le \gamma\}, \qquad \gamma > 0. \tag{10}$$

Both models are evaluated on identical prefixes, and divergences use the complete vocabulary. At exact agreement, $G_\gamma(Q, Q) = 1$: the teacher endpoint is now on the detected side, so the negative-teacher-endpoint premise of Theorem 1 no longer applies to this rule. Moreover, forward KL decreases along the fixed-sequence KD logit flow (Appendix A.2). Thus the new score measures a change directly encouraged by the objective.

This resolves the geometric mismatch, but not yet the statistical problem. Low KL can predate training, and teacher agreement can improve on unseen examples through generalization (Stanton et al., 2021). Establishing a membership signal therefore requires a comparison between members and independent non-members under the same trained model, rather than KL descent on training examples alone.

5.2 Expected Training–Validation KL Separation

To account for generalization, we move from independent logits on one sequence to shared-parameter training on a sampled corpus. Both members and non-members now change under the same parameter

updates. Drawing them from the same population removes a separate source of score differences: systematic dataset mismatch.

Sampling and training. Let $\mathcal{P}$ be a population of length-m documents. Draw the training corpus $D = (X_1, \ldots, X_N) \sim \mathcal{P}^N$ and, independently, an untouched validation set $D_{\text{val}} = (Z_1, \ldots, Z_M) \sim \mathcal{P}^M$. The teacher and initialization θ_0 are fixed independently of these draws; validation examples do not select the model, stopping time, or threshold in this theoretical experiment. The student's parameter domain is a fixed $\Theta \subseteq \mathbb{R}^d$. Its empirical and population KL risks are

$$\widehat{R}_D(\theta) = \frac{1}{N}\sum_i k_\theta(X_i), \qquad R(\theta) = \mathbb{E}_{X\sim\mathcal{P}} k_\theta(X). \tag{11}$$

The same trained parameter θ_t is used for both groups, with

$$\dot{\theta}_t = -\nabla \widehat{R}_D(\theta_t), \qquad \theta_{t=0} = \theta_0. \tag{12}$$

Here t is shared-parameter training time, distinct from the independent-logit time u used above. Define the expected validation-minus-training gap

$$\Gamma_t = \mathbb{E}\left[\frac{1}{M}\sum_{a=1}^{M} k_{\theta_t}(Z_a) - \frac{1}{N}\sum_{i=1}^{N} k_{\theta_t}(X_i)\right] = \mathbb{E}[R(\theta_t) - \widehat{R}_D(\theta_t)]. \tag{13}$$

A positive gap means that members have lower expected KL. Therefore, the following theorem has been derived:

Theorem 2 (Finite-time training–validation KL gap)**.** *In the sampling experiment above, define*

$$\mu_0 = \nabla R(\theta_0), \qquad \sigma_0^2 = \mathbb{E}_{X\sim\mathcal{P}}\|\nabla k_{\theta_0}(X) - \mu_0\|_2^2, \tag{14}$$

$$\Delta_0 = R(\theta_0) - R_{\inf}, \qquad R_{\inf} = \inf_{\theta\in\Theta} R(\theta). \tag{15}$$

Suppose gradients commute with population expectation at θ_0, $R(\theta_0) < \infty$, $\sigma_0^2 < \infty$, and the displayed risks are integrable. For almost every D, suppose the flow in Equation (12) exists on $[0, T]$, remains in Θ, and $\widehat{R}_D$ is twice continuously differentiable near its trajectory, with a deterministic $L_k > 0$ satisfying

$$\nabla^2 \widehat{R}_D(\theta_t) \preceq L_k I_d \qquad (0 \le t \le T). \tag{16}$$

Then

$$\Gamma_t \ge B_t := \frac{1 - e^{-2L_k t}}{2L_k}\left(\|\mu_0\|_2^2 + \frac{\sigma_0^2}{N}\right) - \Delta_0 \qquad (0 \le t \le T). \tag{17}$$

The proof is provided in Appendix A.3. When the lower bound is positive, Theorem 2 shows that *training examples attain lower expected teacher–student KL than independent validation examples, establishing a directional membership signal and implying the existence of a low-KL threshold with positive membership advantage, averaged over training draws.*

5.3 From KL Separation to Practical Policy Auditing

In this section, we propose two KL-based detection methods with different model-access requirements. Specifically, let T denote the teacher, S_0 the initial student, and S_1 the trained student. For prompt x, sample K completions $y^{(k)} = (y_1^{(k)}, \ldots, y_{L_k}^{(k)})$ from S_1 and evaluate the required models on identical prefixes $z_t^{(k)} = (x, y_{<t}^{(k)})$. Then we formalize the two methods as follows:

PD-Div. With the teacher and trained student, the score measures their post-training agreement:

$$\text{PD-Div}(x) = -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{L_k}\sum_{t=1}^{L_k}\text{KL}\Big(T(\cdot \mid z_t^{(k)}) \,\|\, S_1(\cdot \mid z_t^{(k)})\Big). \tag{18}$$

PD-Div-Gain. The initial checkpoint additionally allows calibration for pre-existing teacher agreement by measuring the reduction in divergence:

$$\text{PD-Div-Gain}(x) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{L_k}\sum_{t=1}^{L_k}\Big[\text{KL}\Big(T(\cdot \mid z_t^{(k)}) \,\|\, S_0(\cdot \mid z_t^{(k)})\Big) - \text{KL}\Big(T(\cdot \mid z_t^{(k)}) \,\|\, S_1(\cdot \mid z_t^{(k)})\Big)\Big]. \tag{19}$$

As summarized in Table 3, the choice of method depends on the available models. PD-Div requires the teacher and trained student, measuring their agreement without calibrating for pre-existing similarity. PD-Div-Gain additionally requires the initial student and measures the change relative to initialization.

Table 3: Model access required by each detection method: ✓ required, × not required.

Model	PD-Div	PD-Div-Gain
Teacher T	✓	✓
Initial student S_0	×	✓
Trained student S_1	✓	✓

5.4 Empirical Signal under Privileged Model Access

We now apply these two methods to the same Distillation-MIA students and cached completions used in Section 4.2. Table 4 reports their detection performance under online and offline distillation.

Table 4: The AUROC (%) and low-FPR detection performance (%) of the two proposed detection methods on Distillation-MIA under online and offline distillation. The setting is the same as Table 1.

	AUC					FPR				
Method	MATH	MulDimIF	XSum	AIME26	**AVG**	MATH	MulDimIF	XSum	AIME26	**AVG**
Online										
PD-Div	55.2	56.5	68.9	56.9	59.4	9.5	9.0	19.5	33.4	17.9
PD-Div-Gain	**63.0**	**59.7**	**86.5**	**62.5**	**67.9**	**16.0**	**12.0**	**48.5**	**43.4**	**30.0**
Offline										
PD-Div	55.0	55.0	65.2	51.4	56.7	9.5	11.5	15.0	20.0	14.0
PD-Div-Gain	**64.2**	**59.1**	**80.7**	**55.8**	**65.0**	**18.5**	**15.5**	**40.0**	**26.7**	**25.2**

Teacher access strengthens the detection signal. As shown in Table 4, PD-Div-Gain achieves the best average detection performance, with average AUC results of 67.9% and 65.0% under online and offline distillation. PD-Div, which relies on teacher–student divergence without initialization calibration, achieves 59.4% and 56.7%, *consistently exceeding the best average performance of the conventional detection methods in both settings*. These results indicate that teacher–student divergence provides a stronger signal for detecting distillation exposure than conventional detection methods, while calibrating this divergence against the initial student further improves detection.

The access requirements limit practical deployment. PD-Div assumes the actual teacher is available for full-distribution scoring; PD-Div-Gain additionally needs the student immediately before the relevant distillation stage. These references are controlled in our experiment but may be unavailable in a public-model audit. A released base model may not be the pre-distillation checkpoint, and a larger model from the same family need not be the teacher. Consequently, empirical success with privileged access does not resolve the original detection problem under ordinary model access. The following case study makes this distinction concrete.

5.5 Exploratory Audit of Public Qwen3 Checkpoints

The controlled benchmark evaluates detection against known exposure, whereas a public-model audit must work with incomplete information about training membership and the distillation pipeline. We use Qwen3-8B as a case study because its technical report describes strong-to-weak distillation with off-policy response distillation and on-policy logit alignment for smaller models (Yang et al., 2025, Section 4.5). The audit characterizes cross-dataset score patterns without treating them as verified positive membership labels.

Specifically, we use the public Qwen3-8B-Base as S_0, Qwen3-8B as S_1, and Qwen3-32B as an accessible reference teacher T. We score 30 problems each from AIME 2024, AIME 2025, AIME 2026. The original Qwen3 family was released in April 2025, whereas the 2026 AIME examinations took place in February 2026.[1] If the audited weights are unchanged from that release, and the exact AIME 2026 problems were unavailable before it, AIME 2026 serves as a *temporal negative control* for exact-problem exposure.

[1] Qwen3 release announcement; MAA AIME 2026 schedule.

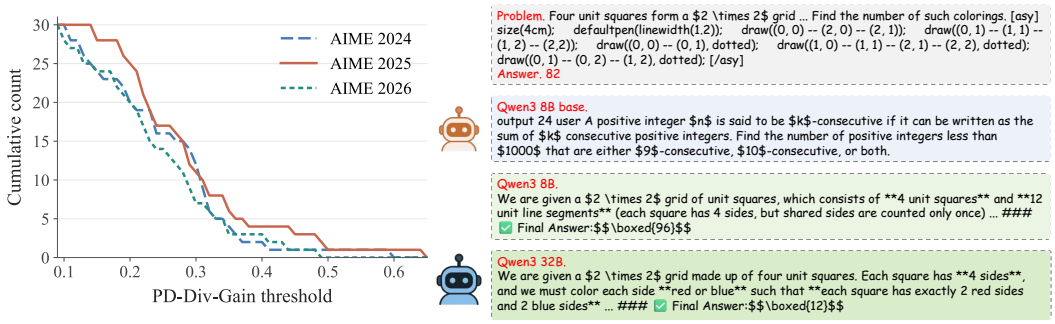

Figure 4: Reverse-cumulative PD-Div-Gain distributions on AIME 2024–2026 for Qwen3-8B. Lines connect counts of problems at or above each threshold on a common grid of width 0.01. Scores use the complete response-token trajectory.

Figure 4 presents the reverse-cumulative PD-Div-Gain distributions for AIME 2024–2026. The AIME 2025 curve lies noticeably above the AIME 2026 curve across much of the score range, indicating more problems with high alignment gains. AIME 2024 follows a broadly similar distribution to AIME 2026, but its curve is also noticeably higher at thresholds between 0.24 and 0.32. Relative to the temporal negative control, these elevated scores are consistent with possible data contamination through policy distillation, motivating further investigation.

Nevertheless, these observations do not establish that AIME 2025 was included in distillation training. First, each dataset contains only 30 problems, limiting statistical power; differences in problem difficulty or response distributions could also account for the observed shifts. Second, even if contamination occurred, it need not affect every problem in a dataset. The substantial overlap between the score distributions provides insufficient separation for reliable sample-level attribution: an elevated score alone cannot determine whether a particular problem was used in distillation. The figure therefore offers an exploratory group-level signal, while both dataset-level confirmation and individual-example detection remain unresolved. The right panel of Figure 4 compares responses from the initial student, trained student, and reference teacher to an AIME 2025 problem, illustrating that the student's response becomes more similar to the teacher's after training, although this similarity alone does not establish that the problem was exposed during training.

6 Conclusion

This work identifies and formalizes a blind spot in data contamination detection introduced by policy distillation. Unlike hard-label training, soft-target optimization can repeatedly expose a student to evaluation corpora without producing the unusually confident predictions targeted by conventional detectors. Under detector regularity, endpoint-margin, and path-length conditions, we show that the entire KL logit-flow trajectory remains on the non-member side, whereas hard-label cross-entropy crosses the detection boundary under an additional hard-target condition. Distillation-MIA experiments on four datasets complement this analysis: conventional detectors achieve near-chance average AUROC under both online and offline distillation despite known exposure. These findings establish a conditional failure mechanism, not universal undetectability, and show why a negative conventional audit cannot reliably certify that evaluation data were excluded from distillation.

Motivated by this analysis, we investigate teacher–student KL as a detection signal and derive a lower bound on its expected non-member-minus-member gap under shared-parameter training. PD-Div and PD-Div-Gain recover part of the signal in controlled experiments, but access to the actual teacher and, for calibration, the pre-distillation student limits their practical use. Our exploratory Qwen3 audit reveals AIME score patterns consistent with possible distillation exposure, without establishing contamination or identifying individual members. Despite demonstrating effectiveness in controlled experiments, there are still practical limitations to the current method, such as the need for teacher models and insufficient classification ability, which are important research goals for the future.

Ethics Statement

This work studies limitations of membership and contamination auditing. Experiments should use controlled or licensed corpora. A negative detector decision does not establish privacy, and evidence of distillation exposure does not by itself establish extractable memorization.

Reproducibility Statement

Notation and theorem conditions appear in Sections 3–5.2; complete proofs appear in Appendix A. Sections 4.2 and 5.4 report the controlled membership construction and detection results. Appendix C specifies the training protocol, model configurations, generation settings, and baseline definitions. The experimental analysis is kept distinct from the certified forward-KL statements proved in the main text and appendix.

AI Use Statement

In this work, we only used generative AI tools for polishing the language to improve readability. We have reviewed all AI-assisted work. All AI-assisted language edits were reviewed by the authors to ensure that they did not alter the technical meaning or scientific claims of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

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 *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *International conference on machine learning*, pp. 1964–1974. PMLR, 2021.

Ziyao Cui, Minxing Zhang, and Jian Pei. On membership inference attacks in knowledge distillation. *ACM SIGKDD Explorations Newsletter*, 28(1):32–40, 2026.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms. *arXiv preprint arXiv:2605.00674*, 2026.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12039–12050, 2024.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership inference attacks against fine-tuned large language models via self-prompt calibration. *Advances in Neural Information Processing Systems*, 37:134981–135010, 2024.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

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

Matthew Jagielski, Milad Nasr, Katherine Lee, Christopher A Choquette-Choo, Nicholas Carlini, and Florian Tramer. Students parrot their teachers: Membership inference on model distillation. *Advances in Neural Information Processing Systems*, 36:44382–44397, 2023.

Chenye Ke, Zirui Liu, Qi Liu, Yan Zhuang, Jintao Zhang, Zhenya Huang, and Shijin Wang. Detecting pretraining data in large language models from a free-energy perspective, 2026a. URL https://arxiv.org/abs/2609.21888.

Chenye Ke, Yan Zhuang, Zirui Liu, and Qi Liu. Rethinking pretraining data detection for LLMs: From local to global. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=ThlRUXqFZ0.

Zirui Liu, Xianquan Wang, Yan Zhuang, Jiatong Li, Qi Liu, Shuanghong Shen, Mingyue Cheng, and Shijin Wang. Fewer battles, more gain: An information-efficient framework for arena-based LLM evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=XUVqFRp9oi.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11330–11343, 2023.

Federico Mazzone, Leander Van Den Heuvel, Maximilian Huber, Cristian Verdecchia, Maarten Everts, Florian Hahn, and Andreas Peter. Repeated knowledge distillation with confidence masking to mitigate membership inference attacks. In *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*, pp. 13–24, 2022.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8332–8347, 2022.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1797–1807, 2018.

Virat Shejwalkar and Amir Houmansadr. Membership privacy for machine learning models through knowledge transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 9549–9557, 2021.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, volume 2024, pp. 51826–51843, 2024.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX security symposium (USENIX security 21)*, pp. 2615–2632, 2021.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in neural information processing systems*, 34: 6906–6919, 2021.

Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. Mitigating membership inference attacks by {Self-Distillation} through a novel ensemble architecture. In *31st USENIX security symposium (USENIX security 22)*, pp. 1433–1450, 2022.

Han Wang, Haoyu Li, Brian Ko, and Huan Zhang. On the fragility of benchmark contamination detection in reasoning models. In *International Conference on Learning Representations*, volume 2026, pp. 100419–100460, 2026.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, et al. Livebench: A challenging, contamination-limited llm benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 91595–91631, 2025.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. Recall: Membership inference via relative conditional log-likelihoods. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8671–8689, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Junjie Ye, Caishuang Huang, Zhuohan Chen, Wenjie Fu, Chenyuan Yang, Leyi Yang, Yilong Wu, Peng Wang, Meng Zhou, Xiaolong Yang, et al. Muldimif: A multi-dimensional constraint framework for evaluating and improving instruction following in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2078–2104, 2026.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Li. Min-k%++: Improved baseline for pre-training data detection from large language models. In *International Conference on Learning Representations*, volume 2025, pp. 64845–64862, 2025a.

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Pretraining data detection for large language models: A divergence-based calibration method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5263–5274, 2024.

Ziqi Zhang, Ali Shahin Shamsabadi, Hanxiao Lu, Yifeng Cai, and Hamed Haddadi. Membership and memorization in llm knowledge distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20085–20095, 2025b.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, et al. A survey of large language models. *Frontiers of Computer Science*, 20(12):2012627, 2026.

Junxiang Zheng, Yongzhi Cao, and Hanpin Wang. Resisting membership inference attacks through knowledge distillation. *Neurocomputing*, 452:114–126, 2021.

Yan Zhuang, Qi Liu, Zachary Pardos, Patrick C. Kyllonen, Jiyun Zu, Zhenya Huang, Shijin Wang, and Enhong Chen. Position: AI evaluation should learn from how we test humans. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=MxCJbuJhWG.

A Complete Proofs

The appendix follows the argument in the main text: first the failure of a fixed conventional detector, then descent of the teacher-aware score, and finally its expected training–validation gap. We also derive the Fisher–Rao reference path and establish the score regularity used in Section 3.

A.1 Logit-flow separation between KD and CE

We prove Theorem 1 for independent logits evaluated on a fixed sequence. Every statement about the two paths starts from the same strictly positive initial output.

Step 1: derive the probability flows. For one token, $\nabla_z \mathrm{KL}(q\|\operatorname{softmax}(z)) = p - q$ and $\nabla_z(-\log p_y) = p - e_y$. The derivative of softmax is $J(p) = \operatorname{diag}(p) - pp^\top$. Negative logit gradient flow therefore induces

$$\dot{p}^{\mathrm{KD}} = J(p^{\mathrm{KD}})(q - p^{\mathrm{KD}}), \qquad \dot{p}^{\mathrm{CE}} = J(p^{\mathrm{CE}})(e_y - p^{\mathrm{CE}}). \tag{20}$$

For the average token losses, a common factor $1/m$ is absorbed into u. Both logit vector fields are smooth and globally bounded, so finite initial logits yield solutions at every finite time and strictly positive outputs.

Step 2: establish convergence to the teacher and finite path length. For each position write $D_j(u) = \mathrm{KL}(q_j\|p_j^{\mathrm{KD}}(u))$. The logit gradient and its negative flow give

$$\dot{D}_j(u) = -\|p_j^{\mathrm{KD}}(u) - q_j\|_2^2 \le 0. \tag{21}$$

Let $H(q_j) = -\sum_v q_{j,v}\log q_{j,v}$. Because every term in the cross-entropy is nonnegative,

$$-q_{j,v}\log p_{j,v}^{\mathrm{KD}}(u) \le D_j(u) + H(q_j) \le D_j(0) + H(q_j). \tag{22}$$

Strict positivity of Q therefore supplies a uniform bound

$$p_{j,v}^{\mathrm{KD}}(u) \ge \exp\left(-\frac{D_j(0) + H(q_j)}{q_{j,v}}\right), \qquad \varepsilon := \min_{j,v} \exp\left(-\frac{D_j(0) + H(q_j)}{q_{j,v}}\right) > 0. \tag{23}$$

No additional interior assumption is needed. For a tangent vector a with $\sum_v a_v = 0$ and a distribution p with $p_v \ge \varepsilon$,

$$a^\top J(p)a = \min_{c\in\mathbb{R}} \sum_v p_v(a_v - c)^2 \ge \varepsilon \min_{c\in\mathbb{R}} \sum_v (a_v - c)^2 = \varepsilon\|a\|_2^2. \tag{24}$$

Also $0 \preceq J(p) \preceq I$, since its quadratic form is a variance bounded above by $\sum_v p_v a_v^2 \le \|a\|_2^2$. Writing $r_j = p_j^{\mathrm{KD}} - q_j$, whose coordinates sum to zero, gives

$$\frac{d}{du}\frac{1}{2}\|P^{\mathrm{KD}}(u) - Q\|_2^2 = -\sum_j r_j^\top J(p_j^{\mathrm{KD}})r_j \le -\varepsilon\|P^{\mathrm{KD}}(u) - Q\|_2^2. \tag{25}$$

Consequently,

$$\|P^{\mathrm{KD}}(u) - Q\|_2 \le e^{-\varepsilon u}\|P_0 - Q\|_2, \qquad \|\dot{P}^{\mathrm{KD}}(u)\|_2 \le \|P^{\mathrm{KD}}(u) - Q\|_2. \tag{26}$$

Thus $P^{\mathrm{KD}}(u) \to Q$, and its total length is finite:

$$\int_0^\infty \|\dot{P}^{\mathrm{KD}}(u)\|_2\, du \le \frac{\|P_0 - Q\|_2}{\varepsilon} < \infty. \tag{27}$$

This estimate establishes finiteness only. It is not substituted into the theorem's path-length condition, which concerns the actual traveled length. The convergence is asymptotic; it does not assert exact teacher matching at a finite checkpoint. If $P_0 = Q$, the KD trajectory is constant.

Step 3: use both endpoint margins to exclude a crossing. Suppose some finite time u satisfies $s(P^{\mathrm{KD}}(u)) \geq \tau$. The initial margin and Lipschitz continuity imply

$$\rho_0 \leq s(P^{\mathrm{KD}}(u)) - s(P_0) \leq L_s \|P^{\mathrm{KD}}(u) - P_0\|_2 \leq L_s \int_0^u \|\dot{P}^{\mathrm{KD}}(v)\|_2 \, dv. \tag{28}$$

The teacher margin similarly gives

$$\rho_Q \leq s(P^{\mathrm{KD}}(u)) - s(Q) \leq L_s \|P^{\mathrm{KD}}(u) - Q\|_2 \leq L_s \int_u^\infty \|\dot{P}^{\mathrm{KD}}(v)\|_2 \, dv. \tag{29}$$

The last inequality follows by integrating the velocity to a finite terminal time and passing to the limit established in Step 2. Adding the inequalities contradicts Equation (7). Therefore no crossing occurs. This argument uses only the Lipschitz property and the length of a convergent curve; it does not assume a straight path or a convex non-member region.

Step 4: prove that CE approaches the hard target. Fix one token with correct label y and put $a(u) = p_y^{\mathrm{CE}}(u)$. Direct substitution into the CE probability flow yields

$$\dot{a}(u) = a(u) \left((1 - a(u))^2 + \sum_{i \neq y} (p_i^{\mathrm{CE}}(u))^2 \right) \geq a(u)(1 - a(u))^2. \tag{30}$$

Thus $a(u)$ is increasing and bounded above by one, so it has a limit $a_\infty \leq 1$. If $a_\infty < 1$, then $\dot{a}(u) \geq a(0)(1 - a_\infty)^2 > 0$ for all u, contradicting boundedness. Therefore $a(u) \to 1$. All other coordinates tend to zero, and applying this argument to each position gives $P^{\mathrm{CE}}(u) \to E$.

Step 5: obtain finite-time classification separation. The endpoint continuity condition gives $s(P^{\mathrm{CE}}(u)) \to s(E) > \tau$. Choose $b = (s(E) - \tau)/2 > 0$. For all sufficiently large finite u, $|s(P^{\mathrm{CE}}(u)) - s(E)| < b$, hence $s(P^{\mathrm{CE}}(u)) > \tau$. Choose $u_\star > 0$ in this range. The KD decision remains zero by Step 3, which proves the theorem. Optionally reindexing progress by $\alpha = 1 - e^{-u}$ gives a crossing value $\alpha_\star = 1 - e^{-u_\star} \in (0, 1)$. This is only a change of time parameter, not a straight-path representation of logit flow.

A.2 KL SCORE DESCENT UNDER LOGIT OPTIMIZATION

Because $\partial k / \partial z_j = (p_j - q_j)/m$ and $\dot{z}_j = q_j - p_j$, the chain rule gives

$$\frac{d}{du} d_{\mathrm{KL}}(P^{\mathrm{KD}}(u), Q) = \frac{1}{m} \sum_j (p_j - q_j)^\top (q_j - p_j) = -\frac{1}{m} \sum_j \|p_j - q_j\|_2^2. \tag{31}$$

This is a fixed-sequence optimization identity; training–validation separation requires the following dataset-level argument.

A.3 EXPECTED TRAINING–VALIDATION KL LOWER BOUND

We prove Theorem 2, retaining the dependence of the trained model on the dataset throughout.

Step 1: condition on the training sample. Given D, the gradient-flow output θ_t is fixed. Every validation document is independent of D, so

$$\mathbb{E}\left[\frac{1}{M} \sum_a k_{\theta_t}(Z_a) \,\middle|\, D \right] = R(\theta_t). \tag{32}$$

For the threshold interpretation, let $X_{\mathrm{m}} = X_I$ and $X_{\mathrm{n}} = Z_A$, where I and A are independent uniform indices in the training and validation sets, independent of those datasets. Then

$$\mathbb{E}[k_{\theta_t}(X_{\mathrm{m}}) \mid D] = \widehat{R}_D(\theta_t), \tag{33}$$

$$\mathbb{E}[k_{\theta_t}(X_{\mathrm{n}}) \mid D] = R(\theta_t). \tag{34}$$

Taking expectation proves Equation (13). Only the independent validation draw is independent of the trained model; training members are not treated as independent of it.

Step 2: separate population improvement and empirical fitting. Because θ_0 and the teacher are fixed before drawing D, $\mathbb{E}\widehat{R}_D(\theta_0) = R(\theta_0)$. Hence

$$\Gamma_t = \mathbb{E}[R(\theta_t) - R(\theta_0)] + \mathbb{E}[\widehat{R}_D(\theta_0) - \widehat{R}_D(\theta_t)]. \tag{35}$$

The population term may be negative: validation KL can improve. But $\theta_t \in \Theta$ implies

$$R(\theta_t) - R(\theta_0) \geq R_{\inf} - R(\theta_0) = -\Delta_0. \tag{36}$$

No population minimizer needs to exist for this step.

Step 3: quantify finite-time empirical fitting. Define $g_D(t) = \nabla \widehat{R}_D(\theta_t)$. The chain rule along the empirical parameter flow gives

$$\frac{d}{dt}\widehat{R}_D(\theta_t) = -\|g_D(t)\|_2^2, \qquad \frac{d}{dt}g_D(t) = -\nabla^2 \widehat{R}_D(\theta_t) g_D(t). \tag{37}$$

Therefore the upper Hessian bound gives

$$\frac{d}{dt}\|g_D(t)\|_2^2 = -2g_D(t)^\top \nabla^2 \widehat{R}_D(\theta_t) g_D(t) \geq -2L_k \|g_D(t)\|_2^2. \tag{38}$$

Equivalently, $\frac{d}{dt}[e^{2L_k t}\|g_D(t)\|_2^2] \geq 0$, so $\|g_D(t)\|_2^2 \geq e^{-2L_k t}\|g_D(0)\|_2^2$. Integrating yields

$$\widehat{R}_D(\theta_0) - \widehat{R}_D(\theta_t) = \int_0^t \|g_D(v)\|_2^2 \, dv \geq \frac{1 - e^{-2L_k t}}{2L_k}\|g_D(0)\|_2^2. \tag{39}$$

This is a property of the actual finite-time trajectory, not a comparison with an empirical minimizer. Negative Hessian eigenvalues do not invalidate it.

Step 4: compute the initial empirical gradient energy. Let $a_i = \nabla k_{\theta_0}(X_i) - \mu_0$. These vectors are independent, centered, and have $\mathbb{E}\|a_i\|_2^2 = \sigma_0^2$. Since $g_D(0) = \mu_0 + N^{-1}\sum_i a_i$,

$$\mathbb{E}\|g_D(0)\|_2^2 = \|\mu_0\|_2^2 + \frac{1}{N^2}\sum_i \mathbb{E}\|a_i\|_2^2 \tag{40}$$

$$= \|\mu_0\|_2^2 + \frac{\sigma_0^2}{N}. \tag{41}$$

Cross terms vanish by independence and centering. Token independence within a document is never required.

Step 5: combine the bounds. Substitute Equations (36), (39), and (41) into Equation (35) to obtain

$$\Gamma_t \geq \frac{1 - e^{-2L_k t}}{2L_k}\left(\|\mu_0\|_2^2 + \frac{\sigma_0^2}{N}\right) - \Delta_0. \tag{42}$$

At $t = 0$ the true gap is exactly zero; the bound $-\Delta_0$ can be loose. If the right-hand side is nonpositive, this result alone does not decide the sign of the true gap.

Step 6: derive a threshold with positive advantage. Put $Y_{\mathrm{m}} = k_{\theta_t}(X_{\mathrm{m}})$ and $Y_{\mathrm{n}} = k_{\theta_t}(X_{\mathrm{n}})$. Both are nonnegative and integrable under the theorem's conditions. Let $C_{\mathrm{m}}(a) = \Pr(Y_{\mathrm{m}} \leq a)$ and $C_{\mathrm{n}}(a) = \Pr(Y_{\mathrm{n}} \leq a)$, with probability over the full membership experiment. The tail-integral identity gives

$$\Gamma_t = \int_0^\infty \big(C_{\mathrm{m}}(a) - C_{\mathrm{n}}(a)\big)\, da. \tag{43}$$

If $\Gamma_t \geq B_t > 0$, the integrand must be positive for some $a > 0$. Choosing such an a as γ gives

$$\Pr\{k_{\theta_t}(X_{\mathrm{m}}) \leq \gamma\} > \Pr\{k_{\theta_t}(X_{\mathrm{n}}) \leq \gamma\}. \tag{44}$$

This establishes existence in the averaged experiment, not a lower bound at an arbitrary preselected threshold.

Population-optimal initialization. For the special case in Theorem 2, interior population optimality gives $\mu_0 = 0$ and $\Delta_0 = 0$. Substituting these into Equation (17) yields

$$\Gamma_t \geq \frac{1 - e^{-2L_k t}}{2L_k} \frac{\sigma_0^2}{N},$$

which is positive for $t > 0$ if $\sigma_0^2 > 0$. Capacity restrictions can leave this residual gradient variation even at population optimality. In contrast, an exactly matching teacher and student have zero KL and zero sample gradients, consistent with the absence of a strictly positive universal gap.

B Alternative Information-Geometric Analysis

The main text uses the probability dynamics induced by Euclidean logit optimization. Here we consider a different metric, for which the path is straight in probability coordinates. Its total length equals the distance between its endpoints, so the main text's path-length certificate reduces exactly to a two-endpoint distance condition.

For the product of categorical simplices, define the Fisher–Rao metric

$$g_P(U, W) = \frac{1}{m} \sum_{j,v} \frac{U_{j,v} W_{j,v}}{p_{j,v}}, \qquad \sum_v U_{j,v} = \sum_v W_{j,v} = 0. \tag{45}$$

Write $\mathcal{S}_x = [P_0, Q]$ for the mixture segment and $d_x = \|P_0 - Q\|_2$ for its length.

Theorem 3 (Fisher–Rao non-detection and hard-label separation)**.** *For fixed strictly positive P_0, Q, suppose s is L_s-Lipschitz on $\mathcal{S}_x$ and*

$$\rho_0 > 0, \qquad \rho_Q > 0, \qquad L_s d_x < \rho_0 + \rho_Q. \tag{46}$$

Fisher–Rao gradient flow of $\mathcal{L}_{\mathrm{KD}}$ satisfies

$$P_{\mathrm{FR}}^{\mathrm{KD}}(u) = e^{-u} P_0 + (1 - e^{-u}) Q, \qquad F_\tau(P_{\mathrm{FR}}^{\mathrm{KD}}(u)) = 0 \quad (u \geq 0). \tag{47}$$

The corresponding CE flow is $P_{\mathrm{FR}}^{\mathrm{CE}}(u) = e^{-u} P_0 + (1 - e^{-u}) E$. If s has a continuous extension to E along this CE path and $s(E) > \tau$, the CE detector becomes positive after finite time.

Natural-gradient calculation. For one token let $h(p) = \mathrm{KL}(q\|p)$, and write $g_p(v, w) = \sum_i v_i w_i / p_i$ for the unaveraged single-token metric. For a tangent vector v with $\sum_i v_i = 0$,

$$dh_p[v] = -\sum_i \frac{q_i}{p_i} v_i = \sum_i \frac{(p_i - q_i) v_i}{p_i} = g_p(p - q, v). \tag{48}$$

The vector $p - q$ is tangent, so it is the natural gradient. Averaging both the objective and the product metric preserves this result. Solving $\dot{p} = q - p$ gives $p(u) = q + e^{-u}(p_0 - q)$. Replacing q by the one-hot target gives the stated CE solution.

Path length and the two-endpoint certificate. The explicit solution gives

$$\int_0^\infty \|\dot{P}_{\mathrm{FR}}^{\mathrm{KD}}(u)\|_2 \, du = \int_0^\infty e^{-u} \|Q - P_0\|_2 \, du = d_x. \tag{49}$$

Equation (46) is therefore exactly the path-length condition. The geometric argument in Step 3 of Appendix A.1 applies unchanged, proving non-detection along the FR path. Equivalently, set $\alpha = 1 - e^{-u}$ and $P_\alpha = (1 - \alpha) P_0 + \alpha Q$. Then

$$\|P_\alpha - P_0\|_2 = \alpha d_x, \qquad \|P_\alpha - Q\|_2 = (1 - \alpha) d_x. \tag{50}$$

If $s(P_\alpha) \geq \tau$, Lipschitz continuity would require $\rho_0 \leq L_s \alpha d_x$ and $\rho_Q \leq L_s (1 - \alpha) d_x$, contradicting their sum. In particular, no path-length inflation factor is introduced for FR dynamics. The CE part follows from its convergence to E and continuity of the score, as in Step 5 of Appendix A.1.

Discrete probability updates. The Euler update $p^+ = (1 - \eta) p + \eta q$ with $0 \leq \eta \leq 1$ remains on the mixture segment. After n constant-size steps,

$$p_n = (1 - \eta)^n p_0 + \big(1 - (1 - \eta)^n\big) q. \tag{51}$$

This identity covers a single probability-space update, but is not an identity for a discrete logit update.

KL contraction on the mixture segment. Convexity in the student distribution yields

$$d_{\mathrm{KL}}(P_\alpha, Q) \le (1-\alpha) d_{\mathrm{KL}}(P_0, Q). \tag{52}$$

For $0 \le \alpha_1 < \alpha_2 \le 1$, $P_{\alpha_2} = cP_{\alpha_1} + (1-c)Q$ with $c = (1-\alpha_2)/(1-\alpha_1) \in [0,1]$. Applying the same bound proves monotonicity between any two progress values.

For the experimental JSD variant, let $d_{\mathrm{JS}}(P,Q) = m^{-1}\sum_j \mathrm{JSD}(p_j, q_j)$, where $\mathrm{JSD}(p,q) = \frac{1}{2}\mathrm{KL}(p\|M) + \frac{1}{2}\mathrm{KL}(q\|M)$ and $M = (p+q)/2$. Joint convexity of KL gives the same mixture-path contraction for JSD. This does not establish a JSD training–validation gap for forward-KL parameter training.

B.1 Regularity of conventional scores

Write $r_j = p_j(x_j \mid x_{<j})$. On the convex domain $r_j \ge \varepsilon > 0$, average log-probability is

$$s_{\mathrm{avg}}(r) = \frac{1}{m}\sum_j \log r_j, \qquad \|\nabla s_{\mathrm{avg}}(r)\|_2 \le \frac{1}{\varepsilon\sqrt{m}}. \tag{53}$$

Integrating the gradient along a segment proves this Lipschitz bound. The coordinate selection $P \mapsto r$ is nonexpansive for the flattened Euclidean norm, so the same constant applies to the full output tuple.

For $1 \le k \le m$, the Min-K% score with k selected positions is

$$s_{\min k}(r) = \min_{\substack{A \subseteq \{1,\dots,m\} \\ |A|=k}} \frac{1}{k}\sum_{j\in A} \log r_j. \tag{54}$$

Each fixed-subset score has gradient norm at most $1/(\varepsilon\sqrt{k})$. Since $|\min_i f_i(r) - \min_i f_i(r')| \le \max_i |f_i(r) - f_i(r')|$, the same constant holds through subset changes and ties.

Both average log-probability and Min-K% Prob extend continuously to the hard-label endpoint E, where all observed-token probabilities equal one and their scores equal zero. Thus their CE endpoint condition in Theorem 1 holds for thresholds $\tau < 0$.

Shannon entropy has gradient components $-(1+\log p_i)$ and is Lipschitz on a convex domain $p_i \ge \varepsilon$. Label confidence is linear; maximum confidence is 1-Lipschitz in Euclidean norm. For a feature score with fixed coefficient vectors a, b and intercept c,

$$f(p) = \sum_i a_i \log p_i + \sum_i b_i p_i \log p_i - c, \qquad \frac{\partial f}{\partial p_i} = \frac{a_i}{p_i} + b_i(1+\log p_i), \tag{55}$$

the same interior restriction bounds the gradient. A combination of component scores is Lipschitz with constant no larger than the sum of their constants weighted by the absolute coefficients. These results establish Equation (3) for the listed scores, not for every confidence-based detector. Normalized or ratio-based variants require separate control of their denominators. These arguments require a region containing the relevant paths. For the independent-logit KD flow with a fixed positive teacher, a uniform interior bound follows from the objective itself. For each position, let $D_j(0) = \mathrm{KL}(q_j\|p_j(0))$ and $H(q_j) = -\sum_v q_{j,v}\log q_{j,v}$. Tokenwise KL descent gives

$$-q_{j,v}\log p_{j,v}(u) \le -\sum_w q_{j,w}\log p_{j,w}(u) \le D_j(0) + H(q_j). \tag{56}$$

Consequently every student coordinate satisfies

$$p_{j,v}(u) \ge \exp\left(-\frac{D_j(0)+H(q_j)}{q_{j,v}}\right) > 0 \qquad (u \ge 0). \tag{57}$$

The same coordinatewise bound holds for Q, since $-q_{j,v}\log q_{j,v} \le H(q_j)$. Taking the minimum over coordinates supplies an ε for a convex domain containing both endpoints and the entire KD path. Thus these scores have a finite uniform Lipschitz constant in this model. The bound can be very conservative for small teacher probabilities; finiteness alone does not ensure that the margin condition in Theorem 1 holds.

C Experimental Details

In this section, we provide details of the Distillation-MIA benchmark construction, model and training configurations, distillation procedures, and detection protocols.

C.1 Distillation-MIA: A Benchmark for Controlled Distillation Exposure

Datasets and membership construction. We construct Distillation-MIA using MATH competition problems (Hendrycks et al., 2021), MulDimIF multi-constraint instruction following (Ye et al., 2026), XSum single-sentence news summarization (Narayan et al., 2018), and AIME 2026 mathematical reasoning problems (Dekoninck et al., 2026), covering diverse task types.

Given computational constraints, we construct a distillation training corpus of 500 examples for each of MATH, MulDimIF, and XSum, consisting of 100 designated member examples and 400 background training examples. We additionally hold out 100 candidate examples as non-members. Thus, members (label $= 1$) participate in distillation together with other training examples, whereas non-members (label $= 0$) are never used for training. This construction captures a controlled contamination setting in which target examples are mixed into a broader post-training corpus rather than trained in isolation. For AIME 2026, we use all 30 problems, with 15 serving as members and the remaining 15 as non-members. In future work, we will scale the benchmark to substantially larger post-training corpora that more closely reflect real-world data contamination scenarios.

To reduce dependence on a particular candidate split, we construct two disjoint candidate groups, A and B, and train separate students S_A and S_B. For MATH, MulDimIF, and XSum, each group contains 100 candidate examples; for AIME 2026, each contains 15 problems. For S_A, group A serves as members and group B as non-members, whereas the assignments are reversed for S_B. Thus, the same candidate groups appear in both membership roles across independently trained students, reducing confounding from sample-specific characteristics.

C.2 Experimental Setup

Models and shared training setup. The initial student S_0 is Qwen2.5-1.5B-Instruct, and the frozen teacher T is Qwen2.5-7B-Instruct (Yang et al., 2024). Each post-training student $S_1 \in \{S_A, S_B\}$ is represented by a rank-16 LoRA adapter ($\alpha = 32$, dropout 0.05) applied to the query, key, value, output, gate, up, and down projections. Within each distillation setting, all eight adapters start from the same seed-2026 initialization. The data splits, model pair, LoRA configuration, optimization schedule, and completion-only JSD objective are held fixed across the two distillation settings.

Implementation details. Students are trained with no LoRA bias, and only the adapter parameters are trainable. Every run uses seed 2026. Both distillation settings use AdamW with learning rate 10^{-4}, zero weight decay, linear warmup ratio 0.05, gradient clipping at 1.0, micro-batch size 1, and gradient accumulation over 8 micro-steps. Training runs for three epochs. The context window is 2,048 tokens, and the per-prompt generation budget is

$$\min\{L_{\text{dataset}},\, 2048 - L_{\text{prompt}} - 1\}, \tag{58}$$

where $L_{\text{dataset}} = 1{,}024$ for MATH, MulDimIF, and AIME 2026 and $L_{\text{dataset}} = 128$ for XSum. Generation stops early at EOS. Training rollouts use temperature 0.5 and top-p $= 0.9$ in both distillation settings. On-policy training samples a fresh student completion at each prompt visit, whereas off-policy training samples one teacher completion per prompt before optimization and reuses it across epochs.

On-policy distillation. Training follows on-policy GKD (Agarwal et al., 2024). For each prompt, the current student samples one completion and is optimized only at completion positions to match the teacher. At a prefix $z_t = (x, y_{<t})$, let $p_t^S = S_1(\cdot \mid z_t)$ and $p_t^T = T(\cdot \mid z_t)$. The practical loss is

$$m_t = \tfrac{1}{2}(p_t^S + p_t^T), \tag{59}$$

$$\mathcal{L}_{\text{OPD}}(x) = \frac{1}{L}\sum_{t=1}^{L}\left[\tfrac{1}{2}\mathrm{KL}(p_t^S \| m_t) + \tfrac{1}{2}\mathrm{KL}(p_t^T \| m_t)\right]. \tag{60}$$

The divergence is computed over the complete shared vocabulary. No reference answer, supervised completion, sequence-level KD objective, or prompt-token loss is used.

Off-policy distillation. Before optimization, the frozen teacher samples one completion $y^T = (y_1^T, \ldots, y_L^T)$ for each training prompt. This trajectory is cached and reused in every epoch. The student never generates a training response in this setting. At $z_t^T = (x, y_{<t}^T)$, the objective is

$$\mathcal{L}_{\text{off}}(x) = \frac{1}{L} \sum_{t=1}^{L} \text{JSD}\big(S_1(\cdot \mid z_t^T), T(\cdot \mid z_t^T)\big) . \tag{61}$$

Teacher logits are recomputed on the fixed prefixes rather than stored, which does not alter the off-policy state distribution. Thus, the two settings differ only in whether prefixes come from the evolving student or a fixed teacher trajectory.

Fixed detection trajectories. For each candidate prompt, its trained student generates $K = 4$ stochastic completions with temperature 0.5 and top-$p = 0.9$. All detectors reuse these cached S_1 trajectories. The prompt is conditioning context only: every reported score averages completion-token prediction positions, first within a rollout and then across the four rollouts. Neither S_0 nor T regenerates a response.

Baseline methods. We evaluate eight conventional detection methods on the same cached student completions. **Loss** uses the average log-likelihood of the generated completion as the membership signal (Carlini et al., 2021). **Zlib** normalizes the completion likelihood using its zlib compression length, while **Lowercase** calibrates likelihood using a lowercased version of the completion (Carlini et al., 2021). **Reference** calibrates the target student's likelihood using a smaller reference model from the same model family (Carlini et al., 2021); in our experiments, we use Qwen2.5-0.5B-Instruct as the reference model. **Min-K%** selects the $k\%$ of completion tokens with the lowest log probabilities and averages their log probabilities (Shi et al., 2024), whereas **Min-K%++** further standardizes token log probabilities using vocabulary-level statistics of the target model (Zhang et al., 2025a). Following prior work, we use the default selected-token ratio of $k = 20\%$ for both methods. **ETD** combines token prediction loss with predictive entropy to construct an entropy-adjusted detection score from a free-energy perspective (Ke et al., 2026a). **AECA** treats the token-probability sequence as a dynamic signal, calibrates token probabilities using corpus-based self-information, and applies convolutional filtering to capture global volatility patterns associated with memorization (Ke et al., 2026b). Each method is first computed per rollout and then averaged over $K = 4$.

Evaluation metrics. We report area under the receiver operating characteristic curve (AUROC) and true positive rate (TPR) at low false-positive rates (FPRs). AUROC measures ranking quality over all thresholds. For MATH, MulDimIF, and XSum, we report TPR@5%FPR. Since AIME 2026 contains only 15 non-members per student, we instead report TPR@2FP, corresponding to TPR@13.3%FPR. For each dataset–setting combination, we average the results of S_A and S_B.

C.3 Audit Details and Theoretical Scope

Real-world Qwen3 audit. The exploratory case study in Figure 4 uses Qwen3-8B-Base (S_0), Qwen3-8B (S_1), and Qwen3-32B (reference T). The same saved S_1 prompt and greedy completion token IDs are evaluated by all three models; the prompt provides context but contributes no score tokens. Generation uses Qwen3's non-thinking mode with a 4,096-token response cap. For each problem, S_1 generates a single greedy completion, and the full-vocabulary forward-KL difference is averaged over all completion positions. Each AIME set contains all 30 problems. The score medians are 0.274, 0.277, and 0.228 for AIME 2024, AIME 2025, and AIME 2026, respectively.

Relation to the theoretical sampling model. The theory-aligned protocol draws the student's corpus and independent validation documents from the same population, with teacher and initialization fixed independently of that draw. A difficulty-matched holdout is an additional stress test rather than a replacement for this sampling model. Distillation-MIA supplies known prompt membership but does not certify this complete sampling model: in particular, its AIME groups preserve source order. The cross-dataset Qwen3 audit is not a training–validation draw of the kind assumed in Theorem 2.

Interpreting numerical path diagnostics. Checkpoint distances lower-bound the traveled arc length; a short sampled path does not certify Equation (7). Certification requires upper bounds on between-checkpoint motion and the unobserved tail. Sampled score gradients and Hessian eigenvalues are likewise estimates, not automatically valid upper bounds. Certified coverage must distinguish verified bounds from estimates, and $R_{\inf}$ requires separate analysis rather than substitution of observed training loss.

D Detailed Experimental Results

In this section, we report detailed results separately for the two student models. Tables 5–6 present the SFT-training results, while Tables 7–10 present the distillation-training results.

D.1 SFT Training

Table 5: AUROC (%) results of the eight conventional detection methods on Distillation-MIA for students S_A and S_B under SFT. AVG averages all eight dataset–student combinations.

	$\mathbf{S_A}$				$\mathbf{S_B}$				
Method	MATH	MulDimIF	XSum	AIME26	MATH	MulDimIF	XSum	AIME26	**AVG**
Loss	59.3	91.7	97.9	84.4	64.7	89.9	98.2	72.4	82.3
Zlib	57.1	93.3	98.3	84.0	64.6	91.1	99.2	74.2	82.7
Reference	75.4	58.2	96.0	61.8	57.8	61.5	97.7	60.0	71.0
Lowercase	66.5	72.9	93.5	80.0	55.1	74.8	94.9	76.4	76.8
Min-K%	61.9	93.1	98.3	85.3	66.6	90.7	98.6	72.9	83.4
Min-K++	71.3	92.4	98.4	93.8	71.0	86.3	98.4	84.9	87.0
ETD	67.4	93.7	99.0	96.4	64.1	91.0	99.1	92.4	87.9
AECA	61.4	80.4	96.2	68.4	61.8	75.7	96.5	64.4	75.6

Table 6: Low-FPR detection results (%) of the eight conventional detection methods on Distillation-MIA for students S_A and S_B under SFT. MATH, MulDimIF, and XSum use TPR@5%FPR; AIME26 uses TPR@2FP (13.3% FPR). AVG averages all eight dataset–student values.

	$\mathbf{S_A}$				$\mathbf{S_B}$				
Method	MATH	MulDimIF	XSum	AIME26	MATH	MulDimIF	XSum	AIME26	**AVG**
Loss	13.0	56.0	90.0	73.3	6.0	49.0	89.0	40.0	52.0
Zlib	11.0	60.0	92.0	66.7	6.0	58.0	98.0	40.0	54.0
Reference	37.0	19.0	76.0	40.0	8.0	4.0	89.0	26.7	37.5
Lowercase	12.0	37.0	82.0	73.3	13.0	34.0	80.0	66.7	49.8
Min-K%	14.0	58.0	92.0	73.3	6.0	52.0	92.0	40.0	53.4
Min-K++	17.0	54.0	91.0	100.0	8.0	0.0	93.0	60.0	52.9
ETD	24.0	69.0	93.0	100.0	3.0	34.0	94.0	80.0	62.1
AECA	17.0	23.0	83.0	33.3	12.0	3.0	84.0	26.7	35.2

D.2 DISTILLATION TRAINING

Table 7: AUROC (%) results of the evaluated detection methods on Distillation-MIA for students S_A and S_B under online distillation. AVG averages all eight dataset–student combinations. Best results are in **bold**; second-best results are underlined.

	$\mathbf{S_A}$				$\mathbf{S_B}$				
Method	MATH	MulDimIF	XSum	AIME26	MATH	MulDimIF	XSum	AIME26	**AVG**
Loss	56.4	56.0	60.9	53.8	48.3	48.5	51.9	53.3	53.6
Zlib	53.9	56.5	59.0	54.7	49.5	49.7	52.0	52.4	53.5
Reference	47.5	49.5	49.4	43.6	50.3	49.4	47.6	46.7	48.0
Lowercase	55.8	54.1	57.6	56.4	49.5	47.0	55.5	61.8	54.7
Min-K%	56.5	54.1	61.2	52.4	48.3	46.7	48.6	55.6	52.9
Min-K++	52.5	53.9	56.1	58.2	51.6	51.5	48.6	49.8	52.8
ETD	54.6	53.2	58.4	57.3	48.7	48.1	48.2	57.8	53.3
AECA	53.8	53.8	56.9	44.4	50.6	48.1	48.4	59.1	51.9
PD-Div	57.9	**58.4**	71.9	55.6	52.4	54.5	65.8	58.2	59.3
PD-Div-Gain	**63.4**	53.2	**82.8**	62.7	**62.6**	**66.1**	**90.2**	**62.2**	**67.9**

Table 8: Low-FPR detection results (%) of the evaluated detection methods on Distillation-MIA for students S_A and S_B under online distillation. MATH, MulDimIF, and XSum use TPR@5%FPR; AIME26 uses TPR@2FP (13.3% FPR). AVG averages all eight dataset–student values. Best results are in **bold**; second-best results are underlined.

	$\mathbf{S_A}$				$\mathbf{S_B}$				
Method	MATH	MulDimIF	XSum	AIME26	MATH	MulDimIF	XSum	AIME26	**AVG**
Loss	2.0	12.0	7.0	33.3	10.0	7.0	7.0	20.0	12.3
Zlib	8.0	14.0	6.0	40.0	7.0	8.0	6.0	6.7	12.0
Reference	7.0	15.0	4.0	20.0	3.0	1.0	4.0	6.7	7.6
Lowercase	1.0	**17.0**	2.0	**46.7**	11.0	5.0	11.0	26.7	15.0
Min-K%	2.0	11.0	9.0	33.3	10.0	7.0	10.0	20.0	12.8
Min-K++	3.0	13.0	5.0	**46.7**	5.0	1.0	8.0	6.7	11.0
ETD	5.0	12.0	9.0	33.3	7.0	1.0	8.0	33.3	13.6
AECA	5.0	13.0	17.0	26.7	6.0	3.0	10.0	20.0	12.6
PD-Div	5.0	9.0	20.0	**46.7**	14.0	9.0	19.0	20.0	17.8
PD-Div-Gain	**17.0**	12.0	**43.0**	**46.7**	**15.0**	**12.0**	**54.0**	**40.0**	**30.0**

Table 9: AUROC (%) results of the evaluated detection methods on Distillation-MIA for students S_A and S_B under offline distillation. AVG averages all eight dataset–student combinations. Best results are in **bold**; second-best results are underlined.

	$\mathbf{S_A}$				$\mathbf{S_B}$				
Method	MATH	MulDimIF	XSum	AIME26	MATH	MulDimIF	XSum	AIME26	**AVG**
Loss	56.8	54.1	55.5	51.6	47.1	45.5	53.3	44.9	51.1
Zlib	53.4	56.1	54.1	47.6	49.1	47.1	55.5	56.4	52.4
Reference	46.5	51.4	54.6	48.4	52.4	51.0	50.7	50.2	50.7
Lowercase	57.6	53.1	52.4	49.8	46.1	45.5	58.7	56.9	52.5
Min-K%	57.0	54.3	55.1	51.6	46.9	44.8	51.2	46.7	50.9
Min-K++	52.1	52.0	51.4	55.1	48.3	49.4	47.8	52.9	51.1
ETD	57.3	51.7	52.2	54.2	46.0	47.5	52.0	44.9	50.7
AECA	52.5	52.9	56.2	44.0	49.7	46.7	51.2	55.1	51.0
PD-Div	58.1	**58.8**	68.4	44.9	51.8	51.2	62.0	57.8	56.6
PD-Div-Gain	**66.2**	52.0	**77.9**	**61.3**	**62.1**	**66.1**	**83.5**	50.2	**64.9**

Table 10: Low-FPR detection results (%) of the evaluated detection methods on Distillation-MIA for students S_A and S_B under offline distillation. MATH, MulDimIF, and XSum use TPR@5%FPR; AIME26 uses TPR@2FP (13.3% FPR). AVG averages all eight dataset–student values. Best results are in **bold**; second-best results are underlined.

	$\mathbf{S_A}$				$\mathbf{S_B}$				
Method	MATH	MulDimIF	XSum	AIME26	MATH	MulDimIF	XSum	AIME26	**AVG**
Loss	4.0	8.0	4.0	13.3	5.0	5.0	13.0	13.3	8.2
Zlib	12.0	10.0	6.0	33.3	1.0	4.0	15.0	13.3	11.8
Reference	10.0	15.0	5.0	33.3	3.0	0.0	8.0	0.0	9.3
Lowercase	0.0	12.0	4.0	20.0	7.0	3.0	18.0	20.0	10.5
Min-K%	4.0	10.0	6.0	13.3	5.0	4.0	15.0	13.3	8.8
Min-K++	5.0	**18.0**	5.0	26.7	2.0	0.0	16.0	13.3	10.8
ETD	5.0	17.0	5.0	26.7	4.0	0.0	16.0	20.0	11.7
AECA	3.0	**18.0**	6.0	6.7	5.0	2.0	5.0	20.0	8.2
PD-Div	2.0	14.0	11.0	20.0	**17.0**	9.0	19.0	20.0	14.0
PD-Div-Gain	**23.0**	**18.0**	**33.0**	**40.0**	14.0	**13.0**	**47.0**	13.3	**25.2**