COVER THE TEACHER, STAY CLOSE TO THE STUDENT: MODE-COVERING TRUST REGIONS FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation should preserve teacher-supported alternatives at student-generated prefixes, yet its usual reverse-KL signal is mode-seeking: alternatives assigned little probability by a capacity-limited student receive weak supervision and can disappear. Proximal Target Distillation (PTD) makes mode covering explicit. It projects the teacher onto a student-relative KL ball, preserving teacher-supported modes while bounding target movement, then fits that target with forward KL to train underrepresented tokens directly. On the normalized teacher-top-K support, the projection follows a closed-form normalized Lambert-W path whose radius bounds the initial conditional-logit gradient. In a misspecified Gaussian study, repeated PTD updates recover substantial probability in a teacher region initially nearly absent from the student. Controls identify forward fitting as the recovery mechanism and show that proximal target geometry controls the recovery trajectory. Across math-only and teacher-routed multi-domain language-model experiments, PTD delivers the highest Macro best-of-n point estimate among trained student methods. Its advantage over top-K reverse KL widens with the sampling budget and overtakes linear interpolation at the larger budget. The controlled density recovery and the growing multi-sample advantage jointly show how a mode-covering objective preserves useful alternatives that a mode-seeking update can lose.

1 INTRODUCTION

Knowledge distillation transfers a teacher model's predictive distribution to a smaller student (Hinton et al., 2015). For autoregressive models, the prefixes used for training matter: a student may encounter very different contexts when generating its own responses than when learning from a fixed corpus (Kim & Rush, 2016; Lin et al., 2020). On-policy distillation (OPD) addresses this mismatch by sampling student trajectories and querying the teacher at the resulting prefixes (Agarwal et al., 2024).

On-policy data determines where supervision is collected; the objective determines how the teacher signal reaches a capacity-limited student. Matching the teacher directly can demand a large target movement; damping that target helps only if the subsequent fit retains the teacher mass that motivated it. Target construction and target fitting must therefore be considered together.

Proximal Target Distillation (PTD) makes the compromise explicit. At each student-generated prefix, it chooses the distribution nearest to the teacher in forward KL among those inside a student-relative KL ball, then fits that target with forward KL. We define "mode covering" through the distributional $D_{\mathrm{KL}}(q\|r)$ objective, which penalizes omitted teacher mass, and evaluate its sequence-level consequence through multi-sample correct-answer coverage. The projection meters target movement; the fitting loss gives low-probability target tokens a direct gradient.

The construction operates entirely on the normalized teacher-top-K support of Revisiting OPD (Fu et al., 2026). The resulting mixed-KL centroid has a normalized Lambert-W form (Nock et al., 2008; Nielsen, 2013). We derive the path, bound the initial conditional-logit fitting gradient, and relate target- fitting error to token-level event mass. Under a fixed prefix and support, exact target replacement converges to the teacher; the inexact result allows summable additive teacher-KL tracking error.

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.

The experiments connect the controlled density mechanism to language-model outcomes. A misspecified two-component Gaussian student acquires substantial mass in a teacher region it initially undercovers. With best@16 for AIME and best@4 for AMC and science, PTD attains the highest Macro point estimate among the trained alternatives in both a 120-update math-only study and a separate teacher-routed multi-domain study. In the routed setting, the ordering between PTD and linear interpolation reverses at four samples, where PTD leads by 0.23 percentage points.

The main contributions are:

- a mode-covering target defined by a teacher-forward KL projection in a student-relative trust region, paired with forward fitting;
- a normalized Lambert-W path, with conditional-gradient, fitting-error, and fixed-support tracking guarantees; and
- a controlled density study and two on-policy language-model comparisons in which PTD attains the highest Macro best-of-n point estimate, together with a budget sweep that reveals a crossover from LI at low sample counts to PTD at four samples.

2 Related Work

On-policy objectives and support. Sequence-level and imitation-based distillation connect supervision to the student's contexts (Kim & Rush, 2016; Lin et al., 2020). GKD compares divergences on student-generated prefixes (Agarwal et al., 2024), while MiniLLM motivates reverse KL under a teacher–student capacity gap (Gu et al., 2024). Revisiting OPD matches separately normalized teacher and student distributions on teacher-top-K tokens (Fu et al., 2026); Rethinking OPD examines support alignment and model compatibility (Li et al., 2026). PTD keeps this support construction but fits a constrained intermediate distribution over all selected logits.

Closest stabilized OPD methods. Recent methods intervene at different stages. VETO builds a geometric bridge in logit space and suppresses low-confidence gradients (Jang et al., 2026). EOPD adds forward KL to reverse KL where teacher entropy is high (Jin et al., 2026). TrOPD gates on-policy supervision by teacher reliability and adds continuation from teacher prefixes (Xing et al., 2026). REOPOLD changes reward clipping, token sampling, and the exploration-to-refinement schedule (Ko et al., 2026), whereas TRB changes the warmup rollout policy but retains the per-prefix reverse-KL loss (Plyusov et al., 2026). TOP-D constructs an arithmetic proximal teacher and uses sampled response-token rewards with PPO-style optimization (Xie et al., 2026). By contrast, PTD computes a mixed-KL projection on the normalized teacher-top-K distribution and fits every selected logit with forward KL. Our linear-interpolation arm is therefore a target control under the opposite student-side constraint orientation. TOP-D differs in its sampled-token reward and PPO-style optimization; the support-based target comparisons share teacher-top-K.

Trust regions and information geometry. Classical TRPO and PPO constrain or clip policy updates (Schulman et al., 2015; 2017). PTD instead constrains an intermediate target distribution; its neural update depends on fitting and the optimizer. Mixed Bregman and Jeffreys centroids supply the Lambert-W structure (Nielsen & Nock, 2009; Nock et al., 2008; Veldhuis, 2002; Nielsen, 2013; Nielsen et al., 2014). Our contribution is the normalized teacher–student path on an OPD support together with its coupling to student fitting.

3 Method

3.1 Teacher-top-K on-policy supervision

A student snapshot π_{roll} samples responses $y \sim \pi_{\text{roll}}(\cdot \mid x)$ for prompts $x \sim \mathcal{D}$. Each response prefix $s = (x, y_{<t})$ becomes a fixed training input. Let $p_\theta^{\mathcal{V}}$ and $q^{\mathcal{V}}$ denote full-vocabulary student and teacher distributions. Following Fu et al. (2026), we select $S(s) = \mathrm{TopK}(q^{\mathcal{V}}(\cdot \mid s), K)$ and normalize both models on this support:

$$p_\theta(a \mid s, S) = \frac{p_\theta^{\mathcal{V}}(a \mid s)}{\sum_{b \in S} p_\theta^{\mathcal{V}}(b \mid s)}, \qquad q(a \mid s, S) = \frac{q^{\mathcal{V}}(a \mid s)}{\sum_{b \in S} q^{\mathcal{V}}(b \mid s)}. \tag{1}$$

These are separate softmax operations on logits gathered at the teacher's selected token indices. We sum over all K tokens, so a visited prefix receives supervision regardless of which next token the student sampled. The direct reverse-KL baseline is

$$\mathcal{L}_{\mathrm{topK,R}}(\theta) = \mathbb{E}_{s\sim\mu_{\mathrm{roll}}}\big[D_{\mathrm{KL}}(p_\theta(\cdot \mid s,S)\|q(\cdot \mid s,S))\big], \tag{2}$$

where μ_{roll} averages over valid response positions. At optimizer step k, the detached current student $p_k = \mathrm{sg}(p_{\theta_k}(\cdot \mid s,S))$ anchors the target. This anchor is refreshed after every optimizer step, including on reused rollout batches. Below, p, q, r denote positive distributions on S with $K \geq 2$. All KL quantities and guarantees are conditional on the selected support S.

3.2 Proximal target construction

At a fixed prefix, PTD solves

$$r^\star_\delta = \arg\min_{r\in\Delta(S)} D_{\mathrm{KL}}(q\|r) \quad \text{subject to} \quad D_{\mathrm{KL}}(r\|p_k) \leq \delta. \tag{3}$$

The objective weights errors by teacher probability and penalizes missing teacher mass. The constraint limits movement from the current student. At $\delta = 0$, the target is p_k; once $\delta \geq D_{\mathrm{KL}}(q\|p_k)$, it is the teacher q. Between these endpoints, strict convexity gives a unique optimizer with an active constraint (Boyd & Vandenberghe, 2004). The radius therefore indexes a continuous path from student to teacher.

For an active constraint, its multiplier $\lambda > 0$ yields

$$r_\lambda = \arg\min_{r\in\Delta(S)} \{D_{\mathrm{KL}}(q\|r) + \lambda D_{\mathrm{KL}}(r\|p_k)\}. \tag{4}$$

This mixed-KL centroid admits a normalized scalar solution (Nock et al., 2008; Nielsen, 2013). In the following theorem, write $p = p_k$ to simplify notation.

Theorem 1 (Proximal path)**.** *For every active solution of Eq.* (3)*, there is a scalar $t > 0$ such that*

$$r_t(a) = \frac{q(a)/W_0(tq(a)/p(a))}{\sum_{b\in S} q(b)/W_0(tq(b)/p(b))}. \tag{5}$$

The parameter is chosen to satisfy $D_{\mathrm{KL}}(r_t\|p) = \delta$. The path is continuous and obeys

$$\lim_{t\downarrow 0} r_t = p, \qquad \lim_{t\to\infty} r_t = q.$$

Here W_0 is the nonnegative branch of the Lambert function, $W(z)e^{W(z)} = z$ (Corless et al., 1996). Evaluating $D_{\mathrm{KL}}(r_t\|p)$ and searching over t gives the prescribed radius without optimizing over the K-token simplex. Tokenwise movement depends on the likelihood ratio $q(a)/p(a)$. In particular, the stationarity condition for Eq. (4) can be written as

$$-\frac{q(a)}{r(a)} + \lambda\left(\log\frac{r(a)}{p_k(a)} + 1\right) + \nu = 0, \tag{6}$$

where ν enforces normalization. The teacher term pushes up tokens that are underrepresented in r, while the student term charges for increasing their probability relative to p_k. The shared multiplier couples these tokenwise adjustments through one distribution-level budget, unlike a fixed interpolation weight between p_k and q. Appendix C derives the path and its endpoint limits.

Other KL orientations. Swapping the teacher objective or student constraint gives three related paths: geometric interpolation, linear interpolation, and swapped PTD. Table 1 separates the teacher-side objective from the student-side constraint. For our comparison, fitting follows the orientation of the student constraint. Thus the $D_{\mathrm{KL}}(r\|p_k)$ rows use forward fitting, whereas the $D_{\mathrm{KL}}(p_k\|r)$ rows use reverse fitting.

Linear interpolation shares PTD's teacher-forward objective but reverses the student-side constraint; the geometric target shares PTD's constraint but reverses the teacher objective. Locally, geometric targets reproduce a scaled reverse-KL gradient, while linear interpolation is tangent to the PTD path. Their finite-radius targets and optimizer trajectories remain distinct. Section 4 crosses geometry and fitting in a restricted family; Appendices A and B derive the identities.

Target	Teacher objective	Student constraint	Normalized path
PTD	$D_{\mathrm{KL}}(q\|r)$	$D_{\mathrm{KL}}(r\|p_k) \le \delta$	$r \propto q/W_0(tq/p_k)$
Linear interpolation	$D_{\mathrm{KL}}(q\|r)$	$D_{\mathrm{KL}}(p_k\|r) \le \delta$	$r = (1-\eta)p_k + \eta q$
Geometric	$D_{\mathrm{KL}}(r\|q)$	$D_{\mathrm{KL}}(r\|p_k) \le \delta$	$r \propto p_k^{1-\eta} q^{\eta}$
Swapped PTD	$D_{\mathrm{KL}}(r\|q)$	$D_{\mathrm{KL}}(p_k\|r) \le \delta$	$r \propto p_k/W_0(tp_k/q)$

Table 1: **Four student-relative target paths.** Equal numerical radii under opposite KL orientations define different feasible sets. The fitting rule follows the displayed student constraint in our matched comparison.

Algorithm 1 PTD on normalized teacher top-K support

Require: Student π_θ, frozen teacher π_T, support size K, path coefficient $B > 0$ or radius $\delta \ge 0$

for each rollout batch **do**
  Sample student responses and collect valid response prefixes
  Query and normalize teacher top-K logits to obtain q
  for each optimizer step on the batch **do**
    Gather current student logits; set $p_k = \mathrm{sg}(p_\theta)$
    Construct normalized r_k by Eq. (10), or solve for the requested radius
    Detach r_k, sum $D_{\mathrm{KL}}(r_k\|p_\theta)$ over valid prefixes, and update the student
  end for
end for

3.3 Forward fitting and implementation

We detach the target and minimize

$$\mathcal{L}_{\mathrm{fit}}(\theta; s) = D_{\mathrm{KL}}(\mathrm{sg}(r_k)\|p_\theta(\cdot \mid s, S)), \qquad \mathcal{L}_{\mathrm{PTD}}(\theta) = \mathbb{E}_{s\sim\mu_{\mathrm{roll}}}\mathcal{L}_{\mathrm{fit}}(\theta; s). \tag{7}$$

The target entropy is constant, so this is soft-label cross-entropy on the selected logits. For logit z_a, the gradient is simply

$$\frac{\partial \mathcal{L}_{\mathrm{fit}}}{\partial z_a} = p_\theta(a) - r_k(a). \tag{8}$$

A token whose probability is underestimated receives a gradient proportional to its target deficit. In contrast, fitting the same target with reverse KL gives

$$g_a^{\mathrm{R}} = p_\theta(a)\left[\log\frac{p_\theta(a)}{r_k(a)} - D_{\mathrm{KL}}(p_\theta\|r_k)\right]. \tag{9}$$

For a fixed positive target, this gradient vanishes as $p_\theta(a) \to 0$ when the remaining reverse-KL contribution stays bounded, whereas the forward gradient approaches $-r_k(a)$. Forward fitting therefore directly supervises target mass in regions the student rarely predicts.

We evaluate the target in log space using $t = Be^B$, with $B > 0$. For each selected token, the implementation computes

$$\begin{aligned} v_a &= \log W_0(\exp(\log B + B + \log q(a) - \log p_k(a))), \\ u_a &= \log q(a) - v_a, \qquad \log r_k(a) = u_a - \mathrm{LSE}_{b\in S}\, u_b. \end{aligned} \tag{10}$$

The final log-sum-exp is the normalization required by Eq. (5); omitting it changes the target weights and therefore the forward-fitting gradient. Appendix F describes a stable evaluation of $\log W_0(e^x)$ that never forms the potentially overflowing Lambert argument. A shared B selects a path point with effective radius $\delta_B(s) = D_{\mathrm{KL}}(r_k\|p_k)$ at each prefix; a scalar search instead enforces a specified δ. Our LLM experiments use the shared-B parameterization, yielding a prefix-dependent radius $\delta_B(s)$. Targets are refreshed at each optimizer step, gradients flow only through the current student, and loss is averaged over valid response tokens. Algorithm 1 summarizes the training loop; stable numerical details appear in Appendix F.

Choosing the target step. The path can be controlled either by a radius δ or by a shared coefficient B. A fixed radius enforces the same KL budget at every prefix. A shared B avoids a per-prefix

radius choice and lets the effective budget grow with teacher–student disagreement. Locally, $r_t = p + t(q-p) + O(t^2)$ and $D_{\mathrm{KL}}(r_t\|p) = \frac{1}{2}t^2\chi^2(q\|p) + O(t^3)$. Thus linear interpolation is tangent to PTD, while a shared coefficient induces more movement when teacher mass lies on low-student-probability tokens. Appendix E gives the full statement and proof for fixed positive distributions.

Target geometry interacts with the fitting loss. With a refreshed, detached linear-interpolation target and forward fitting, the gradient is exactly η times direct teacher-forward KL at the anchor (Eq. (23)). PTD shares that direction locally but uses a nonlinear target at finite steps. In our LLM comparison, LI uses reverse fitting, as specified in Section 3.2; the forward-fitting variant is included among the Gaussian controls.

3.4 From target guarantees to student updates

Projection and fitting form a two-stage update. The following propositions quantify how target proximity and fitting error control the learned conditional distribution on a fixed finite support.

Proposition 1 (Radius-dependent gradient bound). *If $D_{\mathrm{KL}}(r\|p_k) \le \delta$, then at the fitting initialization $p_\theta = p_k$,*

$$\|\nabla_z D_{\mathrm{KL}}(r\|p_\theta)\|_1 = \|p_k - r\|_1 \le \min\{2, \sqrt{2\delta}\}. \tag{11}$$

If additionally $\|\nabla_\theta z_a\|_2 \le G$ for each selected logit at this initialization, the corresponding parameter-gradient norm is at most $G\min\{2, \sqrt{2\delta}\}$.

Equation (8) and Pinsker's inequality give the bound directly; the parameter version follows by the triangle inequality. The radius therefore sets the initial conditional fitting-gradient budget; with a shared B, that budget uses the per-prefix radius $\delta_B(s)$.

Define $\mathrm{TV}(a,b) = \frac{1}{2}\|a-b\|_1$.

Proposition 2 (Consequences of forward fitting error). *Suppose $D_{\mathrm{KL}}(r\|p_k) \le \delta$ and a fitted student distribution p_+ satisfies $D_{\mathrm{KL}}(r\|p_+) \le \epsilon$. Then*

$$\mathrm{TV}(p_k, p_+) \le \sqrt{\delta/2} + \sqrt{\epsilon/2}, \tag{12}$$

$$|p_+(A) - r(A)| \le \sqrt{\epsilon/2} \quad \textit{for every } A \subseteq S. \tag{13}$$

The first bound combines target proximity and fitting error by the triangle inequality; the event characterization of total variation gives the second. Small fitting error transfers both proximity and event mass to the learned student on the selected token support. Section 5 evaluates the resulting autoregressive best-of-n coverage.

3.5 Convergence of target tracking

Fix one prefix, its positive teacher distribution q, and its top-K support. For shared $B > 0$, let $T_B^q(p)$ denote the target in Eq. (5) with $t = Be^B$, and write $D_q(p) = D_{\mathrm{KL}}(q\|p)$ and $\Delta_B^q(p) = D_q(p) - D_q(T_B^q(p))$.

Theorem 2 (Exact and inexact target tracking). *For every positive p, $\Delta_B^q(p) \ge 0$, with equality if and only if $p = q$. Consequently, $p_{k+1} = T_B^q(p_k)$ converges to q from any positive initialization. More generally, suppose positive iterates satisfy*

$$D_q(p_{k+1}) \le D_q(T_B^q(p_k)) + \varepsilon_k, \qquad \varepsilon_k \ge 0. \tag{14}$$

If $\sum_{k=0}^{\infty}\varepsilon_k < \infty$, then $p_k \to q$. Here ε_k is additive teacher-KL tracking error.

Corollary 1 (Local contraction rate). *For a zero-sum perturbation $h \to 0$ around a fixed positive q,*

$$T_B^q(q+h) = q + \frac{1}{1+B}h + O(\|h\|^2), \tag{15}$$

$$D_q(T_B^q(q+h)) = \frac{1}{(1+B)^2}D_q(q+h) + O(\|h\|^3). \tag{16}$$

Theorem 2 establishes fixed-prefix convergence for exact replacement and summable teacher-KL tracking error; Appendix D gives the proof. Section 4 evaluates the repeated-target mechanism under misspecified restricted capacity.

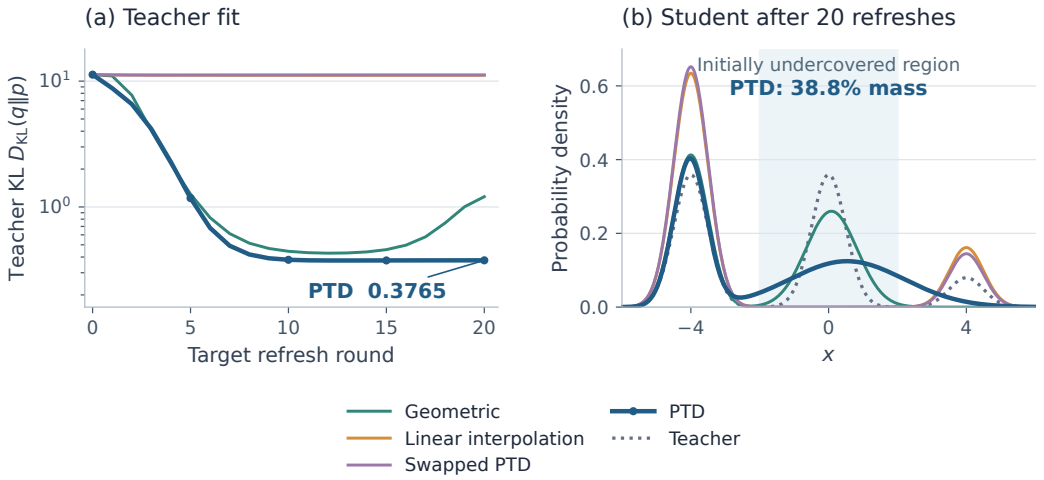


Figure 1: **PTD learns mass in an initially undercovered region.** Left: teacher-forward KL during repeated target construction and fitting. Right: final student densities after 20 target refreshes; shading marks the middle region. Each method uses its native student-relative KL constraint with numerical radius 0.1; opposite KL orientations define different feasible sets.

4 Gaussian synthetic analysis

Consider a misspecified student family. Let $\phi_j = \mathcal{N}(\mu_j, 0.5^2)$ with $(\mu_1, \mu_2, \mu_3) = (-4, 0, 4)$ and teacher $q = 0.45\phi_1 + 0.45\phi_2 + 0.10\phi_3$. The student starts at $p_0 = 0.49\phi_1 + 0.51\phi_3$ and has only two Gaussian components. Its trainable weights, means, and variances form a misspecified approximation to the three-component teacher. The teacher's middle region $[-2, 2)$ initially has almost no student mass.

At each of 20 rounds, we construct a target at native KL radius 0.1 and optimize within the two-component student family until the projected gradient norm falls below 10^{-6}; the resulting student becomes the next anchor. Deterministic quadrature removes sampling noise. Full protocols and convergence checks appear in Appendix I.

Target refresh and undercovered mass. After 20 refreshes, PTD reduces teacher-forward KL to 0.3765 and assigns 38.78% probability to the middle region (Figure 1). Its second component broadens to cover this initially undercovered region ($\mu \approx 0.54$, $\sigma \approx 1.66$), accommodating the teacher's three-component distribution with two student components. Optimizing only the initial PTD target instead plateaus at teacher-forward KL 8.7876, even after 5,000 steps. Once the student acquires some middle-region mass, a refreshed target continues moving toward the teacher, while the fixed-target trajectory remains at its plateau.

Native-radius paths and fitting loss. After 20 rounds, the geometric target reaches teacher-forward KL 1.2069. Linear interpolation and swapped PTD reach 11.1394 and 11.1951, respectively, and leave only about 0.003% probability in the middle region. Their stationary endpoints retain this low middle mass. PTD gives the lowest teacher KL among these four *native-radius* formulations. Opposite KL orientations define different feasible sets at the same numerical radius.

To separate the target path from the fitting loss, we pair each path with the other loss as well. Fitting a linear interpolation target with $D_{\mathrm{KL}}(r\|p_\theta)$ gives teacher-forward KL 0.3708, close to PTD's 0.3765; matching its student-relative radius to PTD gives 0.6392. Optimizing directly against the teacher with the same loss also reaches 0.3708. Under the shared undercovered initialization, forward fitting drives the recovery; PTD geometry meters that signal through an implicit student-relative budget and determines the path and overall teacher fit. With the same fitting loss, the geometric target assigns more middle-region mass (48.09%) but has worse overall teacher fit. The appendix reports the full

target–loss grid and fixed-target controls. The LLM study evaluates the coupled target-and-fitting designs; the Gaussian controls separate their two components.

Free targets at a common radius. A final check removes restricted-family optimization. To compare free target geometry while keeping all three mixture weights positive, it starts from $p = 0.49\phi_1 + 0.01\phi_2 + 0.50\phi_3$; the restricted-family case has two trainable components concentrated in the outer regions. Within $D_{\mathrm{KL}}(r\|p) \leq 0.1$, PTD obtains teacher-forward KL 0.5925, compared with 0.5988 for radius-matched linear interpolation and 0.6570 for the geometric target. PTD therefore attains the minimum on this shared feasible set: 0.5925 versus 0.5988 and 0.6570. The restricted-family case illustrates how one trajectory represents such targets with a smaller model.

5 Reasoning distillation experiments

We study a 120-update math-only setting and a separate 120-update multi-domain setting. Together they include eight trained students, the initial model, and two teachers, each evaluated on four benchmarks.

5.1 Experimental setup

Models and training. Both settings use DeepSeek-R1-Distill-Qwen-1.5B as the student (DeepSeek-AI et al., 2025). Math-only training uses Skywork-OR1-Math-7B and DAPO math prompts (He et al., 2025; Yu et al., 2025); its four arms are evaluated at step 120. The multi-domain study contains 1,280 human prompts from OpenThoughts3-1.2M (Guha et al., 2025), evenly split between math and science. We preserve prompt text and alternate task-homogeneous batches by domain. Each routed arm starts from the same initial student and runs for 120 updates—three passes over the data—with seed 42. All objectives use AdamW (Loshchilov & Hutter, 2019) at learning rate 2×10^{-6}, with 32 prompts per batch and four rollouts per prompt. Support-based objectives use the teacher's normalized top-32 support. We evaluate every arm after 120 updates.

Comparisons. Top-K reverse KL minimizes $D_{\mathrm{KL}}(p_\theta\|q)$ on the teacher support. Linear interpolation (LI) constructs $r_k = 0.8p_k + 0.2q$ and uses reverse fitting. PTD constructs its target with $B = -\log(0.8)$ and uses forward fitting. The two target methods follow their student-constraint orientations as described in Section 3.2. PTD is the coupled target-plus-fitting design. Gaussian controls separately vary target geometry and fitting orientation. Under the shared coefficient convention, the two targets realize different per-prefix KL radii. Both studies report a sampled-token reverse-KL reference with detached $\log q - \log p_{\mathrm{old}}$ advantages and the PPO ratio surrogate.

Evaluation. We sample 32 responses per question on AIME24 (30 questions), AIME25 (30), and AMC23 (40), and four on GPQA Diamond (198) (Rein et al., 2023). All models use temperature 1, top-$p = 0.9$, and a 15,360-token response limit. Our primary metric is best-of-n success: a question succeeds if at least one response is correct. AIME uses fixed draw indices 0–15, AMC uses 0–3, and GPQA uses all four draws. Math averages the three math benchmarks, and Macro avg. equally averages Math and GPQA. This metric measures multi-sample correct-answer success. Its value combines per-sample accuracy, formatting reliability, and the sampling budget. Table 9 reports budget sensitivity; Appendix J gives bootstrap details.

5.2 Results

Math-only training. At step 120, PTD reaches 73.06% Math and 70.87% Macro avg. (Table 2), 3.61 and 5.85 percentage points (pp) above the initial student. Its Macro score exceeds sampled reverse KL by 7.99 pp ($[3.80, 12.32]$). In the math-only setting, GPQA quantifies science retention and cross-domain transfer.

Teacher-routed multi-domain training. At step 120, PTD reaches 65.81% Macro avg., 0.78 pp above the initial student, while its 70.00% Math score is 0.56 pp above the initial model. Its Macro score exceeds the sampled reverse-KL reference by 5.44 pp ($[0.58, 10.37]$). PTD therefore preserves the same Macro ordering under teacher routing and improves both aggregate components over the sampled reference.

	AIME: best@16		AMC: best@4	Science: best@4	Aggregates	
Method	AIME24	AIME25	AMC23	GPQA	Math [95% CI]	Macro avg. [95% CI]
Reference models						
Initial student	66.67	56.67	85.00	60.61	69.44 [60.28, 78.06]	65.03 [59.37, 70.58]
Math teacher	83.33	66.67	95.00	69.70	81.67 [73.89, 88.89]	75.68 [70.64, 80.51]
General teacher	80.00	70.00	95.00	63.64	81.67 [73.89, 89.17]	72.65 [67.50, 77.60]
Math-only training						
Top-K reverse KL	73.33	43.33	90.00	60.10	68.89 [60.28, 77.22]	64.49 [58.99, 69.89]
Sampled reverse KL	66.67	43.33	90.00	59.09	66.67 [57.78, 75.28]	62.88 [57.29, 68.37]
Linear interpolation	66.67	**53.33**	**95.00**	61.62	71.67 [63.06, 80.00]	66.64 [61.16, 71.99]
PTD (ours)	**76.67**	50.00	92.50	**68.69**	**73.06** [64.72, 81.11]	**70.87** [65.53, 76.04]
Multi-domain training						
Top-K reverse KL	66.67	**46.67**	87.50	57.58	66.94 [58.06, 75.56]	62.26 [56.59, 67.80]
Sampled reverse KL (reference)	70.00	40.00	82.50	56.57	64.17 [55.28, 73.06]	60.37 [54.67, 65.96]
Linear interpolation	70.00	**46.67**	**92.50**	60.61	69.72 [61.11, 78.06]	65.16 [59.67, 70.57]
PTD (ours)	**73.33**	**46.67**	90.00	**61.62**	**70.00** [61.39, 78.33]	**65.81** [60.30, 71.19]

Table 2: **Best-of-n success after on-policy distillation.** Percentage solved at fixed budgets (AIME 16, AMC 4, GPQA 4), all at step 120. Bold marks the best trained student per setting and column; shading marks PTD. Aggregate brackets are 95% item-bootstrap intervals (200,000 replicates; Appendix J).

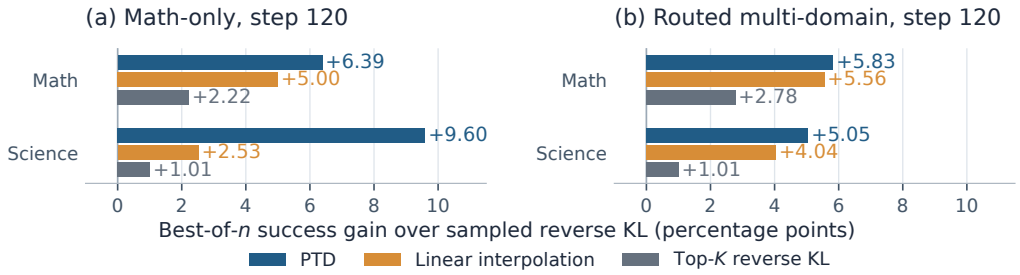


Figure 2: **Best-of-n success gains over sampled reverse KL.** Bars show method-minus-reference percentage points. The matched routed comparison comprises top-K reverse KL, LI, and PTD.

Benchmark-level structure. At best@4, AMC separates the methods and favors LI. In math-only training, LI leads AMC23 at 95.00%, followed by PTD at 92.50% and the two reverse-KL variants at 90.00%. The same ordering appears in the multi-domain setting: LI reaches 92.50%, PTD 90.00%, top-K reverse KL 87.50%, and sampled reverse KL 82.50%. PTD's Math advantage instead comes from AIME. In the math-only setting it leads AIME24 by 3.34 pp over top-K reverse KL and by 10.00 pp over LI, while LI leads AIME25 by 3.33 pp. In the routed setting, PTD leads AIME24 by 3.33 pp over LI and 6.66 pp over top-K reverse KL, with all three methods tied on AIME25. GPQA supplies the other consistent advantage: PTD reaches 68.69% in math-only training and 61.62% with routing, the largest point estimate in each setting. PTD's AIME coverage and GPQA retention or transfer drive the aggregate ordering, whereas AMC favors LI.

Gains over sampled reverse KL. Figure 2 uses sampled reverse KL as the reference in both settings. In math-only training, PTD gains 6.39 pp on Math and 9.60 pp on GPQA over sampled reverse KL.

LI gains 5.00 and 2.53 pp, while top-K reverse KL gains 2.22 and 1.01 pp. The remaining PTD–LI gap is therefore 1.39 pp on Math and 7.07 pp on GPQA. With multi-domain training, PTD gains 5.83 pp on Math and 5.05 pp on GPQA over the sampled baseline; the corresponding LI gains are 5.56 and 4.04 pp. The PTD–LI gap contracts to 0.28 pp on Math and 1.01 pp on GPQA. Routing improves the balance of LI's gains across the two domains, while PTD retains the highest Macro point estimate. The large math-only separation on GPQA also explains why its Macro gain is substantially larger than its Math gain.

Sampling budget and answer validity. For the routed models, estimated Macro pass@1 is 36.15% for PTD, 37.57% for LI, and 35.22% for reverse KL; at pass@4 the values are 61.42%, 61.19%, and 59.05%. At pass@2, the corresponding values are 48.05%, 49.30%, and 46.52%. LI's margin over PTD falls from 1.42 pp at one sample to 1.25 pp at two, then reverses to a 0.23 pp PTD lead at four. PTD's margin over reverse KL grows from 0.93 to 1.53 to 2.37 pp across the same budgets. At the primary mixed budgets, PTD, LI, and reverse KL reach 65.81%, 65.16%, and 62.26%. The method ordering thus changes with the number of attempts: LI is strongest for single-sample success, while PTD is strongest at four draws and at the primary mixed budgets. Accuracy and formatting both contribute to this profile: the GPQA invalid-output rates for PTD, LI, and reverse KL are 18.6%, 21.2%, and 26.4%, respectively. The widening margin over top-K reverse KL and the crossover with LI locate PTD's advantage in the multi-sample regime targeted by mode covering.

Effect decomposition and uncertainty. Because Macro avg. gives equal weight to Math and GPQA, each component's contribution is half of its benchmark-level difference. Relative to sampled reverse KL in math-only training, top-K reverse KL, LI, and PTD gain 1.62, 3.76, and 7.99 pp Macro, respectively. Their paired 95% intervals are $[-2.90, 6.19]$, $[-1.10, 8.72]$, and $[3.80, 12.32]$; PTD has both the largest point estimate and the only interval with a positive lower bound. In routed training, the corresponding Macro gains are 1.89, 4.80, and 5.44 pp, with paired intervals $[-2.87, 6.69]$, $[0.05, 9.63]$, and $[0.58, 10.37]$. The two target-construction methods have positive lower bounds, and PTD retains the largest point estimate.

Appendix J reports the full routed budget sensitivity and paired bootstrap results; Appendix K collects limitations and scope.

6 Conclusion

Proximal Target Distillation turns mode covering into a controlled student-relative update: its teacher-forward projection preserves teacher-supported alternatives within a trust region, while forward fitting directly trains underrepresented tokens. In the Gaussian study, repeated updates assign 38.78% probability to a teacher region initially nearly absent from the misspecified student. The controls isolate forward fitting as the recovery mechanism and show that proximal geometry controls its trajectory.

In language models, PTD records the highest Macro best-of-n point estimate among trained student methods in both math-only and teacher-routed multi-domain settings. Its paired gains over sampled reverse KL are 7.99 and 5.44 percentage points, with positive lower confidence bounds. In the routed sweep, its margin over top-K reverse KL grows from 0.93 points at one draw to 2.37 at four, where it also overtakes linear interpolation. This widening margin localizes PTD's strongest gains to multi-sample coverage. Together, the results show that PTD's distribution-level mode-covering objective yields higher best-of-n success across both training settings.

The budget sweep also supplies a practical diagnostic for mode-covering distillation: PTD's coverage advantage grows as sampling gives the student multiple opportunities to express learned alternatives. This gain is accompanied by the lowest GPQA invalid-output rate among the routed comparisons. Read together with the Gaussian controls, the evidence traces a common mechanism across scales: the proximal target preserves teacher-supported mass, forward fitting transfers it into the capacity-limited student, and autoregressive sampling turns retained mass into additional valid opportunities for a correct answer. This end-to-end view makes coverage an operational design criterion for on-policy distillation under a teacher–student capacity gap.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: generate synthetic data sets for the controlled Gaussian study; assist in the writing of proofs; design or provide feedback on research methodology or experiments; implement methods. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, assist with translation, interpret results, or support qualitative and thematic data analysis. The authors conceived the core method, hypotheses, and experimental design, and the rest of the required disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools for tasks with recommended disclosure: create or modify scientific figures or images; create or edit software code; creation of artifacts, including tables and experiment and evaluation monitoring artifacts; draft parts of a research paper; summarize or analyse existing literature; sourcing/searching for information; edit a research paper to improve readability; identify relevant literature; format references. We have reviewed all AI-assisted work. The authors inspected the cited primary sources, verified the mathematical arguments, tested the implementations, regenerated the numerical outputs and artifacts, and reviewed and edited the final 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, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

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

Robert M. Corless, Gaston H. Gonnet, D. E. G. Hare, David J. Jeffrey, and Donald E. Knuth. On the Lambert W function. *Advances in Computational Mathematics*, 5:329–359, 1996. doi: 10.1007/BF02124750.

DeepSeek-AI et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York, 1993. doi: 10.1007/978-1-4899-4541-9.

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. URL https://arxiv.org/abs/2603.25562.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Etash Guha, Ryan Marten, Sedrick Keh, et al. OpenThoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025. URL https://arxiv.org/abs/2506.04178.

Jujie He, Jiacai Liu, Chris Yuhao Liu, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025. URL https://arxiv.org/abs/2505.22312.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42217–42227. Association for Computational Linguistics, 2026. doi: 10.18653/

v1/2026.findings-acl.2094. URL https://aclanthology.org/2026.findings-acl.2094/.

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 *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://openreview.net/forum?id=J5i09faOOf.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016. doi: 10.18653/v1/D16-1139.

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. URL https://arxiv.org/abs/2603.11137.

Hynek Kydlíček. Math-Verify: Math verification library. Software, 2026. URL https://github.com/huggingface/Math-Verify. Version 0.9.0.

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. *arXiv preprint arXiv:2604.13016*, 2026. URL https://arxiv.org/abs/2604.13016.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6121–6133, 2020. doi: 10.18653/v1/2020.emnlp-main.494.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Frank Nielsen. Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms. *IEEE Signal Processing Letters*, 20(7):657–660, 2013. doi: 10.1109/LSP.2013.2260538. URL https://arxiv.org/abs/1303.7286.

Frank Nielsen and Richard Nock. Sided and symmetrized Bregman centroids. *IEEE Transactions on Information Theory*, 55(6):2882–2904, 2009. doi: 10.1109/TIT.2009.2018176.

Frank Nielsen, Richard Nock, and Shun-ichi Amari. On clustering histograms with k-means by using mixed α-divergences. *Entropy*, 16(6):3273–3301, 2014. doi: 10.3390/e16063273.

Richard Nock, Panu Luosto, and Jyrki Kivinen. Mixed Bregman clustering with approximation guarantees. In *Machine Learning and Knowledge Discovery in Databases*, pp. 154–169. Springer, 2008. doi: 10.1007/978-3-540-87481-2_11.

Daniil Plyusov, Alexey Gorbatovski, Alexey Malakhov, Nikita Balagansky, Boris Shaposhnikov, Daria Korotyshova, and Daniil Gavrilov. Trust-region behavior blending for on-policy distillation. *arXiv preprint arXiv:2605.31159*, 2026. URL https://arxiv.org/abs/2605.31159.

David Rein, Betty Li Hou, Asa Cooper Stickland, et al. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. URL https://arxiv.org/abs/2311.12022.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1889–1897, 2015. URL https://arxiv.org/abs/1502.05477.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Raymond N. J. Veldhuis. The centroid of the symmetrical Kullback–Leibler distance. *IEEE Signal Processing Letters*, 9(3):96–99, 2002. doi: 10.1109/97.995827.

Zhengpeng Xie, Li Lyna Zhang, Zeke Xie, and Mao Yang. Trust region policy distillation. *arXiv preprint arXiv:2607.04751*, 2026. URL https://arxiv.org/abs/2607.04751.

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

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

A ALTERNATIVE KL ORIENTATIONS

The teacher objective and student constraint each admit two KL orientations. Let p_k be the current student and q the teacher on the shared support:

$$\begin{aligned} T_q(r) &\in \{D_{\mathrm{KL}}(q\|r), D_{\mathrm{KL}}(r\|q)\}, \\ C_{p_k}(r) &\in \{D_{\mathrm{KL}}(p_k\|r), D_{\mathrm{KL}}(r\|p_k)\}, \\ r_k^\star &= \arg\min_{r\in\Delta(S)} T_q(r) \quad \text{subject to} \quad C_{p_k}(r) \le \delta. \end{aligned} \tag{17}$$

For this comparison, we match the fitting divergence to the student constraint: substitute p_θ for its student argument and detach the target,

$$\mathcal{L}_k(\theta) = C_{p_\theta}\big(\mathrm{sg}(r_k^\star)\big) = \begin{cases} D_{\mathrm{KL}}(p_\theta\|\,\mathrm{sg}(r_k^\star)), & C_{p_k}(r) = D_{\mathrm{KL}}(p_k\|r), \\ D_{\mathrm{KL}}(\mathrm{sg}(r_k^\star)\|p_\theta), & C_{p_k}(r) = D_{\mathrm{KL}}(r\|p_k). \end{cases} \tag{18}$$

Appendix I lists the four normalized paths. Matching the divergence orientation provides a consistent comparison convention. The fitted student's realized KL can differ from the target's budget under finite capacity. Section 4 tests these alternatives and crosses the target paths with the other fitting rule as a control.

Local relationship to reverse KL. The combination $D_{\mathrm{KL}}(r\|q)$ with a $D_{\mathrm{KL}}(r\|p_k)$ constraint has path

$$r_\eta(a) = \frac{p_k(a)^{1-\eta} q(a)^\eta}{Z_\eta}, \qquad 0 \le \eta \le 1. \tag{19}$$

For the active penalized problem, $\eta = 1/(1+\lambda)$, where λ is the constraint multiplier. Under Eq. (18), its fitting loss is forward KL. Let z be the conditional student logits and $\ell(a) = \log(q(a)/p_k(a))$. At the anchor, its exact gradient and small-η expansion are

$$\begin{aligned} \nabla_z D_{\mathrm{KL}}(r_\eta\|p_\theta)\big|_{p_\theta=p_k} &= p_k - r_\eta \\ &= \eta\, p_k\big(\mathbb{E}_{p_k}\ell - \ell\big) + O(\eta^2) \\ &= \eta\, \nabla_z D_{\mathrm{KL}}(p_\theta\|q)\big|_{p_\theta=p_k} + O(\eta^2). \end{aligned} \tag{20}$$

For fixed positive p_k, q, this choice locally reproduces a rescaled reverse-KL direction. At finite radius, higher-order terms change the direction, especially for large likelihood ratios. An exact anchor-gradient rescaling holds if the geometric target is fitted with *reverse* KL, but that changes the fitting rule (Appendix H). Adaptive-optimizer trajectories also depend on moments, weight decay, and subsequent steps away from the anchor.

The linear interpolation and swapped-PTD alternatives instead constrain $D_{\mathrm{KL}}(p_k\|r)$ and use reverse fitting under the same convention. Their targets can assign mass to regions that the student underrepresents, with retention determined by the restricted-family fit. Section 4 gives a concrete example. PTD combines the teacher-coverage objective $D_{\mathrm{KL}}(q\|r)$, the proximity constraint $D_{\mathrm{KL}}(r\|p_k)$, and forward fitting. The comparisons separate these choices and serve as mechanism controls across KL orientations.

B LINEAR INTERPOLATION TARGETS AND FITTING CONTROLS

The main linear-interpolation comparison uses reverse fitting (Section 3.2). As a control, we also fit the same target with forward KL, deliberately breaking the matched-orientation rule. A linear interpolation target has the form

$$r_\eta^{\mathrm{lin}} = (1-\eta)p_k + \eta q, \qquad \eta \in [0,1]. \tag{21}$$

For $0 < \eta < 1$, this is the minimizer of $D_{\mathrm{KL}}(q\|r) + \lambda D_{\mathrm{KL}}(p_k\|r)$ with $\lambda = (1-\eta)/\eta$. It changes the student-side geometry relative to Eq. (4). Under forward fitting, it also satisfies the exact identity

$$\begin{aligned} D_{\mathrm{KL}}(q\|p_\theta) + \lambda D_{\mathrm{KL}}(p_k\|p_\theta) &= (1+\lambda) D_{\mathrm{KL}}(r_\eta^{\mathrm{lin}}\|p_\theta) + C, \\ C &= D_{\mathrm{KL}}(q\|r_\eta^{\mathrm{lin}}) + \lambda D_{\mathrm{KL}}(p_k\|r_\eta^{\mathrm{lin}}). \end{aligned} \tag{22}$$

The constant C is independent of θ. The identity holds even for a restricted student family: forward fitting of the interpolated target is a teacher-forward KL objective plus a forward-KL penalty to the old student.

If the linear interpolation target is rebuilt at every optimizer step and detached, its gradient at the current anchor obeys

$$\nabla_\theta D_{\mathrm{KL}}(r_\eta^{\mathrm{lin}} \| p_\theta)\big|_{\theta=\theta_k} = \eta\, \nabla_\theta D_{\mathrm{KL}}(q \| p_\theta)\big|_{\theta=\theta_k}. \tag{23}$$

For a shared constant η, this is a scaled direct-forward gradient. For plain SGD it is equivalent to scaling the learning rate. Adaptive moments, weight decay, and additional steps away from the anchor modify the resulting trajectory.

Proposition 3 shows that linear interpolation is tangent to the Lambert target path at the student. Under forward fitting, their initial gradients agree to first order when $\eta = t$. The expansion assumes fixed positive distributions and bounded likelihood ratios. Farther from the anchor, the target geometry can differ. We compare the two families at matched $D_{\mathrm{KL}}(r\|p_k)$ to isolate this difference. The coefficient relation $B = -\log(1-\eta)$ leaves their target radii and forward-fitting gradients unmatched.

C Proof of Theorem 1

The coordinate-wise inversion follows the same Lambert-W mechanism as mixed-KL and Jeffreys centroid calculations (Nock et al., 2008; Nielsen, 2013); we provide the complete constrained derivation to identify the radius parameter and normalized top-K target used here.

For the active case, the Lagrangian is

$$\mathcal{J}(r,\lambda,\nu) = -\sum_a q(a)\log r(a) + \lambda\left(\sum_a r(a)\log\frac{r(a)}{p(a)} - \delta\right) + \nu\left(\sum_a r(a) - 1\right),$$

up to terms independent of r. Slater's condition holds for $\delta > 0$ because $r = p$ is strictly feasible. Convexity therefore makes the KKT conditions sufficient (Boyd & Vandenberghe, 2004). Since $0 < \delta < D_{\mathrm{KL}}(q\|p)$, the unconstrained minimizer q is infeasible, so $\lambda > 0$.

Differentiating with respect to $r(a)$ yields

$$-\frac{q(a)}{r(a)} + \lambda\left(\log\frac{r(a)}{p(a)} + 1\right) + \nu = 0.$$

Let

$$y(a) = \frac{r(a)}{p(a)}, \qquad \rho(a) = \frac{q(a)}{p(a)}, \qquad c = 1 + \frac{\nu}{\lambda}.$$

Then

$$y(a)(\log y(a) + c) = \frac{\rho(a)}{\lambda}.$$

With $v(a) = \log y(a) + c$, we have $y(a) = e^{v(a)-c}$ and hence

$$v(a)e^{v(a)} = \frac{e^c}{\lambda}\rho(a).$$

Defining $t = e^c/\lambda > 0$ gives

$$v(a) = W_0(t\rho(a)).$$

Using $e^{W_0(z)} = z/W_0(z)$,

$$y(a) = \frac{\rho(a)}{\lambda W_0(t\rho(a))}, \qquad r(a) = \frac{q(a)}{\lambda W_0(tq(a)/p(a))}.$$

The condition $\sum_a r(a) = 1$ sets

$$\lambda = \sum_b \frac{q(b)}{W_0(tq(b)/p(b))},$$

which proves Eq. (5). Strict convexity of the objective on the positive simplex gives uniqueness.

For the first endpoint, the expansion $W_0(z) = z - z^2 + O(z^3)$ gives

$$\frac{q(a)}{W_0(tq(a)/p(a))} = \frac{p(a)}{t} + q(a) + O(t).$$

After normalization, $r_t(a) \to p(a)$. For the second endpoint, $W_0(t\rho)/W_0(t) \to 1$ for each fixed $\rho > 0$ as $t \to \infty$. Consequently, the common asymptotic factor cancels in Eq. (5), and $r_t(a) \to q(a)$. Continuity then gives a point on the path for every active radius between 0 and $D_{\mathrm{KL}}(q\|p)$.

The radius is also monotone along the path. Its penalty multiplier $\lambda(t) = \sum_a q(a)/W_0(tq(a)/p(a))$ is strictly decreasing in t. Comparing the optimality inequalities for two positive multipliers shows that the minimizer's $D_{\mathrm{KL}}(r\|p)$ is nonincreasing in λ. For $p \neq q$, equality at distinct multipliers would give the same unique minimizer; subtracting its stationarity equations forces $r = p$, which in turn forces $q = p$. Thus the radius is strictly increasing in t when $p \neq q$, and the active radius has a unique scalar solution.

D Proof of Theorem 2 and Corollary 1

Fix one prefix, its teacher top-K support, a positive teacher q, and $B > 0$. Set $t = Be^B$, write $T_B^q(p) = r_t(p, q)$, and let $D_q(p) = D_{\mathrm{KL}}(q\|p)$. Define

$$\Delta_B^q(p) = D_q(p) - D_q(T_B^q(p)). \tag{24}$$

Projection improvement. Fix p and let $r = T_B^q(p)$. The KKT construction in Appendix C also works in the converse direction: for each $t = Be^B > 0$, Eq. (5) satisfies the KKT conditions for radius $\delta_B(p) = D_{\mathrm{KL}}(r\|p)$. In particular, define

$$\lambda_t = \sum_a \frac{q(a)}{W_0(tq(a)/p(a))}, \quad c_t = \log(t\lambda_t), \quad \nu_t = \lambda_t(c_t - 1). \tag{25}$$

These multipliers recover the stationarity equation in Appendix C; normalization and complementary slackness hold by construction. Thus, r is the unique minimizer of $D_q(v) = D_{\mathrm{KL}}(q\|v)$ over $\{v : D_{\mathrm{KL}}(v\|p) \leq \delta_B(p)\}$. Because p is feasible,

$$D_q(T_B^q(p)) \leq D_q(p). \tag{26}$$

Moreover, $T_B^q(p) = p$ only if $p = q$. Indeed, the Lambert identity rewrites the unnormalized target as $q(a)/W_0(tq(a)/p(a)) = p(a)e^{W_0(tq(a)/p(a))}/t$. If normalization made the target equal to p, then $W_0(tq(a)/p(a))$ would be constant in a. Strict monotonicity of W_0 would make $q(a)/p(a)$ constant, and normalization would force that constant to be one. Strict convexity then makes Eq. (26) strict whenever $p \neq q$, proving the first claim.

Exact iteration. Set $D_k = D_q(p_k)$. Exact updates satisfy $D_{k+1} = D_k - \Delta_B^q(p_k)$, so D_k decreases to a finite limit and $\Delta_B^q(p_k) \to 0$. Because every component of q is positive, each forward-KL sublevel set $\{p : D_q(p) \leq C\}$ is a compact subset of the simplex interior. To see the separation from the boundary, use $q(b)\log(q(b)/p(b)) \geq q(b)\log q(b)$ for $p(b) \leq 1$; a finite upper bound on $D_q(p)$ then gives a positive lower bound on every $p(a)$.

The map T_B^q and the gap Δ_B^q are continuous on this compact sublevel set. Every accumulation point $\bar{p}$ of the iterates therefore satisfies $\Delta_B^q(\bar{p}) = 0$, which by the first part implies $\bar{p} = q$. Since the sequence has no other accumulation point, $p_k \to q$.

Inexact iteration. Equation (14) implies

$$D_{k+1} \leq D_k - \Delta_B^q(p_k) + \varepsilon_k. \tag{27}$$

If $\sum_k \varepsilon_k < \infty$, then $D_k \leq D_0 + \sum_i \varepsilon_i$, so all iterates remain in one compact forward-KL sublevel set. Summing Eq. (27) through step n gives

$$\sum_{k=0}^{n} \Delta_B^q(p_k) \leq D_0 - D_{n+1} + \sum_{k=0}^{n} \varepsilon_k \leq D_0 + \sum_{k=0}^{\infty} \varepsilon_k. \tag{28}$$

The nonnegative gaps are summable and hence converge to zero. The same continuity and accumulation-point argument proves $p_k \to q$.

Local contraction. Let $p = q + h$, where $\sum_a h(a) = 0$, and let $h \to 0$. Since $q(a)/p(a) = 1 - h(a)/q(a) + O(\|h\|^2)$ and

$$\left.\frac{\partial W_0(t\rho)}{\partial \log \rho}\right|_{\rho=1} = \frac{B}{1+B}, \tag{29}$$

we have

$$W_0\left(t\frac{q(a)}{p(a)}\right) = B - \frac{B}{1+B}\frac{h(a)}{q(a)} + O(\|h\|^2). \tag{30}$$

It follows that the unnormalized numerator in Eq. (5) is

$$\frac{q(a)}{W_0(tq(a)/p(a))} = \frac{q(a)}{B} + \frac{h(a)}{B(1+B)} + O(\|h\|^2). \tag{31}$$

Its sum is $1/B + O(\|h\|^2)$ because h has zero total mass. Normalization proves Eq. (15). Finally, the standard local expansion

$$D_q(q+h) = \frac{1}{2}\sum_a \frac{h(a)^2}{q(a)} + O(\|h\|^3) \tag{32}$$

together with Eq. (15) proves Eq. (16).

E Local relationship to linear interpolation targets

Proposition 3 (Local path expansion)**.** *For fixed positive p, q on a finite support, as $t \downarrow 0$,*

$$r_t(a) = p(a) + t(q(a) - p(a)) + O(t^2), \tag{33}$$

$$D_{\mathrm{KL}}(r_t\|p) = \frac{t^2}{2}\chi^2(q\|p) + O(t^3), \qquad \chi^2(q\|p) = \sum_{a\in S}\frac{(q(a)-p(a))^2}{p(a)}. \tag{34}$$

Using $W_0(z) = z - z^2 + O(z^3)$,

$$\frac{1}{W_0(z)} = \frac{1}{z} + 1 + O(z).$$

Substituting $z = tq(a)/p(a)$ gives

$$\frac{q(a)}{W_0(tq(a)/p(a))} = \frac{p(a)}{t} + q(a) + O(t).$$

Summing over a yields $t^{-1} + 1 + O(t)$. Dividing the two expansions gives

$$r_t(a) = p(a) + t(q(a) - p(a)) + O(t^2).$$

For a zero-sum perturbation h, the local expansion of KL is

$$D_{\mathrm{KL}}(p + th\|p) = \frac{t^2}{2}\sum_a \frac{h(a)^2}{p(a)} + O(t^3).$$

Setting $h = q - p$ proves Eq. (34).

F Numerical evaluation of normalized top-K targets

Path parameterization. Equation (10) implements a shared path coefficient $B > 0$ with $t = Be^B$. A shared B chooses a path point whose effective radius $\delta_B(s) = D_{\mathrm{KL}}(r_k\|p_k)$ varies across prefixes. Alternatively, a scalar search on each prefix enforces a specified radius δ, returning q if it is already feasible. Neither computation requires probabilities outside S after the support has been selected.

Stable Lambert evaluation. For each selected token, let $x_a = \log t + \log q(a) - \log p_k(a)$. Instead of forming e^{x_a}, compute $v_a = \log W_0(e^{x_a})$ by solving

$$v_a + e^{v_a} = x_a. \tag{35}$$

Newton iterations take the form

$$v^{(j+1)} = v^{(j)} - \frac{v^{(j)} + e^{v^{(j)}} - x_a}{1 + e^{v^{(j)}}}.$$

An upper-bound initialization is $v^{(0)} = x_a$ for $x_a \le 0$ and $v^{(0)} = \log(1 + x_a)$ otherwise. The implementation uses 12 iterations and preserves float64 inputs. Retaining $\log W$ avoids underflow when the Lambert argument is extremely small. Set $u_a = \log q(a) - v_a$ and subtract $\mathrm{LSE}_{b\in S}\, u_b$ to obtain normalized $\log r_k$.

For a fixed radius, first compute $D_{\mathrm{KL}}(q\|p_k)$. Return q when this quantity is at most δ, and return p_k for $\delta = 0$. Otherwise bracket $\log t$, expand the upper bracket when necessary, and bisect the monotone radius $D_{\mathrm{KL}}(r_t\|p_k)$ in float64. Near the teacher endpoint the path can approach q slowly, so a fixed narrow bracket is insufficient. Check the achieved radius and normalization before accepting the target, and surface any infeasible result as an error.

G Forward-fitting identities

For a frozen target r and selected logits z,

$$D_{\mathrm{KL}}(r\|p_\theta) = \sum_{a\in S} r_a \log r_a - \sum_{a\in S} r_a z_a + \log \sum_{b\in S} e^{z_b}.$$

Differentiation gives Eq. (8). Selected logits have Hessian $\mathrm{diag}(p_\theta) - p_\theta p_\theta^\top$; the conditional loss depends exclusively on these logits. The nonlinear parameterization determines the curvature of the neural-network objective.

To verify Eq. (22), collect the terms depending on the fitted distribution:

$$-\sum_a (q_a + \lambda p_{k,a}) \log p_{\theta,a} = -(1+\lambda) \sum_a r^{\mathrm{lin}}_{\eta,a} \log p_{\theta,a}.$$

The remaining entropy terms give the stated constant. At the anchor, $p_k - r^{\mathrm{lin}}_\eta = \eta(p_k - q)$, which proves Eq. (23) after applying the logit Jacobian.

H Geometric targets and reverse-KL rescaling

Minimizing $D_{\mathrm{KL}}(r\|q) + \lambda D_{\mathrm{KL}}(r\|p_k)$ on the simplex gives $r_\eta \propto p_k^{1-\eta} q^\eta$, where $\eta = 1/(1+\lambda)$. Writing $\ell = \log(q/p_k)$ yields

$$r_\eta = p_k \frac{e^{\eta\ell}}{\mathbb{E}_{p_k} e^{\eta\ell}} = p_k\{1 + \eta(\ell - \mathbb{E}_{p_k}\ell)\} + O(\eta^2)$$

for fixed positive p_k, q as $\eta \to 0$. Forward fitting, required by the $D_{\mathrm{KL}}(r\|p_k)$ constraint, has anchor logit gradient $p_k - r_\eta$. The direct reverse-KL gradient is $p_k(\mathbb{E}_{p_k}\ell - \ell)$, giving Eq. (20). Uniformity requires bounded likelihood ratios. At $\eta = 1$, for example, $r_\eta = q$ and matching fitting becomes direct forward KL, showing the finite-path departure from reverse-gradient rescaling.

For comparison, if one instead fits the geometric target with reverse KL,

$$D_{\mathrm{KL}}(p_\theta\|r_\eta) = \eta D_{\mathrm{KL}}(p_\theta\|q) + (1-\eta) D_{\mathrm{KL}}(p_\theta\|p_k) + \log Z_\eta. \tag{36}$$

Here p_k, q, η, Z_η are detached. At $p_\theta = p_k$, the gradient of the second term vanishes and the update is exactly η times the direct reverse-KL gradient. This identity concerns a supplementary fitting control. The constraint-consistent geometric formulation in Section 3.2 uses forward fitting. Holding a target fixed for multiple steps also moves the student away from the anchor where the scaling holds.

I GAUSSIAN PROTOCOLS AND ADDITIONAL RESULTS

Four projections. The target is a free probability density on the quadrature grid; the student family has two Gaussian components. The four normalized paths and their KL orientations are listed in Table 1. Each path parameter is solved separately to reach its native budget $\delta = 0.1$; the swapped-PTD path runs from q toward p as t increases. Unrestricted minimization of $D_{\mathrm{KL}}(r\|q)$ gives $r = q$. Reverse-KL mode selection arises through restricted-family fitting or limited optimization. The strictly positive Gaussian densities leave nonzero student density in every region.

Target	Left (%)	Middle (%)	Right (%)	$D_{\mathrm{KL}}(q\|r)$	$D_{\mathrm{KL}}(r\|p)$	$D_{\mathrm{KL}}(p\|r)$
Initial p	49.00	1.00	50.00	1.5133	0.0000	0.0000
Teacher q	45.00	45.00	10.00	0.0000	1.5133	0.8083
PTD	51.48	7.71	40.80	0.5925	0.1000	0.0569
Linear interpolation (native)	47.97	12.28	39.75	0.4176	0.2065	0.1000
Geometric	60.60	5.94	33.47	0.6570	0.1000	0.0788
Swapped PTD	54.41	10.34	35.25	0.4505	0.1751	0.1000
Linear interpolation (matched)	48.36	7.98	43.66	0.5988	0.1000	0.0534

Table 3: Free-target experiment with $p = (0.49, 0.01, 0.50)$ and $q = (0.45, 0.45, 0.10)$ mixture weights on the three fixed components. Region boundaries are -2 and 2; overlapping components make region masses different from mixture weights. The radius-matched linear interpolation coefficient is $\eta = 0.15848665$.

Targets use 3,600 midpoint cells on $[-9, 9]$. Native constraint residuals are below 10^{-10}; fitted KKT multipliers are positive and stationarity residuals on numerically active cells are below 10^{-9}. Repeating with 7,200 cells changes the reported teacher KL and region masses by at most 1.14×10^{-9}.

Target	Round 1	Round 5	Round 20	Middle (%)	GD steps
Geometric	10.9960	1.2654	1.2069	48.0897	36066
Linear interpolation (native)	11.2248	11.1530	11.1394	0.0033	18832
Swapped PTD	11.1913	11.1751	11.1951	0.0032	18304
PTD	8.7876	1.1757	0.3765	38.7832	26669

Table 4: Four native-radius formulations: teacher-forward KL after converged inner optimization, final mass in $[-2, 2)$, and cumulative GD steps. The numerical radius is 0.1, but opposite KL orientations define different feasible sets.

Restricted-family optimization. For the two-component student, optimize $\theta = (\operatorname{logit} w, \mu_a, \mu_b, \log \sigma_a, \log \sigma_b)$ with bounds $[-8, 8]$ on the first three coordinates and $\sigma \in [0.2, 4]$. Projected gradient descent uses maximum step size 0.05 and Armijo backtracking with coefficient 10^{-4}. Each new target is fitted from the previous parameters, without restarting or increasing model capacity. The stopping criterion is $\|\theta - \mathrm{clip}(\theta - \nabla L)\|_2 \le 10^{-6}$. All recorded inner fits meet it and their fitting losses are nonincreasing within 10^{-10} numerical tolerance. We report cumulative GD steps because compute varies across targets under the shared convergence tolerance.

The fixed-target experiment uses 6,400 midpoint cells on $[-32, 32]$ and extends the same trajectories to 5,000 steps. A 12,800-cell endpoint check changes key metrics by at most 1.45×10^{-7}. The repeated-target experiment uses 1,600 cells and verifies each endpoint and reconstructed target on 6,400 cells; the maximum KL/gradient-norm difference is 5.84×10^{-9}. This error bound covers KL values and gradient norms; region-mass quadrature also depends on boundary alignment.

All reported trajectories share the initially undercovered student and the same student geometry. The stopping rule verifies stationarity of each inner fit along these trajectories.

Target	Fit	100 steps	500	1,000	5,000
PTD	forward	8.8104	8.7876	8.7876	8.7876
PTD	reverse	11.2357	11.2292	11.2292	11.2292
Linear interpolation (native)	forward	1.1874	1.0822	1.0802	1.0802
Linear interpolation (native)	reverse	11.2319	11.2249	11.2248	11.2248
Geometric	forward	11.0034	10.9959	10.9960	10.9960
Geometric	reverse	11.1733	11.1717	11.1718	11.1718
Swapped PTD	forward	1.2761	1.1742	1.1712	1.1712
Swapped PTD	reverse	11.2033	11.1913	11.1913	11.1913
Direct teacher	forward	0.8478	0.7094	0.7039	0.7005
Direct teacher	reverse	11.1647	11.2060	11.2084	11.2084
Linear interpolation (matched)	forward	8.8363	8.8227	8.8226	8.8226
Linear interpolation (matched)	reverse	11.2579	11.2574	11.2574	11.2574

Table 5: Fixed initial target: teacher KL $D_{\mathrm{KL}}(q\|p_\theta)$ as fitting continues. More steps toward one target are different from refreshing it. Direct teacher-forward is 0.7005 at 5,000 steps and reaches its stationary value 0.3708 after 7,673 steps.

Target	Fit	Round 1	Round 5	Round 20	GD steps
PTD	forward	8.7876	1.1757	0.3765	26669
PTD	reverse	11.2292	11.1629	11.1501	18509
Linear interpolation (native)	forward	1.0802	0.3776	0.3708	18679
Linear interpolation (native)	reverse	11.2248	11.1530	11.1394	18832
Geometric	forward	10.9960	1.2654	1.2069	36066
Geometric	reverse	11.1718	11.2083	11.2084	7240
Swapped PTD	forward	1.1712	0.4176	0.3938	28197
Swapped PTD	reverse	11.1913	11.1751	11.1951	18304
Linear interpolation (matched)	forward	8.8226	2.0227	0.6392	31818
Linear interpolation (matched)	reverse	11.2574	11.2443	11.1995	12031
Direct teacher	forward	0.3708	0.3708	0.3708	7673
Direct teacher	reverse	11.2084	11.2084	11.2084	1565

Table 6: Supplementary fitting controls: all four target geometries, matched-radius linear interpolation, and the direct-teacher reference, each with both fitting directions. Fitting directions are varied independently here. Entries are teacher-forward KL; native budgets in opposite directions differ.

J LLM EVALUATION PROTOCOL AND REPRODUCIBILITY

J.1 TRAINING CONFIGURATION

Setting	Value
Objective support	Separately normalized teacher top-32; sampled: rollout token
Optimizer / learning rate	AdamW / 2×10^{-6}, constant
Batch / rollouts per prompt	32 prompts / 4 responses
Training input / response limits	1,024 / 7,168 tokens
Final step / training seeds	Math-only: 120; routed multi-domain: 120 / one (42)
PTD coefficient / linear interpolation weight	$B = -\log(0.8)$ / $\eta = 0.2$

Table 7: Core training settings. Target radii and realized student KL depend on the method.

J.2 EVALUATION SUMMARIES AND UNCERTAINTY

Bootstrap protocol. We use a paired nonparametric item bootstrap (Efron & Tibshirani, 1993). We first reduce each question to a binary best-of-n success indicator at the fixed primary budget: 16 draws for AIME, four for AMC, and four for GPQA. Each of 200,000 nonparametric bootstrap replicates resamples questions with replacement independently within each benchmark. The same resampled question indices are applied to every model, preserving the pairing induced by the shared test set. For each replicate, Math equally averages the AIME24, AIME25, and AMC23 benchmark means; Macro then equally averages Math and GPQA. The intervals in Table 2 are the 2.5th and 97.5th percentiles of the absolute-score replicates. Table 8 subtracts sampled reverse KL from each method within every replicate, thereby preserving paired question outcomes. We use seed 2027 and NumPy's linear percentile interpolation. These intervals condition on the evaluated models, fixed generations, and grading pipeline, and quantify benchmark-item sampling uncertainty. Training seeds and new generations form separate sources of variation.

Setting	Method	ΔMath [95% CI]	ΔGPQA	ΔMacro [95% CI]
Math-only	Top-K reverse KL	+2.22 [-2.50, +7.22]	+1.01	+1.62 [-2.90, +6.19]
Math-only	LI	+5.00 [-1.11, +11.39]	+2.53	+3.76 [-1.10, +8.72]
Math-only	PTD	+6.39 [+1.94, +11.67]	+9.60	+7.99 [+3.80, +12.32]
Routed	Top-K reverse KL	+2.78 [-3.33, +9.17]	+1.01	+1.89 [-2.87, +6.69]
Routed	LI	+5.56 [+0.83, +10.83]	+4.04	+4.80 [+0.05, +9.63]
Routed	PTD	+5.83 [+0.00, +11.94]	+5.05	+5.44 [+0.58, +10.37]

Table 8: Best-of-n success gains over sampled reverse KL in percentage points. Math and Macro report paired 95% item-bootstrap intervals (200,000 replicates; seed 2027); GPQA reports the point estimate. AIME uses fixed draws 0–15, AMC uses 0–3, and GPQA uses all four draws. Each difference uses the sampled reverse-KL reference for its displayed group.

Method	Macro p@1	Macro p@2	Macro p@4	Macro b@n	Math b@n	GPQA b@4
Top-K reverse KL	35.22	46.52	59.05	62.26	66.94	57.58
LI	37.57	49.30	61.19	65.16	69.72	60.61
PTD (ours)	36.15	48.05	61.42	65.81	70.00	61.62

Table 9: Routed sampling-budget sensitivity. Macro p@k averages Math and GPQA pass@k; Macro b@n uses the primary mixed budgets (AIME b@16, AMC b@4, and GPQA b@4). These metrics report correct-answer coverage across sampling budgets.

J.3 Data and evaluation protocol

Models and data. All runs use DeepSeek-R1-Distill-Qwen-1.5B as the student. The routed dataset contains 640 training prompts each from math and science. Math prompts use Skywork-OR1-Math-7B, and science prompts use Skywork-OR1-7B. Prompts longer than 1,024 student-template tokens are excluded, and domain-homogeneous batches are alternated deterministically. Benchmark strings are filtered during preparation.

Generation and grading. All models use the same benchmark questions, sampling budgets, and deterministic per-response seed construction. Math responses are scored with Math-Verify (Kydlíček, 2026), and GPQA responses use option-based final-answer grading. Invalid outputs count as incorrect, and aggregation requires a resolved grade for every reported response.

K Limitations and scope

The language-model evidence uses one seed and 120 updates per setting. Repeated responses reduce decoding noise, while the bootstrap covers benchmark-item uncertainty. Best-of-n measures correct-answer coverage at a fixed sampling budget and incorporates correctness, formatting, and truncation; raw generations support complementary lexical and semantic diversity analyses.

The theory assumes fixed teacher support, teacher logits, and a shared vocabulary. Its guarantees concern conditional probability on that support. The Gaussian study directly measures recovered probability mass, while the language-model study evaluates its downstream consequence through multi-sample correct-answer coverage.