MI-SD: MUTUAL-INFORMATION-GUIDED MULTI-REFERENCE ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation improves LLM reasoning by aligning supervision with student-generated trajectories. However, single-reference supervision can reduce trajectory diversity and neglect alternative valid solution paths, particularly in complex reasoning tasks. We introduce **MI-SD**, a multi-reference on-policy self-distillation framework that supervises student rollouts using multiple valid reference solutions. A shared frozen teacher evaluates each student rollout under distinct privileged reference contexts and provides corresponding supervision. What's more, MI-SD can quantify token-level reference sensitivity using the generalized Jensen–Shannon divergence across teacher distributions. At positions where references provide distinct guidance, MI-SD applies MI-gated max-support supervision. We further introduce a canonical-solution anchor to stabilize training. Across Qwen3 models with 1.7B, 4B, and 8B parameters, MI-SD consistently improves average accuracy on AIME24, AIME25, and HMMT25 over OPSD by 1.3, 0.7 and 1.4 percentage points, respectively. These results demonstrate the benefit of structured multi-reference supervision for on-policy self-distillation across the evaluated model scales. Our code will be made publicly available soon.

1 INTRODUCTION

Post-training has emerged as a central lever for improving the reasoning capability of large language models (LLMs) (OpenAI, 2025; Meta AI, 2025; Yang et al., 2025; Guo et al., 2025). The dominant paradigms fall into three families: supervised fine-tuning(Zelikman et al., 2022; Yu et al., 2024), reinforcement learning from verifiable rewards (RLVR) (Shao et al., 2024; Guo et al., 2025), and knowledge distillation (Hinton et al., 2015). Among these, knowledge distillation provides dense per-token supervision that complements the sparse outcome signals of RLVR, and thus becomes more popular recently. As a representative knowledge distillation method, *on-policy* distillation (OPD) aims to supervise the student on trajectories sampled from its own current policy (Lu & Lab, 2025; Agarwal et al., 2024; Gu et al., 2024). It mitigates the distribution mismatch inherent in traditional offline distillation. However, due to its full vocabulary-level supervision, OPD requires a teacher with an accessible vocabulary, such as an open-weight model. This constraint excludes API-only teachers and limits the applicability of OPD in real-world scenarios. On-policy self-distillation (OPSD) addresses this limitation by using the same model in two asymmetric roles. In the student role, the model generates a solution given only the problem. In the teacher role, a frozen copy instantiated from the same base model evaluates that trajectory with access to a privileged reference solution (Zhao et al., 2026a). This design eliminates the need for an external teacher while preserving the on-policy benefit of aligning supervision with the student's own rollout distribution.

As self-distillation training progresses on reasoning tasks, a concerning phenomenon emerges. Models tend to produce increasingly confident and concise responses, suppressing epistemic hedges, shortening reasoning chains, and taking more direct paths to final answers (Kim et al., 2026). This collapse in response diversity can be detrimental for complex problems in mathematics, logic, and science, where extended deliberation, alternative-route exploration, and explicit verification are often necessary. In OPSD, since supervision is conditioned on a single reference solution, the teacher consistently reinforces one reasoning perspective. It offers no signal from alternative valid routes, thus further amplifying this homogenization.

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.

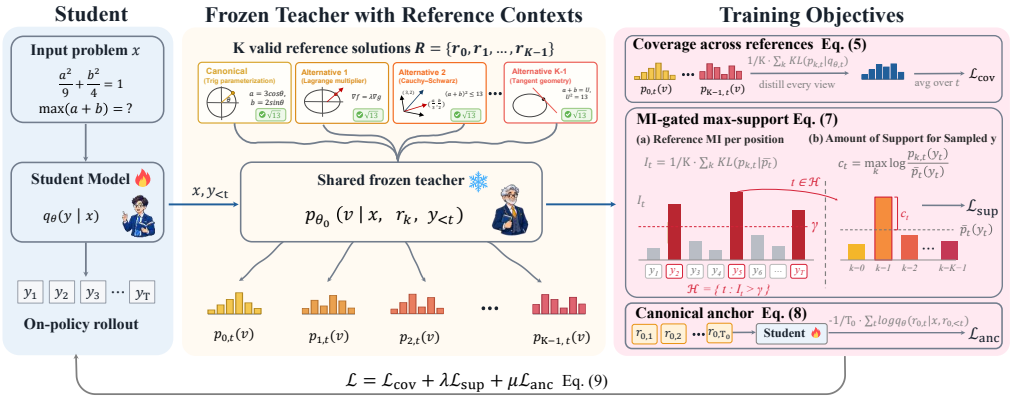

Figure 1: **MI-SD method overview.** The student receives only the problem and generates an on-policy rollout. A shared frozen teacher then evaluates the rollout under K privileged reference contexts. Their generalized Jensen–Shannon divergence yields a token-level mutual-information score for reference-sensitive positions. The MI score selects positions for max-support supervision, complementing multi-reference coverage at all rollout positions and a canonical-solution anchor.

This observation motivates two questions. First, can multiple valid solution routes serve as complementary privileged supervision? Second, can the mutual-information between the reference and the teacher's next token guide better distillation? Mathematical problems often admit multiple solution paths (Hendrycks et al., 2021; Lewkowycz et al., 2022), including different algebraic manipulations, geometric constructions, or proof strategies that lead to the same answer. Conditioning the teacher on only one reference restricts supervision to a single privileged perspective and leaves information from alternative routes unused. Moreover, the student and teacher operate under asymmetric information: the student cannot access the privileged context available to the teacher. This asymmetry induces an irreducible mutual-information gap in the distillation loss (Yang et al., 2026). A single-reference teacher does not provide a token-level mechanism to identify or exploit this gap.

We introduce **MI-SD** (Mutual-Information-guided Self-Distillation), a multi-reference extension of on-policy self-distillation. Given a student rollout generated from the problem alone, a shared frozen teacher evaluates that rollout under K reference contexts, each corresponding to a valid solution route. These K reference-conditioned teacher distributions provide diverse but coherent supervision without requiring additional model parameters. Furthermore, we can identify *reference-sensitive* positions based on the conditional mutual information (MI) between reference solutions and the predicted next-token distribution. This token-level MI signal drives an adaptive supervision objective. All positions receive multi-reference coverage through the uniform mixture of the K teacher distributions. At reference-sensitive positions, where the MI exceeds a threshold, an additional max-support term emphasizes tokens that are strongly supported by at least one teacher. A lightweight canonical-solution anchor is further included to stabilize training. We evaluate MI-SD on Qwen3-1.7B, 4B, and 8B across three mathematical reasoning benchmarks, namely AIME24, AIME25, and HMMT25. With $K = 4$ reference teachers, MI-SD improves average accuracy over OPSD by 1.3, 0.7, and 1.4 percentage points at three model scales.

The main contributions of this paper can be summarized as follows:

- We introduce multi-reference on-policy self-distillation for LLM reasoning, in which a shared frozen teacher evaluates each student rollout under K reference contexts. This design extends conventional single-reference OPSD with multi-perspective reasoning supervision, mitigating its tendency toward overconfident and homogeneous solution trajectories.
- We propose a reference-conditioned mutual-information score, computed as equal-weight generalized JSD over selected reference slots. This signal unifies multi-reference coverage, MI-gated max-support supervision, and canonical-solution anchor in a single training framework.
- We ran comprehensive experiments across three Qwen3 model scales. Results show consistent gains from structured multi-reference supervision and isolate effects of coverage, MI-gated max-support, and canonical anchoring via matched-data and exposure-matched comparisons.

Table 1: Comparison of representative training paradigms in their standard formulations. Multi-view references denote distinct valid solution routes for the same problem; MI guidance denotes token-level reference-conditioned mutual information used to allocate supervision.

Method	Sampling	Signal	Teacher	Multi-view References	MI Guidance
Distillation	× off-policy	✓ dense	× external	—	× none
On-policy distillation	✓ on-policy	✓ dense	× external	—	× none
RLVR (e.g., GRPO)	✓ on-policy	× sparse	—	—	× none
OPSD	✓ on-policy	✓ dense	✓ self	× single	× none
MI-SD (ours)	✓ on-policy	✓ dense	✓ self	✓ multiple	✓ MI-gated

2 Related Work

Distillation with privileged information. Knowledge distillation transfers soft predictive distributions from a teacher to a student, providing richer supervision than hard labels alone (Hinton et al., 2015). Generalized distillation connects this framework to learning under privileged information, where auxiliary information is available during training but withheld at inference (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016). In language models, such privileged information can condition the teacher while the deployed student remains unconditioned, including settings where the teacher observes solution traces or action trajectories unavailable at test time (Penaloza et al., 2026). A complementary line of work addresses the distribution mismatch between offline distillation and autoregressive inference. On-policy distillation (OPD) evaluates teacher and student distributions on trajectories sampled from the current student policy, providing dense token-level supervision on the states the student actually visits (Lu & Lab, 2025; Agarwal et al., 2024; Gu et al., 2024). However, standard OPD requires continued access to a separate teacher model. On-policy self-distillation (OPSD) removes this requirement by assigning teacher and student roles to the same base model, with the teacher gaining an informational advantage through privileged conditioning rather than additional capacity (Furlanello et al., 2018; Zhao et al., 2026a). The student generates rollouts from the problem alone, while the teacher evaluates the same trajectories. MI-SD follows this self-distillation paradigm but extends the teacher from a single privileged solution to multiple valid solution views.

Verifiable rewards and adaptive token-level supervision. Reinforcement learning from verifiable rewards (RLVR), including GRPO-based training and DeepSeek-R1 (Shao et al., 2024; Guo et al., 2025), improves reasoning policies using outcome-level correctness signals without requiring human preference annotation. Its supervision is nevertheless sparse: a final reward indicates whether a trajectory succeeds but provides limited guidance about which intermediate token decisions should change. Distillation can complement RLVR with dense per-token targets, and recent work has shown benefits from combining the two paradigms (Li et al., 2026). Several methods further adapt the strength or form of this token-level supervision. Entropy-aware OPD (Jin et al., 2026a) modulates distillation according to teacher uncertainty; RLCSD (Pan et al., 2026) contrasts correct and incorrect privileged hints to reduce privilege-induced bias; DASH (Hou et al., 2026) adjusts supervision horizons using the evolution of teacher–student divergence; and UniSD (Jin et al., 2026b) combines multiple teachers, EMA stabilization, and token-level contrastive learning. These approaches primarily improve how supervision is weighted or stabilized given a teacher context. MI-SD instead asks a complementary question: when multiple valid privileged solutions are available, which tokens are actually sensitive to the choice of view and therefore benefit most from multi-reference supervision?

Multi-reference and multi-view privileged conditioning. Recent self-distillation methods have begun to exploit variation across privileged contexts. AVSD (Nguyen et al., 2026) decomposes cross-view predictions into shared consensus and view-specific residuals; MOPD (Yu et al., 2026a) uses successful and failed peer rollouts from the same prompt to exploit within-batch solution diversity; and SGSD (Huang et al., 2026) constructs a pool of teachers conditioned on retrieved skill–mistake pairs and gates their contributions using verifier feedback. Related studies on teacher exposure further show that the content and granularity of privileged information can materially affect self-distillation quality (Han et al., 2026; Zhao et al., 2026b; Ichihara et al., 2026). MI-SD is closely related to this line of multi-view self-distillation but differs in its treatment of privileged contexts. Instead of relying on separate teachers, peer routing, or explicit consensus–residual decomposition,

it conditions a single frozen teacher on multiple valid solution references for the same student trajectory. The resulting cross-reference variation is used to determine where multi-view supervision is most informative, while the student itself requires no privileged information at inference.

These three lines of work highlight complementary design dimensions. Conventional distillation provides dense supervision but is typically off-policy and teacher-dependent; OPD makes supervision on-policy, while OPSD replaces the external teacher with privileged conditioning. RLVR instead provides sparse outcome supervision, and adaptive distillation refines token-level supervision without explicitly modeling multiple valid solution routes. Multi-view methods incorporate richer privileged contexts but do not use reference-conditioned mutual information to allocate token-level supervision. Table 1 summarizes these distinctions and situates MI-SD as combining on-policy sampling, dense self-distillation, multi-view references, and MI-guided supervision.

3 METHOD

The proposed MI-SD uses mutual information (MI) to guide on-policy self-distillation from multiple valid solution references. A shared frozen teacher evaluates each student rollout under every reference, and MI identifies positions where its predictions depend strongly on the reference. At these positions, we reinforce sampled tokens using their maximum pointwise MI across references. This selective reinforcement complements a coverage loss over all reference views and a supervised anchor on the canonical solution.

3.1 PROBLEM FORMULATION

Let $\mathcal{D}$ be a dataset in which each problem x is paired with K valid reference solutions, $\mathcal{R} = \{r_0, r_1, \ldots, r_{K-1}\}$. Here r_0 is the canonical solution, and the remaining references provide alternative solution routes. The student model q_θ conditions only on x, and samples an on-policy trajectory $y = (y_1, \ldots, y_T)$ of length T: $y \sim q_\theta(\cdot \mid x)$. Its next-token distribution at position t is

$$q_{\theta,t}(v) = q_\theta(v \mid x, y_{<t}), \tag{1}$$

where $y_{<t} = (y_1, \ldots, y_{t-1})$ is the sampled prefix and v denotes a vocabulary token.

The teacher is a frozen base model p_{θ_0} that additionally receives a reference r_k:

$$p_{k,t}(v) = p_{\theta_0}(v \mid x, r_k, y_{<t}). \tag{2}$$

Thus, all K teacher distributions evaluate the same student-generated prefix under different reference contexts. During optimization, sampled trajectories are held fixed, and gradients flow only through the student. Unless otherwise specified, expectations are over $(x, \mathcal{R}) \sim \mathcal{D}$ and $y \sim q_\theta(\cdot \mid x)$.

3.2 MUTUAL INFORMATION ACROSS REFERENCES

Different references may lead the teacher to favor different continuations of the same student prefix. We quantify this reference dependence using MI to guide the allocation of additional supervision. For a fixed problem, reference set, and prefix, we first assign equal weight to all references and form their mean teacher distribution:

$$\bar{p}_t(v) = \frac{1}{K} \sum_{k=0}^{K-1} p_{k,t}(v). \tag{3}$$

Using $\bar{p}_t(v)$ as a baseline, we measure how much reference-conditioned predictions differ from it:

$$I_t = \frac{1}{K} \sum_{k=0}^{K-1} \mathrm{KL}(p_{k,t} \,\|\, \bar{p}_t) = \frac{1}{K} \sum_{k=0}^{K-1} \sum_{v} p_{k,t}(v) \log \frac{p_{k,t}(v)}{\bar{p}_t(v)}. \tag{4}$$

Under the uniform reference prior, I_t is the mutual information between the reference index and the teacher's next token, equivalently the equal-weight generalized Jensen–Shannon divergence among the teacher distributions. A larger I_t therefore indicates stronger reference dependence, while $I_t = 0$ when all reference-conditioned distributions coincide.

This dependence guides where we apply additional supervision. At high-MI positions, averaging may dilute support for continuations favored by a subset of references. We therefore use I_t to select positions where max-support reinforcement supplements the averaged distillation signal.

Algorithm 1 Training of the Proposed MI-SD. Clipping, microbatch thresholding, and loss reductions are specified in Appendix B.

Require: Batch $\mathcal{B} = \{(x, \mathcal{R})\}$; student q_θ; frozen teacher p_{θ_0}; threshold γ; coefficients λ, μ; optional token-filtering rule

1: **for** $(x, \mathcal{R}) \in \mathcal{B}$ **do**
2: Sample on-policy rollout $y = (y_1, \ldots, y_T) \sim q_\theta(\cdot \mid x)$
3: Evaluate $q_{\theta,t}$ and all reference-conditioned teacher distributions $\{p_{k,t}\}_{k=0}^{K-1}$ along y
4: Compute $\bar{p}_t$ and I_t using Eqs. (3–4)
5: Construct m_t using the optional filtering rule; set $m_t = 1$ if disabled
6: $\mathcal{H} \leftarrow \{t \in \{1, \ldots, T\} : I_t > \gamma,\ m_t = 1\}$
7: $c_t \leftarrow \max_k \log \frac{p_{k,t}(y_t)}{\bar{p}_t(y_t)}$ for $t \in \mathcal{H}$
8: Compute $\mathcal{L}_{\text{cov}}$, $\mathcal{L}_{\text{sup}}$, and $\mathcal{L}_{\text{anc}}$ from the per-example terms in Eqs. (5, 7, 8), with $\mathcal{L}_{\text{sup}} = 0$ if $\mathcal{H} = \varnothing$
9: $\mathcal{L}(x, \mathcal{R}) \leftarrow \mathcal{L}_{\text{cov}} + \lambda \mathcal{L}_{\text{sup}} + \mu \mathcal{L}_{\text{anc}}$
10: **end for**
11: $\mathcal{L}_{\mathcal{B}} \leftarrow \widehat{\mathcal{L}}_{\text{cov}}(\mathcal{B}) + \frac{1}{|\mathcal{B}|} \sum_{(x,\mathcal{R}) \in \mathcal{B}} [\lambda \mathcal{L}_{\text{sup}}(x, \mathcal{R}) + \mu \mathcal{L}_{\text{anc}}(x, \mathcal{R})]$
12: Update the student model θ using $\nabla_\theta \mathcal{L}_{\mathcal{B}}$, keep the teacher fixed

3.3 MI-GUIDED DISTILLATION

Coverage across references. We first provide supervision at every rollout position by averaging the teacher-to-student forward KL divergence across references:

$$\mathcal{L}_{\text{cov}} = \mathbb{E}\left[\frac{1}{TK} \sum_{t=1}^{T} \sum_{k=0}^{K-1} \text{KL}(p_{k,t} \,\|\, q_{\theta,t})\right]. \tag{5}$$

This coverage term incorporates all reference views. Clipping and batch reduction are detailed in Appendix B. We then supplement it with MI-guided reinforcement to retain reference-specific support that averaging may weaken.

Max-support weighting. Under the joint distribution defined above, the pointwise MI between reference k and token v is $\log[p_{k,t}(v)/\bar{p}_t(v)]$. For the student's sampled token y_t, we take its maximum across references:

$$c_t = \max_k \log \frac{p_{k,t}(y_t)}{\bar{p}_t(y_t)} = \log \frac{\max_k p_{k,t}(y_t)}{\bar{p}_t(y_t)}. \tag{6}$$

The contrast c_t measures the sampled token's strongest reference support relative to the mean. It complements I_t, which averages pointwise MI over references and tokens to identify positions where predictions depend on the reference. At these positions, c_t takes the maximum pointwise MI across references, giving greater reinforcement to tokens with a larger support advantage.

MI-gated reinforcement. Given an MI threshold γ, we retain positions with $I_t > \gamma$. An optional mask $m_t \in \{0, 1\}$ excludes predefined style and surface-form tokens; when disabled, $m_t = 1$ at every valid position. The retained set is $\mathcal{H} = \{t \in \{1, \ldots, T\} : I_t > \gamma,\ m_t = 1\}$. We then define

$$\mathcal{L}_{\text{sup}} = -\mathbb{E}\left[\frac{1}{|\mathcal{H}|} \sum_{t \in \mathcal{H}} c_t \log q_{\theta,t}(y_t)\right], \tag{7}$$

with the per-trajectory loss set to zero when $\mathcal{H} = \varnothing$. This objective increases the probability of sampled tokens with greater relative support from at least one reference, specifically at positions where MI indicates reference-dependent predictions.

3.4 TRAINING OBJECTIVE

Canonical anchor. The coverage and max-support losses supervise the student on its own generated prefixes. We complement this supervision by training directly on the canonical solution r_0, which

provides a complete valid trajectory:

$$\mathcal{L}_{\text{anc}} = -\mathbb{E}_{(x,\mathcal{R})\sim\mathcal{D}}\left[\frac{1}{T_0}\sum_{t=1}^{T_0}\log q_\theta(r_{0,t} \mid x, r_{0,<t})\right], \tag{8}$$

where T_0 is the non-padding length of r_0. This anchor provides a fixed reference target when teacher predictions on the sampled trajectory are ambiguous.

Combined objective. We combine the canonical anchor with the coverage and MI-guided max-support losses:

$$\mathcal{L} = \mathcal{L}_{\text{cov}} + \lambda\mathcal{L}_{\text{sup}} + \mu\mathcal{L}_{\text{anc}}, \tag{9}$$

where $\lambda, \mu \geq 0$ control the contributions of max-support reinforcement and the canonical anchor. Within $\mathcal{L}_{\text{sup}}$, the MI score I_t selects positions for additional supervision, and c_t determines the reinforcement weight at each selected position. We optimize the student's LoRA parameters, denoted by θ, while keeping the teacher fixed. Algorithm 1 summarizes the training procedure.

4 Experiments

4.1 Setup

Backbones, Baselines, and Benchmarks. We evaluate the proposed method based on the instruct-tuned Qwen3 models (Yang et al., 2025) at 1.7B, 4B, and 8B parameters on AIME24, AIME25, and HMMT25. The evaluation metric is Avg@12, the fraction of correct generations among 12 independently sampled responses per problem. The overall score is the macro-average across the three benchmarks. We also evaluate MI-SD$_{\text{cov}}$, a coverage-only variant that optimizes $\mathcal{L}_{\text{cov}}$ alone, without MI-gated max-support reinforcement or the canonical anchor. For both variants, we fine-tune the base model with LoRA (rank 64, $\alpha = 128$) using AdamW with a learning rate of 5×10^{-6} and an effective batch size of 32. Objective-specific settings are listed in Appendix B. For both methods, we report the highest score obtained on each benchmark among checkpoints saved at steps 20, 40, 60, 80, and 100. Checkpoint selection is performed independently for each benchmark. The baselines are SFT (Ouyang et al., 2022), GRPO (Shao et al., 2024), PBSD (Yu et al., 2026b), OPSD (Zhao et al., 2026a), SDFT (Shenfeld et al., 2026), SRPO (Li et al., 2026), and SGSD (Huang et al., 2026). SGSD is the closest multi-teacher counterpart to MI-SD. We train it using the same corpus, LoRA and optimizer configurations, and checkpoint-selection protocol as MI-SD.

Training-data construction. Training uses an OpenThoughts-derived corpus of mathematics problems. Unless otherwise specified, we use $K = 4$ reference contexts for both MI-SD and MI-SD$_{\text{cov}}$. Each retained record provides one canonical solution and three alternative solution texts. To encourage variation across references, candidate generation varies the requested reasoning perspective and presentation style. Candidates are subsequently expanded with instructions to preserve their main solution methods, and screened using final-answer agreement, length constraints, lexical-overlap criteria, and verifier assessments of correctness and method preservation. Dataset statistics and construction details are reported in Appendix E.

4.2 Main results

Table 2 compares OPSD with the proposed multi-reference variants across three model scales. MI-SD improves the three-benchmark macro-average over OPSD by 1.3, 0.7, and 1.4 percentage points for Qwen3-1.7B, 4B, and 8B, respectively. The gains are consistent across scales but vary in magnitude: they are largest at 8B, smaller at 4B, and increase again at 1.7B. At 1.7B, MI-SD improves AIME24, AIME25, and HMMT25 by 1.7, 1.7, and 0.5 points, while the corresponding gains at 8B are 1.1, 1.1, and 2.0 points. The contribution of the auxiliary objectives is scale dependent. At 1.7B, the full MI-SD objective raises from 42.4 to 44.7, outperforming the coverage-only variant by 2.3 points. In contrast, at 4B and 8B, MI-SD$_{\text{cov}}$ exceeds the full objective by 0.2 points.

These results suggest that the auxiliary objectives are most beneficial for smaller students under the evaluated setting. One possible explanation is that smaller models have greater difficulty reconciling competing reference-conditioned teacher distributions. When different references favor distinct solution modes, uniform multi-reference supervision can produce a diffuse learning signal. The max-support term may sharpen supervision at reference-sensitive positions, while the canonical anchor

Table 2: Main results (Avg@12, %). For MI-SD, MI-SD$_{\text{cov}}$, and SGSD, we select the best checkpoint per benchmark; results taken from prior work follow their original training and checkpoint-selection protocols. **Bold**: best; underline: second best.

Model	Method	AIME24	AIME25	HMMT25	Average
Qwen3-8B	Base (Yang et al., 2025)	75.8	65.6	43.9	61.8
	SFT (Ouyang et al., 2022)	72.3	64.2	42.9	59.8
	GRPO (Shao et al., 2024)	76.4	68.9	46.7	64.0
	OPSD (Zhao et al., 2026a)	77.8	70.8	45.8	64.8
	SDFT (Yu et al., 2026b)	75.9	69.1	44.2	63.1
	SRPO (Yu et al., 2026b)	77.1	69.6	43.9	63.5
	SGSD (Huang et al., 2026)	77.2	69.7	47.2	64.7
	PBSD (Yu et al., 2026b)	78.4	71.0	46.1	65.2
	MI-SD$_{\text{cov}}$	78.6	**72.8**	**47.8**	**66.4**
	MI-SD	**78.9**	71.9	**47.8**	66.2
Qwen3-4B	Base (Yang et al., 2025)	74.9	66.4	42.2	61.2
	SFT (Ouyang et al., 2022)	70.2	62.3	43.4	58.6
	GRPO (Shao et al., 2024)	75.6	68.1	44.4	62.7
	OPSD (Zhao et al., 2026a)	76.4	68.3	46.1	63.6
	SDFT (Yu et al., 2026b)	75.6	67.1	44.7	62.5
	SRPO (Yu et al., 2026b)	76.3	67.8	45.5	63.2
	SGSD (Huang et al., 2026)	75.8	69.2	45.3	63.4
	PBSD (Yu et al., 2026b)	77.3	69.0	45.6	64.0
	MI-SD$_{\text{cov}}$	76.7	**70.0**	**46.9**	**64.5**
	MI-SD	**77.8**	69.2	45.8	64.3
Qwen3-1.7B	Base (Yang et al., 2025)	51.5	36.7	23.1	37.1
	SFT (Ouyang et al., 2022)	48.4	36.3	22.7	35.8
	GRPO (Shao et al., 2024)	51.1	38.3	23.7	37.7
	OPSD (Zhao et al., 2026a)	57.2	43.9	29.2	43.4
	SDFT (Yu et al., 2026b)	57.0	43.0	28.7	42.9
	SRPO (Yu et al., 2026b)	58.1	44.1	29.4	43.9
	SGSD (Huang et al., 2026)	58.6	42.5	29.0	43.4
	PBSD (Yu et al., 2026b)	58.5	44.4	30.0	44.3
	MI-SD$_{\text{cov}}$	55.8	40.8	**30.6**	42.4
	MI-SD	**58.9**	**45.6**	29.7	**44.7**

provides a stable reference trajectory. Larger students may already exploit the soft multi-reference signal effectively, reducing the marginal benefit of these auxiliary terms.

4.3 Discussion

To assess the effectiveness of MI-SD and MI-SD$_{\text{cov}}$, we perform a comprehensive set of experiments and conduct ablation studies to address the following five key questions:

Question 1: How do max-support reinforcement and the canonical anchor contribute to performance? To examine the roles of the two auxiliary terms, we compare the full MI-SD objective with variants that remove either max-support or the canonical anchor, while retaining multi-reference coverage. Figure 2 shows that the full objective achieves a higher macro Avg@12 than either two-component variant at all three scales. Comparison with coverage-only training in Table 2 reveals that the auxiliary terms are most effective when used together at 1.7B: adding either term alone lowers macro Avg@12, whereas combining them improves it by 2.3 percentage points. This suggests complementarity between reference-sensitive token reinforcement and supervision on a complete canonical solution, with neither auxiliary term alone accounting for the gain. At 4B and 8B, however, coverage-only training slightly outperforms the full objective, even though removing either auxiliary term individually reduces performance (Appendix C). Thus, the single-removal ablations show that each term improves the objective when the other is present, while the coverage-only comparison shows that adding both is beneficial only at 1.7B under the evaluated configurations.

Question 2: How does the number of reference contexts affect performance? To examine the effect of reference count, we compare $K \in \{2, 3, 4\}$ on Qwen3-8B. Table 3 shows that macro Avg@12 increases monotonically with K for both objectives. From $K = 2$ to $K = 4$, MI-SD improves from 65.5% to 66.2%, while MI-SD$_{\text{cov}}$ improves from 65.2% to 66.4%. Even with two references, both methods exceed the reported OPSD average of 64.8% (Zhao et al., 2026a). Coverage-only training benefits more from additional references. Since its supervision relies on the averaged teacher distribution, additional references can enrich the target with information from alternative solution paths. Full MI-SD also uses max-support to strengthen signals supported by individual references. This

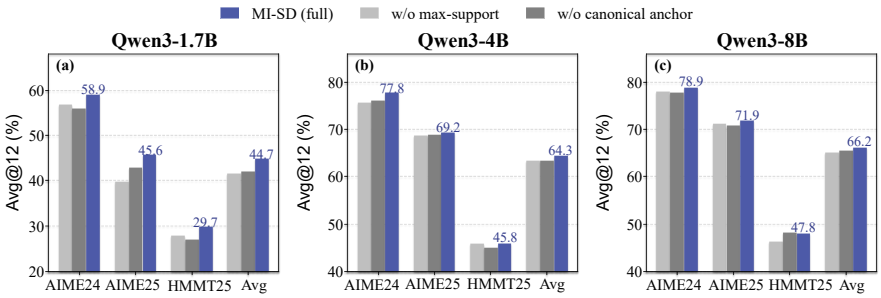

Figure 2: **Component ablation across scales.** Avg@12 (%) on three benchmarks and their macro average. *w/o max-support* removes MI-gated token reinforcement; *w/o canonical anchor* removes supervised training on the canonical solution.

Table 3: K ablation (Qwen3-8B, Avg@12, %).

Method	AIME24	AIME25	HMMT25	Avg
OPSD	77.8	70.8	45.8	64.8
MI-SD$_{\text{cov}}$ (K=2)	78.6	70.0	46.9	65.2
MI-SD$_{\text{cov}}$ (K=3)	78.6	71.1	47.5	65.7
MI-SD$_{\text{cov}}$ (K=4)	78.6	72.8	47.8	66.4
MI-SD (K=2)	76.7	70.6	49.2	65.5
MI-SD (K=3)	78.1	71.1	48.1	65.7
MI-SD (K=4)	78.9	71.9	47.8	66.2

Table 4: γ ablation (Qwen3-1.7B, Avg@12, %).

γ	AIME24	AIME25	HMMT25	Avg
0.05 (ours)	58.9	45.6	29.7	44.7
0.10	55.8	42.8	29.2	42.6
0.15	56.9	41.1	28.6	42.2
0.20	56.1	43.3	27.8	42.4
0.25	56.4	41.7	26.4	41.5
q85	55.8	43.1	28.3	42.4
q90	57.8	41.7	28.9	42.8
q95	56.7	43.9	28.6	43.1

may provide sharper guidance when the reference pool is small, leaving less room for improvement as more references are introduced. Further details are provided in Appendix C.2.

Question 3: How sensitive is MI-SD to the MI threshold? We investigate threshold sensitivity on Qwen3-1.7B using both absolute and percentile-based MI thresholds. All tested alternatives underperform the default $\gamma = 0.05$ on every benchmark, reducing macro Avg@12 by 1.6–3.2 percentage points (Table 4). Under either thresholding scheme, a stricter cutoff concentrates max-support reinforcement on more reference-sensitive positions but may exclude useful supervision at moderately sensitive ones. The excluded positions still receive coverage supervision, but their reference-specific support may be diluted by averaging without the additional reinforcement. For the evaluated 1.7B configuration, the default $\gamma = 0.05$ provides the best performance among the tested choices.

Question 4: Do the gains persist when reference access and exposure are controlled? We construct OPSD-flat, a single-reference baseline using the same corpus and four-reference pool. It cycles through the four references, using one per problem per epoch and retaining the original OPSD objective, whereas MI-SD and MI-SD$_{\text{cov}}$ aggregate supervision from all four for the same student rollout. Table 5 matches the base models, optimizer-step budget, checkpoint grid, and evaluator, with checkpoints selected independently for each benchmark. To match cumulative reference-solution exposure, we pair multi-reference checkpoint s with OPSD-flat checkpoint $4s$ at the same effective batch size, without benchmark-wise checkpoint selection. Figure 3 reports macro Avg@12 at five exposure levels up to 12,800 reference-solution instances. Early performance varies across methods and scales, but both MI-SD and MI-SD$_{\text{cov}}$ outperform OPSD-flat at all three scales from 7,680 instances onward. At higher exposure levels, the gap widens mainly because

Table 5: Matched-data comparison (Avg@12, %). Methods share the corpus, reference pool, base models, checkpoint grid, and evaluator.

Model	Method	AIME24	AIME25	HMMT25	Average
Qwen3-8B	OPSD-flat	76.7	70.6	47.5	64.9
	MI-SD$_{\text{cov}}$	78.6	**72.8**	**47.8**	**66.4**
	MI-SD	**78.9**	71.9	**47.8**	66.2
Qwen3-4B	OPSD-flat	75.8	69.7	45.6	63.7
	MI-SD$_{\text{cov}}$	76.7	**70.0**	**46.9**	**64.5**
	MI-SD	**77.8**	69.2	45.8	64.3
Qwen3-1.7B	OPSD-flat	55.8	41.7	**30.6**	42.7
	MI-SD$_{\text{cov}}$	55.8	40.8	**30.6**	42.4
	MI-SD	**58.9**	**45.6**	29.7	**44.7**

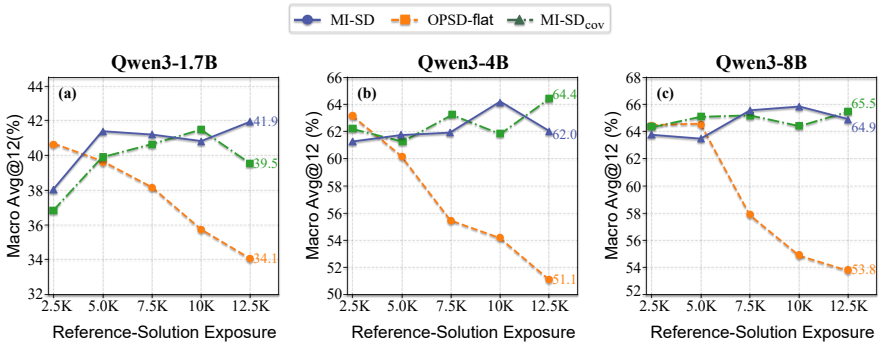

Figure 3: **Exposure matching.** Accuracy versus cumulative reference-solution exposure, evaluated at five paired exposure levels up to 12,800 reference-solution instances.

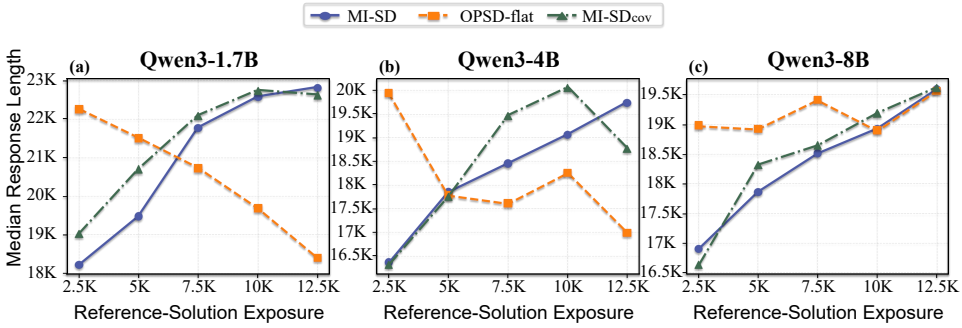

Figure 4: **Response-length dynamics.** Median length versus reference exposure; generation is capped at 38,912 tokens.

OPSD-flat's scores decline while the multi-reference methods remain comparatively stable. Greater cumulative reference exposure alone therefore does not explain this advantage.

Question 5: How does multi-reference training affect reasoning behavior? Prior work reports that self-distillation can shorten responses and reduce trajectory diversity (Kim et al., 2026). We therefore track median evaluation response length across checkpoints (Figure 4). OPSD-flat shortens substantially at 1.7B and 4B but fluctuates at 8B, whereas MI-SD maintains stable or increasing length at all three scales. MI-SD starts with shorter responses but matches or exceeds OPSD-flat in median length at higher reference exposure, indicating better length preservation at 1.7B and 4B over the evaluated range. We measure normalized route entropy over 32 sampled responses per problem using four teacher-derived routes. At 8B, MI-SD raises entropy by 0.053 over OPSD (positive on 12/16 problems) and the hybrid-route rate by 4.7 percentage points. Appendix G provides metric definitions and paired-bootstrap confidence intervals.

5 Conclusion

We introduced MI-SD, a multi-reference extension of on-policy self-distillation that uses a shared frozen teacher conditioned on multiple solution routes. Reference-conditioned disagreement provides a token-level signal of where solution perspectives meaningfully affect teacher predictions, which MI-SD uses to combine multi-reference coverage with MI-gated max-support supervision and a canonical anchor. Across three mathematical reasoning benchmarks and three Qwen3 model scales, MI-SD consistently improves over OPSD and remains competitive under matched-data and exposure-matched controls. The coverage objective accounts for much of the gain at larger scales, while the full objective provides the largest benefit for the smallest model. These results support multi-reference self-distillation as an effective way to exploit diverse valid reasoning paths. Future work can extend MI-SD beyond mathematical reasoning, explore richer forms of privileged references such as solution sketches or verification traces, and study the effect of larger reference sets.

AI USE STATEMENT

We used DeepSeek-v4-flash for data construction, including generating alternative reference solutions, as described in Appendix E. We also used large language models to annotate the reasoning routes of model-generated responses for the analysis reported in Appendix G. In addition, large language models assisted with manuscript preparation, including language polishing, translation, and grammar checking. We take full responsibility for the final content of this work, including its claims, results, and AI-assisted materials.

ETHICS STATEMENT

This work uses OpenThoughts-derived mathematical problems and AI-generated reference solutions to study multi-reference self-distillation. Appendix E describes the automated screening of reference solutions. Appendix D reports checks for overt claims of reference access and verbatim overlap. These checks do not rule out memorization or benchmark contamination. Use and redistribution of the data and generated references are subject to the applicable dataset licenses and model terms.

REPRODUCIBILITY STATEMENT

We describe the training objective and algorithm in Section 3 and Algorithm 1, and derive the mutual-information formulation in Appendix A. Section 4.1 specifies the evaluation benchmarks, Avg@12 metric, and benchmark-wise checkpoint-selection protocol. Appendix B reports training and inference settings and computational resources. Appendix E documents the construction and filtering of the 3,305-record training corpus. Appendix G defines the reasoning-route metrics and the paired-bootstrap procedure used to estimate confidence intervals.

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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1607–1616. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/furlanello18a.html.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8ac015d409635f196f9e3e9dcfb9a94e-Paper-Conference.pdf.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Zihao Han, Tiangang Zhang, Huaibin Wang, and Yilun Sun. Adaptive teacher exposure for self-distillation in LLM reasoning, 2026. URL https://arxiv.org/abs/2605.11458.

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 J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

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

ZhiYan Hou, Xinyu Tang, Hongyan An, Jianjin Zhang, Weizhen Wang, Yunyun Han, Gengsheng Li, Xiangzhao Hao, Haiyun Guo, Wenbin Hu, Jinqiao Wang, and Yafeng Deng. DASH: Divergence-adaptive supervision horizons for on-policy self-distillation of reasoning models, 2026. URL https://arxiv.org/abs/2608.06243.

Jiazhen Huang, Xiao Chen, Xiao Luo, Yong Dai, Senkang Hu, and Yuzhi Zhao. Skill-conditioned gated self-distillation for LLM reasoning, 2026. URL https://arxiv.org/abs/2605.28791. Accepted to Findings of EMNLP 2026, as reported by the authors on arXiv v4.

Yuki Ichihara, Naoto Iwase, Mohammad Atif Quamar, and Junpei Komiyama. Privileged solutions or context-induced teacher behavior? Dissecting on-policy self-distillation, 2026. URL https://arxiv.org/abs/2608.09228.

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, 2026a. URL https://openreview.net/forum?id=J5i09faOOf.

Yiqiao Jin, Yiyang Wang, Lucheng Fu, Yijia Xiao, Yinyi Luo, Haoxin Liu, B. Aditya Prakash, Josiah Hester, Jindong Wang, and Srijan Kumar. UniSD: Towards a unified self-distillation framework for large language models, 2026b. URL https://arxiv.org/abs/2605.06597.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of LLMs? In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=Az7guis46K.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857. Curran Associates, Inc., 2022. doi: 10.52202/068431-0278. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=P2OuWwZspP.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.03643.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Meta AI Blog, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Duy Nguyen, Hanqi Xiao, Archiki Prasad, Zaid Khan, Anirban Das, Austin Zhang, Sambit Sahu, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. AVSD: Adaptive-view self-distillation by balancing consensus and teacher-specific privileged signals, 2026. URL https://arxiv.org/abs/2605.20643.

OpenAI. GPT-5 system card. OpenAI, August 2025. URL https://openai.com/index/gpt-5-system-card/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. RLCSD: Reinforcement learning with contrastive on-policy self-distillation, 2026. URL https://arxiv.org/abs/2606.11709.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.04942.

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

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=qA6FgH0nnZ.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

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, 2025. URL https://arxiv.org/abs/2505.09388.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR, 2026. URL https://arxiv.org/abs/2604.03128.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. MetaMath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=N8N0hgNDRt.

Weichen Yu, Xiaomin Li, Yizhou Zhao, Xiaoze Liu, Ruowang Zhang, Haixin Wang, Yinyi Luo, Chen Henry Wu, Gaurav Mittal, Matt Fredrikson, and Yu Hu. Multi-rollout on-policy distillation via peer successes and failures, 2026a. URL https://arxiv.org/abs/2605.12652.

Xin Yu, Liuchen Liao, Yiwen Zhang, Yingchen Yu, Lingzhou Xue, and Qinzhen Guo. Preference-based self-distillation: Beyond KL matching via reward regularization, 2026b. URL https://arxiv.org/abs/2605.05040.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022. doi: 10.52202/068431-1126. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/639a9a172c044fbb64175b5fad42e9a5-Paper-Conference.pdf.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=Jpxfof0EaS.

Xuyang Zhao, Liting Zhang, Zichen Xu, Zhihu Wang, Xu Caiyue, Shiwan Zhao, and Qicheng Li. Is more privileged information better? From solution traces to problem-solving structure in self-distilled reasoning, 2026b. URL https://arxiv.org/abs/2608.01589.

A STUDENT–TEACHER DIVERGENCE CONFIGURATIONS

Following the OPSD implementation (Zhao et al., 2026a), the student–teacher divergence configuration uses forward KL $\mathrm{KL}(p_{k,t}\|q_{\theta,t})$ at $\beta = 0$ and reverse KL $\mathrm{KL}(q_{\theta,t}\|p_{k,t})$ at $\beta = 1$. For $0 < \beta < 1$, define $m_{k,t}^{(\beta)} = \beta p_{k,t} + (1-\beta)q_{\theta,t}$ and

$$\mathrm{JSD}_\beta(p_{k,t}, q_{\theta,t}) = \beta\,\mathrm{KL}\Big(p_{k,t}\|m_{k,t}^{(\beta)}\Big) + (1-\beta)\,\mathrm{KL}\Big(q_{\theta,t}\|m_{k,t}^{(\beta)}\Big). \tag{10}$$

The endpoint cases are separate implementation branches, not substitutions into Eq. (10). All reported experiments use $\beta = 0$; the interior and reverse-KL configurations were not evaluated. This student–teacher divergence is distinct from the teacher-only, equal-weight K-way JSD used to compute I_t.

A.1 PROOF: MUTUAL INFORMATION EQUALS GENERALIZED JSD

We formalize the uniform-reference construction in the main text and show that I_t equals the conditional mutual information between the reference index and a teacher token.

Definitions. Fix the context $c = (x, \mathcal{R}, y_{<t})$, where $\mathcal{R} = \{r_0, \ldots, r_{K-1}\}$ is the reference set. Let R be a reference index with $P(R = k \mid c) = 1/K$, and let Z_t be a teacher token satisfying

$$P(Z_t = v \mid R = k, c) = p_{k,t}(v) = p_{\theta_0}(v \mid x, r_k, y_{<t}).$$

The joint distribution and token marginal are therefore

$$P(R = k, Z_t = v \mid c) = \frac{p_{k,t}(v)}{K}, \qquad P(Z_t = v \mid c) = \bar{p}_t(v) = \frac{1}{K}\sum_{k=0}^{K-1} p_{k,t}(v).$$

Here Z_t is a teacher random variable over the entire vocabulary; y_t denotes the student's already sampled token. The student prefix is held fixed when defining this joint distribution. All logarithms in the MI and max-support scores are natural, so these scores and absolute MI thresholds are measured in nats. The score in the main text is

$$I_t = I(R; Z_t \mid c) = \mathbb{E}_R[\mathrm{KL}(p_{R,t}\|\bar{p}_t)]\,. \tag{11}$$

Derivation. Expanding the definition gives

$$\begin{aligned} I_t &= \frac{1}{K}\sum_{k=0}^{K-1}\sum_v p_{k,t}(v)\log\frac{p_{k,t}(v)}{\bar{p}_t(v)} \\ &= \frac{1}{K}\sum_{k=0}^{K-1}\sum_v p_{k,t}(v)\log p_{k,t}(v) - \sum_v \bar{p}_t(v)\log\bar{p}_t(v) \\ &= H(\bar{p}_t) - \frac{1}{K}\sum_{k=0}^{K-1} H(p_{k,t}), \end{aligned} \tag{12}$$

where $H(p) = -\sum_v p(v)\log p(v)$, with $0\log 0 = 0$. Equivalently,

$$I_t = \frac{1}{K}\sum_{k=0}^{K-1}\mathrm{KL}(p_{k,t}\|\bar{p}_t), \tag{13}$$

which is precisely the equal-weight generalized Jensen–Shannon divergence among K distributions.

Interpretation. Since $I_t = H(R \mid c) - H(R \mid Z_t, c)$ and $H(R \mid c) = \log K$, we have $0 \le I_t \le \log K$. The lower bound is attained when all reference-conditioned teacher distributions coincide. The upper bound is attained when the teacher token determines the reference index, equivalently when these distributions have pairwise disjoint supports. For $K > 1$, strictly positive softmax distributions can approach this upper bound but cannot attain it exactly. A larger I_t indicates stronger dependence of the teacher's next-token prediction on the reference.

B HYPERPARAMETERS AND IMPLEMENTATION DETAILS

Table 6 summarizes the hyperparameter settings for the main MI-SD and MI-SD$_{\text{cov}}$ experiments.

Table 6: Hyperparameter settings for the main MI-SD and MI-SD$_{\text{cov}}$ experiments. Settings are shared unless marked as MI-SD only.

Parameter	**Value**	**Notes**
Model and training		
Base models	Qwen3-1.7B, 4B, 8B	Instruct checkpoint
LoRA rank	64	
LoRA alpha (α)	128	
LoRA target modules	q, k, v, o, gate, up, down	7 projection layers
Learning rate	5×10^{-6}	AdamW
Gradient clipping	0.1	Maximum gradient norm
Precision	bf16	
Per-device microbatch size	4 (4B/8B)	Rollout examples per GPU
Gradient accumulation steps	2 (4B/8B)	
Effective batch size	32	
Max input length	20,000 tokens	
Max rollout length	1,024 tokens	Student completion only
Training steps	100	Checkpoints at 20, 40, 60, 80, 100
Number of references K	4	1 canonical + 3 alternatives
Generation (training)		
Temperature	1.1	
Top-p	0.95	
Top-k	20	
Objective settings		
Coverage weight	1.0	Implicit
Max-support weight λ	0.03 (1.7B), 0.005 (4B/8B)	MI-SD only; warmup at 4B/8B
Anchor weight μ	0.01	MI-SD only
β (KL direction)	0	Forward KL
Divergence-term cap	0.05 (1.7B/4B), 0.06 (8B)	Upper bound per vocabulary entry
MI scoring (MI-SD only)		
MI threshold γ	0.05 nats (1.7B); 90th percentile (4B/8B)	Quantile per GPU microbatch at 4B/8B
Surface-token filter	Disabled (1.7B); enabled (4B/8B)	Deterministic token-ID exclusion; max-support only
Evaluation		
Temperature	1.0	
Top-p	0.95	
Top-k	Disabled	
Max generation length	38,912 tokens	
Samples per problem	12	
Checkpoint selection	Benchmark-wise best	
Infrastructure		
GPUs	4× NVIDIA A800-SXM4-80GB	
Framework	Accelerate + DeepSpeed ZeRO-2	
Rollout engine	vLLM 0.11.0	

Score and warmup conventions. MI and max-support use natural logarithms. The MI and coverage terms use the same reference-conditioned distributions $p_{k,t}$; max-support evaluates these distributions at the sampled student token y_t. Writing λ for the target max-support weight, the linear warmup schedule in the implementation is

$$\lambda_s = \lambda \begin{cases} 0, & s < s_0, \\ 1, & s \geq s_0 \text{ and } W \leq 0, \\ \min\{1, (s - s_0 + 1)/W\}, & s \geq s_0 \text{ and } W > 0, \end{cases}$$

where s is the trainer's global step, s_0 is the activation step, and W is the warmup duration in steps. At 1.7B, we use $s_0 = W = 0$, so max-support is active from the start at its full weight. At 4B and 8B, we use $s_0 = 50$ and $W = 50$: $\lambda_s = 0$ for $s < 50$, $\lambda_{50} = 0.02\lambda$, and $\lambda_{99} = \lambda$. This schedule applies to the max-support coefficient; the coverage and anchor coefficients remain fixed.

Coverage-loss implementation. The implementation applies elementwise clipping and token normalization to the coverage objective. Let $\mathcal{B}$ denote the local batch passed to the loss computation, and let $a_{b,t} \in \{0, 1\}$ indicate a valid rollout position. Prompt and padding positions have $a_{b,t} = 0$. For the forward-KL setting ($\beta = 0$), define the vocabulary-wise contribution

$$d_{k,b,t,v} = p_{k,b,t}(v) \log \frac{p_{k,b,t}(v)}{q_{\theta,b,t}(v)}.$$

Each contribution is upper-clipped at δ before vocabulary summation. The implemented coverage loss is

$$\widehat{\mathcal{L}}_{\text{cov}}(\mathcal{B}) = \frac{1}{K N_{\text{valid}}} \sum_{b,t} a_{b,t} \sum_{k=0}^{K-1} \sum_{v} \min\{d_{k,b,t,v}, \delta\}, \tag{14}$$

where $N_{\text{valid}} = \sum_{b,t} a_{b,t}$. Thus, each valid rollout token receives equal weight within the local batch. The cap values are listed in Table 6. Clipping has no lower bound and remains part of the autograd graph. The MI score and max-support weights are computed from the original teacher distributions without this clipping. The coverage-only variant uses the same coverage computation with $\lambda = \mu = 0$.

Token filtering and percentile thresholding. The surface-token filter is a deterministic mask on sampled token IDs. For the sampled token $y_{b,t}$, let $m_{b,t} = 0$ if its ID is excluded and $m_{b,t} = 1$ otherwise. For Qwen3-1.7B, we use the absolute threshold $\gamma = 0.05$ nats and disable filtering.

For Qwen3-4B and Qwen3-8B, each GPU processes four rollouts per microbatch. With four GPUs and two gradient-accumulation steps, the effective update batch size is 32. We compute a separate threshold for each per-device microbatch $\mathcal{B}$, using all valid rollout positions before surface-token filtering:

$$\gamma_{\mathcal{B}} = Q_{0.90}(\{I_{b,t} : a_{b,t} = 1\}),$$

where $Q_{0.90}$ denotes the empirical 90th percentile. The quantile does not pool positions across GPU ranks or gradient-accumulation steps. The selected positions for sample b are

$$\mathcal{H}_b = \{t : a_{b,t} = 1,\ I_{b,t} > \gamma_{\mathcal{B}},\ m_{b,t} = 1\}.$$

Positions equal to the threshold are excluded.

The exclusion set combines token IDs obtained from single-token encodings of configured function words and literal surface strings with tokenizer-native IDs for whitespace, digit-only, punctuation, and Markdown/LaTeX surface forms. This mask affects only max-support reinforcement; coverage supervision continues to include all valid rollout positions identified by $a_{b,t}$.

C Additional Ablation Results

C.1 Per-Benchmark Component Ablation

Table 7 reports per-benchmark scores and macro averages for the component ablations in Section 4.3. The three components are multi-reference coverage (A), MI-gated max-support (B), and canonical anchor (C). We compare the full objective with variants that remove either B or C.

Table 7: Per-benchmark component ablation (Avg@12, %). Each configuration uses a single training seed. Checkpoints are selected independently for each benchmark. Avg is the macro average of the selected benchmark scores. Full MI-SD rows retain the one-decimal precision of Table 2; the ablation rows retain the available two-decimal scores.

Model	Configuration	AIME24	AIME25	HMMT25	Avg
Qwen3-1.7B	A+B+C (MI-SD)	58.9	45.6	29.7	**44.7**
	A+C (w/o B)	56.7	39.7	27.8	41.4
	A+B (w/o C)	55.8	42.8	26.9	41.9
Qwen3-4B	A+B+C (MI-SD)	77.8	69.2	45.8	**64.3**
	A+C (w/o B)	75.6	68.6	45.8	63.3
	A+B (w/o C)	76.1	68.9	45.0	63.3
Qwen3-8B	A+B+C (MI-SD)	78.9	71.9	47.8	**66.2**
	A+C (w/o B)	78.1	71.1	46.1	65.1
	A+B (w/o C)	77.8	70.8	48.1	65.6

The full objective achieves a higher macro Avg@12 than either two-component variant at all three scales. At 1.7B, removing B lowers the score by approximately 3.3 percentage points, compared with 2.9 percentage points when removing C. At 8B, the corresponding reductions are approximately 1.1 and 0.6 percentage points. These differences are calculated from the reported macro averages. Removing B therefore has a larger effect at these two scales. At 4B, both ablations reach the same macro average of 63.33, although their per-benchmark results differ. The effects of both deletions are largest at 1.7B.

These comparisons measure the effect of removing one auxiliary term while retaining the other. Relative to coverage-only training, the full objective performs better at 1.7B but scores slightly lower at 4B and 8B (Table 2).

C.2 Reference-Count Ablation: Per-Benchmark Details

We evaluate $K \in \{2, 3, 4\}$ for both the full MI-SD objective and the coverage-only variant on Qwen3-8B, using one training seed per configuration. Tables 8 and 9 report per-benchmark scores and macro averages for the reference-count ablation on Qwen3-8B. We evaluate $K \in \{2, 3, 4\}$ for both the full MI-SD objective and the coverage-only variant. Figure 5 visualizes the same comparison. The $K = 4$ rows retain the one-decimal precision reported in Table 2.

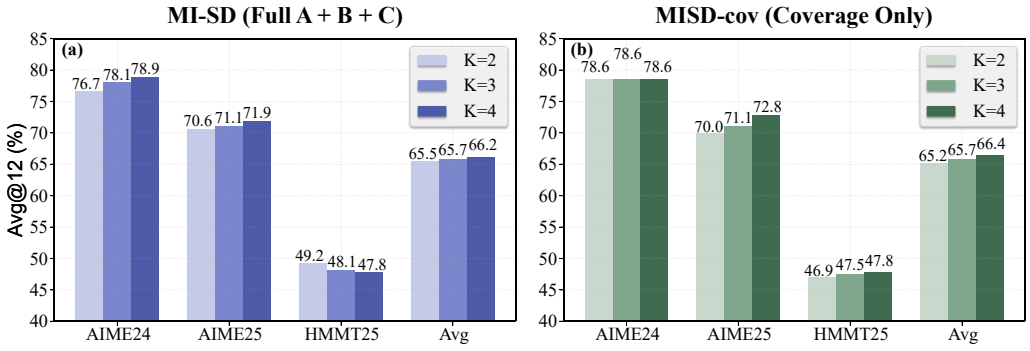

Figure 5: **Reference-count ablation on Qwen3-8B.** Avg@12 (%) for (a) full MI-SD and (b) MI-SD$_{\text{cov}}$, with $K \in \{2, 3, 4\}$. Values are rounded to one decimal place; Tables 8 and 9 report the detailed scores. Checkpoints are selected independently for each benchmark, using a single training seed per configuration. Macro averages improve with K, although full MI-SD declines on HMMT25.

Table 8: Per-benchmark reference-count ablation for full MI-SD (A+B+C) on Qwen3-8B (Avg@12, %). Each configuration uses a single training seed. Checkpoints are selected independently for each benchmark.

K	AIME24	AIME25	HMMT25	Avg
2	76.67	70.56	**49.17**	65.46
3	78.06	71.11	48.06	65.74
4	**78.9**	**71.9**	47.8	**66.2**

Table 9: Per-benchmark reference-count ablation for MI-SD$_{\text{cov}}$ (A only) on Qwen3-8B (Avg@12, %). Each configuration uses a single training seed. Checkpoints are selected independently for each benchmark.

K	AIME24	AIME25	HMMT25	Avg
2	78.61	70.00	46.94	65.18
3	78.61	71.11	47.50	65.74
4	78.6	**72.8**	**47.8**	**66.4**

Macro Avg@12 increases monotonically with K for both objectives. From $K = 2$ to $K = 4$, MI-SD$_{\text{cov}}$ gains approximately 1.2 percentage points, compared with 0.7 percentage points for full MI-SD. The per-benchmark trends differ. For full MI-SD, AIME24 and AIME25 improve as K increases, while HMMT25 declines. For MI-SD$_{\text{cov}}$, AIME24 remains nearly unchanged, while AIME25 and HMMT25 improve. Among the tested reference counts, $K = 4$ achieves the highest macro Avg@12 for both objectives.

D Training Dynamics and Behavioral Diagnostics

Response-length dynamics. Table 10 reports median completion length at five checkpoints for MI-SD and MI-SD$_{\text{cov}}$. No persistent late-stage response-length shortening is observed within the audited interval. Responses that reach the 38,912-token generation cap are right-censored; the maximum reported censoring rate is 6.67% for MI-SD$_{\text{cov}}$ at 4B checkpoint 100 (†). This cap censors the upper tail of the length distribution; all reported medians lie below it.

Table 10: Median completion length (tokens) at checkpoints 20, 40, 60, 80, and 100. † indicates a censoring rate above 5%.

Method	Model	Ckpt 20	Ckpt 40	Ckpt 60	Ckpt 80	Ckpt 100
MI-SD	Qwen3-1.7B	18,215	19,477	21,773	22,573	22,817
MI-SD	Qwen3-4B	16,362	17,853	18,456	19,064	19,742
MI-SD	Qwen3-8B	16,903	17,870	18,514	18,934	19,597
MI-SD$_{\text{cov}}$	Qwen3-1.7B	19,018	20,696	22,084	22,735	22,619
MI-SD$_{\text{cov}}$	Qwen3-4B	16,310	17,755	19,468	20,060	$18,762^{\dagger}$
MI-SD$_{\text{cov}}$	Qwen3-8B	16,634	18,326	$18,649^{\dagger}$	19,192	$19,625^{\dagger}$

Teacher attribution. The attribution diagnostic for 4B and 8B MI-SD evaluates up to 1,024 deterministically spaced non-padding positions across each complete rollout, using 40 questions and four responses per question. The recorded GT-nearest rate is approximately 29–32% across checkpoints, and the recorded GT-dominant trajectory rate increases from 47–51% at checkpoint 20 to 68–70% at checkpoint 100. A trajectory is labeled GT-dominant when the canonical teacher has the largest mean soft responsibility across the sampled positions in that rollout; this criterion does not require its responsibility to exceed 50%. At checkpoint 20, the mean soft responsibilities for the canonical reference and the three alternatives are 25.43%, 24.69%, 24.85%, and 25.03% at 4B, and 25.42%,

24.78%, 24.75%, and 25.05% at 8B, respectively. These summaries show approximately balanced recorded attribution across the four reference slots. They describe token-level teacher attribution, not semantic reasoning-route diversity.

Privileged-reference leakage. Across 16,200 evaluation responses, no response contained an affirmative claim of accessing a reference or answer key. Strict verbatim overlap detection found zero leakage hits out of 1,600 sampled trajectories. A looser high-overlap screen flagged 109/1,600 responses; manual inspection of the flagged spans found shared mathematical notation and no verbatim copying. This detector does not test for latent memorization or paraphrase-level contamination.

E Training Data Construction

The training corpus was constructed from OpenThoughts30K in two stages: initial candidate generation, followed by expansion and verification. Each of the 3,305 final records contains one canonical solution and three alternative solution texts, providing the four reference slots used by MI-SD and MI-SD$_{\mathrm{cov}}$.

Initial candidate generation. We used DeepSeek-v4-flash to generate three additional candidate solutions for each of 7,200 mathematics problems from OpenThoughts30K. The generation instructions varied both the requested reasoning perspective and presentation style. Method-oriented instructions encouraged algebraic or structural and symmetry-based approaches where applicable, while style-oriented instructions varied the level of concision. All candidates were required to produce a single final `\boxed{}` answer. This initial stage yielded a candidate pool of 3,362 problems, each with three candidate solution texts in addition to its canonical solution. This pool served as the input to the subsequent expansion and verification stage.

Expansion and verification. We expanded the initial candidates by adding intermediate equations, substitutions, and case distinctions. The expansion instructions required preserving each candidate's main method whenever it was compatible with the canonical answer. Automated checks required a single final boxed answer, agreement with the canonical answer, and bounds on length (180–520 words) and superficial copying. A verifier prompt additionally assessed correctness, reasoning depth, and whether the expansion overwrote the original solution route. Candidates failing either the automated checks or the verifier assessment were regenerated; unresolved records were omitted. Table 11 summarizes both construction stages.

Table 11: Stages of multi-reference data construction, from initial candidate generation to expansion and verification. Counts refer to problems. "Rerun records" denotes the subset with at least one candidate requiring repair.

Stage	Records	Operation
Source problems	7,200	Input for initial candidate generation
Initial candidate pool	3,362	Three candidate texts per problem
Candidate expansion	3,355	Expand existing candidate solutions
Verifier screening	3,355	Check expanded candidates
Rerun records	1,624	Repair candidates flagged by checks
Final export	3,305	Canonical solution + three alternatives

In the final export, all 9,915 alternatives contain exactly one boxed answer and have word counts between 180 and 520 (median 283). We use "alternative solution" to denote a selected reference text that satisfies the answer-consistency, length, and verification filters. The verifier provides an automated correctness-screening signal rather than a formal guarantee of stepwise mathematical correctness.

Reference diversity. A deterministic check of the final export finds no identical alternatives within a problem and no alternative identical to its canonical solution after lowercasing and whitespace normalization. This check detects text duplication; it does not assess whether the solutions use distinct

mathematical methods. The resulting references provide varied formulations and intermediate reasoning steps for the same problem. MI-SD uses these four solution texts as separate conditioning contexts, allowing the shared teacher to score the student rollout under each reference.

F Full Fixed-Exposure Comparison

Table 12 reports macro Avg@12 at all five pre-specified exposure levels. Δ_{cov} is MI-SD$_{\text{cov}}$ minus OPSD-flat; the MI-SD column reports absolute scores, from which the corresponding difference against OPSD-flat can be read directly. Δ_{cov} is computed from unrounded scores and may differ from the displayed columns by 0.01 points. Δ_{full} is computed from the displayed scores.

Table 12: Fixed-exposure comparison across Qwen3 model scales (macro Avg@12, %). Exposure counts cumulative reference-solution instances. Δ_{cov} and Δ_{full} denote the differences of MI-SD$_{\text{cov}}$ and MI-SD relative to OPSD-flat, respectively, in percentage points.

Model	Exposure	OPSD-flat	MI-SD$_{\text{cov}}$	Δ_{cov}	MI-SD	Δ_{full}
Qwen3-1.7B	2,560	40.65	36.85	-3.80	38.06	-2.59
	5,120	39.63	39.91	$+0.28$	41.39	$+1.76$
	7,680	38.15	40.65	$+2.50$	41.20	$+3.05$
	10,240	35.74	41.48	$+5.74$	40.83	$+5.09$
	12,800	34.07	39.54	$+5.46$	41.94	$+7.87$
Qwen3-4B	2,560	63.15	62.22	-0.93	61.30	-1.85
	5,120	60.19	61.30	$+1.11$	61.76	$+1.57$
	7,680	55.46	63.24	$+7.78$	61.94	$+6.48$
	10,240	54.17	61.85	$+7.69$	64.17	$+10.00$
	12,800	51.11	64.44	$+13.33$	62.04	$+10.93$
Qwen3-8B	2,560	64.44	64.35	-0.09	63.80	-0.64
	5,120	64.54	65.09	$+0.56$	63.52	-1.02
	7,680	57.87	65.19	$+7.31$	65.56	$+7.69$
	10,240	54.91	64.44	$+9.54$	65.83	$+10.92$
	12,800	53.80	65.46	$+11.67$	64.91	$+11.11$

G Semantic Reasoning-Route Diversity

Metric definition. For each question q, we collect 32 student responses and use four teacher-derived reasoning routes. Let $\mathcal{V}_q$ denote this four-route taxonomy, and let $\mathcal{R}_{q,i} \subseteq \mathcal{V}_q$ be the `matched_routes` annotation for response i. Each response with at least one matched route contributes unit mass, divided equally among its matched routes. A single-route response therefore contributes all its mass to one route; a response matching two, three, or four routes contributes $1/2$, $1/3$, or $1/4$ to each, respectively. Unmatched responses contribute zero mass. HYBRID responses participate in this calculation through their matched routes and are not excluded on the basis of their primary label.

Let N_q be the number of responses with at least one matched route, and let $a_q = N_q/32$ be the attribution rate. For $N_q > 0$, we normalize the fractional route masses over attributed responses:

$$p_{q,r} = \frac{1}{N_q} \sum_{i:\,|\mathcal{R}_{q,i}|>0} \frac{\mathbf{1}\{r \in \mathcal{R}_{q,i}\}}{|\mathcal{R}_{q,i}|}. \tag{15}$$

Thus $\sum_{r\in\mathcal{V}_q} p_{q,r} = 1$. The normalized route entropy is

$$H_q = \frac{-\sum_{r\in\mathcal{V}_q} p_{q,r} \log_2 p_{q,r}}{\log_2 4}, \tag{16}$$

with $0 \log_2 0 = 0$. This metric measures the dispersion of fractional route mass among attributed responses and lies in $[0, 1]$. It is undefined when $N_q = 0$; all 16 evaluated questions have $N_q > 0$

under both methods. We report attribution rates separately. For an evaluated question set $\mathcal{Q}$, the macro entropy is $H_{\text{macro}} = |\mathcal{Q}|^{-1} \sum_{q \in \mathcal{Q}} H_q$.

Hybrid-response rate. Let $\ell_{q,i}$ denote the primary label of response i. The hybrid-response rate is

$$h_q = \frac{1}{32} \sum_{i=1}^{32} \mathbf{1}\{\ell_{q,i} = \text{HYBRID}\}.$$

Each response labeled HYBRID counts once, regardless of whether it matches two, three, or four routes. Its fractional route contributions to H_q are computed separately from this binary count. The macro hybrid rate gives each evaluated question equal weight.

Statistical protocol. We compare canonical OPSD and MI-SD at checkpoint 80. For route entropy, the per-question paired difference is $\Delta H_q = H_q^{\text{MISD}} - H_q^{\text{OPSD}}$. We compute percentile-based 95% confidence intervals for the mean paired difference using 10,000 bootstrap resamples of questions with seed 20260812. The same paired-bootstrap procedure is applied to the per-question hybrid-response and attribution rates. The intervals reflect question-sampling variability at a single checkpoint and a single training run per method; they do not quantify variability across training seeds.

The same protocol was applied at 4B and 8B. We report the 8B comparison in detail below. At 4B, the mean paired entropy difference is -0.005 (95% CI: $[-0.051, 0.038]$), with 8 of 16 questions positive, and the hybrid-response change is -1.2 percentage points (95% CI: $[-4.5, 2.3]$). An increase is supported at 8B but is not established at 4B. These separate intervals do not establish a difference in effects between scales.

Results. At 8B, the macro normalized route entropy is 0.4616 for canonical OPSD and 0.5144 for MI-SD. The mean paired increase is $+0.053$ (95% CI: $[0.017, 0.090]$), with 12 of 16 evaluated questions showing a positive ΔH_q (Figure 6(a,b)). The hybrid-response rate increases by 4.7 percentage points (95% CI: $[2.0, 8.0]$), with 10 of 16 questions showing an increase (Figure 6(c)). Attribution rates are 98.24% for canonical OPSD and 99.61% for MI-SD, a difference of 1.37 percentage points (95% CI: $[0.00, 2.93]$).

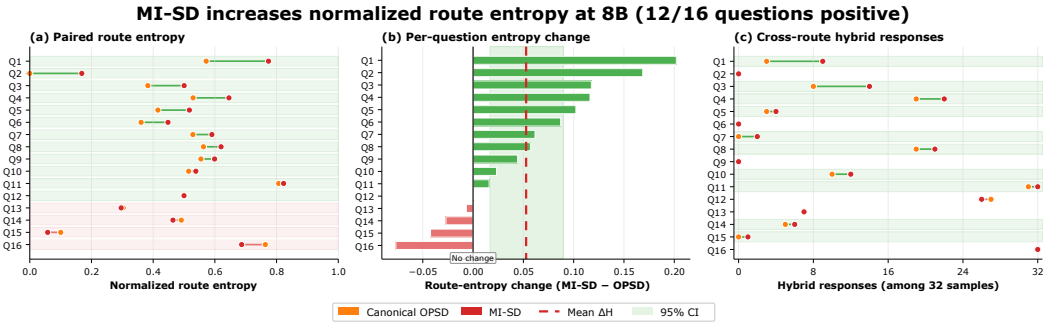

Figure 6: **Normalized reasoning-route entropy and hybrid responses at 8B checkpoint 80.** (a) Per-question normalized route entropy among attributed responses for canonical OPSD (orange) and MI-SD (red), with paired values connected. (b) Per-question entropy differences (MI-SD minus OPSD): positive differences are green and negative differences are red. The black line marks zero change. The dashed red line marks the mean difference of 0.053, and the shaded band shows its 95% paired-bootstrap confidence interval $[0.017, 0.090]$. Twelve of the 16 questions have positive entropy differences. (c) Per-question numbers of responses with primary label HYBRID among 32 samples, counting each response once. The macro hybrid-response rate increases by 4.7 percentage points (95% paired-bootstrap CI: $[2.0, 8.0]$). All panels use the same question order, sorted by descending entropy difference.