TRANSFER-STRATIFIED ON-POLICY DISTILLATION FOR RL-IMPROVED REASONING TEACHERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning can substantially improve a reasoning teacher, but it is unclear which of those improvements survive when the teacher supervises a smaller on-policy student. We study this question in mathematical reasoning by comparing teacher lineages before and after GRPO, multiple student scales, direct GRPO, and several on-policy distillation objectives. The central finding is that transfer is structured rather than scalar: teacher strength alone does not make dense distillation competitive, while an RL-improved teacher creates useful but metric-dependent student gains. This motivates Transfer-Stratified On-Policy Distillation (TS-OPD), which screens training problems by the joint sampled success of the student and teacher, routes acquisition problems to gated forward KL, routes consolidation problems to gated reverse KL, and adds an entropy brake to protect sampled coverage. Across the main comparison, TS-OPD is the strongest student objective for macro average correctness with the GRPO-improved teacher, while pass@K remains more mixed. Ablations show that the gains come from routing and token gating rather than skipping problems. These results support a transfer-aware view of OPD: stronger teachers help when the supervision direction and token budget match the student's observed ability, not merely because the teacher endpoint is stronger. We submit our code as supplementary material and will publicly release it upon acceptance.

1 INTRODUCTION

On-policy distillation (OPD) is the dominant paradigm for transferring reasoning capabilities from large models to smaller ones (Agarwal et al., 2024; Gu et al., 2024). It trains students on their own generated states under token-level supervision from a frozen teacher, reducing the training–inference distribution mismatch. A teacher improved through reinforcement learning with verifiable rewards, such as GRPO (Shao et al., 2024; Guo et al., 2025), should therefore provide stronger supervision. Prior work indeed shows that a GRPO-trained teacher produces a stronger student than the original teacher (Xu et al., 2026b). Yet our initial observations on MATH (Hendrycks et al., 2021) show that student gains can fall short of the teacher's GRPO improvement. We call this shortfall the **transfer gap** and ask: **which teacher improvements transfer to the student, which do not, and how can more of them be made to transfer?**

Our problem-level analysis on MATH shows that this gap is not uniform across training problems (Figure 1). Whether a teacher improvement can transfer depends on how the GRPO-trained teacher (hereafter, the *GRPO teacher*) and student perform on each problem. Where the teacher outperforms the student, two distinct learning needs arise. If the student never solves the problem, it faces a **knowledge gap**: it needs guidance toward correct choices it has never explored, yet standard OPD supervises only its own responses, making such gains especially difficult to transfer. If the student solves the problem only sometimes, it faces a **precision gap**: it needs targeted correction of specific bad choices, whereas indiscriminate distillation may transfer only part of the gain or suppress correct behavior already acquired.

Conversely, when the GRPO teacher does not outperform the student, it has little improvement to transfer. When their performance is comparable or both fail, the teacher offers no better direction; when the student is stronger, matching the teacher may be harmful. Standard OPD nevertheless

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.

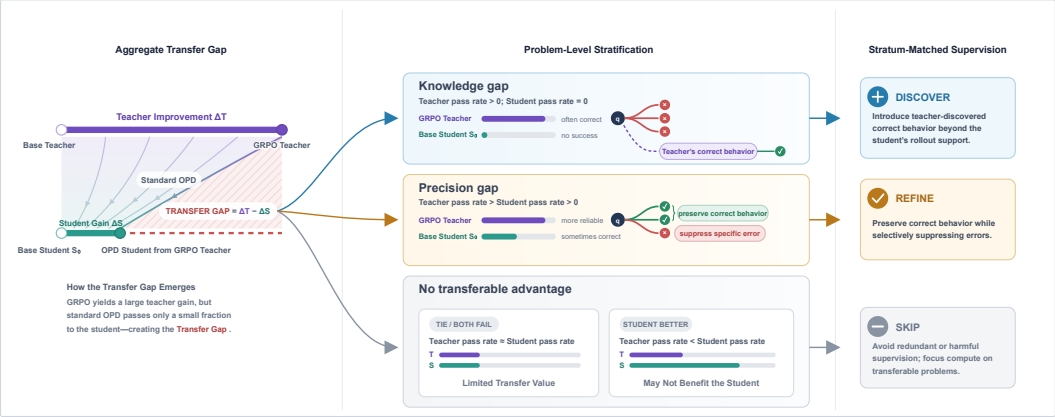


Figure 1: **Overview of the transfer gap and its problem-level decomposition. Left:** GRPO yields a teacher improvement ΔT, but standard OPD transfers only part of it into a student gain ΔS, leaving a transfer gap of $\Delta T - \Delta S$. **Middle:** Joint teacher–student pass rates identify knowledge gaps (teacher $>$ 0, student $=$ 0), precision gaps (teacher $>$ student $>$ 0), and problems with no transferable teacher advantage (teacher $\leq$ student). **Right:** TS-OPD routes these three strata to matched actions—discover, refine, and skip, respectively.

supervises these problems uniformly. Skipping them concentrates the training budget on problems where transfer is possible.

Existing OPD methods adapt supervision at the problem or token level (Lin et al., 2026; Xu et al., 2026c;a; Zhang et al., 2026). Most closely related, Sparse-to-Dense distills from a GRPO-improved teacher using a common pipeline across problems (Xu et al., 2026b). In contrast, we use the joint success of the teacher and raw student to identify each problem's learning need and match both its KL direction and token-level supervision.

We therefore propose **Transfer-Stratified On-Policy Distillation (TS-OPD)**. It first stratifies problems by the joint pass rates of the GRPO teacher and student, then routes knowledge and precision gaps to matched objectives with stratum-specific token gating, skipping problems without a teacher advantage. We evaluate TS-OPD across two teacher lineages (Qwen3-8B before and after GRPO), three student scales (0.6B, 1.7B, and 4B), four distillation objectives, and teacher-free direct GRPO.

Our contributions are as follows:

1. We characterize and decompose the transfer gap in distillation from GRPO-improved teachers into distinct problem-level learning needs.
2. We propose **TS-OPD**, which stratifies and selects problems by observed teacher–student success, routes distillation objectives, and gates token-level supervision to address knowledge and precision gaps.
3. Experiments with a GRPO-improved teacher show that TS-OPD achieves the highest macro-average correctness and smallest benchmark-averaged transfer gap among evaluated methods at all three student scales, while distilling only on selected problems.

2 Related Work

RL-improved teachers. DeepSeek-R1 distills an RL-trained model by fine-tuning on its responses (Guo et al., 2025). Sparse-to-Dense (Xu et al., 2026b) applies GRPO (Shao et al., 2024) to a larger teacher, then a forward-KL warmup and reverse-KL OPD uniformly across problems, and shows that the RL-improved teacher yields a student stronger than direct GRPO; it compares final students, not how much of the teacher's GRPO gain is inherited, or where. Direct-OPD instead transfers the RL-induced policy shift, using the post-RL to pre-RL teacher log-ratio as a dense reward (Feng et al., 2026), while Li et al. (2026) find that OPD requires compatible thinking patterns and teacher capabilities new to the student (Li et al., 2025; Zhang et al., 2025; Yue et al., 2025).

On-policy objectives. GKD trains the student on its own samples and admits divergences beyond forward KL (Agarwal et al., 2024), and MiniLLM optimizes reverse KL on-policy (Gu et al., 2024). AKL, ToDi, EOPD, and TrOPD combine forward and reverse KL through adaptive weights, per-token probability ratios, teacher entropy, or token-level trust regions (Wu et al., 2024; Jung et al., 2025; Jin et al., 2026; Xing et al., 2026); none chooses the direction from per-problem teacher and student performance (Ko et al., 2024; 2025; Wang et al., 2025a).

Selective distillation. PACED weights each problem by $p(1-p)$ of the student pass rate p and applies a forward-then-reverse KL schedule to all problems (Xu et al., 2026a); TREK applies forward KL on verified teacher proposals for prompts the student rarely solves, then returns to GRPO (Xu et al., 2026d; Yan et al., 2025); TGOPD routes each prompt to OPD or GRPO using verifier-scored teacher probes (Zhang et al., 2026). Within responses, OPDVR gates the distillation reward by trajectory correctness (Lin et al., 2026), and TIP selects tokens by student entropy and teacher–student divergence (Xu et al., 2026c; Wang et al., 2025b; Lin et al., 2025). Unlike these student-only or teacher-only signals, TS-OPD uses the joint success of the GRPO teacher and the raw student to choose each problem's KL direction (forward for knowledge gaps, reverse for precision gaps) and adds response-local token gating, rather than introducing problem-level selection itself.

3 Problem Definition and Transfer-Gap Diagnosis

Building on our observations on MATH, we formalize the transfer gap and identify knowledge and precision gaps on MATH training problems. We examine these strata alongside gain shortfalls on held-out competition benchmarks to motivate differentiated OPD supervision.

3.1 Evaluation Metrics and the GRPO Transfer Gap

For each problem x_i, we sample K responses from model M under fixed decoding settings and score them with a verifier $c_i(y) \in \{0, 1\}$. The sampled success rate is $\widehat{p}_{M,i} = K^{-1} \sum_{j=1}^{K} c_i(y_{M,i,j})$. For a test set of N problems, define

$$\text{avg@}\,K(M) = \frac{1}{N} \sum_{i=1}^{N} \widehat{p}_{M,i}, \qquad \text{pass@}\,K(M) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\widehat{p}_{M,i} > 0]. \tag{1}$$

avg@K measures average single-response correctness, our primary performance metric; **pass@K measures solution coverage within the sampling budget** (Chen et al., 2021).

Let T_0 and T_1 denote the teacher before and after GRPO, S_0 the raw student, and S_{OPD} the student distilled from T_1. With test accuracy $a(M) = 100\,\text{avg@}\,K(M)$, define the teacher and student gains and the **transfer gap**:

$$\Delta_T = a(T_1) - a(T_0), \qquad \Delta_S = a(S_{\text{OPD}}) - a(S_0), \qquad G_{\text{transfer}} = \Delta_T - \Delta_S. \tag{2}$$

All three quantities are measured in percentage points (pp). A positive transfer gap means that the student's gain falls short of the teacher's GRPO gain, not that their final accuracies differ by this amount. Here Δ_S is the total student improvement from S_0, not the difference between students distilled from T_1 and T_0. It may include knowledge already present in T_0, so the gap does not isolate the loss of GRPO-acquired knowledge. With the teacher and student initialization fixed, our goal is to reduce this gap and outperform S_{GRPO}, trained directly with GRPO from the same S_0.

3.2 Two Learning Gaps on Training Problems

We hypothesize that incomplete transfer partly reflects **supervision mismatched to problem-specific learning needs**. A stronger teacher need not outperform the student on every problem. Where it does, the guidance needed depends on whether the student can already generate a correct response. Before distillation, we therefore sample $K_{\text{cal}} = 8$ responses per training problem from both T_1 and S_0 and classify their joint success rates.

A: knowledge gap, $\widehat{p}_{S_0,i} = 0 < \widehat{p}_{T_1,i}$. The student needs to discover correct choices absent from its sampled rollouts, not merely increase the frequency of existing correct responses. Zero observed success at a finite budget does not imply zero ability.

C: precision gap, $0 < \widehat{p}_{S_0,i} < \widehat{p}_{T_1,i}$. The student already produces correct responses but less reliably, calling for targeted correction while preserving existing correct behavior.

P: student advantage, $\widehat{p}_{S_0,i} > \widehat{p}_{T_1,i}$; **N: no observed gap**, equal rates, including joint failures. Neither shows an observed teacher advantage, so neither is prioritized for distillation. Let $\mathcal{D}_{\text{train}}$ be the training-problem index set, with subsets $\mathcal{D}_A$ and $\mathcal{D}_C$ for the knowledge and precision gaps. Using $|\cdot|$ for set cardinality and $\mathbf{1}[\cdot]$ for the indicator function, define the **teacher-advantage problem fraction**:

$$r_{\text{adv}} = \frac{|\mathcal{D}_A| + |\mathcal{D}_C|}{|\mathcal{D}_{\text{train}}|} = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{i \in \mathcal{D}_{\text{train}}} \mathbf{1}[\widehat{p}_{T_1,i} > \widehat{p}_{S_0,i}]. \tag{3}$$

This is the fraction of problems with an observed teacher advantage before training, not its magnitude. All four stratum fractions use the full training set as the denominator and sum to one.

3.3 Training Strata and Test Gain Shortfalls

We use Qwen3-8B before and after GRPO as the teacher and Qwen3-0.6B/1.7B as students (Yang et al., 2025). Stratification uses 7,496 MATH training problems (Hendrycks et al., 2021), with eight responses per problem from each of T_1 and S_0; test diagnosis uses AIME24/25 and HMMT-Feb/Nov25, totaling 120 problems with 16 responses per model per problem (Art of Problem Solving, n.d.; HMMT, 2025a;b).

Table 1: Training strata and test gain shortfalls. (a) Fractions of all 7,496 MATH training problems, classified using eight responses per model before training. A/C denote knowledge/precision gaps; P/N denote student advantage/no observed gap. (b) Test avg@16 accuracy on AIME24/25 and HMMT-Feb/Nov25 (120 problems), with gains and transfer gaps in pp. Both students share the Qwen3-8B teacher checkpoints; S_{OPD} uses standard reverse-KL OPD from S_0 with T_1. Values are rounded independently to two decimal places.

(a) Training-problem fractions (%); teacher T_1, student S_0

Student	A	C	P	N	**A+C: r_{adv}**
0.6B	23.24	53.92	3.55	19.29	**77.16**
1.7B	7.39	36.13	7.68	48.80	**43.52**

(b) Test accuracy and gain shortfalls; $G_{\text{transfer}} = \Delta_T - \Delta_S$ (pp)
Shared teacher: $a(T_0) = 15.21\%$, $a(T_1) = 41.15\%$, $\Delta_T = 25.94$ pp

Student	$a(S_0)$ (%)	$a(S_{\text{OPD}})$ (%)	Δ_S (pp)	**G_{transfer} (pp)**
0.6B	1.04	8.91	7.87	**18.08**
1.7B	6.15	19.58	13.44	**12.50**

Table 1(a) shows teacher-advantage fractions of 77.16% and 43.52%, while panel (b) reports the corresponding test gain shortfalls of 18.08 and 12.50 pp for the 0.6B and 1.7B students, respectively. The larger fraction accompanies the larger shortfall, but student scale and stratum distribution both vary. The ordering also reverses across all eight test sets, where macro-average shortfalls are 5.49 and 6.14 pp. We therefore use the strata to describe supervision needs, not as an established predictor of test gain shortfalls.

3.4 Implications for OPD Supervision

For A, correct responses may be rare in student rollouts. Because OPD queries the teacher along student-generated prefixes (Agarwal et al., 2024; Gu et al., 2024), guidance along complete correct solutions may be limited. Feedback at incorrect prefixes remains available but may not suffice to reach a correct solution, motivating guidance toward teacher-preferred continuations the student rarely selects.

For C, uniform supervision may miss decisive errors and spend limited budget on P and N, leaving learnable problems undertrained. This motivates targeted correction while preserving existing correct behavior. These mechanism hypotheses require fixed-budget stratification controls and token-level analysis.

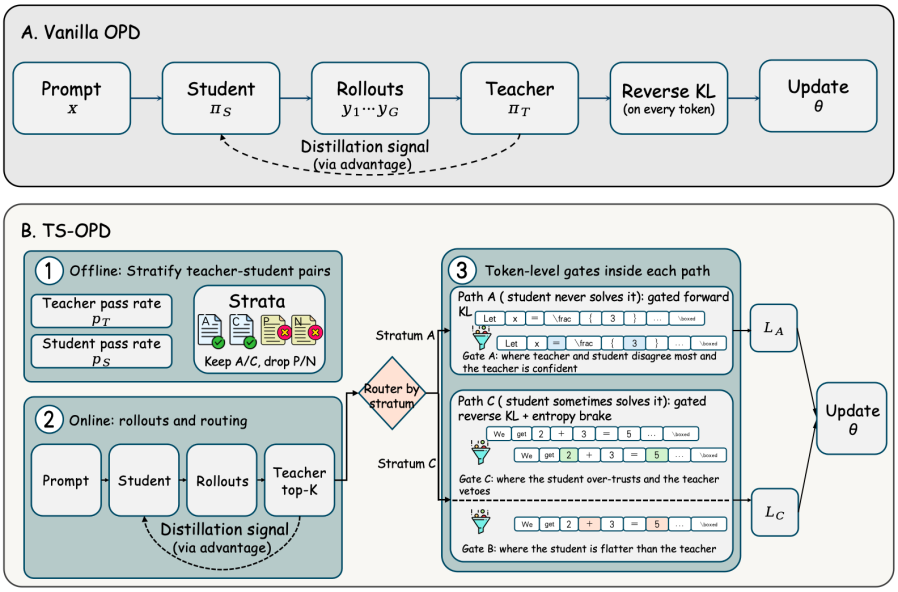


Figure 2: Overview of TS-OPD. Before training, each problem is labeled by the sampled success of the teacher and the student. Problem level: P and N problems are skipped, A problems use forward KL, and C problems use reverse KL. Token level: gates select positions within a response, and C adds an entropy brake.

We therefore prioritize A and C with supervision matched to knowledge acquisition and precision improvement, deferring teacher matching on P and N. Strata are assigned before training from joint teacher–student samples; rollouts remain student-generated.

4 Transfer-Stratified On-Policy Distillation

TS-OPD labels training problems by the teacher and student success observed before training and works at two levels (Figure 2). At the **problem level**, it trains only on A and C problems (**filtering**) and applies forward KL to A and reverse KL to C (**routing**). At the **token level**, it updates only selected positions within each response and adds an entropy brake on C (**gating**). TS-OPD (routing only) keeps the problem level and drops the token level; TS-OPD uses both.

4.1 Strata and KL Direction

Strata. We use the labels $\mathcal{D}_A$ and $\mathcal{D}_C$ from Section 3, computed once per teacher–student pair from $K_{\text{cal}} = 8$ samples of T_1 and S_0 and fixed during training. Filtering keeps only $\mathcal{D}_+ = \mathcal{D}_A \cup \mathcal{D}_C$.

Notation. For a training problem x, the student generates a response $y = (y_1, \ldots, y_{|y|})$ of length $|y|$. The prefix before the t-th token is $s_t = (x, y_1, \ldots, y_{t-1})$. At s_t, the student and the teacher (T_1) assign next-token probabilities $\pi_S(v \mid s_t)$ and $\pi_T(v \mid s_t)$ to each token v in the vocabulary $\mathcal{V}$. OPD pulls π_S toward π_T on these prefixes (Agarwal et al., 2024; Gu et al., 2024).

Why forward KL for A and reverse KL for C. At one position (we drop s_t), reverse KL is $D_{\text{R}} = \sum_v \pi_S(v) \log \frac{\pi_S(v)}{\pi_T(v)}$ and forward KL is $D_{\text{F}} = \sum_v \pi_T(v) \log \frac{\pi_T(v)}{\pi_S(v)}$. Let z_v be the student's logit for token v, so $\pi_S = \text{softmax}(z)$. The gradients are

$$\frac{\partial D_{\text{R}}}{\partial z_v} = \pi_S(v)\Big(\log \frac{\pi_S(v)}{\pi_T(v)} - D_{\text{R}}\Big), \qquad \frac{\partial D_{\text{F}}}{\partial z_v} = \pi_S(v) - \pi_T(v). \tag{4}$$

Training moves z_v against its gradient. For a token the teacher prefers but the student almost never picks ($\pi_S(v) \approx 0$), the reverse-KL gradient carries the factor $\pi_S(v)$ and barely moves it, while

the forward-KL gradient is close to $-\pi_T(v)$ and raises it. A problems need behavior that has not appeared in the student's responses, so we use forward KL. C problems only need to drop choices the teacher rejects, so we use reverse KL (Jung et al., 2025). This choice per stratum is the routing. This fixed-prefix argument only motivates the loss; it does not describe the gated update or guarantee a correct response (Sheng et al., 2026).

Token cells and position selection. The token level (gating) works on a compressed teacher distribution. At each position we keep the teacher's top $K_{\text{tok}} = 512$ tokens plus one tail cell for the rest, indexed by j in decreasing teacher probability. $P_{t,j}$ and $Q_{t,j}$ are the teacher and student probabilities of cell j at position t. Both layers below select positions within a response. Let $\text{qt}_\tau(u)$ be the τ-quantile of a score u over the positions of one response (so long responses cannot take most selections); with $\tau = 0.8$, a position passes when its score is in the top 20% of its response.

4.2 A Layer: Gated Forward KL

Two kinds of positions help little on A problems. Where the student already matches the teacher, there is nothing to learn; where the teacher spreads its probability over many tokens, it cannot teach a definite choice. We measure disagreement by total variation $\text{TV}_t = \frac{1}{2}\sum_j |P_{t,j} - Q_{t,j}| \in [0,1]$, and teacher confidence by $\text{mass@}n(t) = \sum_{j=1}^{n} P_{t,j}$, the probability the teacher puts on its top n tokens. We keep a position only if both conditions hold, and the A loss averages the forward KL on the cells, $D(P_t\|Q_t) = \sum_j P_{t,j}\log(P_{t,j}/Q_{t,j})$, over the kept positions:

$$m_t^A = \mathbf{1}\big[\text{TV}_t \geq \text{qt}_\tau(\text{TV})\big] \cdot \mathbf{1}\big[\text{mass@}n(t) > 0.9\big], \qquad \mathcal{L}^A = \frac{\sum_t m_t^A\, D(P_t\|Q_t)}{\sum_t m_t^A}. \tag{5}$$

Here $\mathbf{1}[\cdot]$ is 1 if the condition holds and 0 otherwise. The first condition keeps the 20% of positions with the largest disagreement. The second requires the teacher to put over 90% of its probability on a few candidates, that is, to be confident. We do not use $n = 1$ because a reasoning step often has several equivalent forms; we set $n = \max(\lfloor \ln K_{\text{tok}} \rfloor, 5) = 6$.

4.3 C Layer: Gated Reverse KL and Entropy Brake

On C problems, we only correct sampled tokens y_t that the teacher rejects and leave the rest unchanged:

$$r_t = \log \frac{\pi_S(y_t \mid s_t)}{\pi_T(y_t \mid s_t)}, \quad \rho_t = \max\Big(0,\, 1 - \frac{\pi_T(y_t \mid s_t)}{\pi_S(y_t \mid s_t)}\Big), \quad m_t^C = \mathbf{1}\big[\rho_t \geq \text{qt}_\tau(\rho)\big] \cdot \mathbf{1}[\rho_t > 0]. \tag{6}$$

$r_t > 0$ means the student is more confident in y_t than the teacher; $\rho_t \in [0,1)$ is 0 when the teacher gives y_t at least the student's probability and grows toward 1 as the teacher's falls below it. m_t^C keeps the top 20% of positions with $\rho_t > 0$; the second factor is needed because over half of the ρ_t are 0, so the quantile can be 0.

Lowering reverse KL with sampled tokens. Reverse KL is the average of $\log \frac{\pi_S(v)}{\pi_T(v)}$ over tokens drawn from the student, so r_t on the sampled token y_t is a one-sample estimate of it. We use r_t as a penalty: the larger r_t, the more we lower the probability of y_t. This needs only the student's own response and no correctness reward (Gu et al., 2024) (derivation in Appendix A.2). With $\omega_t = -m_t^C r_t$ held constant, the reverse-KL loss is

$$\mathcal{L}^{\text{RKL}} = -\frac{1}{|y|}\sum_t \omega_t \log \pi_S(y_t \mid s_t). \tag{7}$$

Unselected positions have $\omega_t = 0$ and are not updated.

Entropy brake. Reverse KL concentrates probability on a few tokens, which can remove other correct solutions and lower pass@K. Let $H_S(t) = -\sum_j Q_{t,j}\log Q_{t,j}$ and $H_T(t) = -\sum_j P_{t,j}\log P_{t,j}$ be the student and teacher entropies. Where $H_S - H_T$ is largest, reverse KL sharpens the student the most, so we raise the student entropy there, at positions picked by m_t^B:

$$m_t^B = \mathbf{1}\big[H_S(t) - H_T(t) \geq \text{qt}_\tau(H_S - H_T)\big], \qquad \mathcal{L}^B = -\beta\,\frac{\sum_t m_t^B H_S(t)}{\sum_t m_t^B}. \tag{8}$$

We set β once per pair before training so that the brake is about a tenth of the penalty in magnitude, a heuristic scale match rather than a gradient ratio, with 0.1 as a design choice (Appendix A.1).

Full objective and training loop. The C loss is $\mathcal{L}^C = \mathcal{L}^{\mathrm{RKL}} + \mathcal{L}^B$, and a response to problem x uses the loss of its stratum:

$$\mathcal{L} = \mathcal{L}^A \, \mathbf{1}[x \in \mathcal{D}_A] + \mathcal{L}^C \, \mathbf{1}[x \in \mathcal{D}_C]. \tag{9}$$

Problems outside $\mathcal{D}_+$ give no loss. Each step, the student answers a batch of problems from $\mathcal{D}_+$ once, the teacher scores the answers, and we make one update; gates get no gradient (Algorithm 1). TS-OPD (routing only) removes gating by setting $m_t^A \equiv m_t^C \equiv 1$ and $\beta = 0$: $\mathcal{L}^A$ becomes forward KL averaged over all positions, and $\mathcal{L}^C$ becomes the sampled reverse KL on all positions.

5 Experiments

Models and arms. The teachers are T_0 (base-8B) and T_1 (GRPO-8B); the students are Qwen3-0.6B, 1.7B, and 4B (Yang et al., 2025). Every run trains for one epoch on the 7,496 MATH training problems (Hendrycks et al., 2021). Direct GRPO trains the same student with GRPO and no teacher; it is the S_{GRPO} that Section 3 sets out to beat. We compare it with OPD-r and OPD-f (uniform reverse and forward KL on all problems), and TS-OPD (routing only) and TS-OPD, which train on A∪C only and so take about 91, 51, and 27 of the 118 OPD steps. Three controls, Direct GRPO (A∪C), OPD-r (A∪C), and OPD-f (A∪C), use the same problems and steps as TS-OPD; only the loss differs. Details are in Appendix B.

Benchmarks and metrics. We sample K responses per problem on eight benchmarks: MATH-test (5,000 problems, $K = 1$) and MATH500 ($K = 8$) (Hendrycks et al., 2021; Lightman et al., 2023), Minerva (Lewkowycz et al., 2022) and OlympiadBench (He et al., 2024) ($K = 8$), and AIME24/25 (Art of Problem Solving, n.d.) and HMMT-Feb/Nov25 (HMMT, 2025a;b) ($K = 16$). We report macro avg@K and pass@K over all eight. With $K = 1$, a MATH-test problem is only right or wrong and cannot show a change in reliability, so G_{transfer} (Section 3) and the per-problem analysis use the other seven benchmarks (1,566 problems).

Outcome categories. To trace gains per test problem, we compare a model's sampled accuracy p_0 before and p_1 after training: A if $p_0 = 0 < p_1$ (newly solved), C if $0 < p_0 < p_1$ (more reliable), P if $p_0 > p_1$ (worse), and N otherwise (unchanged); "before" is S_0 for a student and T_0 for the teacher. Test problems are also split into strata by S_0 and T_1 with the rule of Section 3.

5.1 Teacher GRPO Is Necessary but Not Sufficient

Table 2: Main results. avg@K and pass@K are macro averages over the eight benchmarks (%). G_{transfer} is the transfer gap in percentage points, computed on seven benchmarks. ↑/↓: higher/lower is better. Direct GRPO uses no teacher. In each row, bold marks the best value and underline the second best.

		base-8B teacher				GRPO-8B teacher				
				TS-OPD				TS-OPD		Direct
Student	Metric	OPD-r	OPD-f	routing only	full	OPD-r	OPD-f	routing only	full	GRPO
0.6B	avg@K ↑	19.61	19.15	20.60	21.27	27.00	26.57	<u>27.62</u>	**28.88**	26.04
	pass@K ↑	32.80	32.91	34.14	36.22	44.11	41.96	**44.83**	<u>44.57</u>	40.46
	G_{transfer} ↓	15.10	15.51	13.84	13.06	7.39	8.22	<u>6.93</u>	**5.68**	8.71
1.7B	avg@K ↑	29.22	27.57	29.29	29.71	39.04	39.25	<u>39.84</u>	**40.50**	39.53
	pass@K ↑	44.21	44.90	48.49	45.76	56.83	57.57	<u>59.34</u>	**59.78**	57.33
	G_{transfer} ↓	17.40	19.25	17.27	16.71	7.28	7.15	<u>6.43</u>	**5.77**	6.94
4B	avg@K ↑	36.07	36.12	36.53	36.79	52.27	<u>52.83</u>	52.76	**54.28**	51.57
	pass@K ↑	53.93	54.06	52.81	53.31	70.71	<u>72.43</u>	**73.68**	71.04	68.44
	G_{transfer} ↓	20.06	19.99	19.53	19.21	2.65	<u>1.98</u>	2.09	**0.46**	3.45

With the base-8B teacher, every OPD objective trails Direct GRPO on both metrics at every size, and the transfer gap is 13 to 20 points (Table 2). The teacher bounds what OPD can reach: Direct

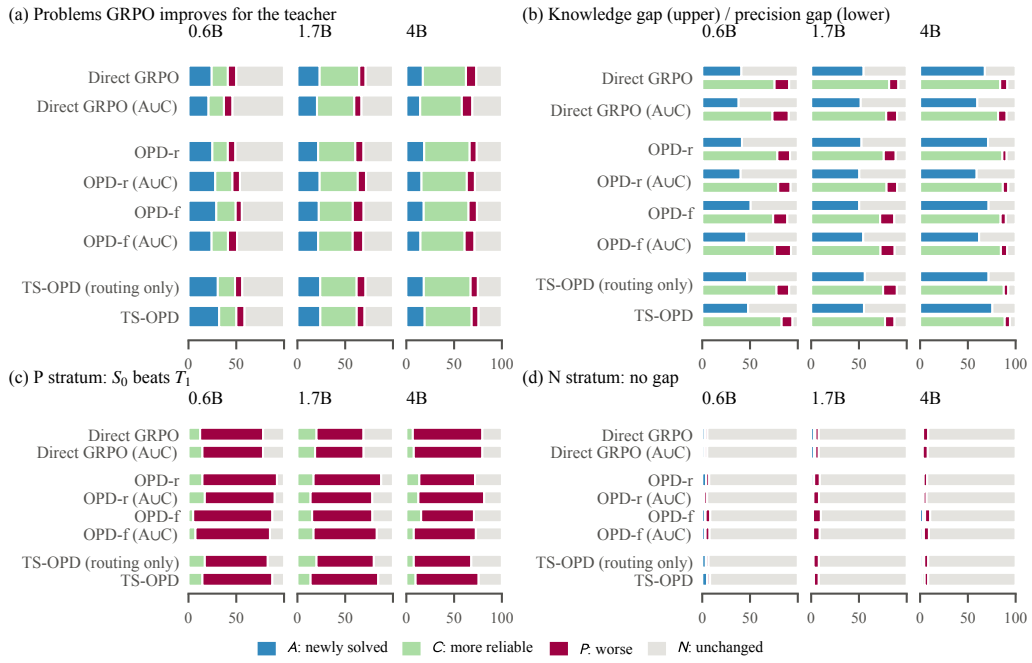


Figure 3: Outcome categories of test problems in each student. (a) The 567 problems GRPO improves for the teacher. (b) Knowledge-gap (upper bar) and precision-gap (lower bar) problems. (c, d) Problems where T_1 is not better than S_0. Numbers: Tables 9, 11, and 10.

GRPO at 1.7B and 4B (39.53 and 51.57) already beats the base-8B teacher itself (38.03). With the GRPO-8B teacher, TS-OPD has the best avg@K and the smallest gap at every size. It is 2.84, 0.97, and 2.71 points above Direct GRPO, and its gap is 5.68, 5.77, and 0.46 points. pass@K is also highest for TS-OPD or TS-OPD (routing only) at every size. Uniform OPD does not reliably help: at 1.7B, both OPD-r and OPD-f fall below Direct GRPO and have a larger gap. The gap nearly closes at 4B because the raw 4B student is almost as strong as T_0 (30.73 vs. 31.54).

5.2 Only Part of the Teacher's GRPO Gain Reaches the Student

G_{transfer} cannot tell whether what GRPO taught the teacher reaches the student (Section 3), so Figure 3(a) follows the 567 test problems that GRPO improves for the teacher. TS-OPD improves 50.1%, 62.1%, and 68.4% of them. Even at 4B, where the gap is almost zero, a third do not improve. The teacher is what makes the difference on these problems: Direct GRPO (A∪C), trained on the same problems without a teacher, improves only 37.1%, 59.5%, and 58.1%, and TS-OPD newly solves the largest share (A) at every size.

Figure 3(b) shows where the teacher helps. Compared with Direct GRPO, TS-OPD newly solves 7.0, 0.4, and 7.9 points more knowledge-gap problems. On precision-gap problems the picture is mixed: TS-OPD is ahead at 0.6B and 4B, but at 1.7B Direct GRPO makes the most problems reliable (81.6%) and the fewest worse (9.7%). This fits Section 3: GRPO can only sharpen what the student already samples, which is what a precision gap needs, and it explains why TS-OPD leads Direct GRPO by only 0.97 points at 1.7B.

The loss comparisons on the same A∪C problems support the choice of KL direction in Section 4.1. Forward KL newly solves more knowledge-gap problems than reverse KL (OPD-f (A∪C) 46.2%, 54.5%, 61.8% vs. OPD-r (A∪C) 40.2%, 50.4%, 59.2%), because it can raise tokens the student rarely picks. Reverse KL makes fewer precision-gap problems worse (13.0%, 11.0%, 5.7% vs. 16.9%, 14.8%, 6.5%), because it only lowers tokens the teacher rejects and leaves the rest of the student's solution alone. Routing uses each loss where it helps and newly solves more knowledge-gap problems than either loss alone (47.0%, 56.2%, 71.7%). Gates then mainly help precision-gap

Table 3: Adding one component at a time: macro avg@K (%) / G_{transfer} (points), GRPO-8B teacher. All A∪C rows use the same problems and steps.

Step	Method	0.6B	1.7B	4B
No teacher	Direct GRPO	26.04 / 8.71	39.53 / 6.94	51.57 / 3.45
+ A∪C only	Direct GRPO (A∪C)	24.91 / 9.34	37.82 / 8.16	47.48 / 4.21
Start	OPD-r	27.00 / 7.39	39.04 / 7.28	52.27 / 2.65
+ filtering	OPD-r (A∪C)	25.66 / 9.02	37.51 / 9.12	47.70 / 7.87
Start	OPD-f	26.57 / 8.22	39.25 / 7.15	52.83 / 1.98
+ filtering	OPD-f (A∪C)	24.87 / 10.17	36.52 / 10.34	47.19 / 8.45
+ routing	TS-OPD (routing only)	27.62 / 6.93	39.84 / 6.43	52.76 / 2.09
+ gating	TS-OPD	28.88 / 5.68	40.50 / 5.77	54.28 / 0.46

problems: TS-OPD makes 5.4, 2.0, and 1.0 points more of them reliable than TS-OPD (routing only), which matches their design of correcting only a few rejected tokens.

5.3 Filtering Helps Only with Routing and Gating

Figure 3(c, d) checks whether the skipped P and N problems matter. On the P stratum, where S_0 beats T_1, every method makes about half or more of the problems worse, including Direct GRPO without a teacher (49.2% to 72.8%). These problems were picked by samples on which S_0 did well, so much of this drop is regression to the mean; training on A∪C only does not change it consistently. Gates lose more of them than TS-OPD (routing only) (73.2% vs. 65.9% at 0.6B), likely because they pull the student harder toward the teacher. On the larger N stratum, more than 89% stay unchanged for every method, and skipping changes the share that gets worse by at most 1.3 points.

Skipping, however, is not free. Table 3 shows that filtering alone hurts: with the same problems and steps as TS-OPD, OPD-r (A∪C) and OPD-f (A∪C) lose 1.3 to 5.6 points of avg@K, and their gap grows at every size. Figure 3(b) locates the loss. On knowledge-gap problems, filtering lowers the newly solved share for all three losses at 0.6B and 4B, by 7.9 to 11.9 points at 4B, where the A∪C runs take only 27 of 118 steps. On precision-gap problems, it does not hurt OPD: the share made reliable stays or rises (by 0.3 to 2.9 points), since the remaining updates fall on problems the student already partly solves. Direct GRPO (A∪C) drops on both, because the student never succeeds on A problems and GRPO gets no reward there. Acquiring a new solution thus needs more supervision than sharpening an existing one.

Routing makes up for this. Forward KL on A problems raises the newly solved share of knowledge-gap problems by 6.8, 5.8, and 12.5 points over OPD-r (A∪C), and reaches the level of OPD on all problems at 1.7B and 4B with fewer steps. As a result, routing lowers the gap by 2.09, 2.69, and 5.78 points, and TS-OPD (routing only) beats OPD-r on all problems in both avg@K and G_{transfer}. Gating lowers the gap by another 1.25, 0.66, and 1.63 points, mainly through precision-gap problems; the 0.6B ablations in Table 7 show that each gate setting and the brake matter.

6 Conclusion

Teacher improvement and student transfer are different objects. In mathematical reasoning, a GRPO-improved teacher can provide useful dense supervision, but only when OPD distinguishes acquisition gaps from consolidation gaps and avoids spending budget on no-gap or student-advantage problems. TS-OPD implements this transfer-aware view with problem-level routing, response-local token gates, and an entropy brake. Across two teacher lineages and three student sizes, the results show that the routed objective is most useful with a GRPO-improved teacher and that its gains depend on the full design. The broader lesson is that stronger teachers should be evaluated not only by endpoint accuracy, but by how their improvements decompose into teachable strata for the student.

LIMITATIONS

Scope. All experiments use mathematical reasoning with verifiable answers, one model family (Qwen3), one teacher size (8B), and one RL algorithm (GRPO) for improving the teacher. The students go up to 4B. We have not tested whether the same strata and routing carry over to other domains, to students close to the teacher in size, or to teachers improved by other RL methods.

Statistical strength. Most runs use a single training seed, so small differences may not be robust. We therefore read the per-problem outcome categories as trends rather than exact rates.

Strata and budget. The strata come from $K_{\mathrm{cal}} = 8$ samples per training problem, taken once before training. With this budget, a problem with no observed student success may still be solvable, and a problem can move between strata as the student improves, but TS-OPD does not relabel it.

Design choices and cost. The hyperparameter ablations use only the 0.6B student with the GRPO-8B teacher. Screening needs $2K_{\mathrm{cal}}$ sampled responses per training problem before training, and every update needs the teacher's top K_{tok} probabilities at each position. Finally, G_{transfer} measures a gain shortfall; it does not isolate how much GRPO-acquired knowledge is lost (Section 3).

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, 2024. URL https://arxiv.org/abs/2306.13649.

Art of Problem Solving. AIME problems and solutions. AoPS Wiki, n.d. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions. Community-maintained problem archive. Accessed September 23, 2026.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

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

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen,

Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 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.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL https://arxiv.org/abs/2402.14008.

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, 2021. URL https://arxiv.org/abs/2103.03874.

HMMT. Archive of february 2025, 2025a. URL https://www.hmmt.org/www/archive/282. Official problem and solution archive. Accessed September 23, 2026.

HMMT. Archive of november 2025, 2025b. URL https://www.hmmt.org/www/archive/291. Official problem and solution archive. Accessed September 23, 2026.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models, 2026. URL https://arxiv.org/abs/2603.07079.

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. Todi: Token-wise distillation via fine-grained divergence control, 2025. URL https://arxiv.org/abs/2505.16297.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models, 2024. URL https://arxiv.org/abs/2402.03898.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms, 2025. URL https://arxiv.org/abs/2503.07067.

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, 2022. URL https://arxiv.org/abs/2206.14858.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe, 2026. URL https://arxiv.org/abs/2604.13016.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners, 2025. URL https://arxiv.org/abs/2502.12143.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Wenze Lin, Jiale Zhao, Xitai Jiang, Songde Rao, Yining Li, Shenzhi Wang, Bingxiang He, and Gao Huang. On-policy distillation with verifiable reward, 2026. URL https://arxiv.org/abs/2608.24696.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Rho-1: Not all tokens are what you need, 2025. URL https://arxiv.org/abs/2404.07965.

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.

Huanxin Sheng, Zhiling Ye, Haonan Wang, Jian Wang, Jinjie Gu, and Jian Kang. 1distillation, 2026. URL https://arxiv.org/abs/2609.24432.

Guanghui Wang, Zhiyong Yang, Zitai Wang, Shi Wang, Qianqian Xu, and Qingming Huang. Abkd: Pursuing a proper allocation of the probability mass in knowledge distillation via α-β-divergence, 2025a. URL https://arxiv.org/abs/2505.04560.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025b. URL https://arxiv.org/abs/2506.01939.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models, 2024. URL https://arxiv.org/abs/2404.02657.

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

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, and Zhipeng Wang. Paced: Distillation and on-policy self-distillation at the frontier of student competence, 2026a. URL https://arxiv.org/abs/2603.11178.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Beyond grpo and on-policy distillation: An empirical sparse-to-dense reward principle for language-model post-training, 2026b. URL https://arxiv.org/abs/2605.12483.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation, 2026c. URL https://arxiv.org/abs/2604.14084.

Yuanda Xu, Zhengze Zhou, Kayhan Behdin, Jelena Markovic-Voronov, Hejian Sang, Xiaomin Li, Wenhui Zhu, Xinchen Du, Aida Rahmattalabi, Ran He, Sen Na, Zhipeng Wang, and Alborz Geramifard. Trek: Distill to explore, reinforce to refine, 2026d. URL https://arxiv.org/abs/2607.05339.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance, 2025. URL https://arxiv.org/abs/2504.14945.

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.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL https://arxiv.org/abs/2504.13837.

Chen Zhang, Qiuchi Li, Dawei Song, Zheyu Ye, Yan Gao, and Yan Hu. Towards the law of capacity gap in distilling language models, 2025. URL https://arxiv.org/abs/2311.07052.

Zhiwei Zhang, Zechen Sun, Fei Zhao, Kang Peng, Bin Liang, Huayu Deng, Yao Hu, Kam-Fai Wong, and Mu Chuan. Verify before you distill: Prompt-level teacher gating for on-policy distillation, 2026. URL https://arxiv.org/abs/2609.02998.

A METHOD DETAILS

A.1 CALIBRATING THE ENTROPY BRAKE

The coefficient β in Eq. 8 is set by

$$\beta = 0.1 \cdot \frac{\text{mean}_{\,m_t^B=1} |r_t|}{\text{mean}_{\,m_t^B=1} H_S(t)}. \tag{10}$$

We compute it once before training for each teacher–student pair and keep it fixed. We take up to 50 student responses on C problems, use only the first 400 tokens of each, and skip responses shorter than 40 tokens. No new text is generated. Both models score these responses in a forward pass: the teacher gives its top 512 tokens plus the tail cell, and the student gives its probabilities on the same 513 cells. From these we compute H_S, H_T, $H_S - H_T$, and $|r_t|$ at every position. We then take the positions whose $H_S - H_T$ is in the top 20%, average $|r_t|$ and H_S over them, and apply Eq. 10. Because H_S is computed on $K_{\text{tok}} + 1$ cells, a larger K_{tok} gives a larger H_S and a smaller β; β therefore depends on both the pair and K_{tok}. Table 4 lists the values used.

Table 4: Entropy-brake coefficient β for each teacher–student pair.

Student	base-8B teacher	GRPO-8B teacher
0.6B	0.078	0.079
1.7B	0.058	0.063
4B	0.053	0.054

A.2 REVERSE KL FROM SAMPLED TOKENS

At one position, write $r(v) = \log \frac{\pi_S(v)}{\pi_T(v)}$, so $D_{\text{R}} = \sum_v \pi_S(v)\, r(v)$ and $r_t = r(y_t)$. Let θ be the student parameters; the teacher is fixed. By the product rule,

$$\nabla_\theta D_{\text{R}} = \sum_v \nabla_\theta \pi_S(v)\, r(v) + \sum_v \pi_S(v)\, \nabla_\theta r(v). \tag{11}$$

The second term is zero. Because π_T does not depend on θ, $\nabla_\theta r(v) = \nabla_\theta \log \pi_S(v)$, and $\sum_v \pi_S(v) \nabla_\theta \log \pi_S(v) = \sum_v \nabla_\theta \pi_S(v) = \nabla_\theta \sum_v \pi_S(v) = \nabla_\theta 1 = 0$. In the first term we use $\nabla_\theta \pi_S = \pi_S \nabla_\theta \log \pi_S$, which gives

$$\nabla_\theta D_{\text{R}} = \sum_v \pi_S(v)\, r(v)\, \nabla_\theta \log \pi_S(v) = \mathbb{E}_{v \sim \pi_S}\big[r(v)\, \nabla_\theta \log \pi_S(v)\big]. \tag{12}$$

So for a sampled token y_t, the product $r_t \nabla_\theta \log \pi_S(y_t \mid s_t)$ is a one-sample estimate of $\nabla_\theta D_{\text{R}}$. With r_t held constant, Eq. 7 has exactly this gradient at the selected positions, so minimizing it lowers reverse KL there.

A.3 TRAINING ALGORITHM

Algorithm 1 TS-OPD training

Require: training set $\mathcal{D}_{\text{train}}$, teacher T_1, raw student S_0

1: Sample K_{cal} responses per problem from T_1 and S_0; build $\mathcal{D}_A$, $\mathcal{D}_C$, and $\mathcal{D}_+ = \mathcal{D}_A \cup \mathcal{D}_C$
2: Calibrate β on student responses to C problems (Appendix A.1)
3: **for** each step **do**
4: Draw a batch from $\mathcal{D}_+$; the student generates one response y per problem
5: The teacher scores y: top-K_{tok} probabilities and the tail cell at each position
6: **if** $x \in \mathcal{D}_A$ **then** compute m_t^A and $\mathcal{L}^A$ (Eq. 5)
7: **else** compute m_t^C, ω_t, m_t^B, and $\mathcal{L}^C$ (Eqs. 6–8)
8: **end if**
9: Stop gradients through the gates; average the losses over the batch and make one update
10: **end for**

B Experimental Setup and Compute

Models and data. Students are Qwen3-0.6B, Qwen3-1.7B, and Qwen3-4B; the teachers are Qwen3-8B (base-8B) and the same model after one epoch of GRPO (GRPO-8B). All training runs use the MATH training set of 7,496 problems. Direct GRPO, OPD-r, and OPD-f train on the full set, and their (A∪C) variants train on the same retained pool as TS-OPD; TS-OPD (routing only) and TS-OPD train only on the retained pool $\mathcal{D}_{\mathrm{A}} \cup \mathcal{D}_{\mathrm{C}}$ of the corresponding teacher–student pair, screened with $K_{\mathrm{cal}} = 8$ samples per model. Every run makes one pass over its training data, and we evaluate the final checkpoint.

Framework. Training uses slime with a Megatron-LM training backend and SGLang for rollout and teacher scoring. The chat template is applied with thinking mode disabled for both training and evaluation.

Evaluation. MATH-test (5,000 problems) uses one sample per problem; MATH500, Minerva, and OlympiadBench use eight; AIME24, AIME25, HMMT-Feb25, and HMMT-Nov25 use sixteen. Sampling uses temperature 1.0 with a response limit of 8,192 tokens for the first four benchmarks and 16,384 tokens for the competition sets. Truncated or unparsable responses count as incorrect.

Compute. All training and evaluation runs in this paper used 1,728 NVIDIA B300 GPU hours in total.

C Training Configurations

Table 5: Training hyperparameters. Shared settings apply to all runs.

Shared	
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$), weight decay 0.1
Learning rate	1×10^{-6}, constant
Responses per update	64
Rollout	temperature 1.0, max response length 8,192
Policy clipping	$\epsilon_{\text{low}} = 0.2$, $\epsilon_{\text{high}} = 0.28$
KL / entropy regularizer	none (coefficients 0)
Direct GRPO	
Batch	16 prompts $\times$ 4 responses
Reward	binary answer correctness
Steps	469 (one epoch)
OPD-r and OPD-f	
Batch	64 prompts $\times$ 1 response
Objective	OPD-r: sampled-token reverse-KL penalty (coefficient 1.0); OPD-f: forward KL on the teacher top-K_{tok} partition
Steps	118 (one epoch)
TS-OPD (routing only) and TS-OPD	
Batch	64 prompts $\times$ 1 response
Training pool	$\mathcal{D}_{\mathrm{A}} \cup \mathcal{D}_{\mathrm{C}}$ of each pair, one epoch
Objective	forward KL on A, sampled reverse KL on C; TS-OPD adds token gates and the entropy brake
Token support	teacher top $K_{\mathrm{tok}} = 512$ plus one tail cell
Gates (TS-OPD only)	within-response quantile 0.8; mass@$n > 0.9$ with $n = \max(\lfloor \ln K_{\mathrm{tok}} \rfloor, 5) = 6$; brake coefficient β measured per pair (Eq. 10)

D Experimental Supplement

Table 6: Training arms. All arms except the two Direct GRPO rows use the GRPO-8B teacher; Table 2 also reports the four OPD objectives with the base-8B teacher.

Arm	Training problems	Loss
Direct GRPO	all	correctness reward, no teacher
Direct GRPO (A∪C)	A∪C	same
OPD-r	all	reverse KL
OPD-f	all	forward KL
OPD-r (A∪C)	A∪C	reverse KL
OPD-f (A∪C)	A∪C	forward KL
TS-OPD (routing only)	A∪C	forward KL on A, reverse KL on C; no gates, no brake
TS-OPD	A∪C	same, plus token gates and the entropy brake

Table 7: Ablations with the 0.6B student and the GRPO-8B teacher (macro averages over eight benchmarks, %). Each row changes one setting of the default TS-OPD.

Setting	avg@K	pass@K
$\tau = 0.60$	27.17	41.45
$\tau = 0.95$	24.71	36.41
20% of positions by student entropy	26.28	40.54
20% of positions at random	26.36	41.34
mass@$n > 0.0$	26.70	39.93
mass@$n > 0.5$	26.75	42.59
mass@$n > 0.7$	26.66	41.11
$n = 1$	25.76	38.26
$n = 3$	27.27	42.70
$n = 10$	25.82	38.53
$\beta = 0$	26.87	41.38
TS-OPD (default)	**28.88**	**44.57**

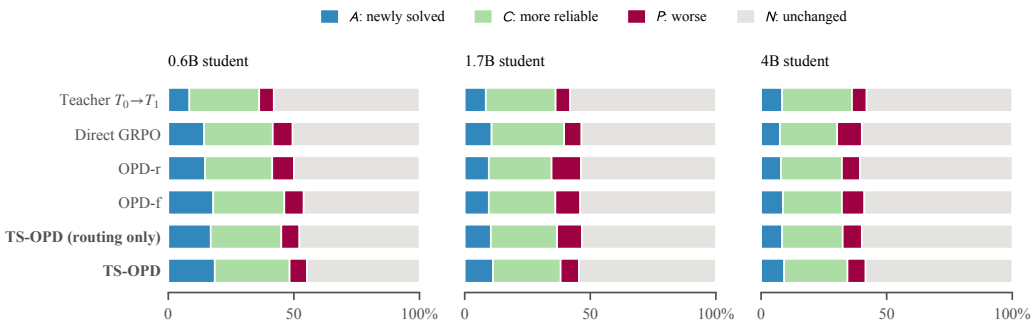


Figure 4: Outcome categories of all 1,566 test problems after training. Students are compared with S_0; the first row shows the teacher from T_0 to T_1.

Table 8: Outcome categories of all 1,566 benchmark problems after training (%), relative to the raw student S_0; the teacher row is $T_0 \to T_1$.

	0.6B				1.7B				4B			
Method	A	C	P	N	A	C	P	N	A	C	P	N
Teacher $T_0 \to T_1$	8.4	27.8	5.8	58.0	8.4	27.8	5.8	58.0	8.4	27.8	5.8	58.0
Direct GRPO	14.3	27.3	7.9	50.6	10.7	28.9	6.9	53.6	7.5	22.8	9.8	59.8
Direct GRPO (A∪C)	13.5	27.0	8.0	51.5	10.2	28.3	7.2	54.3	7.0	22.3	9.8	60.9
OPD-r	14.6	26.8	8.6	50.0	9.6	25.0	11.7	53.7	7.9	24.3	7.2	60.7
OPD-r (A∪C)	15.8	28.7	7.3	48.1	10.0	27.3	9.1	53.6	7.0	24.1	7.9	61.0
OPD-f	17.9	28.3	7.7	46.2	9.6	26.5	9.9	54.0	8.7	23.5	8.9	58.9
OPD-f (A∪C)	14.2	27.2	9.3	49.3	9.1	25.4	10.9	54.6	7.2	23.1	9.0	60.7
TS-OPD (routing only)	16.9	28.1	7.2	47.8	10.4	26.4	9.9	53.4	8.4	24.1	7.6	60.0
TS-OPD	18.5	29.8	6.9	44.9	11.3	26.9	7.3	54.5	9.2	25.2	7.2	58.4

Table 9: Outcome categories, in the student, of the 567 problems that GRPO improves for the teacher (%).

	0.6B				1.7B				4B			
Method	A	C	P	N	A	C	P	N	A	C	P	N
Direct GRPO	24.3	16.9	8.8	49.9	23.3	41.4	6.7	28.6	17.3	45.1	10.8	26.8
Direct GRPO (A∪C)	21.0	16.1	9.0	54.0	20.5	39.0	7.6	33.0	14.5	43.6	11.0	31.0
OPD-r	25.0	16.2	7.9	50.8	21.7	38.9	8.6	30.8	18.7	47.6	7.4	26.3
OPD-r (A∪C)	28.0	18.0	7.9	46.0	23.3	39.9	8.6	28.2	15.9	47.4	8.5	28.2
OPD-f	29.1	20.3	6.5	44.1	22.4	35.6	11.3	30.7	18.2	46.9	8.8	26.1
OPD-f (A∪C)	24.2	17.1	9.5	49.2	21.7	36.2	11.1	31.0	14.8	46.2	10.4	28.6
TS-OPD (routing only)	30.7	18.0	7.6	43.7	23.9	38.1	9.3	28.7	18.3	49.0	7.6	25.0
TS-OPD	32.1	18.0	8.6	41.3	24.0	38.1	8.2	29.8	19.2	49.2	7.2	24.3

Table 10: Outcome categories (%) of test problems where the teacher is not better (Figure 3(c, d)). P stratum: S_0 more accurate than T_1 (41, 65, and 103 problems for 0.6B, 1.7B, and 4B). N stratum: equally accurate (453, 711, and 909 problems). Each row sums to 100 within a size; "-" marks a category that cannot occur. All arms except the two Direct GRPO rows use the GRPO-8B teacher.

	0.6B				1.7B				4B			
Method	A	C	P	N	A	C	P	N	A	C	P	N
P stratum: S_0 more accurate than T_1												
Direct GRPO	-	12.2	65.9	22.0	-	20.0	49.2	30.8	-	6.8	72.8	20.4
Direct GRPO (A∪C)	-	14.6	63.4	22.0	-	18.5	50.8	30.8	-	7.8	71.8	20.4
OPD-r	-	14.6	78.0	7.3	-	16.9	70.8	12.3	-	13.6	58.3	28.2
OPD-r (A∪C)	-	17.1	73.2	9.8	-	13.8	64.6	21.5	-	12.6	68.9	18.4
OPD-f	-	4.9	82.9	12.2	-	15.4	63.1	21.5	-	15.5	55.3	29.1
OPD-f (A∪C)	-	7.3	78.0	14.6	-	16.9	66.2	16.9	-	7.8	65.0	27.2
TS-OPD (routing only)	-	17.1	65.9	17.1	-	20.0	60.0	20.0	-	7.8	60.2	32.0
TS-OPD	-	14.6	73.2	12.2	-	13.8	70.8	15.4	-	9.7	66.0	24.3
N stratum: S_0 and T_1 equally accurate												
Direct GRPO	2.6	0.4	2.4	94.5	3.1	1.4	3.5	92.0	1.7	1.4	5.4	91.5
Direct GRPO (A∪C)	2.4	0.4	2.4	94.7	3.0	1.3	3.7	92.1	1.5	1.3	5.3	91.9
OPD-r	4.0	0.2	3.1	92.7	2.4	0.7	6.0	90.8	1.5	2.2	3.7	92.5
OPD-r (A∪C)	2.0	0.2	2.9	94.9	1.7	0.6	5.9	91.8	1.8	1.9	3.2	93.2
OPD-f	3.1	0.4	4.9	91.6	1.5	0.4	8.3	89.7	3.1	2.1	5.4	89.4
OPD-f (A∪C)	3.3	0.2	4.2	92.3	1.4	0.6	7.0	91.0	2.5	1.5	5.3	90.6
TS-OPD (routing only)	3.8	0.4	2.4	93.4	1.9	0.6	5.6	91.9	2.4	1.9	4.2	91.5
TS-OPD	5.1	0.7	2.4	91.8	2.2	0.7	5.1	92.0	2.3	2.5	3.7	91.4

Table 11: Outcome categories (%) of knowledge-gap and precision-gap test problems (Figure 3(b)). All arms except the two Direct GRPO rows use the GRPO-8B teacher; the A∪C rows share the same problems and steps. Knowledge-gap problems start at 0, so they can only be A or N; precision-gap problems start above 0, so they can only be C, P, or N. There are 517 and 555 such problems for 0.6B, 264 and 526 for 1.7B, and 152 and 402 for 4B.

	0.6B				1.7B				4B			
Method	A	C	P	N	A	C	P	N	A	C	P	N
Knowledge-gap problems: S_0 at 0, T_1 above 0												
Direct GRPO	41.0	-	-	59.0	54.9	-	-	45.1	67.8	-	-	32.2
Direct GRPO (A∪C)	38.1	-	-	61.9	51.9	-	-	48.1	59.9	-	-	40.1
OPD-r	41.6	-	-	58.4	52.7	-	-	47.3	71.1	-	-	28.9
OPD-r (A∪C)	40.2	-	-	59.8	50.4	-	-	49.6	59.2	-	-	40.8
OPD-f	50.7	-	-	49.3	50.3	-	-	49.7	71.7	-	-	28.3
OPD-f (A∪C)	46.2	-	-	53.8	54.5	-	-	45.5	61.8	-	-	38.2
TS-OPD (routing only)	47.0	-	-	53.0	56.2	-	-	43.8	71.7	-	-	28.3
TS-OPD	48.0	-	-	52.0	55.3	-	-	44.7	75.7	-	-	24.3
Precision-gap problems: $0 < S_0 < T_1$												
Direct GRPO	-	75.7	15.3	9.0	-	81.6	9.7	8.7	-	83.8	7.5	8.7
Direct GRPO (A∪C)	-	73.5	17.1	9.4	-	78.5	11.4	10.1	-	81.6	9.0	9.5
OPD-r	-	78.6	13.3	8.1	-	76.0	12.5	11.5	-	86.1	4.5	9.5
OPD-r (A∪C)	-	79.5	13.0	7.6	-	78.9	11.0	10.1	-	86.6	5.7	7.7
OPD-f	-	74.2	14.8	11.0	-	72.2	14.8	12.9	-	84.3	5.5	10.2
OPD-f (A∪C)	-	76.0	16.9	7.0	-	72.6	14.8	12.5	-	84.6	6.5	9.0
TS-OPD (routing only)	-	77.7	13.3	9.0	-	75.3	14.4	10.4	-	87.6	4.7	7.7
TS-OPD	-	83.1	11.4	5.6	-	77.3	10.0	12.7	-	88.6	5.5	6.0

Table 12: Transfer gap G_{transfer} (percentage points) on each benchmark with $K \geq 8$. The last row per student is the macro average used in Table 2.

		base-8B teacher				GRPO-8B teacher				No teacher
Student	Benchmark	OPD-r	OPD-f	TS-OPD (routing only)	TS-OPD	OPD-r	OPD-f	TS-OPD (routing only)	TS-OPD	Direct GRPO
0.6B	MATH500	-3.05	-1.45	-5.92	-6.64	-14.42	-12.17	-15.22	-18.07	-12.55
	Minerva	6.20	5.87	5.32	4.31	-1.52	-0.74	-0.47	-0.79	-0.38
	OlympiadBench	4.63	5.58	3.11	2.26	-4.62	-0.99	-4.14	-4.94	-1.12
	AIME24	30.62	28.75	29.58	27.08	24.16	23.54	21.46	22.50	24.58
	AIME25	26.05	26.46	23.96	22.92	15.00	14.17	13.55	12.09	14.79
	HMMT-Feb25	15.21	15.84	14.59	14.59	8.55	10.00	8.75	5.42	10.42
	HMMT-Nov25	26.05	27.51	26.26	26.89	24.59	23.76	24.59	23.55	25.22
	Macro	15.10	15.51	13.84	13.06	7.39	8.22	6.93	5.68	8.71
1.7B	MATH500	6.42	9.63	6.63	5.70	-2.12	-1.00	-1.95	-2.45	-3.58
	Minerva	8.45	10.52	7.81	6.80	1.47	0.96	1.14	0.50	-0.19
	OlympiadBench	11.09	13.52	10.18	10.07	1.57	3.04	1.22	0.87	-1.65
	AIME24	27.08	30.83	30.20	27.91	12.49	12.70	9.58	9.16	12.08
	AIME25	24.17	24.59	24.80	24.17	11.47	9.17	11.88	10.21	11.05
	HMMT-Feb25	16.88	16.88	14.59	15.21	6.67	6.25	5.63	6.04	8.75
	HMMT-Nov25	27.71	28.75	26.67	27.09	19.38	18.96	17.50	16.05	22.09
	Macro	17.40	19.25	17.27	16.71	7.28	7.15	6.43	5.77	6.94
4B	MATH500	9.13	10.30	9.10	8.37	0.28	0.78	-0.15	-0.78	0.78
	Minerva	7.99	7.58	7.76	7.90	0.64	1.92	1.19	-1.89	1.51
	OlympiadBench	15.15	16.41	14.24	13.43	3.47	3.65	3.37	3.17	5.21
	AIME24	34.58	33.75	32.91	32.29	5.62	2.91	6.04	4.99	5.83
	AIME25	27.08	27.29	27.50	27.71	3.75	1.88	5.21	-1.25	5.62
	HMMT-Feb25	16.67	16.88	16.04	15.21	0.42	-0.62	-3.33	-1.25	0.83
	HMMT-Nov25	29.80	27.71	29.17	29.59	4.38	3.34	2.30	0.21	4.38
	Macro	20.06	19.99	19.53	19.21	2.65	1.98	2.09	0.46	3.45

Table 13: Per-benchmark results for the 0.6B student (%). Direct GRPO uses no teacher. Macro averages weight each benchmark equally.

			Teacher: base-8B				Teacher: GRPO-8B			
Benchmark	Metric	Direct GRPO	OPD-r	OPD-f	TS-OPD (routing only)	TS-OPD	OPD-r	OPD-f	TS-OPD (routing only)	TS-OPD
MATH500	Avg@8	62.58	53.08	51.48	55.95	56.67	64.45	62.20	65.25	68.10
	Pass@8	82.00	78.80	78.00	82.80	81.60	85.00	82.60	84.40	85.40
Minerva	Avg@8	16.87	10.29	10.62	11.17	12.18	18.01	17.23	16.96	17.28
	Pass@8	30.88	25.74	27.21	28.31	26.84	33.46	34.93	32.72	32.35
OlympiadBench	Avg@8	28.86	23.11	22.16	24.63	25.48	32.36	28.73	31.88	32.68
	Pass@8	47.03	44.07	45.10	45.85	48.66	52.23	46.88	50.89	50.15
AIME24	Avg@16	7.50	1.46	3.33	2.50	5.00	7.92	8.54	10.62	9.58
	Pass@16	30.00	10.00	20.00	13.33	20.00	36.67	33.33	43.33	30.00
AIME25	Avg@16	14.38	3.12	2.71	5.21	6.25	14.17	15.00	15.62	17.08
	Pass@16	33.33	20.00	13.33	20.00	23.33	36.67	30.00	40.00	33.33
HMMT-Feb25	Avg@16	6.25	1.46	0.83	2.08	2.08	8.12	6.67	7.92	11.25
	Pass@16	20.00	10.00	10.00	10.00	16.67	20.00	16.67	20.00	26.67
HMMT-Nov25	Avg@16	4.79	3.96	2.50	3.75	3.12	5.42	6.25	5.42	6.46
	Pass@16	13.33	13.33	10.00	13.33	13.33	23.33	23.33	20.00	30.00
MATH-test	Avg@1	67.10	60.42	59.60	59.52	59.34	65.54	67.92	67.30	68.62
Macro Avg	Avg@k	26.04	19.61	19.15	20.60	21.27	27.00	26.57	27.62	28.88
	Pass@k	40.46	32.80	32.91	34.14	36.22	44.11	41.96	44.83	44.57

Table 14: Per-benchmark results for the 1.7B student (%). Direct GRPO uses no teacher. Macro averages weight each benchmark equally.

			Teacher: base-8B				Teacher: GRPO-8B			
Benchmark	Metric	Direct GRPO	OPD-r	OPD-f	TS-OPD (routing only)	TS-OPD	OPD-r	OPD-f	TS-OPD (routing only)	TS-OPD
MATH500	Avg@8	81.38	71.38	68.17	71.17	72.10	79.92	78.80	79.75	80.25
	Pass@8	93.00	90.40	89.60	90.40	89.80	91.80	91.20	92.00	92.00
Minerva	Avg@8	29.00	20.36	18.29	21.00	22.01	27.34	27.85	27.67	28.31
	Pass@8	44.49	37.50	35.66	37.50	37.13	42.28	42.65	44.12	44.12
OlympiadBench	Avg@8	48.94	36.20	33.77	37.11	37.22	45.72	44.25	46.07	46.42
	Pass@8	65.28	59.20	57.57	60.39	60.09	62.91	61.57	63.80	63.35
AIME24	Avg@16	29.58	14.58	10.83	11.46	13.75	29.17	28.96	32.08	32.50
	Pass@16	66.67	33.33	50.00	50.00	40.00	60.00	66.67	66.67	63.33
AIME25	Avg@16	23.54	10.42	10.00	9.79	10.42	23.12	25.42	22.71	24.38
	Pass@16	53.33	33.33	26.67	30.00	33.33	43.33	46.67	53.33	50.00
HMMT-Feb25	Avg@16	11.67	3.54	3.54	5.83	5.21	13.75	14.17	14.79	14.38
	Pass@16	26.67	10.00	16.67	23.33	16.67	33.33	33.33	30.00	40.00
HMMT-Nov25	Avg@16	9.58	3.96	2.92	5.00	4.58	12.29	12.71	14.17	15.62
	Pass@16	26.67	16.67	10.00	23.33	16.67	40.00	36.67	43.33	43.33
MATH-test	Avg@1	82.54	73.28	73.04	72.94	72.40	81.00	81.82	81.48	82.12
Macro Avg	Avg@k	39.53	29.22	27.57	29.29	29.71	39.04	39.25	39.84	40.50
	Pass@k	57.33	44.21	44.90	48.49	45.76	56.83	57.57	59.34	59.78

Table 15: Per-benchmark results for the 4B student (%). Direct GRPO uses no teacher. Macro averages weight each benchmark equally.

			Teacher: base-8B				Teacher: GRPO-8B			
Benchmark	Metric	Direct GRPO	OPD-r	OPD-f	TS-OPD (routing only)	TS-OPD	OPD-r	OPD-f	TS-OPD (routing only)	TS-OPD
MATH500	Avg@8	88.47	80.12	78.95	80.15	80.88	88.97	88.47	89.40	90.03
	Pass@8	94.80	93.60	93.40	93.40	93.20	95.40	94.60	95.60	95.20
Minerva	Avg@8	35.80	29.32	29.73	29.55	29.41	36.67	35.39	36.12	39.20
	Pass@8	48.53	43.75	44.49	43.38	41.54	48.53	51.84	47.79	52.21
OlympiadBench	Avg@8	55.10	45.16	43.90	46.07	46.88	56.84	56.66	56.94	57.14
	Pass@8	67.66	65.28	65.88	66.91	66.47	68.55	70.03	69.58	70.33
AIME24	Avg@16	48.54	19.79	20.62	21.46	22.08	48.75	51.46	48.33	49.38
	Pass@16	76.67	46.67	50.00	43.33	53.33	80.00	83.33	83.33	73.33
AIME25	Avg@16	39.38	17.92	17.71	17.50	17.29	41.25	43.12	39.79	46.25
	Pass@16	73.33	50.00	43.33	46.67	43.33	73.33	73.33	73.33	73.33
HMMT-Feb25	Avg@16	24.17	8.33	8.12	8.96	9.79	24.58	25.62	28.33	26.25
	Pass@16	40.00	20.00	23.33	23.33	20.00	46.67	56.67	66.67	53.33
HMMT-Nov25	Avg@16	31.25	5.83	7.92	6.46	6.04	31.25	32.29	33.33	35.42
	Pass@16	56.67	30.00	30.00	23.33	26.67	63.33	60.00	63.33	60.00
MATH-test	Avg@1	89.86	82.12	82.04	82.10	81.96	89.86	89.64	89.84	90.56
Macro Avg	Avg@k	51.57	36.07	36.12	36.53	36.79	52.27	52.83	52.76	54.28
	Pass@k	68.44	53.93	54.06	52.81	53.31	70.71	72.43	73.68	71.04

Table 16: Per-benchmark results of the raw students and the two teachers (%).

Benchmark	Metric	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	base-8B	GRPO-8B
MATH500	Avg@8	42.23	70.00	81.45	82.30	90.10
	Pass@8	72.60	89.40	93.40	93.60	95.40
Minerva	Avg@8	8.64	20.96	29.46	30.06	37.91
	Pass@8	25.74	35.66	41.54	42.65	50.00
OlympiadBench	Avg@8	15.37	34.92	47.94	47.59	59.96
	Pass@8	37.09	56.82	66.02	67.21	72.26
AIME24	Avg@16	1.04	10.62	23.33	25.42	56.46
	Pass@16	10.00	30.00	46.67	60.00	86.67
AIME25	Avg@16	1.25	6.67	17.08	17.29	45.21
	Pass@16	13.33	23.33	40.00	36.67	73.33
HMMT-Feb25	Avg@16	0.21	3.96	8.54	9.79	26.25
	Pass@16	3.33	13.33	16.67	20.00	50.00
HMMT-Nov25	Avg@16	1.67	3.33	7.29	8.33	36.67
	Pass@16	10.00	13.33	26.67	30.00	63.33
MATH-test	Avg@1	49.70	71.16	82.68	83.46	91.48
Macro Avg	Avg@k	15.01	27.70	37.22	38.03	55.51
	Pass@k	27.72	41.63	51.71	54.20	72.81