ON-POLICY DISTILLATION GUIDED BY EVOLUTION OF INTERNAL PREDICTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation lets a student generate the trajectories on which it receives dense token-level supervision, adapting the training contexts to its own policy and outperforming off-policy post-training methods on reasoning tasks. This student-centric view motivates us to investigate another source of model-specific information: how student predictions evolve across depth. Inspired by prior work on internal policy changes in large language models and depth-contrast decoding, we propose *Internal Guidance Policy Distillation* (IGPD), which combines a frozen teacher with the detached change from a calibrated shallow readout to the final policy. At each sampled prefix, the resulting logit offset modifies the teacher anchor to form an adaptive target. Training jointly aligns the shallow readout with the teacher and distills the final policy toward this target. Because the offset is derived from the current model, it evolves throughout training without requiring a separate reference-policy forward pass. Across two strong-to-weak settings and competition-mathematics benchmarks, IGPD consistently improves over matched OPD and achieves the highest macro-average among the evaluated methods. We hope these findings encourage further exploration of internal prediction dynamics in on-policy distillation.

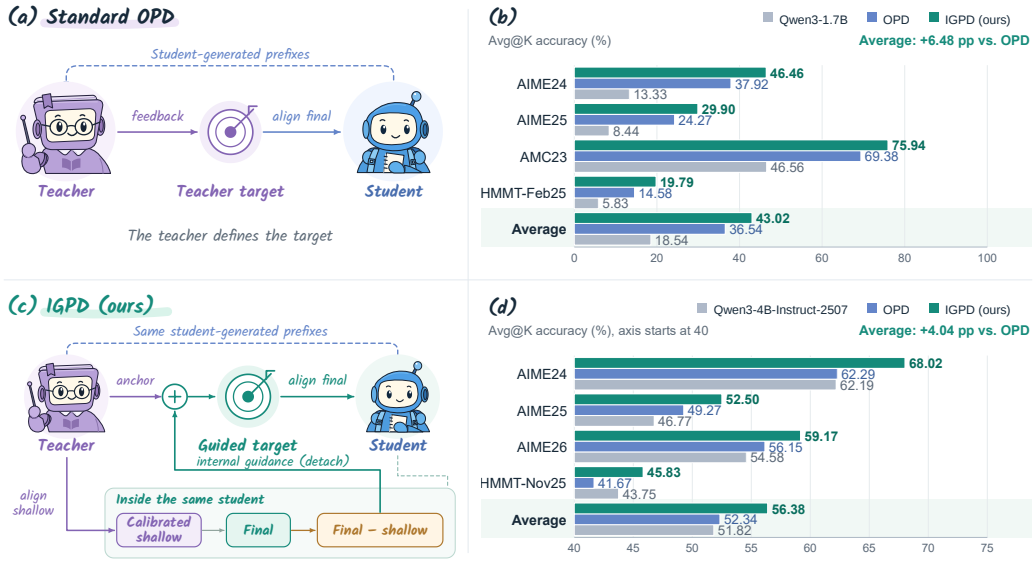

Figure 1: **Method and results overview. Left:** Standard OPD aligns the final student with a teacher target (a); IGPD adds a detached final-minus-shallow offset from the same student and aligns the shallow readout with the teacher anchor (c). **Right:** Avg@K for Qwen3-1.7B in non-thinking mode (b) and Qwen3-4B-Instruct-2507 (d). IGPD improves every displayed benchmark over OPD, with macro-average gains of 6.48 and 4.04 points.

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.

1 INTRODUCTION

Knowledge distillation transfers a stronger teacher's predictive distribution to a smaller student, but conventional offline distillation trains on a fixed set of contexts that need not match those visited by the student at inference (Hinton et al., 2015). On-policy distillation (OPD) reduces this exposure mismatch by sampling trajectories from the student and querying the teacher on the resulting prefixes (Agarwal et al., 2024; Li et al., 2026b). This is particularly useful for mathematical reasoning: a teacher distribution provides dense token-level supervision throughout a long student-generated solution rather than reducing the entire trajectory to a single outcome. This token-level supervision differs from the outcome rewards used in reinforcement-learning recipes for mathematical reasoning (Shao et al., 2024; DeepSeek-AI, 2025; Yu et al., 2025).

Matching the teacher's distribution is a natural starting point, but recent work shows that it need not yield the strongest student. ExOPD improves OPD through reward extrapolation relative to a reference policy (Yang et al., 2026b), while On-Policy Delta Distillation (OPD^2) transfers the difference between a teacher and its pre-tuning base model (Heo et al., 2026). These results motivate studying the supervision target itself. Reference-based constructions also introduce a practical cost: scoring a separate frozen reference or teacher-base model requires additional forward computation on student-generated contexts.

A complementary line of work uses predictive differences during decoding. Contrastive Decoding contrasts an expert with a weaker amateur (Li et al., 2023); DoLa instead contrasts mature final-layer and premature intermediate-layer predictions within one model (Chuang et al., 2024). Both modify inference-time token selection without updating the backbone, so their benefits depend on applying the decoding procedure again at inference. This suggests investigating whether a within-student contrast can inform the training target and improve the student under ordinary decoding.

We propose *Internal Guidance Policy Distillation* (IGPD), which uses the detached change from a calibrated shallow policy to the final policy as a depth offset for the teacher-anchored OPD target. A shallow alignment loss also trains the auxiliary policy toward the teacher. The two student predictions share a backbone forward pass: target construction adds a shallow readout, but requires neither a separate negative-sample model nor another full-model forward pass. By incorporating the contrast into optimization, IGPD trains the backbone with this guidance rather than applying contrastive decoding at deployment. The auxiliary readout is removed after training, leaving ordinary student inference unchanged.

Across two teacher–student configurations, IGPD improves accuracy over matched OPD on all reported competition-math benchmarks, with macro-average gains of 6.48 and 4.04 percentage points. It also exceeds ExOPD in both math macro-averages. The gains accompany different response-length patterns: one student produces longer solutions, while the other improves accuracy with shorter responses. Coefficient and exit-depth ablations examine how the shallow policy contributes to these improvements.

2 INTERNAL GUIDANCE POLICY DISTILLATION

2.1 ON-POLICY DISTILLATION

Let $x \sim \mathcal{D}$ be a prompt and $y = (y_1, \ldots, y_{|y|}) \sim \pi_S(\cdot \mid x)$ a response sampled from the student. At every student-generated token context $s_t = (x, y_{<t})$, standard OPD compares the student policy π_S with a frozen teacher policy π_T (Agarwal et al., 2024; Li et al., 2026b; Yang et al., 2026b). With reverse KL as the token-level divergence, its objective is

$$\mathcal{L}_{\mathrm{OPD}}(\pi_S) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_S(\cdot|x)} \left[\sum_{t=1}^{|y|} D_{\mathrm{KL}}(\pi_S(\cdot \mid s_t) \,\|\, \pi_T(\cdot \mid s_t)) \right]. \tag{1}$$

Every context in Equation 1 comes from the student's own rollout. The loss is normalized by the number of response tokens in the batch.

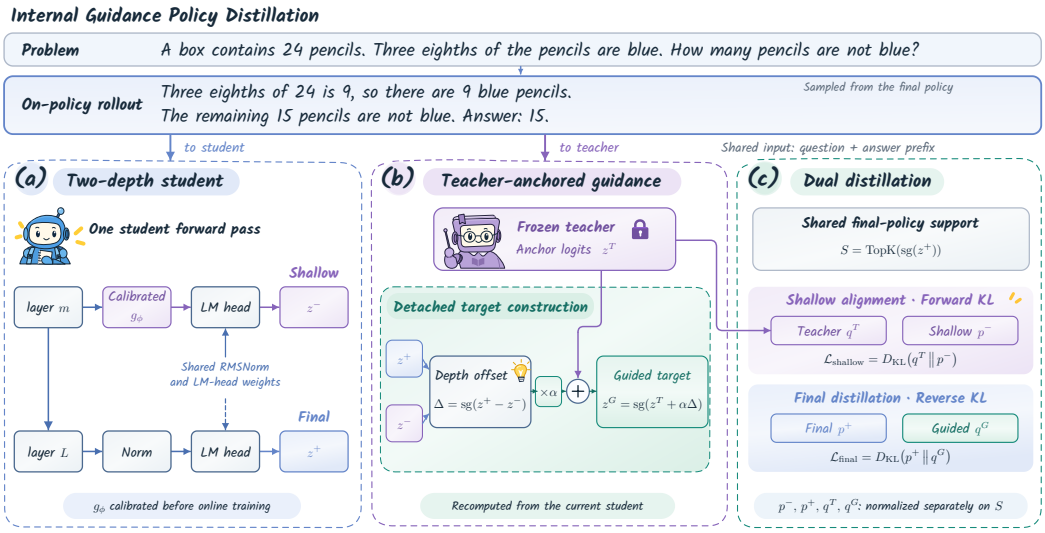

Figure 2: **Detailed overview of IGPD.** Responses are sampled on-policy from the final student policy. A calibrated shallow policy and the final policy define a detached depth offset that modifies the teacher-anchored target. The $-$ and $+$ superscripts denote the shallow and final roles in the offset.

Following the usual trajectory notation, a sampled-token policy-gradient form of OPD is

$$\nabla_\theta \mathcal{L}_{\mathrm{OPD}}^{\mathrm{PG}} = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_S(\cdot|x)}\left[\sum_{t=1}^{|y|} \log \frac{\pi_S(y_t \mid x, y_{<t})}{\pi_T(y_t \mid x, y_{<t})} \nabla_\theta \log \pi_S(y_t \mid x, y_{<t})\right], \tag{2}$$

where the sampled-token log-ratio acts as a token-level coefficient (Yang et al., 2026b). Our experiments differentiate the full categorical reverse KL instead. Treating the sampled contexts, teacher, and support as fixed, the resulting centered gradient is

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\mathrm{OPD}} = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_S(\cdot|x)}\Bigg[\sum_{t=1}^{|y|}\sum_{v} \pi_S(v \mid x, y_{<t}) \Bigg(\log \frac{\pi_S(v \mid x, y_{<t})}{\pi_T(v \mid x, y_{<t})} \\ - \sum_{u} \pi_S(u \mid x, y_{<t}) \log \frac{\pi_S(u \mid x, y_{<t})}{\pi_T(u \mid x, y_{<t})} \Bigg) \nabla_\theta \log \pi_S(v \mid x, y_{<t})\Bigg].\end{aligned} \tag{3}$$

The second sum is the student-probability-weighted mean log ratio, which centers the token-level coefficient. In our implementation, the sums over u and v run over the shared top-k support introduced in Section 2.4, and automatic differentiation is applied to the complete reverse-KL loss. Appendix H gives the corresponding logit gradients for both terms in IGPD.

2.2 Shallow and Final Internal Policies

Let $\boldsymbol{h}_t^\ell$ be the student's hidden state after layer ℓ at the position predicting token y_t. We define the final-layer logits and policy as the final internal policy, denoted by the $+$ superscript:

$$\boldsymbol{z}_t^+ = \boldsymbol{W}_{\mathrm{lm}} \,\mathrm{Norm}(\boldsymbol{h}_t^L), \qquad \pi_t^+ = \mathrm{softmax}(\boldsymbol{z}_t^+). \tag{4}$$

The final policy generates every training trajectory and is the student policy optimized by OPD.

Intermediate representations can be out of distribution for the language-model head, which is trained on final-layer representations. Directly applying this head to intermediate states can therefore yield miscalibrated predictions (Belrose et al., 2023). At exit layer m, we therefore attach a shallow readout

adapter g_ϕ, implemented as a residual SwiGLU transformation:

$$\boldsymbol{u}_t^{\mathrm{sh}} = \mathrm{Norm}(\boldsymbol{h}_t^m), \tag{5}$$

$$\widetilde{\boldsymbol{h}}_t^{\mathrm{sh}} = g_\phi(\boldsymbol{h}_t^m) = \boldsymbol{u}_t^{\mathrm{sh}} + \boldsymbol{W}_{\mathrm{down}} \left[\mathrm{SiLU}(\boldsymbol{W}_{\mathrm{gate}}\boldsymbol{u}_t^{\mathrm{sh}}) \odot (\boldsymbol{W}_{\mathrm{up}}\boldsymbol{u}_t^{\mathrm{sh}})\right], \tag{6}$$

$$\boldsymbol{z}_t^- = \boldsymbol{W}_{\mathrm{lm}}\widetilde{\boldsymbol{h}}_t^{\mathrm{sh}}, \qquad \pi_t^- = \mathrm{softmax}(\boldsymbol{z}_t^-). \tag{7}$$

We call $(\boldsymbol{z}_t^-, \pi_t^-)$ the shallow internal policy. The $-$ and $+$ superscripts denote the shallow and final roles in the depth offset. Both policies belong to the same student and share the final RMSNorm and bias-free language-model head.

Before OPD, we calibrate g_ϕ on a fixed corpus $\mathcal{C}_0$ generated by the initial student. The student backbone, final norm, and language-model head are frozen. Let N_{cal} be the number of response-token contexts in $\mathcal{C}_0$, and let $\mathrm{sg}(\cdot)$ denote stop-gradient. The calibration objective is the token-mean, full-vocabulary forward KL:

$$\mathcal{L}_{\mathrm{cal}} = \frac{1}{N_{\mathrm{cal}}} \sum_{(x,y)\in\mathcal{C}_0} \sum_{t=1}^{|y|} D_{\mathrm{KL}}\left(\mathrm{sg}(\pi_{0,t}^+) \,\|\, \pi_t^-\right), \tag{8}$$

The shallow readout parameters ϕ are the only trainable parameters in this stage. The calibrated readout supplies the shallow policy used in the online depth offset.

2.3 Depth offset and teacher-anchored targets

For each token context, we define the detached shallow-to-final depth offset

$$\boldsymbol{\Delta}_t = \mathrm{sg}\left(\boldsymbol{z}_t^+ - \boldsymbol{z}_t^-\right). \tag{9}$$

Let $f_T(s_t)$ denote the frozen teacher logits at the same token context. We construct the anchor logits and guided logits as

$$\boldsymbol{z}_t^T = \mathrm{sg}(f_T(s_t)), \qquad \boldsymbol{z}_t^G = \mathrm{sg}(\boldsymbol{z}_t^T + \alpha\boldsymbol{\Delta}_t), \tag{10}$$

where $\alpha \geq 0$ controls the strength of internal guidance. The shallow policy is aligned with the anchor $\boldsymbol{z}^T$, while the final policy is aligned with $\boldsymbol{z}^G$. This is a logit-space construction rather than an interpolation between probability distributions. Stop-gradient on the entire target prevents either student branch from changing the target through the depth-offset term within the same update.

Equation 10 computes the offset from two depths of the current student. It therefore changes as the student is updated and adds only the shallow readout to the usual student–teacher OPD pair.

2.4 Final-policy top-k support

Both branches use one support selected from the student's final-layer logits:

$$\mathcal{S}_{K,t} = \mathrm{TopK}_K\left(\mathrm{sg}(\boldsymbol{z}_t^+)\right). \tag{11}$$

For any logits $\boldsymbol{z}$, restriction means gathering the logits indexed by $\mathcal{S}_{K,t}$ and then applying softmax over those K values:

$$\mathcal{R}_{\mathcal{S}}(\boldsymbol{z})(v) = \frac{\exp z_v}{\sum_{u\in\mathcal{S}} \exp z_u}, \qquad v \in \mathcal{S}. \tag{12}$$

The shallow policy, final policy, teacher target, and guided target are independently normalized on the same final-policy top-k support:

$$\begin{aligned} q_t^T &= \mathcal{R}_{\mathcal{S}_{K,t}}(\boldsymbol{z}_t^T), \qquad & q_t^G &= \mathcal{R}_{\mathcal{S}_{K,t}}(\boldsymbol{z}_t^G), \\ p_t^- &= \mathcal{R}_{\mathcal{S}_{K,t}}(\boldsymbol{z}_t^-), \qquad & p_t^+ &= \mathcal{R}_{\mathcal{S}_{K,t}}(\boldsymbol{z}_t^+). \end{aligned} \tag{13}$$

Our default uses $K = 1024$ and $\alpha = 1$.

2.5 INTERNAL-GUIDANCE OPD OBJECTIVE

Our matched OPD baseline uses reverse KL from the final student policy to the teacher distribution, both restricted and renormalized on the support above:

$$\mathcal{L}_{\text{OPD}}^{\text{top-}k}(\pi_S) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_S(\cdot|x)}\left[\sum_{t=1}^{|y|} D_{\text{KL}}(p_t^+ \,\|\, q_t^T)\right]. \tag{14}$$

For shallow alignment, we use forward KL from the detached teacher anchor to the shallow student policy:

$$D_{\text{KL}}(q_t^T \,\|\, p_t^-) = \sum_{v\in\mathcal{S}_{K,t}} q_t^T(v) \log \frac{q_t^T(v)}{p_t^-(v)}. \tag{15}$$

The target distribution q_t^T is detached, so gradients pass only through the shallow student policy p_t^-. This direction is consistent with the full-vocabulary forward-KL objective used to calibrate the readout offline.

The complete IGPD objective aligns the shallow policy with the teacher anchor using forward KL and the final policy with the guided target using reverse KL:

$$\mathcal{L}_{\text{IGPD}}(\pi_S) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_S(\cdot|x)}\left[\sum_{t=1}^{|y|} \left(\lambda_{\text{anchor}} D_{\text{KL}}(q_t^T \,\|\, p_t^-) + D_{\text{KL}}(p_t^+ \,\|\, q_t^G)\right)\right]. \tag{16}$$

Relative to Equation 14, IGPD changes the target of the final student policy from the teacher anchor to the guided target and adds shallow alignment while retaining OPD's reverse-KL direction. The deep term has unit weight; the shallow weight is $\lambda_{\text{anchor}} = 1$ by default. Appendix H gives both logit gradients.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Models and datasets. We evaluate two strong-to-weak settings: a Qwen3-1.7B student distilled from Qwen3-4B-Instruct-2507, and a Qwen3-4B-Instruct-2507 student distilled from Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025). Both configurations use strict non-thinking prompts and draw their training questions from DAPO-Math-17K (Yu et al., 2025). Following BuPO's analysis of shallow-policy exploration in Qwen (Tan et al., 2025), we place the calibrated shallow exit after layer 6.

Training settings. All training experiments use four NVIDIA A100 80GB GPUs. Online runs use a global batch of 128, one rollout per prompt, a 16,384-token response limit, and a 90-step budget. OPD, ExOPD, and IGPD use the same training questions, rollout policy, sampling settings, update budget, and final-policy top-1,024 support. Matched OPD and IGPD also use the same reverse-KL direction for the final policy; IGPD changes its target and adds the shallow alignment term. Following the strong-to-weak ExOPD setting (Yang et al., 2026b), ExOPD uses G-OPD with the frozen initial student as π_{ref} and $\lambda = 1.25$.

Before online training, we generate 12,800 fixed responses from each initial student and fit only g_ϕ for three passes with the full-vocabulary forward-KL objective in Equation 8; the backbone, final RMSNorm, and shared language-model head remain frozen. During each online update, the final student generates the responses, the shallow and final student policies and frozen teacher score the resulting contexts, and the shared support and detached targets are then recomputed before optimizing Equation 16. The teacher and shallow readout are removed after training. We use verl/HybridFlow (Sheng et al., 2025) for training and vLLM (Kwon et al., 2023) for rollout generation. Appendix C provides optimizer, readout, and implementation details.

Table 1: **Competition-math results.** The student is Qwen3-1.7B, and the teacher is Qwen3-4B-Instruct-2507. All methods use strict non-thinking training; the lower panel enables thinking only at evaluation. Bold and underlined entries denote the best and second-best distinct values within each panel. Len. is the macro-average response length over the four benchmarks, in thousands.

	AIME24		AIME25		AMC23		HMMT-Feb25		Average	Len. ↓
Method	Avg@32	Pass@32	Avg@32	Pass@32	Avg@16	Pass@16	Avg@16	Pass@16	Avg.	(k)
Strict non-thinking evaluation										
Initial student	13.33	53.33	8.44	33.33	46.56	<u>90.00</u>	5.83	23.33	18.54	**2.07**
OPD	37.92	70.00	24.27	60.00	69.38	**95.00**	14.58	<u>40.00</u>	36.54	<u>7.65</u>
ExOPD	<u>40.10</u>	<u>73.33</u>	<u>29.27</u>	**66.67**	<u>74.69</u>	**95.00**	<u>17.92</u>	36.67	<u>40.49</u>	9.59
IGPD	**46.46**	**83.33**	**29.90**	<u>63.33</u>	**75.94**	**95.00**	**19.79**	**46.67**	**43.02**	12.93
Δ vs. OPD	+8.54	+13.33	+5.63	+3.33	+6.56	+0.00	+5.21	+6.67	+6.48	+5.28
Thinking-mode evaluation										
Initial student	<u>47.60</u>	**80.00**	<u>34.90</u>	**70.00**	<u>84.22</u>	<u>95.00</u>	<u>25.21</u>	**60.00**	<u>47.98</u>	**15.56**
IGPD	**52.29**	**80.00**	**37.81**	**70.00**	**85.00**	**97.50**	**26.25**	<u>53.33</u>	**50.34**	<u>19.78</u>
Δ vs. Student	+4.69	+0.00	+2.91	+0.00	+0.78	+2.50	+1.04	-6.67	+2.36	+4.22

Table 2: **Out-of-distribution results.** The student is Qwen3-1.7B in non-thinking mode, and the teacher is Qwen3-4B-Instruct-2507. GPQA-D and Minerva report Avg@16; MMLU-Redux reports accuracy over its full 5,700-example test set, and IFEval reports strict-prompt accuracy. The final column averages all four scores. Bold and underlined entries denote the best and second-best distinct scores.

Method	GPQA-D	Minerva	MMLU-R	IFEval	Average
Initial student	30.02	29.89	65.68	67.10	48.17
OPD	<u>35.92</u>	36.74	66.35	67.65	<u>51.67</u>
ExOPD	34.88	<u>36.93</u>	<u>66.44</u>	<u>68.21</u>	51.62
IGPD	**37.25**	**39.11**	**69.79**	**68.39**	**53.64**
Δ vs. OPD	+1.33	+2.37	+3.44	+0.74	+1.97

Evaluation. Across the two configurations, we select main benchmarks from AIME24, AIME25, AIME26, AMC23, HMMT-Feb25, and HMMT-Nov25 (Mathematical Association of America, 2026; Harvard–MIT Mathematics Tournament, 2026). Qwen3-1.7B uses AIME24, AIME25, AMC23, and HMMT-Feb25; Qwen3-4B-Instruct-2507 uses AIME24, AIME25, AIME26, and HMMT-Nov25. AIME benchmarks use Avg@32 and Pass@32, while AMC23 and HMMT use Avg@16 and Pass@16; the macro-average uses Avg@K. Additional evaluations use GPQA-D (Rein et al., 2023), Minerva (Lewkowycz et al., 2022), MMLU-Redux (Gema et al., 2025), IFEval (Zhou et al., 2023), AutoLogi (Zhu et al., 2025), and TheoremQA (Chen et al., 2023); their metrics are specified in Tables 2 and 4. All methods share the same evaluation code. Decoding uses temperature 1.0, top-$p = 1.0$, and a 32,768-token response limit. A fixed 256-problem set from DeepMath-103K (He et al., 2025), deduplicated against the training set and reported benchmarks, guides checkpoint and hyperparameter selection. Unless explicitly marked as thinking-mode evaluation, all 1.7B results use strict non-thinking decoding. Appendix C provides the complete evaluation protocol.

3.2 Main Results

Table 1 shows that IGPD improves Avg@K over matched OPD on all four strict non-thinking benchmarks, raising the macro-average by 6.48 points. The gains are consistent across difficulty levels, ranging from 5.21 points on HMMT-Feb25 to 8.54 points on AIME24. Relative to ExOPD, IGPD raises Avg@K on every benchmark and improves the macro-average by 2.53 points. Pass@K is less uniform: IGPD leads ExOPD on AIME24 and HMMT-Feb25, ties it on AMC23, and trails it by 3.34 points on AIME25. When the same non-thinking-trained checkpoint is evaluated with thinking enabled, IGPD raises the Avg@K macro-average over the initial student by 2.36 points; this setting is reported as an additional evaluation rather than a matched training comparison.

Table 3: **Competition-math results with an instruction-tuned student.** The student is Qwen3-4B-Instruct-2507, and the teacher is Qwen3-30B-A3B-Instruct-2507. Training and evaluation use the same protocol as the strict non-thinking panel of Table 1. Len. is the macro-average response length over the four benchmarks, in thousands.

	AIME24		AIME25		AIME26		HMMT-Nov25		Average	Len. ↓
Method	Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	Pass@32	Avg@16	Pass@16	Avg.	(k)
Initial student	62.19	86.67	46.77	76.67	54.58	**83.33**	43.75	**73.33**	51.82	8.76
OPD	62.29	90.00	49.27	76.67	56.15	**83.33**	41.67	63.33	52.34	7.73
ExOPD	66.67	86.67	**53.96**	80.00	57.60	**83.33**	43.96	70.00	55.55	9.40
IGPD	**68.02**	**93.33**	52.50	**83.33**	**59.17**	**83.33**	**45.83**	**73.33**	**56.38**	**7.30**
Δ vs. OPD	+5.73	+3.33	+3.23	+6.67	+3.02	+0.00	+4.16	+10.00	+4.04	-0.43

The improvements also extend beyond the competition-math sets. Table 2 shows gains over OPD of 1.33 points on GPQA-D, 2.37 on Minerva, 3.44 on MMLU-Redux, and 0.74 on IFEval. Their consistency raises the out-of-distribution macro-average by 1.97 points over OPD and 2.02 points over ExOPD.

The second setting shows gains over matched OPD on all four benchmarks (Table 3), with a macro-average increase of 4.04 points. IGPD exceeds ExOPD on AIME24, AIME26, and HMMT-Nov25, producing a 0.83-point macro-average advantage; ExOPD retains a 1.46-point lead on AIME25 Avg@32. The Pass@K results show a similar pattern: IGPD improves on OPD for AIME24, AIME25, and HMMT-Nov25 and ties it on AIME26. On the additional tasks in Table 4, IGPD improves over OPD on all four benchmarks and has the highest macro-average among the distilled models. The initial student remains slightly higher in the aggregate because of its AutoLogi score. Appendix E reports the full Pass@K curves for both student configurations.

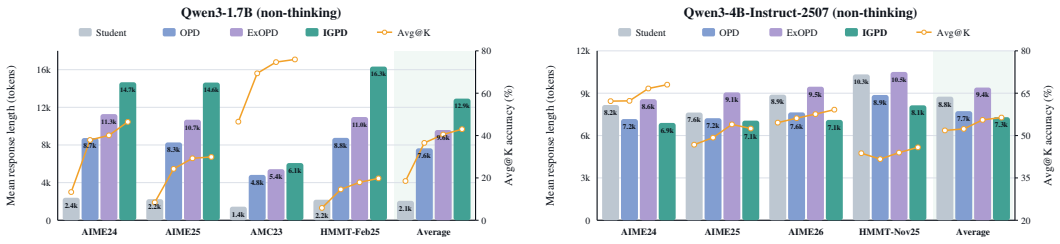

Figure 3: **Response length and accuracy reveal different patterns across students. Left:** Qwen3-1.7B (non-thinking) on AIME24, AIME25, AMC23, HMMT-Feb25, and their average. **Right:** Qwen3-4B-Instruct-2507 on its four main benchmarks. Both panels use strict non-thinking evaluation. Bars show mean response length, and lines show Avg@K.

Accuracy and response length. Figure 3 shows different generation patterns under strict non-thinking evaluation. In the first setting, across AIME24, AIME25, AMC23, and HMMT-Feb25, mean response length increases from 2.07k tokens for the initial student to 7.65k for OPD, 9.59k for ExOPD, and 12.93k for IGPD. Over the same tasks, IGPD attains the highest mean accuracy, 43.02 compared with 36.54 for OPD.

The second setting exhibits the opposite pattern. IGPD produces the shortest responses on every benchmark, with a 7.30k-token average that is 5.5% shorter than OPD, 22.3% shorter than ExOPD, and 16.6% shorter than the initial student. At the same time, its macro-average increases from 52.34 for OPD to 56.38. The effect on response length is therefore model-dependent: the best 1.7B non-thinking result coincides with more extended solutions, whereas the 4B-Instruct result combines higher accuracy with greater concision.

3.3 ABLATIONS

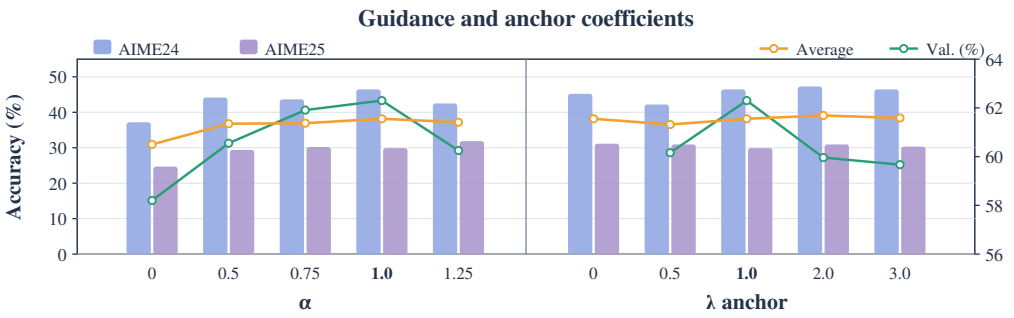

Figure 4: **Guidance and anchor coefficients.** AIME24, AIME25, their mean, and validation accuracy across α and λ_{anchor}. Validation was not recorded for $\lambda_{\text{anchor}} = 0$.

The default coefficients maximize validation accuracy, while $\lambda_{\text{anchor}} = 2$ gives the highest AIME mean. Removing the depth offset ($\alpha = 0$) reduces performance; varying the shallow alignment weight also changes the outcome. The two terms therefore offer distinct ways to influence training.

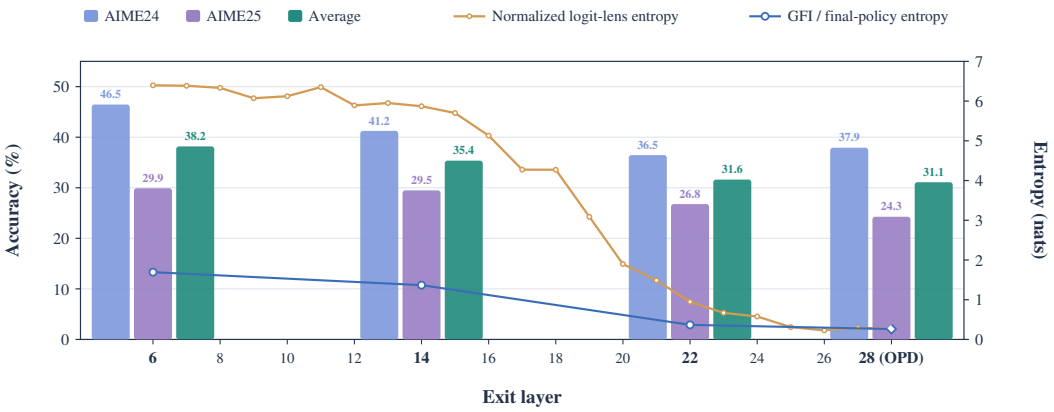

Figure 5: **Internal predictions at different exit depths.** Bars show AIME24, AIME25, and their mean for exits at layers 6, 14, and 22. Layer 28 is the matched OPD reference. Lines show normalized logit-lens entropy and calibrated-readout entropy; the layer-28 endpoint uses the final policy.

The choice of internal policy matters. Figure 5 compares exits at layers 6, 14, and 22 with the same coefficients and training step. Their AIME means are 38.18, 35.36, and 31.61, respectively. Layer 6 improves over the OPD reference by 7.09 points, layer 14 by 4.27, and layer 22 by 0.52. Both AIME24 and AIME25 follow this ordering, so the decline is not driven by only one benchmark.

The later calibrated readouts have lower entropy: 1.69, 1.37, and 0.37 nats at the three exits. The normalized logit lens also becomes more concentrated toward the final layers. Together, these observations favor an earlier calibrated policy in this setting, where the internal prediction can differ substantially from the final policy. If the auxiliary policy is replaced by the final policy itself, the depth offset is zero and the guided final-policy target reduces to the teacher target. The layer-28 bar reports matched OPD as this no-offset reference, not a separately trained layer-28 adapter. Entropy measures concentration rather than agreement between policies. Appendix J also shows next-token predictions at these depths on a fixed erroneous continuation, alongside the final policy and the extrapolated predictions.

3.4 Training dynamics

Appendix B reports the full loss, response-length, and support-mass curves. In both configurations, final reverse KL decreases and the selected support retains over 99.9% of the guided-target mass. The shallow loss need not decrease monotonically as the backbone changes. These curves complement the exit comparison by tracking optimization of the two policies throughout training.

4 Related Work

Distillation for autoregressive models. Knowledge distillation transfers softened teacher outputs to a smaller student (Hinton et al., 2015); sequence- and rationale-level variants instead use generated targets (Kim & Rush, 2016; Hsieh et al., 2023). Intermediate-layer supervision transfers more than final outputs (Sun et al., 2019), while MiniLLM and f-divergence analyses study the choice of distributional objective (Gu et al., 2024; Wen et al., 2023).

On-policy and self-distillation. GKD queries the teacher on student-generated sequences to address the shift from supervised to model-generated contexts (Agarwal et al., 2024). Recent OPD work studies divergence direction, teacher strength, efficient supports, and policy-optimization views, with relaxed imitation or trust-region filtering for difficult teacher–student gaps (Li et al., 2026b; Zhu et al., 2026; Ko et al., 2026; Jang et al., 2026; Xing et al., 2026). Teacher-free variants instead condition on privileged contexts or contrast correct and incorrect contexts (Hübotter et al., 2026; Pan et al., 2026). IGPD obtains an additional target direction from two depths of the same student on its on-policy contexts.

Directional and contrastive targets. G-OPD interprets OPD as KL-regularized policy optimization with an implicit teacher–reference reward, while ExOPD scales that reward relative to its reference-policy constraint (Yang et al., 2026b). Weak-to-Strong OPD, Direct-OPD, and OPD^2 transfer logit or policy shifts between related model states (Yu et al., 2026; Feng et al., 2026; Heo et al., 2026); proxy tuning uses a small model's log-probability shift to steer a larger one (Liu et al., 2024). IGPD derives its offset from the current student's calibrated shallow and final policies and applies it to the teacher.

Intermediate predictions and internal policies. The logit lens exposes intermediate predictions, Tuned Lens calibrates them, and layer-wise analyses track their evolution with depth (Nostalgebraist, 2020; Belrose et al., 2023; He & Su, 2024). At inference, contrastive and early-exit methods compare depths or adapt compute (Li et al., 2023; Chuang et al., 2024; Schuster et al., 2022; Elhoushi et al., 2024), while training-time approaches optimize a layer-induced policy or align intermediate states (Tan et al., 2025; Liu et al., 2026; Yang et al., 2026a). IGPD instead uses the calibrated shallow-to-final logit change to construct the OPD target.

5 Conclusion

We introduced Internal Guidance Policy Distillation, using the student's shallow-to-final prediction change to adapt a teacher-anchored OPD target. A calibrated shallow readout receives direct teacher supervision during training and is removed at inference, leaving the student architecture unchanged.

Across two strong-to-weak settings, IGPD improves matched OPD on every reported main and out-of-distribution benchmark. Relative to ExOPD, IGPD obtains the higher main-task macro-average in both settings, although individual Avg@K and Pass@K comparisons remain benchmark dependent.

In the second setting, the accuracy gains also coincide with more concise generation: IGPD has the shortest mean response on each main benchmark and reduces average length by 5.5% relative to OPD while increasing the Avg@K macro-average. Together, these results identify depth-wise prediction evolution as a useful additional input to on-policy target construction and motivate its exploration in broader model families and post-training tasks.

AI USE STATEMENT

Generative AI tools assisted with technical research and manuscript preparation, including literature review, feedback on experimental design, implementation, assistance with derivations, result analysis and interpretation, and revisions to text, code, and figures. The authors verified the assisted outputs and take responsibility for this work.

REPRODUCIBILITY STATEMENT

Sections 2 and 3 specify the method and experimental protocol; the appendix provides implementation, gradient, overhead, and qualitative details, including training hyperparameters and evaluation settings.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7889–7901, 2023. doi: 10.18653/v1/2023.emnlp-main.489. URL https://aclanthology.org/2023.emnlp-main.489/.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

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

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. LayerSkip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12622–12642. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.681. URL https://aclanthology.org/2024.acl-long.681/.

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

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. Are we done with MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5069–5096. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.262. URL https://aclanthology.org/2025.naacl-long.262/.

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=yNyPawI528.

Harvard–MIT Mathematics Tournament. Problems and results. https://www.hmmt.org/www/archive/problems, 2026. Accessed September 2026.

Hangfeng He and Weijie J. Su. A law of next-token prediction in large language models. *arXiv preprint arXiv:2408.13442*, 2024. URL https://arxiv.org/abs/2408.13442.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. DeepMath-103K: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning, 2025. URL https://arxiv.org/abs/2504.11456.

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 *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran Associates, Inc., 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. On-policy delta distillation. *arXiv preprint arXiv:2607.15161*, 2026. URL https://arxiv.org/abs/2607.15161.

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

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026. URL https://arxiv.org/abs/2601.20802.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42217–42227. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.2094. URL https://aclanthology.org/2026.findings-acl.2094/.

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

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180.

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 *Advances in Neural Information Processing Systems*, 2022.

Ming Li, Chenrui Fan, Yize Cheng, Soheil Feizi, and Tianyi Zhou. Schoenfeld's anatomy of mathematical reasoning by language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32773–32802. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-long.1513. URL https://aclanthology.org/2026.acl-long.1513/.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://arxiv.org/abs/2210.15097.

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

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. Tuning language models by proxy. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=dribhnhm1i.

Xinyu Liu, Darryl Cherian Jacob, Yang Zhou, Jindong Wang, and Pan He. OISD: On-policy internal self-distillation of language models. *arXiv preprint arXiv:2605.29089*, 2026. URL https://arxiv.org/abs/2605.29089.

Mathematical Association of America. American mathematics competitions. https://maa.org/student-programs/amc/, 2026. Accessed September 2026.

Nostalgebraist. Interpreting GPT: The logit lens. Online essay published on LessWrong, August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

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. *arXiv preprint arXiv:2606.11709*, 2026. URL https://arxiv.org/abs/2606.11709.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6fac9e316a4ae75ea244ddcef1982c71-Abstract-Conference.html.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297. ACM, 2025. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 4323–4332. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1441. URL https://aclanthology.org/D19-1441/.

Yuqiao Tan, Minzheng Wang, Shizhu He, Huanxuan Liao, Chengfeng Zhao, Qiunan Lu, Tian Liang, Jun Zhao, and Kang Liu. Bottom-up policy optimization: Your language model policy secretly contains internal policies. *arXiv preprint arXiv:2512.19673*, 2025. URL https://arxiv.org/abs/2512.19673.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 10817–10834. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.605. URL https://aclanthology.org/2023.acl-long.605/.

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

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

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, and Gang Chen. OPRD: On-policy representation distillation. *arXiv preprint arXiv:2606.06021*, 2026a. URL https://arxiv.org/abs/2606.06021.

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

Fangxu Yu, Weijia Xu, Michael Xu, Tianyi Zhou, and Zinan Lin. Weak-to-strong on-policy distillation. *arXiv preprint arXiv:2607.26246*, 2026. URL https://arxiv.org/abs/2607.26246.

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

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Qin Zhu, Fei Huang, Runyu Peng, Keming Lu, Bowen Yu, Qinyuan Cheng, Xipeng Qiu, Xuanjing Huang, and Junyang Lin. AutoLogi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models. *arXiv preprint arXiv:2502.16906*, 2025.

Wenhong Zhu, Ruobing Xie, Rui Wang, and Pengfei Liu. Hybrid policy distillation for LLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2604.20244.

A Additional Out-of-Distribution Results

Table 4: **Additional out-of-distribution results.** The student is Qwen3-4B-Instruct-2507 and the teacher is Qwen3-30B-A3B-Instruct-2507. GPQA-D and Minerva report Avg@16; AutoLogi and TheoremQA report single-sample accuracy, with AutoLogi evaluated over its English and Chinese test sets. The final column averages the four scores. Bold and underlined entries denote the best and second-best distinct scores.

Method	GPQA-D	Minerva	AutoLogi	TheoremQA	Average
Initial student	61.27	**45.43**	**81.05**	48.81	**59.14**
OPD	62.15	45.08	73.76	48.19	57.30
ExOPD	<u>62.88</u>	44.12	70.06	**49.81**	56.72
IGPD	**63.19**	<u>45.40</u>	<u>78.88</u>	<u>49.06</u>	<u>59.13</u>
Δ vs. OPD	+1.04	+0.32	+5.12	+0.87	+1.83

B Training Dynamics

Figures 6 and 7 track the losses, response length, and retained top-k mass for 50 steps.

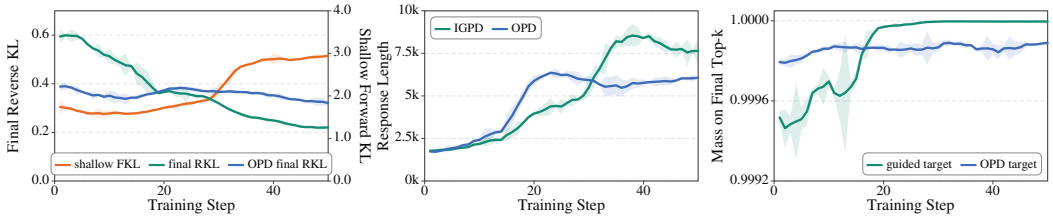

Figure 6: **Training dynamics in the first setting.** The student is Qwen3-1.7B in non-thinking mode. **Left:** shallow forward KL and final reverse KL, with matched OPD. **Middle:** mean response length. **Right:** target mass on the final-policy top-k support.

Shallow forward KL first decreases and then rises, while final reverse KL continues to decrease. The support retains over 99.9% of the guided-target mass, and response length remains below the training cap.

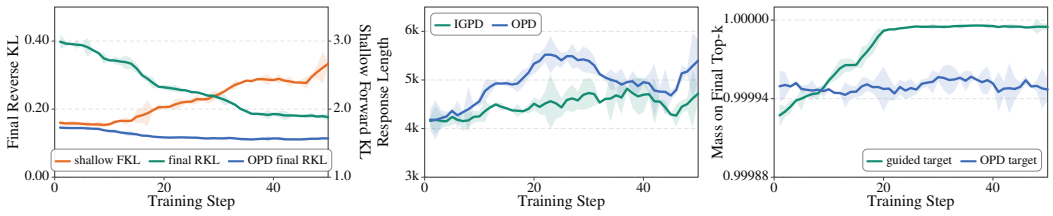

Figure 7: **Training dynamics in the second setting.** The student is Qwen3-4B-Instruct-2507, distilled from Qwen3-30B-A3B-Instruct-2507. The **left**, **middle**, and **right** panels follow Figure 6.

Final reverse KL falls below matched OPD after the initial transient, shallow forward KL remains bounded, and the support retains essentially all guided-target mass. Response length remains comparable to OPD.

C Complete Hyperparameters

Our implementation is based on a verl development snapshot (package version 0.9.0.dev0), with vLLM 0.21.0. Each training run uses four NVIDIA A100-SXM4-80GB GPUs.

For both model pairs, the shallow policy exits after the sixth transformer block (counting blocks from one). Its readout g_ϕ is a residual SwiGLU adapter with intermediate width 3,072.

Table 5: **Online training settings.** Shared by both model pairs.

Setting	Value
Batch size	128
Rollout n	1
Prompt / response limit	2,048 / 16,384
Temperature / top-p	1.0 / 1.0
Optimizer	AdamW
Learning rate	10^{-6}
Schedule	Constant
Gradient clipping	1.0
Training steps	90
Final-policy support K	1,024
Depth-offset scale α	1.0
λ_{anchor}	1

Table 6: **GFI readout calibration.** Shared by both model pairs.

Setting	Value
Readout	Residual SwiGLU
Intermediate width	3,072
Default exit layer	6
Fixed responses	12,800
Batch size	128
Objective	Full-vocabulary FKL
Target	Frozen final student
Optimizer	AdamW
Learning rate	10^{-4}
Weight decay	0.01
Gradient clipping	1.0
Training steps	300

Before online training, we fit only g_ϕ using full-vocabulary forward KL to the frozen final-student policy. For the 1.7B and Qwen3-4B-Instruct-2507 students, we use 12,800 fixed responses and train for three passes (300 updates). Both use batch size 128, learning rate 10^{-4}, and gradient clipping at 1.0. The backbone, final RMSNorm, and shared LM head remain frozen. Evaluation uses temperature 1.0, top-$p = 1.0$, and a maximum response length of 32,768 tokens. We use Avg@32 and Pass@32 for AIME 2024–2026, and Avg@16 and Pass@16 for AMC 2023 and the HMMT 2025 sets. The macro-average includes only Avg@K. We monitor response length and truncation separately. Training uses verl/HybridFlow (Sheng et al., 2025); rollout and evaluation generation use vLLM (Kwon et al., 2023).

For out-of-distribution evaluation, GPQA-Diamond and Minerva Math use Avg@16. Qwen3-1.7B is also evaluated by single-sample accuracy on the full 5,700-example MMLU-Redux test set (Gema et al., 2025) and strict-prompt accuracy on IFEval. Qwen3-4B-Instruct-2507 uses single-sample accuracy on the combined English and Chinese AutoLogi test sets.

The matched OPD baseline and IGPD share student rollouts, final-policy top-k support, final-policy reverse-KL direction, response limits, update budgets, and evaluation code. Following the default strong-to-weak ExOPD setting (Yang et al., 2026b), we use the G-OPD objective, the frozen initial student as π_{ref}, and $\lambda = 1.25$, without a teacher pre-RL checkpoint for reward correction. Our matched top-k realization of this objective is detailed in Section F. The validation set contains 256 competition-math problems from DeepMath-103K (He et al., 2025); we remove normalized exact matches and near-duplicates with DAPO-Math-17K and all test benchmarks.

D Validation Split and Checkpoint Selection

We construct the fixed validation split from the original DeepMath-103K collection. After normalizing question text, we remove exact overlaps with DAPO-Math-17K, duplicate source questions, and overlaps with every benchmark used for final reporting. We then retain self-contained, numeric-answer competition problems and apply deterministic format and consistency filters. This produces an eligible pool of 7,305 problems, from which 256 are sampled once with a fixed seed. The validation examples are never used for gradient updates.

Evaluation on this frozen split provides a reference for checkpoint and hyperparameter selection. Validation evaluations use four independently sampled responses per problem, temperature and top-p both set to 1.0, and a 16,384-token response limit, with the same answer verifier used for test evaluation. The split manifest records the source-row indices, filtering counts, excluded benchmark files, and random seed to keep the validation data fixed across evaluations.

Table 7: **Validation-set record.** Data and settings used for validation evaluations.

Item	Value
Source collection	DeepMath-103K (103,022 rows)
Eligible pool after deterministic filtering	7,305 problems
Frozen validation split	256 problems
Responses per problem	4
Temperature / top-p	1.0 / 1.0
Maximum response length	16,384 tokens
Sampling seed	20260915

E PASS@K RESULTS

For a problem with c correct responses among n samples, we estimate Pass@K as $1 - \binom{n-c}{K}/\binom{n}{K}$ and average across problems. Figures 8 and 9 report the resulting sampling curves for both student configurations. AIME benchmarks use $n = 32$; AMC23 and HMMT use $n = 16$. Accordingly, the AIME panels extend to Pass@32, while AMC23, HMMT, and the four-benchmark macro-average end at Pass@16.

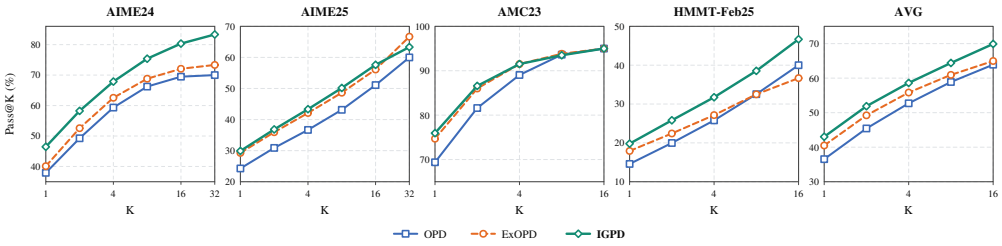

Figure 8: **Pass@K curves in the first setting.** The student is Qwen3-1.7B in non-thinking mode. The first four panels show benchmark-specific curves; the final panel reports their unweighted macro-average at each shared sampling budget.

IGPD leads throughout the AIME24 curve and through Pass@16 on AIME25; ExOPD is higher at Pass@32 on AIME25. On HMMT-Feb25, the IGPD margin grows with the sampling budget, reaching 46.67 Pass@16 compared with 40.00 for OPD and 36.67 for ExOPD. AMC23 is nearly saturated at larger budgets: IGPD improves Pass@1, while all three trained methods solve 95% of the problems with 16 samples.

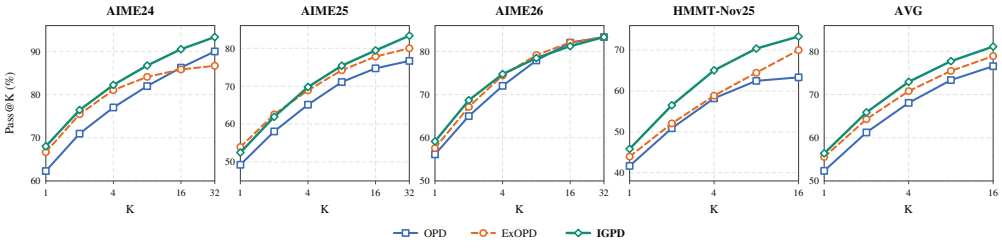

Figure 9: **Pass@K curves in the second setting.** The student is Qwen3-4B-Instruct-2507. The first four panels show benchmark-specific curves; the final panel reports their unweighted macro-average at each shared sampling budget.

In Figure 9, IGPD has the highest Pass@K throughout AIME24 and HMMT-Nov25. On AIME25, ExOPD leads at one and two samples, while IGPD leads from four samples onward. The three

methods are closer on AIME26: IGPD leads through Pass@4, ExOPD leads at Pass@8 and Pass@16, and the three methods tie at Pass@32.

F Implementation Details

Matched ExOPD implementation. The official ExOPD implementation applies a policy-gradient loss to the sampled token (Yang et al., 2026b). In the single-teacher G-OPD setting, its detached token-level coefficient can be written as

$$A_t^{\mathrm{Ex}}(v) = \lambda \log \frac{\pi_T(v \mid s_t)}{\pi_{\mathrm{ref}}(v \mid s_t)} - \log \frac{\pi_S(v \mid s_t)}{\pi_{\mathrm{ref}}(v \mid s_t)}, \tag{17}$$

where π_{ref} is the frozen initial student. Rearranging gives

$$A_t^{\mathrm{Ex}}(v) = \log \frac{\pi_T(v \mid s_t)^{\lambda} \pi_{\mathrm{ref}}(v \mid s_t)^{1-\lambda}}{\pi_S(v \mid s_t)}. \tag{18}$$

After normalization, the distributional optimum induced by this coefficient is therefore

$$q_t^{\mathrm{Ex}}(v) = \frac{\pi_T(v \mid s_t)^{\lambda} \pi_{\mathrm{ref}}(v \mid s_t)^{1-\lambda}}{\sum_u \pi_T(u \mid s_t)^{\lambda} \pi_{\mathrm{ref}}(u \mid s_t)^{1-\lambda}}. \tag{19}$$

Indeed, at a fixed context, the score-function identity gives

$$\mathbb{E}_{v \sim \pi_S(\cdot \mid s_t)} \left[A_t^{\mathrm{Ex}}(v) \nabla_\theta \log \pi_S(v \mid s_t) \right] = -\nabla_\theta D_{\mathrm{KL}} \left(\pi_S(\cdot \mid s_t) \,\|\, q_t^{\mathrm{Ex}} \right). \tag{20}$$

The context-dependent log normalizer vanishes from this expectation because it is independent of v. To match the full-categorical OPD and IGPD implementations, we use the induced optimum as a fixed target on the same detached final-policy support $\mathcal{S}_{K,t}$. If $\boldsymbol{z}_t^{\mathrm{ref}}$ denotes the frozen-reference logits, the implemented target is

$$q_{t,K}^{\mathrm{Ex}} = \mathcal{R}_{\mathcal{S}_{K,t}} \left(\lambda \boldsymbol{z}_t^T + (1-\lambda) \boldsymbol{z}_t^{\mathrm{ref}} \right) = \mathcal{R}_{\mathcal{S}_{K,t}} \left(\boldsymbol{z}_t^T + (\lambda - 1)(\boldsymbol{z}_t^T - \boldsymbol{z}_t^{\mathrm{ref}}) \right), \tag{21}$$

and the student minimizes $D_{\mathrm{KL}}(p_t^+ \,\|\, q_{t,K}^{\mathrm{Ex}})$ by differentiating the complete categorical loss. Thus, our ExOPD baseline changes the official sampled-token gradient estimate to the same complete top-K categorical KL used by OPD and IGPD, while retaining the target implied by the official objective. Equation 21 reduces to the matched OPD target when $\lambda = 1$.

The offline forward-KL calibration objective is exact over the complete vocabulary. Projection is chunked over token positions only to bound memory; the vocabulary is not truncated during this stage. During on-policy distillation, both branches use the same support selected from the detached final student logits as defined in Equation 11. The implementation first gathers the shallow, final, anchor-target, and guided-target logits on those indices and then applies softmax separately to each gathered tensor. Consequently, all four distributions are independently normalized on the same support.

G Computational and Training Overhead

Additional computation. Standard OPD already performs the final-student and frozen-teacher forward passes on every on-policy trajectory. Relative to this baseline, IGPD adds the shallow readout g_ϕ, a shared-LM-head projection for the shallow branch, and the shallow forward-KL term. The readout g_ϕ contains 18.9M parameters for Qwen3-1.7B and 23.6M parameters for Qwen3-4B-Instruct-2507. These parameters are used only during training and are removed together with the shallow branch at inference time.

Training hardware. All training experiments, including offline readout calibration, use four NVIDIA A100-SXM4-80GB GPUs.

Offline readout calibration. The fixed calibration corpus and shallow readout are prepared once per student and can be reused across online runs. Excluding corpus generation, one offline readout update takes 5.7 seconds for Qwen3-1.7B and 23.6 seconds for Qwen3-4B-Instruct-2507 on the same four-GPU node. Summing the 300 logged update durations gives 28.6 and 117.9 minutes, respectively. This one-time optimization is performed before online training. As shown in Figure 10, the Qwen3-1.7B full-vocabulary forward-KL loss reaches final-ten-update averages of 1.45, 1.17, and 0.13 at exit layers 6, 14, and 22, respectively. For Qwen3-4B-Instruct-2507 at layer 6, the calibration loss decreases from 13.75 to a final-ten-update average of 1.60.

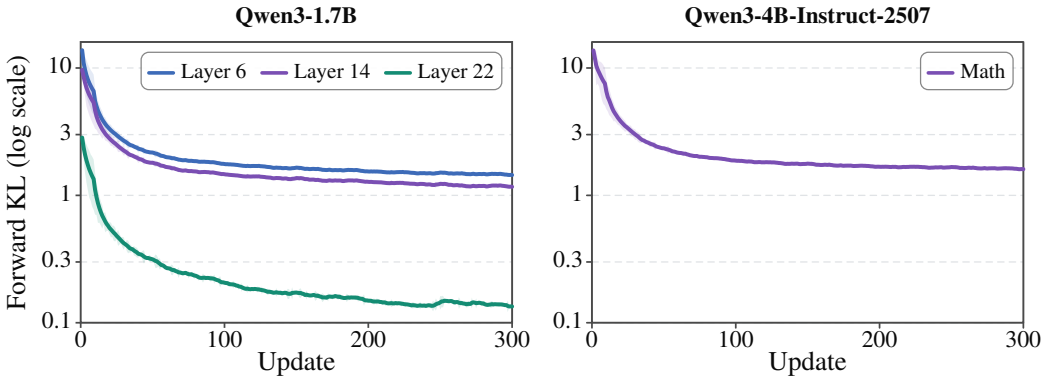

Figure 10: **Offline shallow-readout calibration.** Full-vocabulary forward-KL loss (logarithmic vertical scale) over three passes through the fixed 12,800-response corpus. **Left:** Qwen3-1.7B at exit layers 6, 14, and 22. **Right:** Qwen3-4B-Instruct-2507 at exit layer 6, using math responses.

H Gradient Derivations

We derive the gradients at one token context s_t; the full objective averages these terms over all valid response tokens. Let $j \in \mathcal{S}_{K,t}$. Because the support $\mathcal{S}_{K,t}$, teacher anchor q_t^T, and guided target q_t^G are detached, they are constant in the following derivatives.

Shallow forward KL. The weighted shallow term is

$$\ell_t^- = \lambda_{\text{anchor}} \sum_{v \in \mathcal{S}_{K,t}} q_t^T(v) \log \frac{q_t^T(v)}{p_t^-(v)}. \tag{22}$$

Using the Kronecker delta δ_{vj}, the log-softmax Jacobian is $\partial \log p_t^-(v)/\partial z_{t,j}^- = \delta_{vj} - p_t^-(j)$. The target-entropy part $\sum_v q_t^T(v) \log q_t^T(v)$ is constant, so

$$\frac{\partial \ell_t^-}{\partial z_{t,j}^-} = -\lambda_{\text{anchor}} \sum_{v \in \mathcal{S}_{K,t}} q_t^T(v)\big(\delta_{vj} - p_t^-(j)\big) \tag{23}$$

$$= -\lambda_{\text{anchor}} q_t^T(j) + \lambda_{\text{anchor}} p_t^-(j) \sum_{v \in \mathcal{S}_{K,t}} q_t^T(v) \tag{24}$$

$$= \lambda_{\text{anchor}} \big(p_t^-(j) - q_t^T(j)\big). \tag{25}$$

Thus, the shallow term has the usual soft-target cross-entropy gradient. For $j \notin \mathcal{S}_{K,t}$, the derivative is zero because restriction gathers only the detached support before applying softmax.

Final-policy reverse KL. The final term has fixed unit coefficient:

$$\ell_t^+ = \sum_{v \in \mathcal{S}_{K,t}} p_t^+(v) \log \frac{p_t^+(v)}{q_t^G(v)}. \tag{26}$$

Since $\partial p_t^+(v)/\partial z_{t,j}^+ = p_t^+(v)(\delta_{vj} - p_t^+(j))$ and q_t^G is detached, the product rule gives

$$\frac{\partial \ell_t^+}{\partial z_{t,j}^+} = \sum_{v \in \mathcal{S}_{K,t}} p_t^+(v)\big(\delta_{vj} - p_t^+(j)\big)\left(\log \frac{p_t^+(v)}{q_t^G(v)} + 1\right) \tag{27}$$

$$= p_t^+(j)\left(\log \frac{p_t^+(j)}{q_t^G(j)} + 1\right) - p_t^+(j) \sum_{v \in \mathcal{S}_{K,t}} p_t^+(v)\left(\log \frac{p_t^+(v)}{q_t^G(v)} + 1\right) \tag{28}$$

$$= p_t^+(j)\left[\log \frac{p_t^+(j)}{q_t^G(j)} - \sum_{v \in \mathcal{S}_{K,t}} p_t^+(v) \log \frac{p_t^+(v)}{q_t^G(v)}\right] \tag{29}$$

$$= p_t^+(j)\left[\log \frac{p_t^+(j)}{q_t^G(j)} - D_{\mathrm{KL}}(p_t^+ \,\|\, q_t^G)\right]. \tag{30}$$

The second term in brackets is the exact centering term produced by the softmax Jacobian when the complete reverse KL is differentiated.

Combined parameter gradient. For a batch containing N_{tok} valid response-token contexts, the implemented token-mean gradient is

$$\nabla_\theta \mathcal{L}_{\mathrm{IGPD}} = \frac{1}{N_{\mathrm{tok}}} \sum_t \sum_{v \in \mathcal{S}_{K,t}} \Bigg[\lambda_{\mathrm{anchor}}\big(p_t^-(v) - q_t^T(v)\big)\nabla_\theta z_{t,v}^- + p_t^+(v)\left(\log \frac{p_t^+(v)}{q_t^G(v)} - D_{\mathrm{KL}}(p_t^+ \,\|\, q_t^G)\right)\nabla_\theta z_{t,v}^+\Bigg]. \tag{31}$$

The two direct logit gradients add on student parameters shared by the shallow and final branches. Stop-gradient on $\boldsymbol{z}_t^G = \mathrm{sg}(\boldsymbol{z}_t^T + \alpha(\boldsymbol{z}_t^+ - \boldsymbol{z}_t^-))$ prevents an additional path through the target construction, and detaching Equation 11 prevents gradients through top-k selection.

I Reasoning-Episode Composition

We assign sentences in 200 fixed MATH-500 traces (Hendrycks et al., 2021; Lightman et al., 2024) to the eight ThinkARM reasoning episodes (Li et al., 2026a) using a local frozen Qwen3-30B-A3B-Instruct-2507 classifier and the published protocol (Yu et al., 2026). Each token inherits its sentence label. Figure 11 ranks realized response tokens by three scores: the teacher–final gap, the final–shallow depth offset, and the resulting guided–final gap. For each score, it shows the reasoning-episode composition within the global top 1%; the first row gives the unfiltered reference distribution. We compute this statistic after training.

	Read	Analyze	Plan	Implement	Explore	Verify	Monitor	Answer
All tokens	8.0%	27.2%	4.0%	43.0%	6.9%	5.1%	2.6%	3.1%
Teacher-final	6.0%	**37.8%**	5.3%	**21.2%**	12.8%	7.5%	5.6%	3.8%
Final-shallow	7.8%	**35.0%**	4.0%	**31.9%**	8.5%	6.9%	3.4%	2.5%
Guided-final	6.0%	**37.4%**	5.6%	**19.9%**	**14.0%**	8.0%	5.6%	3.6%

Lower share — Higher share

Change relative to all tokens

Figure 11: **Reasoning-episode composition of high-score token subsets.** Cells report episode shares (%) among all 515,249 realized tokens or among the global top 1% ranked by each score. Shading indicates the percentage-point change relative to all tokens; each row sums to 100%.

All three selected subsets contain fewer *Implement* tokens and more *Analyze* tokens than the full pool. The guided–final ranking shows the largest shift: 19.9% versus 43.0% for *Implement*, and 37.4% versus 27.2% for *Analyze*.

J TOKEN-LEVEL VISUALIZATION

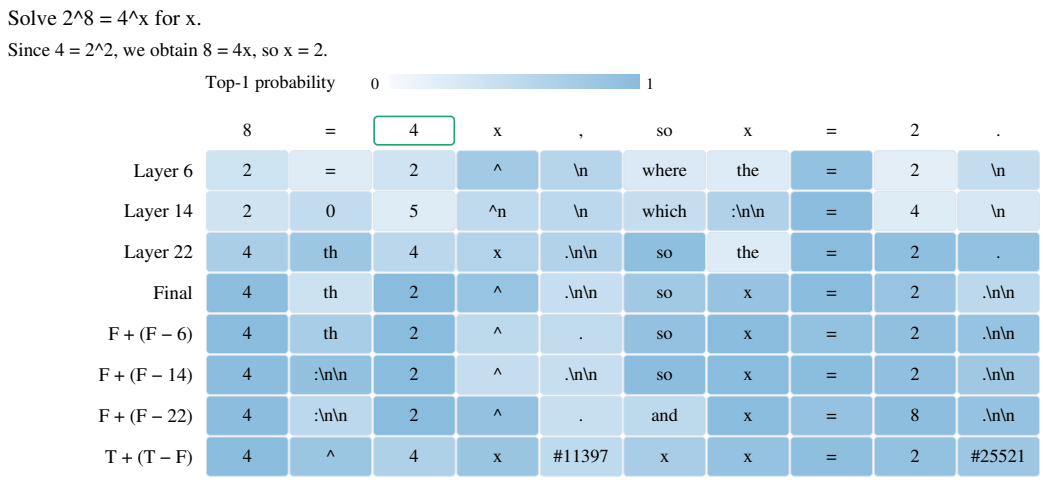

Figure 12: **Layer-wise next-token predictions on a fixed erroneous continuation.** Columns share the supplied prefix above the grid. All rows use the same final-student top-1,024 support at each position and are independently softmax-normalized at temperature 1. Cells show the top-1 token within this support, colored by its probability on a shared 0–1 scale. The initial Qwen3-1.7B non-thinking student uses layer-6/14/22 GFI readouts calibrated for 300 steps: final RMSNorm, residual adapter, then shared LM head. With F denoting layer 28, extrapolated rows use $z^F + (z^F - z^m)$ for $m \in \{6, 14, 22\}$ and $z^T + (z^T - z^F)$, with unit offsets. Teacher T is Qwen3-4B-Instruct-2507 with its native template. Green marks the first error in this constructed continuation ($8 = 2x$ is correct). Space markers and space-only columns are omitted; \n denotes newlines and # prefixes non-Latin token IDs.

Figure 13 shows centered teacher–final and final–shallow logit differences on two fixed erroneous responses. It omits the outer $p_t^+(y_t)$ factor; boxes mark the first errors.

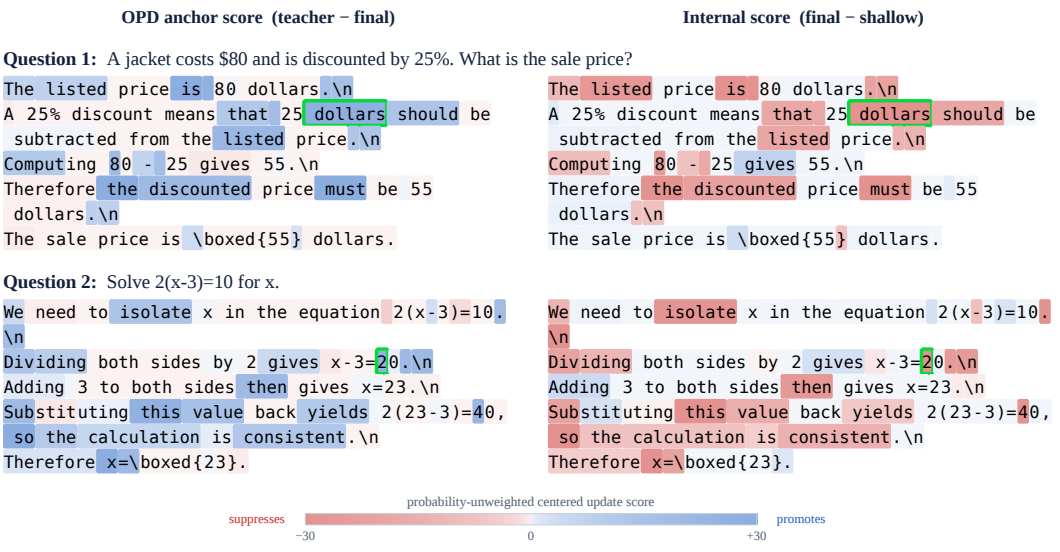

Figure 13: **Probability-unweighted centered token update scores. Left:** the centered OPD anchor score. **Right:** the centered internal score. Blue promotes and red suppresses the realized token; gray lies outside the final-policy top-k support.

K QUALITATIVE EXAMPLES

Figures 14 and 15 compare decisive reasoning steps and final answers on AIME24. Displayed rates are per-problem Avg@32.

Problem **Ground truth: 540**

Find the largest possible real part of $(75+117i)z+\frac{96+144i}{z}$, where $|z|=4$.

Initial student response **Avg@32: 59.38%**

The response sets $z=x+iy$ and correctly reduces the real part to $81x-108y$ under $x^2+y^2=16$. It writes the stationarity conditions as $81=2\lambda x$ and $-108=2\lambda y$, but solves $18225/(4\lambda^2)=16$ as $\lambda^2=18225/1024$ instead of $18225/64$. This produces $(x,y)=(9.6,-12.8)$, which does not satisfy the circle constraint, and the response nevertheless substitutes these values into the objective.
Final answer: 2144.

OPD response **Avg@32: 62.50%**

The response again reaches $81x-108y$ with $x^2+y^2=16$. It then simplifies the ratio of the stationarity equations incorrectly to $x/y=-1/2$ and hence $y=-2x$; the correct relation is $x/y=-3/4$. Enforcing the circle constraint along this wrong direction gives $(x,y)=(4/\sqrt{5},-8/\sqrt{5})$. The response then substitutes that point into the objective and even treats it as aligned with the gradient $(81,-108)$.
Final answer: $1188\sqrt{5}/5$**.**

IGPD response **Avg@32: 96.88%**

The response writes the objective as $81x-108y$ and applies Cauchy–Schwarz directly. Since the coefficient vector has norm $\sqrt{81^2+108^2}=135$ and (x,y) lies on the radius-4 circle, the objective is at most $135\cdot 4=540$. It also identifies the equality point $(x,y)=(12/5,-16/5)$, verifies that it satisfies the constraint, and obtains the same value by substitution.
Final answer: 540.

Figure 14: **AIME24 reasoning comparison in the first setting.** Qwen3-1.7B uses non-thinking mode; panels summarize the same sampled slot.

Problem **Ground truth: 104**

Let $ABCD$ be a tetrahedron with $AB=CD=\sqrt{41}$, $AC=BD=\sqrt{80}$, and $BC=AD=\sqrt{89}$. Its inradius is $m\sqrt{n}/p$, where $\gcd(m,p)=1$ and n is square-free. Find $m+n+p$.

Initial student response **Avg@32: 50.00%**

The response uses $r=3V/S$. Its disphenoid calculation gives $V=80/3$ from the three pairs of opposite edges. It correctly finds that all four faces have side lengths $\sqrt{41},\sqrt{80},\sqrt{89}$ and area $6\sqrt{21}$, hence $S=24\sqrt{21}$. The volume formula, however, is missing a factor of two: the correct volume is $160/3$. Using the smaller volume in $r=3V/S$ gives $r=10\sqrt{21}/63$.
Final answer: $10+21+63=94$**.**

OPD response **Avg@32: 87.50%**

The response constructs symmetric coordinates but obtains $a^2=8$, $b^2=9/4$, and $c^2=12$ from an incorrect correspondence between the coordinate differences and the three edge pairs. Those coordinates yield $V=16\sqrt{6}$ and face area $2\sqrt{141}$ within its own construction, rather than the geometry specified by the problem. It consequently uses $S=8\sqrt{141}$ and obtains $r=6\sqrt{94}/47$.
Final answer: $6+94+47=147$**.**

IGPD response **Avg@32: 96.88%**

The response takes $A=(5/2,2,4)$, $B=(-5/2,-2,4)$, $C=(5/2,-2,-4)$, and $D=(-5/2,2,-4)$, which satisfy all six edge constraints. The scalar triple product of three edge vectors has magnitude 320, giving $V=160/3$. A representative face cross product has norm $12\sqrt{21}$, so every face has area $6\sqrt{21}$ and $S=24\sqrt{21}$. It then computes $r=3V/S=20\sqrt{21}/63$ and verifies the required coprimality and square-free conditions.
Final answer: $20+21+63=104$**.**

Figure 15: **AIME24 reasoning comparison in the second setting.** With Qwen3-4B-Instruct-2507, only IGPD reaches the reference answer in this example.