NOT ALL DISAGREEMENT IS EQUALLY TEACHABLE: TOKEN TEACHABILITY IN ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains a student on its own rollouts with token-level teacher supervision. Recent selective OPD methods exploit the non-uniformity of OPD signals by prioritizing high-entropy or high-disagreement tokens. We revisit this principle and ask: *which teacher signals are more teachable under the student's current predictive state?* Using a fixed-context diagnostic that measures same-context teacher–student KL reduction, we show that raw disagreement conflates two qualitatively distinct cases: support-aligned corrections, where teacher mass lies within the student's top-K candidates, and support-misaligned corrections, which place relatively more mass beyond its local support. We formalize this distinction as **token teachability** and show that it reveals heterogeneity within high-disagreement regions. The analysis shows that support-conditioned disagreement improves sparse-selection utility, while acting as a refinement rather than a replacement for KL disagreement. Motivated by these findings, we propose **Teachability-Aware OPD** (TA-OPD), which applies OPD loss to high-teachability positions without reward models or verifiers. Across Qwen2.5 and Qwen3 settings, TA-OPD improves over entropy- and divergence-based selectors, often matching or surpassing full-token OPD by selecting 5%-10% high-teachability tokens. These results show that OPD depends on how disagreement aligns with the student's local support, not merely dense supervision or high-disagreement tokens. Code and infrastucture are released. [1]

1 INTRODUCTION

Knowledge distillation has long been used to transfer the behavior of a large teacher model to a smaller student model (Hinton, 2014; Zhou et al., 2026; Fang et al., 2026). For large language models (LLMs), a particularly effective variant is on-policy distillation (OPD), which trains a student policy on its own rollouts using token-level supervision from a teacher policy (Agarwal et al., 2024; Gu et al., 2026; Zhang et al., 2025). Compared with off-policy distillation (Zhou et al., 2025; Wang et al., 2026b), OPD supervises the student on states that it actually visits, reducing distribution mismatch between training and deployment. However, this exposes a central challenge: the teacher provides dense supervision at every token, but not every token-level signal is equally useful.

Recent work shows that OPD supervision is highly non-uniform (Jin et al., 2026; Wang et al., 2026a). Selective OPD therefore selects salient positions for supervision, such as high-entropy tokens or low-entropy tokens with large teacher–student disagreement (Guo et al., 2026; Xie et al., 2026; Xu et al., 2026). However, these criteria measure uncertainty or disagreement, not the learning effect of the induced token-level supervision. In on-policy training, two tokens with similar KL disagreement can lead to very different updates, depending on how the teacher distribution interacts with the student's current predictive state (Fig 1.A). This raises a basic question: *which token-level teacher signals are more useful under the student's current predictive state?*

To answer this question, we need to separate token-level learning value from the rollout in which a token appears. OPD performance entangles token supervision with sampling noise, downstream context shifts, and interactions across positions. We therefore introduce a *fixed-context diagnostic*: we collect student-generated prefixes, freeze them as a context bank, and rescore both the initial and

[1] The code is available at `https://anonymous.4open.science/r/TA-OPD-5707`.

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.

trained students against the same teacher distribution. For each token, we measure the same-context reduction in teacher–student KL and ask which pre-update disagreement signals predict this local improvement. This diagnostic allows us to study whether a token-level teacher signal is aligned with useful learning, rather than merely salient under entropy or raw KL.

Our diagnostic shows that high KL divergence, which means large teacher–student disagreement, is heterogeneous. Within matched high-disagreement regions, student-support coverage separates corrections with different realized gains, exposing structure beyond KL magnitude alone. Selector analysis further shows that this refinement is context dependent and does not universally improve fixed-context ranking over KL-only selection. We call this relative, state-dependent property *token teachability* (as shown in Fig. 1.A). Teachability ranks local OPD utility; it does not label tokens as intrinsically learnable or unlearnable.

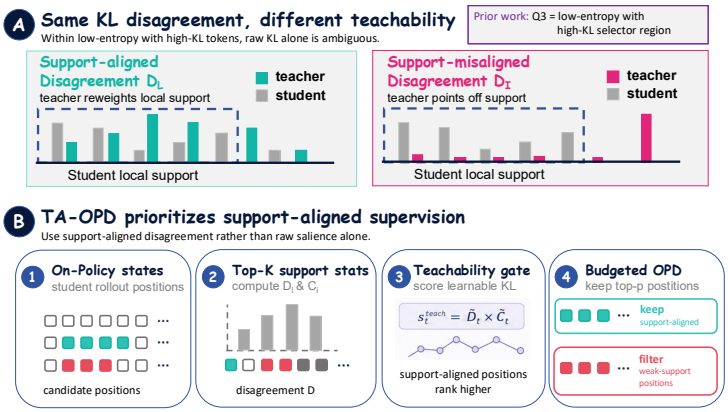


Figure 1: Token teachability. A: Low-entropy, high-KL tokens can contain support-aligned disagreement D^L and support-misaligned disagreement D^I. B: TA-OPD computes disagreement D_t and compatibility C_t, ranks positions by $\tilde{D}_t\tilde{C}_t$, keeps high-teachability signals.

Motivated by these diagnostic, we use support alignment as a soft allocation prior rather than a replacement for disagreement. We instantiate this principle as **Teachability-Aware OPD** (TA-OPD), which ranks response positions by support-conditioned disagreement and applies the OPD loss only to selected positions. TA-OPD uses teacher and student token probabilities already available during OPD and requires no reward model or verifier (Fig. 1.B).

Across Qwen3 and Qwen2.5 teacher–student settings varying in scale and reasoning ability, TA-OPD shows that supervision allocation can outweigh supervision density under constrained token budgets. With as few as 5% selected positions, TA-OPD often matches or surpasses full-token OPD and other selective methods by retaining locally teachable signals rather than dense supervision. These results support our claim: *OPD gains depend on how supervision is allocated, not merely on dense coverage or high disagreement.*

Our contributions are summarized as follows.
(i) We introduce a fixed-context diagnostic that measures same-context KL disagreement and links pre-update token signals to local OPD improvement.
(ii) Using this diagnostic, we formalize *token teachability* as support-conditioned disagreement, revealing that raw KL conflates distinct corrections, and that proxy-optimal selection does not guarantee downstream utility.
(iii) Motivated by this finding, we propose **Teachability-Aware OPD**, a selection method applying OPD loss to high-teachability positions without reward models or verifiers.
(iv) We validate across extensive settings, showing that support-aware sparse supervision can match or improve dense OPD with substantially fewer supervised positions.

2 Related Work

LLM distillation. Knowledge distillation transfers a stronger teacher model's behavior to a smaller student by matching outputs, intermediate representations, or generated trajectories (Hinton, 2014; Zhou et al., 2026; Fang et al., 2026). For LLMs, distillation compresses reasoning, instruction-following, and alignment capabilities under limited compute (Zhou et al., 2025; Wang et al., 2026b). Off-policy distillation trains on teacher- or human-generated trajectories, which may diverge from student-visited states, while on-policy distillation supervises the student on its own rollouts with token-level teacher distributions (Agarwal et al., 2024; Gu et al., 2026; Zhang et al., 2025). We

focus on this OPD setting to examine which teacher signals are truly teachable.

Selective OPD. OPD provides dense token-level supervision, but recent work shows that useful signal is highly non-uniform (Yang et al., 2026; Huang et al., 2026). Entropy-based criteria prioritize uncertain tokens (Jin et al., 2026; Guo et al., 2026), and divergence-based criteria highlight positions where the teacher strongly disagrees with the student (Xie et al., 2026; Wang et al., 2026a). TIP (Xu et al., 2026) formalizes this with a two-axis taxonomy of student entropy and teacher–student divergence for selective OPD. Our work complements this line by conditioning disagreement on the student's local predictive support, revealing support heterogeneity within high-divergence positions.

Teacher–student compatibility. Distillation quality depends on both teacher strength and compatibility with the student policy. Recent OPD analyses (Li et al., 2026) show that shared high-probability vocabulary tokens and compatible reasoning patterns dominate successful OPD. Self-distillation (Zhao et al., 2026; Agarwal et al., 2024; Ding, 2026) and context-conditioned distillation (Ye et al., 2026; Zhang et al., 2026) further emphasize state-dependent supervision. These works restrict supervision within each position to shared high-probability vocabulary tokens. We instead allocate supervision across response using the teacher probability mass covered by the student's local support. The two mechanisms operate at different granularities and are potentially composable.

3 Diagnosing Token Teachability

This section studies which token-level teacher signals have higher local OPD utility. Using TIP's entropy–KL plane as a reference, we focus on Q3, the low-entropy with high-KL region of confident teacher–student disagreement (Xu et al., 2026). We show that Q3 is heterogeneous: raw KL mixes support-aligned corrections with off-support mismatch. We first define a fixed-context gain diagnostic, then decompose disagreement by local support. We also test whether this conditional structure also improves global fixed-context ranking.

3.1 Fixed-Context Gain Diagnostic

OPD trains on student-visited contexts $c_t = (x, y_{<t})$, the prefix generated by the current student. KL at c_t does not fully determine downstream learning value, since end-to-end performance also depends on rollout noise and downstream context changes. We therefore freeze a bank of on-policy prefixes and rescore checkpoints on the same states. For $p_\theta^{i,t} = p_\theta(\cdot \mid c_{i,t})$, define the fixed-context KL reduction

$$G_{i,t}^{\mathrm{fix}} = D_{\mathrm{KL}}\left(p_T^{i,t} \,\|\, p_{\theta_0}^{i,t}\right) - D_{\mathrm{KL}}\left(p_T^{i,t} \,\|\, p_{\theta_\tau}^{i,t}\right). \tag{1}$$

Positive $G_{i,t}^{\mathrm{fix}}$ means the trained student is closer to the teacher on the same prefix. This is a token-indexed diagnostic, not a causal influence estimate for an isolated token update: it measures whether pre-update token features predict later same-context KL reduction after OPD training.

To motivate why high KL alone is insufficient, consider an idealized one-token update.

Proposition 1 *Let $\mathcal{L}_{\mathrm{fix}}$ be differentiable and β-smooth. Let ℓ_t be a token OPD loss. For the update induced by ℓ_t, let G_t be the fixed-context loss reduction. Then*

$$\begin{aligned} G_t &= \eta \,\langle \nabla_\theta \mathcal{L}_{\mathrm{fix}}(\theta), \nabla_\theta \ell_t(\theta) \rangle + R_t, \\ |R_t| &\le \frac{\beta \eta^2}{2} \|\nabla_\theta \ell_t(\theta)\|_2^2. \end{aligned} \tag{2}$$

Thus, even in this local view, useful supervision is not determined by KL magnitude alone; the token loss must align with fixed-context improvement. We use G^{fix} as a diagnostic target for evaluating which teacher signals predict such improvement. The diagnostic and proof are in App. A and I.

3.2 Support-Conditioned Disagreement Decomposition

We now ask what local structure is not captured by KL alone. At context c_t, define the student's *local support* as its top-K token set $S_t^S(K)$; these are the candidates the student currently considers plausible. Let $S_t^T(K)$ be the teacher top-K set and $U_t = S_t^S(K) \cup S_t^T(K)$. Let $\bar{p}^{U_t}$ renormalizes $p(\cdot \mid c_t)$ on U_t, we measure disagreement:

$$D_t = D_{\mathrm{KL}}\left(\bar{p}_T^{U_t} \,\|\, \bar{p}_\theta^{U_t}\right), \tag{3}$$

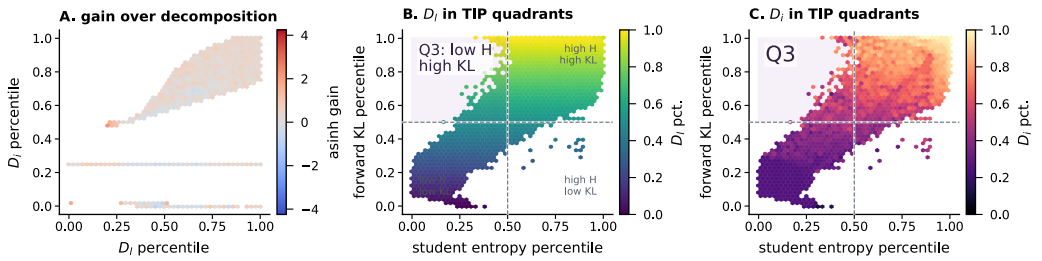


Figure 2: Local-support decomposition. A: fixed-context gain over higher-teachability disagreement D^L and lower-teachability disagreement D^I. B–C: D^L and D^I projected onto TIP's entropy–KL plane; Q3 denotes the low-entropy/high-KL region. Teachability separates corrections with stronger versus weaker local support coverage within Q3.

We measure student-support coverage by the teacher probability mass on the student's local support:

$$C_t = \sum_{v \in S_t^S(K)} p_T(v \mid c_t). \tag{4}$$

High C_t means the teacher mostly reweights tokens already in the student's local support; low C_t means the teacher points off-support. After normalization to $[0, 1]$, we decompose disagreement as

$$D_t^L = \widetilde{D}_t \widetilde{C}_t, \qquad D_t^I = \widetilde{D}_t (1 - \widetilde{C}_t). \tag{5}$$

D_t^L is the higher-teachability component: disagreement with strong local support coverage. D_t^I is the lower-teachability component: disagreement with weaker coverage. These continuous, state-dependent scores do not imply binary learnability; low-D^L positions may still yield positive updates. Fig. 2 links this split to fixed-context gain and shows that Q3 mixes both components. The support construction and normalization are detailed in App. C.

3.3 Support-Dependent Heterogeneity

We next test whether this decomposition predicts $G_{i,t}^{\text{fix}}$. For each K, we fit standardized token-level regressions that replace raw disagreement with D^L and D^I, controlling for student entropy, token position, and teacher entropy. Confidence intervals use prompt-cluster bootstrap.

Fig 3A summarizes the results. Across $K = 8, 16, 32$, D^L has roughly twice the coefficient of D^I (0.086–0.087 vs. 0.043–0.045), with positive bootstrap gaps. This analysis shows that raw KL mixes support-aligned correction with off-support mismatch, but does not establish universal ranking superiority. App. F and B give details and report a positive yet small compatibility interaction.

Following TIP analysis, we use Q3 to denote the low-entropy with high-KL quadrant, where the student is confident yet disagrees with the teacher. This region is a natural target for selective OPD, but our decomposition asks whether its disagreement is uniformly teachable. Fig 3B shows that **it is not**. In a Q3-restricted intervention with matched selected-token ratios, high-D^L tokens yield positive gain, whereas low-D^L tokens are negative and high-D^I tokens are weak. Thus, Q3 is informative but coarse: teachability separates teachable high-KL supervision from off-support mismatch. Exact top-N and support-proxy controls in App. C show the same pattern under matched token counts and alternative support definitions. These indicate that Q3 benefits from a teachability filter rather than a larger token ratio.

3.4 Validation and Boundary Conditions

Finally, we test whether the separation depends on a single prompt shard or teacher. We repeat the fixed-context diagnostic on held-out math prompts, GSM8K-COT prompts (Cobbe et al., 2021; Gao et al., 2024), and stronger 8B/14B teachers. Fig 3C shows positive high-support gaps across the development diagnostics, teachers, and support sizes, with all reported prompt-cluster intervals above zero. App. E and B report the full table and the Full-OPD analysis over held-out Math and GSM8K contexts, respectively. The latter confirms a positive average interaction, and sometimes raw local disagreement remains a stronger ranker.

Selector-neutral boundary. We further evaluate all selectors on context banks after the same Full-OPD treatment to avoid selector-specific coupling. TA-OPD improves sparse-selection utility over

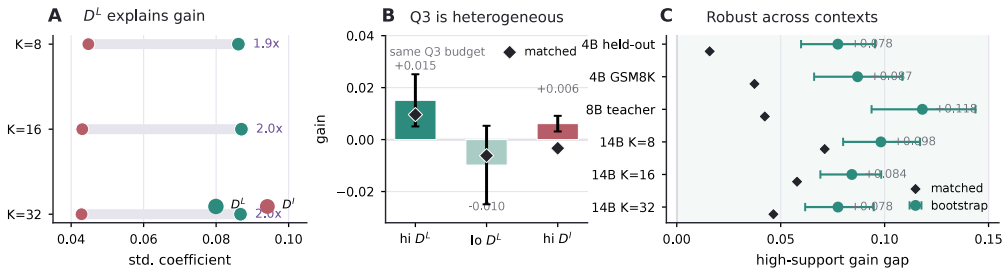


Figure 3: Fixed-context evidence for token teachability. A: D^L has about twice the standardized coefficient of D^I. B: within Q3, high-D^L positions yield larger gains than low-D^L or high-D^I positions. C: high-support gain gaps remain positive across held-out contexts, GSM8K-COT, larger teachers, and support sizes.

entropy, TIP, C-only, and random selection. However, KL-only remains the stronger ranker of absolute fixed-context KL reduction, which is expected because the diagnostic target itself measures forward-KL reduction. At 5%, TA-OPD trails KL-only by 0.161, 0.161, and 0.096 prompt-balanced Gain points on the 4B-to-1.7B Math, 4B-to-1.7B GSM8K, and 14B-to-4B Math banks, respectively. Therefore, support conditioning should be viewed as a complementary allocation prior rather than a universal replacement for disagreement magnitude. Full results are reported in App. B.

4 Teachability-Aware OPD

The diagnostic motivates using support alignment as a soft allocation prior over KL disagreement. We therefore rank positions by teachability, support-conditioned disagreement, while keeping student entropy separate as a baseline and optional auxiliary signal ((shown in Fig 1.B)).

4.1 Selective OPD Objective

For a rollout batch, let $\mathcal{I}$ be all valid response positions after padding and trainer masks. At position t, let $c_t = (x, y_{<t})$ denote the on-policy context, and let $p_\theta(\cdot \mid c_t)$ and $p_T(\cdot \mid c_t)$ be the student and teacher next-token distributions over vocabulary V. Full reverse-KL OPD uses

$$\begin{aligned} \ell_t^{\mathrm{OPD}} &= D_{\mathrm{KL}}(p_\theta(\cdot \mid c_t) \,\|\, p_T(\cdot \mid c_t)) \\ &= \sum_{v \in V} p_\theta(v \mid c_t)\big[\log p_\theta(v \mid c_t) - \log p_T(v \mid c_t)\big]. \end{aligned} \tag{6}$$

Sampled-token OPD does not directly minimize the token log-ratio. Following common OPD trainers, the teacher–old-policy log-prob gap is used as a policy-gradient advantage correction, $A_t^{\mathrm{OPD}} = \beta\big[\log p_T(y_t \mid c_t) - \log p_{\mathrm{old}}(y_t \mid c_t)\big]$, where p_{old} is the rollout policy. Given a binary mask $m_t \in \{0, 1\}$, selective OPD optimizes

$$\mathcal{L}_m(\theta) = \frac{1}{\sum_{t \in \mathcal{I}} m_t} \sum_{t \in \mathcal{I}} m_t \, \ell_t^{\mathrm{train}}(\theta), \tag{7}$$

where ℓ_t^{train} is either the full reverse-KL loss in Eq. 6 or the trainer's sampled-token policy-gradient surrogate using A_t^{OPD}. TA-OPD only changes which response positions receive OPD supervision. Notably, selected-token ratios refer to supervised response positions, not computation efficiency.

4.2 Teachability Score

For each position, define the student and teacher top-K sets

$$\begin{aligned} S_t^S(K) &= \mathrm{TopK}\left(p_\theta(\cdot \mid c_t), K\right), \\ S_t^T(K) &= \mathrm{TopK}\left(p_T(\cdot \mid c_t), K\right), \end{aligned} \tag{8}$$

and let $U_t = S_t^S(K) \cup S_t^T(K)$. We measure local teacher–student disagreement on U_t:

$$\begin{aligned} D_t &= D_{\mathrm{KL}}\left(\bar{p}_T^{U_t} \,\|\, \bar{p}_\theta^{U_t}\right), \\ \bar{p}^{U_t}(v \mid c_t) &= \frac{p(v \mid c_t)}{\sum_{u \in U_t} p(u \mid c_t)}, \; v \in U_t. \end{aligned} \tag{9}$$

Motivated by TIP's analysis (Xu et al., 2026), we use forward KL only as a diagnostic ranking geometry because it is sensitive to teacher-preferred alternatives underweighted by a confident student. Selected positions are still optimized with the standard sampled reverse-KL OPD objective; ranking and optimization therefore serve complementary roles. We define *student-support coverage* as the teacher probability mass on the student's top-K support:

$$C_t = \sum_{v \in S_t^S(K)} p_T(v \mid c_t). \tag{10}$$

High C_t indicates stronger support alignment; low C_t indicates weaker support coverage.

We put all token-level scores on a common scale using batch-wise robust normalization. Let $\mathcal{I}_\mathcal{B}$ be the valid response positions in rollout batch $\mathcal{B}$. For a raw score z_t at position t, such as D_t or C_t, let $z_\mathcal{B} = \{z_j : j \in \mathcal{I}_\mathcal{B}\}$. We define

$$\mathrm{Norm}_\mathcal{B}(z_t) = \mathrm{clip}\left(\frac{z_t - Q_{0.05}(z_\mathcal{B})}{Q_{0.95}(z_\mathcal{B}) - Q_{0.05}(z_\mathcal{B}) + \epsilon}, 0, 1 \right), \tag{11}$$

where Q_q is the q-quantile and $\epsilon > 0$ prevents division by zero. We apply it to disagreement and compatibility:

$$\widetilde{D}_t = \mathrm{Norm}_\mathcal{B}(D_t), \qquad \widetilde{C}_t = \mathrm{Norm}_\mathcal{B}(C_t).$$

The teachability score is

$$s_t^{\mathrm{teach}} = D_t^L = \widetilde{D}_t \, \widetilde{C}_t. \tag{12}$$

The complementary diagnostic term is $D_t^I = \widetilde{D}_t(1 - \widetilde{C}_t)$, which captures disagreement with weaker local support coverage. The teachability score therefore acts as a soft support gate on local forward-KL disagreement. We use it as a supervision-allocation rule and do not assume that it is a calibrated or universally optimal predictor of G^{fix}.

4.3 Teachability-Aware Position Selection

For rollout batch $\mathcal{B}$, let $n_\mathcal{B} = \lceil \rho |\mathcal{I}_\mathcal{B}| \rceil$, where ρ is the supervision ratio. TA-OPD retains the $n_\mathcal{B}$ valid positions with the largest teachability scores:

$$\begin{aligned} \mathcal{T}_{\rho,\mathcal{B}}^{\mathrm{teach}} &= \mathrm{Top}_{n_\mathcal{B}} \left((s_t^{\mathrm{teach}})_{t \in \mathcal{I}_\mathcal{B}} \right), \\ m_t &= \mathbf{1}[t \in \mathcal{T}_{\rho,\mathcal{B}}^{\mathrm{teach}}]. \end{aligned} \tag{13}$$

Here, Top_n returns the indices of the n largest entries. The mask is applied to the selective OPD objective in Eq. 7.

TA-OPD changes only supervision allocation: all response tokens remain in rollout generation and model forward passes, while only selected positions receive nonzero OPD loss. Thus, ρ is a supervision ratio rather than a compute-saving factor. Unless otherwise stated, $K = 16$; alternative support proxies are evaluated in App. C.

To test whether predictive uncertainty provides complementary information, we additionally use

$$\begin{aligned} s_t^{\mathrm{H+teach}} &= \widetilde{H}_t + s_t^{\mathrm{teach}} - \widetilde{H}_t s_t^{\mathrm{teach}}, \\ \widetilde{H}_t &= \mathrm{Norm}_\mathcal{B}(H_t), \qquad H_t = H(p_\theta(\cdot \mid c_t)). \end{aligned} \tag{14}$$

TA-OPD uses s_t^{teach} by default; the entropy mixture is reported only as an auxiliary variant.

5 Experiments

5.1 Setup

Models. We evaluate four pairs: (1) Qwen3-4B to 1.7B; (2) Qwen3-8B-GRPO ((Shao et al., 2024)) to Qwen3-4B; (3) Qwen3-14B to Qwen3-4B; (4) DeepSeek-R1-Distill-Qwen-14B (Guo et al., 2025) to Qwen2.5-3B (Qwen: et al., 2025). The settings vary *scale, capability, and architecture*.
Training data and benchmarks. Training prompts are sampled from DAPO (Yu et al., 2026). We

Table 1: Main benchmark results at a 10% selected-token ratio. We evaluate Qwen3 teacher–student pairs and a cross-backbone DeepSeek-R1-Distill-Qwen-14B to Qwen2.5-3B pair. Entries report mean±standard deviation across two run-level means; each run mean averages five decoding seeds. Avg. is the average score over six benchmarks. TA-OPD$_{+Ent.}$ denotes adding entropy.

Method	Avg.	AIME24	AIME25	GPQA-D.	HumanEval	IFEval	MATH-500
Qwen3-4B → Qwen3-1.7B							
Base	40.77	10.00±3.33	10.00±3.33	31.56±0.76	60.97±1.83	59.52±1.29	72.60±1.00
Pure OPD	42.37	11.67±5.00	**15.00±1.67**	30.55±0.25	59.15±1.22	62.85±0.74	75.00±1.20
Entropy	41.46	13.34±3.34	8.33±5.00	32.58±1.77	59.45±0.30	60.63±0.01	74.40±1.00
TIP	43.05	18.34±1.66	11.67±5.00	29.29±2.02	**62.50±0.30**	60.91±0.27	**75.60±1.00**
TA-OPD	**44.89**	**20.00±0.01**	**15.00±5.00**	**34.59±2.27**	60.98±0.01	**64.05±1.02**	74.70±0.50
TA-OPD$_{+Ent.}$	42.32	15.00±1.67	8.34±1.67	32.58±1.77	61.59±0.61	61.73±0.92	74.70±0.10
Qwen3-8B (GRPO)→ Qwen3-4B							
Base	52.98	21.66±1.66	13.33±0.01	**47.97±0.51**	77.75±0.30	73.57±2.22	83.60±0.20
Pure OPD	53.71	20.00±3.33	21.67±1.67	47.22±1.26	78.36±0.92	70.33±1.20	84.70±1.30
Entropy	56.69	26.33±3.10	23.33±3.30	47.47±1.01	77.65±1.21	**79.85±0.37**	85.50±0.10
TIP	56.81	26.33±3.60	23.67±3.33	47.47±0.49	**80.49±0.61**	76.71±1.11	86.20±0.30
TA-OPD	**56.87**	**30.00±1.40**	21.31±1.32	47.47±0.51	79.88±0.60	75.13±5.30	**87.40±0.80**
TA-OPD$_{+Ent.}$	56.52	**30.00±3.33**	**25.00±1.67**	45.70±2.78	78.05±1.22	73.38±2.22	87.00±0.40
Qwen3-14B → Qwen3-4B							
Base	52.98	21.66±1.66	13.33±0.01	47.97±0.51	77.75±0.30	73.57±2.22	83.60±0.20
Pure OPD	54.64	**23.34±3.34**	21.66±1.66	42.17±0.76	78.05±0.61	**79.11±0.74**	83.50±0.10
Entropy	53.39	21.67±5.00	15.00±1.67	46.46±0.01	79.27±0.61	73.94±4.06	**84.00±0.80**
TIP	53.62	20.00±0.01	16.66±3.33	**47.98±2.53**	78.05±0.61	76.62±1.57	82.40±0.01
TA-OPD	**54.65**	**23.34±3.34**	18.34±1.66	45.20±1.77	**79.57±0.92**	78.93±0.01	82.50±1.70
TA-OPD$_{+Ent.}$	54.10	21.67±5.00	**21.67±1.67**	42.17±0.25	78.35±1.53	77.54±1.02	83.20±0.60
DeepSeek-R1 Distill-Qwen-14B → Qwen2.5-3B							
Base	29.98	3.34±3.33	1.67±1.66	27.81±0.25	64.63±1.22	**26.16±0.27**	**56.31±0.20**
Pure OPD	28.76	5.00±1.67	1.67±1.66	27.78±1.51	62.20±1.21	23.48±0.18	52.50±0.90
Entropy	19.94	1.67±1.66	1.67±1.66	23.98±1.02	54.88±3.66	24.68±0.83	12.83±1.40
TIP	28.11	3.33±0.01	1.67±1.66	27.55±0.01	58.54±0.60	**26.16±0.28**	51.40±0.50
TA-OPD	**30.62**	5.00±1.66	1.67±1.66	29.59±0.51	**66.77±0.31**	25.42±1.20	55.31±0.60
TA-OPD$_{+Ent.}$	30.28	**8.34±1.60**	**3.33±0.01**	**31.63±0.51**	62.81±1.22	23.39±1.02	52.20±0.70

evaluate math, coding, factual QA, and instruction following on AIME24 (Zhang & Math-AI, 2024), AIME25 (Zhang & Math-AI, 2025), GPQA-Diamond (Rein et al., 2023), HumanEval (Chen et al., 2021), IFEval (Zhou et al., 2023), and MATH-500 (Hendrycks et al., 2021), based on the EvalScope (Team, 2024). Each checkpoint is evaluated with five seeds.

Baselines and supervision ratios. We compare TA-OPD with full OPD, entropy selection, and the original TIP selector. Selector ablations additionally include KL-only, which ranks by $\widetilde{D}_t$, C-only, Q3-only, and Q3+high-C. All selective methods share the same OPD trainer, rollout configuration, and number of supervised response positions. Unless otherwise stated, $K = 16$.

Implementation. Each training job uses 8 NVIDIA H800 GPUs; the full sweep uses up to 64 GPUs concurrently. Additional settings, results, and cost appear in App. J, H, and K, respectively.

5.2 Main Results

Table 1 compares all methods at a 10% selected-token ratio. Across the four settings, TA-OPD is strongest for Qwen3-4B to Qwen3-1.7B, and is numerically comparable to the strongest baselines in the two 4B-student settings. This shows that support-aware selection can preserve or improve OPD quality with substantially fewer supervised positions.

For Qwen3-4B to Qwen3-1.7B, TA-OPD improves the average over Full OPD from 42.37 to 44.89 and leads on AIME24, GPQA-Diamond, and IFEval. For different architecture, Qwen2.5 and Qwen3, TA-OPD improves over both Base and Full OPD in average score, while TA-OPD+Entropy

Table 2: Selected-ratio sweep for the Qwen3-4B student settings. Benchmark entries report mean±std; Avg. is the mean of the six benchmark means, excluding the ± terms. TA-OPD denotes Teachability-Aware OPD; TA-OPD$_{+Ent.}$ adds the entropy axis.

Selected ratio	Selector	Avg.	AIME24	AIME25	GPQA-D.	HumanEval	IFEval	MATH-500
Qwen3-8B (GRPO)→ Qwen3-4B								
5%	Entropy	56.95	31.66±1.66	26.66±0.01	47.47±0.15	76.83±1.22	73.47±0.64	85.60±0.20
	TIP	56.20	21.65±8.35	24.98±1.68	**48.23±0.25**	78.66±0.01	**76.80±0.28**	86.90±0.50
	TA-OPD	57.35	31.66±1.67	26.67±0.01	46.47±0.51	78.97±0.92	72.46±0.37	**87.90±1.10**
	TA-OPD$_{+Ent.}$	**57.89**	**33.34±6.67**	**28.33±1.67**	47.22±2.28	**79.58±0.92**	72.37±0.28	86.50±0.30
10%	Entropy	56.69	26.33±3.10	23.33±3.30	**47.47±1.01**	77.65±1.21	**79.85±0.37**	85.50±0.10
	TIP	56.81	26.33±3.60	23.67±3.33	47.47±0.49	**80.49±0.61**	76.71±1.11	86.20±0.30
	TA-OPD	**56.87**	**30.00±1.40**	21.31±1.32	**47.47±0.51**	79.88±0.60	75.13±5.30	**87.40±0.80**
	TA-OPD$_{+Ent.}$	56.52	**30.00±3.33**	**25.00±1.67**	45.70±2.78	78.05±1.22	73.38±2.22	87.00±0.40
30%	Entropy	57.56	28.33±5.00	25.00±5.00	48.49±1.52	79.58±0.31	77.45±3.33	86.50±0.10
	TIP	**58.46**	**31.67±5.00**	28.33±1.66	45.70±0.25	**79.88±0.01**	78.56±1.85	**86.60±1.00**
	TA-OPD	57.28	28.34±1.67	28.33±1.67	**49.19±0.30**	77.75±0.31	74.96±3.24	85.10±0.50
	TA-OPD$_{+Ent.}$	57.46	25.00±1.67	**28.34±1.67**	47.98±3.03	78.97±0.92	**78.75±0.93**	85.70±0.50
50%	Entropy	53.90	20.00±3.33	20.00±3.33	43.68±1.76	78.35±0.91	75.69±4.34	85.70±1.50
	TIP	55.74	23.33±3.33	24.99±1.66	44.69±0.75	78.96±0.91	76.15±2.40	**86.33±0.33**
	TA-OPD	**57.90**	**31.67±4.34**	23.34±0.01	**47.73±0.76**	**78.97±0.92**	**79.67±1.11**	86.00±1.80
	TA-OPD$_{+Ent.}$	57.25	23.33±0.01	**30.00±3.33**	46.97±1.01	78.05±0.61	78.93±0.19	86.20±0.40
Qwen3-14B → Qwen3-4B								
5%	Entropy	53.56	**23.34±3.34**	16.67±0.01	43.94±2.02	77.75±0.92	75.97±5.36	**83.70±0.30**
	TIP	53.83	23.33±0.01	16.67±0.01	44.70±1.77	**79.88±0.61**	75.50±2.49	82.90±0.70
	TA-OPD	53.58	18.34±1.66	**18.34±1.66**	45.70±1.27	77.75±0.30	**79.66±0.93**	81.70±1.30
	TA-OPD$_{+Ent.}$	**54.47**	21.66±1.66	16.67±0.01	**48.23±2.27**	77.75±0.91	79.02±0.65	83.50±0.70
10%	Entropy	53.39	21.67±5.00	15.00±1.67	46.46±0.01	79.27±0.61	73.94±4.06	**84.00±0.80**
	TIP	53.62	20.00±0.01	16.66±3.33	**47.98±2.53**	78.05±0.61	76.62±1.57	82.40±0.01
	TA-OPD	**54.65**	**23.34±3.34**	18.34±1.66	45.20±1.77	**79.57±0.92**	**78.93±0.01**	82.50±1.70
	TA-OPD$_{+Ent.}$	54.10	21.67±5.00	**21.67±1.67**	42.17±0.25	78.35±1.53	77.54±1.02	83.20±0.60
30%	Entropy	53.62	20.00±3.33	**21.66±1.66**	**46.97±0.01**	78.97±0.92	70.52±0.46	**83.60±0.20**
	TIP	**53.93**	**26.67±0.01**	**21.66±5.00**	43.18±1.26	79.57±0.30	69.97±0.09	82.50±0.50
	TA-OPD	52.81	21.66±8.33	15.00±1.67	43.18±0.76	79.57±0.30	**74.77±3.42**	82.70±0.10
	TA-OPD$_{+Ent.}$	53.15	20.00±3.33	18.34±1.66	45.70±2.77	**79.88±0.01**	72.45±1.30	82.50±0.10
50%	Entropy	52.76	18.34±1.66	18.33±1.67	44.95±0.51	78.66±0.01	73.38±2.59	82.90±0.50
	TIP	52.12	**21.66±1.66**	13.33±0.01	**45.95±1.52**	78.05±0.01	70.34±0.27	83.40±1.40
	TA-OPD	52.16	13.33±0.01	**18.34±1.66**	45.20±0.25	77.75±0.31	73.75±3.51	**84.60±0.20**
	TA-OPD$_{+Ent.}$	**52.95**	18.34±1.66	16.66±3.34	44.19±2.78	**78.97±0.91**	**76.43±1.94**	83.10±0.10

gives the best AIME24, AIME25, and GPQA-Diamond scores. These results suggest that the benefit of support-aware selection depends on the teacher-student pair and benchmark.

For stronger Qwen3 teachers with a 4B student, the results have similar trend. With the GRPO-tuned 8B teacher, TA-OPD has the best average and leads on AIME24 and MATH-500, while TIP or entropy remain competitive on HumanEval and IFEval. With the 14B teacher, TA-OPD nearly matches Full OPD in average score and leads HumanEval, while TA-OPD+Entropy is strongest on AIME25. Thus, teachability is not a universal per-benchmark booster; it is a token-allocation rule that improves the quality of supervised positions under a constrained selected-token ratio.

5.3 Sensitivity Analysis

Table 2 sweeps 5%, 10%, 30%, and 50% selected-token ratios for the two Qwen3-4B student settings. The results are not monotonic in selected-token ratio, indicating that more KL-supervised tokens do not always yield better performance.

For Qwen3-8B-GRPO to Qwen3-4B, low selected-token ratios are already competitive: TA-OPD+Entropy reaches the best 5% average score (57.89), while TA-OPD gives the best MATH-500 score at 5% and the best average among teachability-based selectors at 50% (57.90). TIP performs best at 30% average, showing that uncertainty and divergence can still help on some benchmark mix-

Table 3: Downstream selector ablation for Qwen3-14B to 4B at a 10% selected-token ratio. TA-OPD achieves the highest average among the selectors, although KL-only remains a stronger fixed-context KL-reduction ranker (detailed in the Supplement).

Selector	Avg.	AIME24	AIME25	GPQA-D.	HumanEval	IFEval	MATH-500
KL Divergence	53.76	21.66±1.66	20.00±0.01	44.95±1.52	78.97±0.30	73.38±2.22	83.60±0.40
C-only	54.19	21.66±1.66	18.33±8.33	43.94±3.03	77.75±0.91	**80.13±0.83**	83.30±0.30
Q3-only	52.46	21.66±1.66	15.00±1.67	44.70±1.27	**79.57±0.91**	70.52±0.28	83.30±0.10
Q3+high-C	53.32	21.67±5.00	16.66±3.33	44.69±3.29	78.97±0.30	75.05±5.36	82.90±1.10
Entropy	53.39	21.67±5.00	15.00±1.67	46.46±0.01	79.27±0.61	73.94±4.06	**84.00±0.80**
TIP	53.62	20.00±0.01	16.66±3.33	**47.98±2.53**	78.05±0.61	76.62±1.57	82.40±0.01
TA-OPD$_{+Ent.}$	54.10	21.67±5.00	**21.67±1.67**	42.17±0.25	78.35±1.53	77.54±1.02	83.20±0.60
TA-OPD	**54.65**	**23.34±3.34**	18.34±1.66	45.20±1.77	**79.57±0.92**	78.93±0.01	82.50±1.70

tures. However, the strong 5% results indicate that much of the useful OPD signal is concentrated in a small set of teachable positions.

For Qwen3-14B to Qwen3-4B, the best average appears at 10% with TA-OPD (54.65), while TA-OPD+Entropy is strongest at 5% (54.47). Increasing the selected-token ratio to 30% or 50% does not consistently improve the average, and several metrics regress relative to the 5–10% settings. Together with the fixed-context analyses, this non-monotonic trend suggests that supervision allocation matters in addition to token count. App. H provides detailed fixed-context selected-ratio curves, average selected-ratio summaries, and robustness of K.

5.4 Ablation Study

Table 3 provides the selector ablation on Qwen3-14B to Qwen3-4B under the same 10% selected-token ratio. We compare TA-OPD with KL divergence, C-only, Q3-only, and Q3+high-C. Here, C-only is the position-level analogue of overlap-based compatibility; Q3+high-C applies the same signal within TIP's hard Q3 region. TA-OPD combines disagreement magnitude with support conditioning and obtains the best macro-average (54.65), suggesting that combining disagreement magnitude with support conditioning can improve downstream supervision allocation in this setting. Q3-only performs worst; adding high-C improves it but remains below TA-OPD, showing that the TIP quadrant is heterogeneous and a hard Q3 filter is insufficient.

Diagnostic Utility versus Downstream Utility: Table 4 reveals an inversion: KL-only best reduces the immediate proxy, yet TA-OPD achieves higher downstream macro-average (54.65 vs. 53.76). This proxy–objective gap is not a weakness of TA-OPD, but rather a key empirical observation: the best local proxy selector need not yield the best final performance—suggesting selective OPD as an allocation problem where support-aware priors, though weaker on the proxy, may induce more stable training dynamics.

Table 4: **Diagnostic vs. downstream utility.** Bold marks the higher value. Fixed-context results use Gain@5%; results use matched 10% selection.

Setting	KL-only	TA-OPD	$\Delta_{\mathrm{TA-KL}}$
Immediate utility: fixed-context Gain@5%			
4B→1.7B Math	**0.720**	0.558	−0.161
4B→1.7B GSM8K	**0.744**	0.582	−0.161
14B→4B Math	**0.577**	0.481	−0.096
Downstream utility: 14B→4B at 10%			
Reasoning avg.	**42.55**	42.35	−0.21
Code (HumanEval)	78.97	**79.57**	+0.60
Instruction (IFEval)	73.38	**78.93**	+5.55
Macro avg.	53.76	**54.65**	+0.89

6 Conclusion

In this work, we studied why equally large teacher–student disagreement can yield different OPD utility. We propose a fixed-context diagnostic and reveal that raw disagreement is heterogeneous, mixing support-aligned corrections with off-support mismatch. We formalized this relative, state-dependent signal as *token teachability* and introduced TA-OPD, which allocates supervision to high-teachability positions without reward models or verifiers. While the resulting support-conditioned selector does not universally outperform raw KL on the diagnostic proxy itself, it consistently delivers stronger downstream performance, exposing a fundamental gap between local proxy optimization and benchmark performance. Across Qwen3 and Qwen2.5, TA-OPD matches or surpasses dense and selective OPD baselines in several settings with substantially fewer supervised positions.

AI USE STATEMENT

We used generative AI tools to polish the manuscript and figures, help test experiments, and check code and infrastructure. The authors reviewed the AI-assisted work, verified the reported results and claims, and take full responsibility for the submission.

REPRODUCIBILITY STATEMENT

The anonymized code and infrastructure are available at `https://anonymous.4open.science/r/TA-OPD-5707`. Section 5.1 specifies the training settings and benchmarks; Appendices A, J, and B document the diagnostic protocol, training configuration, and validation procedure. The proof of the proposition is in Appendix I.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ken Ding. Hdpo: Hybrid distillation policy optimization via privileged self-distillation. *arXiv preprint arXiv:2603.23871*, 2026.

Luyang Fang, Xiaowei Yu, Jiazhang Cai, Yongkai Chen, Shushan Wu, Zhengliang Liu, Zhenyuan Yang, Haoran Lu, Xilin Gong, Yufang Liu, et al. Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions. *Artificial Intelligence Review*, 59(1):17, 2026.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL `https://zenodo.org/records/12608602`.

Yuxian Gu, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Miniplm: Knowledge distillation for pre-training language models. In *The Thirteenth International Conference on Learning Representations*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yiju Guo, Wenkai Yang, Zexu Sun, Ning Ding, Zhiyuan Liu, and Yankai Lin. Learning to focus: Causal attention distillation via gradient-guided token pruning. *Advances in Neural Information Processing Systems*, 38:24921–24948, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

G Hinton. Distilling the knowledge in a neural network. In *Deep Learning and Representation Learning Workshop in Conjunction with NIPS*, 2014.

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. Selectkd: Selective token-weighted knowledge distillation for llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19287–19297, 2026.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *The 1st Workshop on Scaling Post-training for LLMs*, 2026.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Qwen:, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

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. In *First Conference on Language Modeling*, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

ModelScope Team. EvalScope: Evaluation framework for large models, 2024. URL `https://github.com/modelscope/evalscope`.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *Advances in Neural Information Processing Systems*, 38:115452–115486, 2026a.

Yuanyi Wang, Zhaoyi Yan, Yiming Zhang, Qi Zhou, Yanggan Gu, Fei Wu, and Hongxia Yang. Infigfusion: Graph-on-logits distillation via efficient gromov-wasserstein for model fusion. *Advances in Neural Information Processing Systems*, 38:119677–119713, 2026b.

Xurong Xie, Zhucun Xue, Jiafu Wu, Jian Li, Yabiao Wang, Xiaobin Hu, Yong Liu, and Jiangning Zhang. Llm-oriented token-adaptive knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34070–34078, 2026.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026.

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*, 2026.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

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. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Songming Zhang, Xue Zhang, Tong Zhang, Bojie Hu, Yufeng Chen, and Jinan Xu. Alignditil: Token-level language model alignment as adaptive policy distillation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19791–19807, 2025.

Xinsen Zhang, Zhenkai Ding, Tianjun Pan, Run Yang, Chun Kang, Xue Xiong, and Jingnan Gu. Opsdl: On-policy self-distillation for long-context language models. *arXiv preprint arXiv:2604.17535*, 2026.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2024, 2024.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

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.

Qi Zhou, Yiming Zhang, Yanggan Gu, Yuanyi Wang, Zhijie Sang, Zhaoyi Yan, Zhen Li, Shengyu Zhang, Fei Wu, and Hongxia Yang. Democratizing ai through model fusion: A comprehensive review and future directions. *Nexus*, 2025.

Qi Zhou, Yiming Zhang, Yanggan Gu, Yuanyi Wang, Zhaoyi Yan, Zhen Li, Chi Yung Chung, and Hongxia Yang. Model fusion for scalable and sustainable artificial intelligence: A review and outlook. *Journal of Modern Power Systems and Clean Energy*, 14(1):37–49, 2026.

Table 6: **Fresh selector-neutral fixed-context validation at 5%.** TA and Local-D report prompt-balanced raw forward-KL reduction after Full OPD; remaining columns are paired differences. Intervals use a hierarchical seed–prompt bootstrap with 10,000 resamples. Local-D ranks by normalized local disagreement, and TopK-RKL is the exact student-top-K reverse-KL control.

Setting	Tokens	TA	Local-D	$\Delta_{\mathrm{TA-D}}$ [95% CI]	TA–TIP	TA–Ent.	TA–C	TA–Rand.	TA–TopK-RKL
4B→1.7B Math	57,600	0.558	0.720	**-0.161 [-0.187, -0.134]**	0.089	0.446	0.538	0.495	0.016
4B→1.7B GSM8K	57,596	0.582	0.744	**-0.161 [-0.183, -0.141]**	0.129	0.469	0.574	0.517	0.036
14B→4B Math	57,600	0.481	0.577	**-0.096 [-0.120, -0.073]**	0.087	0.391	0.469	0.431	0.022

A Fixed-Context Diagnostic Protocol

Context bank. For each prompt, we sample student rollouts and store valid response prefixes $c_{i,t} = (x_i, y_{i,<t})$. All checkpoints are scored on this same bank, so diagnostic differences cannot come from resampling different rollouts.

Support metrics. For support size K, $S_t^S(K)$ and $S_t^T(K)$ denote the student and teacher top-K token sets. We compute disagreement on $S_t^S(K) \cup S_t^T(K)$, and compatibility $C_t^{(K)}$ as teacher mass on $S_t^S(K)$. Scalars are robustly normalized within a batch by clipping the 5–95 percentile range to $[0, 1]$.

Fixed-context gain. When full-vocabulary probabilities are available, we compute the fixed-context gain defined in the main paper. Otherwise, top-K support diagnostics are used only for robustness checks. Because token gains are heavy-tailed and correlated within a rollout, all confidence intervals use prompt-cluster bootstrap.

Gradient view. The gradient-alignment proposition gives the local optimization view used by the fixed-context diagnostic. The proof is included below.

B Selector-Neutral Confirmatory Validation

The development analyses below use selector-specific interventions and are therefore informative about heterogeneity, but they can couple the scoring rule to the trained checkpoint. We add a selector-neutral audit in which every score is evaluated after the same Full-OPD treatment. For each of three settings, we freeze 300 fresh prompts, train Full-OPD and rescore 57,596–57,600 positions. Scores use batch-wise 5–95 percentile normalization and exact within-batch top-N selection at $p \in \{1, 3, 5, 10, 20, 50\}\%$. Our primary statistic is prompt-balanced raw forward-KL reduction; uncertainty uses 10,000 hierarchical seed–prompt bootstrap resamples.

Table 5: **Realized-gain audit over 15 historical checkpoints.** Differences compare TA-trained checkpoints with selector-matched checkpoints on a fresh 300-prompt bank (three seeds per selector). Confidence intervals are shown for TA minus Local-D; both cross zero.

Eval. positions	Comparator	Δ gain	95% CI
TA-OPD	Local-D	+0.019	$[-0.011, +0.042]$
	TIP	+0.063	–
	Entropy	+0.117	–
	Random	+0.130	–
Local-D	Local-D	+0.007	$[-0.018, +0.027]$
	TIP	+0.063	–
	Entropy	+0.139	–
	Random	+0.162	–

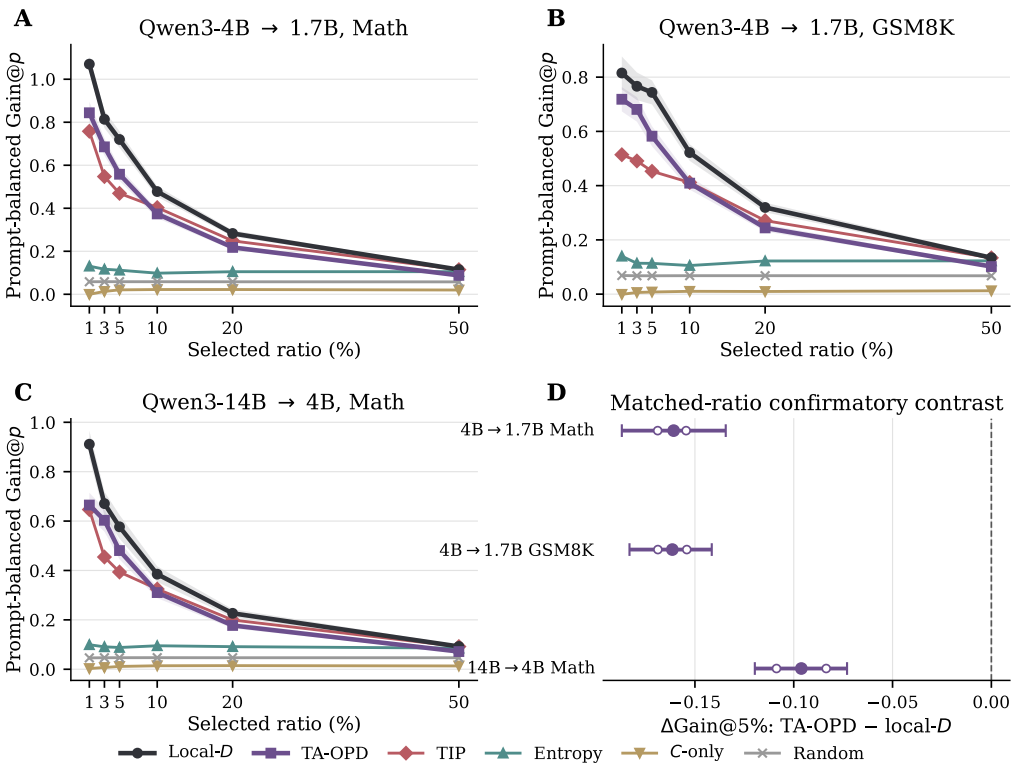


Figure 4: **Selector-neutral validation.** A–C: prompt-balanced Gain@p after the same Full-OPD treatment. D: paired TA-OPD minus KL-only Gain@5% with 95% hierarchical seed–prompt intervals; open circles are seed-specific estimates. TA-OPD consistently improves over TIP, entropy, C-only, and random selection, but KL-only ranks fixed-context gain more strongly.

Matched-ratio selection. Fig. 4 and Table 6 expose a consistent boundary. At 5%, TA-OPD exceeds TIP by 0.087–0.129 and also exceeds entropy, C-only, random selection, and the tested exact student-top-K reverse-KL control. However, KL-only is stronger in all six fresh runs: the hierarchical TA-minus-Local-KL differences are -0.161 on 4B→1.7B Math, -0.161 on GSM8K, and -0.096 on 14B→4B Math, with all intervals below zero. KL-only and full forward KL select nearly identical positions (Jaccard > 0.999 at 5%), so the fresh banks favor disagreement magnitude over its compatibility product. This directly contradicts a universal claim that support conditioning must improve fixed-context ranking.

Prompt-held-out ranking. We split by prompt into five folds, keeping all tokens from one prompt in the same fold. Table 7 reports fold-averaged rank correlation, AUROC for positive gain, NDCG, and positive-gain precision. TA-OPD remains substantially stronger than TIP, entropy, C-only, and random selection in sparse-selection utility, but KL-only has the best Spearman, NDCG@5%, and Precision$_+$@5% in every setting. Accordingly, teachability is useful for sparse ranking relative to salience baselines, not a full token-gain reconstruction model or a universal replacement for raw disagreement.

Transparent interaction test. To separate disagreement from compatibility, we fit

$$G_{i,t}^{\text{fix}} = \beta_D \widetilde{D}_t + \beta_C \widetilde{C}_t + \beta_{DC} \widetilde{D}_t \widetilde{C}_t + \gamma^\top X_t + \epsilon_{i,t}, \tag{15}$$

where X_t contains student entropy, teacher entropy, and normalized position. Training uses four prompt folds and evaluation uses the held-out fold.

The interaction coefficient is positive in all three settings, but its absolute contribution is small: OOS ΔR^2 over $D + C$ ranges from -0.0005 to 0.0018. Moreover, 96.0–97.1% of positions have

Table 7: **Prompt-held-out ranking quality.** Metrics are averaged over two Full-OPD seeds and five prompt-group folds. Precision$_+$@5% is the fraction of selected positions with positive fixed-context gain. Gray rows mark the two central disagreement selectors; bold denotes the best value within each setting.

Setting	Selector	Spearman ρ	Kendall τ	AUROC$_+$	NDCG@5%	Precision$_+$@5%
4B→1.7B Math	Local-D	**0.421**	**0.332**	**0.536**	**0.622**	**0.881**
	TA-OPD	0.292	0.241	0.510	0.542	0.862
	TIP	0.312	0.234	0.459	0.378	0.768
	Entropy	0.207	0.136	0.412	0.137	0.642
	C-only	-0.284	-0.212	0.472	0.020	0.478
	Random	0.000	0.000	0.500	0.057	0.664
4B→1.7B GSM8K	Local-D	**0.614**	**0.543**	**0.522**	**0.680**	**0.924**
	TA-OPD	0.450	0.425	0.501	0.575	0.896
	TIP	0.490	0.436	0.419	0.399	0.779
	Entropy	0.349	0.274	0.343	0.154	0.646
	C-only	-0.456	-0.345	0.490	0.009	0.678
	Random	0.000	0.000	0.500	0.067	0.793
14B→4B Math	Local-D	**0.491**	**0.413**	0.493	**0.635**	**0.907**
	TA-OPD	0.350	0.314	0.473	0.563	0.880
	TIP	0.376	0.311	0.409	0.391	0.791
	Entropy	0.266	0.199	0.355	0.140	0.665
	C-only	-0.335	-0.253	**0.507**	0.014	0.647
	Random	-0.000	-0.000	0.500	0.057	0.731

Table 8: **Compatibility interaction and support saturation.** OOS ΔR^2 compares $D + C + D{\times}C$ with $D + C$ under five-fold prompt splits. High-D ΔG_C contrasts high- versus low-compatibility cells in the highest disagreement quintile using winsorized gain. Jaccard compares TA-OPD with Local-D selections. The shaded GSM8K row marks the cross-domain surface reversal.

Setting	$\beta_{D\times C}$ [95% CI]	OOS ΔR^2 (s1/s2)	High-D ΔG_C (s1/s2)	$C \geq .999$ (%)	Jaccard@5%
4B→1.7B Math	0.0021 $[0.0011, 0.0030]$	0.0018/0.0012	+0.014/+0.016	97.1	0.464
4B→1.7B GSM8K	0.0018 $[0.0001, 0.0030]$	0.0013/-0.0005	-0.065/-0.062	96.0	0.477
14B→4B Math	0.0013 $[0.0008, 0.0018]$	0.0017/0.0009	+0.002/+0.011	96.2	0.463

$C \geq 0.999$, leaving limited dynamic range for compatibility. The highest-disagreement cells show positive high-C contrasts on Math but reverse on GSM8K. Thus, compatibility can moderate disagreement, yet the effect is saturated and context dependent rather than a separate high-capacity predictor.

Realized-gain replication. We additionally evaluate 15 historical 5%-ratio checkpoints—TA-OPD, KL-only, TIP, entropy, and random, each is evaluated with three seeds—on a common fresh 300-prompt bank. As Table 5 shows, TA-trained checkpoints improve over TIP-, entropy-, and random-trained checkpoints. Their advantage over KL-only-trained checkpoints is positive but uncertain under both TA-selected positions ($+0.019$, 95% CI $[-0.011, 0.042]$) and KL-only-selected positions ($+0.007$, $[-0.018, 0.027]$).

Interpretation. The fresh audit refines the development evidence rather than discarding it. The earlier coefficient, Q3, and downstream ablations establish meaningful heterogeneity and motivate support-aware selection. The confirmatory results show that this heterogeneity does not imply that multiplying by compatibility always improves token ranking: raw local disagreement can be the stronger fixed-context selector when teacher mass already saturates the student's top-K support.

Table 9: Section 3 diagnostic datasets. The short diagnostic run is retained only as a boundary check.

Dataset	Contexts	Tokens	Use
4B → 1.7B held-out 300	300	57,600	main
4B → 1.7B GSM8K-COT 300	300	57,600	main
8B → 1.7B 300	300	57,600	main
14B → 1.7B 300 K=8	300	57,600	main
14B → 1.7B 300	300	57,600	main
14B → 1.7B 300 K=32	300	57,600	main
8B → 1.7B short diagnostic	8	1,024	boundary only

Table 10: Support proxy audit inside the low-entropy with high-divergence (Q3) region. Different support definitions produce positive high–low gain gaps.

Setting	Proxy	Tokens	Gap	95% interval
14B → 1.7B, K=16	C_t mass	3342	+0.086	[+0.064, +0.104]
14B → 1.7B, K=16	CBC	3342	+0.105	[+0.078, +0.125]
14B → 1.7B, K=16	Jaccard	3342	+0.038	[+0.020, +0.055]
14B → 1.7B, K=16	overlap	3342	+0.038	[+0.020, +0.055]
14B → 1.7B, K=32	C_t mass	3342	+0.090	[+0.074, +0.106]
14B → 1.7B, K=32	CBC	3342	+0.105	[+0.078, +0.125]
14B → 1.7B, K=32	Jaccard	3342	+0.052	[+0.032, +0.069]
14B → 1.7B, K=32	overlap	3342	+0.052	[+0.032, +0.069]

Relation to downstream utility. The matched Qwen3-14B-to-Qwen3-4B downstream ablation yields a different ordering at a 10% selected-token ratio: TA-OPD reaches a six-benchmark macro-average of 54.65, compared with 53.76 for Local KL. Per-benchmark directions are mixed, and we do not claim a paired confidence interval for this macro difference. Together with the selector-neutral audit, this establishes an objective mismatch—the best ranker of immediate fixed-context KL reduction need not maximize downstream utility—rather than universal dominance by either selector.

C Local-Support Decomposition

Top-K supports. For each fixed context c_t, we store the student and teacher top-K sets $S_t^S(K)$ and $S_t^T(K)$. The union $U_t = S_t^S(K) \cup S_t^T(K)$ is used for local disagreement. For $v \in U_t$, the restricted distribution is

$$\bar{p}^{U_t}(v \mid c_t) = \frac{p(v \mid c_t)}{\sum_{u \in U_t} p(u \mid c_t)}.$$

This keeps the diagnostic focused on locally salient alternatives rather than full-vocabulary tail noise.

Normalization. All scalar token scores are normalized within a rollout batch:

$$\widetilde{z}_t = \mathrm{clip}\left(\frac{z_t - Q_{0.05}(z_{\mathcal{B}})}{Q_{0.95}(z_{\mathcal{B}}) - Q_{0.05}(z_{\mathcal{B}}) + \epsilon}, 0, 1 \right),$$

where Q_q is the q-quantile. We use the same normalization for entropy, disagreement, and compatibility-derived scores.

Support proxies. The main compatibility mass is $C_t = \sum_{v \in S_t^S(K)} p_T(v \mid c_t)$. We also audit simpler proxies: top-K overlap fraction, top-K Jaccard, shared teacher top-K mass, and a contrastive binary compatibility score (CBC). These proxies ask the same question: how much of the teacher correction remains near the student's local support?

D Low-Entropy with High-Divergence Heterogeneity Diagnostics

Live quadrant intervention. We train three masks restricted to the low-entropy with high-divergence (Q3) region: high D^L, low D^L, and high D^I. All keep roughly the same token ratio. The high-D^L mask is the only clearly positive intervention; low-D^L is negative, and high-D^I is weak.

Table 11: Low-entropy with high-divergence live intervention. Even inside this TIP region, high teachability is beneficial while low teachability is harmful.

Q3-region mask	Keep %	Boot. gain	Match. gain	D^L	D^I	C
Q3+D^L high	3.26	0.015 ± 0.010	0.010	0.072	0.005	0.972
Q3+D^L low	3.24	-0.010 ± 0.015	-0.006	0.015	0.011	0.933
Q3+D^I high	3.23	0.006 ± 0.003	-0.003	0.053	0.012	0.919

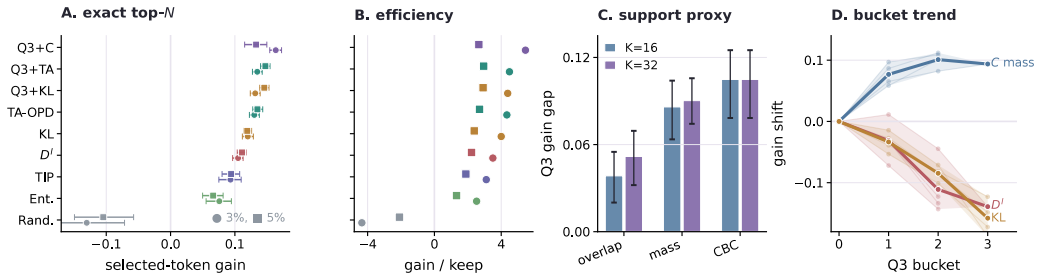


Figure 5: **Q3 controls.** A–B: exact top-N comparisons under matched token counts; B reports gain per kept token. C: support proxies yield positive high–low gain gaps inside Q3 at $K = 16, 32$. D: bucket trends (thin: individual diagnostics; bold: mean) show support mass tracks gain, whereas raw KL and D^I decline.

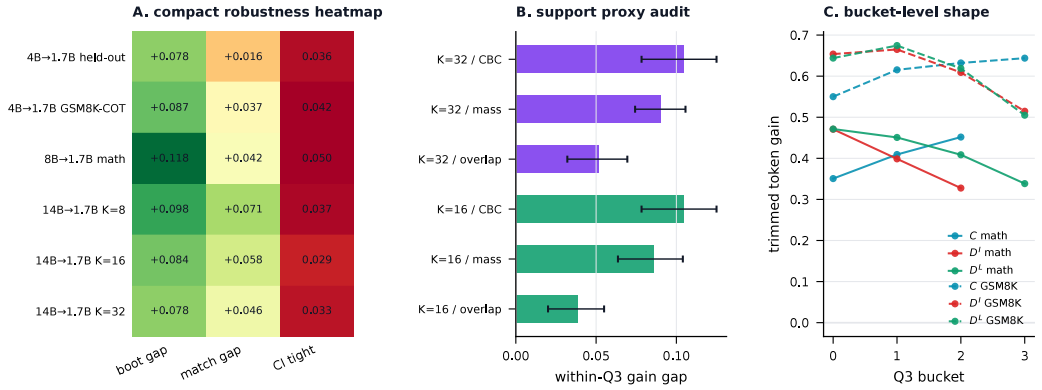


Figure 6: Additional low-entropy with high-divergence and robustness evidence. The panels visualize the support-proxy audit, compact robustness statistics, and bucket-level gain shapes used in the main diagnostic analysis.

Exact-budget fixed-context controls. Fig 5 summarizes the main Q3 control evidence. Here we report the full exact top-N table behind Panels A–B. Every selector keeps the same number of target tokens; support-aligned selectors remain positive, while random selection is consistently negative. Appendix Fig 6C shows the bucket-level gain shapes used as a nonlinear sanity check.

Support-definition audit. The quadrant result is stable across support proxies. Teacher mass on student support, top-K overlap, Jaccard overlap, and CBC all produce positive high–low support gaps inside the low-entropy with high-divergence region. Appendix Fig 6A gives a compact heatmap of the robustness gaps in Table 18. Appendix Fig 6B visualizes the support-proxy audit.

E Context and Teacher-Scale Robustness

We repeat the development-bank diagnostic on held-out math prompts, GSM8K-COT prompts, and stronger 8B/14B teachers. The table shows positive high-support gaps across these development

Table 12: MATH-hard category deltas relative to the base model. Values are percentage-point changes. This table is best treated as boundary analysis rather than the central claim.

Model	Alg.	C&P	Geo.	IntAlg.	Num.	PreAlg.	PreCalc.
D^L	+1.6	0.0	+3.0	-1.4	-3.9	-0.5	+1.5
TIP	+2.9	+0.8	+2.3	+3.2	-7.1	+3.1	+0.7
D^I	+1.0	+4.9	+0.8	+1.8	-3.9	+2.6	0.0
Random	+2.0	-0.8	-0.8	+0.4	-3.9	+2.1	0.0
Entropy+D^L	0.0	+0.8	+3.8	+1.4	-5.2	0.0	+0.7
Entropy	0.0	0.0	+2.3	-2.1	-2.0	+0.5	-0.7

Table 13: Exact top-N matched fixed-context analysis at $K = 16$. Useful as a confound check: all selectors retain exactly the same number of tokens.

Selector	Ratio	Gain	95% interval	Gain/keep	Q3 frac.	H	D	C
D^L	3	0.130	[0.123, 0.138]	4.34	7.4	0.31	1.00	0.99
D^L	5	0.135	[0.128, 0.143]	2.70	5.5	0.37	0.94	0.97
KL	3	0.120	[0.112, 0.129]	4.00	6.1	0.40	1.00	0.66
KL	5	0.119	[0.113, 0.126]	2.38	5.4	0.41	1.00	0.65
TIP-soft	3	0.093	[0.075, 0.110]	3.11	2.6	0.70	0.69	0.59
TIP-soft	5	0.094	[0.080, 0.107]	1.88	2.6	0.69	0.69	0.61
Entropy	3	0.076	[0.055, 0.095]	2.53	0.0	1.00	0.42	0.54
Entropy	5	0.066	[0.050, 0.082]	1.32	0.0	1.00	0.42	0.56
Random	3	-0.130	[-0.200, -0.072]	-4.35	4.8	0.18	0.13	0.93
Random	5	-0.105	[-0.149, -0.057]	-2.09	4.8	0.19	0.13	0.93
D^I	3	0.105	[0.096, 0.113]	3.50	1.8	0.64	0.87	0.03
D^I	5	0.111	[0.104, 0.117]	2.22	1.6	0.66	0.72	0.11

contexts and support sizes. The selector-neutral audit above is stricter: it preserves a positive average interaction but reveals that ranking superiority over KL-only does not transfer universally.

F Development-Bank Regression Diagnostics

Specification. For each token, the dependent variable is the fixed-context gain $G^{\text{fix}}_{i,t}$ defined in the main paper. All scalar predictors are standardized within the diagnostic dataset. The baseline includes student entropy, local disagreement, their interaction, normalized token position, and teacher entropy. We then compare two extensions: adding compatibility C_t alone, or replacing raw disagreement with the decomposed pair D^L_t, D^I_t. Confidence intervals use prompt-cluster bootstrap.

$$G^{\text{fix}}_{i,t} = \beta_L z(D^L_t) + \beta_I z(D^I_t) + \gamma^\top z(X_t) + \epsilon_{i,t}, \tag{16}$$

where X_t includes student entropy, token position, and teacher entropy. Raw disagreement D_t is not included together with D^L_t and D^I_t.

Why coefficients, not only R^2. Token gains are heavy-tailed and noisy because a single local update can affect later prefixes and formatting tokens. We therefore use standardized coefficients and prompt-cluster gaps as the main diagnostic; R^2 is reported only as an incremental sanity check.

Nonlinear sanity check. We also bin residualized gains by D^L, D^I, and C_t using the same fixed-context bank. The binned curves do not indicate that the learnable–incompatible separation is an artifact of a purely linear model.

G Development-Bank Selector Interventions

Before downstream evaluation, we test whether the diagnostic score can serve as a training mask on the development bank. The table below compares selective OPD selectors on the same fixed-context

Table 14: Prompt-cluster bootstrap regression decomposition. Support-aligned disagreement has a consistently larger standardized coefficient than support-misaligned disagreement. $\Delta R^2_{\text{decomp}}$ is reported in 10^{-3} units relative to the entropy+divergence baseline.

K	Q3 tokens	β_L	β_I	Gap	95% interval	$\Delta R^2_{\text{decomp}}$
8	2393	0.086	0.045	0.041	[0.039, 0.044]	0.897
16	2393	0.087	0.043	0.044	[0.041, 0.047]	0.790
32	2393	0.087	0.043	0.044	[0.041, 0.047]	0.743

Table 15: **Fixed-context selector check.** $G_{3\%}$ and $G_{5\%}$ are same-context KL reductions at nominal selected-token ratios; G/K normalizes by actual retained ratio (Keep). TA-OPD gives the largest reduction per retained token. C is teacher mass on student support, and Q3 is TIP's low-entropy/high-divergence region.

Selector	Keep	$G_{3\%}$	$G/K_{3\%}$	$G_{5\%}$	$G/K_{5\%}$	Q3 frac.
TA-OPD	3.4	**0.047**	**1.40**	**0.049**	**0.95**	4.4
Raw KL	3.4	0.033	0.97	0.036	0.69	2.7
Incompat. D^I	3.5	0.025	0.72	–	–	1.4
TIP	3.3	0.023	0.69	0.026	0.50	1.6
Random	3.3	0.021	0.65	0.021	0.41	3.3
TA-OPD$_{+Ent.}$	3.3	0.021	0.63	0.016	0.31	1.0
Entropy	3.3	0.018	0.53	0.009	0.18	0.3
Q3+high-C	3.2	0.016	0.51	0.009	0.23	94.6
Q3+low-C	3.2	0.010	0.32	–	–	94.7
Q3-only	3.2	0.010	0.31	0.002	0.05	91.7

bank. TA-OPD gives the largest KL reduction and gain per kept token at both 3% and 5% selected-token ratios in this intervention. Because the fresh selector-neutral audit favors KL-only, we treat this table as development evidence rather than an out-of-sample ranking claim.

H Selected-Ratio and Downstream Boundary Checks

Selected-ratio sweep. We sweep the selected-token ratio for TA-OPD and low-entropy with high-divergence compatibility controls. The fixed-context gain peaks around 3–5% and saturates at 10%, suggesting that selected supervision is quality-limited rather than simply ratio-limited.

Average selected ratio. Table 17 compresses the Qwen3-4B student selected-ratio sweep from the main paper into benchmark macro-averages. The trend is not monotonic: TA-OPD or TA-OPD$_{+Ent.}$ is strongest at several low selected-ratio points, while larger selected-token ratios can help or hurt depending on the teacher and benchmark mix.

Support-size sensitivity. Table 19 varies the local-support size from $K = 4$ to 64 for Qwen3-14B to Qwen3-4B at a 10% selected-token ratio. The macro-average is stable around the default: $K = 8$ scores 54.87 and $K = 16$ scores 54.65, while per-benchmark optima differ. This supports using $K = 16$ as a fixed default rather than a benchmark-tuned choice.

Early downstream checks. Before the full benchmark suite, we ran small downstream smoke checks on GSM8K-COT, MATH-hard, and capped AIME. GSM8K-COT followed the fixed-context trend, while MATH-hard and capped AIME were low-resolution boundary checks; we therefore report the complete downstream evidence only in the main benchmark tables and omit the sparse smoke table.

Table 16: Fixed-context selected-ratio sweep for TA-OPD. The teachability selector peaks around 3–5% selected-token ratio and saturates at 10%.

Selector	Nominal %	Actual %	Bootstrap gain	Matching gain	Seeds
TA-OPD	1	1.89	0.034 ± 0.016	0.017 ± 0.015	3
TA-OPD	3	3.35	0.047 ± 0.006	0.030 ± 0.003	3
TA-OPD	5	5.14	0.049 ± 0.002	0.027 ± 0.003	3
TA-OPD	10	10.16	0.035 ± 0.013	0.022 ± 0.007	3
Q3+high-C	1	1.93	0.009 ± 0.004	0.002 ± 0.002	2
Q3+high-C	3	3.20	0.016 ± 0.001	0.007 ± 0.006	3
Q3+high-C	5	3.93	0.009 ± 0.015	0.002 ± 0.009	3
Q3+low-C	3	3.23	0.010 ± 0.003	0.006 ± 0.006	3
D^I	3	3.52	0.025 ± 0.000	0.014 ± 0.005	3

Table 17: Macro-average selected-ratio sweep on the two Qwen3-4B student settings. Scores average the six downstream benchmark means from the main-paper selected-ratio sweep; the full per-benchmark results are kept in the main table. TA-OPD denotes Teachability-Aware OPD, and TA-OPD$_{+Ent.}$ adds the entropy axis.

Setting	Selected ratio	Entropy	TIP	TA-OPD	TA-OPD$_{+Ent.}$
Qwen3-8B (GRPO)→4B	5%	56.95	56.20	57.35	**57.89**
	10%	56.69	56.81	56.87	56.52
	30%	57.56	**58.46**	57.28	57.46
	50%	53.90	55.74	**57.90**	57.25
Qwen3-14B →4B	5%	53.56	53.83	53.58	**54.47**
	10%	53.39	53.62	**54.65**	54.10
	30%	53.62	**53.93**	52.81	53.15
	50%	52.76	52.12	52.16	**52.95**

I Proof of the Proposition

Let $g_t = \nabla_\theta \ell_t(\theta)$ and $\Delta_t = -\eta g_t$. By β-smoothness of $\mathcal{L}_{\text{fix}}$,

$$\mathcal{L}_{\text{fix}}(\theta + \Delta_t) = \mathcal{L}_{\text{fix}}(\theta) + \langle \nabla_\theta \mathcal{L}_{\text{fix}}(\theta), \Delta_t \rangle + r_t,$$

with

$$|r_t| \le \frac{\beta}{2} \|\Delta_t\|_2^2.$$

Substituting $\Delta_t = -\eta g_t$ gives

$$\mathcal{L}_{\text{fix}}(\theta_t^+) = \mathcal{L}_{\text{fix}}(\theta) - \eta \, \langle \nabla_\theta \mathcal{L}_{\text{fix}}(\theta), g_t \rangle + r_t.$$

Rearranging,

$$G_t = \eta \, \langle \nabla_\theta \mathcal{L}_{\text{fix}}(\theta), g_t \rangle - r_t.$$

Setting $R_t = -r_t$ yields the result.

J Training Configuration

Table 20 summarizes the training recipe used by TA-OPD and all selected-ratio-matched baselines. In this work, all methods share the same training seed sets. All selectors share the same rollout, optimization, and hardware settings; only the token-position scoring rule and retained-token ratio change. Here, a selected-token ratio means the fraction of response positions with nonzero OPD loss, while the full sequence is still used for rollout, context modeling, and log-probability computation.

Main baselines. Entropy selection ranks positions by normalized student entropy. For TIP, we reproduce the original student-to-teacher reverse-KL divergence

$$\delta_t^{\text{TIP}} = D_{\text{KL}} \left(p_\theta(\cdot \mid c_t) \,\|\, p_T(\cdot \mid c_t) \right)$$

Table 18: **Context and teacher-scale robustness.** The gap compares high- vs. low-support tokens inside Q3, the low-entropy/high-KL diagnostic region. Bootstrap intervals are prompt-clustered; matched gap denotes the prompt-matched estimate. All main 300-context diagnostics have positive intervals.

Setting	Context / support	Contexts	Tokens	Bootstrap gap	95% interval	Matched gap
Qwen3-4B→1.7B	held-out math, $K = 16$	300	57,600	+0.078	[+0.060, +0.095]	+0.016
Qwen3-4B→1.7B	GSM8K-COT, $K = 16$	300	57,600	+0.087	[+0.066, +0.109]	+0.037
Qwen3-8B→1.7B	held-out math, $K = 16$	300	57,600	+0.118	[+0.094, +0.144]	+0.042
Qwen3-14B→1.7B	held-out math, $K = 8$	300	57,600	+0.098	[+0.080, +0.117]	+0.071
Qwen3-14B→1.7B	held-out math, $K = 16$	300	57,600	+0.084	[+0.069, +0.098]	+0.058
Qwen3-14B→1.7B	held-out math, $K = 32$	300	57,600	+0.078	[+0.062, +0.095]	+0.046

Table 19: **Support-size sensitivity for Qwen3-14B to Qwen3-4B at a 10% selected-token ratio.** Entries are mean±variation from the evaluation runs; Avg. averages the six point estimates. $K = 8$ has the best macro-average (54.87), while the default $K = 16$ is within 0.22 points (54.65). Bold marks column maxima.

K	Avg.	AIME24	AIME25	GPQA-D.	HumanEval	IFEval	MATH-500
4	53.30	**25.00±8.33**	13.34±3.34	**45.46±0.51**	78.36±0.31	74.96±5.27	82.70±0.70
8	**54.87**	23.34±6.67	**21.67±1.67**	44.95±1.52	79.27±1.22	77.36±2.68	82.60±0.80
16	54.65	23.34±3.34	18.34±1.66	45.20±1.77	79.57±0.92	**78.93±0.01**	82.50±1.70
32	53.56	23.33±0.01	16.67±3.34	45.20±1.77	78.66±1.22	74.12±2.96	**83.40±1.20**
64	52.99	20.00±0.01	16.67±3.34	43.44±2.53	**79.88±1.22**	75.05±0.56	82.90±1.10

and its Soft-OR score

$$s_t^{\mathrm{TIP}} = \widehat{H}_t^{\mathrm{TIP}} + \widehat{\delta}_t^{\mathrm{TIP}} - \widehat{H}_t^{\mathrm{TIP}}\widehat{\delta}_t^{\mathrm{TIP}},$$

using TIP's original normalization. Full OPD supervises every valid response position. All selective methods retain the same batch-level token budget and share the same OPD trainer and seed sets.

Component ablations. KL divergence and C-only rank positions by $\widetilde{D}_t$ and $\widetilde{C}_t$, respectively. Q3-only and Q3+high-C test whether compatibility remains useful inside TIP's low-entropy, high-divergence region. Random selection samples the same number of valid positions uniformly.

K Computation Cost

Table 21 reports the measured computational cost for the Qwen3-14B to Qwen3-4B setting. We separate teacher serving from student-side computation and covers all selective OPD variants and selected-token ratios used in the downstream distillation. TA-OPD is not designed as sequence-level compute pruning: all methods still generate full rollouts, query the teacher on the full sequence, and train the student on full contexts with a token-level loss mask. Accordingly, selected-token ratios should be interpreted as supervision ratios, not proportional runtime savings. At the 10% selected-token ratio, TA-OPD uses 6.49 GPU-hours, compared with 5.13 for full OPD, 7.31 for entropy selection, 6.02 for TIP, and 8.33 for TA-OPD$_{+Ent.}$. Across ratios, the teacher accounts for roughly 25–32% of total GPU-hours; the remaining cost is dominated by student actor training and rollout. We did not log separate tokens/s, peak-memory, or score-kernel timers in this audit, so we report end-to-end GPU-hours rather than isolated kernel-level overhead.

Table 20: **Training hyperparameters.** All selector variants use the same sampled-token reverse-KL OPD recipe. Selected-token ratios indicate the fraction of response positions assigned nonzero OPD loss, not sequence-level compute pruning.

Group	Hyperparameter	Value
Models and data		
Teacher–student pairs	Backbones	Qwen3-4B→1.7B; Qwen3-8B(GRPO)→4B; Qwen3-14B→4B; DeepSeek-R1-Distill-Qwen-14B→Qwen2.5-3B
Training data	Prompts	DAPO math prompts
Runtime		
Compute	GPUs	8 H800 per job; 64 H800 for the full sweep
Parallelism	Teacher / student TP	2 / 2
GPU allocation	Teacher / actor / rollout	0–1 / 2–5 / 6–7
Teacher serving	Backend; memory fraction	SGLang; 0.30–0.55
Rollout and sampling		
Rollouts	Number	256
Batching	Prompts × samples	2×4
Batch size	Global / micro	8 / 1
Generation	Max response length; temperature	16k; 1.0
Optimization		
Objective	Loss	Sampled-token reverse-KL OPD
Optimizer	Adam (β_1, β_2); weight decay	$(0.9, 0.98)$; 0.1
Learning rate	Schedule	1×10^{-6}; constant
Regularization	Attention / hidden dropout	0.0 / 0.0
Checkpointing	Interval	Every 50 rollouts
Selectors and diagnostics		
Support size	K	8,16,32
Main selectors	Methods	Full OPD; Entropy; TIP; TA-OPD; TA-OPD+Entropy
Ablations	Methods	Raw KL; C-only; Q3-only; Q3+high-C; Random
Selected-token ratio	Retained ratios	5%, 10%, 30%, 50%; Full OPD: 100%
Fixed-context diagnostic	Contexts / positions / K	300 / 57,600 / $\{8,16,32\}$
Evaluation	Seeds	5
Training	Seeds	2

Table 21: **End-to-end compute audit for Qwen3-14B to Qwen3-4B.** All runs use the same 8-H800 layout: teacher serving on 2 GPUs, student actor training on 4 GPUs, and student rollout on 2 GPUs. Student GPU-hours combine actor training and rollout. Wall time is the maximum component duration. The selector score computation is included in the end-to-end student-side time rather than timed as a separate micro-kernel.

Selector	Selected ratio (%)	Teacher GPU-h	Student GPU-h	Total GPU-h	Teacher share (%)	Wall min	Rel. to Full
Full OPD	–	1.42	3.71	5.13	27.7	43	1.00×
Entropy	5	2.33	6.53	8.86	26.3	70	1.73×
TIP	5	2.09	5.91	8.00	26.1	63	1.56×
TA-OPD	5	2.40	6.41	8.81	27.2	72	1.72×
TA-OPD$_{+Ent.}$	5	2.58	7.38	9.96	25.9	77	1.94×
Entropy	10	1.95	5.37	7.31	26.7	58	1.42×
TIP	10	1.61	4.41	6.02	26.7	48	1.17×
TA-OPD	10	1.68	4.81	6.49	25.9	50	1.27×
TA-OPD$_{+Ent.}$	10	2.20	6.13	8.33	26.4	66	1.62×
Entropy	30	2.08	5.77	7.86	26.5	62	1.53×
TIP	30	1.45	4.05	5.51	26.3	44	1.07×
TA-OPD	30	2.65	7.54	10.18	26.0	80	1.98×
TA-OPD$_{+Ent.}$	30	1.63	4.52	6.15	26.5	49	1.20×
Entropy	50	2.13	6.04	8.18	26.0	64	1.59×
TIP	50	2.68	7.33	10.01	26.8	80	1.95×
TA-OPD	50	2.29	5.91	8.20	27.9	69	1.60×
TA-OPD$_{+Ent.}$	50	1.68	4.64	6.32	26.6	50	1.23×