MINI-OPD: ON-POLICY RELATION DISTILLATION

Anonymous authors
Paper under double-blind review



Figure 1: **Headline results of Mini-OPD. Left:** AIME 2024, LiveCodeBench, and SWE-Bench scores and gains over OPD (Table 1). **Right:** output reverse KL in a representative run. Mini-OPD reaches the marked threshold $3.41\times$ earlier in training steps and attains 66% lower KL at the comparison point. These single-run diagnostics are not wall-clock speedups.

ABSTRACT

On-policy distillation (OPD) brings teacher feedback to student-generated trajectories. What the teacher should convey on those trajectories is a separate design choice. We introduce **Mini-OPD**, which supplements output reverse KL with response-only Q–Q, K–K, and V–V relation matching at one teacher–student layer pair. Temperature-scaled cosine relations provide a token-position interface that accommodates unequal feature widths without learned projectors. We characterize the angular geometry preserved under direct position alignment and formulate position transport for different tokenizers. In separate domain-specific 4B experiments, Mini-OPD improves all five mathematics, code, and software-engineering scores over output-only OPD by 1.0–4.2 percentage points. For Qwen3-4B $\rightarrow$ Llama-3.2-3B, it improves all six mathematics and instruction-following scores over SimpleOPD by 0.8–3.7 points. A four-response code diagnostic shows a further 54–56% reduction in residual relation KL at the supervised layer. Together, these results support response-state relations as a useful addition to output supervision on student trajectories.

1 INTRODUCTION

An autoregressive student must continue from the history it creates, including its own mistakes. Conventional knowledge distillation trains on fixed or teacher-generated responses (Hinton et al., 2015; Kim & Rush, 2016), whereas on-policy distillation (OPD) queries a teacher on fresh student rollouts (Agarwal et al., 2024; Gu et al., 2024). This brings feedback to the prefixes encountered by the student. It leaves a separate design question: *what should the teacher communicate at those states beyond its next-token distribution?*

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.

Output KL directly supervises predictions, while changing intermediate representations only through their effect on those predictions. It does not explicitly match how response states relate to one another. Direct hidden-state supervision offers one alternative, but heterogeneous feature spaces require compatible coordinates or a fitted bridge, as in OPRD (Yang et al., 2026). Relations offer another: MiniLMv2 compares Q–Q, K–K, and V–V affinities over token positions rather than hidden coordinates (Wang et al., 2021). The question is whether this position-based interface supplies useful additional supervision on the student's own trajectories, not whether output probabilities necessarily discard reasoning information.

We introduce **Mini-OPD**, which combines output reverse KL with response-only relation supervision on the *same student-generated response* (Figure 2). At one teacher–student layer pair, normalized Q/K/V features define temperature-scaled cosine relations; the student matches the teacher's relation distributions with forward KL. Each row distributes mass over response positions, allowing unequal feature widths without a learned projector. Both targets follow the evolving student policy, without aligning independently generated teacher and student answers. Feature extraction adds training computation, but no teacher rollout or inference-time module. These component self-relations are distinct from native QK attention.

The formulation makes the transferred structure explicit. Under direct position alignment, unit-normalized rows and a shared temperature, exact relation matching recovers the cosine Gram matrix. A high-temperature expansion further identifies the leading dependence on the relation weight and temperature. Different tokenizers introduce a separate position-matching problem, for which we formulate probability transport; the direct-alignment geometry result does not automatically survive that transport.

Our experiments cover domain-specific 4B mathematics, code, and software-engineering runs, 32B-to-7B distillation, and Qwen3-4B $\rightarrow$ Llama-3.2-3B transfer in mathematics and instruction following. On common saved responses, output-only OPD already moves relations toward the teacher; Mini-OPD reduces the remaining discrepancy across Q/K/V, with the largest effect near the supervised layer. This diagnostic characterizes target attainment, while downstream comparisons test utility and the complete layer sweep exposes placement sensitivity. The evidence does not equate lower relation KL with a causal explanation of every task gain.

Our main contributions are:

- We introduce **Mini-OPD**, which augments output distillation with response-only Q–Q/K–K/V–V supervision on the same student rollouts. The target accommodates unequal feature widths without learned projectors or changes to the student's inference architecture.
- We show that relation matching recovers cosine Gram geometry for directly aligned unit-normalized rows at a shared temperature, and derive the leading λ/τ^2 supervision scale at high temperature.
- We demonstrate 1.0–4.2-point gains over OPD on five domain-specific 4B evaluations and 0.8–3.7-point gains over SimpleOPD on six tasks for Qwen3-4B $\rightarrow$ Llama-3.2-3B. A 32B-to-7B comparison extends the empirical scope; common-response diagnostics and a full layer sweep characterize alignment and sensitivity.

2 Background

2.1 On-Policy Distillation

For $x \sim \mathcal{D}$, a detached student snapshot $\bar{\theta} = \mathrm{sg}(\theta)$ samples $\hat{\mathbf{y}} \sim \pi_{\bar{\theta}}(\cdot \mid x)$. With a shared vocabulary $\mathcal{V}$, teacher feedback on these prefixes defines $p_t(v) = \pi_\theta(v \mid x, \hat{\mathbf{y}}_{<t})$ and $q_t(v) = \pi_{\mathrm{T}}(v \mid x, \hat{\mathbf{y}}_{<t})$. We use output reverse KL (Agarwal et al., 2024; Gu et al., 2024):

$$\mathcal{L}_{\mathrm{out}}(\theta) \coloneqq \mathbb{E}_{x\sim\mathcal{D},\, \hat{\mathbf{y}}\sim\pi_{\bar{\theta}}(\cdot|x)} \left[\frac{1}{T} \sum_{t=1}^{T} \sum_{v\in\mathcal{V}} p_t(v) \log \frac{p_t(v)}{q_t(v)} \right]. \tag{1}$$

Here T is the valid response length; logarithms are natural unless noted.

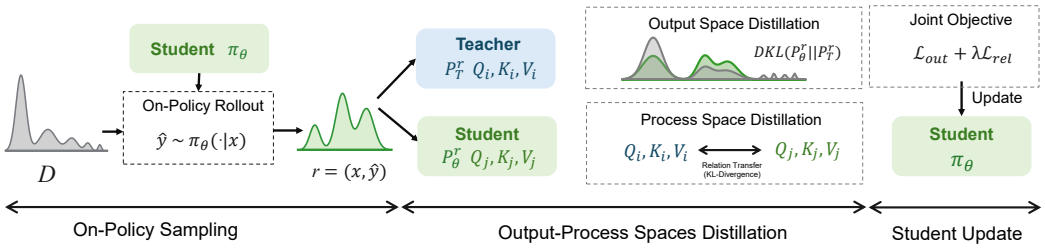


Figure 2: **Mini-OPD.** A stopped-gradient student snapshot samples a response. The frozen teacher and trainable student process that same sequence. The upper branch matches output distributions; the lower branch transfers Q–Q, K–K, and V–V response-state relations at one layer pair. Both losses update only the student. Relation transfer does not match native $\mathrm{QK}^\top$ attention.

2.2 Relation Distillation

MiniLMv2 forms Q–Q/K–K/V–V distributions over token positions, with relation heads decoupled from native attention heads (Wang et al., 2021). Unlike hidden-coordinate matching (Jiao et al., 2020), this permits unequal feature widths when positions align. For teacher row $\mathbf{q}$ and student row $\mathbf{p}$, our directional convention is

$$D_{\mathrm{rel}}^{\xi}(\mathbf{q}, \mathbf{p}) \coloneqq \begin{cases} D_{\mathrm{KL}}(\mathbf{q} \| \mathbf{p}), & \xi = \mathrm{FKL}, \\ D_{\mathrm{KL}}(\mathbf{p} \| \mathbf{q}), & \xi = \mathrm{RKL}. \end{cases} \tag{2}$$

Both act on component self-relations, not native $\mathrm{QK}^\top$ attention. Reported runs use FKL; RKL is analytical only (Appendix A.2).

3 Method

3.1 One Rollout, Two Supervision Targets

We start with directly aligned positions under a shared tokenizer; distinct tokenizations are treated in Section 3.5. For $x \sim \mathcal{D}$, a detached student snapshot samples a fresh response:

$$\bar{\theta} \coloneqq \mathrm{sg}(\theta), \qquad \hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_T) \sim \pi_{\bar{\theta}}(\cdot \mid x). \tag{3}$$

The frozen teacher and trainable student then process the same realized sequence

$$\mathbf{z} = (\mathbf{x}, \hat{\mathbf{y}}) = (z_1, \ldots, z_N), \qquad N = P + T, \tag{4}$$

where P is the prompt length. The two targets use different index shifts: output KL uses the T predictive states at positions $P, \ldots, N-1$, whereas relation KL selects the valid response-token states at positions $P+1, \ldots, N$.

Both targets are computed on the same realized response, rather than on separate teacher and student trajectories. Feature extraction requires auxiliary teacher and student forwards, but no teacher-generated response.

3.2 Response-State Relation Operator

Let $\mathcal{M} = \{\mathrm{T}, \mathrm{S}\}$ index teacher and student and $\mathcal{C} = \{\mathrm{Q}, \mathrm{K}, \mathrm{V}\}$ index attention components. At selected layer ℓ^M, let $\mathbf{Z}_{c,a}^M \in \mathbb{R}^{T \times d_{c,a}^M}$ contain the valid response-token states of native head a. In the Qwen3 runs, Q and K are extracted after per-head QK RMSNorm and rotary embedding; V is the raw value projection. Concatenate native heads, then repartition the flattened width into a common number H of contiguous *relation heads*:

$$\begin{aligned} \mathbf{X}_c^M &\coloneqq [\mathbf{Z}_{c,1}^M \mid \cdots \mid \mathbf{Z}_{c,n_c^M}^M] \in \mathbb{R}^{T \times D_c^M}, \\ (\mathbf{X}_{c,1}^M, \ldots, \mathbf{X}_{c,H}^M) &\coloneqq \mathrm{Split}_H(\mathbf{X}_c^M), \qquad \mathbf{X}_{c,h}^M \in \mathbb{R}^{T \times d_c^M}, \quad d_c^M = D_c^M / H. \end{aligned} \tag{5}$$

The teacher and student share H, but neither native head counts nor d_c^{T} and d_c^{S} need match. We require $H \mid D_c^M$ for every model and component; native heads need not be paired one-to-one.

Each split-head row is ℓ_2-normalized in float32. Writing $\widehat{\mathbf{X}}_{c,h}^M$ for the row-normalized matrix, Mini-OPD maps its normalized Gram geometry to a row-stochastic Gibbs kernel with $\tau > 0$:

$$\begin{aligned} \widehat{\mathbf{x}}_{c,h}^M(i) &\coloneqq \frac{\mathbf{x}_{c,h}^M(i)}{\max(\|\mathbf{x}_{c,h}^M(i)\|_2, \varepsilon_{\mathrm{n}})}, \\ \mathbf{K}_{c,h}^M &\coloneqq \widehat{\mathbf{X}}_{c,h}^M (\widehat{\mathbf{X}}_{c,h}^M)^\top, \\ \mathbf{R}_{c,h}^M &\coloneqq \mathfrak{G}_\tau(\mathbf{X}_{c,h}^M) = \mathrm{softmax}_{\mathrm{row}}(\mathbf{K}_{c,h}^M/\tau) \in (0,1]^{T\times T}. \end{aligned} \tag{6}$$

Here $\varepsilon_{\mathrm{n}} = 10^{-12}$ is the numerical norm floor; rows above it lie on the unit sphere. Thus $R_{c,h}^M(i,j) \propto \exp(\langle \widehat{\mathbf{x}}_{c,h}^M(i), \widehat{\mathbf{x}}_{c,h}^M(j)\rangle/\tau)$; there is no $1/\sqrt{d}$ factor. Q–Q, K–K, and V–V are component self-relations rather than native attention weights. The decoder states remain causally computed, but the relation itself is non-causal over the completed valid response: every row is normalized over all T response positions. In implementation, KL is evaluated from float32 log-softmax values rather than explicit probability ratios.

3.3 On-Policy Relation Objective

Let $\mathbf{r}_{c,h}^M(i) = \mathbf{R}_{c,h}^M[i,:]$ and define the frozen teacher and trainable student rows as $\mathbf{q}_{c,h,i} = \mathrm{sg}[\mathbf{r}_{c,h}^{\mathrm{T}}(i)]$ and $\mathbf{p}_{c,h,i} = \mathbf{r}_{c,h}^{\mathrm{S}}(i)$. Both lie in the interior of the shared token-position simplex Δ^{T-1} even when their source feature widths differ. The relation objective is

$$\mathcal{L}_{\mathrm{rel}}^\xi \coloneqq \mathbb{E}_{\mathrm{on}}\left[\frac{1}{|\mathcal{C}|HT}\sum_{c\in\mathcal{C}}\sum_{h=1}^{H}\sum_{i=1}^{T} D_{\mathrm{rel}}^\xi(\mathbf{q}_{c,h,i}, \mathbf{p}_{c,h,i})\right], \tag{7}$$

where $\mathbb{E}_{\mathrm{on}}$ averages prompts and rollouts from equation 3. The primary configuration uses teacher-to-student KL ($\xi = \mathrm{FKL}$). The conceptual joint objective is

$$\boxed{\mathcal{L}_{\mathrm{M}}^\xi(\theta) = \underbrace{\mathcal{L}_{\mathrm{out}}(\theta)}_{\text{action alignment}} + \lambda \underbrace{\mathcal{L}_{\mathrm{rel}}^\xi(\theta)}_{\text{relation alignment}}}, \qquad \lambda \geq 0. \tag{8}$$

Here $\mathcal{L}_{\mathrm{out}}$ denotes the full-vocabulary idealization; the original runs use a teacher-top-k generalized-KL surrogate, not an unbiased full-KL estimate (Appendix A.5). Writing $\mathcal{L}_{\mathrm{M}}^\xi(\theta;\bar\theta)$ makes the stopped-trajectory update explicit:

$$\mathbf{g}_{\mathrm{upd}} = \partial_\theta \mathcal{L}_{\mathrm{M}}^{\mathrm{FKL}}(\theta;\bar\theta)\big|_{\bar\theta\ \mathrm{fixed}}. \tag{9}$$

Teacher targets and sampled tokens are detached; no score-function gradient passes through generation. Appendix A.4 specifies finite-batch weighting.

For a relation row $\mathbf{p} = \mathrm{softmax}(\mathbf{a})$, FKL gives the logit gradient $\mathbf{p} - \mathbf{q}$. The Fisher geometry, RKL comparison, and support qualifications are derived in Appendix A.2; they are not a direction ablation.

3.4 What the Relation Preserves

Proposition 1 (Recovery of angular geometry) *Fix one response, component, and relation head. If teacher and student rows have unit norm and use the same $\tau > 0$, then*

$$\begin{aligned} K_{ij}^M &= 1 + \tau \log \frac{R_{ij}^M}{R_{ii}^M}, \\ \mathbf{R}^{\mathrm{T}} = \mathbf{R}^{\mathrm{S}} \quad &\Longleftrightarrow \quad \mathbf{K}^{\mathrm{T}} = \mathbf{K}^{\mathrm{S}}. \end{aligned} \tag{10}$$

The self-similarity $K_{ii}^M = 1$ anchors each row: dividing by R_{ii}^M cancels the softmax normalizer, recovering the cosine similarities (Appendix A.1). The target therefore preserves within-response

angular geometry without fixing a feature basis. For positive diagonal $\mathbf{D}$ and orthogonal $\mathbf{U}$ within a relation head,

$$\widetilde{\mathbf{X}} = \mathbf{D}\mathbf{X}\mathbf{U} \quad \Longrightarrow \quad \mathfrak{G}_\tau(\widetilde{\mathbf{X}}) = \mathfrak{G}_\tau(\mathbf{X}). \tag{11}$$

Both statements require inactive norm floors. Comparing unequal widths is well-defined, but exact matching may be impossible when the student's width is below the teacher Gram rank. These are properties of directly aligned relations, not guarantees for arbitrary transported targets.

Cost. For directly aligned responses, total width $D_\Sigma = \sum_{M,c} D_c^M$, and response-row chunk size $B \le T$, relation construction has

$$\mathrm{Time}_{\mathrm{rel}} = \mathcal{O}(T^2 D_\Sigma), \qquad \mathrm{Workspace}_{\mathrm{rel}} = \mathcal{O}(T D_\Sigma + HBT). \tag{12}$$

Chunking with recomputation reduces the quadratic workspace term by T/B, not arithmetic, and leaves the objective unchanged. These bounds exclude feature-forward activations and total training memory. Unselected layers can retain FlashAttention (Dao et al., 2022); any required eager feature capture is confined to the selected pair, not cost-free. Appendix A.6 details kernel dispatch.

3.5 Alignment Across Tokenizers

When the same response text yields T_{T} teacher tokens and T_{S} student tokens, matching token indices is invalid. Let $\mathcal{A}$ be the aligned student rows, $a(i)$ the teacher state at the same post-token text boundary, and $\mathbf{\Pi}_i \in \mathbb{R}_{\ge 0}^{T_{\mathrm{T}} \times T_{\mathrm{S}}}$ a row-stochastic column map. Treating distributions as column vectors, use

$$\widetilde{\mathbf{q}}_{c,h,i} = \mathrm{sg}\left[\mathbf{\Pi}_i^\top \mathbf{r}_{c,h}^{\mathrm{T}}(a(i))\right], \qquad \mathbf{\Pi}_i \mathbf{1} = \mathbf{1}, \quad i \in \mathcal{A}. \tag{13}$$

This preserves probability mass and permits FKL over matched rows without equating feature coordinates. It does not solve output-vocabulary alignment, which requires a separate compatible objective. Appendix C gives a byte-overlap reference construction, coverage checks, and the limits of the geometry claims. The construction is a mathematical interface, distinct from the reported cross-family comparisons in Section 4.3.

4 Experiments

We test whether relations improve output-only OPD within and across model families and tokenizers: domain-specific 4B runs, 32B-to-7B distillation, and Qwen3-to-Llama transfer. We also compare OPRD recipes. Section 5 studies optimization, layer sensitivity, and fixed-response relation matching.

4.1 Experimental Design

Models and domain-specific data. Table 1 contains separate domain-specific runs, not one checkpoint evaluated everywhere. Mathematics uses a Qwen3-4B student (Qwen Team, 2025) and Polaris-4B-Preview teacher (An et al., 2025),[1] both 36-layer models with the same tokenizer and grouped-query layout. Code uses the same initialization and a code-specialized Qwen3-4B teacher from RL step 390. SWE-Bench uses a separately trained Qwen3-4B agent pair.

Baselines and comparison scope. For the 4B runs, OPD and Mini-OPD share the reported initialization, prompts, sampler, rollout budget, output estimator, schedule, and teacher-scoring protocol; Mini-OPD adds the relation loss. Each samples its own evolving policy, not identical trajectories. The frozen teacher supplies a reference. OPRD (Yang et al., 2026) uses hidden-state alignment on its own rollouts, with the 1–35 layer setting where available.

Optimization and checkpointing. The 4B recipe uses learning rate 5×10^{-6}, no warm-up, and teacher-top-256 output supervision. Prompts are capped at 2,048 tokens and mathematics/code responses at 16,384 tokens. Relation rows are chunked with recomputation. Reported mathematics and code checkpoints are at step 100 with $\lambda = 10^4$ and distilled layers 20 and 26. Appendix B specifies the separate controlled relation-analysis runs.

[1] Public checkpoints: `https://huggingface.co/Qwen/Qwen3-4B` and `https://huggingface.co/POLARIS-Project/Polaris-4B-Preview`.

Table 1: **Main results** (%). Macro is the unweighted mean of five displayed primary scores, excluding pass. Bold: best student, including ties. Δ: Mini-OPD minus OPD (percentage points). AMC pass@4 uses four saved responses per question. Primary gains retain supplied precision; pass gains use displayed scores.

Method	Math					Code		Agentic	Summary
	AIME 24		AMC		Olym.	LCB		SWE-Bench	Macro
	Avg@32	pass@32	Avg@4	pass@4	Avg@1	Avg@4	pass@4	Avg@1	Avg@1
Separate domain-specific 4B teacher–student pairs									
Teacher	77.4	93.3	93.7	97.6	66.2	54.6	62.7	55.0	69.4
Initial student	73.3	90.0	87.0	94.0	62.1	49.5	60.6	40.4	62.5
OPD	74.3	90.0	92.5	95.2	65.3	53.6	60.2	49.2	67.0
OPRD	75.5	90.0	92.2	94.0	64.3	54.0	61.3	50.4	67.3
Mini-OPD	**76.4**	**93.3**	**93.7**	**97.6**	**66.4**	**55.7**	**62.7**	**53.4**	**69.1**
Δ vs. OPD	+2.1	+3.3	+1.2	+2.4	+1.0	+2.2	+2.5	+4.2	+2.1

Benchmarks and metrics. We report exact final-answer accuracy on all 30 AIME 2024 problems, an 83-problem AMC subset, and 675 English text-only OlympiadBench problems (He et al., 2024); execution accuracy on 279 LiveCodeBench v5 problems (Jain et al., 2024); and issue resolution on SWE-Bench Verified (Jimenez et al., 2024). SWE-Bench uses the same R2E-Gym scaffold and Podman environments for every checkpoint; a patch is resolved iff the benchmark harness returns reward one. Its Avg@1 scores one complete agent trajectory per issue; no pass metric is reported.

The 4B mathematics/code decoding uses temperature 1.0, top-$p = 0.7$, disabled top-k, and a 65,536-token cap. Avg@n averages binary correctness over n responses per problem, estimating single-generation accuracy; pass@N measures any-success within N attempts (Chen et al., 2021). SWE-Bench uses temperature 1.0, at most 100 agent steps, and 8,192 tokens per completion.

4.2 Main Results

Table 1 compares 4B runs under these domain-specific settings.

Relations improve an already effective output baseline. OPD raises Macro from 62.5 to 67.0; Mini-OPD further raises it to 69.1 and improves all five primary scores in Table 1. Reported gains average 1.4 points in mathematics and reach 2.2 on code and 4.2 on issue resolution. AIME pass@32 and AMC/LCB pass@4 also improve by 3.3, 2.4, and 2.5 points. Thus the aggregate gain is not driven by one benchmark or by trading single-generation accuracy for additional attempts.

Comparison with representation matching. Mini-OPD exceeds OPRD on all five primary scores, including SWE-Bench (53.4 versus 50.4); their Macros are 69.1 and 67.3. The advantage in these same-width pairs shows that the relational target is useful beyond accommodating width mismatch. Different layer coverage and unverified OPRD output-loss settings prevent an isolated loss-level attribution; the OPD comparison matches updates, not compute.

Teacher-relative performance. Mini-OPD matches or exceeds displayed teacher scores on AMC, OlympiadBench, and LiveCodeBench, but remains below on AIME and SWE-Bench. This uneven headroom argues against treating teacher accuracy as a universal ceiling; the point estimates do not establish general student superiority.

Distilling a larger teacher into a smaller student. Table 2 extends the comparison to DeepSeek-R1-Distill-Qwen-32B → 7B. Mini-OPD reaches 64.7 on AIME 24, 94.0 on MATH-500, and 44.9 on LCB, exceeding OPD by 2.6, 0.7, and 2.4 points. Macro increases from 66.0 to 67.9. It also exceeds OPRD-Bridge on all three primary scores, by 3.4, 0.3, and 0.7 points. This extends the positive pattern beyond equal-size pairs, without requiring a learned feature projector.

Table 2: **DeepSeek-R1-Distill-Qwen: 32B teacher → 7B student** (%). Macro is the unweighted mean of three displayed primary scores, excluding pass. Bold: best student, including ties. Δ: Mini-OPD minus OPD (percentage points). AIME 24 reports Avg@32 and pass@32; LCB v5 reports Avg@1.

Method	Math			Code	Summary
	AIME 24		MATH-500	LCB v5	Macro
	Avg@32	pass@32	Avg	Avg@1	Avg@1
DeepSeek-R1-Distill-Qwen-32B ⟶ DeepSeek-R1-Distill-Qwen-7B					
Teacher	80.8	93.3	94.3	57.7	77.6
Initial student	52.1	80.0	92.8	35.4	60.1
OPD (9k)	62.1	86.7	93.3	42.5	66.0
OPRD-Bridge	61.3	84.3	93.7	44.2	66.4
Mini-OPD	**64.7**	**87.6**	**94.0**	**44.9**	**67.9**
Δ vs. OPD	+2.6	+0.9	+0.7	+2.4	+1.9

AIME pass@32 increases from 86.7 for OPD to 87.6 for Mini-OPD, also exceeding OPRD-Bridge's 84.3 and the initial student's 80.0. Both per-generation accuracy and reported multi-attempt success therefore improve in this setting. Mini-OPD is the strongest student on every reported metric, while remaining below the teacher. These point estimates support transfer across the tested size gap, not a claim of complete teacher recovery or matched-compute superiority.

4.3 Cross-Tokenizer Transfer

We distill Qwen3-4B into Llama-3.2-3B in separate Math and IF runs. Math covers AIME 2025/2026 and MATH-500; IF covers IFEval, AdvanceIF, and the multi-turn MultiChallenge (Sirdeshmukh et al., 2025). Both family and tokenization change. SimpleOPD (He et al., 2026) is the identical-span output baseline; its published stabilizers are described in Appendix C. OPRD-Bridge uses projected features; Mini-OPD uses response-state relations.

Higher scores in the cross-family setting. Mini-OPD has the highest reported student primary score on all six tasks (Table 3). Gains over SimpleOPD are 2.6–3.4 points in Math and 0.8–3.7 in IF; descriptive Macro rises from 26.0 to 28.8 (+2.8), versus 24.4 for OPRD-Bridge. MATH-500 improves from 55.4 to 58.0, IFEval from 70.1 to 73.3, and MultiChallenge from 15.4 to 19.1. Gains therefore extend beyond low-baseline AIME scores to instruction following and multi-turn dialogue. The student nevertheless remains below the teacher throughout.

Mean and pass@5 gains differ. Pass@5 increases by 5.4, 0.6, and 1.1 points on AIME 2025, AIME 2026, and IFEval. Their unequal changes again separate average accuracy from multi-attempt coverage. Together, the three tables support relational supervision across the tested domains, size gaps, and tokenizers, not a universal transfer guarantee. Family and tokenizer changes are coupled; protocol and aggregation limits remain in Appendix C.

5 Analysis

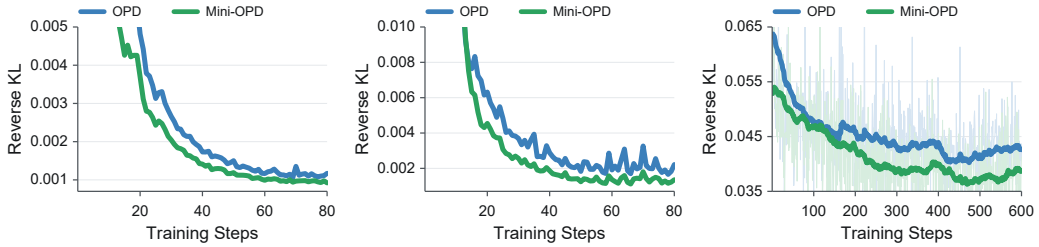
We examine output-loss dynamics, layer sensitivity, and attainment of the relational target, separating policy-specific training curves from fixed-response diagnostics.

5.1 Output-Space Optimization Dynamics

Mini-OPD reaches lower logged output reverse KL in all three domains, with the largest separation in code (Figure 3). Each panel fixes the model pair, output objective, sampler, and estimator, but policies visit different evolving prefixes. These curves are not a common-prefix evaluation.

Table 3: **Cross-tokenizer transfer** (%). Macro is the unweighted mean of six displayed primary scores, excluding pass. Bold: best student, including ties. Δ: Mini-OPD minus SimpleOPD (percentage points). AIME and IFEval report Avg@5 and pass@5; the other tasks have no pass metric. MultiCh.: MultiChallenge.

Method	Math					Instruction Following				Summary
	AIME 25		AIME 26		MATH-500	IFEval		AdvanceIF	MultiCh.	Macro
	Avg@5	pass@5	Avg@5	pass@5	Avg	Avg@5	pass@5	Avg@1	Avg@1	Avg@1
Teacher: Qwen3-4B ⟶ **Student:** Llama-3.2-3B										
Teacher	14.7	23.3	16.0	30.0	82.2	80.8	88.5	13.5	20.3	37.9
Initial student	0	0	0.7	3.3	45.6	67.5	84.8	12.2	14.2	23.4
SimpleOPD	1.3	3.3	1.3	6.7	55.4	70.1	86.0	12.6	15.4	26.0
OPRD-Bridge	1.0	3.3	1.3	6.7	47.1	67.6	86.3	12.4	16.9	24.4
Mini-OPD	**4.3**	**8.7**	**4.7**	**7.3**	**58.0**	**73.3**	**87.1**	**13.4**	**19.1**	**28.8**
Δ	+3.0	+5.4	+3.4	+0.6	+2.6	+3.2	+1.1	+0.8	+3.7	+2.8



(a) Mathematical reasoning (b) Code generation (c) SWE-Bench

Figure 3: **Output reverse-KL convergence.** Lower is better. Each panel compares matched OPD and Mini-OPD runs; axes are independently scaled, so comparisons are within panel. In (c), faint lines are raw observations and saturated lines are a time-weighted EMA with 20-step half-life.

Figure 1 marks a representative $3.41\times$ reduction in steps to the chosen divergence threshold and 66% lower KL at the comparison point. These single-run, step-based measurements do not account for extra feature forwards and are not wall-clock speedups.

5.2 Sensitivity to the Distilled Layer

Layer placement matters. Of 32 mathematics configurations, 30 have complete means ranging from 76.80 to 78.80; 18 exceed OPD's 77.36. Layer 20 leads, but gains vary across layers and tasks. This is a selected maximum, not an unbiased layer-selection estimate or independent-seed evidence; Appendix D reports the full sweep.

5.3 Direct Relation Alignment

We score teacher and students on identical saved responses using teacher-to-student relation KL and the training response mask. This is a separate step-100 code run at layer 16 with $H = 16$, $\tau = 5$, and $\lambda = 10^4$, not the main table's layer-26 checkpoint.

OPD already moves relations toward the teacher; Mini-OPD reduces the residual further across Q/K/V. Lower relation KL is expected because it is optimized directly.

The informative pattern is consistency across components and a peak near the supervised layer. The four-response component comparison and three-response depth probe show local target attainment, not a causal mechanism. Appendix B includes all six placements and the temperature diagnostic, retaining the degenerate layer-28 run.

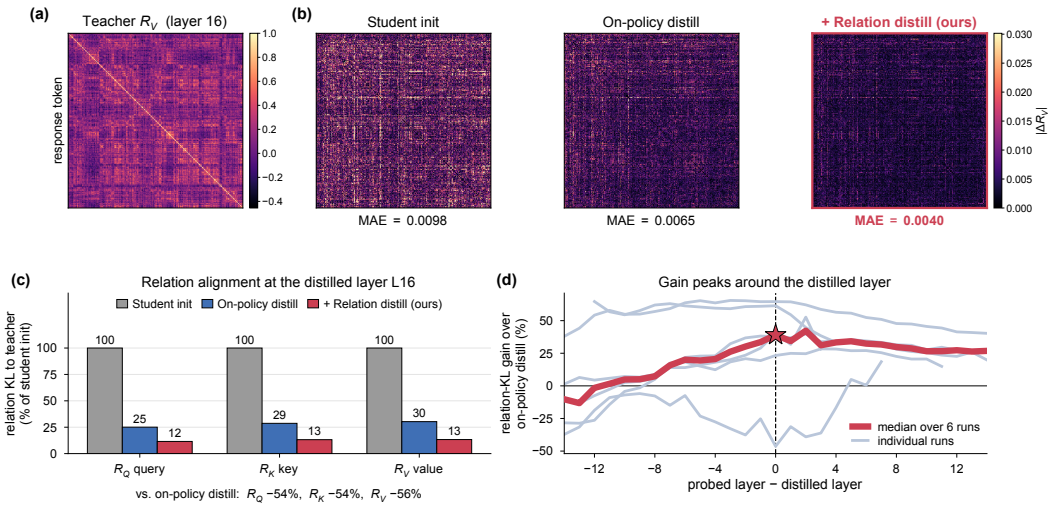


Figure 4: **Direct relation alignment** in the controlled code run at step 100 and layer 16. (a–b) Exported teacher and student relation-map visualizations. Quantitative comparisons below use the row-wise KL in (c), not the heatmap's MAE annotations. (c) Q–Q/K–K/V–V KL residuals are $25/29/30\%$ for OPD and $12/13/13\%$ for Mini-OPD, reductions of $54/54/56\%$. (d) Across six layer-placement runs, median gain peaks at the supervised-layer offset (39.0% versus 4.9% at offset -10).

6 Related Work

Output supervision on student trajectories. GKD and MiniLLM established student-rollout supervision (Agarwal et al., 2024; Gu et al., 2024). Recent work examines signal quality along student prefixes (Li et al., 2026a; Liu et al., 2026) and adapts supervision to teacher uncertainty (Jin et al., 2026). Mini-OPD adds an internal relational target rather than modifying output feedback alone.

Internal supervision. MiniLMv2 transfers Q–Q/K–K/V–V relations on offline inputs (Wang et al., 2021). OPRD aligns on-policy hidden coordinates, fitting and freezing a low-rank bridge for heterogeneous models (Yang et al., 2026). PHF transfers normalized hidden-state transitions and their geometry from a privileged EMA self-teacher (Li et al., 2026b). Mini-OPD instead adds response-only cosine QKV relations at one layer pair to external-teacher output supervision. Neither internal nor relational supervision alone is new.

Cross-tokenizer interfaces. SimpleOPD retains output feedback on identical text spans (He et al., 2026); SimCT constructs output preferences over shared tokens and jointly tokenizable continuation units (Sun et al., 2026). Mini-OPD's position transport concerns internal relation distributions, not output-token probabilities, and does not replace output alignment. Appendix E gives further comparisons.

7 Conclusion

Mini-OPD broadens supervision on student trajectories from next-token behavior to response-state relations. Its one-layer Q–Q/K–K/V–V objective avoids learned feature projectors, with a precise geometric interpretation under direct alignment and a separate position-transport formulation. Reported 4B, 32B-to-7B, and cross-tokenizer comparisons favor Mini-OPD on primary scores; diagnostics show closer local relation matching. The evidence supports this additional target within the tested settings, rather than a general efficiency or causal claim. Mini-OPD thus connects two design choices in distillation: where the student learns and which teacher structure it learns.

AI USE STATEMENT

Generative AI tools assisted with manuscript preparation and revision, including language polishing and presentation. The authors are responsible for reviewing and validating all content and take full responsibility for the final submission.

REPRODUCIBILITY STATEMENT

The method and appendices specify the relation objective, geometric results, Qwen3 feature extraction, loss reduction, and computational cost. Section 4.1 gives the reported training and evaluation settings; the appendices retain the complete layer sweep and diagnostic protocols. The reference cross-tokenizer construction is distinguished from the reported runs, whose implementation and aggregation records are incomplete.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Chenxin An, Zhihui Xie, Xiaonan Li, Lei Li, Jun Zhang, Shansan Gong, Ming Zhong, Jingjing Xu, Xipeng Qiu, Mingxuan Wang, and Lingpeng Kong. POLARIS: A post-training recipe for scaling reinforcement learning on advanced reasoning models. Technical report and model release, 2025.

Mohammadreza Armandpour, Fatih Ilhan, David Harrison, Ajay Jaiswal, Duc N. M. Hoang, Fartash Faghri, Yizhe Zhang, Minsik Cho, and Mehrdad Farajtabar. Unmasking on-policy distillation: Where it helps, where it hurts, and why. *arXiv preprint arXiv:2605.10889*, 2026. URL https://arxiv.org/abs/2605.10889v2.

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. URL https://arxiv.org/abs/2107.03374.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022. URL https://arxiv.org/abs/2205.14135.

DeepSeek-AI. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.08543.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2402.14008.

Haonan He, Haodi Lei, Yun Luo, Haoran Zhang, Shunkai Zhang, Yizhuo Li, Shengji Tang, Zhilin Wang, Runzhe Zhan, Lei Bai, Ganqu Cui, Fangchen Yu, Yafu Li, Peng Ye, Ning Ding, and Yu Cheng. SimpleOPD: Simple tokenizer-agnostic on-policy distillation for long-context reasoning. *arXiv preprint arXiv:2608.14277*, 2026. URL https://arxiv.org/abs/2608.14277v1.

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. NeurIPS Deep Learning Workshop.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14852–14882. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.830. URL https://aclanthology.org/2023.acl-long.830/.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024. URL https://arxiv.org/abs/2403.07974.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.372. URL https://aclanthology.org/2020.findings-emnlp.372/.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06770.

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

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, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

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*, 2026a. URL https://arxiv.org/abs/2604.13016v2.

Yuhan Li, Mingxu Zhang, Dazhong Shen, and Ying Sun. PHF: Privileged hidden flow for on-policy self-distillation. *arXiv preprint arXiv:2606.29340*, 2026b. URL https://arxiv.org/abs/2606.29340v1.

Yanjiang Liu, Jie Lou, Xinyan Guan, Yuqiu Ji, Hongyu Lin, Ben He, Xianpei Han, Le Sun, Xing Yu, and Yaojie Lu. Your teacher can't help you here: Combating supervision fidelity decay in on-policy distillation. *arXiv preprint arXiv:2605.30833*, 2026. URL https://arxiv.org/abs/2605.30833v1.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training. *arXiv preprint arXiv:2606.30406*, 2026. URL https://arxiv.org/abs/2606.30406.

Jingyu Peng, Maolin Wang, Hengyi Cai, Yuchen Li, Kai Zhang, Shuaiqiang Wang, Dawei Yin, and Xiangyu Zhao. AdaSwitch: Balancing exploration and guidance in knowledge distillation via adaptive switching. *arXiv preprint arXiv:2510.07842*, 2025. URL https://arxiv.org/abs/2510.07842.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. *arXiv preprint arXiv:2501.17399*, 2025. URL https://arxiv.org/abs/2501.17399v2.

Jie Sun, Mao Zheng, Mingyang Song, Qiyong Zhong, Yilin Cheng, Bichuan Feng, Pengfei Liu, Junfeng Fang, and Xiang Wang. SimCT: Recovering lost supervision for cross-tokenizer on-policy distillation. *arXiv preprint arXiv:2605.07711*, 2026. URL https://arxiv.org/abs/2605.07711v2.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. Knowledge fusion of large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.10491.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2140–2151. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.188. URL https://aclanthology.org/2021.findings-acl.188/.

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 (Volume 1: Long Papers)*, pp. 10817–10834. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.605. URL https://aclanthology.org/2023.acl-long.605/.

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*, 2026. URL https://arxiv.org/abs/2606.06021.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025. URL https://arxiv.org/abs/2511.10643.

Songming Zhang, Xue Zhang, Tong Zhang, Bojie Hu, Yufeng Chen, and Jinan Xu. AlignDistil: 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)*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.972. URL https://aclanthology.org/2025.acl-long.972/.

A Additional Method Details

A.1 Geometry Preserved by the Relation Operator

Proof of Proposition 1. Fix one response, component, and relation head; teacher and student have the same T positions but may have different feature widths. Assume the norm floor is inactive, so every normalized row has unit norm and $K^M_{ii} = 1$. Dividing a softmax entry by its diagonal entry cancels the row normalizer:

$$\frac{R^M_{ij}}{R^M_{ii}} = \frac{\exp(K^M_{ij}/\tau)}{\exp(K^M_{ii}/\tau)} = \exp\left(\frac{K^M_{ij} - 1}{\tau}\right). \tag{14}$$

Taking logarithms yields equation 10. Equal relation matrices therefore imply equal cosine Gram matrices; the converse follows immediately by applying the same row softmax. Equivalently, softmax identifies logits only up to a row constant, and the unit diagonal fixes that constant. The statement holds in exact arithmetic at any finite positive temperature.

Invariance and representability. For positive diagonal $\mathbf{D}$ and orthogonal $\mathbf{U}$ acting within one relation head, exact normalization gives $\mathcal{N}(\mathbf{DXU}) = \mathcal{N}(\mathbf{X})\mathbf{U}$. Multiplication by its transpose recovers the same Gram matrix, proving equation 11. This is not invariance to arbitrary mixing between relation heads. Nor does a width-independent comparison guarantee realizability: a necessary condition for exact teacher matching is

$$\operatorname{rank}(\mathbf{K}^{\mathrm{T}}) = \operatorname{rank}(\mathbf{K}^{\mathrm{S}}) \le d^{\mathrm{S}}. \tag{15}$$

The condition is necessary for feature matrices, not sufficient for a parameterized student network to attain the target. If the norm floor is active, the diagonal need not be one; neither exact recovery nor arbitrary positive-rescaling invariance is asserted for those rows.

Small KL and numerical conditioning. Exact identifiability should not be confused with a temperature-independent stability guarantee. Since $K_{ij} \in [-1, 1]$, every relation entry satisfies

$$0 < a_{T,\tau} \coloneqq \frac{e^{-2/\tau}}{T} \le R_{ij} \le \min\left(1, \frac{e^{2/\tau}}{T}\right). \tag{16}$$

For unit-row teacher and student matrices, the logarithm is $1/a_{T,\tau}$-Lipschitz on their probability range. Applying equation 10 and Pinsker's inequality to row i gives, for $i \neq j$,

$$|K^{\mathrm{T}}_{ij} - K^{\mathrm{S}}_{ij}| \le \frac{\tau}{a_{T,\tau}} \|\mathbf{r}^{\mathrm{T}}(i) - \mathbf{r}^{\mathrm{S}}(i)\|_1 \le \frac{\tau}{a_{T,\tau}} \sqrt{2 D_{\mathrm{KL}}(\mathbf{r}^{\mathrm{T}}(i) \| \mathbf{r}^{\mathrm{S}}(i))}. \tag{17}$$

The diagonal difference is zero. This bound can be loose and depends on both length and temperature; a small numerical KL at high temperature alone need not imply a small geometry error.

A.2 Directional Relation-KL Derivatives

On the interior of the probability simplex, KL is the Bregman divergence generated by negative Shannon entropy:

$$\begin{aligned} \varphi(\mathbf{u}) &\coloneqq \sum_j u_j \log u_j, \\ B_\varphi(\mathbf{u}, \mathbf{v}) &\coloneqq \varphi(\mathbf{u}) - \varphi(\mathbf{v}) - \langle \nabla\varphi(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle = D_{\mathrm{KL}}(\mathbf{u} \| \mathbf{v}). \end{aligned} \tag{18}$$

For a fixed response row, suppress (c, h, i) and write $\mathbf{q} = \mathrm{sg}[\mathbf{r}^{\mathrm{T}}]$, $\mathbf{p} = \mathbf{r}^{\mathrm{S}} = \mathrm{softmax}(\mathbf{a})$, and $\delta^\xi = D^\xi_{\mathrm{rel}}(\mathbf{q}, \mathbf{p})$, with $p_j, q_j > 0$. The categorical Fisher matrix is the softmax Jacobian:

$$\mathbf{F}(\mathbf{p}) = \mathrm{Diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top, \qquad \mathbf{F}(\mathbf{p})\mathbf{1}_T = \mathbf{0}. \tag{19}$$

The probability gradients are $-\mathbf{q} \oslash \mathbf{p}$ for FKL and $\mathbf{1}_T + \log(\mathbf{p} \oslash \mathbf{q})$ for RKL. Applying the chain rule and $\mathbf{F}(\mathbf{p})\mathbf{1}_T = 0$ gives

$$\begin{aligned} \nabla_{\mathbf{a}} \delta^{\mathrm{FKL}} &= \mathbf{F}(\mathbf{p})(-\mathbf{q} \oslash \mathbf{p}) = \mathbf{p} - \mathbf{q}, \\ \nabla_{\mathbf{a}} \delta^{\mathrm{RKL}} &= \mathbf{F}(\mathbf{p}) \log(\mathbf{p} \oslash \mathbf{q}) \\ &= \mathbf{p} \odot [\log(\mathbf{p} \oslash \mathbf{q}) - D_{\mathrm{KL}}(\mathbf{p} \| \mathbf{q}) \mathbf{1}_T]. \end{aligned} \tag{20}$$

Here $\odot$ and $\oslash$ denote elementwise multiplication and division. The Fisher matrix is positive semidefinite with $\mathrm{Null}(\mathbf{F}(\mathbf{p})) = \mathrm{span}\{\mathbf{1}_T\}$ for interior $\mathbf{p}$. Hence both directions preserve softmax gauge symmetry and share a unique distributional fixed point:

$$\begin{aligned} \nabla_{\mathbf{a}}\delta^{\xi} \in T_{\mathbf{p}}\Delta^{T-1} = \mathbf{1}_T^{\perp}, \quad \mathbf{1}_T^{\top}\nabla_{\mathbf{a}}\delta^{\xi} = 0, \\ \nabla_{\mathbf{a}}\delta^{\xi} = \mathbf{0} \iff \mathbf{p} = \mathbf{q}, \qquad\qquad \xi \in \{\mathrm{FKL}, \mathrm{RKL}\}. \end{aligned} \tag{21}$$

Their behavior differs when the student nearly ignores a teacher-supported position. Hold $\mathbf{q}$ fixed, let the remaining student coordinates converge in the simplex interior, and take $p_k \to 0^+$ with $q_k > 0$. Then

$$\begin{aligned} [\nabla_{\mathbf{a}}\delta^{\mathrm{FKL}}]_k &= -q_k + \mathcal{O}(p_k), \\ [\nabla_{\mathbf{a}}\delta^{\mathrm{RKL}}]_k &= \mathcal{O}(p_k|\log p_k|), \\ \therefore \quad \lim_{p_k\to 0^+} \left([\nabla_{\mathbf{a}}\delta^{\mathrm{FKL}}]_k, [\nabla_{\mathbf{a}}\delta^{\mathrm{RKL}}]_k\right) &= (-q_k, 0). \end{aligned} \tag{22}$$

where the second equality uses $\lim_{u\to 0^+} u\log u = 0$. This is a generic categorical-softmax limit: the teacher-weighted branch retains a finite correction while the reverse branch vanishes. It is *not* a limit available to the fixed-T, fixed-τ cosine-relation family, whose entries have the positive lower bound in equation 16. It therefore does not establish FKL's empirical superiority in Mini-OPD.

A.3 Temperature and the Local Quadratic Objective

Fix finite T and fixed cosine-logit rows $\mathbf{k}^{\mathrm{T}}, \mathbf{k}^{\mathrm{S}} \in [-1, 1]^T$. Let $\beta = 1/\tau$, $\mathbf{u} = \mathbf{1}_T/T$, and $\mathbf{C}_T = \mathbf{I} - \mathbf{1}_T\mathbf{1}_T^{\top}/T$. For directly aligned rows with softmaxes $\mathbf{q}_\tau$ and $\mathbf{p}_\tau$,

$$D_{\mathrm{KL}}(\mathbf{q}_\tau\|\mathbf{p}_\tau) = \frac{\|\mathbf{C}_T(\mathbf{k}^{\mathrm{T}} - \mathbf{k}^{\mathrm{S}})\|_2^2}{2T\tau^2} + \mathcal{O}(\tau^{-3}), \qquad \tau \to \infty. \tag{23}$$

This fixed-length expansion identifies λ/τ^2 as the leading loss scale, not an equivalence of finite-temperature training runs. Expanding softmax about zero gives

$$\mathrm{softmax}(\beta\mathbf{k}) = \mathbf{u} + \frac{\beta}{T}\mathbf{C}_T\mathbf{k} + \mathcal{O}(\beta^2). \tag{24}$$

For an explicit KL derivation, write $A(\mathbf{a}) = \log\sum_j e^{a_j}$, $\mathbf{a}_{\mathrm{T}} = \beta\mathbf{k}^{\mathrm{T}}$, and $\mathbf{a}_{\mathrm{S}} = \beta\mathbf{k}^{\mathrm{S}}$. The exact identity is

$$D_{\mathrm{KL}}(\mathbf{q}_\tau\|\mathbf{p}_\tau) = A(\mathbf{a}_{\mathrm{S}}) - A(\mathbf{a}_{\mathrm{T}}) - \nabla A(\mathbf{a}_{\mathrm{T}})^{\top}(\mathbf{a}_{\mathrm{S}} - \mathbf{a}_{\mathrm{T}}). \tag{25}$$

Because $\nabla^2 A(\mathbf{0}) = \mathbf{C}_T/T$ and $\mathbf{C}_T^2 = \mathbf{C}_T$, Taylor expansion of this Bregman divergence yields equation 23; third-order terms contribute $\mathcal{O}(\tau^{-3})$. The leading term compares *centered* cosine rows. The remainder is not uniform over arbitrary response lengths, nor does the expansion establish an empirical tuning rule.

A.4 Variable-Length and Microbatch Reduction

For a nonempty response b of valid length T_b, define its implemented teacher-to-student relation loss by

$$\ell_{\mathrm{rel},b} = \frac{1}{|\mathcal{C}|HT_b}\sum_{c,h}\sum_{i=1}^{T_b} D_{\mathrm{KL}}(\mathbf{q}_{c,h,i,b}\|\mathbf{p}_{c,h,i,b}). \tag{26}$$

The relation kernel first averages these per-response values within each microbatch $\mathcal{B}_m$, rather than pooling all relation rows across unequal-length responses. In the inspected actor, the gradient contribution is scaled by that microbatch's valid response-token count:

$$\widehat{\mathcal{L}}_{\mathrm{rel}} = \sum_m w_m \frac{1}{|\mathcal{B}_m|}\sum_{b\in\mathcal{B}_m}\ell_{\mathrm{rel},b}, \qquad w_m = \frac{\sum_{b\in\mathcal{B}_m} T_b}{N_{\mathrm{step}}}. \tag{27}$$

Here N_{step} is the valid-token normalization supplied by the training driver; distributed synchronization follows the training backend. Empty responses contribute zero and remain in the microbatch denominator. The output kernel uses its masked token reduction separately. Consequently, unequal-length microbatch partitioning can affect the relative weighting and must be held fixed when comparing loss coefficients. The inspected relation implementation computes FKL only; the RKL equations above analyze an alternative, not an additional reported training branch.

A.5 Output Reverse KL Estimation

The conceptual output objective in equation 1 is the full-vocabulary reverse KL. The teacher-top-k path instead evaluates a generalized-KL surrogate on $S_t = \mathrm{TopK}_k(q_t)$:

$$\ell_{\mathrm{out},t}^{(k)} = \sum_{v \in S_t} \left[p_t(v) \log \frac{p_t(v)}{q_t(v)} - p_t(v) + q_t(v) \right]. \tag{28}$$

Both probabilities are evaluated on the same teacher-selected token IDs using full-vocabulary normalization; they are not renormalized over S_t. The linear terms correct the generalized divergence on truncated mass, not the omitted tail. In particular,

$$0 \le \ell_{\mathrm{out},t}^{(k)} \le D_{\mathrm{KL}}(p_t \| q_t), \qquad \ell_{\mathrm{out},t}^{(|\mathcal{V}|)} = D_{\mathrm{KL}}(p_t \| q_t). \tag{29}$$

Each summand is nonnegative, and the linear terms cancel over the full vocabulary. This surrogate is not an unbiased estimate of full-vocabulary KL. Using the same estimator and k controls its implementation across methods, but does not rule out different tail errors as their policies evolve. The logged curves do not measure this tail error: they use their training estimator, not an independent full-vocabulary common-prefix evaluation.

A.6 Single-Layer Execution

Only the selected teacher and student layers expose projected Q/K/V. If the backend supplies these before fused attention, every layer retains the optimized kernel; otherwise only the selected pair needs eager fallback. Unselected layers remain optimized. This limits kernel disruption, not the quadratic arithmetic in equation 12; row chunking reduces relation workspace, not feature-forward activations.

A.7 Evaluation Protocol Notes

Macro averages five/three/six displayed primary scores in Tables 1/2/3, excluding pass. Its Avg@1 subheader denotes an aggregate of per-attempt performance estimates, not a common one-sample evaluation budget across benchmarks. Means and differences are rounded after aggregation; they are descriptive, not pooled accuracy. Table 1 mathematics/code gains retain supplied pre-rounding precision, including 1.0 on OlympiadBench and 2.2 on LiveCodeBench; its SWE/pass gains and all Table 2 gains use displayed rates. These descriptive summaries do not establish evaluation uncertainty.

32B-to-7B evaluation scope. Table 2 reports AIME 24 Avg@32/pass@32 and LCB v5 Avg@1; MATH-500's sampling count is unspecified. The @8 evaluation in Liu et al. (2026) does not establish these @32 measurements or their decoding settings. The OPD (9k) row retains its reported recipe identifier; the matched 4B configuration is not assumed for this model pair.

All five AIME average scores are compatible with a 30-problem, 32-response binary evaluation after rounding. However, OPRD-Bridge's 84.3 and Mini-OPD's 87.6 pass@32 do not equal rounded single-repeat any-success fractions over 30 problems. Evaluation repeats, candidate-pool estimators, and denominators remain distinct protocol questions; we retain the reported values without inferring their aggregation or uncertainty.

B Relation-Alignment Analysis

Models and controlled runs. The analysis uses the code-domain family from Section 4.1. The Qwen3-4B initialization and the code-specialized Qwen3-4B teacher from RL step 390 have the same architecture: 36 layers, 32 query heads, 8 key–value heads, hidden size 2,560, and head dimension 128. Neither uses RoPE scaling. We compare all arms at optimization step 100. The principal Mini-OPD run applies the relation loss at layer 16 with $H = 16$, $\tau = 5$, and $\lambda = 10^4$; the layer analysis repeats the same recipe at layers $\{12, 16, 20, 24, 28, 32\}$. These controlled code runs use the reported 16K-response-cap recipe, teacher top-$k = 256$, batch size 32, learning rate 5×10^{-6} without warm-up, and two responses per prompt. They are distinct from the layer-26 code checkpoint in Table 1; placement variants are not independent training seeds. These settings describe the existing Qwen3-4B diagnostics, not the separate DeepSeek 32B-to-7B results.

Measured states and relations. Extraction, float32 normalization, and full-response support follow Section 3.2 and equation 6. The Qwen3 flattened widths are $D_{\mathrm{Q}} = 4096$ and $D_{\mathrm{K}} = D_{\mathrm{V}} = 1024$. Causally computed Q/K/V states are compared through non-causal relations over valid response positions.

Alignment metric and data. Teacher and student are evaluated teacher-forced on identical saved on-policy responses. For an analysis batch $\mathcal{B}$ with response lengths $\{n_b\}_{b\in\mathcal{B}}$, we measure the training direction directly,

$$\mathcal{D}_c(\ell;\mathcal{B}) \coloneqq \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}}\frac{1}{Hn_b}\sum_{(h,i)\in[H]\times[n_b]} D_{\mathrm{KL}}\left(\mathbf{r}^{\mathrm{T},\ell}_{c,h,b}(i)\,\|\,\mathbf{r}^{\mathrm{S},\ell}_{c,h,b}(i)\right),$$

and report the three components separately or their mean. The relation maps and layer-16 KL use four responses of lengths 1,073, 1,073, 833, and 833. The layer sweep uses three responses of length 576 and probes all 36 layers. Heatmaps show relation head 0 and the first 384 response tokens. The exported heatmaps serve as qualitative illustrations; our quantitative alignment comparison uses the post-softmax row-wise KL defined above, not the maps' MAE annotations.

Layer specificity. For a run distilled at $\ell^\star$, define its gain over OPD at probed layer ℓ as

$$\overline{\mathcal{D}}_m(\ell) \coloneqq \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\mathcal{D}_{c,m}(\ell),$$

$$g_{\ell^\star}(\ell) \coloneqq 1 - \frac{\overline{\mathcal{D}}_{\ell^\star}(\ell)}{\overline{\mathcal{D}}_{\mathrm{OPD}}(\ell)}.$$

Thus the bar averages Q–Q, K–K, and V–V, and positive g means lower relation KL than OPD. We align runs by $\ell - \ell^\star$ and report their median; all six runs are retained, including the degenerate layer-28 run.

Temperature diagnostic. For a teacher relation row $\mathbf{r}^{\mathrm{T}}_{c,h}(i;\tau)$, we quantify target structure by its information departure from the uniform relation $\mathbf{u}_n = \mathbf{1}_n/n$:

$$\mathcal{S}_\tau \coloneqq \frac{1}{\ln 2}\mathbb{E}_{c,h,i}\left[D_{\mathrm{KL}}(\mathbf{r}^{\mathrm{T}}_{c,h}(i;\tau)\|\mathbf{u}_n)\right] = \log_2 n - \mathbb{E}_{c,h,i}\left[H_2(\mathbf{r}^{\mathrm{T}}_{c,h}(i;\tau))\right] \quad \text{bits.}$$

Here $\log_2 n$ is the maximum possible value for a distribution over n response positions; $1/\ln 2$ converts the natural-log KL to bits. The expectation averages rows, relation heads, and components.

In Figure 5(d), retained structure is 0.41, 0.019, and approximately 7×10^{-4} bits at $\tau = 0.25$, 1, and 5. At $\tau = 5$, cosine logits lie in $[-0.2, 0.2]$, explaining the nearly uniform target. These measurements characterize the target scale, not an optimal tuning rule.

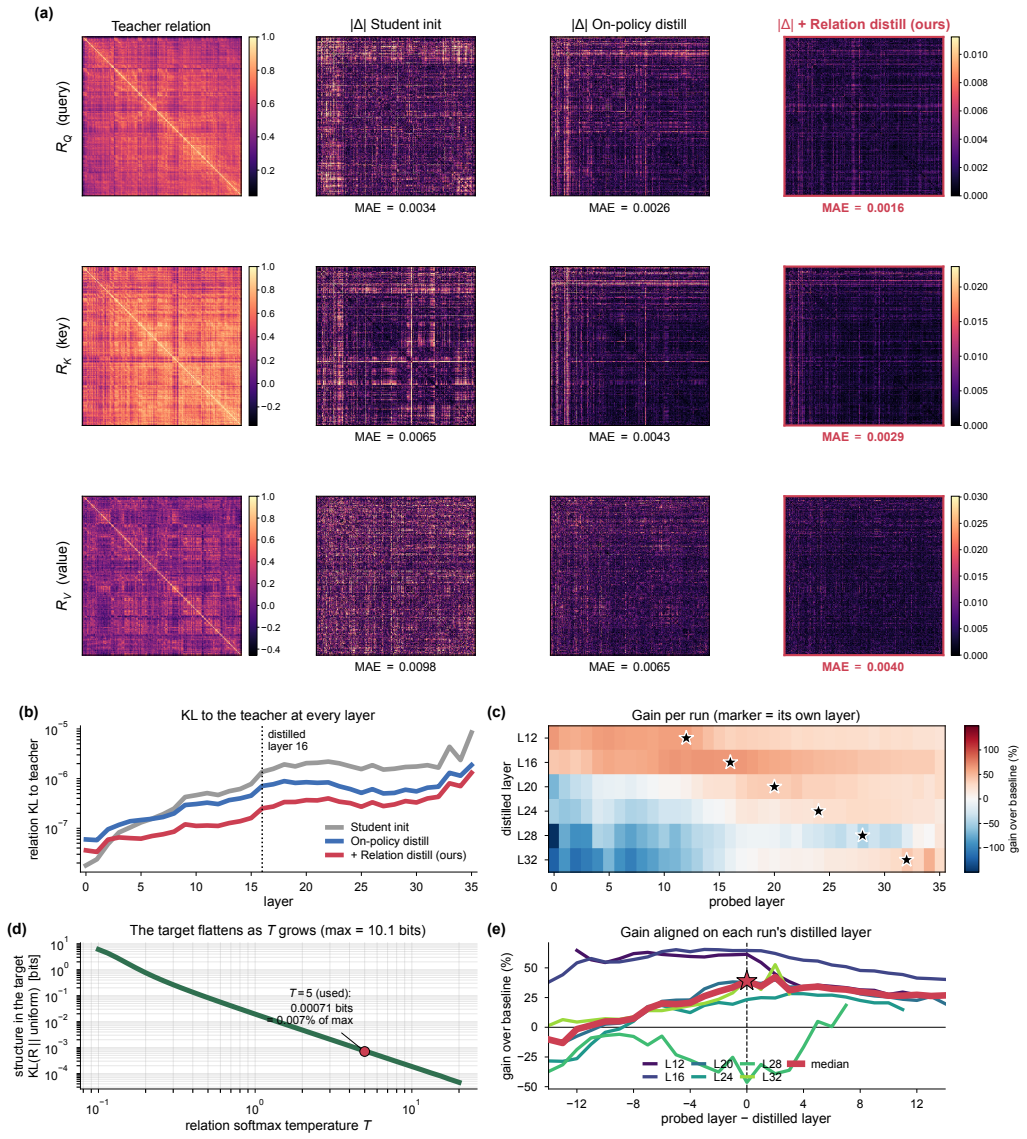


Figure 5: **Full relation-alignment analysis.** (a) Exported Q–Q, K–K, and V–V maps at layer 16, showing the teacher, initialization, OPD, and Mini-OPD. Map annotations are not used as probability-scale error estimates. (b) Relation KL across all 36 layers for the layer-16 run. The largest reduction occurs near the supervised layer and persists above it. (c) Gain over OPD for each layer-placement run; stars mark the layer receiving supervision. (d) Structure retained by the teacher target, measured as $D_{\mathrm{KL}}(\mathbf{R}\|\mathcal{U})$ in bits. At the training temperature $\tau = 5$ (denoted T in the panel), the target retains only about 7×10^{-4} bits, or 0.007% of the 10.1-bit maximum. (e) Per-run gains aligned by each run's distilled layer; the red curve is the median and no run is excluded.

C Cross-Tokenizer Alignment and Evaluation

Different tokenizers represent the same student-generated response with T_{S} student tokens and T_{T} teacher tokens. Relation distillation must align both query rows and their column support. We give a byte-overlap reference construction for the interface in Section 3.5. The available run records do not establish that the reported Qwen3-4B → Llama-3.2-3B comparison uses this specific construction.

C.1 Query Alignment and Probability Transport

Query rows. Assign positive-length, half-open UTF-8 byte spans to both tokenizations of the same decoded response. Let $\mathcal{A} \subseteq [T_{\mathrm{S}}]$ contain student rows whose right boundary matches a teacher row $a(i)$. These are states *after consuming* the tokens, not predictive states. Prompt templates and response-start offsets determine the correspondence; equal token indices do not establish equal response boundaries.

Relation columns. Treat each row distribution as a column vector in this subsection. For each eligible query, a nonnegative row-stochastic matrix transports teacher relation mass into the student's response-position simplex:

$$\begin{aligned} \boldsymbol{\Pi}_i \in \mathbb{R}_{\geq 0}^{T_{\mathrm{T}} \times T_{\mathrm{S}}}, \qquad & \boldsymbol{\Pi}_i \mathbf{1}_{T_{\mathrm{S}}} = \mathbf{1}_{T_{\mathrm{T}}}, \\ \Phi_i(\mathbf{u}) \coloneqq \boldsymbol{\Pi}_i^{\top} \mathbf{u}, \qquad & \Phi_i : \Delta^{T_{\mathrm{T}}-1} \longrightarrow \Delta^{T_{\mathrm{S}}-1}. \end{aligned} \tag{30}$$

One concrete realization uses teacher spans I_j^{T} and student spans I_k^{S} that both partition the same retained response bytes:

$$(\Pi_i)_{jk} = \frac{|I_j^{\mathrm{T}} \cap I_k^{\mathrm{S}}|}{|I_j^{\mathrm{T}}|}, \qquad |I_j^{\mathrm{T}}| > 0. \tag{31}$$

Here $|\cdot|$ counts bytes. Complete coverage makes each row sum to one. This construction shares the column map across queries; i permits a general row-dependent map. Query matching and one-to-many column transport are distinct, and columns cover the completed response as in the base method.

The transported target in equation 13 is nonnegative and normalized, since $\mathbf{1}^{\top} \boldsymbol{\Pi}_i^{\top} \mathbf{u} = (\boldsymbol{\Pi}_i \mathbf{1})^{\top} \mathbf{u} = 1$ for every teacher probability vector $\mathbf{u}$. For a nonempty aligned set, the corresponding per-response FKL is

$$\ell_{\mathrm{rel}}^{\mathrm{XT}} = \frac{1}{|\mathcal{C}|H|\mathcal{A}|} \sum_{c \in \mathcal{C}} \sum_{h=1}^{H} \sum_{i \in \mathcal{A}} D_{\mathrm{KL}}\left(\widetilde{\mathbf{q}}_{c,h,i} \,\|\, \mathbf{r}_{c,h}^{\mathrm{S}}(i)\right). \tag{32}$$

This expression is defined for nonempty $\mathcal{A}$; empty-alignment handling and microbatch weighting are separate implementation choices. The teacher, sampled response, and alignment map are detached; gradients update only the student. For identical tokenizations, $a(i) = i$, $\mathcal{A} = [T]$, and $\boldsymbol{\Pi}_i = \mathbf{I}$ recover the base relation loss exactly.

Coverage and failure handling. Unmatched query boundaries are omitted from the row mean, not from other rows' column support. Non-round-tripping text, ambiguous special tokens, zero-length spans, or uncovered bytes require rejection or an explicit handling rule. Query coverage is

$$C_{\mathrm{query}} = \frac{\sum_b |\mathcal{A}_b|}{\sum_b T_{\mathrm{S},b}}. \tag{33}$$

Rejected responses contribute zero aligned rows and remain in the denominator. Dropping columns and renormalizing instead requires reporting retained mass *before* renormalization: the resulting conditional target differs from full-coverage transport.

C.2 Support, Output Alignment, and Claim Boundary

FKL permits zero transported target mass using $0 \log(0/p) = 0$. For a general transport with zero columns, relation RKL would require a full-support target; an analytical option is

$$\widetilde{\mathbf{q}}_{c,h,i}^{(\varepsilon)} = (1-\varepsilon)\widetilde{\mathbf{q}}_{c,h,i} + \frac{\varepsilon}{T_{\mathrm{S}}} \mathbf{1}_{T_{\mathrm{S}}}, \qquad 0 < \varepsilon < 1. \tag{34}$$

Smoothing and relation RKL are analytical alternatives, not evaluated configurations. Under full-coverage positive-length byte partitions and strictly positive teacher rows, every student column receives positive mass already, so smoothing is not required for support.

Relation transport aligns *positions*, not output vocabulary IDs. The joint objective, when a valid output alignment is available, is

$$\mathcal{L}_{\mathrm{M}}^{\mathrm{XT}} = \mathcal{L}_{\mathrm{out}}^{\mathrm{aligned}} + \lambda\, \mathbb{E}_{\mathrm{on}}[\ell_{\mathrm{rel}}^{\mathrm{XT}}]. \tag{35}$$

The output term needs its own common support; raw teacher IDs cannot index a different student vocabulary. SimpleOPD (He et al., 2026) requires both start and end offsets to match. Unmatched tokens receive zero teacher-derived advantage but may retain student-reference regularization. Its output coverage therefore differs from C_{query}. Output alignment, reference KL, termination masking, and policy updates are distinct factors in a recipe comparison; relation-only training is not equivalent to the joint objective above.

"Bridge-free" concerns learned feature-coordinate projectors, not vocabulary or position alignment. Proposition 1 applies to directly aligned, unit-diagonal cosine relations; arbitrary transport need not preserve that diagonal or permit recovery of the original teacher Gram matrix. The shared-position temperature expansion likewise does not transfer unchanged. For a fixed row-stochastic map,

$$\lim_{\tau\to\infty} \widetilde{\mathbf{q}}_\tau = \mathbf{\Pi}^\top \frac{\mathbf{1}_{T_{\mathrm{T}}}}{T_{\mathrm{T}}}, \qquad \lim_{\tau\to\infty} \mathbf{p}_\tau = \frac{\mathbf{1}_{T_{\mathrm{S}}}}{T_{\mathrm{S}}}. \tag{36}$$

These limits need not coincide: segmentation alone can induce a nonuniform transported baseline, distinct from task-relevant relation structure.

C.3 Evaluation Protocol and Scope

Table 3 reports Avg@5/pass@5 for AIME and IFEval, Avg for MATH-500, Avg@1 for AdvanceIF, and Score for MultiChallenge. The latter three have no pass metric. The available records do not fully specify model/tokenizer revisions, extraction layers, alignment coverage, stabilization, seeds, or compute. IFEval's strict/loose and prompt/instruction variant, AdvanceIF's scorer, and MultiChallenge's judge/history aggregation (Sirdeshmukh et al., 2025) are also unspecified. Consequently, these scores do not isolate the relation objective from all implementation differences. IFEval pass@5 denotes binary response success, not repeated judging; MATH-500's sample count is unspecified.

Interpretation of AIME aggregates. Under a single 30-problem, five-completion binary evaluation, Avg@5 varies in $100/150$-point increments and direct pass@5 in $100/30$-point increments. AIME 2025 Avg@5 values 1.0 (OPRD-Bridge) and 4.3 (Mini-OPD), and Mini-OPD's AIME 2025/2026 pass@5 values 8.7/7.3, cannot arise under that specific protocol. The reported aggregates are retained, but the available records do not specify their denominators, repeats, and estimators; they do not identify a single-repeat 30×5 evaluation.

Table 4: Polaris 16K mathematics layer sweep at step 100 (%). AIME denotes AIME 2024, and Olym. abbreviates OlympiadBench. AIME uses avg@32, AMC uses avg@4, and Olym. uses avg@1. Training and evaluation response caps are 16K and 64K, respectively. All Mini-OPD rows use $\lambda = 10^4$. Bold scores mark the best Mini-OPD layer; shaded cells identify the layer-20 configuration reported in Table 1. Avg. is the unweighted mean of the three benchmarks and is omitted when a benchmark is unavailable. "–" denotes an unavailable result.

Method	AIME 2024	AMC	Olym.	Avg.
Polaris-4B-Preview (teacher)	77.40	93.67	66.22	79.10
OPD	74.27	92.47	65.33	77.36
OPRD (1–35)	75.52	92.17	64.30	77.33

Layers 1–16					**Layers 17–32**				
Layer	AIME	AMC	Olym.	Avg.	Layer	AIME	AMC	Olym.	Avg.
1	75.62	92.47	64.30	77.46	17	76.15	91.57	65.63	77.78
2	75.42	92.17	65.33	77.64	18	74.17	90.96	**66.37**	77.17
3	**76.88**	91.57	64.44	77.63	19	74.90	92.47	64.59	77.32
4	74.69	90.06	–	–	20 (main)	76.35	**93.67**	**66.37**	**78.80**
5	74.69	90.06	–	–	21	76.04	91.57	64.74	77.45
6	75.73	90.06	65.48	77.09	22	75.42	90.96	65.33	77.24
7	75.42	92.47	64.74	77.54	23	75.83	**93.67**	65.04	78.18
8	75.00	90.66	64.89	76.85	24	73.96	92.17	65.19	77.11
9	75.31	90.66	64.44	76.80	25	76.25	90.96	64.74	77.32
10	75.10	91.87	64.30	77.09	26	76.15	92.77	65.19	78.04
11	76.35	92.17	65.04	77.85	27	76.04	91.87	65.19	77.70
12	76.04	91.87	65.33	77.75	28	**76.88**	91.87	64.59	77.78
13	74.58	92.17	65.33	77.36	29	76.15	92.47	65.33	77.98
14	74.58	92.47	65.19	77.41	30	75.73	91.27	63.85	76.95
15	76.46	92.47	64.30	77.74	31	75.52	91.87	64.59	77.33
16	75.00	92.47	65.04	77.50	32	76.04	91.87	65.33	77.75

D POLARIS 16K MATHEMATICS LAYER SWEEP

Table 4 reports the complete layer sweep underlying the mathematics comparison in the main paper. All student checkpoints are evaluated at step 100 under the same 64K-response protocol. The two missing OlympiadBench entries correspond to runs for which no metric file was produced.

Ranges and task-dependent rankings. The 32 placements span 73.96–76.88 on AIME, 90.06–93.67 on AMC, and 63.85–66.37 on OlympiadBench among 30 complete variants. Layers 3/28 lead AIME, 20/23 lead AMC, and 18/20 lead OlympiadBench. Layer 20 has the best complete Mini-OPD mean (78.80 versus 77.36 for OPD), while complete means span 76.80–78.80. The improvement is broad on AIME (30/32 layers exceed OPD) but sparse on AMC (3/32) and OlympiadBench (4/30); 18 of 30 complete means exceed OPD. These finite-set maxima are descriptive and selection-biased, not an independently validated layer-selection rule.

E Extended Related Work

Output supervision on student trajectories. Classical KD and SeqKD transfer soft targets and teacher-generated sequences (Hinton et al., 2015; Kim & Rush, 2016), with later extensions to sequence divergences, reasoning traces, and teacher fusion (Wen et al., 2023; Ho et al., 2023; DeepSeek-AI, 2025; Wan et al., 2024). GKD and MiniLLM instead supervise student rollouts (Agarwal et al., 2024; Gu et al., 2024). Subsequent methods refine divergences and replay (Ko et al., 2024), adaptive teachers (Zhang et al., 2025), generation switching (Peng et al., 2025), teacher routing (Ma et al., 2026), or black-box feedback (Ye et al., 2025). Recent work examines support and entropy along student prefixes (Li et al., 2026a), adds FKL at high-teacher-entropy positions (Jin et al., 2026), or scores candidate tokens through subsequent teacher confidence (Liu et al., 2026). The gradient-alignment study of Armandpour et al. (2026) is an offline diagnostic, not evidence of improved training. Mini-OPD adds an internal target without claiming to resolve these output-level failure modes.

Cross-tokenizer interfaces. SimpleOPD (He et al., 2026) matches identical response-text spans under the teacher's native template, with student-reference KL and termination-token advantage masking. SimCT (Sun et al., 2026) adds minimal jointly tokenizable continuations to shared tokens, normalizing length-averaged autoregressive scores over this candidate set. Its preference distribution is not an exact marginal of native next-token probabilities. Both concern output supervision; Mini-OPD transports internal response-position relations. Output alignment, relation coverage, and stabilization remain distinct comparison factors.

Internal supervision. TinyBERT matches hidden states and attention maps (Jiao et al., 2020); MiniLMv2 transfers dimension-scaled Q–Q/K–K/V–V relations on offline inputs (Wang et al., 2021). OPRD aligns on-policy hidden states; OPRD-Bridge fits projectors with frozen backbones and freezes them during distillation (Yang et al., 2026). PHF matches normalized hidden-state transitions, within-layer Gram geometry, and adjacent-layer relations alongside output JSD, using a reference-conditioned EMA self-teacher (Li et al., 2026b). Mini-OPD instead combines response-only cosine QKV relations at one layer pair with external-teacher output OPD, without learned projectors or privileged reference conditioning. Neither relation heads nor internal on-policy supervision alone is claimed as new.