VIEW-DECOMPOSED ON-POLICY DISTILLATION FOR ROBUST REASONING IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation reduces training–inference mismatch by supervising large language models (LLMs) on their own generated trajectories. For robust reasoning, this supervision should preserve useful teacher uncertainty while limiting the transfer of prompt-sensitive behavior. We investigate how consistency across task-preserving prompt variations can guide distillation toward more accurate reasoning with less sensitivity to prompt wording. We introduce View-Decomposed On-Policy Distillation (VD-OPD), which compares a frozen teacher's predictions for the same student-generated prefix under two prompt views. Their shared probability mass defines both the learning target and its weight, retaining alternatives supported by both views while attenuating disagreement. VD-OPD uses no answer labels during distillation and adds no inference cost. On six mathematics benchmarks, it raises Qwen3-1.7B-Base's macro-average Avg@8 from 34.8% to 36.7% over EOPD (full), with better unseen-prompt robustness. Gains remain 1.4/1.2 points under two-view/total-FLOPs controls. Cross-view agreement determines VD-OPD's target and weight. Our analysis explains when pooling improves the target and why shared teacher errors can persist.

1 INTRODUCTION

On-policy distillation (OPD) trains a large language model (the student) using feedback from a stronger model (the teacher) (Agarwal et al., 2024). During training, the student generates its own responses, and the teacher guides how it should continue. This reduces training–inference mismatch by teaching the student along the reasoning paths it actually follows.

The quality of this feedback depends on which aspects of the teacher's prediction the student learns. Reverse-KL distillation emphasizes high-probability teacher continuations (Gu et al., 2024), while EOPD adds forward-KL supervision at high-entropy states to preserve broader distributional coverage (Jin et al., 2026). Preserving these alternatives raises a further question: how stable are they under changes in the prompt? A teacher may consistently support several plausible continuations, or confidently favor different continuations under incidental changes in prompt wording. Given the documented sensitivity of language models to prompt formatting (Sclar et al., 2024), suppressing uncertainty can discard useful information, whereas reproducing it indiscriminately can transmit prompt-specific behavior.

Other methods refine teacher–student matching through vocabulary-wise divergence mixing in ToDi (Jung et al., 2025) or an intermediate target in Veto (Jang et al., 2026). Teacher stability across task presentations is a separate supervision property. CROP uses prompt interventions to select tokens by relevance to task conditions (Li et al., 2026), leaving the reconciliation of competing continuations open. The remaining challenge is to preserve useful distributional breadth while limiting the transfer of incidental prompt sensitivity, without requiring answer verification in the distillation update.

We ask whether OPD can support robust LLM reasoning by reducing sensitivity to prompt wording. We propose *View-Decomposed On-Policy Distillation* (VD-OPD; Figure 1), which evaluates a frozen teacher under two task-preserving prompt views at the same student-generated prefix. It separates shared support from disagreement through an unnormalized geometric consensus: the target's shape specifies what to imitate, and its retained mass determines how strongly to imitate it. The resulting objective preserves the teacher distribution when views agree and attenuates supervision when they disagree; inference uses only the student.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

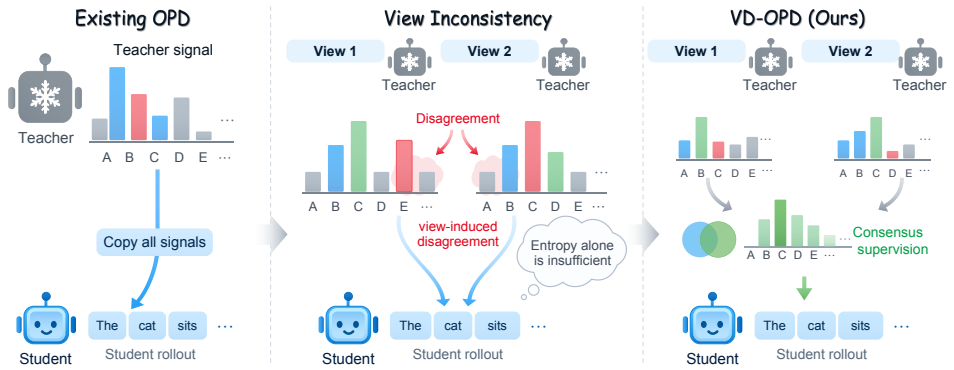

Figure 1: **From teacher imitation to consensus supervision.** Left: matching a single teacher view can transfer prompt-sensitive predictions. Middle: task-preserving prompt views can disagree at the same student prefix, which entropy alone does not reveal. Right: VD-OPD learns what both views support and reduces supervision when they disagree.

Our contributions are threefold:

- **An insight into supervision reliability.** A study of 5,000 paired teacher predictions shows that entropy does not determine consistency across task-preserving views. This distinguishes distributional breadth from sensitivity to how a task is presented.
- **Consensus as a supervision principle.** VD-OPD jointly determines the supervision target and its weight. Our analysis characterizes exactly when its normalized target improves upon the original teacher at fixed prefixes and why shared errors remain.
- **Improved reasoning accuracy and prompt robustness.** Across three student backbones in two model families, VD-OPD improves over EOPD (full). For Qwen3-1.7B-Base, six-benchmark macro-average Avg@8 rises from 34.8% to 36.7%, and unseen-wrapper worst-view accuracy from 28.9% to 30.2%. On the same student, gains persist against two-view EOPD and after matching total training FLOPs. Controlled ablations support the value of both the consensus target and its state-dependent weight.

2 Related Work

On-policy knowledge distillation. Sequence-level KD trains on teacher-generated outputs (Kim & Rush, 2016). GKD matches distributions on student trajectories to reduce training–inference mismatch (Agarwal et al., 2024), while MiniLLM optimizes sequence-level reverse KL (Gu et al., 2024). Subsequent methods address the teacher–student gap through interleaved sampling in SKD (Xu et al., 2025) and contrasting teacher- and student-generated responses in DistiLLM-2 (Ko et al., 2025). These methods focus on trajectory collection and distribution matching. Teacher stability across task-preserving views offers a complementary basis for supervision. VD-OPD retains student rollouts and refines their supervision through paired teacher evaluations.

Adaptive supervision. Discrepancy-based methods adjust correction through vocabulary-wise divergence mixing in ToDi (Jung et al., 2025) or token ranking and entropy-gap-gated matching in TSD-KD (Kim & Baek, 2026). EOPD supplements reverse KL with forward KL at high-entropy teacher states to preserve distributional coverage (Jin et al., 2026). Teacher entropy and teacher–student discrepancy characterize single-prompt predictions; cross-view consistency measures stability across prompts. Outcome-based TGOPD verifies teacher responses before routing prompts to OPD or GRPO (Zhang et al., 2026). VD-OPD uses cross-view overlap to weight supervision without answer verification in the distillation update; shared teacher errors can persist across views.

View consistency and geometric consensus. Prompt-format sensitivity motivates comparing predictions across meaning-preserving inputs (Sclar et al., 2024). CROP rescores fixed student rollouts under paraphrases and counterfactuals to select task-relevant tokens (Li et al., 2026); CoDA constructs privileged self-teaching context from answer-level consensus, with a separate penalty for

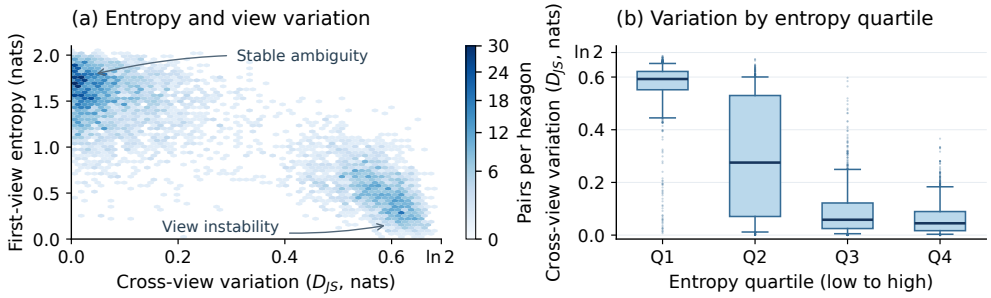

Figure 2: Teacher entropy and cross-view variation across 5,000 paired predictions. (a) Entropy–variation landscape, colored by pairs per hexagon. (b) JS divergence within entropy quartiles. Boxes span the interquartile range, lines indicate medians, whiskers mark the 5th–95th percentile range, and dots show values beyond the whiskers.

minority trajectories (Guo et al., 2026). Geometric aggregation underlies logarithmic opinion pooling (Heskes, 1997), while Veto uses a normalized geometric bridge between teacher and student (Jang et al., 2026). VD-OPD combines two views of the same frozen teacher at identical student prefixes: one unnormalized geometric target specifies what to imitate, and its total mass determines supervision strength.

3 Preliminary Study

Motivation. Entropy-aware on-policy distillation uses forward KL at high-entropy teacher states to preserve distributional coverage (Jin et al., 2026). Entropy characterizes a single distribution, whereas view consistency concerns the relationship between distributions. We examine this distinction using teacher predictions at student-generated prefixes.

Experimental setup. A frozen teacher scores prefixes generated by a fixed student checkpoint under two instruction wrappers, keeping the task content unchanged. We analyze 5,000 paired predictions. For each pair $p^{(1)}, p^{(2)}$, we measure first-view entropy $H(p^{(1)}) = -\sum_a p^{(1)}(a) \log p^{(1)}(a)$ and cross-view Jensen–Shannon (JS) divergence:

$$D_{\mathrm{JS}}(p^{(1)}, p^{(2)}) = \frac{1}{2} D_{\mathrm{KL}}(p^{(1)} \| m) + \frac{1}{2} D_{\mathrm{KL}}(p^{(2)} \| m), \qquad m = \frac{p^{(1)} + p^{(2)}}{2}. \tag{1}$$

The KL divergence is $D_{\mathrm{KL}}(p\|m) = \sum_a p(a) \log[p(a)/m(a)]$, summing over categories a. Both statistics use natural logarithms and are measured in nats; $D_{\mathrm{JS}} \in [0, \ln 2]$. We partition the pairs into four first-view entropy quartiles with 1,250 pairs each.

Observation 1: Entropy does not determine view consistency. Figure 2(a) exhibits high-entropy distributions with low cross-view divergence alongside low-entropy distributions that concentrate probability on conflicting candidates. Thus high entropy does not imply view instability, and low entropy does not imply agreement.

Observation 2: Variation persists within entropy strata. In Figure 2(b), the second quartile covers entropies of approximately 0.652–1.300 nats, yet its 5th and 95th percentiles of JS divergence are 0.012 and 0.600 nats. Entropy stratification leaves markedly different degrees of agreement within the same group. This motivates preserving coverage when views agree and attenuating supervision when they disagree. Section 4 implements this distinction with geometric consensus; entropy and JS divergence serve as diagnostics for this study.

4 Method

Overview. View-Decomposed On-Policy Distillation (VD-OPD) uses agreement between two teacher views to determine what the student learns and how strongly it is supervised. At each student-generated prefix, we form an unnormalized geometric consensus: its shape defines the target, and its total mass sets the loss weight. Figure 3 shows the training procedure.

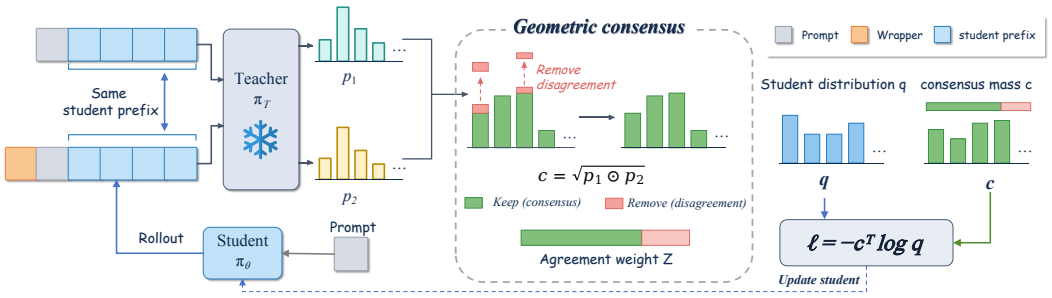

Figure 3: VD-OPD evaluates both teacher views at the same student prefix. The indices i, t identify a response and a position within it; N_k counts valid response positions in iteration k. Square roots, products $\odot$, and logarithms act elementwise. Teacher predictions are detached throughout the student update.

4.1 Teacher views at fixed student prefixes

Following on-policy distillation, we supervise the student on its own generated prefixes (Agarwal et al., 2024). We score each prefix under two teacher prompt views, so their difference captures prompt sensitivity at the same response state (Li et al., 2026).

Let π_T be the frozen teacher and π_θ the student, with a shared tokenizer and vocabulary $\mathcal{A}$ of size d. At iteration k, we sample prompts $x_i \sim \mathcal{D}$ from the training distribution and generate responses $y_i \sim \pi_{\theta_k}(\cdot \mid x_i)$ using the pre-update parameters θ_k. Each response has two prompt views: the original $x_i^{(1)} = x_i$ and $x_i^{(2)} = T_i(x_i)$, formed by uniformly sampling a valid wrapper. Both teacher views condition on the same generated token IDs:

$$\begin{aligned} p_{i,t}^{(m)}(a) &= \pi_T(a \mid x_i^{(m)}, y_{i,<t}), \qquad m \in \{1,2\}, \\ q_{i,t}(a;\theta) &= \pi_\theta(a \mid x_i, y_{i,<t}), \qquad a \in \mathcal{A}. \end{aligned} \tag{2}$$

Here t indexes a position in response i, and $y_{i,<t}$ is its preceding token sequence. The student always uses the original prompt. We denote the vocabulary probability vectors by $\mathbf{p}_{i,t}^{(m)}, \mathbf{q}_{i,t}(\theta) \in \Delta^{d-1}$, where Δ^{d-1} is the probability simplex. All logarithms are natural.

The wrapper set contains two generic solving instructions and one layout heading. We prepend the sampled wrapper to the final user message, preserving all original content, message roles, and the assistant generation prefix. The wrapper stays fixed throughout the response. We check prompt lengths before generation and use the original prompt for both views if no alternative fits. Appendix A.1 gives the exact strings, validity checks, and token alignment. The distillation update uses full-vocabulary teacher probabilities and no answer labels.

4.2 Geometric consensus through view decomposition

Two views that agree on $(1/2, 1/2)$ have the same arithmetic mean as deterministic views selecting opposite tokens. We distinguish these cases in square-root probability space. At a fixed response position, abbreviate $\mathbf{p}_{i,t}^{(m)}$ as $\mathbf{p}_m$ and apply

$$[\mathbf{u}_+ \ \mathbf{u}_-] = [\sqrt{\mathbf{p}_1} \ \sqrt{\mathbf{p}_2}]\mathbf{R}, \qquad \mathbf{R} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{3}$$

The orthogonal matrix $\mathbf{R} \in \mathbb{R}^{2\times 2}$ produces symmetric and differential coordinates $\mathbf{u}_+, \mathbf{u}_- \in \mathbb{R}^d$. Subtracting the squared differential coordinates from the arithmetic mean $\bar{\mathbf{p}} = (\mathbf{p}_1 + \mathbf{p}_2)/2$ gives the *geometric consensus mass*

$$\boxed{\mathbf{c} = \bar{\mathbf{p}} - \mathbf{u}_-^{\odot 2} = \sqrt{\mathbf{p}_1 \odot \mathbf{p}_2}.} \tag{4}$$

Here $\mathbf{u}_-^{\odot 2}$ denotes elementwise squaring. The target satisfies $0 \le c(a) \le \bar{p}(a)$ for every token a, and its total mass is

$$Z = \mathbf{1}^\top \mathbf{c} = \langle \sqrt{\mathbf{p}_1}, \sqrt{\mathbf{p}_2} \rangle = 1 - \|\mathbf{u}_-\|_2^2 \in [0,1], \tag{5}$$

where $\mathbf{1} \in \mathbb{R}^d$ is the all-ones vector. Thus Z is the Bhattacharyya affinity, and $1 - Z$ is the squared Hellinger distance, defined as $\frac{1}{2}\|\sqrt{\mathbf{p}_1} - \sqrt{\mathbf{p}_2}\|_2^2$. Identical views give $\mathbf{c} = \mathbf{p}_1$ and $Z = 1$, preserving the full teacher distribution, including multiple supported continuations. Disjoint supports give zero mass. Appendix D.1 provides derivations and an example.

The target also minimizes the average generalized KL divergence to the two views. For $\mathbf{v} \in \mathbb{R}^d_{\geq 0}$ and $\mathbf{w} \in \mathbb{R}^d_{>0}$, define $D_{\mathrm{I}}(\mathbf{v}\|\mathbf{w}) = \sum_a [v(a)\log(v(a)/w(a)) - v(a) + w(a)]$, with $0\log(0/w(a)) = 0$. For strictly positive teacher probabilities,

$$\mathbf{c} = \operatorname*{arg\,min}_{\mathbf{v} \in \mathbb{R}^d_{\geq 0}} \frac{1}{2} \sum_{m=1}^{2} D_{\mathrm{I}}(\mathbf{v}\|\mathbf{p}_m). \tag{6}$$

Leaving the total mass unconstrained determines the target shape and Z. Normalizing the solution recovers the logarithmic opinion pool (Bandyopadhyay & Chung, 2014). We use the closed-form geometric mean in training; Appendix D.2 proves uniqueness and the associated divergence identities.

4.3 On-policy distillation of consensus mass

We train the student directly on the unnormalized target, retaining its total mass as the supervision weight. Let $\mathcal{T}_k$ be the valid response-position pairs (i, t) in iteration k, with $N_k = |\mathcal{T}_k| > 0$. We exclude prompt tokens and padding and skip empty batches. Holding sampled responses and teacher predictions fixed, we minimize

$$\mathcal{L}_k(\theta) = -\frac{1}{N_k} \sum_{(i,t) \in \mathcal{T}_k} \mathbf{c}_{i,t}^{\top} \log \mathbf{q}_{i,t}(\theta). \tag{7}$$

Because N_k counts positions before weighting, each state contributes proportionally to its retained mass.

At a position with $Z > 0$, define $\mathbf{r} = \mathbf{c}/Z$ and write $\mathbf{q} = \mathrm{softmax}(\mathbf{z})$, where $\mathbf{z}$ denotes the student logits. The loss $\ell(\theta) = -\mathbf{c}^{\top} \log \mathbf{q}$ and its gradient satisfy

$$\ell(\theta) = Z\big[D_{\mathrm{KL}}(\mathbf{r}\|\mathbf{q}) + H(\mathbf{r})\big], \qquad \nabla_{\mathbf{z}} \ell = Z\mathbf{q} - \mathbf{c}. \tag{8}$$

Here $D_{\mathrm{KL}}(\mathbf{r}\|\mathbf{q}) = \sum_a r(a)\log[r(a)/q(a)]$ and $H(\mathbf{r}) = -\sum_a r(a)\log r(a)$. Since $H(\mathbf{r})$ is constant with respect to the student, the update is equivalent to forward-KL distillation toward $\mathbf{r}$ with weight Z. At $Z = 0$, Equation 7 gives zero loss without normalizing $\mathbf{c}$. This weight measures consistency across the selected views; errors shared by both views remain.

Each iteration accumulates the gradient of Equation 7 at θ_k, takes one optimizer step, and collects a fresh student rollout. Gradients pass only through student predictions, with sampled prefixes fixed. Training requires two teacher scoring passes per response; inference uses only the student. Appendix D.3 gives local curvature and gradient bounds, and Appendix A details the algorithm and exact blockwise evaluation.

4.4 Robustness of the consensus target

We analyze the normalized target at a fixed student-generated prefix. Let $p_0 = \mathbf{p}_1 = \mathrm{softmax}(z_0)$ and $p_T = \mathbf{p}_2 = \mathrm{softmax}(z_0 + \delta)$ be the teacher predictions under the original and wrapped prompts. Assume both views share a strictly positive ideal distribution $p^\star$, and define the target error $\mathcal{E}(q) = D_{\mathrm{KL}}(p^\star\|q)$. All logits are finite; expectations are over wrapper selection. The centered shift $\delta = P(\log p_T - \log p_0)$, with $P = I_d - \mathbf{1}\mathbf{1}^{\mathsf{T}}/d$, removes additive logit constants that leave predictions unchanged.

Theorem 1 (Exact improvement criterion)**.** *Assume $\mathbb{E}\|\delta\|_2^2 < \infty$ and let $\mu = \mathbb{E}\delta$. Define the mean correction $\mathcal{G} = \frac{1}{2}\langle p^\star - p_0, \mu\rangle$ and the curvature cost $\mathcal{C} = \mathbb{E}D_{\mathrm{KL}}(p_0\|r)$. Then $r = \mathrm{softmax}(z_0 + \delta/2)$ and $\mathbb{E}\mathcal{E}(r) - \mathcal{E}(p_0) = \mathcal{C} - \mathcal{G}$. Thus pooling strictly improves expected forward KL if and only if $\mathcal{G} > \mathcal{C}$.*

Proof sketch. The normalized geometric pool is the softmax of the average teacher logits. Expanding its log-partition function at the original logits separates a linear correction from a nonnegative KL remainder. Taking expectations gives the stated identity; Appendix D.7 provides the full proof. □

Pooling improves the target when mean correction exceeds curvature cost, but agreement alone cannot remove shared errors. Appendix D.8 bounds this cost; Appendices D.7 and D.9 analyze centered wrapper noise and unidentifiable shared bias. Theory concerns fixed-prefix target quality; experiments assess Z's training benefit.

5 EXPERIMENTS

We examine **RQ1**, reasoning gains across benchmarks, backbones, and compute controls (Section 5.2); **RQ2**, unseen-prompt robustness and cross-domain transfer (Section 5.3); **RQ3**, contributions of consensus targets and state-specific weights (Section 5.4); and **RQ4**, the relationship between view disagreement and local supervision utility (Section 5.5).

5.1 EXPERIMENTAL SETUP

Benchmarks and metrics. Following EOPD (Jin et al., 2026), we report Avg@8 and Pass@8 on MATH500, AMC23, AIME24, AIME25, Minerva Math, and OlympiadBench. Additional evaluations cover unseen-prompt robustness, cross-domain accuracy on GPQA-Diamond and MMLU-Pro, and GSM8K performance. Appendix B.2 specifies dataset subsets, metrics, and decoding.

Comparison protocol. We compare with GKD (FKL/RKL/JSD) (Agarwal et al., 2024), ToDi (Jung et al., 2025), EOPD (Jin et al., 2026), TSD-KD (Kim & Baek, 2026), and Veto (Jang et al., 2026). Core comparisons match teacher, student initialization, training data, evaluation, and rollout/update budgets. Separate controls match two-view teacher access, total FLOPs, or teacher-scoring FLOPs (Appendix B.5.3).

Backbones and implementation. Qwen3-1.7B-Base and Qwen3-4B-Base students use Qwen3-8B (Yang et al., 2025); Llama-3.2-3B-Instruct uses Llama-3.1-8B-Instruct (Meta, 2024). Teachers are frozen. Core runs share the MATH training split and use three seeds. Hardware and implementation details are in Appendix B.

5.2 RQ1: REASONING PERFORMANCE AND COMPUTE CONTROLS

Controlled mathematical reasoning. VD-OPD ranks first among distilled students on all six benchmarks for Qwen3-1.7B-Base (Table 1), exceeding the strongest baseline, EOPD (full) (Jin et al., 2026), by 1.3–2.4 percentage points in zero-shot Avg@8. The six-benchmark mean rises from 34.8% to 36.7% (from displayed rounded entries), with Pass@8 gains of 1.2–2.8 points (Table 6). Frozen teachers are excluded from this ranking.

Backbones and compute controls. On MATH500 (Figure 4), VD-OPD exceeds EOPD (full) in Avg@8 by 1.8, 1.8, and 1.6 points for Qwen3-1.7B-Base, Qwen3-4B-Base, and Llama-3.2-3B-Instruct, respectively; Pass@8 gains are 1.2, 0.8, and 1.2 points. Across all six benchmarks, Avg@8 gains are 1.4–2.2 points for Qwen3-4B-Base and 1.3–3.3 for Llama-3.2-3B-Instruct (Tables 7 and 8). For Qwen3-1.7B-Base, EOPD-2V averages the EOPD (full) objective (Jin et al., 2026) over two teacher views, matching teacher-query, rollout, and update counts (Table 2). VD-OPD gains 1.4/1.5 points in six-math Avg@8/Pass@8; shortening training to EOPD (full)'s 4.7×10^{18} total-FLOPs budget retains gains of 1.2/1.5 points (VD-OPD-Budget; Appendix B.5.3). At four matched teacher-scoring budgets (actual mean FLOPs differ by less than 0.2% per pair), VD-OPD leads by 0.8–1.3 points in six-math Avg@8 and 0.8–1.2 on MATH500, including when EOPD (full) receives twice its default budget (Figure 5).

Complementary training protocols. Within the protocol blocks of Table 11, VD-OPD reaches 67.8/88.7 MATH500 Avg@8/Pass@8 versus DistiLLM-2's 64.2/86.0 (Ko et al., 2025), and 66.5/87.5 versus CROP's 65.0/86.8 (Li et al., 2026) on the common training subset. GRPO is an answer-supervised reference; resource accounting covers local training only.

> **Takeaway (RQ1).** Gains across backbones and under matched two-view access and compute support consensus construction as a source of improvement.

Table 1: **Mathematical reasoning under the common protocol.** Qwen3-8B → Qwen3-1.7B-Base; zero-shot Avg@8 (%, higher is better). All trained students use the same MATH split and matched rollout/update budget. Olymp. denotes the English text-only mathematics subset of OlympiadBench. Mean is the unweighted average of the six scores. Bold and underline mark the best and second-best distilled-student scores; the frozen teacher and initial student are references.

Category	Method	MATH500	AMC23	AIME24	AIME25	Minerva	Olymp.	Mean
Reference models	Frozen teacher	72.8	51.4	16.7	12.5	38.6	43.1	39.2
	Initial student	54.7	32.8	8.3	5.1	25.4	30.2	26.1
Standard distillation	GKD-FKL (Agarwal et al., 2024)	58.9	36.1	9.7	6.2	27.8	33.0	28.6
	GKD-RKL (full) (Agarwal et al., 2024)	59.6	36.8	10.1	6.5	28.4	33.7	29.2
	GKD-JSD (Agarwal et al., 2024)	60.1	37.5	10.4	6.9	28.9	34.2	29.7
	Sampled-token OPD (Jin et al., 2026)	62.4	39.8	11.7	7.8	30.7	36.0	31.4
Adaptive supervision	ToDi-OP (Jung et al., 2025)	63.1	40.5	12.4	8.3	31.5	36.8	32.1
	TSD-KD (Kim & Baek, 2026)	63.8	41.2	12.8	8.9	32.1	37.5	32.7
	EOPD (top-16) (Jin et al., 2026)	65.1	42.8	13.9	9.9	33.7	39.1	34.1
	EOPD (full) (Jin et al., 2026)	<u>66.0</u>	<u>43.6</u>	<u>14.4</u>	<u>10.4</u>	<u>34.5</u>	<u>40.0</u>	<u>34.8</u>
Target reformulation	Veto (Jang et al., 2026)	64.2	41.7	13.1	9.2	32.8	38.0	33.2
	VD-OPD (ours)	**67.8**	**45.9**	**15.8**	**11.7**	**36.9**	**42.1**	**36.7**

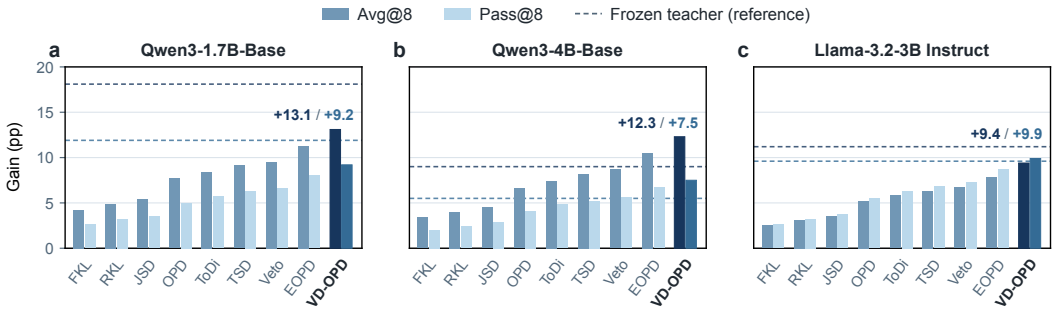

Figure 4: **Generalization across student sizes and model families on MATH500.** Bars show zero-shot gains over the initial student in percentage points; dashed lines show frozen-teacher gains in the corresponding metric color. Dark bars highlight VD-OPD, with gain annotations listing Avg@8 / Pass@8. FKL/RKL/JSD denote GKD variants, OPD denotes sampled-token OPD, and EOPD uses full-vocabulary FKL; method definitions follow Table 1. Absolute scores are in Tables 6–8.

5.3 RQ2: Robustness to prompts and transfer across tasks

Unseen prompt views. Across the six mathematics benchmarks, VD-OPD improves macro-average mean-view, worst-view, and worst-template accuracy over EOPD (full) by 0.9, 1.3, and 1.0 points (Table 3), with the largest gain on the least favorable views. Views are evaluated independently, without inference-time ensembling; Table 10 gives per-benchmark scores.

Transfer beyond the training tasks. The mathematics-trained student exceeds EOPD (full) by 1.5 points in greedy accuracy on GPQA-Diamond and MMLU-Pro (Table 3), by 1.2/0.5 points in GSM8K Avg@8/Pass@8, and by 1.2 points in MMLU-Pro-Math greedy accuracy (Table 9).

> **Takeaway (RQ2).** Consistency across training views benefits single-prompt evaluation, including unseen wrappers and tasks beyond the training domain.

5.4 RQ3: Contributions of target shape and weight assignment

Target shape and supervision weight. Table 4 crosses arithmetic $\bar{\mathbf{p}}_t = (\mathbf{p}_{t,1} + \mathbf{p}_{t,2})/2$ and geometric $\mathbf{r}_t = \mathbf{c}_t/Z_t$ targets with unit or Z_t weights. All six configurations share the two-view construction and loss $-N_k^{-1} \sum_t w_t \mathbf{s}_t^\top \log \mathbf{q}_t$, with normalized target $\mathbf{s}_t$, weight w_t, and unweighted denominator N_k.

Table 2: **Teacher-access and total-compute controls.** Qwen3-8B → Qwen3-1.7B-Base; zero-shot accuracy (%). Six-math macro-averages the six benchmarks. Views counts teacher scoring passes per response; costs are three-seed per-run means. Bold marks the better accuracy within each matched block.

				MATH500		Six-math	
Method	Views	FLOPs (10^{18})	GPU-h	Avg@8	Pass@8	Avg@8	Pass@8
Matched total FLOPs							
EOPD (full)	1	4.7	30.3	66.0	87.5	34.8	55.5
VD-OPD-Budget	2	4.7	30.5	**67.0**	**88.2**	**36.0**	**57.0**
Matched two-view access and rollout/update counts							
EOPD-2V	2	5.8	37.0	66.5	87.9	35.3	56.1
VD-OPD	2	5.8	37.2	**67.8**	**88.7**	**36.7**	**57.6**

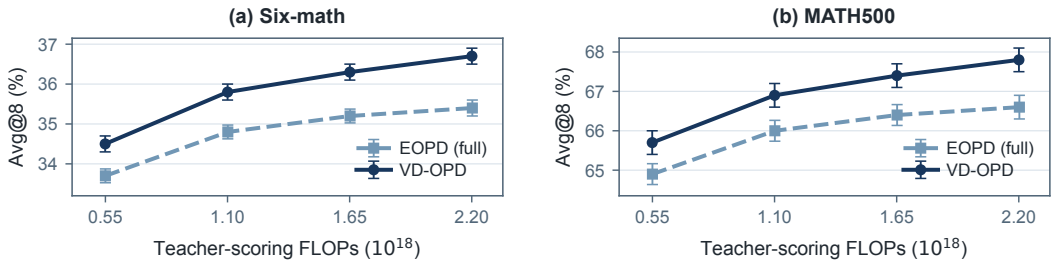

Figure 5: **Accuracy versus teacher-scoring computation.** Qwen3-8B → Qwen3-1.7B-Base; zero-shot Avg@8. Points and error bars show mean ±1 sample SD over seeds 11, 22, and 33. The horizontal axis measures teacher-only FLOPs, distinct from total FLOPs in Table 2.

Table 3: **Robustness and transfer for Qwen3-1.7B-Base.** Wrapper scores (%) are computed per benchmark and macro-averaged over the six mathematics benchmarks. Worst-view accuracy minimizes over wrappers per question; worst-template accuracy minimizes after averaging questions. GPQA-Diamond and MMLU-Pro use greedy accuracy with zero- and five-shot prompts, respectively. Bold and underline mark the best and second-best distilled-student scores per column; reference models are excluded.

		Unseen wrappers (math)			Cross-domain accuracy	
Category	**Method**	Mean view	Worst view	Worst template	GPQA Diamond	MMLU Pro
Reference models	Frozen teacher	35.8	32.4	33.1	42.1	51.7
	Initial student	25.9	21.8	22.7	34.5	41.8
Standard distillation	GKD-FKL (Agarwal et al., 2024)	28.1	23.6	24.5	35.9	43.0
	Sampled-token OPD (Jin et al., 2026)	30.0	25.8	26.7	36.9	44.2
Adaptive supervision	ToDi-OP (Jung et al., 2025)	30.9	26.7	27.8	37.5	44.8
	TSD-KD (Kim & Baek, 2026)	31.5	27.3	28.5	38.0	45.3
	EOPD (full) (Jin et al., 2026)	<u>32.8</u>	<u>28.9</u>	<u>30.1</u>	<u>39.1</u>	<u>46.7</u>
Target reformulation	Veto (Jang et al., 2026)	31.9	27.9	29.0	38.4	45.8
	VD-OPD (ours)	**33.7**	**30.2**	**31.1**	**40.6**	**48.2**

For the two weight controls, π is a seeded random permutation of valid response positions within a rollout batch, and $\bar{Z}_k = N_k^{-1} \sum_t Z_t$. Both retain each position's geometric target; shuffling preserves the weight multiset, and batch-mean weighting preserves total weight.

Geometric pooling raises unit-weight six-math Avg@8 from 35.0% to 36.1%. With Z_t weighting, geometric and arithmetic targets score 36.7% and 35.7%, respectively. VD-OPD exceeds shuffled and batch-mean weighting by 0.5 and 0.3 points in both MATH500 and six-math Avg@8.

Local update utility. Matched branches share a checkpoint, optimizer state, and frozen-checkpoint rollouts (Table 12). On disjoint questions, VD-OPD raises Avg@8 from 54.8% to 56.8% (+2.0

Table 4: **Target-shape and weight ablation.** Qwen3-1.7B-Base; three-seed means (%). Six-math: unweighted six-benchmark mean. Bold/underline: column best/second best; gray subscripts: differences from VD-OPD (percentage points).

Group	Supervision design			MATH500		Six-math mean	
	Configuration	Target	Weight	Avg@8	Pass@8	Avg@8	Pass@8
(a) Target-shape and weight combinations	Arithmetic	$\bar{\mathbf{p}}_t$	1	66.1$_{-1.7}$	87.8$_{-0.9}$	35.0$_{-1.7}$	56.1$_{-1.5}$
	Weighted arithmetic	$\bar{\mathbf{p}}_t$	Z_t	66.8$_{-1.0}$	88.1$_{-0.6}$	35.7$_{-1.0}$	56.8$_{-0.8}$
	Geometric	$\mathbf{r}_t$	1	67.2$_{-0.6}$	88.4$_{-0.3}$	36.1$_{-0.6}$	57.2$_{-0.4}$
	VD-OPD	$\mathbf{r}_t$	Z_t	**67.8**	**88.7**	**36.7**	**57.6**
(b) Weight-assignment controls	Shuffled weight	$\mathbf{r}_t$	$Z_{\pi(t)}$	67.3$_{-0.5}$	88.5$_{-0.2}$	36.2$_{-0.5}$	57.3$_{-0.3}$
	Batch-mean weight	$\mathbf{r}_t$	$\bar{Z}_k$	<u>67.5</u>$_{-0.3}$	<u>88.6</u>$_{-0.1}$	<u>36.4</u>$_{-0.3}$	<u>57.4</u>$_{-0.2}$

points), versus +1.5 for unit-weight geometric pooling, +1.6 for shuffled weights, and +1.8 for batch-mean weights, reproducing the full-training ranking.

Takeaway (RQ3). These point estimates support both target pooling and state-specific weighting beyond uniform loss scaling.

5.5 RQ4: View disagreement and local supervision utility

We relate consensus mass $Z_t = \mathbf{1}^\top \mathbf{c}_t$ to the change in student fit after a VD-OPD update, holding each pre-update prefix and normalized target $\mathbf{r}_t = \mathbf{c}_t / Z_t$ ($Z_t > 0$) fixed:

$$u_t = \ell_t^{\text{cons}}(\theta_{\text{before}}) - \ell_t^{\text{cons}}(\theta_{\text{after}}), \qquad \ell_t^{\text{cons}}(\theta) = -\mathbf{r}_t^\top \log \mathbf{q}_t(\theta). \tag{9}$$

Diagnostic setup. For Qwen3-8B → Qwen3-1.7B-Base, we use 250 reserved problems for branch updates and 250 source-disjoint problems for answer-level measurement (Appendix B.7). Positive u_t indicates improved consensus fit; negative u_t indicates degraded fit. We pool 24,576 valid positions (8,192 per seed; seeds 11, 22, and 33), each inheriting its sampled response's correctness label.

Findings. Figure 6 shows a negative-utility tail at low mass and a narrower, predominantly positive distribution at high mass. Median utility and token-weighted response correctness generally increase with Z_t, indicating less reliable consensus-fit improvements at low mass.

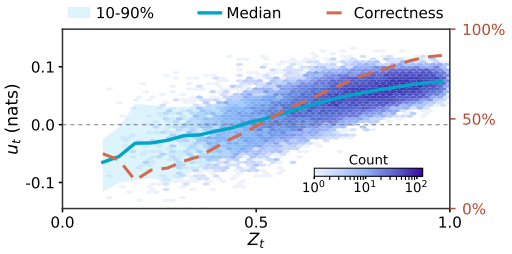

Figure 6: **Consensus mass and utility.** Hexagons: log token density. Solid/band: median and 10th–90th percentiles. Dashed: token-weighted response correctness. Binning and smoothing follow Appendix B.7.

Takeaway (RQ4). Average gains can conceal locally counterproductive updates; agreement and utility help identify where attenuation is useful.

6 Conclusion

We presented VD-OPD to preserve the teacher's consistent support for alternative continuations while attenuating prompt-sensitive supervision. By retaining geometric consensus mass, the method couples target construction with selective weighting in one objective. Experiments show reasoning gains across model families and improved robustness to unseen prompt variants; two-view and compute controls support benefits beyond additional teacher queries, while ablations support contributions from consensus formation and selective weighting. Shared teacher errors remain a limitation because cross-view agreement alone cannot guarantee a correct target. Consistency across task-preserving inputs can guide which teacher behaviors to transfer. How to select complementary prompt views that maximize transfer gains under a fixed teacher-query budget remains open.

AI USE STATEMENT

Generative AI tools, including OpenAI Codex, assisted with checking proofs and refining mathematical expressions. They also supported literature search and synthesis, drafting and revising the English manuscript, checking for typographical errors, preparing LaTeX and references, and developing or editing figures and plotting code. The authors take responsibility for the final content of this work, including its scientific claims, experimental evidence, and AI-assisted text, code, and figures.

REFERENCES

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

Saptarshi Bandyopadhyay and Soon-Jo Chung. Distributed estimation using bayesian consensus filtering. *arXiv preprint arXiv:1403.3117*, 2014. URL https://arxiv.org/abs/1403.3117.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

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://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

Jiaxin Guo, Yanwei Yue, Xuanbo Fan, Chunyu Yang, and Yan Zhang. Learning from consensus and disagreement: Unsupervised on-policy self-distillation with minority-trajectory contrast. *arXiv preprint arXiv:2608.08764*, 2026. doi: 10.48550/arXiv.2608.08764. URL https://arxiv.org/abs/2608.08764.

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 (Volume 1: Long Papers)*, 2024. URL https://arxiv.org/abs/2402.14008.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Tom Heskes. Selecting weighting factors in logarithmic opinion pools. In *Advances in Neural Information Processing Systems*, volume 10, 1997. URL https://papers.nips.cc/paper/1413-selecting-weighting-factors-in-logarithmic-opinion-pools.pdf.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. doi: 10.48550/arXiv.1503.02531. URL https://arxiv.org/abs/1503.02531.

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

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. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64855.

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. ToDi: Token-wise distillation via fine-grained divergence control. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8091, 2025. doi: 10.18653/v1/2025.emnlp-main.409. URL https://aclanthology.org/2025.emnlp-main.409/.

Minsang Kim and Seung Baek. Explain in your own words: Improving reasoning via token-selective dual knowledge distillation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a89331e2ecbbd9d9618bc94717a91155-Abstract-Conference.html.

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, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31044–31062. PMLR, 2025. URL https://proceedings.mlr.press/v267/ko25a.html.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.14858.

Enhan Li, Junhao He, and Hongyang Du. CROP: Task relevance via counterfactuals for selective on-policy distillation. *arXiv preprint arXiv:2608.13387*, 2026. URL https://arxiv.org/abs/2608.13387.

Meta. Llama 3.2 model card. Official model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct.

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*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6c0e99d736da621403018ca7b32b1a4d-Abstract-Conference.html.

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

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.01574.

Wenda Xu, Rujun Han, Zifeng Wang, Long Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a2747a3844ca1e4667fbff3f558eb39b-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Zhiwei Zhang, Zechen Sun, Fei Zhao, Kang Peng, Bin Liang, Huayu Deng, Yao Hu, Kam-Fai Wong, and Mu Chuan. Verify before you distill: Prompt-level teacher gating for on-policy distillation. *arXiv preprint arXiv:2609.02998*, 2026. doi: 10.48550/arXiv.2609.02998. URL https://arxiv.org/abs/2609.02998.

APPENDIX CONTENTS

Page

A Technical Details . . . 14
A.1 Prompt wrappers and alignment . . . 14
A.2 Distribution evaluation and optimization . . . 14
A.3 Complete training algorithm . . . 15

B Experimental Details . . . 16
B.1 Training data and splits . . . 16
B.2 Evaluation protocols and metrics . . . 16
B.3 Baselines and comparison protocols . . . 17
B.4 Optimization, serialization, and model selection . . . 18
B.5 Resource accounting and compute controls . . . 18
B.6 VD-OPD implementation and mechanism ablations . . . 19
B.7 Local-update diagnostic protocol . . . 20
B.8 Target-improvement diagnostic protocol . . . 20

C Additional Experimental Results . . . 22
C.1 Per-benchmark results across model backbones . . . 22
C.2 Supplementary evaluation tasks . . . 25
C.3 Per-benchmark unseen-wrapper robustness . . . 25
C.4 Additional training protocols and resource comparisons . . . 26
C.5 Local-update utility . . . 27
C.6 Empirical validation of the target-improvement criterion . . . 28

D Proofs . . . 30
D.1 Square-root geometry and limiting cases . . . 30
D.2 Generalized KL characterization . . . 30
D.3 Derivatives of the distillation loss . . . 30
D.4 Cross-view setting and scope . . . 31
D.5 Exponential tilting and gauge-invariant curvature . . . 31
D.6 Optimal range and variance envelopes . . . 32
D.7 Exact comparison with the original teacher . . . 34
D.8 Distribution-sensitive curvature and its tightness . . . 35
D.9 Exact minimax risk under bounded binary shared bias . . . 37
D.10 Limits of agreement and unrestricted shared bias . . . 40

A TECHNICAL DETAILS

A.1 PROMPT WRAPPERS AND ALIGNMENT

We use a fixed chat serialization for both models. Each original prompt ends with a user message, followed by the fixed assistant generation prefix. A wrapper prepends one of the strings in Table 5, followed by exactly two newline characters (U+000A), to the final user-message content. All original characters, including task instructions and answer-format requirements, are retained as an unchanged suffix. System messages, earlier dialogue turns, message roles, and the assistant prefix are identical across views. The original view adds no text.

Table 5: Fixed prefixes for the three alternative teacher views. Each prefix is followed by two newline characters and the original final user message.

View	Exact prefix
Instruction 1	`Solve the following problem.`
Instruction 2	`Find a solution to the problem below.`
Layout	`Problem:`

Let $L_{\max}$ be the maximum response length, including EOS. Before sampling a response, we require the original encoded prompt plus $L_{\max}$ tokens to fit both model context limits. Invalid originals are excluded before generation. An alternative is valid when its prompt plus the same response allowance fits the teacher context limit and its serialized messages preserve the original content and roles. We sample uniformly from valid alternatives once per response. An empty valid set selects the identity transformation. All views of a source problem belong to the same data split; view selection never reads reference answers.

After generation, we attach the same response token IDs to each serialized prompt. The teacher distribution for $y_{i,t}$ is read from the output immediately preceding that token, so it conditions on $y_{i,<t}$. Causal masking permits the complete response to be scored in one forward pass per view. This procedure preserves response-token alignment despite different prompt lengths.

A.2 DISTRIBUTION EVALUATION AND OPTIMIZATION

Student rollouts are sampled from the full vocabulary at temperature one and terminate at EOS, with a cap of $L_{\max}$ response tokens. Teacher and student scoring also use temperature one. Dropout is disabled during scoring and gradient evaluation. We compute log-probabilities with full-vocabulary log-softmax and form each target as $\mathbf{c}_{i,t} = \exp\{\frac{1}{2}[\log \mathbf{p}^{(1)}_{i,t} + \log \mathbf{p}^{(2)}_{i,t}]\}$. Target construction and loss reductions use float32. The targets remain detached. Averaging log-probabilities retains $Z_{i,t}$; applying softmax to the averaged teacher logits would normalize this mass away.

We evaluate the full-vocabulary loss in blocks of response positions. The model forward passes provide the hidden states preceding vocabulary projection; the output heads materialize logits for one block at a time. Each block contributes its gradient with the common denominator N_k. Student parameters remain fixed until the complete batch gradient has been accumulated. The position blocks introduce no truncation of vocabulary support. Prompt and padding positions are masked out, and generated EOS tokens contribute to the loss. A batch with no valid response positions is skipped. Nonfinite values invalidate the entire batch: accumulated gradients are discarded and the optimizer state remains unchanged. A batch with exactly zero total consensus mass also skips the optimizer step.

For N_k response positions and vocabulary size d, the consensus construction and loss require $O(N_k d)$ elementwise work. A position block of size b requires $O(bd)$ temporary distribution storage, in addition to hidden states, model activations, parameters, and optimizer state. The two teacher views require two scoring passes for each student response. Algorithm 1 summarizes the update; inference uses the student on the original prompt.

A.3 COMPLETE TRAINING ALGORITHM

Algorithm 1 View-Decomposed On-Policy Distillation

Require: Prompt distribution $\mathcal{D}$; student θ_0; frozen teacher π_T; fixed wrappers; response cap $L_{\max}$; optimizer; iteration budget J.
Ensure: Student parameters θ_J.

for $k = 0, \ldots, J-1$ **do**
  Sample prompts; validate originals and alternative wrappers using Appendix A.1. Exclude invalid originals.
  For each remaining x_i, set $x_i^{(1)} = x_i$ and sample $x_i^{(2)}$ uniformly from its valid alternatives; an empty set selects $x_i^{(2)} = x_i$.
  Generate $y_i \sim \pi_{\theta_k}(\cdot \mid x_i)$ and retain its token IDs for both views.
  Collect valid response-position pairs $\mathcal{T}_k$ and set $N_k = |\mathcal{T}_k|$. If $N_k = 0$, set $\theta_{k+1} = \theta_k$ and continue.
  Score both teacher views and the original student view; initialize the accumulated student gradient and total consensus mass to zero.
  for each position block $\mathcal{C} \subseteq \mathcal{T}_k$ **do**
    Compute full-vocabulary log-probabilities at all $(i,t) \in \mathcal{C}$.
    Form $\mathbf{c}_{i,t} = \exp\{[\log \mathbf{p}_{i,t}^{(1)} + \log \mathbf{p}_{i,t}^{(2)}]/2\}$ and accumulate $\sum_{(i,t)\in\mathcal{C}} \mathbf{1}^\top \mathbf{c}_{i,t}$.
    Accumulate the gradient of $-N_k^{-1} \sum_{(i,t)\in\mathcal{C}} \mathbf{c}_{i,t}^\top \log \mathbf{q}_{i,t}(\theta)$ at θ_k; release block distribution buffers.
  end for
  If the batch is finite with positive total consensus mass, take one optimizer step to obtain θ_{k+1}; otherwise retain $\theta_{k+1} = \theta_k$ and the optimizer state.
  Discard the rollout batch and accumulated gradients.
end for
return θ_J.

B Experimental Details

B.1 Training data and splits

Our primary training source is the official MATH training split (Hendrycks et al., 2021). After deduplication and removal of detected overlaps with evaluation questions, we reserve 500 problems for implementation and mechanism diagnostics and 500 for development; the remaining problems form the optimization set. Splits are stratified by subject and difficulty with split seed 42. Every derived view remains in its source problem's split. Distillation uses question text without reference solutions or answer-based filtering; development answers are used only for configuration selection. We freeze the example IDs, filtering rules, and dataset revisions before final training. All three backbones use this split, holding the training corpus fixed across model-scale comparisons. The common-protocol tables report runs on this MATH split.

B.2 Evaluation protocols and metrics

B.2.1 Benchmark definitions and scoring

We use the canonical 500-question MATH500 subset (Hendrycks et al., 2021); the standard AMC23, AIME24, and AIME25 releases; Minerva Math (Lewkowycz et al., 2022); and the English, text-only, open-ended mathematics subset of OlympiadBench (He et al., 2024). GSM8K uses its official test split (Cobbe et al., 2021). We freeze dataset IDs and evaluator versions, including the symbolic-equivalence rules for mathematical answers. Missing or malformed final answers receive zero credit. Reported mathematical evaluations use zero-shot prompting and eight independent completions per question, with sampling temperature 1.0, top-$p = 0.8$, and an 8,192-token response cap. Prompt wording, stop tokens, and answer extraction are identical across methods within each model family.

For question i, let c_i be the number of correct completions among eight samples. We compute

$$\mathrm{Avg@8} = \frac{1}{8n}\sum_{i=1}^{n} c_i, \qquad \mathrm{Pass@8} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[c_i > 0]. \tag{10}$$

For AIME24 and AIME25, Pass@8 is computed over 30 problems for each seed and then averaged over three seeds. Therefore, the reported values are rounded from multiples of $100/90\% = 1.111\ldots\%$. All reported AIME Pass@8 values are computed from per-question predictions. AMC23 Pass@8 is computed over 40 problems per seed and averaged over three seeds; its reported values are rounded from multiples of $100/120\% = 0.8333\ldots\%$.

Tables 6–8 give the per-dataset results. Any six-benchmark aggregate is an unweighted mean over all six primary mathematics benchmarks; GSM8K and cross-domain scores remain separate. GPQA-Diamond uses zero-shot prompting and MMLU-Pro uses the fixed five-shot validation exemplars (Rein et al., 2024; Wang et al., 2024). Their primary metric is exact option accuracy under greedy decoding with the same response cap. We report the MMLU-Pro mathematics subset separately from the aggregate.

For GPQA-Diamond, greedy decoding itself is deterministic. The evaluation seed affects only a fixed permutation of the multiple-choice options for each question. The option labels are mapped back to the canonical answer before scoring. Evaluation seeds 11, 22, and 33 use three independently generated but method-shared permutations. Thus the frozen teacher and initial student use identical checkpoint weights across the three evaluations, while their scores may differ because of option-order sensitivity. For trained methods, each training seed is paired with the corresponding evaluation-seed permutation. Reported accuracy is the mean over the three evaluations, each covering all 198 questions.

B.2.2 Unseen-view evaluation

Training uses the three wrappers specified in Appendix A.1. We reserve four additional prefixes exclusively for evaluation: `Please solve this problem.`, `Determine the answer to the following problem.`, `Problem statement:`, and `Question:`. Each is followed by two newline characters and the unchanged original final user message. The question, formulas,

options, answer-format requirements, message roles, and assistant boundary remain fixed. We evaluate the student independently on every view, without teacher assistance or answer ensembling. Let $a_{i,m}$ denote Avg@8 for question i under unseen wrapper m, with $M = 4$. We report

$$\mathrm{MeanView} = \frac{1}{nM} \sum_{i,m} a_{i,m}, \tag{11}$$

$$\mathrm{WorstView} = \frac{1}{n} \sum_{i} \min_{m} a_{i,m}, \qquad \mathrm{WorstTemplate} = \min_{m} \frac{1}{n} \sum_{i} a_{i,m}. \tag{12}$$

The last two metrics distinguish per-question vulnerability from the globally weakest template. For Table 3, we compute each metric separately on each of the six mathematics benchmarks and report the unweighted mean. These definitions imply $\mathrm{WorstView} \leq \mathrm{WorstTemplate} \leq \mathrm{MeanView}$, both per benchmark and after macro-averaging. These accuracy measures use answer correctness over the specified wrapper set and sampling budget.

B.3 Baselines and comparison protocols

B.3.1 Core baselines and controlled adaptations

We include the initial student and frozen teacher as reference points. GKD uses fully student-generated trajectories and compares FKL, RKL, and generalized JSD, with the JSD parameter initialized to the published 0.9 setting (Agarwal et al., 2024; Kim & Baek, 2026). Full-vocabulary RKL and sampled-token OPD use distinct objectives. ToDi retains its detached vocabulary-wise divergence weights (Jung et al., 2025); we denote its application to fresh student trajectories by ToDi-OP. TSD-KD retains indirect token ranking, selective direct distillation, and entropy regularization, starting from top-10 candidates, JSD parameter 0.9, and the published selection settings (Kim & Baek, 2026). Veto retains its teacher–student target construction and initializes its linearly decayed parameter at 0.8 (Jang et al., 2026).

EOPD starts from entropy threshold 0.8 nats, FKL weight 1.0, and teacher top-16 FKL (Jin et al., 2026). The full-vocabulary FKL variant isolates vocabulary truncation from the objective change. We evaluate these reimplemented objectives under the common one-update schedule.

B.3.2 Additional comparisons and information access

On the primary Qwen3-1.7B setting, we include SeqKD (Kim & Rush, 2016), off-policy soft-target KD (Hinton et al., 2015), SKD (Xu et al., 2025), and DistiLLM-2 (Ko et al., 2025). The first two use fixed responses from the same frozen teacher; SKD preserves teacher–student token interleaving, and DistiLLM-2 preserves its teacher/student response contrast and refresh schedule. Teacher-generated responses, verification passes, and replay are charged to their respective methods. These methods form a separate comparison that preserves their algorithm-specific trajectory protocols.

CROP (Li et al., 2026), a 2026 preprint, is an auxiliary-resource comparison against VD-OPD, EOPD (full), and sampled-token OPD on an identical validated-triplet training subset. EOPD in this subset uses full-vocabulary FKL with the same entropy threshold of 0.8 nats as in the main controlled comparison. We preserve its paraphrase/counterfactual construction, 20% per-response token-selection budget, and selected-token normalization, and include random selection at the same retention rate. The subset changes training data only; all blocks in Table 11 use the same unchanged MATH500 test set with the common zero-shot decoding protocol. Its published construction pipeline uses GPT-5.1 and human semantic checks. For CROP, the reported FLOPs and GPU-hours include only the local training pipeline after the validated triplets have been constructed. External LLM/API calls and human semantic validation used for triplet construction are excluded from the accounting. Table 11 therefore reports local training-side compute. This subset experiment is evaluated separately from the strictly matched-teacher-compute comparison.

GRPO (Shao et al., 2024), with eight responses per prompt and answer-verifiable rewards, is reported in a separate answer-supervised reference block and excluded from label-free distillation rankings.

B.4 Optimization, serialization, and model selection

Core experiments use eight NVIDIA A100 80GB GPUs. All runs use full-parameter training, BF16 model computation, sharded optimizer states, gradient checkpointing, and float32 probability/consensus reductions. Dropout is disabled. Our default is AdamW with learning rate 3×10^{-6}, $(\beta_1, \beta_2) = (0.9, 0.95)$, numerical epsilon 10^{-8}, zero weight decay, gradient-norm clipping at 1.0, 3% warmup, and cosine decay. A rollout batch contains 128 prompts and one response per prompt, with temperature 1.0, top-$p = 1.0$, no top-k sampling restriction, and a 4,096-token response cap. Microbatches accumulate into one optimizer step, followed immediately by fresh generation. The matched-rollout budget is three passes over the optimization prompts. Learning rates are selected from $\{10^{-6}, 3 \times 10^{-6}, 10^{-5}\}$ on development data; every method receives the same maximum of nine configuration trials and the same pilot-compute allowance, including its published default. The common schedule applies to controlled comparisons; the separately reported trajectory protocols retain their algorithm-specific schedules. Final runs use seeds 11, 22, and 33, with independent recorded streams for training, view selection, and evaluation.

Qwen3 runs use the teacher's non-thinking chat serialization, including the fixed empty thinking block, for both teacher and student. The Llama pair uses one shared instruction-chat serialization. We verify vocabulary indices, special-token IDs, and output-head compatibility within each pair. Distribution alignment and comparisons are performed within model pairs, with separate results for base and instruction-tuned students. Model revisions, serialization strings, software versions, and evaluation seeds are frozen in the run manifest. Invalid or overlong inputs are rejected before generation; the common retained prompt set is used by all methods in a paired comparison. Context checks reserve space for the complete response allowance under every required view.

Model selection. Checkpoints and hyperparameters maximize MATH development Avg@8, with ties resolved by lower training cost; benchmark test answers never determine training, wrappers, or stopping.

B.5 Resource accounting and compute controls

B.5.1 Resource matching

For the controlled OPD comparison, we match source-question order, rollout count, response cap, optimizer-update count, and initialization. Each method samples from its own current student; identical trajectories are used only in frozen-checkpoint local interventions. Generated-token counts and total computation can vary at equal rollout counts. The additional-protocol table is descriptive cost reporting; the matched-total-FLOPs comparison is isolated in Table 2.

B.5.2 Resource accounting

In Tables 2 and 11, FLOPs and GPU-hours are per-run means over seeds 11, 22, and 33. For each run, GPU-hours sum occupied GPU time across the eight A100 80GB devices before averaging over seeds. Reported FLOPs include student generation, teacher generation/scoring, and student forward/backward computation.

The accounting boundary is the local training pipeline. For methods that require externally constructed resources, such as CROP's validated paraphrase/counterfactual triplets, external API inference and human validation are excluded. The measurements cover the training stage after these external resources have been constructed.

B.5.3 Controls for teacher access and training computation

Two-view EOPD. For the primary Qwen3-8B $\rightarrow$ Qwen3-1.7B-Base setting, EOPD-2V uses the same original-prompt and wrapper-view construction as VD-OPD. Both teacher distributions score the same sampled response token IDs at each update. Let $\mathcal{L}_{\mathrm{EOPD}}(\mathbf{p}, \mathbf{q})$ denote the single-view EOPD objective (Jin et al., 2026), using the full-vocabulary FKL variant from the main comparison. We define

$$\mathcal{L}_{\text{EOPD-2V}} = \frac{1}{2}\left[\mathcal{L}_{\mathrm{EOPD}}(\mathbf{p}^{(1)}, \mathbf{q}) + \mathcal{L}_{\mathrm{EOPD}}(\mathbf{p}^{(2)}, \mathbf{q})\right]. \tag{13}$$

Each term preserves the original sampled-token OPD component and teacher-entropy-gated FKL component. The entropy condition $H(\mathbf{p}^{(m)}) > 0.8$ nats is evaluated separately for each view, with FKL weight 1.0 and the complete vocabulary. Both terms use the same student predictions on the original prompt. Their losses are averaged before taking one optimizer step per fresh rollout batch. EOPD-2V and default VD-OPD have the same teacher-query count, two-view construction, rollout count, update count, initialization, data split, and evaluation protocol. Each method generates trajectories from its own current student. The comparison tests independently applying the single-view objective against forming a geometric consensus with state-specific mass weighting.

Total-FLOPs control. VD-OPD-Budget retains two teacher views per response and shortens training to the 4.7×10^{18} total-FLOPs budget of the default EOPD (full) run. It uses the same consensus objective and evaluation protocol. The budget matches accumulated total training FLOPs, including student generation, teacher scoring, and student training; rollout and update counts can differ. The reported GPU-hours are 30.5 and 30.3, respectively; elapsed device time is measured independently of the FLOPs budget. Default VD-OPD and EOPD-2V both report 5.8×10^{18} total FLOPs, with 37.2 and 37.0 GPU-hours. All costs follow the per-run three-seed accounting above. Total teacher-scoring cost depends on both the number of views per response and the duration of training.

Accuracy versus teacher-scoring FLOPs. Figure 5 uses the recorded cumulative teacher-only FLOPs, in units of 10^{18}. These count the frozen teacher's scoring passes for all prompt views and exclude student generation and student training. The four nominal budgets are $0.5\times$, $1.0\times$, $1.5\times$, and $2.0\times$ the default full EOPD teacher-scoring budget. Each point uses its measured cost as the horizontal coordinate. Their paired mean costs differ by less than 0.2%.

For method m, nominal budget b, and seed $s \in \{11, 22, 33\}$, let $C^T_{m,b,s}$ be the recorded teacher-only FLOPs and $a_{m,b,s}$ the Avg@8 score in percent. The point and vertical error bar are

$$\overline{C}^T_{m,b} = \frac{1}{3}\sum_s C^T_{m,b,s}, \quad \overline{a}_{m,b} = \frac{1}{3}\sum_s a_{m,b,s}, \quad s_{m,b} = \sqrt{\frac{1}{2}\sum_s (a_{m,b,s} - \overline{a}_{m,b})^2}. \tag{14}$$

Error bars span $\overline{a}_{m,b} \pm s_{m,b}$ and describe between-seed variation. Each curve joins four measured means by straight segments without smoothing. The six-math score is the unweighted benchmark mean within each seed, followed by the mean across seeds. Rollout and update counts may differ at a matched teacher-scoring budget.

At its default $1\times$ budget, EOPD (full) uses mean teacher-only FLOPs of 1.101333×10^{18} and reaches 34.8/66.0 six-math/MATH500 Avg@8. Beyond this point, its curve uses additional training. Default VD-OPD reaches 36.7/67.8 at mean teacher-only FLOPs of 2.200667×10^{18}, the nominal $2\times$ point. These scores match Table 2. VD-OPD at nominal $1\times$ reaches 35.8/66.9; it is distinct from VD-OPD-Budget, which matches total training FLOPs and reaches 36.0/67.0.

B.6 VD-OPD IMPLEMENTATION AND MECHANISM ABLATIONS

We sample one valid alternative wrapper per response and score the same response token IDs under both teacher views. All scoring temperatures equal one. The detached target is $c_t = \exp\{(\log p_{t,1} + \log p_{t,2})/2\}$, evaluated over the complete vocabulary in blocks of 32 response positions. VD-OPD uses no entropy threshold or mixing coefficient. We retain the unweighted valid-response-token denominator N_k, include generated EOS tokens, and exclude prompt and padding positions. Neither vocabulary truncation nor normalization by $\sum_t Z_t$ is applied. Changing the block size changes memory use while preserving the objective. Identity-view fallback and numerical failures follow Appendix A.2.

The mechanism study uses the primary Qwen3-8B $\rightarrow$ Qwen3-1.7B-Base pair. Table 4 specifies the four target/weight combinations and two weight-assignment controls. It uses the common MATH split, one-update schedule, matched rollout/update budget, and seeds 11, 22, and 33. Each configuration generates fresh trajectories from its own student. Mathematical evaluation uses the six-benchmark sampled protocol. Appendix B.7 specifies the accompanying local-update diagnostic.

B.7 LOCAL-UPDATE DIAGNOSTIC PROTOCOL

Matched branches. This diagnostic uses only the primary Qwen3-1.7B setting and the 500 reserved diagnostic problems, split by source-problem ID into 250 update problems and 250 measurement problems with split seed 42. Diagnostic answers are used only to score the measurement set. The checkpoint, optimizer state, branch-update budget, and batch schedule are shared within each comparison. Each branch reuses the same sampled response token IDs and teacher views and changes only the target/weight configuration in Table 4. These branches are used only for the local diagnostic.

Measurements and aggregation. Table 12(a) summarizes the common pre-update response positions: full-vocabulary teacher entropies $H(p_1), H(p_2)$, student entropy $H(q)$, and retained mass Z. Statistics are token means within each seed, then means over seeds 11, 22, and 33. Part (b) reports Avg@8 before and after each matched branch on the disjoint measurement questions, using the common mathematics decoding protocol and paired evaluation random streams. Utility is the within-seed difference in percentage points, averaged across seeds; it is computed before display rounding. All six configurations share the same pre-update score. The intervention isolates changes to the objective at fixed rollouts; Z quantifies agreement between teacher views.

Token-level consensus-fit diagnostic. Figure 6 uses the VD-OPD branch and measures the decrease in normalized consensus cross-entropy at fixed pre-update response prefixes, as defined in Equation 9. Both teacher distributions and the normalized target remain fixed between student evaluations. The token-level diagnostic contains 8,192 valid positions for each of seeds 11, 22, and 33, giving 24,576 positions in total. Its three-seed mean retained mass is 0.740233, reported as 0.740 in Table 12(a). Each position inherits the final-answer correctness of its sampled response; averaging these labels within a Z bin therefore weights responses by their recorded positions. This token-level fit measure is distinct from the source-disjoint answer-accuracy change in panel (b).

For visualization, we divide $[0, 1]$ into 24 equal-width bins, compute the median and 10th–90th utility percentiles and the token-weighted correctness rate, and apply a three-bin moving average to these summaries. Empty bins remain empty. Hexagonal counts include all recorded positions. The moving average is applied only to the displayed summaries.

B.8 TARGET-IMPROVEMENT DIAGNOSTIC PROTOCOL

We use the reserved MATH diagnostic split. The 250 measurement problems provide four sampled fixed prefixes each, giving 1,000 analysis states; a separate set of 250 problems is used only for matched one-update branches. Each problem has a fixed candidate set containing one reference correct answer and three incorrect answers. The oracle distribution $p^\star$, teacher distributions, and student distributions in this diagnostic are defined over these same four candidates. The answer-derived $p^\star$ is used only for oracle diagnostics and does not enter the VD-OPD training objective. The original teacher view p_0 is paired with four unseen wrappers. We use seeds 11, 22, and 33.

B.8.1 CANDIDATE DISTRIBUTIONS

For candidate $a_k = (a_{k,1}, \ldots, a_{k,L_k})$ at the fixed state $s = (x, y_{<t})$, we use the length-normalized log-likelihood of its complete canonicalized final-answer string:

$$\begin{aligned} z_k^{(m)} &= \frac{1}{L_k} \sum_{\ell=1}^{L_k} \log \pi_T(a_{k,\ell} \mid x^{(m)}, y_{<t}, a_{k,<\ell}), \\ z_{k,\theta} &= \frac{1}{L_k} \sum_{\ell=1}^{L_k} \log \pi_\theta(a_{k,\ell} \mid x, y_{<t}, a_{k,<\ell}). \end{aligned} \tag{15}$$

At temperature one, the induced candidate distributions are

$$p^{(m)}(a_k) = \frac{\exp(z_k^{(m)})}{\sum_{j=1}^4 \exp(z_j^{(m)})}, \qquad q_\theta(a_k) = \frac{\exp(z_{k,\theta})}{\sum_{j=1}^4 \exp(z_{j,\theta})}. \tag{16}$$

Length normalization controls the effect of answer-token length on the scores. Only final-answer tokens are scored; reasoning text is excluded, and no multiple-choice option labels are introduced.

All teacher views and the student use identical candidate strings, tokenization, and EOS handling. The candidates are fixed independently of the prompt view.

B.8.2 Reference target and logit shifts

With correct candidate $a^\star$, we use the smoothed ground-truth distribution

$$p^\star(a_k) = \begin{cases} 1-\epsilon, & a_k = a^\star, \\ \epsilon/(K-1), & a_k \neq a^\star, \end{cases} \qquad K = 4, \quad \epsilon = 0.01. \tag{17}$$

Thus the correct candidate has mass 0.99 and each incorrect candidate has mass $0.01/3$. All teacher, student, and reference distributions belong to the same simplex Δ^3, with $p^\star$ strictly positive as required by Theorem 1. We compute the wrapper shift directly from the candidate logits:

$$\delta_m = P\big(z^{(m)} - z^{(0)}\big), \qquad P = I_4 - \frac{\mathbf{1}\mathbf{1}^\top}{4}. \tag{18}$$

This equals $P(\log p^{(m)} - \log p^{(0)})$, because P removes the additive softmax normalization constant. The resulting correction, curvature cost, held-out target KL, and student utility are all defined on the same fixed candidate space.

B.8.3 Cross-fitting and local-update utility

We perform four-fold leave-one-wrapper-out cross-fitting at each fixed prefix. In each fold, three wrappers estimate the correction and curvature cost, and the fourth is used only for held-out evaluation. Specifically, let $\mathcal{V}_{-h}$ contain the three fitting wrappers and $r_v \propto \sqrt{p_0 \odot p^{(v)}}$. With the shifts in Equation 18, the fitted quantities are

$$\widehat{\mu}_{-h} = \frac{1}{3}\sum_{v\in\mathcal{V}_{-h}} \delta_v, \qquad \widehat{\mathcal{G}} = \frac{1}{2}\langle p^\star - p_0, \widehat{\mu}_{-h}\rangle, \qquad \widehat{\mathcal{C}} = \frac{1}{3}\sum_{v\in\mathcal{V}_{-h}} D_{\mathrm{KL}}(p_0 \| r_v). \tag{19}$$

States are grouped by $\widehat{M}$ from Q1 (lowest) to Q4 (highest). With r^{held} formed from p_0 and the held-out wrapper, the margin and test outcome are

$$\widehat{M} = \widehat{\mathcal{G}} - \widehat{\mathcal{C}}, \qquad \Delta_{\mathrm{KL}}^{\mathrm{held}} = D_{\mathrm{KL}}(p^\star \| p_0) - D_{\mathrm{KL}}(p^\star \| r^{\mathrm{held}}). \tag{20}$$

Positive held-out improvement means that pooling brings the target closer to the diagnostic reference. "Improve" is the fraction of states with $\Delta_{\mathrm{KL}}^{\mathrm{held}} > 0$.

We evaluate student-update utility (RQ4) by comparing geometric-pooling and original-teacher branches under the matched-branch protocol of Appendix B.7. They share their initial checkpoint, optimizer state, rollouts, and one-update budget. Both use unit supervision weights, isolating target choice from the effect of Z. Their utility is evaluated against the common oracle candidate-answer target:

$$u_b = D_{\mathrm{KL}}(p^\star \| q_{\mathrm{before}}) - D_{\mathrm{KL}}(p^\star \| q_{\mathrm{after},b}), \qquad \Delta u = u_{\mathrm{pool}} - u_{\mathrm{original}}. \tag{21}$$

The equivalent oracle cross-entropy decrease gives the same utility because $p^\star$ is fixed. This differs from the consensus-fit loss in Equation 9. The answer-level comparison uses the source-disjoint measurement questions, with positive ΔAvg@8 favoring the pooling branch.

B.8.4 Mean-correction intervention

Corollary 5 predicts a failure case: if $\mu = \mathbb{E}\delta = 0$, pooling cannot improve the original teacher in expectation, and nonzero centered wrapper perturbations strictly increase expected target KL. We test the effect of mean correction by decomposing the wrapper shift into its mean and centered variation and rescaling the mean:

$$\delta = \mu + \epsilon, \qquad \mathbb{E}\epsilon = 0, \qquad \delta_\beta = \beta\mu + \epsilon, \qquad r_\beta = \mathrm{softmax}(z_0 + \delta_\beta/2). \tag{22}$$

For fixed p_0, $p^\star$, and μ, the population correction satisfies $\mathcal{G}_\beta = \beta\mathcal{G}_1$, while the curvature cost can vary nonlinearly with β.

C Additional Experimental Results

C.1 Per-benchmark results across model backbones

Tables 6–8 report per-benchmark Avg@8 and Pass@8 (%) for each teacher–student pair; blank cells denote unreported combinations. Qwen3-1.7B Avg@8 matches Table 1, whose rounded entries define the canonical six-benchmark mean. MATH500 scores underlie Figure 4; EOPD (full) and default VD-OPD scores also match the compute controls in Table 2. Unseen-wrapper aggregates in Table 3 use the same six benchmarks and Appendix B's definitions. Initial-student and frozen-teacher references use fixed checkpoint weights and common decoding; their AIME and AMC23 Pass@8 scores average three evaluation seeds.

Table 6: **Per-benchmark reporting: Qwen3-8B → Qwen3-1.7B-Base.** Avg@8 and Pass@8 (%); common MATH training split and one-update schedule. Avg@8 matches Table 1; MATH500 scores underlie the gains in Figure 4. Bold marks the best distilled student. Blank cells denote unreported combinations. Olymp. uses the subset in Appendix B.

Method	MATH 500	AMC23	AIME24	AIME25	Minerva	Olymp.
			Avg@8			
Frozen teacher	72.8	51.4	16.7	12.5	38.6	43.1
Initial student	54.7	32.8	8.3	5.1	25.4	30.2
GKD-FKL	58.9	36.1	9.7	6.2	27.8	33.0
GKD-RKL (full)	59.6	36.8	10.1	6.5	28.4	33.7
GKD-JSD	60.1	37.5	10.4	6.9	28.9	34.2
Sampled-token OPD	62.4	39.8	11.7	7.8	30.7	36.0
ToDi-OP	63.1	40.5	12.4	8.3	31.5	36.8
TSD-KD	63.8	41.2	12.8	8.9	32.1	37.5
Veto	64.2	41.7	13.1	9.2	32.8	38.0
EOPD (top-16)	65.1	42.8	13.9	9.9	33.7	39.1
EOPD (full)	66.0	43.6	14.4	10.4	34.5	40.0
VD-OPD (ours)	**67.8**	**45.9**	**15.8**	**11.7**	**36.9**	**42.1**
			Pass@8			
Frozen teacher	91.4	78.3	32.2	25.6	60.2	67.8
Initial student	79.5	60.8	18.9	13.3	44.1	50.4
GKD-FKL	82.1	65.0	21.1	15.6	47.5	54.2
GKD-RKL (full)	82.7					
GKD-JSD	83.0					
Sampled-token OPD	84.5	68.3	24.4	18.9	51.1	57.7
ToDi-OP	85.2	69.2	25.6	20.0	52.5	59.0
TSD-KD	85.8	70.0	25.6	21.1	53.3	60.1
Veto	86.1	70.8	26.7	21.1	54.1	61.0
EOPD (top-16)						
EOPD (full)	87.5	73.3	28.9	23.3	56.8	63.0
VD-OPD (ours)	**88.7**	**75.0**	**31.1**	**25.6**	**59.4**	**65.8**

Qwen3-1.7B. In Table 6, VD-OPD improves over EOPD (full) on all six benchmarks by 1.3–2.4 percentage points in Avg@8 and 1.2–2.8 points in Pass@8. Minerva has the largest Avg@8 gain (34.5% to 36.9%), while OlympiadBench has the largest Pass@8 gain (63.0% to 65.8%). The improvements therefore extend to both average sample accuracy and eight-sample coverage.

Table 7: **Per-benchmark reporting: Qwen3-8B → Qwen3-4B-Base.** Avg@8 and Pass@8 (%); common MATH training split and one-update schedule. MATH500 scores underlie the gains in Figure 4; the frozen-teacher reference is shared with Table 6. Bold marks the best distilled student. Blank cells denote unreported combinations. Olymp. uses the subset in Appendix B.

Method	MATH 500	AMC23	AIME24	AIME25	Minerva	Olymp.
			Avg@8			
Frozen teacher	72.8	51.4	16.7	12.5	38.6	43.1
Initial student	63.8	40.2	12.1	8.7	32.4	35.9
GKD-FKL	67.2	43.1	13.9	10.1	35.2	39.0
GKD-RKL (full)	67.8					
GKD-JSD	68.3					
Sampled-token OPD	70.4	46.8	15.6	11.7	37.9	41.8
ToDi-OP	71.2	47.5	16.2	12.4	38.6	42.5
TSD-KD	72.0	48.1	16.8	12.9	39.4	43.2
Veto	72.5	48.8	17.1	13.2	39.9	43.8
EOPD (top-16)	73.6	50.0	18.1	14.2	41.0	45.0
EOPD (full)	74.3	50.8	18.7	14.6	41.9	45.8
VD-OPD (ours)	**76.1**	**53.0**	**20.1**	**16.0**	**44.0**	**48.0**
			Pass@8			
Frozen teacher	91.4	78.3	32.2	25.6	60.2	67.8
Initial student	85.9		26.7	20.0		
GKD-FKL	87.9		27.8	21.1		
GKD-RKL (full)	88.3					
GKD-JSD	88.8					
Sampled-token OPD	90.0		30.0	23.3		
ToDi-OP	90.7		31.1	24.4		
TSD-KD	91.1		32.2	25.6		
Veto	91.5		33.3	26.7		
EOPD (top-16)						
EOPD (full)	92.6		35.6	28.9		
VD-OPD (ours)	**93.4**		**36.7**	**31.1**		

Qwen3-4B. Table 7 shows that VD-OPD retains its advantage with the larger student: Avg@8 exceeds EOPD (full) by 1.4–2.2 percentage points across all six benchmarks. On the three benchmarks with reported student Pass@8 results, the gains are 0.8 points on MATH500, 1.1 on AIME24, and 2.2 on AIME25. VD-OPD also exceeds the frozen-teacher reference in all six Avg@8 columns under this evaluation protocol. The missing Pass@8 cells preclude the corresponding six-benchmark comparison.

Table 8: **Per-benchmark reporting: Llama-3.1-8B-Instruct → Llama-3.2-3B-Instruct.** Avg@8 and Pass@8 (%); common MATH training split and one-update schedule. MATH500 scores underlie the gains in Figure 4. Bold marks the best distilled student. Blank cells denote unreported combinations. Olymp. uses the subset in Appendix B.

Method	MATH 500	AMC23	AIME24	AIME25	Minerva	Olymp.
			Avg@8			
Frozen teacher	49.7					
Initial student	38.5	22.1	5.1	3.3	18.9	22.7
GKD-FKL	41.0	24.5	6.0	4.2	20.7	24.8
GKD-RKL (full)	41.6					
GKD-JSD	42.0					
Sampled-token OPD	43.7	26.9	6.9	4.9	22.8	26.9
ToDi-OP	44.3	27.5	7.4	5.1	23.5	27.8
TSD-KD	44.8	28.1	7.8	5.6	24.0	28.4
Veto	45.2	28.7	8.1	5.8	24.5	29.0
EOPD (top-16)	46.0	29.4	8.6	6.2	25.3	29.8
EOPD (full)	46.3	30.0	9.0	6.5	25.8	30.2
VD-OPD (ours)	**47.9**	**32.1**	**10.4**	**7.8**	**27.8**	**33.5**
			Pass@8			
Frozen teacher	70.8					
Initial student	61.2		13.3	10.0		
GKD-FKL	63.8		15.6	11.1		
GKD-RKL (full)	64.4					
GKD-JSD	64.9					
Sampled-token OPD	66.7		17.8	13.3		
ToDi-OP	67.5		18.9	14.4		
TSD-KD	68.0		20.0	15.6		
Veto	68.5		21.1	16.7		
EOPD (top-16)						
EOPD (full)	69.9		23.3	18.9		
VD-OPD (ours)	**71.1**		**26.7**	**21.1**		

Llama-3.2-3B. The advantage also appears in the Llama pair (Table 8): VD-OPD improves over EOPD (full) by 1.3–3.3 percentage points in Avg@8 across the six benchmarks, with the largest gain on OlympiadBench (30.2% to 33.5%). Pass@8 increases by 1.2, 3.4, and 2.2 points on MATH500, AIME24, and AIME25, respectively. On MATH500, VD-OPD remains below the frozen teacher in Avg@8 (47.9% versus 49.7%) while slightly exceeding it in Pass@8 (71.1% versus 70.8%); eight-sample coverage and average sample accuracy therefore give distinct comparisons.

C.2 SUPPLEMENTARY EVALUATION TASKS

Table 9 reports GSM8K Avg@8/Pass@8 and MMLU-Pro-Math greedy accuracy.

Table 9: **Supplementary evaluation for Qwen3-1.7B-Base.** Scores (%) under the common protocol: zero-shot GSM8K and five-shot MMLU-Pro-Math. GSM8K and the MMLU-Pro mathematics subset are reported separately from the primary six-benchmark average. Bold marks the best distilled student.

	GSM8K		MMLU-Pro-Math
Method	Avg@8	Pass@8	Greedy
Frozen teacher	82.4	96.1	58.7
Initial student	68.5	89.0	44.3
GKD-FKL	72.1	91.2	46.0
GKD-RKL (full)	72.8	91.6	46.7
GKD-JSD	73.2	92.0	47.1
Sampled-token OPD	74.8	92.9	48.3
ToDi-OP	75.4	93.2	48.9
TSD-KD	76.0	93.6	49.4
Veto	76.4	93.8	49.8
EOPD (top-16)	77.1	94.1	50.6
EOPD (full)	77.8	94.5	51.2
VD-OPD (ours)	**79.0**	**95.0**	**52.4**

VD-OPD leads the reported distilled students on both supplementary tasks. Relative to EOPD (full), GSM8K improves from 77.8% to 79.0% in Avg@8 and from 94.5% to 95.0% in Pass@8, while MMLU-Pro-Math greedy accuracy rises from 51.2% to 52.4%. These results extend the observed advantage to an additional free-form mathematics task and a multiple-choice evaluation. A gap to the frozen teacher remains: 3.4 points in GSM8K Avg@8, 1.1 in Pass@8, and 6.3 in MMLU-Pro-Math, with the largest remaining gap on the multiple-choice task.

C.3 PER-BENCHMARK UNSEEN-WRAPPER ROBUSTNESS

Table 10 compares VD-OPD's canonical-prompt accuracy with its mean accuracy over the four unseen wrappers. The six-benchmark unseen MeanView is 33.65% when calculated from the displayed entries, rounding to the 33.7% reported in Table 3. The canonical macro-average is 36.7%; the mean decrease is 3.05 percentage points from the displayed values. Accuracy decreases on every benchmark. Worst-view and worst-template scores in Table 3 are computed separately from per-question wrapper results using Appendix B.

Table 10: **VD-OPD under canonical prompts and unseen wrappers.** Qwen3-1.7B-Base; Avg@8 (%). MeanView averages the four held-out wrappers. Canonical scores repeat Table 1. Macro-averages are calculated from the displayed per-benchmark values.

Benchmark	Canonical Avg@8	Unseen MeanView
MATH500	67.8	64.0
AMC23	45.9	41.7
AIME24	15.8	13.9
AIME25	11.7	10.2
Minerva	36.9	33.1
OlympiadBench	42.1	39.0
Six-benchmark macro-average	36.7	33.7

C.4 Additional training protocols and resource comparisons

Table 11 separates the trajectory comparison—SeqKD (Kim & Rush, 2016), off-policy soft KD (Hinton et al., 2015), SKD (Xu et al., 2025), and DistiLLM-2 (Ko et al., 2025)—from the common validated-triplet training-subset comparison with CROP (Li et al., 2026) and the answer-supervised GRPO reference (Shao et al., 2024). CROP and random selection retain 20% of response tokens; other subset rows retain their dense objectives. Training subsets and information access follow Appendix B. Comparisons are made within each block, separately from the controlled one-update objective ranking.

Under the trajectory-protocol comparison on MATH500, VD-OPD obtains 67.8/88.7 Avg@8/Pass@8, compared with 64.2/86.0 for DistiLLM-2 and 62.4/84.5 for sampled-token OPD. The VD-OPD and sampled-token OPD references match the MATH500 entries in Table 6, which underlie the gains in Figure 4. The validated-triplet subset changes the training data only; all methods are evaluated on the same unchanged MATH500 test set. Within this training-subset block, VD-OPD reaches 66.5/87.5, compared with CROP's 65.0/86.8.

The resource columns quantify local training computation. Within the trajectory-protocol block, VD-OPD's higher accuracy comes with greater cost than DistiLLM-2: 5.8×10^{18} versus 5.2×10^{18} FLOPs and 37.2 versus 33.1 GPU-hours. Within the validated-triplet subset, VD-OPD improves on CROP by 1.5/0.7 Avg@8/Pass@8 points while using 1.3×10^{18} versus 1.5×10^{18} FLOPs and 8.9 versus 10.3 GPU-hours. CROP's external triplet construction and human validation fall outside this measurement boundary. These comparisons concern local training within each protocol block, rather than total data-preparation cost.

Table 11: **Additional comparisons for Qwen3-1.7B-Base.** Avg@8/Pass@8 are zero-shot MATH500 results (%). FLOPs and GPU-hours report local training-side computation, including student generation, teacher generation/scoring, and student forward/backward passes. For CROP, external LLM/API-based triplet construction and human semantic validation are excluded. The resource columns cover this local training stage only.

Method	Avg@8	Pass@8	FLOPs (10^{18})	GPU-h
Teacher/student trajectory protocols				
SeqKD	60.5	83.4	4.0	25.7
Off-policy soft KD	61.3	84.0	4.1	26.2
SKD	63.0	85.1	4.8	30.4
DistiLLM-2	64.2	86.0	5.2	33.1
Sampled-token OPD	62.4	84.5	4.1	26.8
VD-OPD (ours)	**67.8**	**88.7**	5.8	37.2
Common validated-triplet training subset (same MATH500 evaluation)				
Sampled-token OPD	61.8	84.0	1.0	6.8
EOPD (full)	64.7	86.2	1.2	8.1
Random selection (20%)	62.1	84.3	1.0	6.9
CROP (20%)	65.0	86.8	1.5	10.3
VD-OPD (ours)	**66.5**	**87.5**	1.3	8.9
Answer-supervised reference				
GRPO	69.2	89.4	8.6	54.0

C.5 LOCAL-UPDATE UTILITY

Table 12 reports the shared pre-update statistics and matched-branch accuracy changes. The branch construction, measurement split, and token-level visualization protocol are specified in Appendix B.7.

Table 12: **Local-update utility.** Qwen3-8B $\rightarrow$ Qwen3-1.7B-Base; three-seed means. Panel (a) uses shared pre-update rollouts, so its statistics are common to the six branches. Panel (b) uses the same source-disjoint measurement set before and after each branch update. Entropies are in nats, $Z \in [0, 1]$, Avg@8 is in percent, and ΔAvg@8 is in percentage points. Bold marks the largest post-update accuracy and gain.

(a) Shared pre-update diagnostics

$H(p_1)$	$H(p_2)$	$H(q)$	Z
1.18	1.16	1.42	0.740

(b) Matched branch updates

Configuration	Before Avg@8	After Avg@8	ΔAvg@8
Arithmetic	54.8	55.6	+0.8
Weighted arithmetic		56.0	+1.2
Geometric		56.3	+1.5
VD-OPD		**56.8**	**+2.0**
Shuffled weight		56.4	+1.6
Batch-mean weight		56.6	+1.8

Panel (a) provides a common starting point: the student entropy is 1.42 nats, compared with 1.18 and 1.16 for the two teacher views, and their mean affinity is $Z = 0.740$. Because all branches reuse these pre-update rollouts, the statistics describe the shared inputs rather than differences caused by the updates.

In panel (b), geometric pooling improves Avg@8 by 1.5 points, compared with 0.8 for arithmetic pooling. VD-OPD raises the gain to 2.0 points, adding 0.5 points over the unweighted geometric branch. Shuffled and batch-mean weights yield gains of 1.6 and 1.8 points, respectively. This ordering is consistent with a benefit from matching the affinity weight to each state, beyond changing target shape or average update scale. The comparison concerns one local update and reported three-seed means; it does not establish statistical significance or an end-to-end effect size.

C.6 Empirical validation of the target-improvement criterion

Theorem 1 separates the benefit of mean correction, $\mathcal{G}$, from the curvature cost, $\mathcal{C}$: geometric pooling improves expected target KL exactly when $\mathcal{G} - \mathcal{C} > 0$. We examine whether an estimated margin predicts improvement on a held-out teacher view and whether it also tracks the benefit of a student update. This controlled test measures KL over four candidate answers. Main training operates on full-vocabulary token distributions.

Candidate distributions, cross-fitting, and the mean-correction intervention are defined in Appendix B.8.

C.6.1 Cross-fitted target improvement and local-update utility

Table 13: **Cross-fitted validation of the theoretical improvement criterion.** Quartiles are formed using $\widehat{M}$ without the held-out prompt view. The first five numeric columns describe target quality; the last two compare the geometric-pooling and original-teacher student branches. All differences and criterion estimates are in nats, except Improve (%) and ΔAvg@8 (percentage points). Positive differences favor pooling.

	Cross-fitted estimates			Held-out target		Student update	
Quartile	$\widehat{\mathcal{G}}$	$\widehat{\mathcal{C}}$	$\widehat{M}$	$\Delta_{\mathrm{KL}}^{\mathrm{held}}$	Improve (%)	Δu	ΔAvg@8
Q1	0.006	0.038	-0.032	-0.025	18.7	-0.006	-0.3
Q2	0.013	0.021	-0.008	-0.006	39.5	-0.001	$+0.0$
Q3	0.023	0.015	$+0.008$	$+0.006$	65.8	$+0.003$	$+0.4$
Q4	0.047	0.012	$+0.035$	$+0.028$	85.1	$+0.010$	$+1.1$

The cross-fitted criterion predicts target improvement on an unseen teacher view (Table 13). From the lowest to the highest margin quartile, held-out KL improvement changes from -0.025 to $+0.028$ nats, and the fraction of states benefiting from pooling rises from 18.7% to 85.1%. The student-update advantage follows the same ordering: Δu changes from -0.006 to $+0.010$ nats, while source-disjoint ΔAvg@8 changes from -0.3 to $+1.1$ percentage points. Thus a better margin tracks both a better target and a more useful local update in this diagnostic. These associations concern a single update with unit weights. End-to-end gains and the effect of Z are evaluated in the main experiments.

C.6.2 State-level predictive performance

Table 14 reports correlations between $\widehat{M}$ and $\Delta_{\mathrm{KL}}^{\mathrm{held}}$. Sign accuracy compares their signs, and AUROC uses $\widehat{M}$ to rank the binary outcome $\mathbf{1}[\Delta_{\mathrm{KL}}^{\mathrm{held}} > 0]$. Across seeds, the mean Spearman correlation is 0.74 ± 0.03 (mean $\pm$ sample standard deviation, computed from the reported seed-level coefficients), with mean sign accuracy 78.8% and AUROC 0.858. The separate fitting and evaluation views test the predictive validity of the criterion in Theorem 1 on held-out teacher predictions.

Table 14: **State-level prediction of held-out target improvement.** Correlations use the continuous KL improvement. Sign accuracy and AUROC evaluate whether that improvement is positive. The final row is the unweighted mean over the three displayed seeds; accuracy is in percent.

Seed	Spearman ρ	Pearson r	Sign accuracy (%)	AUROC
11	0.72	0.67	77.8	0.845
22	0.77	0.71	80.1	0.872
33	0.74	0.69	78.6	0.858
Mean	**0.74**	**0.69**	**78.8**	**0.858**

C.6.3 MEAN CORRECTION AND CENTERED WRAPPER VARIATION

Table 15: **Mean-correction intervention.** The intervention rescales the mean wrapper correction while retaining its centered component. Criterion estimates and empirical held-out KL improvements are in nats; positive improvements favor pooling. Values are rounded to three decimals.

Mean scale β	$\widehat{\mathcal{G}}$	$\widehat{\mathcal{C}}$	$\widehat{\mathcal{G}} - \widehat{\mathcal{C}}$	$\Delta_{\mathrm{KL}}^{\mathrm{held}}$
0.00	0.000	0.012	-0.012	-0.011
0.50	0.009	0.013	-0.004	-0.004
0.75	0.013	0.013	0.000	-0.001
1.00	0.017	0.014	$+0.003$	$+0.003$
1.50	0.026	0.017	$+0.009$	$+0.008$
2.00	0.034	0.022	$+0.012$	$+0.011$

At $\beta = 0$, the estimated correction vanishes, the curvature cost remains positive, and held-out pooling improvement is -0.011 nats (Table 15), consistent with the centered-noise failure case. The estimated margin and empirical improvement both become positive between $\beta = 0.75$ and 1.00. Increasing the mean component can therefore change the sign of the observed pooling benefit; agreement alone does not determine that sign.

D PROOFS

D.1 SQUARE-ROOT GEOMETRY AND LIMITING CASES

For $\mathbf{p}_1, \mathbf{p}_2 \in \Delta^{d-1}$, each square-root vector has unit Euclidean norm. Equation 3 gives $\mathbf{u}_\pm = (\sqrt{\mathbf{p}_1} \pm \sqrt{\mathbf{p}_2})/\sqrt{2}$. Expanding their elementwise squares yields $\bar{\mathbf{p}} = (\mathbf{u}_+^{\odot 2} + \mathbf{u}_-^{\odot 2})/2$ and $\mathbf{c} = (\mathbf{u}_+^{\odot 2} - \mathbf{u}_-^{\odot 2})/2 = \sqrt{\mathbf{p}_1 \odot \mathbf{p}_2}$. The unit-norm identities also give $\mathbf{u}_+^\top \mathbf{u}_- = 0$ and $\|\mathbf{u}_-\|_2^2 = 1 - \langle \sqrt{\mathbf{p}_1}, \sqrt{\mathbf{p}_2} \rangle$. Nonnegativity and Cauchy–Schwarz establish $0 \le Z \le 1$; coordinatewise AM–GM establishes $c(a) \le \bar{p}(a)$. These identities extend to distributions with zero entries. The geometric mass generally differs from the pointwise minimum of the two distributions.

For a two-token illustration, let both views assign $(0.5, 0.5)$ in the first case, and let them assign $(0.99, 0.01)$ and $(0.01, 0.99)$ in the second case. Both cases yield the arithmetic mean and normalized geometric target $(0.5, 0.5)$. Their retained masses are 1 and $2\sqrt{0.0099} \approx 0.199$, respectively. At the same student prediction, the second case therefore contributes approximately 0.199 times the first case's loss and gradient. If the two views agree on any distribution $\mathbf{p}$, their target is exactly $\mathbf{p}$. If their supports are disjoint, $\mathbf{c} = \mathbf{0}$ and the loss vanishes. Finite softmax logits give strictly positive probabilities, so exact zero overlap occurs only as a limit.

D.2 GENERALIZED KL CHARACTERIZATION

The generalized KL in Section 4.2 extends to zero reference entries by setting $0 \log(0/0) = 0$ and $v \log(v/0) = +\infty$ for $v > 0$.

Proposition 2 (Geometric consensus as an unnormalized barycenter)**.** *Let $\mathbf{p}_1, \mathbf{p}_2$ be strictly positive probability vectors, $\mathbf{c} = \sqrt{\mathbf{p}_1 \odot \mathbf{p}_2}$, and $Z = \mathbf{1}^\top \mathbf{c}$. For every $\mathbf{v} \in \mathbb{R}^d_{\ge 0}$,*

$$F(\mathbf{v}) := \frac{1}{2} \sum_{m=1}^{2} D_{\mathrm{I}}(\mathbf{v} \| \mathbf{p}_m) = D_{\mathrm{I}}(\mathbf{v} \| \mathbf{c}) + 1 - Z. \tag{23}$$

Consequently, $\mathbf{c}$ uniquely minimizes F, with minimum value $1 - Z$.

Proof. The identity $\log c(a) = \frac{1}{2}[\log p_1(a) + \log p_2(a)]$ combines the logarithmic terms in F. The remaining reference-mass term is $\frac{1}{2} \sum_m \sum_a p_m(a) = 1$, whereas the corresponding term in $D_{\mathrm{I}}(\mathbf{v} \| \mathbf{c})$ is Z. Their difference gives Equation 23. Generalized KL is nonnegative and vanishes exactly at $\mathbf{v} = \mathbf{c}$. □

For the normalized target $\mathbf{r} = \mathbf{c}/Z$, restricting the optimization to the probability simplex gives

$$\frac{1}{2} \sum_{m=1}^{2} D_{\mathrm{KL}}(\mathbf{v} \| \mathbf{p}_m) = D_{\mathrm{KL}}(\mathbf{v} \| \mathbf{r}) - \log Z, \qquad \mathbf{v} \in \Delta^{d-1}. \tag{24}$$

This follows by substituting $\log c(a) = \log r(a) + \log Z$ and using $\sum_a v(a) = 1$. The normalized pool uniquely minimizes the constrained objective, with minimum $-\log Z$. For a strictly positive student probability vector $\mathbf{q}$, the difference between $D_{\mathrm{I}}(\mathbf{c} \| \mathbf{q})$ and the per-position loss $-\mathbf{c}^\top \log \mathbf{q}$ is the constant $\sum_a c(a) \log c(a) - Z + 1$. The two objectives therefore have identical student gradients.

D.3 DERIVATIVES OF THE DISTILLATION LOSS

Let $\mathbf{c} \ge \mathbf{0}$ have mass $Z \le 1$, and let $\mathbf{q} = \mathrm{softmax}(\mathbf{z})$. Rewriting the loss as $\ell(\mathbf{z}) = -\mathbf{c}^\top \mathbf{z} + Z \log \sum_a \exp z(a)$ gives the gradient in Equation 8 and

$$\nabla_{\mathbf{z}}^2 \ell = Z\big[\,\mathrm{diag}(\mathbf{q}) - \mathbf{q}\mathbf{q}^\top\big], \qquad \|\nabla_{\mathbf{z}} \ell\|_1 \le 2Z. \tag{25}$$

Here $\mathrm{diag}(\mathbf{q})$ places the entries of $\mathbf{q}$ on a diagonal matrix. The Hessian is Z times a categorical covariance matrix and is positive semidefinite. The norm bound follows from $\|Z\mathbf{q} - \mathbf{c}\|_1 \le Z\|\mathbf{q}\|_1 + \|\mathbf{c}\|_1 = 2Z$. For $Z > 0$, minimizing the loss over the probability simplex gives $\mathbf{q} = \mathbf{c}/Z$. For $Z = 0$, every student distribution has zero loss.

The parameter gradient follows the chain rule: $\nabla_\theta \ell = (\partial \mathbf{z}/\partial \theta)^\top (Z\mathbf{q} - \mathbf{c})$. The optimizer acts on the sum of these contributions across sampled states. The bound in Equation 25 applies to output-logit gradients. Parameter displacement also depends on the model Jacobian and the optimizer.

D.4 Cross-view setting and scope

We condition on an original prompt and a fixed student-generated prefix, as in Section 4.4. Let Δ_d° denote the strictly positive probability simplex, with $d \geq 2$, and let $p^\star \in \Delta_d^\circ$ be the ideal distribution shared by both views. The fixed original teacher is $p_0 = \mathrm{softmax}(z_0)$ and the random wrapper teacher is $p_T = \mathrm{softmax}(z_0 + \delta)$, with finite logits. All expectations are over wrapper selection at this fixed prefix. The canonical shift is $\delta = P(\log p_T - \log p_0)$, where $P = I_d - \mathbf{1}\mathbf{1}^\mathsf{T}/d$, so $P\delta = \delta$. We use $c_a = \sqrt{p_{0,a} p_{T,a}}$, $Z = \sum_a c_a$, $r = c/Z$, and $\mathcal{E}(q) = D_{\mathrm{KL}}(p^\star \| q)$. For a vector w, write $\mathrm{osc}(w) = \max_a w_a - \min_a w_a$ and $\mathrm{Var}_p(w) = \sum_a p_a (w_a - \sum_j p_j w_j)^2$.

The improvement criterion requires favorable mean correction in addition to task preservation. The envelopes are sharp over distributions with the specified moments; a fixed teacher can admit tighter bounds. The results characterize target error at a fixed prefix. Student convergence, sequence-level accuracy, and the training effect of Z lie outside this target-level analysis. Appendix D.10 characterizes the limits imposed by shared teacher errors.

D.5 Exponential tilting and gauge-invariant curvature

Lemma 3 (Tilted-variance representation)**.** *Let $p \in \Delta_d^\circ$, $w \in \mathbb{R}^d$, and define*

$$K_p(w) = \log \sum_a p_a e^{w_a} - \sum_a p_a w_a, \qquad p_{t,a} = \frac{p_a e^{t w_a}}{\sum_j p_j e^{t w_j}}, \quad 0 \leq t \leq 1. \tag{26}$$

Then

$$K_p(w) = \int_0^1 (1-t)\, \mathrm{Var}_{p_t}(w)\, dt, \qquad 0 \leq K_p(w) \leq \frac{\mathrm{osc}(w)^2}{8} \leq \frac{\|Pw\|_2^2}{4}. \tag{27}$$

Moreover, $K_p(w) = 0$ if and only if $Pw = 0$.

Proof. Set $F(t) = \log \sum_a p_a e^{t w_a}$. The finite alphabet permits differentiation term by term, yielding

$$\begin{aligned} F'(t) &= \sum_a p_{t,a} w_a, \\ F''(t) &= \sum_a p_{t,a} w_a^2 - \left(\sum_a p_{t,a} w_a \right)^2 = \mathrm{Var}_{p_t}(w) \geq 0. \end{aligned} \tag{28}$$

Since $F(0) = 0$, the fundamental theorem of calculus, followed by an interchange of two integrals over a compact triangle, gives

$$\begin{aligned} K_p(w) &= F(1) - F(0) - F'(0) \\ &= \int_0^1 \big(F'(s) - F'(0) \big)\, ds \\ &= \int_0^1 \int_0^s F''(t)\, dt\, ds \\ &= \int_0^1 (1-t)\, \mathrm{Var}_{p_t}(w)\, dt. \end{aligned} \tag{29}$$

Write $m = \min_a w_a$, $M = \max_a w_a$, and $R = M - m$. For any probability vector ν, the mean minimizes squared deviation from a constant. Therefore

$$\begin{aligned} 0 \leq \mathrm{Var}_\nu(w) &= \inf_{a_0 \in \mathbb{R}} \sum_j \nu_j (w_j - a_0)^2 \\ &\leq \sum_j \nu_j \left(w_j - \frac{M+m}{2} \right)^2 \\ &\leq \sum_j \nu_j \frac{(M-m)^2}{4} = \frac{R^2}{4}. \end{aligned} \tag{30}$$

If i, j attain M, m, respectively, then $e_i - e_j$ is orthogonal to $\mathbf{1}$. Cauchy–Schwarz and $\|e_i - e_j\|_2^2 \le 2$ imply

$$\begin{aligned} R^2 &= \big((e_i - e_j)^\mathsf{T} w\big)^2 = \big((e_i - e_j)^\mathsf{T} P w\big)^2 \\ &\le \|e_i - e_j\|_2^2 \|Pw\|_2^2 \le 2\|Pw\|_2^2. \end{aligned} \tag{31}$$

Combining these bounds gives

$$0 \le K_p(w) \le \int_0^1 (1-t)\frac{R^2}{4}\,dt = \frac{R^2}{8} \le \frac{\|Pw\|_2^2}{4}. \tag{32}$$

Every p_t has full support. Consequently, a nonconstant w has strictly positive variance for every t, making the integral strictly positive. A constant w gives $K_p(w) = 0$ directly. This proves the final equivalence. □

D.6 Optimal range and variance envelopes

For $R > 0$ and $0 \le S \le R^2/4$, define

$$\begin{aligned} G_R(u) &= \log(1 + u(e^R - 1)) - Ru, \\ \theta_\pm(R,S) &= \frac{1 \pm \sqrt{1 - 4S/R^2}}{2}, \\ F_-(R,S) &= G_R(\theta_+(R,S)), \qquad F_+(R,S) = G_R(\theta_-(R,S)), \\ \Psi(R) &= \log\frac{e^R - 1}{R} - 1 + \frac{R}{e^R - 1}. \end{aligned} \tag{33}$$

At $R = 0$, the only admissible variance is $S = 0$; set $F_-(0,0) = F_+(0,0) = \Psi(0) = 0$.

Lemma 4 (Exact range–variance envelopes). *For $p \in \Delta_d^\circ$, $w \in \mathbb{R}^d$, $R \ge \operatorname{osc}(w)$, and $S = \operatorname{Var}_p(w)$,*

$$F_-(R,S) \le K_p(w) \le F_+(R,S) \le \Psi(R). \tag{34}$$

For every $R > 0$ and $0 < S \le R^2/4$, the first two envelopes are attained by endpoint-valued perturbations and full-support probability vectors. The range-only envelope is also attainable. In addition, Ψ is nondecreasing on $[0,\infty)$ and $\Psi(R) = R^2/8 + O(R^4)$ as $R \downarrow 0$.

Proof. If $R = 0$, all terms vanish. Suppose $R > 0$.

Step 1: an exact range envelope. Let $X_a = (w_a - \min_j w_j)/R \in [0,1]$ and $\mu = \mathbb{E}_p X \in [0,1]$. Convexity of the exponential gives the pointwise secant inequality $e^{Rx} \le 1 + x(e^R - 1)$ for $x \in [0,1]$. Shift invariance of K_p, this inequality, and monotonicity of the logarithm yield

$$\begin{aligned} K_p(w) &= \log \mathbb{E}_p e^{RX} - R\mu \\ &\le \log \mathbb{E}_p\big[1 + X(e^R - 1)\big] - R\mu \\ &= \log\big(1 + \mu(e^R - 1)\big) - R\mu \\ &\le \max_{0 \le u \le 1} \big\{\log\big(1 + u(e^R - 1)\big) - Ru\big\}. \end{aligned} \tag{35}$$

Write the maximized function as $f_R(u)$. Its derivatives satisfy

$$\begin{aligned} f_R'(u) &= \frac{e^R - 1}{1 + u(e^R - 1)} - R, \\ f_R''(u) &= -\frac{(e^R - 1)^2}{(1 + u(e^R - 1))^2} < 0, \\ f_R'(0) &= e^R - 1 - R > 0, \qquad f_R'(1) = 1 - e^{-R} - R < 0. \end{aligned} \tag{36}$$

Thus the unique maximizer lies in $(0,1)$ and equals

$$u_R = \frac{1}{R} - \frac{1}{e^R - 1}, \qquad f_R(u_R) = \log\frac{e^R - 1}{R} - 1 + \frac{R}{e^R - 1} = \Psi(R). \tag{37}$$

Take two nonempty groups of coordinates, set $w = 0$ on one and $w = R$ on the other, and assign total probability u_R to the second group. Distribute each group mass strictly positively within that group. Every inequality in equation 35 is then an equality. This proves exact attainability for every fixed dimension $d \geq 2$.

For each fixed $u \in [0, 1]$, direct differentiation gives

$$\frac{\partial f_R(u)}{\partial R} = \frac{u(1-u)(e^R - 1)}{1 + u(e^R - 1)} \geq 0. \tag{38}$$

The maximum over u is therefore nondecreasing. Finally, the attainable endpoint construction and Lemma 3 imply $\Psi(R) \leq R^2/8$. Choosing $u = 1/2$ in the same maximization gives

$$\log\cosh\frac{R}{2} = f_R(1/2) \leq \Psi(R) \leq \frac{R^2}{8}, \qquad \log\cosh\frac{R}{2} = \frac{R^2}{8} + O(R^4). \tag{39}$$

The displayed expansion follows by integrating $\tanh x = x + O(x^3)$ at zero; it proves the asserted expansion of Ψ without cancellation of singular terms in its closed form.

Step 2: feasible centered support. We next incorporate the variance constraint. If $S = 0$, full support implies constant w and all joint envelopes are zero. Suppose $S > 0$, and center the perturbation: $Y = w_a - \mathbb{E}_p w$ under $a \sim p$. Write its support endpoints as $-a, b$, where $a, b > 0$ and $a + b \leq R$. Its first two moments satisfy

$$\mathbb{E}Y = 0, \qquad \mathbb{E}Y^2 = S, \qquad 0 \leq \mathbb{E}[(Y + a)(b - Y)] = ab - S. \tag{40}$$

In particular, $S \leq ab \leq R^2/4$. Define

$$A = \frac{R - \sqrt{R^2 - 4S}}{2}, \qquad B = \frac{R + \sqrt{R^2 - 4S}}{2}, \qquad A + B = R, \quad AB = S. \tag{41}$$

The inequalities $S \leq ab \leq b(R - b)$ and $S \leq ab \leq a(R - a)$ give $a, b \in [A, B]$.

Step 3: moment-matched quadratic envelopes. Introduce the entire function

$$k(y) = \int_0^1 (1-t)e^{ty}\,dt, \qquad e^y = 1 + y + y^2 k(y), \qquad k'(y) = \int_0^1 t(1-t)e^{ty}\,dt > 0. \tag{42}$$

For any $u > 0$, let Y_u take the values $-S/u$ and u with probabilities $u^2/(S + u^2)$ and $S/(S + u^2)$, respectively. Then

$$\mathbb{E}Y_u = 0, \qquad \mathbb{E}Y_u^2 = S, \qquad \mathbb{E}e^{Y_u} = M_S(u) := \frac{Se^u + u^2 e^{-S/u}}{S + u^2}. \tag{43}$$

To upper-bound $\mathbb{E}e^Y$, put $\ell = -S/b$. Since $y \leq b$, monotonicity of k gives the quadratic majorant

$$\begin{aligned} e^y &= e^\ell\big[1 + (y - \ell) + (y - \ell)^2 k(y - \ell)\big] \\ &\leq e^\ell\big[1 + (y - \ell) + (y - \ell)^2 k(b - \ell)\big] =: Q_b(y). \end{aligned} \tag{44}$$

Equality holds at both $y = \ell$ and $y = b$. Because Y and Y_b have identical first two moments,

$$\mathbb{E}e^Y \leq \mathbb{E}Q_b(Y) = \mathbb{E}Q_b(Y_b) = \mathbb{E}e^{Y_b} = M_S(b). \tag{45}$$

For the lower bound put $u = S/a$. Since $y \geq -a$, the same monotone kernel gives

$$\begin{aligned} e^y &= e^u\big[1 + (y - u) + (y - u)^2 k(y - u)\big] \\ &\geq e^u\big[1 + (y - u) + (y - u)^2 k(-a - u)\big] =: Q_a^-(y). \end{aligned} \tag{46}$$

This quadratic is exact at $y = -a = -S/u$ and $y = u$. Matching the same moments now yields

$$\mathbb{E}e^Y \geq \mathbb{E}Q_a^-(Y) = \mathbb{E}Q_a^-(Y_{S/a}) = \mathbb{E}e^{Y_{S/a}} = M_S(S/a). \tag{47}$$

Step 4: optimizing the centered endpoints. For $u > 0$, differentiation gives

$$M_S'(u) = \frac{Sue^{-S/u}}{(S + u^2)^2}\big[(s - 2)e^s + s + 2\big], \qquad s = u + S/u > 0. \tag{48}$$

Indeed, if $g(s) = (s-2)e^s + s + 2$, then $g(0) = g'(0) = 0$ and $g''(s) = se^s$. Consequently,

$$g(s) = \int_0^s (s-t)te^t\,dt > 0 \quad (s > 0), \qquad M_S'(u) > 0. \tag{49}$$

Since $a, b \le B$ and $A = S/B$, all inequalities combine into

$$\begin{aligned} M_S(A) = M_S(S/B) &\le M_S(S/a) \\ &\le \mathbb{E}e^Y \\ &\le M_S(b) \le M_S(B). \end{aligned} \tag{50}$$

Using $AB = S$ and $A + B = R$, direct substitution gives

$$\begin{aligned} \log M_S(A) &= G_R(B/R) = G_R(\theta_+(R,S)), \\ \log M_S(B) &= G_R(A/R) = G_R(\theta_-(R,S)). \end{aligned} \tag{51}$$

Taking logarithms in equation 50 therefore proves the joint bounds; $F_+(R,S) \le \Psi(R)$ follows from Step 1.

Step 5: exact joint sharpness. Choose a perturbation taking only 0 and R. Assign total probability $\theta_-(R,S)$ to its high level for the upper bound, or $\theta_+(R,S)$ for the lower bound. In both cases,

$$\mathrm{Var}_p(w) = R^2\theta_\pm(1-\theta_\pm) = S, \qquad K_p(w) = G_R(\theta_\pm). \tag{52}$$

For $0 < S \le R^2/4$, both group probabilities are strictly positive. Splitting each group into positive coordinate masses establishes attainment in every fixed dimension $d \ge 2$. At $S = 0$, a constant perturbation attains zero under the constraint $\mathrm{osc}(w) \le R$. □

D.7 Exact comparison with the original teacher

Proof of Theorem 1. Write $A(z) = \log \sum_a e^{z_a}$ and $w = \delta/2$. The geometric target can be evaluated exactly:

$$\begin{aligned} \log c_a &= z_{0,a} + w_a - \frac{A(z_0) + A(z_0 + 2w)}{2}, \\ \log Z &= A(z_0 + w) - \frac{A(z_0) + A(z_0 + 2w)}{2}, \\ \log r_a &= z_{0,a} + w_a - A(z_0 + w). \end{aligned} \tag{53}$$

Hence $r = \mathrm{softmax}(z_0 + w)$ and $D_{\mathrm{KL}}(p_0\|r) = K_{p_0}(w)$. For any fixed $p \in \Delta_d^\circ$, expansion of the log probabilities gives

$$\begin{aligned} D_{\mathrm{KL}}(p\|r) - D_{\mathrm{KL}}(p\|p_0) &= A(z_0 + w) - A(z_0) - \langle p, w\rangle \\ &= \underbrace{A(z_0 + w) - A(z_0) - \langle p_0, w\rangle}_{D_{\mathrm{KL}}(p_0\|r)} + \langle p_0 - p, w\rangle. \end{aligned} \tag{54}$$

The expectations are well defined. Indeed, Lemma 3 and Cauchy–Schwarz yield

$$\begin{aligned} 0 \le \mathbb{E}D_{\mathrm{KL}}(p_0\|r) &\le \frac{1}{4}\mathbb{E}\|Pw\|_2^2 = \frac{1}{16}\mathbb{E}\|\delta\|_2^2 < \infty, \\ \mathbb{E}|\langle p_0 - p, w\rangle| &\le \|p_0 - p\|_2 \mathbb{E}\|w\|_2 \\ &\le \frac{1}{2}\|p_0 - p\|_2 \sqrt{\mathbb{E}\|\delta\|_2^2} < \infty. \end{aligned} \tag{55}$$

Taking $p = p^\star$ and expectations in equation 54 therefore gives

$$\begin{aligned} \mathbb{E}\mathcal{E}(r) - \mathcal{E}(p_0) &= \mathbb{E}D_{\mathrm{KL}}(p_0\|r) + \frac{1}{2}\langle p_0 - p^\star, \mathbb{E}\delta\rangle \\ &= \mathcal{C} - \mathcal{G}. \end{aligned} \tag{56}$$

The necessary and sufficient condition follows directly. □

Corollary 5 (Mean correction and centered wrapper noise)**.** *Under the assumptions of Theorem 1, let* $p_\mu = \mathrm{softmax}(z_0 + \mu/2)$*. Then*

$$\begin{aligned} \mathbb{E}\mathcal{E}(r) &= \mathcal{E}(p_\mu) + \mathbb{E}D_{\mathrm{KL}}(p_\mu \| r), \\ \mathcal{C} &= D_{\mathrm{KL}}(p_0 \| p_\mu) + \mathbb{E}D_{\mathrm{KL}}(p_\mu \| r). \end{aligned} \tag{57}$$

Thus centered fluctuations around the mean wrapper target add nonnegative error. If $\mu = 0$*, pooling cannot improve upon* p_0 *and strictly worsens expected forward KL whenever* $\Pr(\delta \neq 0) > 0$*.*

Proof. Put $\bar{w} = \mu/2$, $e = (\delta - \mu)/2$, and $p_\mu = \mathrm{softmax}(z_0 + \bar{w})$. For any fixed $p \in \Delta_d^\circ$, apply the same expansion at $z_0 + \bar{w}$:

$$D_{\mathrm{KL}}(p\|r) - D_{\mathrm{KL}}(p\|p_\mu) = D_{\mathrm{KL}}(p_\mu\|r) + \langle p_\mu - p, e\rangle. \tag{58}$$

Since $\mathbb{E}e = 0$ and e has finite second moment, the same integrability argument applies. Taking expectations with $p = p^\star$ and $p = p_0$, respectively, gives

$$\begin{aligned} \mathbb{E}\mathcal{E}(r) &= \mathcal{E}(p_\mu) + \mathbb{E}D_{\mathrm{KL}}(p_\mu \| r), \\ \mathcal{C} &= D_{\mathrm{KL}}(p_0 \| p_\mu) + \mathbb{E}D_{\mathrm{KL}}(p_\mu \| r). \end{aligned} \tag{59}$$

If $\mu = 0$, then $p_\mu = p_0$ and the expected risk difference is $\mathcal{C} \geq 0$. By Lemma 3, its integrand vanishes precisely when $P\delta = 0$. Since $P\delta = \delta$, this is equivalent to $\delta = 0$. A nonnegative random variable that is positive with positive probability has strictly positive expectation. This proves the strict deterioration claim. □

D.8 Distribution-sensitive curvature and its tightness

Define $\phi_+(L) = (e^L - 1 - L)/L^2$ and $\phi_-(L) = (L - 1 + e^{-L})/L^2$ for $L > 0$, with $\phi_+(0) = \phi_-(0) = 1/2$. The following bounds depend on the original teacher's categorical curvature.

Corollary 6 (Distribution-sensitive cost of a wrapper)**.** *Suppose* $\mathrm{osc}(\delta/2) \leq L$ *almost surely and let* $W = \mathbb{E}\,\mathrm{Var}_{p_0}(\delta)$*. Then* $\phi_-(L)W/4 \leq \mathcal{C} \leq \phi_+(L)W/4$ *and* $|\mathcal{C} - W/8| \leq Le^L W/24$*. For every* $L > 0$*, the coefficients* $\phi_-(L)$ *and* $\phi_+(L)$ *cannot be improved uniformly over full-support teacher distributions and wrapper perturbations satisfying the range constraint. In particular,* $\mathcal{G} > \phi_+(L)W/4$ *guarantees improvement, whereas* $\mathcal{G} < \phi_-(L)W/4$ *guarantees deterioration.*

Writing $H_0 = \mathrm{diag}(p_0) - p_0 p_0^\mathsf{T}$ *and* $\Sigma = \mathrm{Cov}(\delta)$ *gives* $W = \mu^\mathsf{T} H_0 \mu + \mathrm{tr}(H_0\Sigma)$*. The exact, jointly optimal range–variance envelopes in Lemma 4 further sharpen these bounds without changing the method.*

Proof of Corollary 6. First consider a deterministic w with $\mathrm{osc}(w) \leq L$. Use the tilted probabilities p_t from Lemma 3 and put $S = \mathrm{Var}_p(w)$. Bounds on the normalizing constant give

$$e^{t\min w} \leq \sum_j p_j e^{tw_j} \leq e^{t\max w}, \qquad e^{-tL} \leq \frac{p_{t,a}}{p_a} \leq e^{tL}. \tag{60}$$

Let $m_t = \mathbb{E}_{p_t} w$ and $m_0 = \mathbb{E}_p w$. Using the variational characterization of variance on both sides of the comparison yields

$$\begin{aligned} e^{-tL} S = e^{-tL} \inf_c \sum_a p_a (w_a - c)^2 \\ &\leq e^{-tL} \sum_a p_a (w_a - m_t)^2 \\ &\leq \sum_a p_{t,a} (w_a - m_t)^2 = \mathrm{Var}_{p_t}(w) \\ &\leq \sum_a p_{t,a} (w_a - m_0)^2 \\ &\leq e^{tL} \sum_a p_a (w_a - m_0)^2 = e^{tL} S. \end{aligned} \tag{61}$$

Consequently, with the continuous values at $L = 0$,

$$\begin{aligned}\phi_-(L)S &= S\int_0^1 (1-t)e^{-tL}\,dt \\ &\le \int_0^1 (1-t)\,\mathrm{Var}_{p_t}(w)\,dt = K_p(w) \\ &\le S\int_0^1 (1-t)e^{tL}\,dt = \phi_+(L)S.\end{aligned} \tag{62}$$

Apply this pointwise with $p = p_0$ and $w = \delta/2$. Since $\mathrm{Var}_{p_0}(\delta/2) = \mathrm{Var}_{p_0}(\delta)/4$, expectation gives

$$\frac{\phi_-(L)}{4}W \le \mathcal{C} \le \frac{\phi_+(L)}{4}W. \tag{63}$$

For the explicit quadratic remainder, observe that

$$\begin{aligned}0 \le \frac{1}{2} - \phi_-(L) &\le \phi_+(L) - \frac{1}{2} \\ &= \int_0^1 (1-t)(e^{tL}-1)\,dt \\ &\le \int_0^1 (1-t)tLe^L\,dt = \frac{Le^L}{6}.\end{aligned} \tag{64}$$

The first comparison follows from $e^{tL} + e^{-tL} \ge 2$; the last uses $e^{tL} - 1 = \int_0^{tL} e^s\,ds \le tLe^L$. Combining equation 63 and equation 64 gives

$$\left|\mathcal{C} - \frac{W}{8}\right| \le \frac{Le^L}{24}W. \tag{65}$$

The improvement and deterioration conditions follow by substitution in equation 56.

To identify the distribution dependence, write $H_0 = \mathrm{diag}(p_0) - p_0p_0^\mathsf{T}$. Then $\mathrm{Var}_{p_0}(\delta) = \delta^\mathsf{T}H_0\delta$. Expanding $\delta = \mu + (\delta - \mu)$ and using centering,

$$\begin{aligned}W &= \mathbb{E}\delta^\mathsf{T}H_0\delta \\ &= \mu^\mathsf{T}H_0\mu + 2\mu^\mathsf{T}H_0\mathbb{E}(\delta-\mu) + \mathbb{E}(\delta-\mu)^\mathsf{T}H_0(\delta-\mu) \\ &= \mu^\mathsf{T}H_0\mu + \mathrm{tr}(H_0\Sigma).\end{aligned} \tag{66}$$

Sharpness of the multiplicative coefficients. Fix $L > 0$, let $w = (0, L)$, and choose either $p_\epsilon^+ = (1-\epsilon, \epsilon)$ or $p_\epsilon^- = (\epsilon, 1-\epsilon)$, where $0 < \epsilon < 1$. For a deterministic wrapper use $\delta = 2Pw$; both K_p and variance are invariant to constant shifts. Direct calculation gives

$$\begin{aligned}\mathrm{Var}_{p_\epsilon^\pm}(w) &= \epsilon(1-\epsilon)L^2, \\ K_{p_\epsilon^+}(w) &= \log(1+\epsilon(e^L-1)) - \epsilon L, \\ K_{p_\epsilon^-}(w) &= \log(1+\epsilon(e^{-L}-1)) + \epsilon L, \\ \lim_{\epsilon\downarrow 0}\frac{K_{p_\epsilon^\pm}(w)}{\mathrm{Var}_{p_\epsilon^\pm}(w)} &= \phi_\pm(L).\end{aligned} \tag{67}$$

These limits follow from $\log(1+x)/x \to 1$. They prohibit a smaller upper coefficient or a larger lower coefficient for every prescribed $L > 0$. Splitting the two masses into positive coordinate groups gives the same construction in any fixed dimension $d \ge 2$.

Sharper joint envelopes. For each wrapper let $R_\delta = \mathrm{osc}(\delta/2)$ and $S_\delta = \mathrm{Var}_{p_0}(\delta)/4$. Lemma 4 yields

$$\begin{aligned}\mathbb{E}F_-(R_\delta, S_\delta) \le \mathcal{C} &\le \mathbb{E}F_+(R_\delta, S_\delta) \\ &\le \mathbb{E}\Psi(R_\delta) \le \frac{1}{16}\mathbb{E}\|\delta\|_2^2.\end{aligned} \tag{68}$$

These bounds are finite under the second-moment assumption of Theorem 1; a uniform L is unnecessary. Their sharpness is over distributions with the specified range and variance. A fixed teacher p_0 can admit tighter bounds. □

Proposition 7 (Local Fisher expansion with only second moments). *Let $\delta = tX$, where $PX = X$ and $\mathbb{E}\|X\|_2^2 < \infty$, and let $t \to 0$. For $r_t = \mathrm{softmax}(z_0 + tX/2)$,*

$$\begin{aligned} \mathbb{E}D_{\mathrm{KL}}(p_0\|r_t) &= \frac{t^2}{8}\mathbb{E}\,\mathrm{Var}_{p_0}(X) + o(t^2), \\ \mathbb{E}\mathcal{E}(r_t) - \mathcal{E}(p_0) &= \frac{t}{2}\langle p_0 - p^\star, \mathbb{E}X\rangle + \frac{t^2}{8}\mathbb{E}\,\mathrm{Var}_{p_0}(X) + o(t^2). \end{aligned} \tag{69}$$

Proof. For each realization of X, Lemma 3 gives

$$\begin{aligned} \frac{K_{p_0}(tX/2)}{t^2} &= \frac{1}{4}\int_0^1 (1-s)\,\mathrm{Var}_{p_{s,t}}(X)\,ds, \qquad p_{s,t,a} = \frac{p_{0,a}e^{stX_a/2}}{\sum_j p_{0,j}e^{stX_j/2}}, \\ \lim_{t\to 0}\frac{K_{p_0}(tX/2)}{t^2} &= \frac{1}{8}\,\mathrm{Var}_{p_0}(X). \end{aligned} \tag{70}$$

The limit uses continuity on a finite alphabet and a bounded integrand for each fixed X. For passage through expectation, the same lemma supplies the integrable domination

$$0 \le \frac{K_{p_0}(tX/2)}{t^2} \le \frac{\|PX\|_2^2}{16} = \frac{\|X\|_2^2}{16}. \tag{71}$$

Dominated convergence proves the first expansion. The exact original-view identity proves the second. □

D.9 Exact minimax risk under bounded binary shared bias

Shared bias can make agreeing teacher views inaccurate. The following result quantifies the minimax error of estimating the ideal distribution from these views under a bounded shared bias.

Theorem 8 (Exact minimax risk under bounded shared bias). *Consider $d = 2$, $p^\star = \mathrm{softmax}(z^\star)$, and identical teacher views $p_0 = p_T = \mathrm{softmax}(z^\star + b)$, where $z^\star, b \in \mathbb{R}^2$ and $\mathrm{osc}(b) \le B$. An estimator $\widehat{p}$ observes only (p_0, p_T) and may randomize. Define its minimax forward-KL risk as $\mathfrak{R}(B) = \inf_{\widehat{p}}\sup_{z^\star, b:\,\mathrm{osc}(b)\le B}\mathbb{E}D_{\mathrm{KL}}(p^\star\|\widehat{p}(p_0,p_T))$. For every $B \ge 0$, $\mathfrak{R}(B) = \frac{B}{2}\tanh(B/2) - \log\cosh(B/2)$. In particular, $\mathfrak{R}(B) = B^2/8 + O(B^4)$ near zero, although $Z = 1$ for every admissible observation.*

Throughout this subsection, logarithms are natural, and a Bernoulli parameter denotes the probability of the first class. Define

$$\begin{aligned} \sigma(s) &= \frac{e^s}{1+e^s}, & A(s) &= \log(1+e^s), \\ F(p) &= p\log p + (1-p)\log(1-p), & \mathrm{kl}(p,q) &= F(p) - p\log q - (1-p)\log(1-q). \end{aligned} \tag{72}$$

Here $p \in (0,1)$, while $q \in [0,1]$ is allowed, with infinite KL at either endpoint. The identities used below follow by differentiation:

$$\begin{aligned} A'(s) &= \sigma(s), & \sigma'(s) &= \frac{1}{4\cosh^2(s/2)}, \\ F'(p) &= \log\frac{p}{1-p}, & F''(p) &= \frac{1}{p(1-p)} > 0, \\ \mathrm{kl}(p,\sigma(s)) &= F(p) - ps + A(s). \end{aligned} \tag{73}$$

Proof of Theorem 8. **Step 1: reduction to an observable logit interval.** In the identical-view submodel, both observed teachers are equal. Let their common first-class probability be $\sigma(t)$, where $t \in \mathbb{R}$ is observable. Writing $\theta = z_1^\star - z_2^\star$ and $\gamma = b_1 - b_2$, the binary model and bias constraint become

$$t = \theta + \gamma, \qquad |\gamma| = \mathrm{osc}(b) \le B, \qquad p_1^\star = \sigma(\theta) \in [\sigma(t-B), \sigma(t+B)]. \tag{74}$$

Conversely, every $\theta \in [t-B, t+B]$ is compatible with this same observed teacher: choose $z^\star = (\theta/2, -\theta/2)$ and $b = ((t-\theta)/2, -(t-\theta)/2)$. Thus equation 74 gives the exact compatible interval.

Randomization cannot improve the minimax risk. Indeed, for a random estimator output $Q_t \in [0,1]$, put $\bar{q}_t = \mathbb{E}Q_t$. Convexity of the negative logarithm gives, for every compatible $p \in (0,1)$,

$$\begin{aligned}\mathbb{E}\,\mathrm{kl}(p, Q_t) &= F(p) - p\mathbb{E}\log Q_t - (1-p)\mathbb{E}\log(1-Q_t) \\ &\geq F(p) - p\log\mathbb{E}Q_t - (1-p)\log\mathbb{E}(1-Q_t) \\ &= \mathrm{kl}(p, \bar{q}_t).\end{aligned} \tag{75}$$

The inequality remains valid with extended expectations if the risk is infinite. We can consequently restrict attention to deterministic functions of t.

If $B = 0$, the ideal parameter is known exactly from t, and the estimator $\sigma(t)$ has zero risk. The stated formula also equals zero at $B = 0$. In the remaining steps assume $B > 0$ and set

$$a = t - B, \qquad b = t + B, \qquad l = \sigma(a), \qquad u = \sigma(b), \qquad 0 < l < u < 1. \tag{76}$$

Step 2: the exact information radius of an interval. Since $\mathrm{kl}(p,q)$ is convex in p, for any $p = (1-\lambda)l + \lambda u$ and $\lambda \in [0,1]$,

$$\begin{aligned}\mathrm{kl}(p,q) &\leq (1-\lambda)\mathrm{kl}(l,q) + \lambda\mathrm{kl}(u,q) \\ &\leq \max\{\mathrm{kl}(l,q), \mathrm{kl}(u,q)\}.\end{aligned} \tag{77}$$

It therefore suffices to minimize the maximum of the two endpoint losses. Define the secant slope and its associated Bernoulli parameter by

$$k(t) = \frac{F(u) - F(l)}{u - l}, \qquad m(t) = \sigma(k(t)), \qquad C(t) = \mathrm{kl}(l, m(t)). \tag{78}$$

Strict convexity of F implies $a = F'(l) < k(t) < F'(u) = b$, hence $l < m(t) < u$. The two endpoint losses at $m(t)$ are equal because

$$\begin{aligned}\mathrm{kl}(u,m) - \mathrm{kl}(l,m) &= F(u) - F(l) - (u-l)\log\frac{m}{1-m} \\ &= F(u) - F(l) - (u-l)k(t) = 0.\end{aligned} \tag{79}$$

Put $\alpha = (u-m)/(u-l) \in (0,1)$, so that $\alpha l + (1-\alpha)u = m$. For any $q \in [0,1]$, direct expansion and Jensen's inequality give the following lower-bound chain:

$$\begin{aligned}\max\{\mathrm{kl}(l,q), \mathrm{kl}(u,q)\} &\geq \alpha\mathrm{kl}(l,q) + (1-\alpha)\mathrm{kl}(u,q) \\ &= \alpha\mathrm{kl}(l,m) + (1-\alpha)\mathrm{kl}(u,m) + \mathrm{kl}(m,q) \\ &= C(t) + \mathrm{kl}(m,q) \\ &\geq C(t).\end{aligned} \tag{80}$$

The last inequality follows from

$$\begin{aligned}\mathrm{kl}(m,q) &= -m\log\frac{q}{m} - (1-m)\log\frac{1-q}{1-m} \\ &\geq -\log\left(m\frac{q}{m} + (1-m)\frac{1-q}{1-m}\right) = 0.\end{aligned} \tag{81}$$

At $q = m(t)$, equality is attained in the minimax problem by equation 77 and equation 79. Consequently,

$$\inf_{q\in[0,1]}\sup_{p\in[l,u]}\mathrm{kl}(p,q) = C(t) = F(l) - lk(t) + A(k(t)). \tag{82}$$

All functions in this display are continuous in t for fixed $B > 0$, so $m(t)$ defines an admissible measurable estimator.

Step 3: the secant logit is displaced toward zero. We prove $k(t) \leq t$ whenever $t \geq 0$. Changing variables $p = \sigma(s)$ in the secant integral yields

$$\begin{aligned}k(t) &= \frac{\int_l^u F'(p)\,dp}{u-l} \\ &= \frac{\int_{t-B}^{t+B} s\sigma'(s)\,ds}{\int_{t-B}^{t+B}\sigma'(s)\,ds}, \\ k(t) - t &= \frac{\int_0^B v\,[\sigma'(t+v) - \sigma'(t-v)]\,dv}{\int_{t-B}^{t+B}\sigma'(s)\,ds}.\end{aligned} \tag{83}$$

The function $\sigma'(s) = 1/(4\cosh^2(s/2))$ is even and nonincreasing as $|s|$ grows. Since $|t+v| \geq |t-v|$ for $t, v \geq 0$, every integrand in the last numerator is nonpositive. Hence

$$t \geq 0 \quad \Longrightarrow \quad a < k(t) \leq t < b, \qquad 0 < \frac{k(t)-a}{2} \leq \frac{b-k(t)}{2}. \tag{84}$$

This ordering enables an exact evaluation of the information radius.

Step 4: the worst observed teacher is uniform. We next establish $C'(t) \leq 0$ on $[0, \infty)$. Primes in this step denote differentiation with respect to t. Since $a' = b' = 1$, we have $l' = \sigma'(a)$ and $u' = \sigma'(b)$. Differentiating the secant quotient and then equation 82 gives

$$\begin{aligned} k'(t) &= \frac{u'(b-k) - l'(a-k)}{u-l}, \\ C'(t) &= l'(a-k) + (m-l)k'(t) \\ &= \frac{(u-m)l'(a-k) + (m-l)u'(b-k)}{u-l}. \end{aligned} \tag{85}$$

Let $x = (k-a)/2$ and $y = (b-k)/2$. By equation 84, $0 < x \leq y$ when $t \geq 0$. The elementary identity

$$\sigma(s_2) - \sigma(s_1) = \frac{\sinh((s_2-s_1)/2)}{2\cosh(s_2/2)\cosh(s_1/2)} \qquad (s_1 < s_2) \tag{86}$$

follows by subtracting the two logistic fractions. Applying it twice, and using equation 73, yields

$$\begin{aligned} \frac{m-l}{u-m} &= \frac{\sinh x}{\sinh y}\frac{\cosh(b/2)}{\cosh(a/2)}, \\ \frac{u'}{l'} &= \frac{\cosh^2(a/2)}{\cosh^2(b/2)}. \end{aligned} \tag{87}$$

The function $s \mapsto \sinh(s)/s$ is increasing for $s > 0$, because

$$\frac{d}{ds}\frac{\sinh s}{s} = \frac{s\cosh s - \sinh s}{s^2} = \frac{1}{s^2}\int_0^s v\sinh v\, dv > 0. \tag{88}$$

Moreover, $|a| = |t-B| \leq t+B = |b|$ for $t \geq 0$. The ratio between the positive and negative terms in the numerator of equation 85 therefore satisfies

$$\begin{aligned} \frac{(m-l)u'(b-k)}{(u-m)l'(k-a)} &= \frac{m-l}{u-m}\frac{u'}{l'}\frac{y}{x} \\ &= \frac{\sinh x/x}{\sinh y/y}\frac{\cosh(a/2)}{\cosh(b/2)} \\ &\leq \frac{\cosh(a/2)}{\cosh(b/2)} \\ &\leq 1. \end{aligned} \tag{89}$$

Every denominator is strictly positive. Thus the numerator of equation 85 is nonpositive, proving $C'(t) \leq 0$ for $t \geq 0$.

Reflection changes l and u into $1-u$ and $1-l$, respectively. Since $F(1-p) = F(p)$, it changes k to $-k$ and m to $1-m$. Bernoulli KL is invariant under this simultaneous reflection, so $C(-t) = C(t)$. Together with the derivative bound, this proves

$$\sup_{t\in\mathbb{R}} C(t) = C(0). \tag{90}$$

Step 5: matching global upper and lower bounds. At $t = 0$ the interval endpoints are $\sigma(-B)$ and $\sigma(B)$; their equal entropies give $k(0) = 0$ and $m(0) = 1/2$. Put $x = B/2$. The endpoint probabilities are $e^{-x}/(2\cosh x)$ and $e^{x}/(2\cosh x)$, so

$$\begin{aligned} C(0) &= \mathrm{kl}(\sigma(B), 1/2) \\ &= \frac{e^x}{2\cosh x}\big(x - \log\cosh x\big) + \frac{e^{-x}}{2\cosh x}\big(-x - \log\cosh x\big) \\ &= x\tanh x - \log\cosh x \\ &= \frac{B}{2}\tanh\frac{B}{2} - \log\cosh\frac{B}{2} =: L(B). \end{aligned} \tag{91}$$

The estimator that returns $(m(t), 1-m(t))$ has risk at most $C(t) \le C(0) = L(B)$ for every finite observed logit t and every compatible ideal target. This establishes the global upper bound.

For the lower bound, fix the observed teacher to be uniform. Choose the two parameter pairs $z^{\star,+} = (B/2, -B/2)$, $b^+ = -z^{\star,+}$ and $z^{\star,-} = (-B/2, B/2)$, $b^- = -z^{\star,-}$. Both have $\operatorname{osc}(b^{\pm}) = B$ and produce the same observed pair of uniform teachers. For the estimator output Q_0, equality of the two endpoint losses at $1/2$ gives

$$\begin{aligned}\max_{p\in\{\sigma(-B),\sigma(B)\}} \mathbb{E}\,\mathrm{kl}(p, Q_0) &\ge \frac{1}{2}\mathbb{E}\big[\mathrm{kl}(\sigma(-B), Q_0) + \mathrm{kl}(\sigma(B), Q_0)\big] \\ &= L(B) + \mathbb{E}\,\mathrm{kl}(1/2, Q_0) \\ &\ge L(B).\end{aligned} \tag{92}$$

The identity holds pointwise before expectation, and includes outputs with infinite loss. Taking the infimum over all estimators and combining both bounds proves

$$\mathfrak{R}(B) = \frac{B}{2}\tanh\frac{B}{2} - \log\cosh\frac{B}{2} \qquad \text{for every } B \ge 0. \tag{93}$$

Step 6: local sharpness and boundary behavior. Differentiating $x\tanh x - \log\cosh x$ gives $x\operatorname{sech}^2 x$, and its value at zero is zero. Taylor's theorem applied to the smooth function $\operatorname{sech}^2 x$ yields $\operatorname{sech}^2 x = 1 - x^2 + O(x^4)$ near zero. Hence

$$\begin{aligned}\mathfrak{R}(B) &= \int_0^{B/2} x\operatorname{sech}^2 x\, dx \\ &= \int_0^{B/2} \big(x - x^3 + O(x^5)\big)\, dx \\ &= \frac{B^2}{8} - \frac{B^4}{64} + O(B^6), \\ \lim_{B\downarrow 0}\frac{\mathfrak{R}(B)}{B^2} &= \frac{1}{8}.\end{aligned} \tag{94}$$

This establishes the exact risk at every radius and its attained small-bias coefficient. For every finite B, the lower-bound construction uses finite latent logits and strictly positive ideal targets. □

D.10 Limits of agreement and unrestricted shared bias

The classical average-teacher KL decomposition for logarithmic pooling (Heskes, 1997) describes the reduction in average target error. The following examples show how shared bias limits its relation to correctness.

Proposition 9 (Agreement is not a correctness certificate)**.** *For any positive teachers p_0, p_T and ideal target $p^\star$,*

$$\mathcal{E}(r) = \frac{\mathcal{E}(p_0) + \mathcal{E}(p_T)}{2} + \log Z. \tag{95}$$

Nevertheless, already for $d = 2$, $Z = 1$ can coexist with arbitrarily large $\mathcal{E}(r)$, and arbitrarily small Z can coexist with $\mathcal{E}(r) = 0$. Furthermore, in the identical-view model $p_0 = p_T = \operatorname{softmax}(z^\star + b)$ with unrestricted finite shared bias,

$$\inf_{\widehat{p}} \sup_{z^\star, b\in\mathbb{R}^d} \mathbb{E} D_{\mathrm{KL}}(p^\star \| \widehat{p}(p_0, p_T)) = \log d, \qquad p^\star = \operatorname{softmax}(z^\star). \tag{96}$$

The infimum permits randomized estimators.

Proof. The logarithmic-pool identity follows directly from

$$\begin{aligned}D_{\mathrm{KL}}(p^\star\|r) &= \sum_a p_a^\star\left[\log p_a^\star - \tfrac{1}{2}\log p_{0,a} - \tfrac{1}{2}\log p_{T,a} + \log Z\right] \\ &= \tfrac{1}{2}D_{\mathrm{KL}}(p^\star\|p_0) + \tfrac{1}{2}D_{\mathrm{KL}}(p^\star\|p_T) + \log Z.\end{aligned} \tag{97}$$

For $0 < \epsilon < 1/2$, take $p^\star = (1-\epsilon, \epsilon)$ and $p_0 = p_T = (\epsilon, 1-\epsilon)$. Then

$$Z = 1, \qquad \mathcal{E}(r) = (1-2\epsilon)\log\frac{1-\epsilon}{\epsilon} \longrightarrow \infty. \tag{98}$$

Conversely, take $p^\star = (1/2, 1/2)$, $p_0 = (1-\epsilon, \epsilon)$, and $p_T = (\epsilon, 1-\epsilon)$. Direct multiplication gives

$$Z = 2\sqrt{\epsilon(1-\epsilon)} \longrightarrow 0, \qquad r = (1/2, 1/2) = p^\star, \qquad \mathcal{E}(r) = 0. \tag{99}$$

All distributions have finite logits for every fixed $\epsilon > 0$.

For the minimax statement, fix any observed $s \in \Delta_d^\circ$. Every $p \in \Delta_d^\circ$ is compatible with the same observations (s, s) by choosing

$$z^\star = \log p, \qquad b = \log s - \log p, \qquad p_0 = p_T = s. \tag{100}$$

Let Q be the possibly random estimator output at (s, s). For $u = \mathbf{1}/d$ and $0 < \epsilon < 1$, put $p^{(i,\epsilon)} = (1-\epsilon)e_i + \epsilon u$. These d positive targets have equal entropy H_ϵ, average u, and $H_\epsilon \to 0$. If any coordinate of Q is zero with positive probability, the risk against each positive target is infinite. Otherwise Jensen's inequality gives

$$\begin{aligned}
\sup_{p\in\Delta_d^\circ} \mathbb{E}D_{\mathrm{KL}}(p\|Q) &\geq \frac{1}{d}\sum_i \mathbb{E}D_{\mathrm{KL}}(p^{(i,\epsilon)}\|Q) \\
&= -H_\epsilon - \mathbb{E}\left[\frac{1}{d}\sum_j \log Q_j\right] \\
&\geq -H_\epsilon - \mathbb{E}\log\left(\frac{1}{d}\sum_j Q_j\right) \\
&= \log d - H_\epsilon.
\end{aligned} \tag{101}$$

The calculation is valid with extended cross-entropies as well. Letting $\epsilon \downarrow 0$ proves the lower bound for every estimator. The constant uniform estimator attains the upper bound:

$$D_{\mathrm{KL}}(p\|u) = \log d - H(p) \leq \log d. \tag{102}$$

This establishes the exact minimax risk under unrestricted shared bias. A bias constraint yields the sharper result proved above. □