VISITATION IS NOT SUPERVISION: D-OPTIMALITY-GUIDED ALLOCATION FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains a student with dense teacher feedback along its own generated trajectories. Yet the way these token-level signals are aggregated implicitly decides how the supervision budget is distributed. Standard token averaging assigns more total weight to problems that generate longer responses, and repeated visits to similar reasoning states can continue to accumulate weight even when they induce redundant updates. In this paper, we propose a D-Optimality-guided framework that hierarchically allocates supervision across problems and reasoning states, termed **DOpt-OPD**. At the problem level, we normalize the supervision so that each problem receives equal total loss weight, removing the direct dependence of problem-level supervision on generated length. Within each problem, we pool reasoning spans across rollouts and redistribute the fixed supervision budget using ridge-leverage scores derived from a regularized D-optimal objective. This one-step allocation reduces redundant weighting of similar update directions under a fixed loss-weight budget. Our theoretical analysis shows that the one-step update is near-optimal for small ridge values under orthogonal direction groups, and further provides a computable upper bound for its optimality gap under general, non-orthogonal directions. Across two teacher–student scales and seven mathematical reasoning benchmarks, DOpt-OPD achieves the highest macro-averaged Avg@8 and Pass@8 among the evaluated baselines. These results highlight supervision allocation as a key design choice in on-policy distillation.

1 INTRODUCTION

Modern LLM post-training increasingly integrates supervised fine-tuning, reinforcement learning, and distillation to acquire, specialize, and consolidate model capabilities. Agarwal et al. (2024) formulate on-policy distillation (OPD), in which a student learns from token-level teacher feedback along its own generated trajectories. Recent reports on Qwen3 (Yang et al., 2025), MiMo-V2-Flash (Xiaomi LLM-Core Team, 2026), and GLM-5 (GLM-5 Team, 2026) illustrate its role in strong-to-weak transfer, the integration of specialized teachers, and capability preservation across successive post-training stages. Compared with distillation on fixed teacher-generated data, OPD better aligns training states with the student's inference distribution. Unlike optimization based solely on outcome rewards, it provides dense supervision throughout each generated trajectory.

However, access to dense token-level supervision does not alone determine how these signals should be aggregated into a training objective. A common choice is token averaging, which averages losses over all valid tokens in the sampled batch, assigning each token the same scalar loss coefficient. Prior work shows that teacher supervision's usefulness can vary with token position, trajectory mismatch, student uncertainty, teacher–student disagreement, and whether a correction is learnable by the student (Xie et al., 2026a; Xu et al., 2026; Wang et al., 2026). Other studies identify complementary issues, including reverse-KL mode seeking, degenerate agreement, asymmetric teacher–student mismatch, and unstable or low-quality trajectories (Jin et al., 2026; Yu et al., 2026; Luo et al., 2026). These findings motivate trajectory filtering, token selection or reweighting, and modifications to the distillation objective (Li et al., 2026b; Xie et al., 2026b; Jin et al., 2026; Yu et al., 2026).

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.

We study a complementary question: how should a fixed supervision budget be allocated across problems and collectively across the states visited for each problem? Under token averaging, a problem receives total loss weight proportional to the number of valid tokens generated for it. Longer responses therefore receive a larger share of the supervision budget, even when length is unrelated to supervision quality. Moreover, even after problem-level budgets are equalized, repeatedly visited states can continue to accumulate weight without accounting for whether their update directions are already represented by other states. We therefore distinguish two levels of supervision allocation: determining each problem's total budget and distributing that budget among its visited states while accounting for redundancy among their update directions.

We propose a D-Optimality-guided framework that hierarchically allocates supervision across problems and reasoning states, termed DOpt-OPD. At the problem level, we normalize supervision to give each problem equal total loss weight, removing the direct dependence of problem-level supervision on generated length. Within each problem, we pool reasoning spans across rollouts and perform D-optimal allocation under the fixed supervision budget. The log-determinant objective measures the collective coverage of state-conditioned gradient-proxy directions, so directions already well represented by other spans receive less marginal allocation. Ridge-leverage scores derived from a regularized D-optimal objective provide the one-step allocation weights. This one-step allocation therefore reduces redundant weighting of similar update directions under a fixed problem budget.

We analyze this one-step allocation theoretically and evaluate its effectiveness empirically. Under orthogonal direction groups and sufficiently small ridge values, the one-step update closely approximates the optimal D-optimal allocation; for general, non-orthogonal directions, we provide a computable upper bound on its optimality gap. These results connect the practical one-step allocation to the underlying D-optimal design objective. Empirically, we evaluate DOpt-OPD across two teacher–student scales and seven mathematical reasoning benchmarks against standard OPD and recent baselines. Problem normalization and uniform span allocation controls, hyperparameter sensitivity studies, and recurrent-motif and repeated-suffix diagnostics further examine the proposed allocation mechanism. Across the tested settings, DOpt-OPD achieves the highest macro-averaged Avg@8 and Pass@8 among the evaluated baselines.

We make three contributions:

- **Hierarchical Supervision Allocation.** We identify two allocation issues in OPD: token averaging assigns larger supervision budgets to problems with longer responses, while repeated visits within a problem can overcount redundant update directions. This motivates allocating supervision at both the problem and state levels.
- **D-Optimality-Guided On-Policy Distillation.** We propose DOpt-OPD, which equalizes problem budgets and allocates within-problem supervision using a one-step ridge-leverage update. We establish near-optimality under orthogonal direction groups and a computable optimality-gap bound for general directions.
- **Empirical Gains and Analysis.** Across two teacher–student scales and seven reasoning benchmarks, DOpt-OPD achieves the best macro-averaged Avg@8 and Pass@8 among evaluated baselines. Component ablations, length-stratified analyses, and redundancy diagnostics support the allocation design and examine where its benefits arise.

2 Preliminaries

2.1 On-Policy Distillation

Following Agarwal et al. (2024) and the OPD analysis of Li et al. (2026a), we consider an on-policy distillation setting in which the student is supervised on states visited by its own generations rather than only on teacher-forced prefixes. Let $\{x_i\}_{i=1}^{N}$ be a batch of problem prompts. For prompt i, the student p_θ generates R rollouts $y_{ir} = (y_{ir1}, \ldots, y_{irT_{ir}})$, where $T_{ir} > 0$ counts valid response tokens. The state at token t is $s_{irt} = (x_i, y_{ir,<t})$, with student and teacher distributions $p_\theta(\cdot \mid s_{irt})$ and $p_{\mathrm{T}}(\cdot \mid s_{irt})$.

In the reverse-KL OPD formulation used in our setup, the local objective at a student-visited state s is

$$d_{\mathrm{RKL}}(\theta; s) = D_{\mathrm{KL}}(p_\theta(\cdot \mid s) \,\|\, p_{\mathrm{T}}(\cdot \mid s)) = \mathbb{E}_{v \sim p_\theta(\cdot|s)} \left[\log \frac{p_\theta(v \mid s)}{p_{\mathrm{T}}(v \mid s)} \right]. \tag{1}$$

Treating the teacher as fixed, its gradient can be written as

$$\nabla_\theta d_{\mathrm{RKL}}(\theta; s) = \mathbb{E}_{v\sim p_\theta(\cdot|s)}\left[\log \frac{p_\theta(v \mid s)}{p_{\mathrm{T}}(v \mid s)} \nabla_\theta \log p_\theta(v \mid s)\right], \tag{2}$$

where the constant score term vanishes because $\mathbb{E}_{v\sim p_\theta}[\nabla_\theta \log p_\theta(v \mid s)] = 0$. Thus, OPD provides a local teacher signal at each student-visited state. DOpt-OPD retains this local supervision signal and changes how it is aggregated across problems and visited states.

Let ℓ_{irt} denote the token-level distillation loss at state s_{irt}. Under standard token averaging,

$$\mathcal{L}_{\mathrm{tok}} = \frac{\sum_{i=1}^{N}\sum_{r=1}^{R}\sum_{t=1}^{T_{ir}} \ell_{irt}}{\sum_{i=1}^{N}\sum_{r=1}^{R} T_{ir}}. \tag{3}$$

The total scalar supervision weight assigned to problem i is therefore

$$b_i^{\mathrm{tok}} = \frac{\sum_{r=1}^{R} T_{ir}}{\sum_{u=1}^{N}\sum_{r=1}^{R} T_{ur}}. \tag{4}$$

Hence, problems with longer generated rollouts receive more total loss weight. The local OPD objective determines what teacher signal is applied at each visited state, whereas the aggregation rule determines how much each problem and state contributes to the update. DOpt-OPD addresses this second choice through problem-level normalization and within-problem D-optimal allocation.

2.2 D-Optimal Design

Following the classical treatments of Kiefer & Wolfowitz (1960) and Pukelsheim (2006), D-optimal design allocates a measurement budget to maximize the determinant of an information matrix. Given candidate vectors $\{\boldsymbol{z}_j\}_{j=1}^m \subset \mathbb{R}^d$, let $\boldsymbol{q} \in \Delta_m$ specify their allocation, where $\Delta_m = \{\boldsymbol{q} \in \mathbb{R}^m_{\geq 0} : \sum_{j=1}^m q_j = 1\}$. The regularized design matrix is

$$\boldsymbol{M}(\boldsymbol{q}) = \lambda \boldsymbol{I} + \sum_{j=1}^{m} q_j \boldsymbol{z}_j \boldsymbol{z}_j^\top, \quad \lambda > 0. \tag{5}$$

The corresponding D-optimal objective is

$$\boldsymbol{q}^\star \in \arg\max_{\boldsymbol{q}\in\Delta_m} \Phi(\boldsymbol{q}), \qquad \Phi(\boldsymbol{q}) = \log\det \boldsymbol{M}(\boldsymbol{q}). \tag{6}$$

The ridge term ensures that $\boldsymbol{M}(\boldsymbol{q})$ remains positive definite even when the candidates span only a low-dimensional subspace. The log-determinant favors coverage of complementary rather than redundant directions.

The marginal sensitivity to the allocation of candidate j follows from $d \log\det \boldsymbol{M} = \mathrm{tr}(\boldsymbol{M}^{-1} d\boldsymbol{M})$:

$$\tau_j(\boldsymbol{q}) \equiv \frac{\partial \Phi(\boldsymbol{q})}{\partial q_j} = \boldsymbol{z}_j^\top \boldsymbol{M}(\boldsymbol{q})^{-1} \boldsymbol{z}_j. \tag{7}$$

Thus, $\tau_j(\boldsymbol{q})$ measures the first-order benefit of allocating additional budget to candidate j relative to the directions already represented by the current design.

DOpt-OPD applies these sensitivities to state-conditioned gradient proxies to construct its one-step allocation. Further derivations are provided in Appendix A.

3 Method

DOpt-OPD organizes supervision allocation at two levels within the same OPD update. At the problem level, it assigns an equal total supervision budget to each problem, removing the dependence of problem-level weight on generated length. Within each problem, it then redistributes this fixed budget across reasoning states using D-optimal allocation.

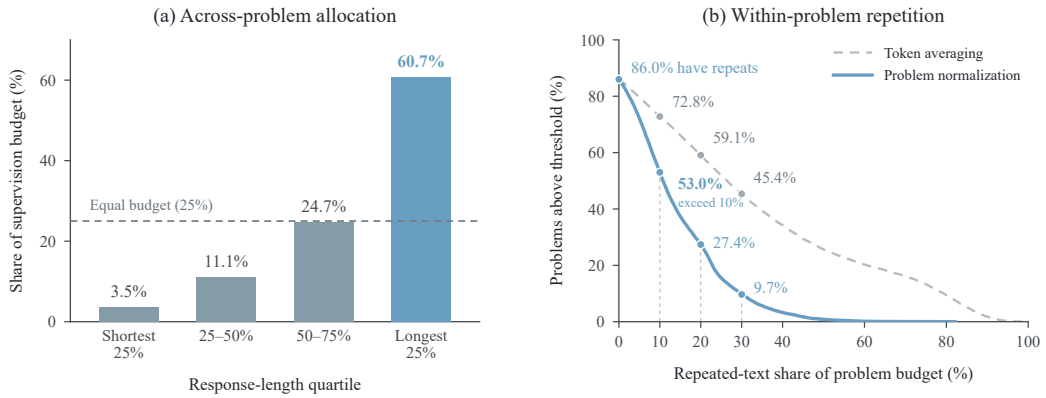

Figure 1: **Supervision allocation on OPD rollouts. (a)** Budget shares by problem-length quartile under token averaging; the dashed line indicates equal problem budgets. **(b)** CCDF of repeated-text budget shares under token averaging and problem normalization, measured by an 8-gram diagnostic.

3.1 Problem-Level Budget Normalization

Under token averaging, a problem's total loss weight scales with its generated token count (4). We remove this dependence by fixing its supervision budget to $b_i = 1/N$. The problem-normalized reference objective weights rollouts equally and averages token losses within each:

$$\mathcal{L}_{\mathrm{PN}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R}\sum_{r=1}^{R}\frac{1}{T_{ir}}\sum_{t=1}^{T_{ir}}\ell_{irt}. \tag{8}$$

Each valid token receives coefficient $1/(NRT_{ir})$, giving each problem total scalar loss weight $1/N$, independently of generated length.

Supervision-allocation audit. We audit fixed standard-OPD rollouts on DAPO-Math-17K (Appendix F.1). Under token averaging, the longest problem-length quartile receives 60.7% of the budget, versus 3.5% for the shortest (Fig. 1(a)). On the same rollouts, PN's equal rollout weighting reduces the median repeated-text budget share from 26.4% to 10.8%. Still, 53.0% of problems allocate over 10% of their budget to repeated text, detected by an 8-gram diagnostic (Fig. 1(b)).

Equal problem budgets do not determine within-problem allocation. Since textual repetition alone does not establish gradient redundancy, DOpt-OPD retains $b_i = 1/N$ and allocates this budget according to the collective coverage of gradient-proxy directions.

3.2 Within-Problem D-Optimal Allocation

Span pooling and gradient proxies. To aggregate local corrections and assess redundancy across rollouts of the same problem, we partition each rollout into nonoverlapping spans of L_{span} valid tokens and pool them within the problem. For problem i, this yields M_i spans, $\{\mathcal{S}_{ij}\}_{j=1}^{M_i}$. We denote the mean token loss over span $\mathcal{S}_{ij}$ by $\bar{\ell}_{ij}$.

We use LM-head gradients of the local OPD score surrogate ℓ_t^{loc} defined in Appendix C.2 to approximate update geometry without computing full-model gradients. Suppressing prompt and rollout indices, let $\boldsymbol{h}_t$ be the student LM-head input and $\boldsymbol{W}$ the output matrix. For logits $\boldsymbol{W}\boldsymbol{h}_t$, define $\boldsymbol{\delta}_t = \nabla_{\boldsymbol{W}\boldsymbol{h}_t}\ell_t^{\mathrm{loc}}$. Under column-major vectorization, the corresponding LM-head gradient and its inner-product structure satisfy

$$\mathrm{vec}(\nabla_{\boldsymbol{W}}\ell_t^{\mathrm{loc}}) = \boldsymbol{h}_t \otimes \boldsymbol{\delta}_t, \quad \langle \boldsymbol{h}\otimes\boldsymbol{\delta}, \boldsymbol{h}'\otimes\boldsymbol{\delta}'\rangle = \langle \boldsymbol{h},\boldsymbol{h}'\rangle\langle\boldsymbol{\delta},\boldsymbol{\delta}'\rangle. \tag{9}$$

Thus, head-gradient similarity depends jointly on hidden-state and correction geometry, motivating a proxy that combines both factors.

Let $\widehat{\boldsymbol{\delta}}_t$ be the sparse logit correction retained by the base OPD estimator. To reduce covariance computation, we sketch hidden states and sparse corrections with fixed Rademacher $\boldsymbol{R}_h$ and CountSketch

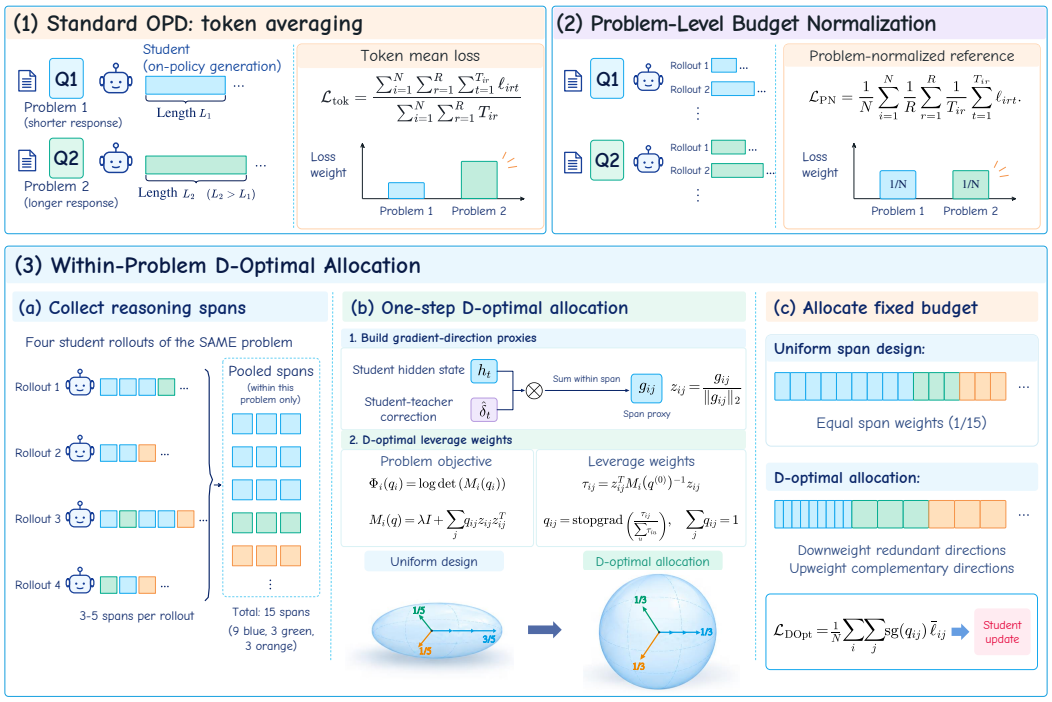

Figure 2: **Overview of DOpt-OPD.** Problem-level normalization fixes equal supervision budgets across problems. Within each problem, D-optimal allocation redistributes this budget across spans according to their gradient-proxy directions. Different colors represent different directions.

$\boldsymbol{R}_\delta$ maps (Charikar et al., 2002), respectively. Summing before normalization allows opposing token corrections to cancel within spans; unit normalization separates directional coverage from proxy magnitude:

$$\tilde{\boldsymbol{g}}_t = (\boldsymbol{R}_h \boldsymbol{h}_t) \otimes (\boldsymbol{R}_\delta \widehat{\boldsymbol{\delta}}_t) \in \mathbb{R}^d, \quad \boldsymbol{g}_{ij} = \sum_{t \in \mathcal{S}_{ij}} \tilde{\boldsymbol{g}}_t, \quad \boldsymbol{z}_{ij} = \frac{\boldsymbol{g}_{ij}}{\|\boldsymbol{g}_{ij}\|_2}, \tag{10}$$

where $d = d_h d_\delta$, and d_h and d_δ denote the hidden-state and correction sketch dimensions, respectively. The unit direction is defined on the active set $\mathcal{A}_i = \{j : \|\boldsymbol{g}_{ij}\|_2 > 0\}$, with $m_i = |\mathcal{A}_i|$.

D-optimal leverage allocation. For a nonempty active set, let $\boldsymbol{q}_i = (q_{ij})_{j \in \mathcal{A}_i}$ denote the within-problem allocation, where $q_{ij} \geq 0$ and $\sum_{j \in \mathcal{A}_i} q_{ij} = 1$. To favor balanced directional coverage under a fixed budget, we apply the D-optimal criterion from Section 2.2:

$$\boldsymbol{M}_i(\boldsymbol{q}_i) = \lambda \boldsymbol{I} + \sum_{j \in \mathcal{A}_i} q_{ij} \boldsymbol{z}_{ij} \boldsymbol{z}_{ij}^\top, \qquad \Phi_i(\boldsymbol{q}_i) = \log\det \boldsymbol{M}_i(\boldsymbol{q}_i). \tag{11}$$

Its sensitivities are

$$\tau_{ij}(\boldsymbol{q}_i) = \frac{\partial \Phi_i}{\partial q_{ij}} = \boldsymbol{z}_{ij}^\top \boldsymbol{M}_i(\boldsymbol{q}_i)^{-1} \boldsymbol{z}_{ij}. \tag{12}$$

To avoid per-prompt iterative optimization, DOpt-OPD takes one multiplicative update from the uniform active-span design $q_{ij}^{(0)} = 1/m_i$, using marginal coverage gains to obtain normalized allocation weights:

$$q_{ij} = q_{ij}^{(1)} = \frac{q_{ij}^{(0)} \tau_{ij}^{(0)}}{\sum_{u \in \mathcal{A}_i} q_{iu}^{(0)} \tau_{iu}^{(0)}} = \frac{\tau_{ij}^{(0)}}{\sum_{u \in \mathcal{A}_i} \tau_{iu}^{(0)}}, \tag{13}$$

where $\tau_{ij}^{(0)} = \tau_{ij}(\boldsymbol{q}_i^{(0)})$. This one-step update assigns larger weights to spans with greater marginal coverage gains, accounting for directions already represented in the pool. Its Bayesian motivation and derivation are given in Appendix A.

For efficient computation, form the summed covariance

$$\boldsymbol{C}_i = \sum_{j \in \mathcal{A}_i} \boldsymbol{z}_{ij} \boldsymbol{z}_{ij}^\top + \lambda_{\text{impl}} \boldsymbol{I}, \qquad \lambda_{\text{impl}} = m_i \lambda. \tag{14}$$

Since $\boldsymbol{M}_i(\boldsymbol{q}_i^{(0)}) = \boldsymbol{C}_i / m_i$, we have $\tau_{ij}^{(0)} = m_i \tilde{\tau}_{ij}$ for $\tilde{\tau}_{ij} = \boldsymbol{z}_{ij}^\top \boldsymbol{C}_i^{-1} \boldsymbol{z}_{ij}$. The common factor cancels in (13). A Cholesky factorization and triangular solves compute these scores without explicitly forming an inverse.

Training objective. For a nonempty active set, we normalize active scores and assign zero mass to inactive spans; if all proxies vanish, we use uniform span weights $q_{ij} = 1/M_i$. The objective is

$$\mathcal{L}_{\text{DOpt}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \text{stopgrad}(q_{ij}) \bar{\ell}_{ij}. \tag{15}$$

Let $n_{ij} = |\mathcal{S}_{ij}|$ denote the number of valid tokens in span j. Each token receives coefficient $q_{ij}/(N n_{ij})$. Since $\sum_j q_{ij} = 1$, each problem retains total scalar mass $1/N$. Only the allocation weights are detached, while the base surrogate retains its ordinary gradient path. The training procedure is summarized in Algorithm 1.

3.3 Theoretical Analysis of One-Step D-Optimal Allocation

The D-optimal problem induced by (11) defines the ideal allocation, whereas we use only one inexpensive update from the uniform span design. We analyze this approximation in two complementary regimes: an orthogonal direction-group model provides an ex ante near-optimality guarantee, while for general directions, the post-update KKT residual yields an ex post, computable certificate for the actual optimality gap.

We first establish a general KKT characterization. For any $\boldsymbol{q}_i \in \Delta_{\mathcal{A}_i} = \left\{ \boldsymbol{q}_i \geq 0 : \sum_{j \in \mathcal{A}_i} q_{ij} = 1 \right\}$, define

$$\overline{\tau}_i(\boldsymbol{q}_i) = \sum_{j \in \mathcal{A}_i} q_{ij} \tau_{ij}(\boldsymbol{q}_i), \qquad G_i(\boldsymbol{q}_i) = \max_{j \in \mathcal{A}_i} \tau_{ij}(\boldsymbol{q}_i) - \overline{\tau}_i(\boldsymbol{q}_i). \tag{16}$$

Lemma 1 (KKT residual)**.** *For $\lambda > 0$, $\boldsymbol{q}_i$ is globally D-optimal if and only if $\tau_{ij}(\boldsymbol{q}_i) \leq \overline{\tau}_i(\boldsymbol{q}_i)$ for all j, with equality whenever $q_{ij} > 0$. Equivalently,*

$$G_i(\boldsymbol{q}_i) = \max_{\boldsymbol{p}_i \in \Delta_{\mathcal{A}_i}} \nabla \Phi_i(\boldsymbol{q}_i)^\top (\boldsymbol{p}_i - \boldsymbol{q}_i) \geq 0, \tag{17}$$

and $G_i(\boldsymbol{q}_i) = 0$ if and only if $\boldsymbol{q}_i$ is globally D-optimal.

We first consider directional redundancy. Partition $\mathcal{A}_i$, with $m_i = |\mathcal{A}_i|$, into r_i direction groups and assume that

$$\boldsymbol{z}_{ij} \boldsymbol{z}_{ij}^\top = \boldsymbol{v}_{ig} \boldsymbol{v}_{ig}^\top \quad (j \in \mathcal{G}_{ig}), \qquad \boldsymbol{v}_{ig}^\top \boldsymbol{v}_{ih} = \delta_{gh}. \tag{18}$$

Thus, multiple spans may share the same effective direction, including directions with opposite signs. Let $m_{ig} = |\mathcal{G}_{ig}|$, $Q_{ig}(\boldsymbol{q}_i) = \sum_{j \in \mathcal{G}_{ig}} q_{ij}$, and $m_{i,\min} = \min_g m_{ig}$.

Theorem 1 (Near-optimality under orthogonal direction groups)**.** *Suppose that (18) holds. Every global solution $\boldsymbol{q}_i^\star$ satisfies $Q_{ig}(\boldsymbol{q}_i^\star) = 1/r_i$. Let $\delta_i = m_i \lambda / m_{i,\min}$. Starting from $q_{ij}^{(0)} = 1/m_i$, the one-step D-optimal allocation satisfies*

$$\max_g \left| Q_{ig}(\boldsymbol{q}_i^{(1)}) - \frac{1}{r_i} \right| \leq \frac{\delta_i}{r_i}, \qquad 0 \leq \Phi_i(\boldsymbol{q}_i^\star) - \Phi_i(\boldsymbol{q}_i^{(1)}) \leq \frac{r_i}{2} (1 + \delta_i)^2 \delta_i^2. \tag{19}$$

Consequently, for fixed group sizes, the group-mass error is $O(\lambda)$, while the objective gap is $O(\lambda^2)$. If all groups have the same size, the uniform initialization is already globally optimal.

Under (18), $\tau_{ij} = (\lambda + Q_{ig})^{-1}$ for $j \in \mathcal{G}_{ig}$, so the KKT condition equalizes mass across effective directions; for general directions, G_i instead provides an ex post certificate.

Theorem 2 (Optimality certificate for general directions)**.** *Let* $\boldsymbol{q}_i^\star \in \arg\max_{\boldsymbol{q}_i \in \Delta_{\mathcal{A}_i}} \Phi_i(\boldsymbol{q}_i)$. *Without any orthogonality assumption,*

$$0 \le \Phi_i(\boldsymbol{q}_i^\star) - \Phi_i(\boldsymbol{q}_i^{(1)}) \le G_i(\boldsymbol{q}_i^{(1)}), \tag{20}$$

and

$$\exp\left(-\frac{G_i(\boldsymbol{q}_i^{(1)})}{d}\right) \le \left(\frac{\det \boldsymbol{M}_i(\boldsymbol{q}_i^{(1)})}{\det \boldsymbol{M}_i(\boldsymbol{q}_i^\star)}\right)^{1/d} \le 1. \tag{21}$$

Proofs and sharper error bounds are provided in Appendix B.

Because G_i uses only post-update sensitivities, a small value certifies near-optimality, whereas a large value is inconclusive. Theorem 1 explains *why* one step mitigates directional redundancy, while Theorem 2 certifies *how close* it is to optimality for general directions.

Table 1: **Main results on Avg@8.** Results are reported in percentages. **Bold** indicates the best performance among student models; underline indicates second-best.

Method	AIME24	AIME25	AMC23	Minerva	Olympiad	CMIMC25	BRUMO25	Avg.
Scale 1 (Large): Qwen3-8B Teacher → Qwen3-4B Student								
Student (Base)	7.50	6.25	30.94	7.90	23.29	0.62	9.58	12.30
OPD	16.25	15.00	53.13	32.35	43.14	5.62	22.08	26.80
TIP	17.08	14.17	**58.44**	32.40	43.51	5.94	24.17	27.96
EOPD	17.08	15.00	55.63	33.13	42.75	6.25	23.33	27.60
FiRe-OPD	19.17	15.42	55.63	**34.19**	**43.55**	**6.88**	24.17	28.43
DOpt-OPD (Ours)	**20.00**	**16.25**	57.19	32.81	42.93	6.56	**27.08**	**28.97**
Δ vs. OPD	+3.75	+1.25	+4.06	+0.46	−0.21	+0.94	+5.00	+2.17
Scale 2 (Small): Qwen3-4B Teacher → Qwen3-1.7B Student								
Student (Base)	2.92	2.08	19.69	7.49	13.26	0.63	4.17	7.18
OPD	7.92	5.42	45.00	19.85	31.34	2.19	16.25	18.28
TIP	11.67	8.33	44.38	**21.92**	34.74	3.75	15.42	20.03
EOPD	11.67	**11.25**	44.38	20.68	34.79	4.38	15.83	20.43
FiRe-OPD	10.42	7.92	**45.94**	20.96	32.01	2.50	13.33	19.01
DOpt-OPD (Ours)	**12.50**	10.00	45.63	21.05	**35.00**	**5.62**	**17.08**	**20.98**
Δ vs. OPD	+4.58	+4.58	+0.63	+1.20	+3.66	+3.43	+0.83	+2.70

4 Experiments

4.1 Experimental Setup

Models and training. We evaluate DOpt-OPD at two teacher–student scales: (1) **Large scale**: Qwen3-8B teacher → Qwen3-4B student; and (2) **Small scale**: Qwen3-4B teacher → Qwen3-1.7B student (Yang et al., 2025). All students are trained on DAPO-Math-17K (Yu et al., 2025) using the VERL framework (Sheng et al., 2024). Following Li et al. (2026a), we use Student-TopK consistently across all methods. Full training and implementation details are provided in the appendix.

Baselines and benchmarks. We compare DOpt-OPD with standard OPD (Agarwal et al., 2024; Li et al., 2026a), TIP (Xu et al., 2026), EOPD (Jin et al., 2026), and FiRe-OPD (Li et al., 2026b). We evaluate on seven mathematical reasoning benchmarks: AIME 2024 and 2025, AMC 2023, Minerva Math, OlympiadBench, CMIMC 2025, and BRUMO 2025, reporting Avg@8 and Pass@8.

4.2 Main Results

DOpt-OPD outperforms the evaluated baselines in seven-benchmark macro-averaged Avg@8 and Pass@8 at both scales (Tables 1 and 2). Over OPD, it gains 2.17/3.55 points for the 4B student and

Table 2: **Main results on Pass@8.** Results are reported in percentages. **Bold** indicates the best performance among student models; underline indicates second-best.

Method	AIME24	AIME25	AMC23	Minerva	Olympiad	CMIMC25	BRUMO25	Avg.
Scale 1 (Large): Qwen3-8B Teacher → Qwen3-4B Student								
Student (Base)	26.67	23.33	65.00	33.46	53.56	5.00	30.00	33.86
OPD	30.00	**33.33**	82.50	47.06	**63.50**	22.50	40.00	45.56
TIP	40.00	**33.33**	77.50	46.69	62.31	22.50	43.33	46.52
EOPD	36.67	30.00	85.00	47.79	63.20	17.50	36.67	45.26
FiRe-OPD	40.00	26.67	85.00	47.43	62.91	22.50	36.67	45.88
DOpt-OPD (Ours)	**43.33**	30.00	**87.50**	**48.53**	62.76	**25.00**	**46.67**	**49.11**
Δ vs. OPD	+13.33	−3.33	+5.00	+1.47	−0.74	+2.50	+6.67	+3.55
Scale 2 (Small): Qwen3-4B Teacher → Qwen3-1.7B Student								
Student (Base)	16.67	10.00	57.50	27.57	39.47	5.00	23.33	25.65
OPD	23.33	13.33	67.50	37.87	50.45	10.00	26.67	32.74
TIP	26.67	16.67	70.00	**40.81**	56.38	10.00	**30.00**	35.79
EOPD	23.33	20.00	70.00	37.87	**57.72**	15.00	26.67	35.80
FiRe-OPD	**30.00**	13.33	72.50	39.34	51.93	12.50	26.67	35.18
DOpt-OPD (Ours)	**30.00**	**23.33**	**77.50**	38.60	54.90	**17.50**	**30.00**	**38.83**
Δ vs. OPD	+6.67	+10.00	+10.00	+0.73	+4.45	+7.50	+3.33	+6.09

Table 3: **Ablation study across OPD response lengths.** Entries report Avg@8 / Pass@8, averaged equally over seven benchmarks. Length groups are defined within each benchmark using OPD responses; grouped values show gains over OPD in percentage points.

Method	Avg.	Δ vs. OPD		
		Shortest 25%	Middle 50%	Longest 25%
OPD	18.28 / 32.74	—	—	—
Problem normalization	20.07 / 36.33	−0.33 / + 0.23	+2.28 / + 0.95	+2.93 / + 12.23
Uniform span allocation	19.73 / 37.18	+0.07 / + 0.37	+1.50 / + 2.83	+2.73 / + 11.73
DOpt-OPD (Ours)	**20.98 / 38.83**	−0.16 / + 1.41	+3.94 / + 4.85	+3.08 / + 13.25

2.70/6.09 points for the 1.7B student (Avg@8/Pass@8). The larger Pass@8 gains indicate improved coverage under eight samples alongside higher average correctness.

4.3 Ablation and Sensitivity Analysis

Component ablations. Table 3 compares four variants with the 1.7B student across seven benchmarks. Problem normalization equalizes problem budgets and averages token losses within equally weighted rollouts; uniform span allocation splits each problem's budget equally across pooled spans; DOpt-OPD uses leverage weights. Normalization improves Avg@8/Pass@8 over OPD by 1.79/3.59 points; DOpt-OPD adds 0.91/2.50 over normalization and 1.25/1.65 over uniform allocation, supporting both components.

Length-stratified performance. Using shared within-benchmark groups defined by OPD's mean response length, DOpt-OPD gains 1.41/4.85/13.25 Pass@8 points in the shortest 25%, middle 50%, and longest 25%, with Avg@8 changes of $-0.16/+3.94/+3.08$ points. In the longest group, its Pass@8 gains over normalization and uniform allocation are 1.02 and 1.52 points.

Hyperparameter sensitivity. Table 4 varies span length and proxy dimensions with the 4B teacher and 1.7B student on AIME 2024/2025 and AMC 2023. Default rows reuse the main-results run; Tokens/Trunc. report mean response length/truncation rate (%). The 32-token, 8×8 default achieves the highest Avg@8/Pass@8 and lowest truncation among tested settings. Shorter responses from

Table 4: **Sensitivity to span length and proxy dimensions.** Proxy shape: hidden-state projection dimension $\times$ correction-sketch dimension. Highlighted rows share the default run (32 tokens, 8×8).

Setting	Avg@8 ↑	Pass@8 ↑	Tokens	Trunc. ↓
Span length (tokens)				
16	20.90	36.94	11,476	26.62
32 (default)	**22.71**	**43.61**	8,246	**15.50**
64	19.44	36.11	7,765	19.75
Proxy dimensions ($d_h \times d_\delta$)				
8×8 (64-D, default)	**22.71**	**43.61**	8,246	**15.50**
16×8 (128-D)	19.62	35.83	10,997	24.88
8×16 (128-D)	21.01	38.61	9,612	20.00
16×16 (256-D)	21.18	39.17	8,746	16.75

64-token spans yield lower accuracy, and larger proxies also underperform; at 128 dimensions, 8×16 exceeds 16×8 by 1.39/2.78 points.

5 Related Work

On-policy distillation and supervision refinement. On-policy distillation trains students with teacher feedback on self-generated trajectories, reducing the mismatch between training and inference states (Agarwal et al., 2024). Subsequent work has examined its empirical behavior and training recipe (Li et al., 2026a), and refined supervision through position, token importance, entropy, teachability, and trust-region criteria (Xie et al., 2026a; Xu et al., 2026; Jin et al., 2026; Wang et al., 2026; Xie et al., 2026b). Our work studies a complementary aspect of OPD: how teacher signals are aggregated across problems and their visited states. DOpt-OPD equalizes problem-level budgets and allocates each fixed budget according to the collective coverage of update directions.

Length, position, and repetition in OPD. Length inflation and unstable trajectories motivate OPD stabilization strategies (Luo et al., 2026), while position-aware weighting addresses variation across generated responses (Xie et al., 2026a). Related work explores tokenizer-agnostic distillation for long-context reasoning (He et al., 2026). We examine update-direction redundancy in recurrent reasoning spans. Rather than explicitly penalizing long responses or repeated text, our within-problem allocation applies the same gradient-coverage criterion to recurrent and nonrecurrent spans.

Optimal design and gradient diversity. D-optimal design allocates measurement weights to maximize the determinant of an information matrix (Kiefer & Wolfowitz, 1960; Pukelsheim, 2006), with related developments in ridge leverage (Cohen et al., 2015; Murray et al., 2023). Gradient embeddings and diversity criteria have also been studied for batch active learning, data selection, and dynamic coreset construction (Ash et al., 2020; 2021; Li et al., 2025; Tang et al., 2026). We apply this geometric perspective to supervision allocation within each problem. State-conditioned gradient proxies represent the visited spans, and sensitivities of a regularized D-optimal objective yield a one-step allocation under a fixed supervision budget.

6 Conclusion

We studied supervision allocation in OPD, where token averaging couples problem-level loss weight to generated length and repeated visits can accumulate supervision along already represented update directions. We introduced DOpt-OPD, combining problem-level normalization with a one-step D-optimal allocation over state-conditioned gradient-proxy directions. Across two teacher–student scales and seven mathematical reasoning benchmarks, DOpt-OPD achieves the highest macro-averaged Avg@8 and Pass@8 among evaluated baselines. Component ablations support the two-level allocation design, while length-stratified results show larger Pass@8 gains on questions with longer standard OPD responses. Controlled repeated-suffix diagnostics show reduced supervision-budget dilution in tested cases. These results highlight supervision allocation across problems and visited states as an important OPD design choice.

AI USE STATEMENT

Generative AI tools assisted with language editing and polishing to improve the clarity and readability of the manuscript. The authors reviewed and revised the resulting text and take full responsibility for the final manuscript and its content.

ETHICS STATEMENT

This work studies language-model distillation using public mathematical-reasoning datasets and involves neither human subjects nor personal data.

REPRODUCIBILITY STATEMENT

Appendix B provides the theoretical proofs, Appendix C details the algorithm and implementation, and Appendix G lists shared training and evaluation hyperparameters. Appendices D–F report supplementary comparisons, controlled theoretical validation experiments, and mechanism diagnostics.

REFERENCES

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

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with Fisher embeddings. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. doi: 10.1214/ss/1177009939.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *Automata, Languages and Programming*, volume 2380 of *Lecture Notes in Computer Science*, pp. 693–703. Springer, 2002. doi: 10.1007/3-540-45465-9_59.

Michael B. Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. *arXiv preprint arXiv:1511.07263*, 2015.

GLM-5 Team. GLM-5: From vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026. URL https://arxiv.org/abs/2602.15763.

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

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Jack Kiefer. General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, 2(5):849–879, 1974. doi: 10.1214/aos/1176342810.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960. doi: 10.4153/CJM-1960-030-4.

Ming Li, Yanhong Li, Ziyue Li, and Tianyi Zhou. How instruction and reasoning data shape post-training: Data quality through the lens of layer-wise gradients. *arXiv preprint arXiv:2504.10766*, 2025.

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

Yuying Li, Leqi Zheng, Yongzi Yu, Wenrui Zhou, Xuchang Zhong, Xing Hu, Jing Jin, Hangjie Yuan, and Tao Feng. Filter, then reweight: Rethinking optimization granularity in on-policy distillation. *arXiv preprint arXiv:2606.02684*, 2026b. URL https://arxiv.org/abs/2606.02684.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying OPD: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*, 2026.

Riley Murray, James Demmel, Michael W. Mahoney, N. Benjamin Erichson, Maksim Melnichenko, Osman Asif Malik, Laura Grigori, Piotr Luszczek, Michał Dereziński, Miles E. Lopes, Tianyu Liang, Hengrui Luo, and Jack Dongarra. Randomized numerical linear algebra: A perspective on the field with an eye to software. *arXiv preprint arXiv:2302.11474*, 2023.

Friedrich Pukelsheim. *Optimal Design of Experiments*, volume 50 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2006. ISBN 978-0-89871-604-7. doi: 10.1137/1.9780898719109.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024.

Tianhao Tang, Haoyang Li, and Lei Chen. GRACE: A dynamic coreset selection framework for large language model optimization. *arXiv preprint arXiv:2604.11810*, 2026.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026.

Xiaomi LLM-Core Team. MiMo-V2-Flash Technical Report. *arXiv preprint arXiv:2601.02780*, 2026. URL https://arxiv.org/abs/2601.02780.

Yan Xie, Sijie Zhu, Tiansheng Wen, Bo Chen, and Yifei Wang. On the position bias of on-policy distillation. *arXiv preprint arXiv:2606.22600*, 2026a.

Zhengpeng Xie, Li Lyna Zhang, Zeke Xie, and Mao Yang. Trust region policy distillation. *arXiv preprint arXiv:2607.04751*, 2026b.

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

An Yang et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Zichao Yu, Chengzhi Yu, Shengze Xu, Yujin Han, Bingqing Jiang, Xu Wang, and Difan Zou. Mismatch matters: On-policy distillation beyond token agreement. *arXiv preprint arXiv:2608.09836*, 2026. URL https://arxiv.org/abs/2608.09836.

A D-OPTIMAL FOUNDATIONS AND ONE-STEP UPDATE

This section derives the DOpt-OPD objective and its one-step update from a local information-allocation model. Fix prompt i, and let $\Delta\theta \in \mathbb{R}^p$ denote a local parameter correction. For span j, define the teacher–student residual

$$\boldsymbol{r}_{ij} = \boldsymbol{t}_{ij} - \boldsymbol{f}_{ij}(\theta).$$

A first-order expansion gives the local linear relation

$$\boldsymbol{r}_{ij} = J_{ij}\Delta\theta + \boldsymbol{\varepsilon}_{ij}, \qquad J_{ij} = \nabla_\theta \boldsymbol{f}_{ij}(\theta). \tag{22}$$

This relation is used only as a local surrogate describing how each span constrains the unknown correction $\Delta\theta$, rather than as a literal model of the training data.

Following approximate-design theory (Kiefer, 1974; Pukelsheim, 2006), suppose conceptually that a prompt-level budget of B_i units is allocated through integer counts κ_{ij}, with $\sum_j \kappa_{ij} = B_i$. Approximate design replaces the normalized counts κ_{ij}/B_i with continuous proportions

$$q_{ij} \geq 0, \qquad \sum_{j\in\mathcal{A}_i} q_{ij} = 1.$$

Their effect is represented by the weighted Gaussian likelihood

$$p(\{\boldsymbol{r}_{ij}\}_{j\in\mathcal{A}_i} \mid \Delta\theta, \boldsymbol{q}_i) \propto \exp\left\{-\frac{1}{2}\sum_{j\in\mathcal{A}_i} q_{ij}\big(\boldsymbol{r}_{ij} - J_{ij}\Delta\theta\big)^\top \Gamma_{ij}^{-1}\big(\boldsymbol{r}_{ij} - J_{ij}\Delta\theta\big)\right\}, \tag{23}$$

where $\Gamma_{ij} \succ 0$ is the local residual covariance before budget weighting. Thus, q_{ij} scales the information contribution of span j. For $q_{ij} > 0$, this is equivalent, as a function of $\Delta\theta$, to using covariance Γ_{ij}/q_{ij}. When $q_{ij} = 0$, the corresponding term vanishes and that span contributes no information to the local surrogate. The common budget scale B_i is absorbed into Γ_{ij}^{-1}.

We place the Gaussian prior

$$\Delta\theta \sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{\Lambda}_0^{-1}), \qquad \boldsymbol{\Lambda}_0 \succ 0, \tag{24}$$

where $\boldsymbol{m}_0$ is the correction expected before using the span-level residuals and $\boldsymbol{\Lambda}_0$ is the baseline precision. In particular, $\boldsymbol{m}_0 = \mathbf{0}$ corresponds to a prior centered at no local correction, while $\boldsymbol{\Lambda}_0$ provides baseline regularization. By standard linear-Gaussian conjugacy, the posterior remains Gaussian with precision

$$\boldsymbol{\Lambda}_{i,\mathrm{post}}(\boldsymbol{q}_i) = \boldsymbol{\Lambda}_0 + \sum_{j\in\mathcal{A}_i} q_{ij} J_{ij}^\top \Gamma_{ij}^{-1} J_{ij}. \tag{25}$$

The residuals affect the posterior mean but not its precision. Since DOpt designs the allocation through posterior uncertainty, only (25) is needed below. This construction follows the standard Bayesian experimental-design perspective (Chaloner & Verdinelli, 1995).

For the scalarized span-loss model, let

$$\boldsymbol{g}_{ij}^{\mathrm{full}} = \nabla_\theta \bar{\ell}_{ij}, \qquad J_{ij} = (\boldsymbol{g}_{ij}^{\mathrm{full}})^\top, \qquad \Gamma_{ij} = \sigma_{ij}^2.$$

The corresponding information contribution is

$$J_{ij}^\top \Gamma_{ij}^{-1} J_{ij} = \omega_{ij}\boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top, \qquad \boldsymbol{v}_{ij} = \frac{\boldsymbol{g}_{ij}^{\mathrm{full}}}{\|\boldsymbol{g}_{ij}^{\mathrm{full}}\|_2}, \qquad \omega_{ij} = \frac{\|\boldsymbol{g}_{ij}^{\mathrm{full}}\|_2^2}{\sigma_{ij}^2}. \tag{26}$$

Here, $\boldsymbol{v}_{ij}$ specifies the full-gradient information direction and ω_{ij} its scale. DOpt focuses on directional complementarity under a fixed prompt-level budget. It therefore removes the scale ω_{ij} by unit normalization and approximates $\boldsymbol{v}_{ij}$ with the sketched unit direction $\boldsymbol{z}_{ij}$ from (10).

Assigning a common information scale $\alpha > 0$ to the proxy directions and using an isotropic baseline precision $\lambda_0 \boldsymbol{I}$ gives

$$\widetilde{\boldsymbol{\Lambda}}_{i,\mathrm{post}}(\boldsymbol{q}_i) = \lambda_0 \boldsymbol{I} + \alpha \sum_{j\in\mathcal{A}_i} q_{ij}\boldsymbol{z}_{ij}\boldsymbol{z}_{ij}^\top = \alpha \boldsymbol{M}_i(\boldsymbol{q}_i), \qquad \lambda = \frac{\lambda_0}{\alpha}, \tag{27}$$

where

$$\boldsymbol{M}_i(\boldsymbol{q}_i) = \lambda \boldsymbol{I} + \sum_{j \in \mathcal{A}_i} q_{ij} \boldsymbol{z}_{ij} \boldsymbol{z}_{ij}^\top. \tag{28}$$

Thus, $\boldsymbol{M}_i(\boldsymbol{q}_i)$ is a low-dimensional, direction-normalized approximation to the full posterior precision. Since the common factor α contributes only $d \log \alpha$ to the log determinant, it does not affect the optimal allocation. This gives the D-optimal problem

$$\max_{\boldsymbol{q}_i \in \Delta_{\mathcal{A}_i}} \Phi_i(\boldsymbol{q}_i), \qquad \Phi_i(\boldsymbol{q}_i) = \log\det \boldsymbol{M}_i(\boldsymbol{q}_i), \tag{29}$$

where

$$\Delta_{\mathcal{A}_i} = \left\{ \boldsymbol{q}_i \geq 0 : \sum_{j \in \mathcal{A}_i} q_{ij} = 1 \right\}.$$

To interpret this objective, consider the ellipsoid

$$\mathcal{E}_i(\boldsymbol{q}_i) = \left\{ \boldsymbol{u} \in \mathbb{R}^d : \boldsymbol{u}^\top \boldsymbol{M}_i(\boldsymbol{q}_i) \boldsymbol{u} \leq 1 \right\}. \tag{30}$$

Its volume satisfies

$$\mathrm{Vol}\big(\mathcal{E}_i(\boldsymbol{q}_i)\big) = V_d \det \boldsymbol{M}_i(\boldsymbol{q}_i)^{-1/2}, \tag{31}$$

where V_d is the volume of the d-dimensional unit ball. Hence, maximizing Φ_i minimizes the corresponding uncertainty volume. Moreover, because $\|\boldsymbol{z}_{ij}\|_2 = 1$,

$$\mathrm{tr}\, \boldsymbol{M}_i(\boldsymbol{q}_i) = d\lambda + 1 \tag{32}$$

is constant over the feasible set. Maximizing the determinant therefore promotes balanced coverage of complementary directions rather than repeated coverage of redundant ones.

By Jacobi's formula, the marginal sensitivity of span j is

$$\tau_{ij}(\boldsymbol{q}_i) = \frac{\partial \Phi_i(\boldsymbol{q}_i)}{\partial q_{ij}} = \boldsymbol{z}_{ij}^\top \boldsymbol{M}_i(\boldsymbol{q}_i)^{-1} \boldsymbol{z}_{ij}. \tag{33}$$

It measures the first-order objective gain obtained by increasing the allocation of that span. Indeed, transferring a sufficiently small mass ε from span k to span j gives

$$\Phi_i(\boldsymbol{q}_i + \varepsilon(\boldsymbol{e}_j - \boldsymbol{e}_k)) - \Phi_i(\boldsymbol{q}_i) = \varepsilon\left(\tau_{ij}(\boldsymbol{q}_i) - \tau_{ik}(\boldsymbol{q}_i)\right) + O(\varepsilon^2). \tag{34}$$

The same sensitivities are related to ridge leverage scores. For

$$\boldsymbol{A}_i(\boldsymbol{q}_i) = \left[\sqrt{q_{ij}} \boldsymbol{z}_{ij}\right]_{j \in \mathcal{A}_i}, \qquad \boldsymbol{M}_i(\boldsymbol{q}_i) = \lambda \boldsymbol{I} + \boldsymbol{A}_i(\boldsymbol{q}_i) \boldsymbol{A}_i(\boldsymbol{q}_i)^\top,$$

the ridge leverage score of the weighted column $\sqrt{q_{ij}} \boldsymbol{z}_{ij}$ is

$$q_{ij} \boldsymbol{z}_{ij}^\top \boldsymbol{M}_i(\boldsymbol{q}_i)^{-1} \boldsymbol{z}_{ij} = q_{ij} \tau_{ij}(\boldsymbol{q}_i). \tag{35}$$

Thus, the sensitivity of a span depends on how well its direction is already represented by the complete prompt-local pool (Cohen et al., 2015; Murray et al., 2023).

DOpt-OPD converts these marginal gains into allocation weights through one multiplicative update. Define

$$\overline{\tau}_i(\boldsymbol{q}_i) = \sum_{j \in \mathcal{A}_i} q_{ij} \tau_{ij}(\boldsymbol{q}_i). \tag{36}$$

The update is

$$q_{ij}^{(k+1)} = \frac{q_{ij}^{(k)} \tau_{ij}(\boldsymbol{q}_i^{(k)})}{\overline{\tau}_i(\boldsymbol{q}_i^{(k)})}. \tag{37}$$

Equivalently, it normalizes the ridge leverage scores in (35). It is also the solution of the KL-regularized reweighting problem

$$\boldsymbol{q}_i^{(k+1)} = \arg\max_{\boldsymbol{q}_i \in \Delta_{\mathcal{A}_i}} \left\{ \sum_{j \in \mathcal{A}_i} q_{ij} \log \tau_{ij}(\boldsymbol{q}_i^{(k)}) - D_{\mathrm{KL}}\big(\boldsymbol{q}_i \| \boldsymbol{q}_i^{(k)}\big) \right\}, \tag{38}$$

where

$$D_{\mathrm{KL}}(\boldsymbol{q}_i \| \boldsymbol{q}_i^{(k)}) = \sum_{j \in \mathcal{A}_i} q_{ij} \log \frac{q_{ij}}{q_{ij}^{(k)}}.$$

The first-order condition gives $q_{ij} \propto q_{ij}^{(k)} \tau_{ij}(\boldsymbol{q}_i^{(k)})$, which yields (37) after normalization.

Starting from the uniform active-span design $q_{ij}^{(0)} = 1/m_i$, the common initial mass cancels, giving

$$q_{ij}^{(1)} = \frac{\tau_{ij}(\boldsymbol{q}_i^{(0)})}{\sum_{u \in \mathcal{A}_i} \tau_{iu}(\boldsymbol{q}_i^{(0)})}. \tag{39}$$

This is the one-step DOpt-OPD allocation used in the main method. It performs a single inexpensive reweighting rather than a fully converged D-optimal solve. The main-text analysis establishes near-optimality under orthogonal direction groups for sufficiently small ridge values and provides a computable upper bound on the optimality gap for general directions.

Finally, the uniform initialization permits an efficient implementation. Let

$$\boldsymbol{C}_i = \sum_{j \in \mathcal{A}_i} \boldsymbol{z}_{ij} \boldsymbol{z}_{ij}^\top + m_i \lambda \boldsymbol{I}. \tag{40}$$

Then

$$\boldsymbol{M}_i(\boldsymbol{q}_i^{(0)}) = \frac{\boldsymbol{C}_i}{m_i}, \qquad \tau_{ij}(\boldsymbol{q}_i^{(0)}) = m_i \boldsymbol{z}_{ij}^\top \boldsymbol{C}_i^{-1} \boldsymbol{z}_{ij}. \tag{41}$$

B Proofs for the One-Step Optimality Analysis

Proof of Lemma 1. Since $\lambda > 0$, the matrix $\boldsymbol{M}_i(\boldsymbol{q}_i)$ is positive definite for every feasible $\boldsymbol{q}_i$. Hence, Φ_i is differentiable and concave on $\Delta_{\mathcal{A}_i}$, with

$$\frac{\partial \Phi_i(\boldsymbol{q}_i)}{\partial q_{ij}} = \tau_{ij}(\boldsymbol{q}_i).$$

For multipliers $\nu_i \in \mathbb{R}$ and $\mu_{ij} \geq 0$, consider

$$\mathcal{L}(\boldsymbol{q}_i, \nu_i, \boldsymbol{\mu}_i) = \Phi_i(\boldsymbol{q}_i) - \nu_i \left(\sum_j q_{ij} - 1 \right) + \sum_j \mu_{ij} q_{ij}.$$

The KKT conditions are

$$\tau_{ij}(\boldsymbol{q}_i) - \nu_i + \mu_{ij} = 0, \qquad \mu_{ij} q_{ij} = 0, \qquad \mu_{ij} \geq 0. \tag{42}$$

Thus, $\tau_{ij}(\boldsymbol{q}_i) \leq \nu_i$, with equality whenever $q_{ij} > 0$. Multiplying the stationarity condition by q_{ij}, summing over j, and using complementary slackness gives

$$\nu_i = \sum_{j \in \mathcal{A}_i} q_{ij} \tau_{ij}(\boldsymbol{q}_i) = \overline{\tau}_i(\boldsymbol{q}_i).$$

This proves the stated KKT characterization.

Moreover, because a linear function over the simplex is maximized at a vertex,

$$\begin{aligned} \max_{\boldsymbol{p}_i \in \Delta_{\mathcal{A}_i}} \nabla \Phi_i(\boldsymbol{q}_i)^\top (\boldsymbol{p}_i - \boldsymbol{q}_i) &= \max_{\boldsymbol{p}_i \in \Delta_{\mathcal{A}_i}} \sum_j \tau_{ij}(\boldsymbol{q}_i)(p_{ij} - q_{ij}) \\ &= \max_j \tau_{ij}(\boldsymbol{q}_i) - \sum_j q_{ij} \tau_{ij}(\boldsymbol{q}_i) = G_i(\boldsymbol{q}_i). \end{aligned} \tag{43}$$

Hence $G_i(\boldsymbol{q}_i) \geq 0$. It vanishes exactly when the KKT conditions hold, which, by concavity of Φ_i, is equivalent to global D-optimality. □

Proof of Theorem 1. For brevity, suppress the prompt index i. Under (18), the design matrix depends on the span-level allocation only through the group masses:

$$\boldsymbol{M}(\boldsymbol{q}) = \lambda \boldsymbol{I} + \sum_{g=1}^{r} Q_g \boldsymbol{v}_g \boldsymbol{v}_g^\top. \tag{44}$$

Therefore, for every $j \in \mathcal{G}_g$,

$$\tau_j(\boldsymbol{q}) = \frac{1}{\lambda + Q_g}. \tag{45}$$

We first characterize the optimum using Lemma 1. Every optimal group must have positive mass. Indeed, if $Q_g = 0$, then $\tau_j = 1/\lambda$ for $j \in \mathcal{G}_g$. Since $\sum_h Q_h = 1$, some group h has $Q_h > 0$, and the sensitivity of any positively weighted span in that group equals

$$\frac{1}{\lambda + Q_h} < \frac{1}{\lambda},$$

contradicting the KKT inequality for group g. Thus $Q_g > 0$ for every g. The KKT equality on the support and (45) then imply

$$\frac{1}{\lambda + Q_1} = \cdots = \frac{1}{\lambda + Q_r},$$

so $Q_1 = \cdots = Q_r$. Since $\sum_g Q_g = 1$,

$$Q_g(\boldsymbol{q}^\star) = \frac{1}{r}, \qquad g = 1, \ldots, r. \tag{46}$$

At the uniform initialization $q_j^{(0)} = 1/m$, the initial mass of group g is $Q_g(\boldsymbol{q}^{(0)}) = m_g/m$. Hence,

$$\tau_j(\boldsymbol{q}^{(0)}) = \frac{1}{\lambda + m_g/m} = \frac{m}{m\lambda + m_g}, \qquad j \in \mathcal{G}_g.$$

Using the one-step update $q_j^{(1)} = \tau_j(\boldsymbol{q}^{(0)}) / \sum_u \tau_u(\boldsymbol{q}^{(0)})$, we obtain

$$Q_g(\boldsymbol{q}^{(1)}) = \frac{m_g/(m\lambda + m_g)}{\sum_{h=1}^{r} m_h/(m\lambda + m_h)}. \tag{47}$$

Define

$$a_g = \frac{m_g}{m\lambda + m_g} = \frac{1}{1 + m\lambda/m_g}, \qquad S = \sum_{h=1}^{r} a_h, \qquad \delta = \frac{m\lambda}{m_{\min}}.$$

Since $m_g \geq m_{\min}$,

$$\frac{1}{1+\delta} \leq a_g \leq 1, \qquad \frac{r}{1+\delta} \leq S \leq r.$$

Together with $Q_g(\boldsymbol{q}^{(1)}) = a_g/S$, this gives

$$\frac{1}{r(1+\delta)} \leq Q_g(\boldsymbol{q}^{(1)}) \leq \frac{1+\delta}{r},$$

and therefore

$$\left| Q_g(\boldsymbol{q}^{(1)}) - \frac{1}{r} \right| \leq \frac{\delta}{r}. \tag{48}$$

It remains to bound the objective gap. The orthogonality of the group directions gives

$$\Phi(\boldsymbol{q}) = (d - r)\log\lambda + \sum_{g=1}^{r} \log(\lambda + Q_g). \tag{49}$$

Let $x_g = Q_g(\boldsymbol{q}^{(1)})$, $\mu = 1/r$, and $f(x) = \log(\lambda + x)$. Since $\sum_g (x_g - \mu) = 0$, Taylor's theorem yields

$$\Phi(\boldsymbol{q}^\star) - \Phi(\boldsymbol{q}^{(1)}) = \frac{1}{2} \sum_{g=1}^{r} \frac{(x_g - \mu)^2}{(\lambda + \xi_g)^2}, \tag{50}$$

where each ξ_g lies between x_g and μ. Since

$$x_g, \mu \geq \frac{1}{r(1+\delta)} \quad \text{and} \quad |x_g - \mu| \leq \frac{\delta}{r},$$

we obtain the sharper bound

$$\Phi(\boldsymbol{q}^\star) - \Phi(\boldsymbol{q}^{(1)}) \leq \frac{\delta^2}{2r\left(\lambda + \frac{1}{r(1+\delta)}\right)^2} \leq \frac{r}{2}(1+\delta)^2\delta^2. \tag{51}$$

Since $\delta = m\lambda/m_{\min} = O(\lambda)$ for fixed group sizes, the group-mass error is $O(\lambda)$ and the objective gap is $O(\lambda^2)$. If all groups have equal size, then $Q_g(\boldsymbol{q}^{(0)}) = 1/r$, so the uniform initialization is already globally optimal. □

Proof of Theorem 2. Let $\widehat{\boldsymbol{q}}_i = \boldsymbol{q}_i^{(1)}$. By concavity of Φ_i,

$$\begin{aligned}\Phi_i(\boldsymbol{q}_i^\star) - \Phi_i(\widehat{\boldsymbol{q}}_i) &\leq \nabla\Phi_i(\widehat{\boldsymbol{q}}_i)^\top(\boldsymbol{q}_i^\star - \widehat{\boldsymbol{q}}_i) \\ &\leq \max_{\boldsymbol{p}_i \in \Delta_{\mathcal{A}_i}} \nabla\Phi_i(\widehat{\boldsymbol{q}}_i)^\top(\boldsymbol{p}_i - \widehat{\boldsymbol{q}}_i) \\ &= G_i(\widehat{\boldsymbol{q}}_i),\end{aligned} \tag{52}$$

where the last equality follows from Lemma 1. The lower bound is zero because $\boldsymbol{q}_i^\star$ is globally optimal. Hence,

$$0 \leq \Phi_i(\boldsymbol{q}_i^\star) - \Phi_i(\boldsymbol{q}_i^{(1)}) \leq G_i(\boldsymbol{q}_i^{(1)}).$$

Finally,

$$\begin{aligned}\left(\frac{\det \boldsymbol{M}_i(\boldsymbol{q}_i^{(1)})}{\det \boldsymbol{M}_i(\boldsymbol{q}_i^\star)}\right)^{1/d} &= \exp\left(\frac{\Phi_i(\boldsymbol{q}_i^{(1)}) - \Phi_i(\boldsymbol{q}_i^\star)}{d}\right) \\ &= \exp\left(-\frac{\Phi_i(\boldsymbol{q}_i^\star) - \Phi_i(\boldsymbol{q}_i^{(1)})}{d}\right).\end{aligned} \tag{53}$$

Applying the objective-gap bound proves

$$\exp\left(-\frac{G_i(\boldsymbol{q}_i^{(1)})}{d}\right) \leq \left(\frac{\det \boldsymbol{M}_i(\boldsymbol{q}_i^{(1)})}{\det \boldsymbol{M}_i(\boldsymbol{q}_i^\star)}\right)^{1/d} \leq 1.$$

□

C Algorithm and Implementation Details

This appendix specifies the training procedure, the construction of the gradient proxy, and the conversion from span allocations to token-level loss coefficients. We use the design-ridge convention of the main text: λ regularizes the probability-weighted design matrix, whereas $\lambda_{\text{impl}} = m_i\lambda$ regularizes the summed covariance of the m_i active spans of problem i.

C.1 Complete training algorithm

Algorithm 1 summarizes a training update. The student generates the trajectories on which the teacher is queried. Allocation is performed separately for each problem, jointly across its rollouts. The resulting coefficients weight the base actor loss and are held fixed during the actor update.

Algorithm 1: DOpt-OPD Training

Input: Prompt distribution $\mathcal{D}$; student p_θ; teacher p_{T}; rollout count R; span length L_{span}; Student-TopK size K; proxy dimensions d_h, d_δ; design ridge $\lambda > 0$

Output: Trained student parameters θ

Initialize fixed hidden-state and correction sketch maps $\boldsymbol{R}_h, \boldsymbol{R}_\delta$;
for *each training update* **do**
  Sample a batch of N token-valid problems from $\mathcal{D}$;
  Generate R student rollouts per problem and obtain base OPD token signals;
  foreach *problem* i **do**
    Partition each rollout into nonoverlapping spans; retain shorter terminal spans;
    Pool the M_i token-valid spans belonging to problem i;
    Sum token proxies within each span to obtain $\boldsymbol{g}_{ij}$;
    Identify $\mathcal{A}_i = \{j : \|\boldsymbol{g}_{ij}\|_2 > 0\}$ and set $m_i = |\mathcal{A}_i|$;
    if $m_i > 0$ **then**
      Normalize active proxies: $\boldsymbol{z}_{ij} = \boldsymbol{g}_{ij}/\|\boldsymbol{g}_{ij}\|_2$;
      Form $\boldsymbol{C}_i = \sum_{j\in\mathcal{A}_i} \boldsymbol{z}_{ij}\boldsymbol{z}_{ij}^\top + m_i\lambda\boldsymbol{I}$;
      Factor $\boldsymbol{C}_i = \boldsymbol{L}_i\boldsymbol{L}_i^\top$ and solve $\boldsymbol{L}_i\boldsymbol{u}_{ij} = \boldsymbol{z}_{ij}$;
      Set $q_{ij} = \|\boldsymbol{u}_{ij}\|_2^2 / \sum_{u\in\mathcal{A}_i} \|\boldsymbol{u}_{iu}\|_2^2$ on $\mathcal{A}_i$;
      Set $q_{ij} = 0$ outside $\mathcal{A}_i$;
    else
      Set $q_{ij} = 1/M_i$ over token-valid spans;
    Assign coefficient $\mathrm{stopgrad}(q_{ij})/(Nn_{ij})$ to each valid token in span j;
  Aggregate the weighted base token losses using the global problem normalization;
  Update θ using the actor optimizer;
return θ;

C.2 Gradient-proxy construction

Student-TopK correction. At a student-visited state s_t, let $p_t(v) = p_\theta(v \mid s_t)$, $p_{\mathrm{T},t}(v) = p_{\mathrm{T}}(v \mid s_t)$, and let $\mathcal{K}_t$ contain the K highest-probability student candidates. The student-probability-weighted direct coefficient is

$$\pi_{tk} = \frac{p_t(k)}{\sum_{u\in\mathcal{K}_t} p_t(u)}, \qquad a_{tk} = \mathrm{stopgrad}\big[\pi_{tk}\big(\log p_{\mathrm{T},t}(k) - \log p_t(k)\big)\big], \quad k \in \mathcal{K}_t. \tag{54}$$

Candidate selection and these direct coefficients are fixed when forming the local proxy. Writing the corresponding local score surrogate as $\ell_t^{\mathrm{loc}} = -\sum_{k\in\mathcal{K}_t} a_{tk}\log p_t(k)$, and setting $A_t = \sum_{k\in\mathcal{K}_t} a_{tk}$, its full logit derivative is

$$\delta_{tv} = p_t(v)A_t - \mathbf{1}\{v \in \mathcal{K}_t\}a_{tv}. \tag{55}$$

The sparse correction used for allocation retains only the candidate coordinates:

$$\widehat{\delta}_{tv} = \begin{cases} p_t(v)A_t - a_{tv}, & v \in \mathcal{K}_t, \\ 0, & v \notin \mathcal{K}_t. \end{cases} \tag{56}$$

The probabilities $p_t(v)$ in both corrections are normalized over the full vocabulary.

LM-head factorization. For hidden state $\boldsymbol{h}_t \in \mathbb{R}^H$ and output matrix $\boldsymbol{W} \in \mathbb{R}^{V\times H}$, logits are $\boldsymbol{W}\boldsymbol{h}_t$ and

$$\nabla_{\boldsymbol{W}}\ell_t^{\mathrm{loc}} = \boldsymbol{\delta}_t\boldsymbol{h}_t^\top, \qquad \mathrm{vec}(\boldsymbol{\delta}_t\boldsymbol{h}_t^\top) = \boldsymbol{h}_t \otimes \boldsymbol{\delta}_t. \tag{57}$$

Consequently, under column-major vectorization,

$$\langle \boldsymbol{h}_t \otimes \boldsymbol{\delta}_t, \boldsymbol{h}_u \otimes \boldsymbol{\delta}_u\rangle = (\boldsymbol{h}_t^\top\boldsymbol{h}_u)(\boldsymbol{\delta}_t^\top\boldsymbol{\delta}_u). \tag{58}$$

This factorization motivates retaining both the hidden state and the teacher correction when approximating update-direction similarity.

Random maps and span aggregation. The hidden-state map $\boldsymbol{R}_h \in \mathbb{R}^{d_h \times H}$ has independent Rademacher entries scaled by $1/\sqrt{d_h}$. The correction map is a CountSketch (Charikar et al., 2002): with bucket map $b : [V] \to [d_\delta]$ and signs $\xi_v \in \{-1, +1\}$,

$$(\boldsymbol{R}_\delta \widehat{\boldsymbol{\delta}}_t)_a = \sum_{v \in \mathcal{K}_t : b(v) = a} \xi_v \widehat{\delta}_{tv}. \tag{59}$$

The maps are fixed across updates. The token proxy and span direction are

$$\widetilde{\boldsymbol{g}}_t = (\boldsymbol{R}_h \boldsymbol{h}_t) \otimes (\boldsymbol{R}_\delta \widehat{\boldsymbol{\delta}}_t), \quad \boldsymbol{g}_{ij} = \sum_{t \in \mathcal{S}_{ij}} \widetilde{\boldsymbol{g}}_t, \quad \boldsymbol{z}_{ij} = \frac{\boldsymbol{g}_{ij}}{\|\boldsymbol{g}_{ij}\|_2}. \tag{60}$$

Summation precedes normalization, so opposing token corrections may cancel within a span. The flattened dimension is $d = d_h d_\delta$. For ideal independent CountSketch buckets and signs and fixed vectors $\boldsymbol{a}, \boldsymbol{b}$, the standard second-moment calculation gives

$$\mathbb{E}\langle \boldsymbol{R}_\delta \boldsymbol{a}, \boldsymbol{R}_\delta \boldsymbol{b} \rangle = \langle \boldsymbol{a}, \boldsymbol{b} \rangle, \qquad \mathrm{Var}[\langle \boldsymbol{R}_\delta \boldsymbol{a}, \boldsymbol{R}_\delta \boldsymbol{b} \rangle] \le \frac{2}{d_\delta} \|\boldsymbol{a}\|_2^2 \|\boldsymbol{b}\|_2^2. \tag{61}$$

C.3 Budget preservation and edge cases

Let n_{ij} be the number of valid tokens in span $\mathcal{S}_{ij}$ and $\bar{\ell}_{ij} = n_{ij}^{-1} \sum_{t \in \mathcal{S}_{ij}} \ell_t$. The training objective is

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \mathrm{stopgrad}(q_{ij}) \bar{\ell}_{ij}. \tag{62}$$

Thus each valid token in span j has coefficient $q_{ij}/(N n_{ij})$, and

$$\sum_{j=1}^{M_i} \sum_{t \in \mathcal{S}_{ij}} \frac{q_{ij}}{N n_{ij}} = \frac{1}{N} \sum_{j=1}^{M_i} q_{ij} = \frac{1}{N}. \tag{63}$$

Spans are formed separately within each rollout, using the actual valid-token count for shorter terminal spans and zero weight for padding. Allocation uses the nonzero span proxies, with uniform weights $q_{ij} = 1/M_i$ over token-valid spans if all proxies vanish. We exclude problems with no valid response tokens and use the retained global problem count for N.

Gradients pass through the student token losses, with the allocation computation and weights detached. If distributed gradient reduction averages over W workers, local summed losses using global coefficients are scaled by W as needed to recover the gradient of (62). Microbatching and gradient accumulation use the same global normalization.

C.4 Efficient computation

For the uniform active-span design $q_{ij}^{(0)} = 1/m_i$,

$$\boldsymbol{C}_i = \sum_{j \in \mathcal{A}_i} \boldsymbol{z}_{ij} \boldsymbol{z}_{ij}^\top + \lambda_{\mathrm{impl}} \boldsymbol{I}, \qquad \lambda_{\mathrm{impl}} = m_i \lambda, \qquad \boldsymbol{M}_i(\boldsymbol{q}_i^{(0)}) = \frac{\boldsymbol{C}_i}{m_i}. \tag{64}$$

Hence $\tau_{ij}^{(0)} = m_i \widetilde{\tau}_{ij}$, and the factor m_i cancels when scores are normalized. With $\boldsymbol{C}_i = \boldsymbol{L}_i \boldsymbol{L}_i^\top$,

$$\widetilde{\tau}_{ij} = \boldsymbol{z}_{ij}^\top \boldsymbol{C}_i^{-1} \boldsymbol{z}_{ij} = \|\boldsymbol{L}_i^{-1} \boldsymbol{z}_{ij}\|_2^2. \tag{65}$$

A single Cholesky factorization followed by triangular solves avoids explicit matrix inversion. Covariance formation and score evaluation cost $O(m_i d^2)$; factorization costs $O(d^3)$. The allocation solve therefore costs $O(m_i d^2 + d^3)$ per problem, with $O(d^2)$ matrix storage and $O(m_i d)$ storage if all active proxies are materialized. These costs exclude construction of token sketches. Hidden states and teacher signals are reused from the distillation forward passes, so no additional teacher forward pass is needed. Norms, solves, and normalized masses are checked for finiteness.

D Additional Benchmark Comparisons and Ablations

D.1 Combination experiments

To examine whether DOpt can be combined with other distillation methods, we apply its allocation rule to TIDE (Yu et al., 2026) and compare OPD, TIDE, and TIDE+DOpt at the two main student scales. Table 5 reports the seven-task results.

Combining TIDE with DOpt. We retain TIDE's loss and construct the gradient proxy from its masked correction coefficients s_{tc} over the sampled token and teacher top-16 candidates. With student probability π_{tc}, the correction is $\delta_{tc}^{\mathrm{TIDE}} = \pi_{tc} \sum_u s_{tu} - s_{tc}$. All valid response tokens are grouped into 32-token spans. Within each problem, DOpt normalizes the span D-leverage scores once and distributes each span's mass uniformly across its valid tokens. These weights multiply the TIDE loss $-\sum_c s_{tc} \log \pi_{tc}$ while preserving the fixed problem budget. TIDE's selection is encoded by zero coefficients; response positions remain in the span pool, and the selected coefficients retain their original scale.

Table 5: **OPD, TIDE, and TIDE+DOpt.** Entries are Avg@8 / Pass@8 in percentages. Bold marks the best Avg@8 and Pass@8 separately within each model scale, including ties.

Method	AIME24	AIME25	AMC23	Minerva	Olympiad	CMIMC25	BRUMO25	Avg.
Qwen3-8B teacher → Qwen3-4B student								
OPD	16.25/30.00	15.00/33.33	53.13/82.50	32.35/47.06	**43.14**/63.50	5.62/**22.50**	22.08/40.00	26.80/45.56
TIDE	15.83/33.33	16.67/**36.67**	**56.88/85.00**	33.69/**48.16**	42.58/63.20	**9.06/22.50**	21.25/33.33	27.99/46.03
TIDE+DOpt	**19.58/36.67**	**18.33/36.67**	53.75/**85.00**	**33.92**/47.43	42.99/**63.95**	6.25/20.00	**24.17/43.33**	**28.43/47.58**
Qwen3-4B teacher → Qwen3-1.7B student								
OPD	7.92/23.33	5.42/13.33	45.00/67.50	19.85/37.87	31.34/50.45	2.19/10.00	16.25/26.67	18.28/32.74
TIDE	11.67/26.67	8.33/16.67	**47.81**/77.50	**21.92/40.44**	34.44/56.23	5.00/15.00	**17.08/33.33**	20.89/37.98
TIDE+DOpt	**13.33/30.00**	**9.58/33.33**	44.38/**82.50**	21.74/38.97	**35.87/57.86**	**5.94/20.00**	15.83/**33.33**	**20.95/42.28**

Relative to TIDE, adding DOpt increases macro Avg@8/Pass@8 by 0.44/1.55 percentage points at 4B and 0.06/4.30 percentage points at 1.7B.

D.2 Evaluation at a smaller student scale

We also evaluate DOpt-OPD with a Qwen3-4B teacher and a Qwen3-0.6B student. Table 6 compares the undistilled Student initialization, standard OPD, and DOpt-OPD across seven benchmarks, using eight sampled responses per question.

Table 6: **Supplementary 0.6B student comparison.** Entries are Avg@8 / Pass@8 (%). Bold marks the best value for each metric, including ties.

Method	AIME24	AIME25	AMC23	Minerva	Olympiad	CMIMC25	BRUMO25	Avg.
Student	0.00/0.00	**0.42/3.33**	5.63/32.50	2.80/14.71	2.86/16.32	0.00/0.00	0.83/6.67	1.79/10.50
OPD	1.25/**6.67**	**0.42/3.33**	11.25/40.00	6.43/23.53	8.94/29.08	**0.31/2.50**	2.08/10.00	4.38/16.44
DOpt-OPD	**2.92/6.67**	**0.42/3.33**	**28.75/52.50**	**11.53/29.04**	**16.75/35.01**	**0.31/2.50**	**5.00/13.33**	**9.38/20.34**

At 0.6B, DOpt-OPD improves macro Avg@8/Pass@8 over OPD by 5.00/3.90 percentage points. The largest gains are on AMC23 (17.50/12.50 percentage points), while AIME25 and CMIMC25 are unchanged.

D.3 Component Ablations

The per-task ablation table expands the aggregate comparison in the main text, comparing OPD, problem normalization, uniform span allocation, and DOpt-OPD.

Table 7: **Per-task component ablations at 1.7B.** Entries are Avg@8 / Pass@8 (%). Avg. gives equal weight to the seven benchmarks. Bold marks the best value for each metric, including ties.

Method	AIME24	AIME25	AMC23	Minerva	Olympiad	CMIMC25	BRUMO25	Avg.
OPD	7.92/23.33	5.42/13.33	45.00/67.50	19.85/37.87	31.34/50.45	2.19/10.00	16.25/26.67	18.28/32.74
Problem normalization	12.08/23.33	**10.00/23.33**	44.06/65.00	20.63/**38.97**	34.77/55.34	3.12/15.00	15.83/**33.33**	20.07/36.33
Uniform span allocation	10.42/23.33	9.58/**23.33**	43.75/70.00	20.40/38.60	34.37/**56.68**	3.75/15.00	15.83/**33.33**	19.73/37.18
DOpt-OPD	**12.50/30.00**	**10.00/23.33**	**45.63/77.50**	**21.05**/38.60	**35.00**/54.90	**5.62/17.50**	**17.08**/30.00	**20.98/38.83**

Table 8: Maximum direction-group mass error before and after one update. Errors are in probability-mass units; reductions are relative to the initial error.

Setting	$E_Q^{(0)}$	$E_Q^{(1)}$	Reduction
Balanced	0	0	—
Moderate	1.3393×10^{-2}	1.1875×10^{-3}	91.1%
Severe	2.0089×10^{-2}	3.2233×10^{-3}	84.0%

DOpt-OPD improves the seven-benchmark Avg@8/Pass@8 over problem normalization by 0.91/2.50 points and over uniform span allocation by 1.25/1.65 points. Its largest Pass@8 advantage over either control occurs on AMC23. Both controls achieve higher Pass@8 on OlympiadBench and BRUMO25, while problem normalization also scores higher on Minerva Pass@8.

E Empirical Validation of the Theoretical Analysis

Theorem 1 predicts that the optimal allocation assigns equal total mass to orthogonal direction groups, regardless of their repeat counts, and that one update approaches this allocation when the ridge is sufficiently small. We examine this behavior by varying the imbalance in repeat counts while keeping the direction set, total span count, and ridge fixed. This construction isolates the effect of directional repetition on the initial allocation and its correction by one DOpt update.

Experimental setup. We generate a Gaussian random matrix $\boldsymbol{A} \in \mathbb{R}^{64\times 64}$ with independent standard-normal entries and compute its QR decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$. The 64 columns of $\boldsymbol{Q}$ define the unit directions for the direction groups, with $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}$. We use $r_i = d = 64$, $m_i = 896$ spans, and design ridge $\lambda = 10^{-3}$. Each span belongs to one direction group. The Balanced setting repeats every direction 14 times. In the Moderate setting, four categories of 16 directions have repeat counts $(26, 16, 9, 5)$. In the Severe setting, the corresponding counts are $(32, 16, 6, 2)$. All three settings therefore have the same dimension, number of distinct directions, and total number of spans. The Balanced setting checks whether an already optimal allocation is preserved; the Moderate and Severe settings test how much one update corrects the allocation bias induced by unequal repeat counts.

Uniform allocation over spans gives group mass $Q_{ig}^{(0)} = m_{ig}/m_i$. For this QR construction, the optimal group allocation is $Q_{ig}^\star = 1/64 = 1.5625\%$. The one-step result is

$$Q_{ig}^{(1)} = \frac{m_{ig}/(m_i\lambda + m_{ig})}{\sum_{h=1}^{64} m_{ih}/(m_i\lambda + m_{ih})}. \tag{66}$$

For $t \in \{0, 1\}$, we summarize allocation error by

$$E_Q^{(t)} = \max_g \left| Q_{ig}^{(t)} - \frac{1}{64} \right|. \tag{67}$$

Each plotted value reports the mass of one direction in the corresponding repeat-count category.

One-step allocation results. In the Balanced setting, the initial allocation is already optimal, and one update preserves the mass of 1.5625% per direction (Figure 3). In the Moderate and Severe settings, uniform span allocation initially gives more mass to frequently repeated directions. One DOpt update redistributes mass toward less frequent directions, reducing the maximum group-mass error by 91.1% and 84.0%, respectively (Table 8). These results show that one-step allocation counteracts repeat-count imbalance and moves the group masses toward the equal-direction optimum predicted by Theorem 1.

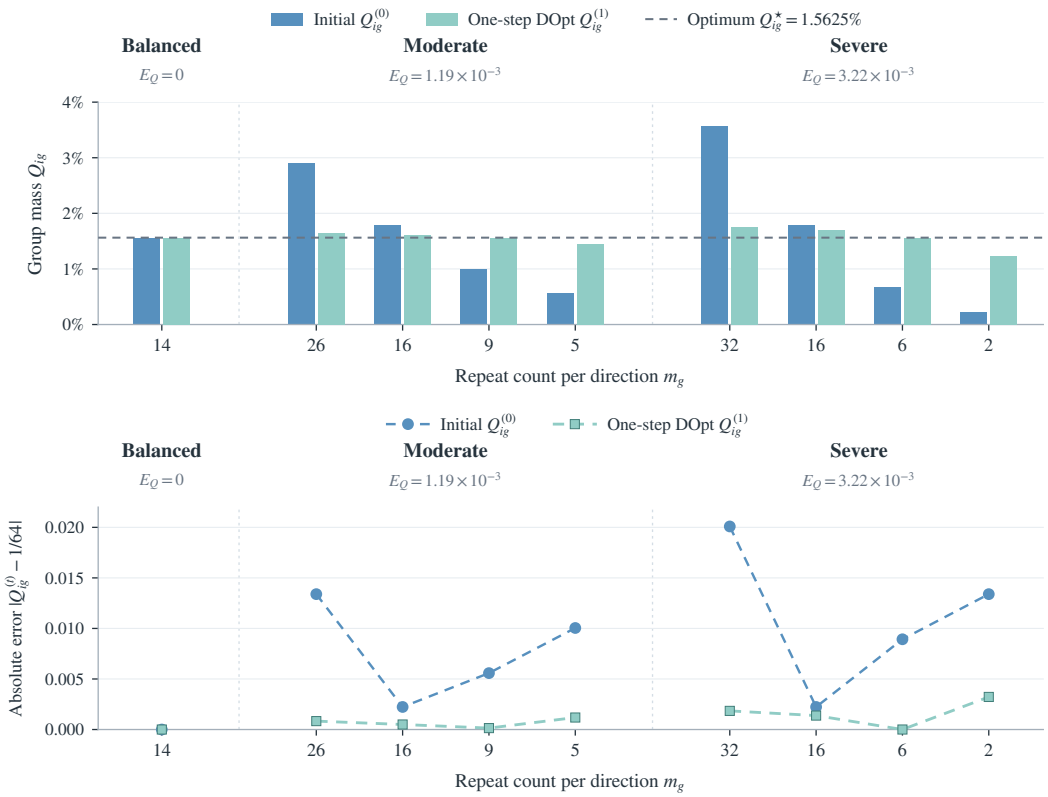

Figure 3: **One-step allocation for QR-generated direction groups.** Top: initial and one-step direction-group masses, with the optimum $1/64 = 1.5625\%$ shown by the dashed line. Bottom: absolute deviation from the optimal group mass. Balanced repeats each of 64 directions 14 times; Moderate and Severe each contain four categories of 16 directions. All settings use $m_i = 896$ and $\lambda = 10^{-3}$. The annotation E_Q is the maximum post-update error in probability-mass units, whereas the upper axis reports percentages.

F Mechanism Analysis and Additional Diagnostics

The benchmark comparisons establish the performance of DOpt-OPD; the analyses below examine the allocation behavior motivating its design. We first quantify length-dependent allocation and the budget assigned to repeated text on fixed OPD rollouts. We then examine whether recurrent content corresponds to overlapping proxy update directions, and test whether DOpt preserves prefix supervision when a repeated suffix grows. We also report response length and truncation as supplementary generation statistics.

F.1 Fixed-rollout supervision-allocation audit

We first quantify the two allocation issues discussed in Section 3.1: how strongly token averaging concentrates supervision on long-response problems, and how much budget repeated text receives after problem budgets are equalized. Evaluating both rules on fixed responses isolates the effect of the weighting rule.

Figure 1 uses a fixed pool of 17,917 DAPO-Math-17K prompts with four rollouts per prompt, generated by the standard-OPD Qwen3-1.7B checkpoint after 139 updates with a Qwen3-4B teacher. Both allocation rules are evaluated on the same 71,668 responses, with a maximum response length of 7,168 tokens.

For the across-problem comparison, let $L_i = \sum_{r=1}^{R} T_{ir}$ be the total valid response length of problem i. We rank problems by L_i and divide them into four nearly equal-count groups $\mathcal{I}_g$. The token-averaging

budget share of group g is

$$B_g = \frac{\sum_{i \in \mathcal{I}_g} L_i}{\sum_{i=1}^{N_\mathrm{a}} L_i}, \qquad N_\mathrm{a} = 17{,}917. \tag{68}$$

The shares are 3.5%, 11.1%, 24.7%, and 60.7% in increasing length order; equal problem budgets give approximately 25% per group. This calculation normalizes over the fixed audit pool.

For the within-problem comparison, let D_{ir} count valid response tokens flagged by the 8-gram repetition diagnostic in rollout r of problem i. The fraction of each problem's budget assigned to flagged tokens under token averaging and problem normalization is, respectively,

$$\rho_i^{\mathrm{tok}} = \frac{\sum_{r=1}^{R} D_{ir}}{\sum_{r=1}^{R} T_{ir}}, \qquad \rho_i^{\mathrm{PN}} = \frac{1}{R} \sum_{r=1}^{R} \frac{D_{ir}}{T_{ir}}. \tag{69}$$

For either allocation rule a, the plotted CCDF is $\widehat{S}_a(x) = N_\mathrm{a}^{-1} \sum_i \mathbf{1}\{\rho_i^a > x\}$, including problems with no flagged tokens. The median share is 26.4% under token averaging and 10.8% under problem normalization; 53.0% of problems have $\rho_i^{\mathrm{PN}} > 0.10$. The audit shows that problem normalization reduces the budget share assigned to repeated text, while substantial repeated-text allocation remains for many problems. This motivates examining the geometry of recurrent content to determine whether such allocation also coincides with overlapping proxy update directions.

F.2 Directional redundancy in recurrent motifs

The fixed-rollout audit measures textual recurrence, whereas DOpt allocates supervision according to update-direction geometry. To connect these two quantities, we test whether recurrent reasoning windows have more similar proxy update directions than matched controls. We measure cosine similarities between hidden states $\boldsymbol{h}_t$, sparse corrections $\widehat{\boldsymbol{\delta}}_t$, and joint LM-head proxies $\boldsymbol{g}_t = \boldsymbol{h}_t \otimes \widehat{\boldsymbol{\delta}}_t$. For aligned positions t and u, define

$$\begin{aligned} C_h(t,u) &= \frac{\boldsymbol{h}_t^\top \boldsymbol{h}_u}{\|\boldsymbol{h}_t\|_2 \|\boldsymbol{h}_u\|_2}, \\ C_\delta(t,u) &= \frac{\widehat{\boldsymbol{\delta}}_t^\top \widehat{\boldsymbol{\delta}}_u}{\|\widehat{\boldsymbol{\delta}}_t\|_2 \|\widehat{\boldsymbol{\delta}}_u\|_2}, \\ C_g(t,u) &= \frac{\boldsymbol{g}_t^\top \boldsymbol{g}_u}{\|\boldsymbol{g}_t\|_2 \|\boldsymbol{g}_u\|_2} = C_h(t,u) C_\delta(t,u). \end{aligned} \tag{70}$$

The identity holds for nonzero factors (Appendix C). Here $\widehat{\boldsymbol{\delta}}_t$ retains the candidate coordinates of the local surrogate's logit correction, so $\boldsymbol{g}_t$ precedes random projection and span aggregation.

We audit 128 DAPO prompts, totaling 60,259 valid spans. The comparison includes 16 recurrent and 16 control trajectories, yielding 834 recurrent-motif pairs and 209 matched-control pairs. Controls are matched without replacement on relative position, student–teacher KL, teacher entropy, and trajectory length to reduce differences associated with position, disagreement, uncertainty, and response length.

For each pair of aligned 32-token windows, we compute C_h, C_δ, and their product C_g at each offset. We take each measure's median over valid offsets, then its median across window pairs.

Table 9: **Directional similarity in recurrent motifs.** Entries are medians across window pairs of within-pair medians over aligned token positions.

Cosine measure	Recurrent motifs
Hidden state: C_h	0.956
Sparse correction: C_δ	0.770
Joint LM-head proxy: C_g	0.716

Table 9 shows high directional similarity in recurrent motifs. Matched controls have a median hidden-state cosine similarity of 0.258, while their median sparse-correction and joint LM-head proxy

cosine similarities are both near zero. The joint-proxy comparison therefore links recurrent content to overlapping proxy update directions in the audited windows, providing a geometric motivation for accounting for redundancy when allocating supervision across spans.

F.3 Budget retention under repeated suffixes

We next test whether the allocation rule preserves supervision of an existing prefix when additional repeated content competes for a fixed problem budget. Under token-uniform allocation, extending a suffix increases the token-count denominator and reduces the prefix's budget share. To measure how DOpt responds to this intervention, we expand a recurrent suffix in one of each problem's four rollouts, keeping the original prefix and the other three rollouts fixed. The suffix occupies approximately 25%, 50%, or 75% of the modified response. We define dose as the suffix token count divided by the combined prefix and suffix length, round repetition counts to integers, and exclude constructions whose actual dose differs from the target by more than 0.05.

For each constructed context, we recompute hidden states, sparse corrections, and teacher signals before allocating the problem budget across all four rollouts. Nine cases are feasible at all three doses and define the common comparison set. All nine have an actual dose of 0.75 at the highest dose.

The token-uniform baseline assigns each valid token weight $1/\sum_{r=1}^{4} T_{ir}(\rho)$ within problem i, corresponding to span mass $q_{ij}(\rho) = n_{ij}(\rho)/\sum_{r=1}^{4} T_{ir}(\rho)$. DOpt-OPD distributes the same unit problem budget across spans pooled from all four rollouts. Fixing this budget focuses the comparison on within-problem redistribution. For either method, the allocated mass of the fixed original prefix P_i and its retention at dose ρ are

$$B_i(\rho) = \sum_j q_{ij}(\rho) \frac{|\mathcal{S}_{ij}(\rho) \cap P_i|}{n_{ij}(\rho)}, \qquad R_i(\rho) = \frac{B_i(\rho)}{B_i(0.25)}. \tag{71}$$

The overlap fraction accounts for spans that straddle the prefix boundary. The 25% dose supplies the reference allocation, with $B_i(0.25) > 0$.

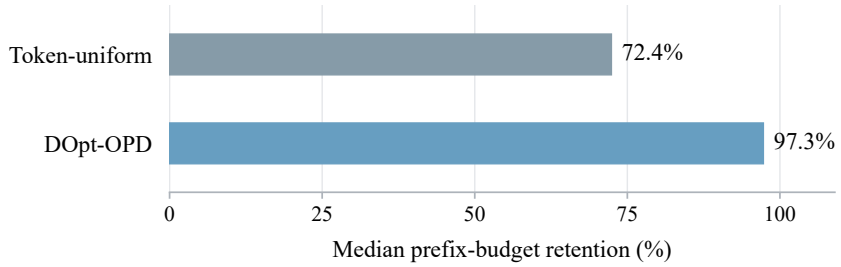

Figure 4: **Prefix-budget retention under repeated suffixes.** Bars show median case-wise retention across nine cases at 75% suffix dose, relative to 25%. Both methods allocate the same unit problem budget across four rollouts, with only the target response modified.

At the 75% dose, DOpt-OPD retains a median of 97.3% of the prefix budget allocated at the 25% dose, compared with 72.4% under token-uniform allocation (Figure 4). Thus, in these cases, DOpt substantially reduces the dilution of existing prefix supervision as the repeated suffix grows. This directly tests the allocator's response to repetition under a fixed problem budget.

F.4 Generation diagnostics

Table 10 reports mean response length and truncation rate, averaged equally over seven benchmarks. Truncation is the percentage of responses that reach the generation limit without natural completion.

DOpt-OPD produces shorter responses with lower truncation than both equal-budget controls. Relative to standard OPD, its mean response length increases from 6,790 to 7,828 tokens, while truncation decreases from 17.09% to 14.83%.

Table 10: **Generation length and truncation.** Tokens denotes mean response length; Trunc. denotes truncation rate (%). Results are averaged equally over seven benchmarks. Bold marks the lowest truncation rate.

Method	Tokens	Trunc.
OPD	6,790	17.09
Problem normalization	8,801	17.92
Uniform span allocation	8,749	18.03
DOpt-OPD	7,828	**14.83**

G Unified Hyperparameter Settings

All models use the unified default hyperparameters listed in Table 11 for training, rollout generation, and evaluation. The default DOpt parameters are also shared across model scales.

Table 11: Default training, evaluation, and DOpt hyperparameters.

Hyperparameter	Value
Data and batching	
Training dataset	DAPO-Math-17K
Prompt batch size	128
Rollouts per prompt	4
Policy mini-batch size	128
Micro-batch size per GPU	1
Training epochs	1
Optimizer steps	139
Optimization	
Optimizer	AdamW
Learning rate	2×10^{-6}
LR schedule / warmup	Constant / 0
Adam coefficients (β_1, β_2)	$(0.9, 0.999)$
Weight decay	0.01
Student-TopK size K	16
Generation and evaluation	
Maximum prompt length	1,024 tokens
Maximum training response length	7,168 tokens
Training temperature / top-p	1.0 / 1.0
Teacher scoring temperature	1.0
Repetition penalty	1.0
Evaluation samples per question	8
Maximum evaluation response length	31,744 tokens
Evaluation temperature / top-p	0.7 / 0.95
DOpt parameters	
Span length L_{span}	32 tokens
Hidden projection dimension d_h	8
Correction sketch dimension d_δ	8
Design ridge λ	10^{-3}
Allocation updates	1

H LIMITATIONS

Our experiments focus on mathematical reasoning with Qwen3 teacher–student pairs; extending evaluation to other model families and task domains would clarify the broader applicability of DOpt-OPD. The allocation relies on low-dimensional sketches of restricted LM-head gradient proxies, whose fidelity affects the representation of directional redundancy. The single-step update also approximates the D-optimal allocation, with accuracy depending on the span geometry and ridge parameter.