MITIGATING CAPABILITY DEGRADATION IN LLM SELF-IMPROVEMENT VIA REPRESENTATION GEOMETRY REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-Policy Self-distillation is a self-improvement paradigm that enables large language models (LLMs) to improve their capabilities by leveraging the same model as both teacher and student under different conditioning contexts. However, we find that OPSD methods can progressively degrade previously acquired capabilities: queries that are correctly responded at earlier training stages may become incorrectly responded as optimization proceeds, even the overall performance continues to improve. Further analysis reveals that this degradation coincides with a deterioration of the geometry of hidden representations, characterized by reduced intra-cluster compactness and diminished inter-cluster separation. Motivated by this observation, we propose Representation Geometry Regularization (RGR), which preserves representation geometry during post-training by encouraging compactness within cluster and separation across cluster. RGR can be seamlessly integrated into existing OPSD methods without altering their core optimization procedures. Extensive experiments on multiple OPSD methods and benchmarks show that RGR generally alleviates capability degradation while retaining capability improvements. Our code is available in `https://anonymous.4open.science/r/RGR-782F`

1 INTRODUCTION

On-Policy Self-Distillation (OPSD) (Zhao et al., 2026; Hübotter et al., 2026; Yang et al., 2026; Li et al., 2026), is a self-improvement paradigm for the post-training of large language models (LLMs) (Brown et al., 2020; Vaswani et al., 2017; Yang et al., 2025; Zeng et al., 2026; Xiao et al., 2026; Xu et al., 2026), which enables LLMs to improve thier capabilities through self-knowledge distillation. Unlike conventional knowledge distillation(Hinton et al., 2015; Sanh et al., 2019), which relies on a separately trained teacher, OPSD uses the same LLM as both teacher and student under different conditioning contexts. Specifically, the student model generates responses to the original query, while teacher model, which share the same parameters with the student model, conditioned on additional privileged information (e.g., verified reasoning traces) to provide supervision for these generated responses.

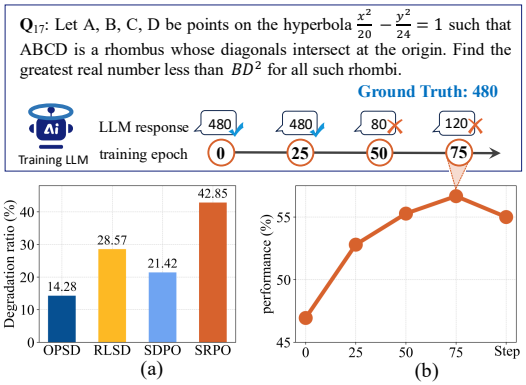


Figure 1: (a) The capability degradation ratio of existing methods; (b) Although the overall performance improves during training, some problems that were previously responded correctly become incorrectly answered as optimization progresses.

Although recent studies (Zhao et al., 2026; Hübotter et al., 2026; Yang et al., 2026; Li et al., 2026; Wang et al., 2026) have empirically demonstrated that optimizing the LLM with OPSD can outperform conventional post-training methods such as SFT (Wei et al., 2021) and GRPO (Shao et al., 2024), benefiting from the dense and fine-grained token-level supervision of OPSD. We observe a progressive reasoning capability degradation phenomenon during self-improvement in existing

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.

OPSD methods. As shown in Figure 1(b), a query that is responded correctly at earlier training stages gradually becomes incorrectly responded as training proceeds, even as overall performance continues to improve.

Furthermore, we examine several representative OPSD methods and find that this phenomenon consistently occurs across these methods, albeit to varying degrees, as it is shown in Figure 1 (a). Such degradation may undermine the stability of self-improvement.

To better understand this phenomenon, we conduct further analysis, and the results reveal that this capability degradation is accompanied by a progressive deterioration in the geometry of hidden representations. When a query is answered correctly at earlier stages of training, the hidden representations of its generated responses exhibit a relatively well-structured clustering pattern, characterized by larger inter-cluster distances and clearer cluster boundaries. As training proceeds and the response to the same query gradually becomes incorrect, the relative distances between clusters shrink and the cluster boundaries become increasingly indistinct. This observation suggests that preserving the geometric structure of the representation space may be important for maintaining previously acquired capabilities during self-improvement.

Based on this finding, we theoretically show that preserving intra-cluster compactness and inter-cluster separation in the hidden representation space helps maintain the capabilities that the LLM has already acquired. Based on this theoretical insight, we develop Representation Geometry Regularization (RGR), which constrains the evolution of hidden representations during self-improvement. RGR incorporates two geometry-preserving constraints called intra-cluster compactness and inter-cluster separation into the OPSD methods, enabling the model to improve targeted capabilities while mitigating the degradation of previously acquired ones. Experiments across four existing SD models and three widely used benchmarks show that RGR substantially reduces correct-to-incorrect transitions during training while retaining its self-improvement gains.

In summary, our contributions are summarized as follows:

- We uncover an overlooked failure mode, where OPSD methods can degrade previously acquired capabilities, causing previously correct predictions to become incorrect.
- We identify representation geometry distortion as a key factor underlying this degradation, and theoretically characterize the importance of hidden representation geometry for prediction stability.
- We propose RGR, which incorporates geometry-preserving constraints into OPSD methods to enable self-improvement while reducing capability degradation.

2 Related works

On-policy self-distillation (OPSD) enables LLMs to improve their capabilities without requiring an external model as a teacher. Recently, it has attracted significant attention as a promising paradigm for scalable and efficient LLM self-improvement. Based on how self-teacher signals are constructed and how external feedback is incorporated, we organize existing OPSD methods into privileged-information-driven methods and external-feedback-driven methods. The privileged-information-driven methods construct self-teacher signals using additional privileged information that is unavailable to the student policy at inference time. As a representative method, OPSD (Zhao et al., 2026) uses the same model to construct self-teacher signals and optimize the student policy under different conditioning contexts: the self-teacher signals are constructed from both the question and a verified reasoning trace, whereas the student policy is conditioned only on the question. The external-feedback-driven methods retain the self-teacher structure while incorporating environmental feedback, verification results, or outcome rewards to construct the supervision for policy updating. Within this category, feedback can be used to construct distillation targets, select training objectives, or modulate token-level advantages. SDPO (Hübotter et al., 2026) conditions the current model on textual or reward-derived feedback to form a self-teacher, whose next-token predictions serve as distillation targets for the student policy. SRPO (Li et al., 2026)routes successful trajectories to reward-driven GRPO updates and failed trajectories to entropy-weighted, token-level self-distillation against a feedback-conditioned teacher distribution. RLSD (Yang et al., 2026) uses verifiable outcomes to determine the direction of policy updates and token-level teacher–student discrepancies to modulate their magnitude, integrating self-distillation into reward-driven learning.

3 PRELIMINARIES

Given a query q, an LLM π_θ typically generates a response $y = (y_1, \ldots, y_T)$ according to $y_t \sim \pi_\theta(\cdot \mid q, y_{<t})$, where $y_{<t} = (y_1, \ldots, y_{t-1})$ denotes the previously generated tokens. OPSD methods improve the LLM by using the same model under different conditioning contexts as the teacher and student to conduct knowledge distillation. Specifically, the student is designed to generate an on-policy response based only on the original query, like the LLM, *i.e.*, $y \sim \pi_\theta(\cdot \mid q)$. While the teacher is designed to generate the response with additional privileged information z, such as a verified reasoning trace, $\pi_\theta(\cdot \mid q, z, y_{<t})$. The OPSD methods optimize the model by minimizing the divergence between the teacher and student token distributions along the student's trajectory:

$$\mathcal{L}_{\text{OPSD}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|q)} \left[\sum_{t=1}^{T} D\left(\pi_\theta(\cdot \mid q, z, y_{<t}) \,\|\, \pi_\theta(\cdot \mid q, y_{<t})\right) \right]. \tag{1}$$

where $D(\cdot|\cdot)$ denotes a divergence measure between the teacher and student token distributions. D can be instantiated by forward KL divergence (Kullback & Leibler, 1951), reverse KL divergence (Minka et al., 2005), Jensen–Shannon divergence (Lin, 1991), or other suitable distribution-matching objectives.

Since the teacher and student share the underlying model parameters, optimizing $\mathcal{L}_{\text{OPSD}}$ updates the LLM itself. Consequently, the model iteratively improves its responses by distilling the additional information available to the teacher into its ordinary, query-conditioned behavior.

4 THE GEOMETRY OF CAPABILITY DEGRADATION

4.1 PROGRESSIVE CAPABILITY DEGRADATION

As discussed previously, although OPSD methods continuously improve the overall performance in the training, some queries that are correctly responded at earlier training stages become incorrectly responded as training proceeds. To better understand this phenomenon, for a fixed query, we generate a response and visualize the hidden representations corresponding to the generated response every 25 training epochs. We observe that changes in the geometry of the hidden representations consistently coincide with changes in response correctness. As illustrated in Figure. 2, when the model generates a correct response at an earlier training stage (Epoch 25), the corresponding hidden representations exhibit a relatively structured geometric organization, with clearly distinguishable clusters and relatively large distances between clusters. As training proceeds and the response becomes incorrect, the relative distances between clusters gradually decrease, while the boundaries between clusters become increasingly indistinct (Epoch 75). We hypothesize that as a query transitions from being correctly to incorrectly responded, the reasoning process in the generated response gradually becomes less coherent. When the query is answered correctly, the reasoning tends to exhibit a clear and consistent logical structure, which is reflected in the hidden representation space as distinct and well-separated clusters corresponding to different reasoning patterns. As the reasoning becomes increasingly disorganized, these patterns become less distinguishable, causing the corresponding representations to become more entangled and eventually leading to an incorrect response. Interestingly, this phenomenon is consistently observed in existing OPSD methods, such as OPSD(Zhao et al., 2026), SDPO(Hübotter et al., 2026), RLSD(Yang et al., 2026), SRPO(Li et al., 2026). These observations suggest a strong connection between the geometry of the hidden representation and the stability of the model's acquired capabilities during self-improvement.

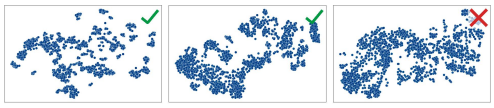


Figure 2: Hidden-state representations and response correctness across training epochs. The query is answered correctly at epochs 25 and 50, but incorrectly at epoch 75.

4.2 THEORETICAL ANALYSIS FOR CAPABILITY DEGRADATION

To further substantiate the relationship between hidden representation geometry and response correctness, we conduct a theoretical analysis.

Suppose the LLM undergoing on-policy self-distillation is denoted by $\pi_\theta(\cdot \mid q, y_{<t})$, where $\pi_\theta \in \Pi_{\text{OPSD}}$, and Π_{OPSD} denote the policy set trained using OPSD.

Let $\pi_{\theta*}(\cdot \mid q, z, y_{<t})$ denote the optimal teacher distribution conditioned on the privileged information z, and let $\pi_{\theta*}(\cdot \mid q, y_{<t})$ denote the corresponding optimal student distribution without access to z. We define the excess distillation risk as follows.

Definition 4.1 (The Excess Distillation Risk)**.** *The excess distillation risk of the student model $\pi_\theta(\cdot \mid q, y_{<t})$ is defined as the difference between its distillation risk with respect to the optimal teacher and that achieved by the optimal student, i.e.,*

$$\mathcal{E}_{KD}(\pi_\theta) = \mathbb{E}_{q\sim\mathcal{D}, y\sim\pi_\theta(\cdot|q)} \Bigg[\sum_{t=1}^{T} D\left(\pi_{\theta^*}(\cdot \mid q, z, y_{<t}) \,\|\, \pi_\theta(\cdot \mid q, y_{<t})\right) - D\left(\pi_{\theta^*}(\cdot \mid q, z, y_{<t}) \,\|\, \pi_{\theta^*}(\cdot \mid q, y_{<t})\right). \tag{2}$$

This quantity measures the gap between the distillation risk of the current student and that of the optimal student, with a smaller value indicating a closer approximation to the optimal.

Let $\hat{\mathcal{H}} = \{\hat{\mathbf{h}}_i\}_{i=1}^T$ and $\mathcal{H}^* = \{\mathbf{h}_i^*\}_{i=1}^T$ denote the hidden representations produced by the current teacher $\pi_\theta(\cdot \mid q, z, y_{<t})$ and the optimal teacher $\pi_{\theta^*}(\cdot \mid q, z, y_{<t})$, respectively. Their midpoint representations are denoted by $\mathcal{H}^\dagger = \{\mathbf{h}_i^\dagger\}_{i=1}^T$, where $\mathbf{h}_i^\dagger = \frac{(\hat{\mathbf{h}}_i + \mathbf{h}_i^*)}{2}$. Suppose that $\mathcal{H}^\dagger$ is partitioned into K clusters, denoted by $\mathcal{C}^\dagger = \{C_1^\dagger, \ldots, C_K^\dagger\}$. The center of is cluster $C_k^\dagger$ is $\mu_k^\dagger = \frac{1}{|C_k^\dagger|} \sum_{\mathbf{h}_i^\dagger \in C_k^\dagger} \mathbf{h}_i^\dagger$. Let $\mathcal{H} = \{\mathbf{h}_i\}_{i=1}^T$ denote the student hidden representations corresponding to the current-teacher representations $\{\hat{\mathbf{h}}_i\}_{i=1}^T$. For each cluster $C_k^\dagger$, we define $\mathcal{X}_k = \{\mathbf{h}_i \mid \mathbf{h}_i^\dagger \in C_k^\dagger\}$. That is, if the midpoint $\mathbf{h}_i^\dagger$ associated with a current-teacher representation $\hat{\mathbf{h}}_i$ belongs to $C_k^\dagger$, its corresponding student representation $\mathbf{h}_i$ is assigned to $\mathcal{X}_k$. The center of is cluster $\mathcal{X}_k$ is $\mu_k^{\mathcal{X}} = \frac{1}{|\mathcal{X}_k|} \sum_{h_i \in \mathcal{X}_k} \mathbf{h}_i$. Let $\mu = \frac{1}{T} \sum_i^T \mathbf{h}_i$ denote the mean of hidden representations. Based on the definition, we have

Theorem 4.1 (Excess Risk Upper Bound)**.**

$$\mathcal{E}_{KD}(\pi_\theta) \leq \frac{2||M||_2}{N|\mathcal{V}|} \sum_N \sup_{\pi_\theta \in \Pi_{\text{OPSD}}} |\Delta| [\sqrt{\frac{Td}{2} + \frac{1}{2} \sum_{|\mathcal{C}^\dagger|} \sum_{\mathbf{h}_i \in \mathcal{X}_k} ||\mathbf{h}_i - \mu_k^{\mathcal{X}}||_2^2 - \frac{1}{2} \sum_{|\mathcal{C}^\dagger|} |\mathcal{C}_k^\dagger| ||\mu_k^{\mathcal{X}} - \mu||_2^2} + \sqrt{\sum_{|\mathcal{C}^\dagger|} |\mathcal{C}_k^\dagger| ||\mu_k^{\mathcal{X}} - \mu_k^\dagger||_2^2 + \sum_{|\mathcal{C}^\dagger|} \sum_{\mathbf{h}_i^\dagger \in \mathcal{X}_k} ||\mathbf{h}_i^\dagger - \mu_k^\dagger||_2^2}] \tag{3}$$

where N denote the number of samples, $|\Delta| = \sum_i ||\hat{\mathbf{h}}_i - \mathbf{h}_i{}^*||_2^2$, $M = W^T(I_{|\mathcal{V}|} - \frac{1}{|\mathcal{V}|} \mathbf{1}_{|\mathcal{V}|} \mathbf{1}_{|\mathcal{V}|}^T) W$. The detail proof could be found in Appendix D

From Theorem 4.1, we observe that the excess risk upper bound is affected by both intra-cluster compactness and inter-cluster separation of the learned representations. Specifically, reducing the intra-cluster variance term $\|\mathbf{h}_i - \mu_k^{\mathcal{X}}\|_2^2$ and increasing the inter-cluster dispersion term $\|\mu_k^{\mathcal{X}} - \mu\|_2^2$ can tighten the excess risk upper bound. Meanwhile, the term $\|\mu_k^{\mathcal{X}} - \mu_k^\dagger\|_2^2$ reflects the deviation between the learned student representation and the teacher and optimal teacher representation. Therefore, preventing capability degradation during OPSD requires not only maintaining the alignment with the teacher representation, but also preserving compact intra-cluster compactness and clear inter-cluster separation in the representation space.

5 METHODOLOGY

As discussed previously, we observe that increasingly blurred boundaries between hidden representation clusters coincide with the progressive degradation of model capabilities during OPSD training. Motivated by this observation, we propose Representation Geometry Regularization (RGR) to preserve clear boundaries between hidden representation clusters. To achieve this goal, we consider two complementary geometric properties: (1) intra-cluster compactness, where representations within each cluster remain sufficiently compact, and (2) inter-cluster separation, where different clusters maintain sufficient separation from each other. In the following, we discuss how to preserve the two properties in detail and derive the corresponding regularization objectives.

5.1 INTRA-CLUSTER COMPACTNESS

According to previous works , the intra-cluster compactness of hidden representations can be evaluated by the Within-Cluster Sum of Squares (WCSS)(Lloyd, 1982), defined as the sum of squared distances between each hidden representation and the center of its corresponding cluster. However, directly minimizing WCSS may lead to excessive compactness within each cluster. In the extreme case, hidden representations within the same cluster may become overly similar, reducing the discriminative information encoded in the representation space. Therefore, we introduce the teacher model as a reference and encourage the WCSS of the student representations within each cluster to remain close to that of the teacher representations.

Given the query q, the hidden representations of the response generated by the student model are denoted as $\{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$, while the corresponding hidden representations of the reference model are denoted as $\{\mathbf{h}_1^{\text{ref}}, \ldots, \mathbf{h}_T^{\text{ref}}\}$. Suppose that a cluster C_k^{ref} in the reference model's representation space corresponds to a cluster C_k^{align} in the student model's representation space, with cluster centers $\mathbf{u}_k^{\text{ref}}$ and $\mathbf{u}_k^{\text{align}}$, respectively. To align the intra-cluster compactness of the student model with that of the reference model, the learning target for intra-cluster compactness is defined as:

$$\mathcal{L}_{\text{intra}} = \sum_{|\mathcal{C}_{\text{ref}}|} \left| \sum_{\mathbf{h}_i \in C_k^{\text{align}},} \left\| \mathbf{h}_i - \boldsymbol{\mu}_k^{\text{align}} \right\|_2^2 - \sum_{\mathbf{h}_i^{\text{ref}} \in C_k^{\text{ref}}} \left\| \mathbf{h}_i^{\text{ref}} - \boldsymbol{\mu}_k^{\text{ref}} \right\|_2^2 \right|, \tag{4}$$

where $|\mathcal{C}_{\text{ref}}|$ denotes the number of clusters.

However, directly optimizing Eq. 17 is difficult. To address this, we stack the hidden representations of the student model as $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_T] \in \mathbb{R}^{T \times d}$, where each hidden representation is the row vector of $\mathbf{H}$. Similarly, we stack the hidden representations of the student model as $\mathbf{H}_{\text{ref}} = [\mathbf{h}_1^{\text{ref}}, \ldots, \mathbf{h}_T^{\text{ref}}]$, where each hidden representation is the row vector of $\mathbf{H}_{\text{ref}}$. Under the above definitions and assumptions, we have:

Theorem 5.1 (Intra-cluster Compactness Upper Bound)**.** *The upper bound of the intra-cluster compactness learning target is*

$$\mathcal{L}_{\text{intra}} \leq (T - |\mathcal{C}_{\text{ref}}|)^{\frac{1}{2}} ||\mathbf{H}\mathbf{H}^\top - \mathbf{H}_{\text{ref}}\mathbf{H}_{\text{ref}}^T||_F, \tag{5}$$

where $|\mathcal{C}_{\text{ref}}|$ denotes the number of clusters of the reference model.

The details of the proof can be found in Appendix E.

5.2 INTER-CLUSTER SEPARATION

We construct a Laplacian matrix(Von Luxburg, 2007) from the pairwise relationships among all hidden representations and use its low-end spectrum, corresponding to the smallest eigenvalues of the Laplacian matrix, to evaluate the separability between clusters. However, directly maximizing these low-end eigenvalues is undesirable, as excessively large inter-cluster separation can over-expand the representation space and distort the relative geometric relationships among clusters. Therefore, we use the teacher model as a reference model to encourage the student model to preserve its inter-cluster separation, thereby preventing undesirable geometric distortion during optimization.

Specifically, given the query q, the hidden representations of the response generated by the student model are denoted as $\{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$, while the corresponding hidden representations of the reference model are denoted as $\{\mathbf{h}_1^{\text{ref}}, \ldots, \mathbf{h}_T^{\text{ref}}\}$. We define the Gaussian kernel matrices of the hidden representations of the student model and the reference model as:

$$A_{ij} = \begin{cases} \exp\left(-\dfrac{\|\mathbf{h}_i - \mathbf{h}_j\|_2^2}{2\sigma^2}\right), & i \neq j, \\ 0, & i = j, \end{cases} \qquad A_{ij}^{ref} = \begin{cases} \exp\left(-\dfrac{\left\|\mathbf{h}_i^{\text{ref}} - \mathbf{h}_j^{\text{ref}}\right\|_2^2}{2\sigma^2}\right), & i \neq j, \\ 0, & i = j, \end{cases} \tag{6}$$

where σ denotes the standard deviation of the Gaussian kernel.

The corresponding degree matrices of the student and reference Gaussian kernel matrices are denoted by $\mathbf{D}$ and $\mathbf{D}^{\text{ref}}$, respectively, where $D_{ii} = \sum_{j=1}^{T} A_{ij}$ and $D_{ii}^{\text{ref}} = \sum_{j=1}^{T} A_{ij}^{\text{ref}}$. The unnormalized Laplacian matrices of the student and reference hidden representations are then defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$ and $\mathbf{L}^{\text{ref}} = \mathbf{D}^{\text{ref}} - \mathbf{A}^{\text{ref}}$, respectively. The corresponding eigenvalue vectors are

$$\begin{aligned} \boldsymbol{\lambda} &= [\lambda_1(\mathbf{L}), \ldots, \lambda_T(\mathbf{L})]^\top, \\ \boldsymbol{\lambda}^{\mathrm{ref}} &= \left[\lambda_1(\mathbf{L}^{\mathrm{ref}}), \ldots, \lambda_T(\mathbf{L}^{\mathrm{ref}})\right]^\top, \end{aligned} \tag{7}$$

where the eigenvalues are arranged in nondecreasing order. The learning target for inter-cluster separation could be represented by the difference between the eigenvalue vectors of the student and the reference, that is

$$\mathcal{L}_{\mathrm{inter}} = \left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^{\mathrm{ref}}\right\|_2. \tag{8}$$

Then, by the Hoffman–Wielandt inequality (Hoffman et al., 1953), we obtain

$$\mathcal{L}_{\mathrm{inter}} \leq (1+\sqrt{T})||\mathbf{L} - \mathbf{L}^{\mathbf{ref}}||_F \tag{9}$$

Under the above definitions and assumptions, we have

Theorem 5.2 (Inter-cluster Separation Upper Bound)**.** *The upper bound of the inter-cluster separation could be induced by applying Taylor expansion and the Newton–Schwarz inequality:*

$$\mathcal{L}_{\mathrm{inter}} \leq \left(\frac{1+\sqrt{T}}{\sigma}\right)^2 \left\|\mathbf{H}\mathbf{H}^\top - \mathbf{H}_{\mathrm{ref}}\mathbf{H}_{\mathrm{ref}}^\top\right\|_F, \tag{10}$$

The proof is provided in Appendix F.

5.3 Optimization

Building upon the previous analysis, preserving the intra-cluster compactness and inter-cluster separation of the hidden representation space requires minimizing $\mathcal{L}_{\mathrm{intra}}$ and $\mathcal{L}_{\mathrm{inter}}$. From the above theoretical analysis, we show that both geometric discrepancies can be controlled by the discrepancy between the Gram matrices of the student and reference hidden representations. Specifically, we have $\mathcal{L}_{\mathrm{intra}} \leq C_{\mathrm{intra}} \left\|\mathbf{H}\mathbf{H}^\top - \mathbf{H}_{\mathrm{ref}}\mathbf{H}_{\mathrm{ref}}^\top\right\|_F$ and $\mathcal{L}_{\mathrm{inter}} \leq C_{\mathrm{inter}} \left\|\mathbf{H}\mathbf{H}^\top - \mathbf{H}_{\mathrm{ref}}\mathbf{H}_{\mathrm{ref}}^\top\right\|_F$, where C_{intra} and C_{inter} are constants independent of the student representations. Therefore, minimizing the Gram matrix discrepancy provides a unified objective for preserving both intra-cluster compactness and inter-cluster separation. We thus define the geometry-preserving objective as

$$\mathcal{L}_{\mathrm{geo}} = \left\|\mathbf{H}\mathbf{H}^\top - \mathbf{H}_{\mathrm{ref}}\mathbf{H}_{\mathrm{ref}}^\top\right\|_F^2. \tag{11}$$

In addition to representation geometry, recent studies (Jin et al., 2026) on on-policy distillation have highlighted the importance of controlling the entropy of the model output distribution during optimization, which motivates us to explicitly incorporate an entropy-related objective into the optimization process. Accordingly, we jointly preserve the hidden representation geometry and regulate the output distribution entropy during OPSD optimization. Combining the original OPSD objective with the geometry-preserving and entropy-related objectives, we formulate the overall training objective as

$$\mathcal{L} = \mathcal{L}_{\mathrm{OPSD}} + \alpha\mathcal{L}_{\mathrm{geo}} + \beta\mathcal{L}_{\mathrm{ent}}, \tag{12}$$

where $\mathcal{L}_{\mathrm{OPSD}}$ denotes the original on-policy self-distillation objective, $\mathcal{L}_{\mathrm{geo}}$ denotes the proposed geometry-preserving regularization, and $\mathcal{L}_{\mathrm{ent}} = \left\|\mathbf{H}^\top\mathbf{H} - \mathbf{H}_{\mathrm{ref}}^\top\mathbf{H}_{\mathrm{ref}}\right\|_F^2$ denotes the entropy-related objective. The coefficients α and β control the contributions of the geometry-preserving and entropy-related terms, respectively.

6 EXEPERIMENT

In this section, we conduct experiments to evaluate the performance of RGR. We aim to answer the following research questions: (1) Is RGR effective for improving the performance of self distillation (RQ1)? (2) How do different components of RGR contribute to the performance gains (RQ2)? (3) Can RGR alleviate the degradation of previously acquired capabilities during OPSD optimization, and how does it affect the evolution of hidden representations (RQ3)? (4) Can RGR generalize to task-transfer settings and alleviate the degradation of previously acquired capabilities when learning new tasks (RQ4)?

6.1 Experiment setup

Datasets. Following the dataset setup of OPSD (Zhao et al., 2026), we sample up to 30K problem-solution pairs containing chain-of-thought reasoning from the mathematical reasoning subset of OpenThoughts(Guha et al.)for model training.. We conduct evaluation on three representative

benchmarks: AIME 2024 (Zhang & Math-AI, 2024), AIME 2025 (Zhang & Math-AI, 2025), and HMMT 2025 (Balunovic et al., 2026). AIME 2024 consists of problems from the 2024 American Invitational Mathematics Examination, covering diverse areas of high-school mathematics. AIME 2025 consists of problems from the 2025 American Invitational Mathematics Examination. HMMT 2025 consists of problems from the Harvard-MIT Mathematics Tournament, covering algebra, combinatorics, geometry, and number theory.

Baseline. To assess the effectiveness of RGR, we integrate it into four representative self-distillation methods, namely, OPSD (Zhao et al., 2026), SDPO (Hübotter et al., 2026), SRPO (Li et al., 2026), and RLSD (Yang et al., 2026), and compare each RGR-enhanced method with its original counterpart. These methods differ in how they optimize the self-distillation objective. Specifically, OPSD employs forward KL divergence (Kullback & Leibler, 1951), with the policy update constrained by token-level KL clipping; SDPO uses Jensen–Shannon divergence (Lin, 1991) to align the student with the teacher; SRPO routes successful responses to reward-based reinforcement and failed responses to self-distillation; and RLSD uses environment-derived rewards to determine the direction of policy updates, while reweighting token-level advantages according to the teacher–student log-probability gap to modulate the update magnitude.

ImplementationDetails. Our experiments are conducted on two foundation LLMs: Qwen3-1.7B and Qwen3-4B(Yang et al., 2025), with thinking mode enabled throughout training and validation. All experiments are performed on four NVIDIA A40 GPUs. All methods use LoRA (Hu et al., 2021) with a rank of 64 and a scaling factor of 128, with the learning rate set to 5×10^{-6}. The teacher model remains frozen throughout training. For our method, we fix the parameter α in Eq.(12) at 2.5×10^{-2} and the parameter β in Eq.(12) at 5×10^{-2} across all benchmarks.

6.2 Main results (RQ1)

Following OPSD (Zhao et al., 2026), we evaluate each model by generating 12 responses for each problem and report the average accuracy over the 12 responses (Avg@12). The results are reported in Table 1. As it is shown in Table 1, we can observe that: (1) RGR generally improves different methods across both Qwen3-1.7B and Qwen3-4B. OPSD, SDPO, and SRPO consistently benefit from RGR, demonstrating that representation geometry preservation provides complementary benefits to different optimization objectives. Although a few settings show slight decreases, RGR improves the majority of cases, supporting its general effectiveness; (2) The gains vary across different methods, with larger improvements generally observed for OPSD and SDPO but limited gains for RLSD. This can be attributed to their different optimization mechanisms: OPSD and SDPO directly rely on self-distillation signals for dense token-level supervision, while RLSD mainly uses self-distillation to determine the update magnitude and obtains the update direction from RLVR. Consequently, RGR has more room to benefit methods whose optimization induces stronger changes in the representation space; (3) The improvements also vary across AIME24, AIME25, and HMMT25. This variation may be attributed to the different difficulty levels and reasoning characteristics of these benchmarks, which result in different distributions of correct and incorrect responses during self-improvement. Overall, RGR remains beneficial across diverse mathematical reasoning benchmarks, while its improvement varies with the characteristics of the target dataset.

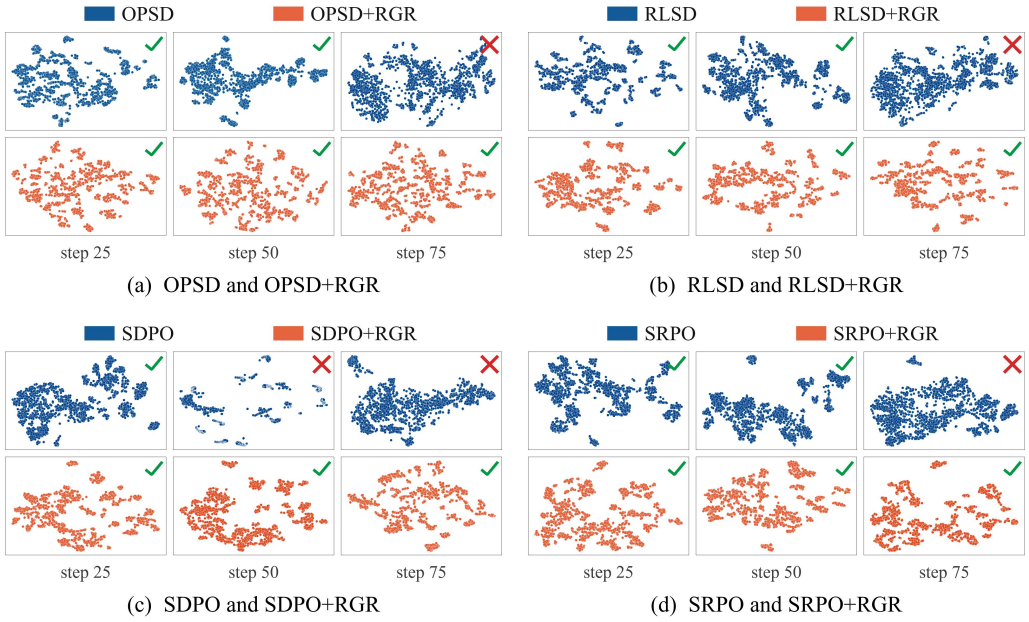
Table 1: Performance gains from RGR. Best results in bold; RGR improvements in green.

Model	Method	AIME24	AIME25	HMMT25	Avg.
Qwen3-1.7B	OPSD	56.67	43.06	28.89	42.87
	OPSD + RGR	**59.89**	**43.60**	**30.00**	**44.50** $_{+1.63}$
	SDPO	38.61	28.61	21.39	29.54
	SDPO + RGR	**41.94**	**31.39**	**21.94**	**31.76** $_{+2.22}$
	RLSD	50.27	39.70	25.83	38.60
	RLSD + RGR	**51.39**	39.17	25.56	**38.70** $_{+0.10}$
	SRPO	37.78	30.00	20.00	29.26
	SRPO + RGR	**39.44**	29.72	**21.67**	**30.28** $_{+1.02}$
Qwen3-4B	OPSD	74.17	68.06	45.00	62.41
	OPSD + RGR	**75.56**	**70.56**	**47.50**	**64.54** $_{+2.13}$
	SDPO	67.22	61.39	38.61	55.74
	SDPO + RGR	**72.50**	**63.61**	**41.11**	**59.07** $_{+3.33}$
	RLSD	75.00	65.00	44.44	61.48
	RLSD + RGR	74.72	**66.11**	**44.72**	**61.85** $_{+0.37}$
	SRPO	72.22	64.16	39.17	58.52
	SRPO + RGR	**73.33**	**66.67**	**41.67**	**60.56** $_{+2.04}$

6.3 Ablation Study (RQ2)

To investigate the contribution of each component to RGR, we conduct an ablation study with the following variants:

- **w/o Ent**: We remove the $\mathcal{L}_{\text{ent}}$ in Eq.(12), removing the constraint on output distribution.



(a) OPSD and OPSD+RGR (b) RLSD and RLSD+RGR

(c) SDPO and SDPO+RGR (d) SRPO and SRPO+RGR

Figure 4: Visualization of hidden state representations across different models and training steps, with response correctness indicated.

- **w/o Geo**: We remove the L_{ics} in Eq.(12), eliminating the constraint on intra-cluster compactness and inter-cluster separation.

The results are shown in Table 2 that RGR achieves the best performance across all three benchmarks. From Table 2, we can also observe that: (1) Removing the geometry constraint (**w/o Geo**) consistently degrades the performance across all benchmarks, compared with the full RGR. This indicates that explicitly preserving intra-cluster compactness and inter-cluster center separation is crucial for maintaining a structured representation space during OPSD optimization. Without this constraint, the hidden representations may become less organized, leading to limited capability improvement.

Table 2: Ablation study of RGR.

Method	AIME24	AIME25	HMMT25
OPSD	56.67	43.06	28.89
w/o Ent	57.22	41.39	28.33
w/o Geo	58.61	42.22	29.17
RGR	**59.89**	**43.60**	**30.00**

(2) Removing the entropy constraint (**w/o Ent**) also leads to performance drops on one benchmarks. This suggests that entropy regularization provides complementary optimization benefits by further stabilizing the output distribution during self-distillation. However, the performance gap between RGR and **w/o Geo** is larger than that between RGR and **w/o Ent**, indicating that representation geometry preservation serves as the primary contributor to the improvements.

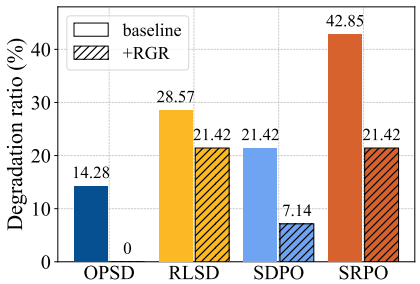


Figure 3: Comparison of the capability degradation ratio between the baseline and the baseline+RGR model on AIME24 (%).

6.4 Training Dynamics Visualization (RQ3)

Effect of RGR on mitigating progressive capability degradation. To evaluate whether RGR alleviates progressive capability degradation, we compare the ratio of queries that transition from correct to incorrect responses before and after incorporating RGR into different OPSD variants. As shown in Figure 3, adding RGR consistently reduces the degradation ratio across all evaluated methods. Specifically, RGR eliminates the degradation in OPSD and reduces the degradation ratios of SDPO, RLSD, and SRPO

by 14.28%, 7.15%, and 21.43%, respectively. These results demonstrate that RGR not only improves self-improvement performance but also helps preserve previously acquired capabilities by maintaining a more stable hidden representation geometry during optimization.

Representation geometry analysis. To further understand how RGR alleviates progressive capability degradation, we visualize the hidden representations of responses corresponding to the degraded queries. Specifically, we select queries that are initially responded correctly but become incorrectly responded during optimization, and compare the representation geometry before and after incorporating RGR. As shown in Figure 4, without RGR, the hidden representations gradually lose their geometric structure, exhibiting reduced inter-cluster center separation and less distinguishable cluster boundaries when the response changes from correct to incorrect. In contrast, RGR maintains more compact clusters with clearer boundaries in most cases. These results are consistent with our theoretical analysis and demonstrate that preserving representation geometry helps stabilize the evolution of hidden representations during OPSD training.

Overall, RGR achieves consistent reductions in capability degradation across OPSD, SDPO, RLSD, and SRPO, despite these methods adopting different self-distillation objectives. This suggests that RGR serves as a general geometry-preserving regularization mechanism that can be integrated into different OPSD methods.

Table 3: Forgetting evaluation on AIME24 (%). **Before** denotes the performance before training on the tool-use dataset, while **After** denotes the performance after tool-use training.

Method	Before	After	Forgetting ↓
OPSD	56.67	54.72	1.95
OPSD + RGR	59.89	58.06	1.83
SDPO	38.61	33.33	5.28
SDPO + RGR	41.94	40.00	1.94
RLSD	50.27	47.22	3.05
RLSD + RGR	51.39	48.00	3.39
SRPO	37.78	34.72	3.06
SRPO + RGR	39.44	38.33	1.11

6.5 Cross-Task Forgetting Study (RQ4)

To further investigate whether RGR can mitigate capability degradation beyond the original task, we conduct a cross-task forgetting experiment. We first post-train a model on OpenThoughts to acquire mathematical problem-solving capabilities, and evaluate the model on AIME24 to measure performance. Then we continue post-training it on a tool-use dataset(Tang et al., 2023), after convergence, we evaluate the model again on AIME24 to measure the performance after tool-use training. We apply this setting to OPSD, SDPO, RLSD, and SRPO, and compare their performance before and after training on the tool-use dataset, as shown in Table 3. From Table 3, we can observe that: (1) Sequential training on a new task leads to forgetting of previously acquired capabilities. After tool-use training, all four methods exhibit performance degradation on AIME24, consistent with previous studies of catastrophic forgetting in continual fine-tuning of LLMs (Luo et al., 2025; Li et al., 2024). (2) RGR generally alleviates cross-task forgetting. RGR reduces the forgetting gap for OPSD, SDPO, and SRPO, with particularly notable improvements for SDPO ($5.28 \rightarrow 1.94$) and SRPO ($3.06 \rightarrow 1.11$). This suggests that preserving representation geometry during subsequent optimization can help retain previously acquired capabilities. (3) The effectiveness of RGR depends on the underlying post-training method. For RLSD, the forgetting gap slightly increases from 3.05 to 3.39 after introducing RGR, suggesting that the effect of representation geometry preservation may vary across different optimization dynamics. Nevertheless, RGR reduces forgetting for three of the four methods, supporting its broader applicability to sequential post-training.

7 Conclusion

In this work, we identify and investigate the progressive capability degradation phenomenon in on-policy self-distillation, where LLM may lose previously acquired capabilities during self-improvement despite overall performance improvements. Through empirical analysis and theoretical investigation, we reveal that this degradation is associated with the deterioration of hidden representation geometry, characterized by reduced intra-cluster compactness and inter-cluster center separation. Based on these findings, we propose Representation Geometry Regularization (RGR), a general regularization framework that preserves intra-cluster compactness and inter-cluster center separation during OPSD optimization. Extensive experiments across multiple OPSD variants and mathematical reasoning benchmarks demonstrate that RGR effectively reduces correct-to-incorrect transitions while maintaining the self-improvement capability of OPSD. Future work will explore more adaptive strategies for preserving representation geometry and extending the proposed framework to broader model optimization paradigms beyond self-distillation.

REFERENCES

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. *Advances in Neural Information Processing Systems*, 38, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models, 2025. *URL https://arxiv. org/abs/2506.04178*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Alan J Hoffman, Helmut W Wielandt, et al. The variation of the spectrum of a normal matrix. *Duke Math. J*, 20(1):37–39, 1953.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 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.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing. *arXiv preprint arXiv:2604.02288*, 2026.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 4297–4308, 2024.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Tom Minka et al. Divergence measures and message passing. 2005.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

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

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Jiaxuan Wang, Xuan Ouyang, Zhiyu Chen, Yulan Hu, Zheng Pan, Xin Li, and Lan-Zhe Guo. Trace: Distilling where it matters via token-routed self on-policy alignment. *arXiv preprint arXiv:2605.10194*, 2026.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

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.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

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

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

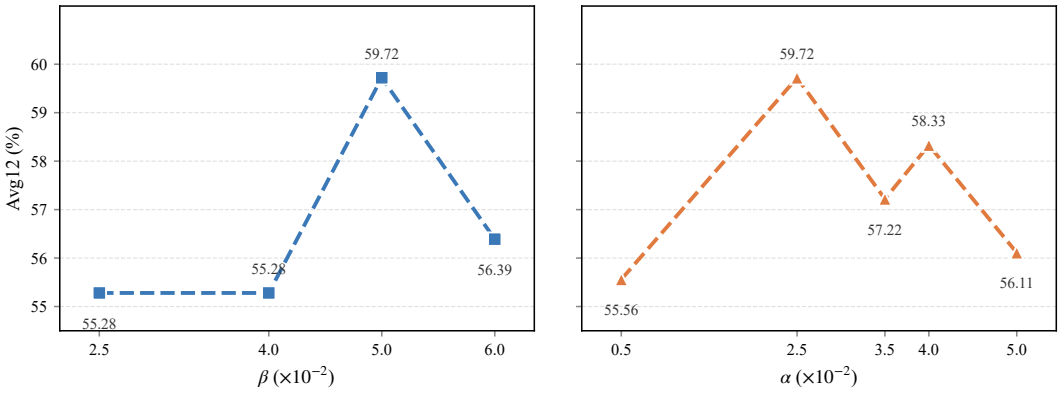
Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

A Sensitivity Analysis

Hyperparameter sensitivity. We investigate the sensitivity of our method to the hyperparameters α and β in Eq. (12). We vary one hyperparameter at a time while fixing the other at its selected value. Specifically, we vary β over $\{2.5, 4.0, 5.0, 6.0\} \times 10^{-2}$ while fixing $\alpha = 2.5 \times 10^{-2}$. Similarly, we vary α over $\{0.5, 2.5, 3.5, 4.0, 5.0\} \times 10^{-2}$ while fixing $\beta = 5 \times 10^{-2}$. As shown in Figure 5, As shown in the figure, the performance of our model follows a smooth increasing-then-decreasing trend with respect to the hyperparameter values. This observation suggests that the proposed method does not require highly precise hyperparameter tuning and remains stable change across a relatively broad range of parameter settings.

B Implementation Details

Table 4 lists the training hyperparameters that are shared across all experiments. All runs use LoRA for parameter-efficient fine-tuning with rank r=64 and α=128, and are optimized with a learning rate of 5×10^{-6} under a maximum gradient norm of 0.1. The maximum sequence length is $20{,}000$ tokens, while completions are capped at $1{,}024$ tokens. Rollout sampling uses a temperature of 1.1, top-p of 0.95, and top-k of 20. Training is performed in BF16 precision.



(a) Varying β ($\alpha = 2.5 \times 10^{-2}$) **(b) Varying α ($\beta = 5.0 \times 10^{-2}$)**

Figure 5: Sensitivity analysis of the hyperparameters α and β on AIME24. Left: performance under different values of β, with α fixed at 2.5×10^{-2}. Right: performance under different values of α, with β fixed at 5×10^{-2}. The stars indicate the selected setting: $\alpha = 2.5 \times 10^{-2}$ and $\beta = 5 \times 10^{-2}$.

Table 4: Hyperparameters used across all experiments.

Hyperparameter	**Value**
Learning rate	5×10^{-6}
Max gradient norm	0.1
LoRA rank r	64
LoRA α	128
Max sequence length	20,000
Max completion length	1,024
Temperature	1.1
Top-p	0.95
Top-k	20
BF16	Enabled

C Notation and Additional Preliminaries

D Proof of Theorem 4.1

By definition 4.1, we obtain

$$\mathcal{E}_{\text{KD}}(\pi_\theta) = \mathbb{E}_{q\sim\mathcal{D}, y\sim\pi_\theta(\cdot|q)}\left[\sum_{t=1}^{T} D\left(\pi_{\theta^*}(\cdot \mid q, z, y_{<t}) \,\|\, \pi_\theta(\cdot \mid q, y_{<t})\right) - D\left(\pi_{\theta^*}(\cdot \mid q, z, y_{<t}) \,\|\, \pi_{\theta^*}(\cdot \mid q, y_{<t})\right)\right] \tag{13}$$

To simplify the notation, we define

$$\begin{aligned} q_t^* &:= \pi_{\theta^*}(\cdot \mid q, z, y_{<t}) \\ p_t &:= \pi_\theta(\cdot \mid q, y_{<t}) \\ q_t &:= \pi_\theta(\cdot \mid q, z, y_{<t}) \\ p_t^* &:= \pi_{\theta^*}(\cdot \mid q, y_{<t}) \end{aligned}$$

Let z_t denote the corresponding logits of p_t, z_t^* denote the corresponding logits of q_t^* and $\hat{z}_t$ denotes the corresponding logit of q_t. For each $f \in \mathcal{F}_{OPSD}$, let p_f denote the policy induced by f.

For one response and $q \sim \mathcal{D}, y \sim \pi_\theta$, let $\mathcal{E}_{q,\theta} = \sum_t (KL(q_t^* || p_t)) - E(KL(q_t^* || p_t^*)$, we have

$$\mathcal{E}_{\text{KD}}(\pi_\theta) = \mathbb{E}_{q\sim\mathcal{D}, y\sim\pi_\theta} \mathcal{E}_{q,\theta}$$

$$\mathcal{E}_{q,\theta} = \sum_t (KL(q_t^*||p_t)) - KL(q_t^*||p_t^*))$$

$$\mathcal{E}_{q,\theta} = \sum_t [KL(q_t^*||p_t) - KL(q_t||p_t) + KL(q_t||p_t) - KL(q_t^*||q_t) + KL(q_t^*||q_t) - KL(q_t^*||p_t^*)]$$

$$\sum_t [KL(q_t^*||p_t) - KL(q_t||p_t)] \leq sup_{\pi_\theta \in \Pi_{\text{OPSD}}} |\sum_t [KL(q_t^*||p_f) - KL(q_t||p_f)]|$$

$$\sum_t (KL(q_t||p_t) - KL(q_t^*||q_t)) \leq 0$$

$$\sum_t (KL(q_t^*||q_t) - KL(q_t^*||p_t^*)) \leq \sup_{\pi_\theta \in \Pi_{\text{OPSD}}} |\sum_t [KL(q_t^*||p_f) - KL(q_t||p_f)]|$$

so we have

$$\mathcal{E}_{q,\theta} \leq 2 * \sup_{\pi_\theta \in \Pi_{\text{OPSD}}} |\sum_t [KL(q_t^*||p_f) - KL(q_t||p_f)]|$$

$$\mathcal{E}_{q,\theta} \leq 2 * \sup_{\pi_\theta \in \Pi_{\text{OPSD}}} |\sum_t [\sum_i q_{t,i}^* log q_{t,i}^* - q_{t,i}^* log p_{t,i} - \hat{q}_{t,i} log q_{t,i}^* + \hat{q}_{t,i} log p_{t,i}]| \tag{14}$$

Consider an arbitrary LLM and fix an input context. Let $\widetilde{\mathbf{z}} = (\widetilde{z}_1, \ldots, \widetilde{z}_V)^\top \in \mathbb{R}^V$ denote the logit vector produced by the model, where V is the vocabulary size. The corresponding next-token probability vector $\mathbf{p} = (p_1, \ldots, p_V)^\top$ is given by

$$p_i := \frac{\exp(\widetilde{z}_i)}{\sum_{j=1}^V \exp(\widetilde{z}_j)}, \qquad i \in [V].$$

Let $\mathbf{1} \in \mathbb{R}^V$ denote the all-ones vector. Since softmax is invariant to a constant shift, for any $c \in \mathbb{R}$,

$$\mathrm{softmax}(\widetilde{\mathbf{z}} + c\mathbf{1}) = \mathrm{softmax}(\widetilde{\mathbf{z}}).$$

Define the centering matrix

$$A := I_V - \frac{1}{V}\mathbf{1}\mathbf{1}^\top$$

and the centered logit vector

$$\mathbf{z} := A\widetilde{\mathbf{z}} = \widetilde{\mathbf{z}} - \left(\frac{1}{V}\mathbf{1}^\top \widetilde{\mathbf{z}}\right)\mathbf{1}.$$

Then

$$\mathbf{1}^\top \mathbf{z} = 0, \qquad A\mathbf{z} = \mathbf{z},$$

and

$$\mathrm{softmax}(\mathbf{z}) = \mathrm{softmax}(\widetilde{\mathbf{z}}).$$

Therefore, without loss of generality, we may work with centered logits and assume throughout that

$$\mathbf{z} = A\mathbf{z}, \qquad \text{or equivalently} \qquad \sum_{j=1}^V z_j = 0.$$

For centered logits satisfying $\|\mathbf{z}\|_2 \to 0$, Taylor expansion gives

$$e^{z_i} = 1 + z_i + \frac{1}{2}z_i^2 + O\big(\|\mathbf{z}\|_2^3\big)\,.$$

Consequently,

$$\begin{aligned}\sum_{j=1}^V e^{z_j} &= V + \sum_{j=1}^V z_j + \frac{1}{2}\sum_{j=1}^V z_j^2 + O\big(\|\mathbf{z}\|_2^3\big) \\ &= V + \frac{1}{2}\|\mathbf{z}\|_2^2 + O\big(\|\mathbf{z}\|_2^3\big)\,,\end{aligned}$$

where the second equality follows from $\sum_{j=1}^{V} z_j = 0$. Therefore,

$$\log\left(\sum_{j=1}^{V} e^{z_j}\right) = \log V + \log\left(1 + \frac{1}{2V}\|\mathbf{z}\|_2^2 + O\big(\|\mathbf{z}\|_2^3\big)\right)$$
$$= \log V + \frac{1}{2V}\|\mathbf{z}\|_2^2 + O\big(\|\mathbf{z}\|_2^3\big)\,.$$

In particular,

$$p_i = \frac{1}{V}\left(1 + z_i + O(\|\mathbf{z}\|_2^2)\right).$$

For Qwen3, let $\mathbf{h} \in \mathbb{R}^d$ denote the final hidden representation at the current token position and let $W \in \mathbb{R}^{V\times d}$ denote the weight matrix of its bias-free language-modeling head. The original logits are $\widetilde{\mathbf{z}} = W\mathbf{h}$. Defining

$$A := I_V - \frac{1}{V}\mathbf{1}\mathbf{1}^\top,$$

the centered logits can be written as

$$\mathbf{z} := A\widetilde{\mathbf{z}} = AW\mathbf{h}.$$

Since $\mathbf{1}^\top\mathbf{z} = 0$ and softmax is invariant to constant shifts, $\mathbf{z}$ induces the same next-token distribution as $\widetilde{\mathbf{z}}$.

Since OPSD employs LoRA without applying it to the output layer, the output-layer weights remain unchanged. Therefore, we obtain:

So we can simlify 14:

$$\mathcal{E}_{q,\theta} \le \frac{2}{|\mathcal{V}|} * \sup_{\pi_\theta\in\Pi_{\text{OPSD}}} |\sum_t [(h_t - \frac{h_t^* + \hat{h}_t}{2})^T W^T (I_{|\mathcal{V}|} - \frac{1}{|\mathcal{V}|} 1_{|\mathcal{V}|} 1_{|\mathcal{V}|}^T) W (h_t^* - \hat{h}_t)]| \tag{15}$$

Throughout our analysis, we assume that the hidden representations are normalized using RMSNorm without learnable scaling parameters.

By the Cauchy–Schwarz inequality, we have

$$\mathcal{E}_{\text{KD}}(\pi_\theta) \le \frac{2}{N|\mathcal{V}|}\sum_N \sup_{\pi_\theta\in\Pi_{\text{OPSD}}} |\Delta|||M||_2[\sqrt{\frac{Td}{2} + \frac{1}{2}\sum_{|\mathcal{C}^\dagger|}\sum_{\mathbf{h}_i\in\mathcal{X}_k} ||\mathbf{h}_i - \mu_k^{\mathcal{X}}||_2^2 - \frac{1}{2}\sum_{|\mathcal{C}^\dagger|}|\mathcal{C}_k^\dagger|||\mu_k^{\mathcal{X}} - \mu||_2^2} +$$
$$\sqrt{\sum_{|\mathcal{C}^\dagger|}|\mathcal{C}_k^\dagger|||\mu_k^{\mathcal{X}} - \mu_k^\dagger||_2^2 + \sum_{|\mathcal{C}^\dagger|}\sum_{\mathbf{h}_i^\dagger\in\mathcal{X}_k} ||\mathbf{h}_i^\dagger - \mu_k^\dagger||_2^2}] \tag{16}$$

E Proof of Theorem 5.1

Proof. For each cluster, we stack all $h_{i,ref} \in \mathcal{C}_k^{ref}$ to form $H_k^{ref} = [(h_1^{ref})^\top, \ldots, (h_{|C_k^{ref}|}^{ref})^\top]^\top$, then we define $G_k^{ref} = H_k^{ref} H_k^{ref\top}$. Accordingly, we stack all $h_{i,align} \in \mathcal{C}_k^{align}$ to form $H_k^{align} = [(h_1^{align})^\top, \ldots, (h_{|C_k^{align}|}^{align})^\top]^\top$, then we define $G_k^{align} = H_k^{align} H_k^{align\top}$. and $\mathbf{1}_{|C_k^{ref}|} \in \mathbb{R}^{|C_k^{ref}|}$ denotes the all-ones column vector.$|\mathcal{C}_{ref}|$ denotes the number of clusters in the reference model.

First, we can simplify the optimization objective of intra-cluster compactness for each cluster:

$$\left|\sum_{\mathbf{h}_i\in C_k^{align},} \left\|\mathbf{h}_i - \boldsymbol{\mu}_k^{align}\right\|_2^2 - \sum_{\mathbf{h}_i^{\text{ref}}\in C_k^{\text{ref}}} \left\|\mathbf{h}_i^{\text{ref}} - \boldsymbol{\mu}_k^{\text{ref}}\right\|_2^2\right|, \le ||I_{|C_{ref}|} - \mathbf{1}_{|C_k^{ref}|}^\top \mathbf{1}_{|C_k^{ref}|}||_F ||G_k^{ref} - G_k||_F \tag{17}$$

Second, G_k^{align} is a principal submatrix of HH^T and G_k^{ref} is a principal submatrix of HH^T. Because the squared Frobenius norm is additive over pairwise disjoint principal submatrices, Therefore, for all clusters, we can obtain this through the Cauchy-Schwarz inequality:

$$\sum_k ||I_{|C_k^{ref}|} - \mathbf{1}^\top_{|C_k^{ref}|}\mathbf{1}_{|C_k^{ref}|}||_F||G_k^{ref} - G_k^{align}||_F \leq (T - |C_{ref}|)^{\frac{1}{2}}||\sum_k (G_k^{ref} - G_k^{align})||_F$$

Third, through the inclusion of the pairwise disjoint principal submatrices, we can get the upper bound of $\mathcal{L}_{intra}$:

$$\leq (T - |\mathcal{C}_{ref}|)^{\frac{1}{2}}||HH^T - H_{ref}H_{ref}^\top||_F \tag{18}$$

□

F PROOF OF PROPOSITION 5.2

Proof. Through the definition of the Laplacian matrix and the Cauchy-Schwarz inequality, we can obtain: Let $\mathbf{1}_T \in \mathbb{R}^{\mathbb{T}}$ denote the all-ones column vector.

$$\begin{aligned} ||L - L^{ref}||_F &= ||diag(A\mathbf{1}_T) - diag(A^{ref}\mathbf{1}_T) + A - A^{ref}||_F \\ &\leq ||diag(A_{i+1} - A_i)\not{\mathbf{1}}||_F + ||A_i - A_{i+1}||_F, \end{aligned}$$

$$||diag(A - A^{ref})\mathbf{1}_T||_F = ||A - A^{ref}||_2 \leq \sqrt{T}||A - A^{ref}||_F$$

$$||diag(A - A^{ref})\mathbf{1}_T||_F + ||A - A^{ref}||_F \leq (\sqrt{T} + 1)||A - A^{ref}||_F$$

$$(\sqrt{T} + 1)||A - A^{ref}||_F \leq \frac{1 + \sqrt{T}}{2\sigma^2}||D - D^{ref}||_F$$

$$||D - D^{ref}||_F = ||diag\{HH^T\}\mathbf{1}_T^T + \mathbf{1}_T diag\{HH^T\} - 2HH^T - diag\{H_{ref}H_{ref}^T\}1^T - 1diag\{H_{ref}H_{ref}^T\} + 2H_{ref}H_{ref}^T||_F$$

Through Cauchy-Schwarz inequality, we have

$$\sqrt{\sum_i |\lambda(L) - \lambda(L^{ref})|^2} \leq (\frac{1 + \sqrt{T}}{\sigma})^2||HH^T - H_{ref}H_{ref}^T||_F$$

□

Table 5: Summary of the main notation used in the paper.

Symbol	Description
q	a query
π_θ	LLM policy parameterized by θ and trained within the OPSD framework.
θ	Parameters of the LLM policy.
y_t	Token generated at decoding step t by π_θ.
$\mathbf{y} = (y_1, \ldots, y_T)$	Response consisting of T tokens generated by π_θ.
$\mathbf{y}_{<t} = (y_1, \ldots, y_{t-1})$	Sequence of tokens generated before step t.
$\pi_\theta(\cdot \mid q, \mathbf{y}_{<t})$	Conditional distribution over the next token given the query and previously generated tokens.
T	Length of the generated response.
$\mathcal{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$	the hidden representations corresponding to the generated response.
$\hat{\mathcal{H}} = \{\hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_T\}$	the hidden representations generated by the teacher model.
$\mathcal{H}^* = \{\mathbf{h}_1^*, \ldots, \mathbf{h}_T^*\}$	the hidden representations generated by the optimal teacher model.
$\mathcal{H}^\dagger = \{\frac{\mathbf{h}_1^* + \hat{\mathbf{h}}_1}{2} + \ldots, \frac{\mathbf{h}_T^* + \hat{\mathbf{h}}_T}{2}\}$	the mixed hidden representations of $\mathcal{H}$ and $\hat{\mathcal{H}}$.
$\mathcal{H}^{ref} = \{\mathbf{h}_1^{ref}, \ldots, \mathbf{h}_T^{ref}\}$	the hidden representations of the reference model.
$\mathcal{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$	the hidden representations corresponding to the generated response.
$\mathcal{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$	the hidden representations corresponding to the generated response.
$\mathcal{C} = \{C_1, \ldots, C_K\}$	$\mathcal{H}$ into K representation clusters.
$\mathcal{C}^\dagger = \{C_1^\dagger, \ldots, C_K^\dagger\}$,	$\mathcal{H}^\dagger$ into K representation clusters.
$\mathcal{C}^{ref} = \{C_1^{ref}, \ldots, C_K^{ref}\}$,	$\mathcal{H}^{ref}$ into K representation clusters.
$\mathcal{X}_k \coloneqq \{\mathbf{h}_i \mid i \in \mathcal{I}_k^\dagger\}$	the associated index set of for each cluster $C_k^\dagger$
$\lvert\mathcal{C}\rvert$	the number of clusters of $\mathcal{C}$.
$\lvert\mathcal{C}^\dagger\rvert$	the number of clusters of $\mathcal{C}^\dagger$.
$\lvert\mathcal{C}_{\text{ref}}\rvert$	the number of clusters of $\mathcal{C}_{\text{ref}}$.
$\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_T] \in \mathbb{R}^{T \times d}$	the hidden representation matrix of the student model.
$\mathbf{H}^{ref} = [\mathbf{h}_1^{ref}, \ldots, \mathbf{h}_T^{ref}] \in \mathbb{R}^{T \times d}$	the hidden representation matrix of the reference model.
μ_k	$\frac{1}{\lvert C_k\rvert} \sum_{\mathbf{h}_i \in C_k} \mathbf{h}_i$.
$\mu_k^\dagger$	$\frac{1}{\lvert C_k^\dagger\rvert} \sum_{\mathbf{h}_i^\dagger \in C_k^\dagger} \mathbf{h}_i^\dagger$.
μ_k^{ref}	$\frac{1}{\lvert C_k^{ref}\rvert} \sum_{\mathbf{h}_i^{ref} \in C_k^{ref}} \mathbf{h}_i^{ref}$.
$\mathbf{L}$	The Laplacian matrix computed from the hidden representation matrix of the student model with a Gaussian kernel.
$\mathbf{L}^{\text{ref}}$	The Laplacian matrix computed from the hidden representation matrix of the reference model with a Gaussian kernel.
$\boldsymbol{\lambda} = [\lambda_1(\mathbf{L}), \ldots, \lambda_n(\mathbf{L})]^\top$	the eigenvalues of the Laplacian matrix $\mathbf{L}$.
$\boldsymbol{\lambda}^{\text{ref}} = \left[\lambda_1(\mathbf{L}^{\text{ref}}), \ldots, \lambda_n(\mathbf{L}^{\text{ref}})\right]^\top$	the eigenvalues of the Laplacian matrix $\mathbf{L}^{\text{ref}}$.
μ_k^{ref}	$\frac{1}{\lvert C_k^{ref}\rvert} \sum_{\mathbf{h}_i^{ref} \in C_k^{ref}} \mathbf{h}_i^{ref}$.
$\mathcal{V}$	the vocabulary of the LLM.
$\lvert\mathcal{V}\rvert$	the size of the vocabulary.
$W \in \mathbb{R}^{\lvert\mathcal{V}\rvert \times d}$	the weight matrix of the final layer for the trained model.
$I_{\lvert\mathcal{V}\rvert} \in \mathbb{R}^{\lvert\mathcal{V}\rvert \times \lvert\mathcal{V}\rvert}$	the identity matrix .
$\mathbf{1}_{\lvert\mathcal{V}\rvert} \in \mathbb{R}^{\lvert\mathcal{V}\rvert}$	the all-ones column vector.
$\lvert\Delta\rvert$	$\sum_i \lvert\lvert\hat{\mathbf{h}}_i - \mathbf{h}_i{}^*\rvert\rvert_2$.
$D(\cdot\lvert\cdot)$	a divergence measure between the teacher and student token distributions.
$\mathcal{E}_{\text{KD}}(\pi_\theta)$	a divergence measure between the teacher and student token distributions.