OPDX: OPTIMAL TEACHER SELECTION FOR MULTI-DOMAIN ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-Policy Distillation (OPD) improves LLM post-training by providing token-level supervision on trajectories sampled from the student policy. However, extending OPD to multiple domains introduces two core challenges: (1) the intra-domain teacher-student distribution mismatch causes gradient spikes and unstable convergence; and (2) distributional inconsistency between domain-specific teachers induces optimization oscillations. To address these challenges, we reframe the problem from a teacher-selection perspective and propose OPDX, a two-stage framework for multi-domain OPD. First, a domain teacher selection method uses the Teacher Suitability Index (TSI) to select suitable single-domain teachers, alleviating convergence issues caused by distribution bias. Second, we propose a joint optimization distillation mechanism that evaluates optimal teacher group using aggregate TSI and the Cross-domain Penalty Factor to determine the optimal teacher group for multi-domain OPD. This mechanism alleviates optimization conflicts between domain-specific teachers. Experiments on three domains with Qwen3.5 models of varying sizes show that OPDX consistently outperforms strong baselines in single-domain precision, multi-domain integration, and scaling efficiency. OPDX establishes a scalable, generalizable approach to stable multi-domain OPD. The code will be released upon acceptance.

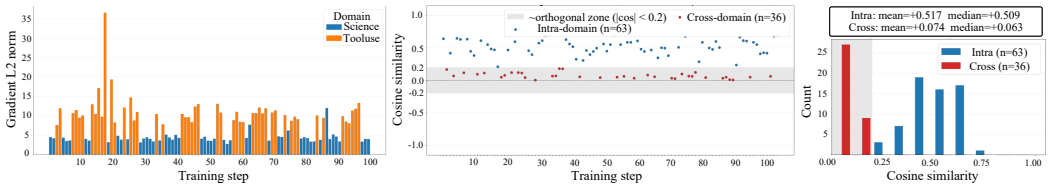
1 INTRODUCTION

On-Policy Distillation (OPD) has emerged as a crucial paradigm in modern large language model (LLM) post-training, significantly enhancing student capabilities by providing dense, token-level distributional supervision on trajectories sampled on-policy by the student Agarwal et al. (2023); Gu et al. (2023); Zeng et al. (2026). Recent flagship models (e.g., Qwen Team (2026), GLM Zeng et al. (2026), Kimi Moonshot AI (2026)) increasingly adopt a two-stage multi-expert distillation strategy: first training specialized domain experts independently and then consolidating their expertise into a single deployable student via multi-teacher OPD. This paradigm avoids the massive infrastructure overhead of running multiple specialized models in production while preserving fine-grained domain reasoning capabilities.

Despite its empirical success, multi-teacher OPD remains prone to substantial training instability, which limits further performance improvements across heterogeneous domains and tasks Chen et al. (2026); Ma et al. (2026). Although prior studies have sought to improve the stability of OPD through objective reformulation, trust-region constraints, loss clipping, etc. Jang et al. (2026); Xing et al. (2026); Luo et al. (2026); Ko et al. (2026); Jin et al. (2026a); Li et al. (2026a), these methods are primarily designed for single-teacher settings and therefore do not adequately account for the heterogeneous and potentially conflicting supervision induced by multiple domain-specific teachers.

Our preliminary gradient analysis in Fig. 1 provides direct evidence of this issue. Fig. 1(a) illustrates the change in gradient norms for batches sampled from different domains, revealing substantial cross-domain variation in gradient magnitudes and indicating that certain domains may dominate the optimization process. Fig. 1(b) further presents cosine similarities between intra-domain and cross-domain gradient pairs throughout training, showing pronounced directional fluctuations. Finally, Fig. 1(c) reports the distribution of pairwise cosine similarities between gradients of different domains, where a considerable proportion of low or negative similarities indicates frequent gradient

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.



(a) Gradient norm change across batches (b) Gradient direction change across batches (c) Cosine similarity distribution

Figure 1: Compressed-gradient diagnostics of multi-domain OPD on the Science and Tooluse domains over 100 training steps. Each mini-batch contains samples from a single domain. For computational efficiency, we reduce each gradient tensor by averaging over its last dimension and concatenate the resulting tensors into a one-dimensional gradient vector. (a) Change of the gradient norms for the two domains. (b) Cosine similarities between intra-domain and cross-domain gradient pairs throughout training. (c) Empirical distributions of the intra-domain and cross-domain cosine similarities. These compressed representations need not preserve full-gradient angles or norms.

misalignment and optimization conflicts. Collectively, these observations suggest that the instability of multi-teacher OPD primarily arises from two critical issues:

- **Intra-domain teacher-student mismatch:** Directly distilling from domain experts trained under distinct optimization trajectories creates sudden distributional shifts. This leads to severe gradient spikes during optimization, causing loss volatility and erratic convergence.
- **Cross-domain teacher inconsistency:** Heterogeneous domain experts often exhibit mismatched output scales and nearly orthogonal optimization directions (with cross-domain cosine similarities averaging as low as 0.074). Consequently, dominant domains overpower shared parameter updates, inducing optimization oscillations and task interference.

Addressing these issues before distillation is highly non-trivial, as it requires joint modeling of single-domain teacher-student suitability and cross-domain mutual compatibility without introducing prohibitive search costs. To bridge this gap, we introduce a two-stage distillation framework that systematically resolves intra- and cross-domain conflicts. To mitigate intra-domain teacher-student mismatch and suppress gradient spikes, we explore diverse teacher optimization trajectories, including SFT, RL, and privileged-information augmentation. We then introduce the TSI to identify suitable candidate teachers for each domain. By selecting teachers whose preferred probability mass lies within the student's effective support, TSI ensures accessible supervision under the current policy. This reduces teacher-student distribution shifts and stabilizes convergence. To address cross-domain inconsistency and directional interference, we propose a joint optimization distillation mechanism. By jointly considering TSI scores and Cross-Domain Penalty Factor (CPF), the combination score jointly favors accessible supervision and lower cross-domain entropy dispersion; its effectiveness is evaluated empirically.

We conduct extensive experiments in three distinct domains (scientific reasoning, tooluse, and telecommunications) using the Qwen3.5-4B and Qwen3.5-9B models, which shows that OPDX consistently outperforms SFT, GRPO, and existing distillation baselines, achieving superior task performance, smoother convergence, and robust multi-domain synergy.

The main contributions of this work are threefold.

- **Systematic Analysis of Multi-teacher OPD:** We perform an empirical gradient and distribution analysis on multi-domain OPD, revealing that intra-domain mismatch causes gradient spikes while cross-domain inconsistency leads to severe gradient scale imbalance and directional interference.
- **Novel Multi-teacher OPD Method:** We introduce the Teacher Suitability Index along with a cross-domain mutually compatible selection mechanism, establishing a principled pre-distillation screening framework for multi-teacher alignment.
- **Extensive Experiments:** We evaluate OPDX in three domains using Qwen3.5 models of different sizes. OPDX consistently outperforms SFT, GRPO, and OPSD in single-domain and multi-domain settings, achieving superior task performance, smoother convergence, and robust multi-domain synergy.

2 Related Work

2.1 Multi-Domain Capability Integration

LLM post-training commonly combines supervised fine-tuning and reinforcement learning to improve instruction following and domain-specific capabilities (Ouyang et al., 2022; Shao et al., 2024). However, joint or sequential training across multiple domains can cause gradient interference or catastrophic forgetting (Yang et al., 2026; Ma et al., 2026). Training domain experts independently alleviates interference during expert development, but their capabilities must still be consolidated into a single deployable model (Ma et al., 2026). We study stable capability integration when independently trained experts differ in their compatibility with the student and with one another.

2.2 On-Policy Distillation

On-policy distillation (OPD) provides token-level supervision on student-generated trajectories, reducing the mismatch between training and inference distributions (Agarwal et al., 2023; Gu et al., 2023), and has also been extended to same-scale self-distillation with privileged information (Zhao et al., 2026). Existing methods improve OPD stability through target reformulation (Jang et al., 2026), trust-region constraints (Xing et al., 2026), rollout-policy regularization (Plyusov et al., 2026; Luo et al., 2026), entropy-aware weighting (Jin et al., 2026a), reward clipping (Ko et al., 2026), teacher-aligned data selection (Li et al., 2026a), and step-wise divergence-aware weighting (Zhong et al., 2026). However, these methods mainly assume a fixed teacher and focus on single-teacher settings. Although conventional multi-teacher distillation aggregates or dynamically selects teachers (You et al., 2017; Yuan et al., 2020), recent multi-teacher OPD methods typically rely on predefined domain teachers (Ma et al., 2026; Chen et al., 2026). In contrast, OPDX selects teachers according to both teacher-student suitability and inter-teacher distribution consistency.

3 Preliminaries

Single-Domain On-Policy Distillation. Let $\mathcal{D}$ denote the prompt distribution of a target domain. Given $x \sim \mathcal{D}$, the student policy p_θ generates a response $y = (y_1, \ldots, y_{|y|}) \sim p_\theta(\cdot \mid x)$. At token position t, the student-visited context is $c_t = (x, y_{<t})$. A single-domain teacher policy q_d provides token-level supervision on the same context (Agarwal et al., 2023; Gu et al., 2023).

For brevity, we define $p_t(v) = p_\theta(v \mid c_t)$ and $q_t(v) = q_d(v \mid c_t)$. The token-level reverse KL divergence is

$$\ell_t^{\mathrm{OPD}} = \sum_{v \in \mathcal{V}} p_t(v) \log \frac{p_t(v)}{q_t(v)}, \tag{1}$$

where $\mathcal{V}$ is the full vocabulary. The single-domain OPD objective is

$$\mathcal{L}_{\mathrm{OPD}}(\theta; \mathcal{D}, q_d) = \mathbb{E}_{\substack{x \sim \mathcal{D} \\ y \sim p_\theta(\cdot|x)}} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} \ell_t^{\mathrm{OPD}} \right], \tag{2}$$

where $|y|$ denotes the number of tokens in the generated response. Responses are sampled from the current student policy and treated as fixed sequences when computing the distillation gradient; we do not backpropagate through the sampling distribution. In OPDX, Eq. 2 is used to train the student after teacher construction and selection, as described in Section 4.2.

Privileged-Information Teacher. A same-scale teacher q_d can be enhanced with privileged information e_d, such as a reference answer or additional context (Zhao et al., 2026; Penaloza et al., 2026). Its token distribution is

$$q_t^e(v) = q_d(v \mid c_t, e_d). \tag{3}$$

The student still observes only c_t. In this setting, $q_t(v)$ in Eq. 1 is replaced by $q_t^e(v)$.

4 Method

Fig. 2 presents OPDX, a two-stage framework for stable multi-domain OPD. We initialize a single student shared across all domains from the original general-model checkpoint used to initialize the

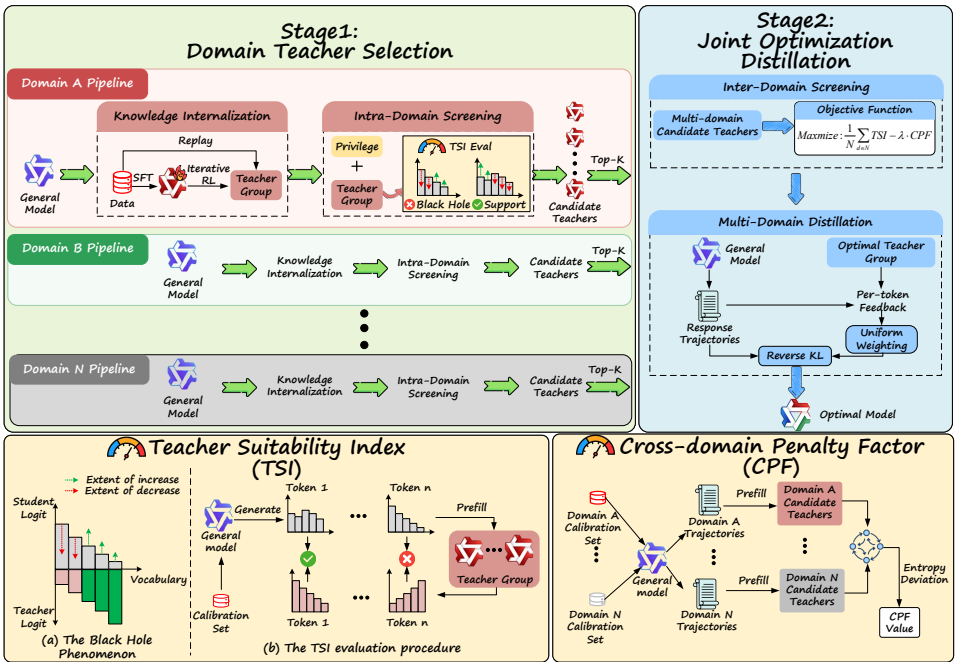


Figure 2: Overview of OPDX. Starting from a general model, we perform domain-specific Knowledge Internalization to construct an initial teacher group for each domain. Each candidate teacher is then augmented with privileged information and evaluated using TSI. Specifically, (a) illustrates the principle of TSI for identifying teachers that mitigate the black-hole phenomenon, while (b) presents the TSI evaluation procedure. The first stage retains the top-k candidates per domain. In the second stage, CPF and TSI are jointly used to select a cross-domain teacher combination containing one teacher from each domain. Finally, the selected teachers supervise the general model through multi-domain OPD under uniform domain sampling, enabling multi-domain knowledge integration.

domain teachers. In the first stage, we construct candidate teachers through domain-specific Knowledge Internalization, and screen them within each domain using the TSI. In the second stage, we jointly select one teacher per domain using TSI and the CPF. The selected teachers then supervise the shared student through OPD under uniform domain sampling.

4.1 Domain Teacher Selection

A teacher with high domain accuracy is not necessarily suitable for OPD Li et al. (2026a). Its output distribution may be poorly matched to the student. We therefore construct domain teachers from Knowledge Internalization and Intra-Domain Screening by evaluating teacher-student compatibility.

Knowledge Internalization. For each domain d, we first fine-tune the general model for one epoch on the domain dataset $\mathcal{D}_d$ using supervised fine-tuning. This step is an offline pass over the domain data and provides the initial domain adaptation for subsequent RL. The resulting SFT checkpoint is used only to initialize RL. It is not included in the teacher candidate pool.

Starting from the SFT checkpoint, we perform R consecutive rounds of GRPO Shao et al. (2024) on the same domain training prompts. Each round continues training from the preceding checkpoint, reusing the prompts while generating fresh responses from the current policy. We retain the checkpoint at the end of each round:

$$\mathcal{R}_d = \left\{ q_d^{(1)}, q_d^{(2)}, \ldots, q_d^{(R)} \right\}. \tag{4}$$

Here, $q_d^{(r)}$ denotes the model after the r-th cumulative training round. These candidates are successive checkpoints of a single training trajectory. The replay prompts come from the same domain training split used for SFT. Replay scheduling and domain-specific reward definitions are detailed in Appendix A.4.

Intra-Domain Screening. As shown in Fig. 2, this module takes the initial checkpoint pool $\mathcal{R}_d$ produced by *Knowledge Internalization* as input. Following the privileged-conditioning design of OPSD (Zhao et al., 2026), we augment each checkpoint with sample-specific information $e_d(x)$: the correct answer label for Science and Telecommunication, or the reference tool-call sequence for Tooluse. No reference reasoning traces are supplied. Only the teacher receives $e_d(x)$; the student observes $c = (x, y_{<t})$. When conditioned on the privileged information $e_d(x)$, each checkpoint $q_d^{(r)}$ induces a privileged teacher policy $\tilde{q}_d^{(r)}(\cdot \mid c, e_d)$. The resulting teacher policies form the candidate teacher pool:

$$\mathcal{Q}_d = \left\{ \tilde{q}_d^{(r)} \right\}_{r=1}^{R}. \tag{5}$$

We write $\widetilde{q}_d \in \mathcal{Q}_d$ for a generic candidate, suppressing r; without explicit arguments, this symbol denotes the conditional policy itself. Prompt templates and examples are provided in Appendix A.4.

Fig. 2(a) illustrates the motivation for TSI: teacher-preferred tokens with very low student probability may receive weak logit-level corrections under reverse KL. Inspired by the exploration black hole in advantage-based OPD (Jia et al., 2026), TSI measures how much teacher-preferred probability mass remains accessible to the student. More gradient analysis is provided in Appendix A.2.

Following Fig. 2(b), we fix a calibration subset $\mathcal{S}_d \subset \mathcal{D}_d$ of five training prompts per domain. The student, initialized from the general model, generates m responses per prompt. Each valid response-token position defines a context $c_{i,j,t} = (x_i, y_{i,j,<t})$. We denote the indexed collection of these contexts by $\mathcal{C}_d$, whose size is

$$|\mathcal{C}_d| = \sum_{i=1}^{|\mathcal{S}_d|} \sum_{j=1}^{m} T_{i,j}, \tag{6}$$

where $T_{i,j}$ is the number of valid response tokens in the j-th rollout for prompt x_i. All candidate teachers are evaluated on the same fixed contexts, using the corresponding privileged information $e_d(x_i)$. For a candidate $\widetilde{q}_d$, we define its preferred-token set as

$$\mathcal{F}_{\widetilde{q}_d}(c) = \{ v \in \mathcal{V} : \widetilde{q}_d(v \mid c, e_d(x)) > p_\theta(v \mid c) \}. \tag{7}$$

TSI is the average fraction of this teacher-preferred mass within the student's effective support:

$$\mathrm{TSI}_d(\widetilde{q}_d) = \frac{1}{|\mathcal{C}_d|} \sum_{c \in \mathcal{C}_d} \frac{\sum_{v \in \mathcal{F}_{\widetilde{q}_d}(c)} \widetilde{q}_d(v \mid c, e) \mathbb{I}[p_\theta(v \mid c) \geq \epsilon]}{\sum_{v \in \mathcal{F}_{\widetilde{q}_d}(c)} \widetilde{q}_d(v \mid c, e)}. \tag{8}$$

We use a relatively conservative support threshold of $\epsilon = 0.1$ throughout; higher TSI indicates greater accessibility of teacher-preferred probability mass.

Candidate Screening. Exhaustive cross-domain selection considers R^N teacher combinations, growing exponentially with the number of domains. To reduce this search space while prioritizing student-compatible teachers, we retain the top-k candidates per domain according to TSI:

$$\widehat{\mathcal{Q}}_d = \mathrm{TopK}_{\widetilde{q}_d \in \mathcal{Q}_d} \mathrm{TSI}_d(\widetilde{q}_d). \tag{9}$$

This reduces the number of combinations to k^N: with three domains, retaining three of five candidates per domain reduces the search from 125 to 27 combinations. The retained candidates are passed to *Inter-Domain Screening* in Stage 2, where TSI and CPF jointly select one teacher per domain.

4.2 Joint Optimization Distillation

In the first stage, a set of student-compatible candidate teachers is selected for each domain. However, independently choosing the teacher with the highest TSI in each domain does not necessarily produce an optimal cross-domain combination. Variations in output sharpness between teachers may cause an optimization imbalance when jointly supervising a shared student. To mitigate this issue, we introduce a joint optimization distillation mechanism.

Inter-Domain Screening. We use the dispersion of teacher output entropy as a lightweight proxy for inconsistency in the cross-domain distribution. For a candidate teacher $\widetilde{q}_d$ in domain d, its average token entropy is

$$\bar{H}_d(\widetilde{q}_d) = \frac{1}{|\mathcal{C}_d|} \sum_{c \in \mathcal{C}_d} H(\widetilde{q}_d(\cdot \mid c, e_d)), \tag{10}$$

where

$$H(q(\cdot)) := -\sum_{v \in \mathcal{V}} q(v) \log q(v), \tag{11}$$

Consider a candidate teacher combination

$$\mathcal{G} = (\widetilde{q}_1, \ldots, \widetilde{q}_N), \qquad \widetilde{q}_d \in \widehat{\mathcal{Q}}_d. \tag{12}$$

We define its Cross-domain Penalty Factor (CPF) as

$$\mathrm{CPF}(\mathcal{G}) = \sqrt{\frac{1}{N} \sum_{d=1}^{N} \left(\bar{H}_d - \bar{H}_{\mathcal{G}}\right)^2}, \tag{13}$$

where

$$\bar{H}_{\mathcal{G}} = \frac{1}{N} \sum_{d=1}^{N} \bar{H}_d. \tag{14}$$

A smaller CPF indicates more consistent mean teacher entropy across domains. We use CPF as a heuristic proxy for gradient-scale dispersion, motivated by the conditional analysis in Appendix A.3. Finally, we jointly consider the TSI of each teacher and the CPF across domains. The combination score is defined as

$$S(\mathcal{G}) = \frac{1}{N} \sum_{d=1}^{N} \mathrm{TSI}_d(\widetilde{q}_d) - \lambda \, \mathrm{CPF}(\mathcal{G}), \tag{15}$$

where λ controls the trade-off between teacher suitability and cross-domain consistency. The final set of teachers is selected by

$$\mathcal{G}^\star = \arg \max_{\mathcal{G} \in \widehat{\mathcal{Q}}_1 \times \cdots \times \widehat{\mathcal{Q}}_N} S(\mathcal{G}). \tag{16}$$

Optimality in Eq. 16 refers only to the selection score over the retained candidate pool, not to downstream accuracy. The search space contains k^N combinations and remains manageable in our setting.

Multi-Domain Distillation. After teacher construction and selection, we obtain $\mathcal{G}^\star = (\widetilde{q}_1^\star, \ldots, \widetilde{q}_N^\star)$, containing one selected teacher for each domain. The selected teachers remain fixed throughout student distillation. For a prompt from domain d, the student generates a response conditioned only on the original prompt, while $\widetilde{q}_d^\star$ evaluates the same generated prefixes with access to the privileged information e_d.

We apply Eq. 2 to each domain, replacing the teacher distribution with that of its selected privileged-information teacher. The shared student is optimized using

$$\mathcal{L}_{\mathrm{MOPD}}(\theta) = \sum_{d=1}^{N} \alpha_d \mathcal{L}_{\mathrm{OPD}}(\theta; \mathcal{D}_d, \widetilde{q}_d^\star), \qquad \alpha_d = \frac{1}{N}. \tag{17}$$

At each optimization step, we uniformly sample a domain and draw a mini-batch from that domain. Only the student parameters are updated.

5 Experiments

5.1 Experimental Setup

We evaluate OPDX across three domains: scientific reasoning on the Chemistry L-3 subset of SciKnowEval Feng et al. (2024), Tooluse on ToolAlpaca Tang et al. (2023), and telecommunications on the TeleQnA standard-specification subset Maatouk et al. (2023). Experiments use Qwen3.5-4B-No-Thinking and Qwen3.5-9B-No-Thinking Team (2026) in single- and multi-domain settings. We evaluate each model on the corresponding benchmark evaluation set and report the pass@1 accuracy for each domain. Unless otherwise stated, the teacher and student have the same parameter count.

Method	Qwen3.5-4B				Qwen3.5-9B			
	Science	Tooluse	Tele	Average	Science	Tooluse	Tele	Average
Base	23.47	40.21	57.75	40.48	23.27	47.42	61.75	44.15
SFT	45.36	64.95	61.75	57.35	55.23	65.98	62.50	61.24
GRPO	56.02	67.01	**64.25**	<u>62.43</u>	56.21	<u>69.07</u>	66.00	63.76
OPSD	<u>56.80</u>	<u>68.04</u>	62.00	62.28	<u>71.01</u>	54.64	**66.75**	<u>64.13</u>
Ours	**75.15**	**71.13**	<u>63.25</u>	**69.84**	**78.30**	**73.20**	<u>66.25</u>	**72.58**

Table 1: Single-domain results on three domain benchmarks with Qwen3.5-4B and Qwen3.5-9B. For each method, we report the results of the model with the best performance after two epochs of training. Average is computed across the three domains. The best and second-best results are shown in bold and underlined, respectively.

Method	Qwen3.5-4B				Qwen3.5-9B			
Train on S&T1	**Science**	**Tooluse**	**Tele**	**Average**	**Science**	**Tooluse**	**Tele**	**Average**
SFT	44.97	59.79	/	52.38	<u>57.00</u>	58.76	/	57.88
GRPO	<u>50.69</u>	<u>61.86</u>	/	<u>56.28</u>	54.24	<u>67.01</u>	/	<u>60.63</u>
OPSD	40.24	50.52	/	45.38	46.94	53.61	/	50.28
Ours	**71.99**	**69.07**	/	**70.53**	**73.37**	**72.16**	/	**72.77**
Train on S&T1&T2	**Science**	**Tooluse**	**Tele**	**Average**	**Science**	**Tooluse**	**Tele**	**Average**
SFT	43.00	<u>59.79</u>	<u>60.50</u>	54.43	<u>53.25</u>	62.89	61.25	<u>59.13</u>
GRPO	49.31	55.67	**60.75**	<u>55.24</u>	40.63	<u>64.95</u>	<u>64.25</u>	56.61
OPSD	<u>54.44</u>	53.61	43.00	50.35	40.43	52.58	63.50	52.17
Ours	**69.63**	**73.20**	60.00	**67.61**	**68.05**	**72.16**	**67.25**	**69.15**

Table 2: Multi-domain performance under two-domain and three-domain joint training. **S** denotes the Science dataset, **T1** denotes the Tooluse dataset, and **T2** denotes the Telecommunications dataset. For each method, model size, and training setting, we report the checkpoint with the best average evaluation performance over the jointly trained domains. Average is computed over the domains included in each training setting. The best and second-best results in each model group are shown in bold and underlined, respectively.

5.2 Implementation Details

We implement all experiments using the ms-swift framework Zhao et al. (2024). For OPDX, we perform one epoch of SFT and then conduct up to $R = 5$ rounds of RL with domain-data replay. Each RL checkpoint is augmented with privileged information to form a candidate teacher. For each domain, we use a fixed five-example calibration set to compute TSI with $\epsilon = 0.1$, and retain the top $k = 3$ candidates. We then select the final teacher combination based on aggregate TSI and CPF. During multi-domain OPD, each mini-batch is drawn from a single domain selected uniformly at random. During evaluation, the student is fixed when the candidate teachers are compared, and all candidates for the domain d are evaluated in the same student-visited contexts $\mathcal{C}_d$.

5.3 Comparative Experiments

Baseline. We compare OPDX with three representative post-training baselines in both single-domain and multi-domain settings: standard supervised fine-tuning (SFT), Group Relative Policy Optimization (GRPO) Shao et al. (2024), and on-policy self-distillation (OPSD) Zhao et al. (2026). SFT directly adapts the general model using the domain training data. GRPO optimizes the model with sequence-level rewards, while OPSD provides token-level supervision using privileged information. For a fair comparison, all methods use the same general models, domain data, and evaluation protocol. Detailed implementations and hyperparameters are provided in Appendix A.4.

Single-Domain Experiments. Table 1 shows that OPDX achieves the highest average performance at both model scales. Compared with SFT and GRPO, OPDX yields greater gains across the three domains. These comparisons assess the full pipelines rather than isolating a single mechanism.

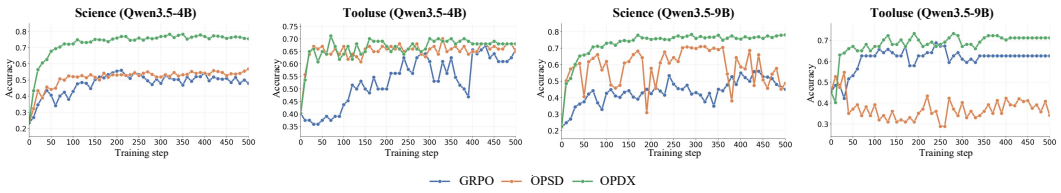


Figure 3: Single-domain training dynamics of GRPO, OPSD, and OPDX on Science and Tooluse with Qwen3.5-4B and Qwen3.5-9B. The four panels show task accuracy over training steps for each model-domain setting.

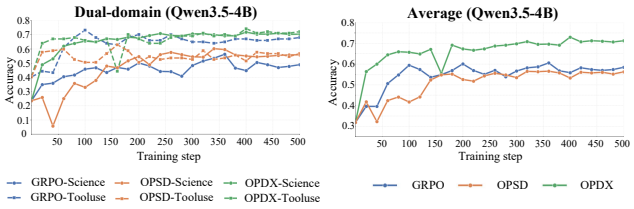


Figure 4: Dual-domain training dynamics of GRPO, OPSD, and OPDX on Science and Tooluse using Qwen3.5-4B. The left panel reports the accuracy of each domain over training steps, while the right panel reports the average accuracy across the two domains.

This suggests that dense token-level supervision from iteratively constructed teachers enables more effective knowledge internalization than offline target fitting or scalar rewards. OPDX also yields more balanced gains than OPSD and avoids severe degradation in weaker domains, indicating that TSI-selected teachers offer more stable and accessible supervision.

Teacher performance does not directly determine teacher suitability. In some domains, GRPO itself provides limited performance gains, yet its checkpoints can still guide the distilled student to substantially better results. This suggests that teacher-student distributional compatibility is more important than teacher accuracy alone. RL checkpoints may also provide smoother token distributions, leading to clearer supervision and more effective student learning.

Multi-Domain Experiments. Table 2 reports the multi-domain results. Under uniform domain sampling, SFT, GRPO, and OPSD yield limited or uneven gains, indicating the difficulty in integrating heterogeneous domain knowledge into a shared model. In contrast, OPDX achieves the highest average performance in both two-domain and three-domain settings across model scales. By selecting teachers with high individual suitability and low CPF, OPDX mitigates optimization imbalance and enables more stable capability integration.

In the three-domain setting, OPDX maintains strong performance across Science, Tooluse, and Telecommunication. Although it does not achieve the best score on every individual task, it obtains the highest overall average for both model sizes. Moreover, performance in some domains exceeds the corresponding single-domain results, suggesting positive cross-domain transfer. These results demonstrate that OPDX can integrate multiple domain capabilities while reducing severe performance trade-offs across domains.

5.4 Training Dynamics Analysis

Fig. 3 presents the single-domain training dynamics of Qwen3.5-4B and Qwen3.5-9B on Science and Tooluse. OPDX converges to strong performance earlier and maintains smoother improvements than GRPO and OPSD. We observe that GRPO exhibits substantial fluctuations, particularly on the Qwen3.5-4B model, while OPSD fails to sustain its early gains and shows pronounced performance degradation on several tasks. Similar instability has been reported in recent OPD studies, showing that teacher-student distribution mismatch can produce unreliable gradients, reduced distributional diversity, and even optimization failure (Jang et al., 2026; Xing et al., 2026; Jia et al., 2026). In contrast, OPDX maintains more consistent improvements across both model sizes, supporting the effectiveness of student-compatible teacher selection.

Method	Qwen3.5-4B				Qwen3.5-9B			
Single Domain	**Science**	**Tooluse**	**Tele**	**Average**	**Science**	**Tooluse**	**Tele**	**Average**
OPDX	**75.15**	**71.13**	**63.25**	**69.84**	**78.30**	**73.20**	**66.25**	**72.58**
w/o-Prompt-OPDX	70.61	70.10	62.00	67.57	70.41	69.07	63.75	67.74
w/o-TSI-OPDX	72.78	70.10	63.00	68.63	73.77	68.04	64.50	68.77
Dual Domain	**Science**	**Tooluse**	**Tele**	**Average**	**Science**	**Tooluse**	**Tele**	**Average**
OPDX	71.99	**69.07**	/	**70.53**	73.37	**72.16**	/	**72.77**
w/o-Prompt-OPDX	**74.36**	64.95	/	69.66	73.57	69.07	/	71.32
w/o-CPF-OPDX	73.96	67.01	/	70.49	**73.96**	67.01	/	70.49

Table 3: Ablation results of OPDX on single-domain and dual-domain benchmarks. *w/o-Prompt*-OPDX removes privileged-information augmentation, *w/o-TSI*-OPDX selects the teacher from the final RL round instead of using TSI-based selection, and *w/o-CPF*-OPDX selects the teacher group with the highest aggregate TSI instead of using CPF-based selection. The best result for each domain under each training setting is highlighted in bold.

Fig. 4 shows the dynamics of Qwen3.5-4B jointly trained on Science and Tooluse. GRPO and OPSD exhibit unbalanced trajectories: an improvement in one domain is often accompanied by stagnation or degradation in the other. This behavior is consistent with previous findings that heterogeneous tasks can induce weakly aligned or interfering gradients when sharing the same model parameters (Chen et al., 2017; Li et al., 2026b). In contrast, OPDX improves both domains more steadily and with fewer oscillations. These results suggest that selecting a low-conflict teacher combination mitigates cross-domain optimization imbalance and stabilizes multi-domain OPD.

5.5 Ablation Studies

For the single-domain evaluation, we consider two ablations: *w/o Prompt*, which removes privileged-information augmentation, and *w/o TSI*, which replaces TSI-based screening with the final RL checkpoint. For the multi-domain evaluation, we examine *w/o Prompt* and *w/o CPF*, where the latter replaces CPF-based selection with the teacher group having the highest aggregate TSI. The complete results are reported in Table 3.

In the single-domain setting, OPDX achieves the highest average performance on both model scales. In Qwen3.5-4B, it outperforms *w/o Prompt* and *w/o TSI* by 2.27 and 1.22 points, respectively. The corresponding gains in Qwen3.5-9B are 4.84 and 3.81 points. These results show that privileged information strengthens token-level supervision, while TSI improves teacher-student compatibility.

Removing CPF increases Science accuracy but degrades Tooluse performance, reducing dual-domain averages by 0.04 and 2.28 points on the 4B and 9B models, respectively. This suggests that CPF improves cross-domain balance. Even without privileged information, OPDX retains domain-specific teachers trained through SFT and RL and outperforms the strongest baselines in average single-domain accuracy (Tables 1 and 3). Privileged conditioning further improves single-domain averages by 2.27 and 4.84 points and dual-domain averages by 0.87 and 1.45 points, respectively. However, dual-domain Science accuracy decreases, indicating that these benefits vary across domains. Additional results are provided in Appendix A.5.

6 Conclusion

In this paper, we propose OPDX, a two-stage framework for stable multi-domain on-policy distillation. OPDX improves optimization stability by internalizing domain knowledge and selecting student-compatible teachers before distillation, thereby reducing teacher-student mismatch. It further incorporates a cross-domain penalty to identify low-conflict teacher combinations, mitigating inter-teacher inconsistency and stabilizing multi-teacher distillation. Experiments across single- and multi-domain settings and different model scales demonstrate that OPDX consistently improves task performance and training stability over strong baselines. We hope that this work highlights the importance of teacher selection for scalable multi-domain OPD and inspires future research on adaptive teacher selection strategies.

LLM USAGE STATEMENT

In this work, we only employed a Large Language Model (LLM) as an auxiliary tool to **polish the manuscript**. Specifically, its use was focused on enhancing the clarity and fluency of the writing to better articulate our core insights and findings. We do not use LLMs for tasks such as reference retrieval or figure generation. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, the hyperparameters and environmental configurations for our main experiments are detailed in the manuscript, specifically within Sections 5.2 and Appendix A.4. Furthermore, the source code will be made publicly available in the next version.

IMPACT STATEMENT

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stańczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2023. URL https://api.semanticscholar.org/CorpusID:263610088.

Tianle Chen, Jiao Ou, Ziyuan Liu, Ruiming Tang, Jian Liang, and Han Li. Counteraction-aware multi-teacher on-policy distillation for general capability recovery with domain preservation. *ArXiv*, abs/2605.27115, 2026. URL https://api.semanticscholar.org/CorpusID:288672850.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. *ArXiv*, abs/1711.02257, 2017. URL https://api.semanticscholar.org/CorpusID:4703661.

Kehua Feng, Xinyi Shen, Weijie Wang, Zhuang Xiang, Yuqi Tang, Qiang Zhang, and Keyan Ding. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. 2024. URL https://api.semanticscholar.org/CorpusID:281887235.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. 2023. URL https://api.semanticscholar.org/CorpusID:259164722.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42217–42227, 2026.

Nan Jia, Haojin Yang, Xingchen Ma, Jiesong Lian, Shuai Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. Asymmetric on-policy distillation: Bridging exploitation and imitation at the token level. *ArXiv*, abs/2605.06387, 2026. URL https://api.semanticscholar.org/CorpusID:288014846.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *ArXiv*, abs/2603.07079, 2026a. URL https://api.semanticscholar.org/CorpusID:286372090.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *ArXiv*, abs/2603.07079, 2026b. URL https://api.semanticscholar.org/CorpusID:286372090.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *ArXiv*, abs/2603.11137, 2026. URL https://api.semanticscholar.org/CorpusID:286489350.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *ArXiv*, abs/2604.13016, 2026a. URL https://api.semanticscholar.org/CorpusID:287472296.

Yu Li, Mingyang Yi, Xiuyu Li, Ju Fan, Fuxin Jiang, Binbin Chen, Peng Li, Jie Song, and Tieying Zhang. Reasoning and tool-use compete in agentic rl: From quantifying interference to disentangled tuning. *arXiv preprint arXiv:2602.00994*, 2026b.

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*, abs/2604.08527, 2026. URL https://api.semanticscholar.org/CorpusID:287256383.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, et al. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026.

Ali Maatouk, Fadhel Ayed, Nicola Piovesan, Antonio De Domenico, Mérouane Debbah, and Zhi-Quan Luo. Teleqna: A benchmark dataset to assess large language models telecommunications knowledge. *IEEE Network*, 40:253–260, 2023. URL https://api.semanticscholar.org/CorpusID:264436416.

Moonshot AI. Kimi K3: Open frontier intelligence. https://www.kimi.com/blog/kimi-k3, 2026. Technical blog, accessed July 27, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *ArXiv*, abs/2602.04942, 2026. URL https://api.semanticscholar.org/CorpusID:285304042.

Daniil Plyusov, Alexey Gorbatovski, Alexey Malakhov, Nikita Balagansky, Boris Shaposhnikov, Daria Korotyshova, and Daniil Gavrilov. Trust-region behavior blending for on-policy distillation. *ArXiv*, abs/2605.31159, 2026. URL https://api.semanticscholar.org/CorpusID:288814565.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. URL https://api.semanticscholar.org/CorpusID:267412607.

Jiabin Shen, Guang Chen, and Chengjun Mao. When top-k misses the decision: Tool-call drift in multi-teacher on-policy distillation. 2026. URL https://api.semanticscholar.org/CorpusID:289973139.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *ArXiv*, abs/2601.19897, 2026. URL https://api.semanticscholar.org/CorpusID:285071839.

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*, abs/2306.05301, 2023. URL https://api.semanticscholar.org/CorpusID:259108190.

Qwen Team. Qwen3.5-omni technical report. *ArXiv*, abs/2604.15804, 2026. URL https://api.semanticscholar.org/CorpusID:287607592.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. 2026. URL https://api.semanticscholar.org/CorpusID:288861606.

Lei Yang, Siyu Ding, and Deyi Xiong. A local perturbation theory for cross-domain interference and recovery in multi-domain rl. 2026. URL https://api.semanticscholar.org/CorpusID:288861340.

Shan You, Chang Xu, and Dacheng Tao. Learning from multiple teacher networks. pp. 1285–1294, 08 2017. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098135.

Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. Reinforced multi-teacher selection for knowledge distillation. In *AAAI Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:228376532.

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.

Ke Zhang, Yunjie Tian, Dongdi Zhao, Yijiang Li, Yuanye Liu, Vishal M. Patel, and Di Fu. On-policy distillation with best-of-n teacher rollout selection. *ArXiv*, abs/2605.09725, 2026. URL https://api.semanticscholar.org/CorpusID:288259621.

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*, abs/2601.18734, 2026. URL https://api.semanticscholar.org/CorpusID:285050713.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning. *ArXiv*, abs/2408.05517, 2024. URL https://api.semanticscholar.org/CorpusID:271854774.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. Sod: Step-wise on-policy distillation for small language model agents. *ArXiv*, abs/2605.07725, 2026. URL https://api.semanticscholar.org/CorpusID:288147873.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Wei Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *ArXiv*, abs/2605.11182, 2026. URL https://api.semanticscholar.org/CorpusID:288259455.

A APPENDIX

A.1 OVERVIEW

This supplementary material provides additional theoretical, implementation, and experimental details of OPDX. **Appendix A.2** motivates the Teacher Suitability Index (TSI) and analyzes its theoretical properties and connection to accessible reverse-KL supervision. **Appendix A.3** analyzes the Cross-domain Penalty Factor (CPF) as a proxy for the gradient-scale imbalance. **Appendix A.4** describes the datasets, the evaluation protocols, and the implementation details. **Appendix A.5** provides additional ablation experiments on teacher selection, mini-batch composition, and domain combinations. These materials provide the theoretical and empirical details required for a comprehensive understanding and reproduction of OPDX.

A.2 MOTIVATION AND PROPERTIES OF THE TEACHER SUITABILITY INDEX

Following the notation in the main text, let $p(v) = p_\theta(v \mid c)$ and $q(v) = q_d^e(v \mid c, e)$ denote the student and teacher distributions, respectively. Following the standard on-policy distillation, we optimize the student in its self-generated contexts using dense token-level teacher supervision (Agarwal et al., 2023; Zhao et al., 2026). The token-level reverse-KL objective is

$$\ell(c) = \sum_{v \in \mathcal{V}} p(v) \log \frac{p(v)}{q(v)}. \tag{18}$$

Let z_v be the student logit of token v. Using the softmax Jacobian, its gradient is as follows.

$$\frac{\partial \ell(c)}{\partial z_v} = p(v) \left(\log \frac{p(v)}{q(v)} - D_{\mathrm{KL}}(p \| q) \right). \tag{19}$$

For a teacher-preferred token satisfying $q(v) > p(v)$, the gradient in Eq. 19 is negative, so the gradient descent increases its student logit. However, its magnitude is multiplied by $p(v)$. Moreover, for any fixed $q(v) > 0$,

$$\lim_{p(v) \to 0} p(v) \log \frac{q(v)}{p(v)} = 0. \tag{20}$$

This support-mismatch effect is consistent with recent analyses showing that teacher–student distribution mismatch can substantially weaken or destabilize on-policy distillation (Zhu et al., 2026). Therefore, when a teacher-preferred token has an extremely low probability under the student, its corrective gradient becomes negligible even if the teacher assigns it a high probability. This motivates measuring how much teacher–preferred probability mass lies within the effective support of the student.

Define the set of teacher-favored tokens and effective student support as

$$\mathcal{F}_q(c) = \{v \in \mathcal{V} : q_v > p_v\}, \tag{21}$$

$$\mathcal{S}_\epsilon(c) = \{v \in \mathcal{V} : p_v \geq \epsilon\}, \tag{22}$$

where ϵ is a support threshold. The accessible teacher-preferred mass is

$$A_\epsilon(c; q) = \frac{\sum_{v \in \mathcal{F}_q(c)} q(v) \mathbb{I}[p(v) \geq \epsilon]}{\sum_{v \in \mathcal{F}_q(c)} q(v)}. \tag{23}$$

Consequently, the TSI in calibration contexts $\mathcal{C}_d$ is

$$\mathrm{TSI}_{d,\epsilon}(q) = \frac{1}{|\mathcal{C}_d|} \sum_{c \in \mathcal{C}_d} A_\epsilon(c; q). \tag{24}$$

Thus, TSI measures the proportion of teacher–preferred probability mass that remains accessible under the current student policy. It satisfies the following requirements.

$$0 \leq \mathrm{TSI}_{d,\epsilon}(q) \leq 1, \tag{25}$$

Domain	source	Train	Calib.	Test
Science	SciKnowEval	2674	5	507
Tooluse	ToolAlpaca	4046	5	97
Telecommunication	TeleQnA	1598	5	400

Table 4: Dataset statistics for the three domains. The calibration prompts are sampled from and included in the corresponding training sets.

Hyperparameter	**SFT**	**GRPO**	**OPSD**	**OPDX**
Learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}
Batch size	32	32	32	32
Max. sequence length	2048 / 1024	2048 / 1024	2048 / 1024	2048 / 1024
Max. gradient norm	1	1	1	1
Generation hyperparameters				
Generations / prompt	–	4	1	1
Rollout temperature	–	0.9	1.2	0.9
Max. generation length	–	2048 / 1024	2048 / 1024	2048 / 1024
OPD-specific hyperparameters				
Distillation objective	–	–	JSD	Reverse KL
Teacher-logit range	–	–	Full vocabulary	Full vocabulary
OPDX-only hyperparameters				
Replay rounds R	–	–	–	5
Calibration prompts / domain	–	–	–	5
TSI threshold ϵ	–	–	–	0.1
Candidates / domain	–	–	–	5
Retained candidates k	–	–	–	3
Penalty coefficient λ	–	–	–	1

Table 5: Key hyperparameters for SFT, GRPO, OPSD, and OPDX. All methods use Qwen3.5-4B and Qwen3.5-9B as general models. The maximum sequence and generation lengths are 2048 for Science and 1024 for Tooluse and Telecommunication. A dash indicates that the corresponding hyperparameter is not applicable. Unless otherwise specified, identical configurations are used across domains and model scales.

where a larger TSI indicates that a larger fraction of the teacher-preferred probability mass lies within the effective support of the student and is therefore more likely to provide usable reverse-KL supervision. In addition, for $\epsilon_1 \leq \epsilon_2$,

$$\mathrm{TSI}_{d,\epsilon_2}(q) \leq \mathrm{TSI}_{d,\epsilon_1}(q), \tag{26}$$

because a larger threshold defines stricter student support. Importantly, TSI depends on both the teacher and the current student; therefore, a teacher with greater task precision is not necessarily a more suitable teacher for policy distillation Zhang et al. (2026). Empirically, we use a fixed $\epsilon = 0.1$ for all candidates to ensure a consistent comparison.

A.3 Derivation and Analysis of the Cross-Domain Penalty Factor

We analyze why the CPF can serve as a local first-order proxy for gradient-scale imbalance Shen et al. (2026). All domains are sampled uniformly and each input contains an explicit domain label that routes it to the corresponding teacher. When candidate teacher groups are compared, the student parameters and the calibration contexts visited by the students are fixed.

For a context c from the domain d, let

$$p(v) = p_\theta(v \mid c), \qquad q_d^e(v) = q_d^e(v \mid c, e_d). \tag{27}$$

The token-level reverse-KL objective is

$$\ell_d(c) = \sum_{v \in \mathcal{V}} p(v) \log \frac{p(v)}{q_d^e(v)}. \tag{28}$$

Define the average parameter gradient and its scale for domain d as

$$\mathbf{g}_d = \frac{1}{|\mathcal{C}_d|} \sum_{c \in \mathcal{C}_d} \nabla_\theta \ell_d(c), \qquad G_d = ||\mathbf{g}_d||_2. \tag{29}$$

For two domains i and j,

$$||\mathbf{g}_i - \mathbf{g}_j||^2 = G_i^2 + G_j^2 - 2G_i G_j \cos\theta_{ij}, \tag{30}$$

where θ_{ij} is the angle between their gradients. Equation 30 separates gradient magnitudes from directional interactions. The analysis below concerns only dispersion of gradient magnitudes; it does not establish alignment of cross-domain gradients.

Define the average teacher entropy in domain d as

$$\bar{H}_d = \frac{1}{|\mathcal{C}_d|} \sum_{c \in \mathcal{C}_d} H\left(q_d^e(\cdot \mid c, e_d)\right), \tag{31}$$

where

$$H(q_d^e) = -\sum_{v \in \mathcal{V}} q_d^e(v) \log q_d^e(v). \tag{32}$$

For a teacher combination $\mathcal{G} = (q_1^e, \ldots, q_N^e)$, let

$$\bar{H}_\mathcal{G} = \frac{1}{N} \sum_{d=1}^{N} \bar{H}_d. \tag{33}$$

Teacher entropy does not uniquely determine the gradient scale in general Jin et al. (2026b). For a conditional local analysis, we assume that the retained candidates admit the following model with a sensitivity κ shared across domains and candidate groups:

$$G_d = G_0 + \kappa\left(\bar{H}_d - \bar{H}_\mathcal{G}\right) + R_d, \tag{34}$$

where κ is the common local entropy sensitivity and R_d collects domain-dependent and higher-order effects.

Let

$$\bar{G} = \frac{1}{N} \sum_{d=1}^{N} G_d, \qquad \bar{R} = \frac{1}{N} \sum_{d=1}^{N} R_d. \tag{35}$$

The centering of Eq. 34 gives

$$G_d - \bar{G} = \kappa\left(\bar{H}_d - \bar{H}_\mathcal{G}\right) + \left(R_d - \bar{R}\right). \tag{36}$$

Define the cross-domain gradient-scale standard deviation as

$$\sigma_G = \sqrt{\frac{1}{N} \sum_{d=1}^{N} (G_d - \bar{G})^2}, \tag{37}$$

and the entropy-dispersion penalty as

$$P(\mathcal{G}) = \sqrt{\frac{1}{N} \sum_{d=1}^{N} \left(\bar{H}_d - \bar{H}_\mathcal{G}\right)^2}. \tag{38}$$

Let us

$$\sigma_R = \sqrt{\frac{1}{N} \sum_{d=1}^{N} (R_d - \bar{R})^2}. \tag{39}$$

Model	Teacher Variant	Science	Tooluse	Average
	Single-Domain			
4B	SFT Teacher	73.57	47.42	60.50
	RL Teacher	46.35	69.07	57.71
	OPDX	**75.15**	**71.13**	**73.14**
	Multi-Domain			
4B	SFT Teacher	67.46	47.42	57.44
	RL Teacher	36.88	69.07	52.98
	OPDX	**71.99**	**69.07**	**70.53**

Table 6: Ablation study on different teacher checkpoint sources. All teachers use the same privileged-information prompt and OPD training configuration.

Model	Batch strategy	Science	Tooluse	Average
4B	Mixed-domain	71.01	65.98	68.50
	Single-domain	71.99	69.07	70.53

Table 7: Effect of mini-batch composition under joint training on Science and Tooluse. Both strategies use uniform domain sampling.

Applying the reverse triangle inequality to Eq. 36 yields

$$|\sigma_G - |\kappa| P(\mathcal{G})| \leq \sigma_R. \tag{40}$$

Equivalently, the cross-domain gradient-scale dispersion is bounded by

$$\max\{0,\ |\kappa| P(\mathcal{G}) - \sigma_R\} \leq \sigma_G \leq |\kappa| P(\mathcal{G}) + \sigma_R. \tag{41}$$

We additionally assume a uniform bound on residual dispersion across candidate teacher groups. This assumption is not implied by TSI screening:

$$\sigma_R(\mathcal{G}) \leq \delta, \qquad \forall \mathcal{G} \in \mathbb{G}, \tag{42}$$

where $\mathbb{G}$ denotes the set of candidate teacher groups and δ is a group-independent residual bound. Combining Eqs. 40 and 42 gives

$$|\sigma_G(\mathcal{G}) - |\kappa| P(\mathcal{G})| \leq \delta, \tag{43}$$

and, in particular,

$$\sigma_G(\mathcal{G}) \leq |\kappa| P(\mathcal{G}) + \delta. \tag{44}$$

Thus, minimizing $P(\mathcal{G})$ minimizes a local upper bound on the cross-domain gradient-scale dispersion. Moreover, for two candidate teacher groups $\mathcal{G}_1$ and $\mathcal{G}_2$, if

$$|\kappa| \left(P(\mathcal{G}_2) - P(\mathcal{G}_1)\right) > 2\delta, \tag{45}$$

then Eqs. 43 and 45 guarantee that

$$\sigma_G(\mathcal{G}_1) < \sigma_G(\mathcal{G}_2). \tag{46}$$

Therefore, when the difference in the dispersion of the entropy between two candidate groups dominates the residual uncertainty, the CPF preserves their ordering in terms of the gradient-scale imbalance.

Under these assumptions, $|\kappa| P(\mathcal{G})$ approximates gradient-scale dispersion with error at most δ. A smaller CPF lowers the conditional upper bound, while an ordering of actual gradient-scale dispersion requires Eq. 45. This analysis motivates the selection heuristic but does not guarantee reduced directional conflict or convergence stability.

Method	Qwen3.5-4B				Qwen3.5-9B			
Train on S&T2	**Science**	**Tooluse**	**Tele**	**Average**	**Science**	**Tooluse**	**Tele**	**Average**
SFT	43.79	/	61.00	52.40	59.96	/	63.00	61.48
GRPO	43.79	/	**61.25**	52.52	43.98	/	62.25	53.12
OPSD	25.44	/	60.50	42.97	38.66	/	63.25	50.96
Ours	**72.78**	/	61.00	**66.89**	**74.56**	/	**66.75**	**70.66**
Train on T1&T2	**Science**	**Tooluse**	**Tele**	**Average**	**Science**	**Tooluse**	**Tele**	**Average**
SFT	/	53.61	59.25	56.43	/	58.76	62.00	60.38
GRPO	/	54.64	59.00	56.82	/	58.76	62.75	60.76
OPSD	/	50.52	48.75	49.64	/	38.14	**63.50**	50.82
Ours	/	**61.86**	**62.25**	**62.06**	/	**62.89**	63.25	**63.07**

Table 8: Main experimental results on Multi-Domain benchmarks. **S** denotes the Science dataset, **T1** denotes the Tooluse dataset, and **T2** denotes the Telecommunications dataset. For each method, model size, and training setting, we report the checkpoint with the best average evaluation performance over the jointly trained domains. Average is computed over the domains included in each training setting. The best results in each group are highlighted in bold.

A.4 Implementation Details

Datasets and Data Splits. We evaluate OPDX in three domains: scientific reasoning, tooluse, and telecommunications. For Science and Tooluse, we follow the experimental setup of Shenfeld et al. (2026), including its predefined training–test partitions and strict-alignment evaluation protocols. Science uses the Chemistry L-3 subset of SciKnowEval Feng et al. (2024), while Tooluse is constructed from ToolAlpaca Tang et al. (2023). To extend domain coverage, we additionally select the TeleQnA standard-specification questions Maatouk et al. (2023) and randomly divide them into training and test sets with an approximately $4{:}1$ ratio.

Dataset statistics are summarized in Table 4. For the computation of TSI and CPF, we fix five calibration prompts from the training split of each domain. The same calibration prompts and student-generated contexts are used to evaluate all candidate teachers for each domain.

Evaluation Protocol. For evaluation, we use greedy decoding and report pass@1 accuracy. A response is considered correct only when its parsed prediction exactly matches the reference answer under the domain-specific evaluation rules; invalid or unparseable outputs are counted as incorrect. We use the benchmark test sets for both checkpoint selection and reporting, without a separate validation split. The reported scores are therefore subject to checkpoint-selection bias and are not independent test estimates.

During training, we evaluate the models every 20 optimization steps. For single-domain training, the checkpoint is selected according to the best performance after two epochs of training and its accuracy is reported once. For multi-domain training, checkpoint selection is based on the average evaluation accuracy across all jointly trained domains.

Training and Implementation Details. We compare OPDX with three representative post-training baselines: SFT, GRPO, and OPSD. All methods use the same Qwen3.5 foundation model Team (2026) and domain-specific training data. OPDX additionally incurs teacher-construction and screening costs. Unless otherwise specified, the learning rate is set to (1×10^{-6}), the effective batch size is set to 32, and the maximum gradient norm is set to 1. The maximum sequence and generation lengths are set to 2048 for the science task and 1024 for the tooluse and telecommunication tasks. Table 5 summarizes the key training hyperparameters. All models are fully fine-tuned using the ms-swift framework Zhao et al. (2024).

SFT directly fine-tunes the general model on the reference responses of each domain using the standard token-level cross-entropy objective. In the multi-domain setting, training examples are uniformly sampled across the jointly trained domains.

GRPO performs reinforcement learning using domain-specific sequence-level rewards Shao et al. (2024). For each prompt, four responses are sampled with a rollout temperature of 0.9, and their relative rewards are used to optimize the policy. Reference targets are extracted from the assistant responses in the SFT training data and stored in the `solution` field: answer-tag contents for Science and Telecommunication, and tool-call blocks for Tooluse.

For Science and Telecommunication, the reward is $r = r_{\text{acc}} + 0.1\, r_{\text{fmt}}$, where both components are binary. The accuracy reward compares the answer extracted from `<answer>` tags with the reference using `math_verify`, with exact string matching as a fallback. The format reward checks for answer tags and either a reasoning section or a nonempty thinking block.

For Tooluse, we use a single custom binary reward with unit weight. Predictions and references are processed by the same parser to extract action names and JSON argument dictionaries. The reward equals one if both the action-name multisets and the merged argument dictionaries match, and zero otherwise. Argument dictionaries are merged across calls, with later values overwriting earlier ones for duplicate keys. Thus, matching ignores call order and the association between arguments and individual calls. No partial credit or additional format reward is used.

OPSD follows the on-policy self-distillation setting, in which the student generates its own trajectories and receives token-level supervision from a privileged-information teacher Zhao et al. (2026). Following its original implementation, OPSD optimizes Jensen–Shannon divergence over full-vocabulary teacher logits. It uses one student rollout per prompt with a sampling temperature of 1.2.

For Science and Telecommunication, the privileged information is only the correct answer label. We use the following template:

```
{question }
You already know the correct answer below.
Reasoning that arrives at that answer.
Correct answer:  <answer>{answer}</answer>
Now briefly write the reasoning and the answer.
```

Tooluse requires structured tool-call prediction and therefore uses a separate template:

```
Given the correct tool calls below, add a reasoning sentence
before each step.
Do not modify the Action/Action Input lines.  Output format:
[reasoning]\nAction:  ...\nAction Input:
...\n[reasoning]\nAction:  ...\nAction Input:  ...\n...\n\n
Correct calls:\n{answer}\n\n
Now write the full sequence with reasoning.
```

No complete reasoning trace or explanatory reference answer is provided in any domain. The privileged information consists only of the answer label or the call to the reference tool.

OPDX follows the training configurations of the corresponding SFT and GRPO baselines, except for the number of optimization iterations during teacher construction. It also adopts the same augmentation prompts as OPSD Zhao et al. (2026). For multi-domain OPD, OPDX uses reverse KL divergence with a temperature of 0.9.

A.5 More Ablation Studies

We conduct additional experiments to examine the effects of teacher construction, mini-batch composition, and domain combinations. Unless otherwise specified, all variants use the same data partitions, training budget, optimization settings, and evaluation protocol as the full OPDX model. The checkpoints are evaluated every 20 optimization steps. For multi-domain experiments, all per-domain results are reported from the checkpoint with the highest average accuracy over the jointly trained domains.

Ablation on SFT and RL Teachers. We first investigate whether iterative RL checkpoints provide more effective supervision than the cold-start SFT model. All compared teachers use the same privileged-information prompt, so the only difference is the source of the teacher checkpoint.

We consider the following variants: *SFT Teacher*, which directly uses the cold-start SFT checkpoint; and *RL Teacher*, which directly uses the Iterative RL checkpoint with the highest task accuracy. Instead, OPDX selects the teacher according to TSI. This comparison examines whether teacher suitability is better captured by TSI than by training stage or teacher accuracy alone.

Ablation on Mini-Batch Composition. We next study whether the composition of each mini-batch affects cross-domain optimization. We compare two strategies while keeping the expected domain sampling probability uniform.

In *Mixed-Domain Batches*, each mini-batch contains an equal number of samples from all jointly trained domains. In *Single-Domain Batches*, a domain is first sampled uniformly, and the entire mini-batch is then drawn from that domain. Therefore, the two strategies use the same expected amount of data from each domain and differ only in whether multiple domains are mixed within one parameter update.

Additional Dual-Domain Experiments. The main experiments primarily consider joint training on Science and Tooluse. To verify that the effectiveness of OPDX is not specific to this pair, we evaluate all three possible two-domain combinations: Science and Tooluse, Science and Telecommunication, and Tooluse and Telecommunication.

For each combination, the two domains are sampled uniformly during training. The reported per-domain results are obtained from the same checkpoint, selected according to the highest average accuracy over the two jointly trained domains.