WEIGHTED ON-POLICY DISTILLATION FOR MULTI-DOMAIN CAPABILITY INTEGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-domain post-training seeks to combine specialist skills while preserving capabilities of the reference model (the starting model). We propose Weighted On-Policy Distillation (WOPD), which evaluates specialists on shared verifiable training prompts and assigns domain-specific teacher weights in the distillation loss based on their success rates. The reference model also serves as a teacher with a fixed weight. All teachers supervise the same student-generated tokens through a weighted reverse-KL objective. In a five-domain study with Qwen3-4B, WOPD exceeds equal weighting on all five in-domain benchmarks by 1.6–15.8 percentage points under the same six-teacher pool. Compared with domain-routing Multi-Teacher On-Policy Distillation (MOPD), it improves the in-domain mean by 1.1 points and the out-of-domain mean by 3.8 points. WOPD matches the reference model on the out-of-domain math benchmark and exceeds it on the other four. Removing the reference model from the teacher pool lowers every out-of-domain score by 3.0–6.6 points. Combining score-based specialist weights with reference supervision improves in-domain performance without out-of-domain losses relative to the reference model.

1 INTRODUCTION

Domain-specific post-training can produce models with strong specialist capabilities. Integrating these capabilities into a single model, however, requires more than improving performance on the adaptation tasks: the resulting model should also retain the capabilities of the model it starts from, which we call the reference model. Gains on one goal can come with losses on the other. In our five-domain study, training directly on mixed verifier rewards improves financial question answering by 31.95 percentage points while reducing out-of-domain math performance by 6.60 points (Figure 1).

On-policy distillation (OPD) provides a way to transfer capabilities by having teachers supervise responses generated by the current student (Agarwal et al., 2024; Gu et al., 2024). With multiple specialists, the method must decide how much each teacher contributes. Multi-Teacher On-Policy Distillation (MOPD) routes each prompt to its in-domain specialist (Ma et al., 2026). Yet a specialist's training domain does not fully describe its capabilities. Our cross-domain evaluation shows that specialists also solve problems outside their own domains, with success rates that vary across teachers and tasks (Section 4.1). Domain routing leaves this additional supervision unused, whereas uniform weighting includes all teachers without accounting for these differences. These observations motivate using measured domain-level capabilities to guide teacher allocation.

We propose **weighted on-policy distillation (WOPD)**, which assigns distinct roles to specialist and reference supervision. Specialists provide the capabilities we seek to integrate, so their relative weights depend on their measured success in each domain. The reference model also serves as a teacher. Its contribution is set independently and keeps the student close to the reference model. Before distillation, we evaluate teachers on shared verifiable training prompts and compute a fixed specialist weight vector for each domain. During training, all frozen teachers compute log-probabilities of the same student-generated tokens, and the prompt's domain determines the weights used in the reverse-KL objective.

We evaluate WOPD on Qwen3-4B, using test sets from the five training domains and five out-of-domain benchmarks. With the same six-teacher pool, training data, optimizer, and training budget, WOPD scores higher than uniform weighting on all five in-domain benchmarks, by 1.60–15.80

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

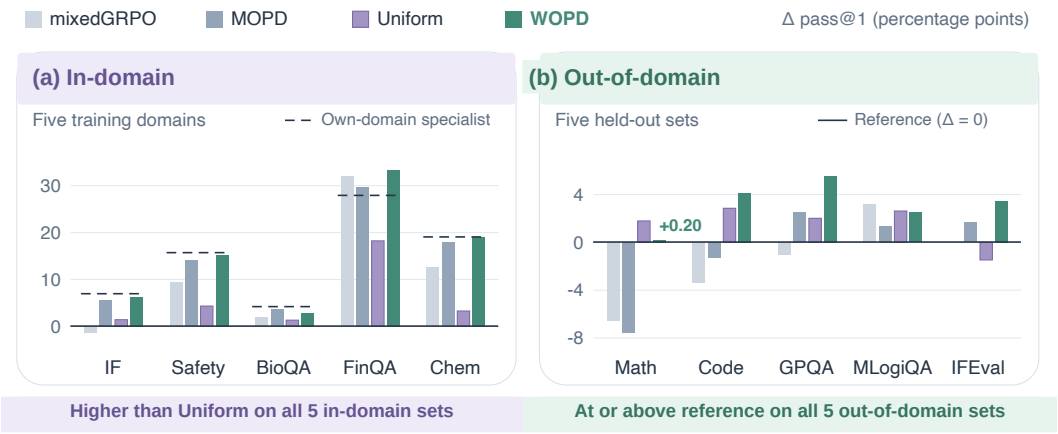

Figure 1: **In-domain gains do not determine out-of-domain performance.** Change in pass@1, in percentage points, relative to the Qwen3-4B reference (solid zero lines; panels use different axis ranges). **(a)** WOPD exceeds Uniform on all five in-domain sets; dashed markers show the own-domain GRPO specialists. **(b)** WOPD matches the reference on Math and exceeds it on the other four out-of-domain sets. Uniform is higher than WOPD on Math and slightly higher on MLogiQA. Uniform weights the same six teachers equally, with the same training budget. MOPD routes to one specialist; mixedGRPO uses verifier rewards without teachers. Both improve most in-domain sets but fall below the reference on Math and Code. These are single-run point estimates (Table 1).

percentage points (Table 1). WOPD also matches the reference model on Math and exceeds it on the other four out-of-domain benchmarks. To examine reference supervision separately, we remove the reference model from the teacher pool while holding the relative specialist weights fixed. This lowers every out-of-domain score by 3.00–6.60 points, while in-domain scores change by at most 1.80 points (Table 2). The reference model therefore contributes separately from the specialist weights.

Our contributions are:

- We introduce WOPD, a multi-teacher on-policy distillation method that derives fixed domain-specific specialist weights from verifier success and independently controls reference supervision (Section 3).
- We compare WOPD with domain routing and uniform weighting, and isolate the reference model's effect with an ablation that removes only this teacher (Section 4.2; Tables 1 and 2).
- We show that, at each state, the weighted reverse-KL objective trains the student toward a single effective teacher built from the teacher distributions. This explains how the weights shape the training signal: equal weights dilute the own-domain specialist, and the reference model adds a KL penalty toward itself (Section 5).

2 Related Work

Teacher allocation in distillation. GKD trains language models from teacher feedback on student-generated responses and also studies mixtures with offline data (Agarwal et al., 2024). MiniLLM optimizes sequence-level reverse KL with policy gradients and a mixture of teacher and student sampling (Gu et al., 2024). Both methods use one teacher and therefore leave teacher allocation outside their scope. Earlier multi-teacher knowledge distillation addresses this choice in supervised settings. RL-KD learns instance-level teacher selection, CA-MKD weights teachers by confidence on each labeled example, and MMKD meta-learns weights that depend on the student and teachers (Yuan et al., 2021; Zhang et al., 2022; 2023). WOPD instead computes one weight vector per domain before distillation and keeps it fixed on student-generated responses.

Recent OPD methods allocate teacher influence at several granularities. MOPD routes each prompt to its in-domain specialist (Ma et al., 2026). LS-MOPD ranks teachers on language-specific validation data and assigns preset weights to the highest-ranked teachers (Xie et al., 2026). DeepSeek-V4 describes a weighted sum of expert-specific reverse-KL losses, with weights reflecting expert importance and task context (DeepSeek-AI et al., 2026). H-OPD uses token-level entropy to form an arithmetic mixture of heterogeneous teacher distributions (Yin et al., 2026). WOPD uses fixed domain-level weights derived from verifier success in a sum of teacher-specific reverse-KL terms. FiRe-OPD and SEAD make a different allocation decision: they choose which responses, tokens, or prompts receive supervision (Li et al., 2026; Lee et al., 2026). REGEN avoids teacher inference during generalist training by learning offline from replay collected during specialist RL (Chen et al., 2026c).

Reference behavior during post-training. Learning without Forgetting trains on new-task inputs while matching the original model's old-task outputs, without requiring old-task data (Li & Hoiem, 2018). In language-model adaptation, KL-regularized policy optimization discourages the policy from drifting from its initialization (Korbak et al., 2022), while BAPO anchors the likelihood of responses generated by the base model during personalized preference optimization (Lee et al., 2024).

CaMOPD places the original general teacher and the specialized teacher in separate branches. The general teacher supervises recovery on proxy general prompts, whereas the specialist supervises preservation on domain prompts; teacher–student gaps select samples within each branch (Chen et al., 2026b). WOPD queries the reference model and all specialists on the same student-generated, in-domain states. Recent studies report less forgetting with on-policy data or reinforcement fine-tuning than with supervised adaptation (Chen et al., 2026a; Lai et al., 2025). WOPD holds the on-policy procedure fixed and measures the added effect of the reference model with an ablation that removes only this teacher.

3 METHOD

WOPD distills domain specialists and the reference model into a single student in two stages. Before distillation, we evaluate the teachers on shared training prompts and compute fixed teacher weights for each domain. During distillation, the student generates responses, and all teachers compute log-probabilities of the sampled tokens at the same states. The prompt domain determines which weights combine their supervision. Figure 2 summarizes these stages. Inference uses only the trained student.

3.1 SPECIALIST TEACHERS AND CAPABILITY SCORES

We consider D training domains. Each domain d has a prompt distribution $\mathcal{D}_d$ and a binary verifier $\mathrm{ver}_d(x, y) \in \{0, 1\}$, which determines whether response y succeeds on prompt x. Starting from the reference model π_0, we train one specialist $\pi_{\mathrm{spec}(d)}$ per domain using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and its verifier, without a KL penalty to the reference model. The teacher pool contains these specialists and the reference: $\mathcal{T} = \{0\} \cup \{\mathrm{spec}(d) : 1 \leq d \leq D\}$. We denote the specialists by $\mathcal{T}^- = \mathcal{T} \setminus \{0\}$. All teachers remain frozen during distillation, and the student π_θ is initialized from π_0. All models share a finite token vocabulary $\mathcal{V}$.

We evaluate every teacher on the same N_d training prompts in each domain d; we call this evaluation the capability exam. For prompt $x_i^{(d)}$, teacher j generates one response $y_{ij} \sim \pi_j(\cdot \mid x_i^{(d)})$. Its capability score is the mean verifier success:

$$S_{dj} = \frac{1}{N_d} \sum_{i=1}^{N_d} \mathrm{ver}_d\Big(x_i^{(d)}, y_{ij}\Big). \tag{1}$$

The resulting matrix $S \in [0, 1]^{D \times |\mathcal{T}|}$ records each teacher's performance in every domain. Specialist scores determine the relative specialist weights; the reference score is reported for comparison.

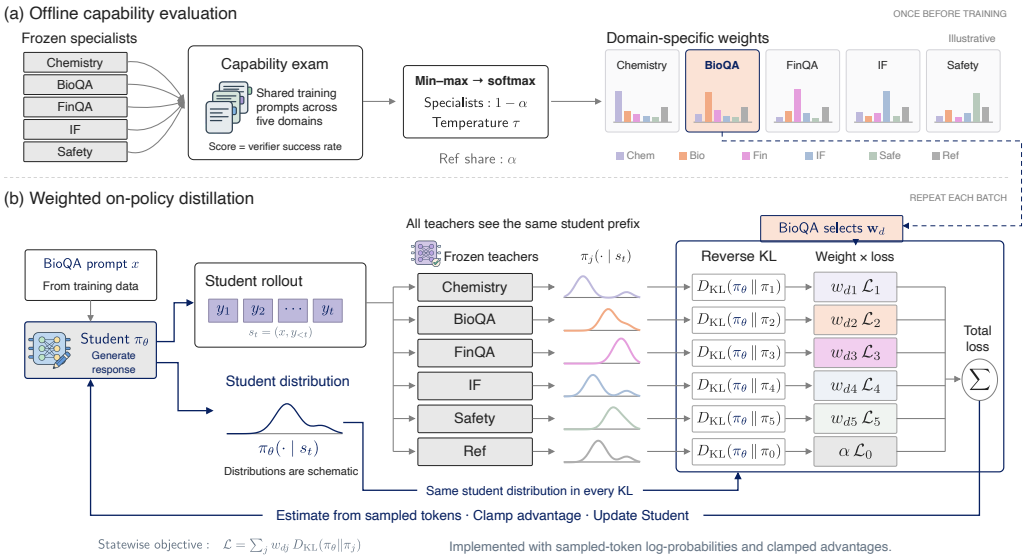

Figure 2: **Overview of weighted on-policy distillation (WOPD). (a)** Frozen specialists answer shared training prompts before distillation. Min–max normalization and softmax of verifier success assign specialist weights with total mass $1 - \alpha$; the reference receives α. **(b)** All six frozen teachers compute log-probabilities of the student's sampled response at the same states, using the prompt domain's weights. The diagram shows the weighted reverse-KL objective at one state. Bars, the BioQA example, and token distributions are schematic.

3.2 Domain-Specific Teacher Weights

We assign greater weight to specialists that score higher in the prompt's domain. Specialist score ranges differ across domains, so applying a shared softmax temperature to raw scores would make weight concentration depend on those ranges. We first normalize the scores within each domain, using only the specialists:

$$R_d = \max_{k \in \mathcal{T}^-} S_{dk} - \min_{k \in \mathcal{T}^-} S_{dk}, \qquad \widehat{S}_{dj} = \frac{S_{dj} - \min_{k \in \mathcal{T}^-} S_{dk}}{R_d}, \quad j \in \mathcal{T}^-. \tag{2}$$

All measured domains have $R_d > 0$. This maps the lowest and highest specialist scores in each domain to 0 and 1.

We then apply softmax to obtain the relative specialist weights:

$$p_{dj} = \frac{\exp(\widehat{S}_{dj}/\tau)}{\sum_{k \in \mathcal{T}^-} \exp(\widehat{S}_{dk}/\tau)}, \qquad j \in \mathcal{T}^-. \tag{3}$$

Smaller temperatures $\tau > 0$ concentrate weight on higher-scoring specialists, while larger temperatures distribute it more evenly. Within each domain, the gaps between normalized scores determine the relative weights.

We retain the reference model π_0 as a teacher and assign it a fixed share α, independently of its capability score. The remaining weight is distributed among specialists:

$$w_{d0} = \alpha, \qquad w_{dj} = (1 - \alpha) p_{dj}, \quad j \in \mathcal{T}^-. \tag{4}$$

Thus, τ controls the relative influence of specialists, while α controls the reference contribution. The primary experiments use $\tau = 0.3$ and $\alpha = 0.1$ in every domain.

3.3 Weighted On-Policy Distillation

For each training batch, we sample prompts from the fixed domain mixture and generate responses with the current student. At step t of a response, all teachers condition on the same state $s_t =$

$(x, y_{<t})$: the prompt followed by the tokens the student has generated so far. We use the weight vector $w_{d\cdot}$ associated with the prompt's domain at every token in that response.

The weights define a sum of teacher-specific reverse-KL losses:

$$\mathcal{K}(\theta; w, \mu) = \mathbb{E}_{(d,s)\sim\mu}\left[\sum_{j\in\mathcal{T}} w_{dj} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s) \| \pi_j(\cdot \mid s))\right]. \tag{5}$$

Here, μ is the distribution of states visited by the student, excluding padding. We hold μ, the teachers, and their weights fixed when differentiating this objective.

For the actual update, we query teacher log probabilities only for sampled student tokens. Let θ_{old} denote the student parameters used to generate the batch. We form the immediate token advantage

$$A_t = \sum_{j\in\mathcal{T}} w_{dj} \log \pi_j(y_t \mid s_t) - \log \pi_{\theta_{\mathrm{old}}}(y_t \mid s_t), \qquad y_t \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid s_t). \tag{6}$$

The implementation clamps each advantage, $\widetilde{A}_t = \mathcal{C}(A_t)$, and treats it as fixed during optimization. The token-level gradient contribution at the update point is $-\widetilde{A}_t \nabla_\theta \log \pi_\theta(y_t \mid s_t)$. All matched runs use the same clamp. Before clamping, the expected gradient contribution equals the gradient of $\mathcal{K}$ at $\theta = \theta_{\mathrm{old}}$ for fixed μ, as derived in Appendix B.1.

We take one gradient step per batch of generated responses, using the full batch, so the policy that generated the responses is the current policy and the importance ratio is one at the update point.

The implemented update thus differs from Equation 5 in two ways: it clamps the advantage, and it does not differentiate through the distribution of visited states. The second makes it a token-level surrogate for sequence-level reverse KL, whose gradient also includes future-token terms (Fu et al., 2026). Section 5 analyzes the objective without these two differences.

4 Experiments and Results

4.1 Experimental Setup

Models and training data. The student, reference model, and all five GRPO specialists use `Qwen/Qwen3-4B` (Yang et al., 2025). The training domains are Chemistry (Zhang et al., 2024a), BioQA (Ziegler et al., 2024), FinQA (Chen et al., 2021), instruction following (IF) (Argilla, 2024), and Safety (Jiang et al., 2024). The fixed mixture contains 30,360 prompts: 5,000 Chemistry, 5,120 each for BioQA, FinQA, and Safety, and 10,000 IF. The capability exam uses all of these prompts, with one response per teacher per prompt, for 182,160 generations.

Teacher capabilities. Each specialist has the highest capability score in its own domain, but other specialists also solve prompts from that domain. On Safety, for example, the FinQA specialist scores 77.46%, compared with 91.21% for the Safety specialist. WOPD uses these differences to assign domain-specific specialist weights. Appendix C gives the full matrix.

Baselines and training budget. Uniform and WOPD share six frozen teachers, training data, optimizer, and a three-epoch budget; only the teacher weights change. WOPD uses a reference share of 0.1, compared with Uniform's $1/6$, and redistributes the remaining weight among specialists. MOPD also trains for three epochs, but queries only the own-domain specialist rather than all six teachers. Equal epochs therefore do not imply equal teacher-inference cost. mixedGRPO trains directly on the domain mixture without teachers or a KL-to-reference penalty. We also report the reference model and the own-domain specialists for comparison with the distilled student. All models use thinking mode and a maximum response length of 32,768 tokens. Appendix E gives the training settings.

Evaluation. We measure in-domain performance on held-out sets from each of the five training domains. The out-of-domain group consists of Math (500 examples, NVIDIA, 2025), Code (769, Open-R1, 2025), GPQA Diamond (198, Rein et al., 2023), MLogiQA (800, Zhang et al., 2024b),

Table 1: **Five-domain main comparison**. Multi-domain pass@1 (%). Ref. is the reference model. Uniform shares WOPD's six-teacher pool. Its reference share is $1/6$, compared with WOPD's 0.1; relative specialist weights also differ. Specialist is the own-domain GRPO model and does not apply to out-of-domain rows. Bold and underlined values indicate the best and second-best results in each row, respectively, excluding the Specialist column. The GRPO teacher matrix is in Appendix H.

	Dataset	Ref.	mixedGRPO	MOPD	Uniform	WOPD	Specialist
In-domain	IF	90.90	89.50	<u>96.45</u>	92.35	**97.20**	97.80
	Safety	70.85	80.30	<u>84.90</u>	75.20	**86.00**	86.50
	BioQA	70.35	72.35	**74.10**	71.70	<u>73.30</u>	74.50
	FinQA	39.95	<u>71.90</u>	69.70	58.20	**73.20**	67.80
	Chem	29.30	42.05	<u>47.25</u>	32.60	**48.40**	48.30
Out-of-domain	Math	68.80	62.20	61.20	**70.60**	<u>69.00</u>	-
	Code	31.08	27.70	29.78	<u>33.94</u>	**35.20**	-
	GPQA	53.54	52.53	<u>56.06</u>	55.56	**59.10**	-
	MLogiQA	63.50	**66.75**	64.88	<u>66.13</u>	66.00	-
	IFEval	82.62	82.62	<u>84.29</u>	81.15	**86.10**	-

and IFEval (541, Zhou et al., 2023). Math and Code use randomly sampled subsets; the other three benchmarks are used in full. No evaluation prompts from either group are used for training or weighting. The out-of-domain group extends evaluation beyond the five domain-specific test sets, but some capabilities overlap: IFEval also tests instruction following, using a different dataset from IF. Tables report pass@1 percentages at decoding temperature 0.6. We report one training run per setting; the tables do not estimate variation across training seeds.

4.2 In-Domain and Out-of-Domain Performance

In-domain performance. WOPD scores higher than MOPD on four of the five in-domain sets (Table 1). The largest gain is on FinQA, where WOPD scores 73.20 versus MOPD's 69.70. BioQA is the exception, with 73.30 for WOPD and 74.10 for MOPD. WOPD also exceeds the own-domain specialist on FinQA and Chemistry and trails by at most 1.20 points on the other three sets.

Compared with Uniform, WOPD scores higher on all five in-domain sets (Figure 1(a)). The largest differences are on Chemistry and FinQA, at 15.80 and 15.00 points, where Uniform is far below the own-domain specialist. Uniform also scores below MOPD on all five in-domain sets.

Out-of-domain performance. WOPD matches the reference model on Math and exceeds it on the other four benchmarks (Figure 1(b)). It also scores higher than MOPD on all five. On Math, MOPD scores 61.20, below the reference's 68.80, while WOPD scores 69.00. On Code, WOPD scores 35.20, compared with 29.78 for MOPD and 31.08 for the reference. WOPD thus avoids the declines that MOPD shows on Math and Code.

Uniform scores higher than WOPD on Math (70.60 versus 69.00) and MLogiQA (66.13 versus 66.00), while WOPD is higher on Code, GPQA, and IFEval.

4.3 Reference-Model Ablation

To examine the reference model's contribution as a teacher, we compare two WOPD runs with the same five specialists, relative specialist weights p_d., and training budget. The full model assigns 0.1 to the reference and scales the specialist weights by 0.9; the no-reference model assigns all weight to the specialists.

Including the reference raises all five out-of-domain scores by 3.00–6.60 points (Table 2), with the largest difference on GPQA: 59.10 with the reference versus 52.50 without it. In-domain scores change less. Four are higher with the reference by 1.10–1.80 points, while FinQA is 0.30 lower. The reference thus improves out-of-domain performance with little change in in-domain performance. Section 5.3 explains the reference model's contribution to the training objective.

Table 2: **Reference-model ablation**. Pass@1 (%). The two WOPD columns use the same training budget and relative specialist weights; the reference share changes from 0.1 to zero.

	Dataset	WOPD (full pool)	WOPD (no reference)
In-domain	IF	97.20	96.00
	Safety	86.00	84.50
	BioQA	73.30	72.20
	FinQA	73.20	73.50
	Chem	48.40	46.60
Out-of-domain	Math	69.00	65.20
	Code	35.20	31.90
	GPQA	59.10	52.50
	MLogiQA	66.00	63.00
	IFEval	86.10	83.00

Table 3: **Single-domain in- and out-of-domain performance** (pass@1, %). Each block trains on one domain from the reference model. GRPO optimizes verifier reward; OPD distills that domain's specialist; WOPD distills it together with the reference model. Bold marks the best score in each block and column.

		Own domain	Out-of-domain		
Train domain	Method		Math	GPQA	MLogiQA
BioQA	GRPO	**74.50**	71.00	52.53	61.62
	OPD	72.30	69.40	52.53	**64.88**
	WOPD	74.45	**72.00**	**55.56**	63.62
FinQA	GRPO	67.80	59.80	38.38	47.13
	OPD	64.65	50.80	28.28	35.00
	WOPD	**72.05**	**61.20**	**42.42**	**63.13**
IF	GRPO	**97.80**	**68.60**	53.54	61.75
	OPD	97.50	64.40	52.02	61.00
	WOPD	96.55	66.80	**56.57**	**63.25**
Safety	GRPO	86.50	68.60	54.55	**66.38**
	OPD	86.00	**70.40**	52.53	65.00
	WOPD	**87.35**	69.40	**55.05**	63.75
Chem	GRPO	**48.30**	69.20	**51.52**	66.50
	OPD	43.15	33.15	33.03	56.06
	WOPD	47.35	**72.40**	51.01	**66.88**

4.4 Single-Domain In- and Out-of-Domain Performance

Table 3 reports five single-domain comparisons. Each WOPD run uses only the own-domain specialist and the reference model, with weights 0.9 and 0.1. The paired GRPO and OPD scores are those used in this comparison. The teacher-pool evaluations are reported separately in Appendix H.

Relative to OPD, WOPD scores higher in four of five own-domain columns, all five out-of-domain means, and 12 of 15 out-of-domain cells. Two OPD models also have large out-of-domain regressions against their paired GRPO baselines. The Chemistry run scores 33.15 on Math, 36.1 points below its paired GRPO run, and 18.5 and 10.4 points below it on GPQA and MLogiQA. The FinQA run is 9.0, 10.1, and 12.1 points below its paired GRPO run on the same three sets.

Weight concentration. For each of Chemistry and FinQA, five students differ only in $\tau \in \{0.05, 0.1, 0.2, 0.4\}$ or equal weights, with learning rate 10^{-5} and batch size 300 (Table 4). The pool contains all five specialists and no reference model.

At $\tau = 0.1$ and $\tau = 0.05$, the own-domain specialist receives more than 0.99 of the weight, and the in-domain score rises in both domains: Chemistry from 32.75 under equal weights to 44.35 and 48.00, and FinQA from 66.95 to 74.90 and 72.10. Chemistry's out-of-domain mean stays between

Table 4: **Weight concentration with five specialists** (pass@1, %). The pool excludes the reference model; only τ changes. Out-of-domain mean averages the five out-of-domain benchmarks. Bold marks the largest own-domain score in each block.

		Equal	τ=0.4	τ=0.2	τ=0.1	τ=0.05
Chemistry	Own-specialist weight	0.2000	0.7136	0.9606	0.9995	> 0.9999
	Chemistry	32.75	35.75	38.50	44.35	**48.00**
	Math	72.20	71.00	71.40	70.20	73.80
	IF	91.15	91.50	91.10	91.00	92.20
	Out-of-domain mean	62.20	61.90	62.00	61.42	61.88
FinQA	Own-specialist weight	0.2000	0.6295	0.8911	0.9907	0.9999
	FinQA	66.95	69.10	74.45	**74.90**	72.10
	IF	91.50	90.05	88.60	81.60	77.60
	IFEval	80.22	76.16	77.08	73.75	69.87
	Out-of-domain mean	60.38	59.09	58.41	57.52	54.27

61.42 and 62.20 in all settings. In FinQA, the out-of-domain mean falls from 60.38 to 57.52 and 54.27, and IF falls from 91.50 to 81.60 and 77.60. Appendix G varies the reference share in single-domain runs.

5 What the Weights Change

We analyze the distillation objective to understand which token distribution the student is encouraged to match and how the teacher weights change that distribution. Throughout this section, we fix a domain and a state s. We assume a finite vocabulary with positive teacher probabilities and consider the unclamped token-level objective.

5.1 The Effective Teacher Distribution

Write $\pi_j(a) = \pi_j(a \mid s)$. Let q be a candidate next-token distribution, and let $v_j \geq 0$ be teacher weights that sum to one. Expanding the weighted KL loss gives

$$\sum_j v_j D_{\mathrm{KL}}(q\|\pi_j) = D_{\mathrm{KL}}(q\|q_v^\star) - \log Z^v, \qquad q_v^\star(a) = \frac{\prod_j \pi_j(a)^{v_j}}{Z^v}, \tag{7}$$

where Z^v normalizes the product. The effective teacher is therefore a normalized geometric mean of the teacher distributions, known as a logarithmic opinion pool (Heskes, 1997).

Since Z^v does not depend on q, the distribution $q_v^\star$ uniquely minimizes the loss over all next-token distributions at this state. With the distribution of visited states held fixed, the expected weighted loss likewise equals the student's expected KL to these effective teachers plus a constant.

5.2 How Teacher Weights Change Token Preferences

The normalizer Z^v cancels in probability ratios, so Equation 7 gives, for any two tokens a and b,

$$\log \frac{q_v^\star(a)}{q_v^\star(b)} = \sum_j v_j \log \frac{\pi_j(a)}{\pi_j(b)}. \tag{8}$$

The effective teacher's log-odds are a weighted average of the teachers' log-odds. A teacher's weight is the share of its token preference that reaches the effective teacher.

This gives one reason why Uniform trails WOPD on the in-domain sets. Uniform gives the own-domain specialist a weight of $1/6$; WOPD gives it 0.65–0.76 (Appendix C). Suppose the specialist strongly prefers token a over b and the other teachers are indifferent. Uniform's effective teacher then keeps one sixth of the specialist's log-odds, and WOPD's keeps about four times as much. The in-domain gap is 10.80–15.80 points in Chemistry, FinQA, and Safety (Table 1), where the specialist's capability score exceeds the mean of the other five teachers by 20.30–30.17 points (Table 5). In IF

and BioQA, that lead is 7.19–12.46 points and the gap is 1.60–4.85 points. Section 4.4 tests this by varying only the specialist weights.

The capability scores provide a response-level criterion for assigning teacher weights. The resulting token preferences depend on the probabilities that the teachers assign at a shared state. Appendix A.3 characterizes conditions under which these changes improve the effective teacher's value under a fixed continuation rule.

5.3 The Reference Model Adds a KL Penalty

The reference-removal ablation holds the relative specialist weights p_j fixed and changes the reference share. Let $\mathcal{L}^-$ denote the specialist-only loss and $\mathcal{L}^+$ the loss with reference share $\alpha \in [0, 1)$. Then

$$\begin{aligned}\mathcal{L}^+(q) &= (1-\alpha)\mathcal{L}^-(q) + \alpha D_{\mathrm{KL}}(q\|\pi_0) \\ &= (1-\alpha)\left[\mathcal{L}^-(q) + \frac{\alpha}{1-\alpha} D_{\mathrm{KL}}(q\|\pi_0)\right].\end{aligned} \tag{9}$$

Thus, reference supervision adds an explicit KL penalty toward the reference model on the same states. At $\alpha = 0.1$, its coefficient relative to the specialist loss is $1/9$. The overall factor 0.9 preserves the loss minimizers but can change finite-step optimization.

Reference inclusion also changes the effective teacher from $q^\star_-$ to

$$q^\star_+(a) \propto \pi_0(a)^\alpha q^\star_-(a)^{1-\alpha}. \tag{10}$$

Consequently,

$$\log \frac{q^\star_+(a)}{q^\star_+(b)} - \log \frac{\pi_0(a)}{\pi_0(b)} = (1-\alpha)\left[\log \frac{q^\star_-(a)}{q^\star_-(b)} - \log \frac{\pi_0(a)}{\pi_0(b)}\right]. \tag{11}$$

Every log-odds difference between the effective teacher and the reference model shrinks by a factor of 0.9 in the primary configuration.

With the specialist weights held fixed, increasing the reference share also brings the effective teacher closer to the reference model in both KL directions. To see this, consider the path $q_\eta(a) \propto \pi_0(a)^{1-\eta} q^\star_-(a)^\eta$, for $\eta \in [0, 1]$. Its endpoints are the reference model and the specialist-only effective teacher, while reference share α selects $\eta = 1 - \alpha$. Define $h(a) = \log(q^\star_-(a)/\pi_0(a))$ and $\psi(\eta) = \log \mathbb{E}_{a\sim\pi_0}[e^{\eta h(a)}]$. Differentiating the normalizer gives $\psi''(\eta) = \mathrm{Var}_{a\sim q_\eta}(h(a)) \geq 0$, and hence

$$\begin{aligned}\frac{d}{d\eta} D_{\mathrm{KL}}(q_\eta\|\pi_0) &= \eta\psi''(\eta) \geq 0, \\ \frac{d}{d\eta} D_{\mathrm{KL}}(\pi_0\|q_\eta) &= \psi'(\eta) - \psi'(0) \geq 0.\end{aligned} \tag{12}$$

Both divergences are therefore nonincreasing as the reference share grows. For any positive reference share, the decrease relative to the specialist-only effective teacher is strict unless it already equals the reference model. Appendix B gives the full proof.

Reference supervision therefore constrains the effective teacher distribution at the states visited during training. The matched reference-removal comparison in Section 4.3 measures the corresponding change in student performance.

6 Conclusion

WOPD combines domain-specific specialist weights with reference supervision on the same student-generated responses. On the five-domain Qwen3-4B study, it improves every in-domain score over equal weighting. It matches the reference model on Math and exceeds it on the other four out-of-domain benchmarks. The matched ablation shows that including the reference model as a teacher raises every out-of-domain score by 3.00–6.60 points and changes each in-domain score by at most 1.80 points.

WOPD separates specialist allocation from reference supervision: a shared training-domain exam determines the relative specialist weights, while the reference model supplies supervision for maintaining existing capabilities without a separate stream of preservation prompts.

AI USE STATEMENT

AI tools assisted with literature search and reference verification, manuscript organization and editing, consistency checks of reported results and mathematical exposition, and preparation of explanatory figures. AI tools were not used to run experiments or produce experimental results. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, and figures produced with the aid of generative AI.

REFERENCES

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

Argilla. ifeval-like-data. https://huggingface.co/datasets/argilla/ifeval-like-data, 2024. Synthetic instruction-following corpus, config `filtered`.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. In *International Conference on Machine Learning*, 2026a. URL https://arxiv.org/abs/2510.18874.

Tianlei 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 preprint arXiv:2605.27115*, 2026b. URL https://arxiv.org/abs/2605.27115.

Yunjie Chen, Zihao Chen, Xiaoxin Chen, and Fang Wang. REGEN: Replay-recycling for expert-to-generalist distillation with offline reinforcement learning. *arXiv preprint arXiv:2607.19450*, 2026c. URL https://arxiv.org/abs/2607.19450.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

DeepSeek-AI, Anyi Xu, Bangcai Lin, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. URL https://arxiv.org/abs/2606.19348.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026. URL https://arxiv.org/abs/2603.25562.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/html/2306.08543v3.

Tom Heskes. Selecting weighting factors in logarithmic opinion pools. In *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. WildTeaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *arXiv preprint arXiv:2406.18510*, 2024. Introduces the WildJailbreak safety dataset.

Tomasz Korbak, Ethan Perez, and Christopher L Buckley. RL with KL penalties is better viewed as Bayesian inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022. URL https://arxiv.org/abs/2205.11275.

Song Lai, Haohan Zhao, Rong Feng, Changyi Ma, Wenzhuo Liu, Hongbo Zhao, Xi Lin, Dong Yi, Qingfu Zhang, Hongbin Liu, Gaofeng Meng, and Fei Zhu. Reinforcement fine-tuning naturally mitigates forgetting in continual post-training. *arXiv preprint arXiv:2507.05386*, 2025. URL https://arxiv.org/abs/2507.05386.

Chia-Hsuan Lee, Zelei Cheng, Yu Wang, Renkun Ni, Sambit Sahu, Shi-Xiong Zhang, and William Campbell. SEAD: Competence-aware on-policy distillation via entropy-guided supervision. *arXiv preprint arXiv:2606.28562*, 2026. URL https://arxiv.org/abs/2606.28562.

Gihun Lee, Minchan Jeong, Yujin Kim, Hojung Jung, Jaehoon Oh, Sangmook Kim, and Se-Young Yun. BAPO: Base-anchored preference optimization for overcoming forgetting in large language models personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. URL https://arxiv.org/abs/2407.00693.

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

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. doi: 10.1109/TPAMI.2017.2773081.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021. URL https://arxiv.org/abs/2105.03023.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. Tuning language models by proxy. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2401.08565.

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

NVIDIA. AceReason-Math. https://huggingface.co/datasets/nvidia/AceReason-Math, 2025. Revision a5cc41c5; source of the Math out-of-domain subset.

Open-R1. CodeForces. https://huggingface.co/datasets/open-r1/codeforces, 2025. Revision fbe3f6e9; source of the Code out-of-domain subset.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

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

Yuan Xie, Jiaqi Song, Xianliang Wang, Ming Lei, Jie Gao, and Jie Wu. Language-specialized multi-teacher on-policy distillation for multilingual LLM-based ASR. *arXiv preprint arXiv:2608.03610*, 2026. URL https://arxiv.org/abs/2608.03610.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qixiang Yin, Huanjin Yao, Yuchen Cai, Jianghao Chen, Ziyi Wang, Min Yang, Fei Su, and Zhicheng Zhao. H-OPD: Confidence aware heterogeneous multi-teacher multimodal on-policy distillation. *arXiv preprint arXiv:2607.02592*, 2026. URL https://arxiv.org/abs/2607.02592.

Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. Reinforced multi-teacher selection for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. URL https://arxiv.org/abs/2012.06048.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. ChemLLM: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024a. Instruction-tuning corpus distributed as AI4Chem/ChemData700K.

Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022. URL https://arxiv.org/abs/2201.00007.

Hailin Zhang, Defang Chen, and Can Wang. Adaptive multi-teacher knowledge distillation with meta-learning. In *IEEE International Conference on Multimedia and Expo*, 2023. URL https://arxiv.org/abs/2306.06634.

Yidan Zhang, Yu Wan, Boyi Deng, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, Fei Huang, and Jingren Zhou. P-MMEval: A parallel multilingual multitask benchmark for consistent evaluation of LLMs. *arXiv preprint arXiv:2411.09116*, 2024b. Source of the MLogiQA split used here.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Ingo Ziegler, Abdullatif Köksal, Desmond Elliott, and Hinrich Schütze. CRAFT your dataset: Task-specific synthetic dataset generation through corpus retrieval and augmentation. *arXiv preprint arXiv:2409.02098*, 2024. BioQA split distributed as ingoziegler/CRAFT-BioQA.