LEARN FROM WHOEVER IS RIGHT: PER-SAMPLE ANSWER VERIFICATION FOR MULTI-TEACHER ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern large language models (LLMs) rely on reinforcement learning to build strong capabilities in individual domains, but integrating those capabilities into a single deployable model remains challenging. By routing each sample to the teacher whose domain matches it, existing approaches let a domain label decide which teacher provides supervision. However, domain expertise holds only on average: the matched teacher is not always correct on a given sample, while a teacher from another domain sometimes is. The reliable teacher therefore has to be identified per sample, not per domain. In this paper, we introduce **M**ulti-**T**eacher **S**elf-**D**istillation **P**olicy **O**ptimization (MT-SDPO), an on-policy distillation method that unifies several frozen teachers into one student model. MT-SDPO consists of three components: (1) *self-anchors*, where a rollout is supervised by a correct rollout from its own group; (2) *answer-verified eligibility*, where a teacher may supervise a sample only if its own answer passes a verifier; and (3) *privileged distillation*, which merges the anchor and the feedback of all eligible teachers into one context that an exponential moving average self-teacher reads and the student does not, thereby keeping one policy at deployment. On three scientific domains with verifiable answers, MT-SDPO outperforms domain routing in worst-domain accuracy for all five students from three model families. On Qwen3-8B, it lifts worst-domain accuracy by 14.79 points and narrows its domain gap by 74.7%, a better balance than serving one matched teacher per domain. Verified reliability, not domain membership, should decide who teaches.

1 INTRODUCTION

> *"When three walk together, there is always a teacher among them: I choose what is good and follow it; what is not good, I correct."*
>
> — *Confucius,* The Analects

Post-training a modern large language model (LLM) is organized by domain: verifiable-answer reinforcement learning (RL) for mathematical reasoning (Shao et al., 2024; Guo et al., 2025), software-evolution rewards (Wei et al., 2025) and executable SWE environments (Jain et al., 2025), human-preference rewards for instruction following (Ouyang et al., 2022), and interactive environments for search agents (Nakano et al., 2021; Jin et al., 2025). Each recipe needs its own data and rollout protocol, so teachers are trained per domain. A deployed system, however, answers every request with one policy: a teacher pool adds overhead and requests cross domain boundaries. Multi-domain post-training therefore ends by asking how to fold the teachers back into one deployable model.

Among the routes to this problem, we study on-policy integration, which keeps the teachers frozen and lets one student learn on its own rollouts, so what remains open is which teacher supervises which rollout. Multi-Teacher On-Policy Distillation (MOPD) (Ma et al., 2026) develops domain teachers independently, then routes each rollout to the teacher whose domain matches the prompt, assuming that teacher is right for every sample in its domain.

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.

Domain expertise, however, is an average-case property: domain boundaries overlap, teachers trained from a shared source repeat the same errors, and a teacher from another domain sometimes succeeds anyway. Figure 1 quantifies this gap for the Qwen3-8B teachers: the matched teacher is correct on only 52.9% of training samples, while per-sample selection raises the share with a correct teacher to 65.7%. The reliable teacher therefore varies per sample, not per domain: **whom should the student trust?**

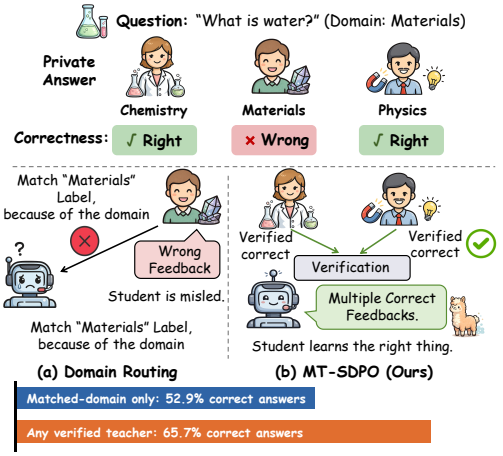


Figure 1: Sample-level teacher reliability gap on the Qwen3-8B teacher pool. A domain label can select a teacher wrong on the sample; MT-SDPO verifies private answers and uses all eligible teachers.

We answer with an allocation rule that drops the domain label: on each sample, any source whose answer is verified correct is a legitimate teacher, and no unverified source is. The student is one such source: its own correct rollout is the cheapest verified teacher, so frozen teachers are consulted only where it fails. Supervision flows exactly where someone is known to be right.

We instantiate the rule as **Multi-Teacher Self-Distillation Policy Optimization (MT-SDPO)** on top of Self-Distillation Policy Optimization (SDPO) (Hübotter et al., 2026), which turns textual feedback into a dense token-level signal through a privileged self-teacher. Three technical questions remain, and one component answers each: anchoring where the student is already right, coverage where it is not and the anchor is missing, and merging where anchor and feedback must become one signal. Figure 2 traces one training step through the three components. Self-anchors let a correct rollout supervise the incorrect peers of its own group. Answer-verified eligibility admits a teacher's feedback only when its private answer passes a verifier that no model sees, and takes every teacher that qualifies. Privileged distillation merges the anchor and the feedback of all eligible teachers into one context that only an exponential moving average (EMA) self-teacher reads, and distills that self-teacher into the student token by token. Training updates only the student, which is all that deployment keeps.

Built on per-sample answer verification and privileged self-distillation, MT-SDPO not only redistributes capability toward the weakest domain, but can also outperform a three-model reference that serves one matched teacher per domain. We evaluate on the chemistry, materials science, and physics domains of the Science Q&A subset in SciKnowEval L3 (Feng et al., 2024) with five student models from three families, each under its own multi-domain initialization. MT-SDPO exceeds Domain Routing, which differs only in the allocation rule, in worst-domain accuracy for all five students. On Qwen3-8B (Yang et al., 2025), MT-SDPO gains 4.64 points in Macro accuracy and 14.79 in worst-domain accuracy, and cuts the domain gap from 20.96 to 5.30. The gains are not tied to one model: every Qwen3 scale becomes more balanced, though the aggregate improves only at 8B, and all three metrics improve together on OLMo-3-7B-Instruct (Team Olmo, 2025), a second family with a different teacher generator. Our contributions are as follows:

- **Sample-level diagnosis.** We quantify the reliability gap between domain-matched and per-sample teacher selection, separating matched coverage, cross-domain rescue, and failure across all teachers.
- **MT-SDPO.** Answer-verified multi-teacher self-distillation that grants eligibility per sample and merges all eligible teachers with the self-anchor into one privileged self-teacher.
- **Controlled evidence.** Same-initialization ablations isolating answer verification, per-sample selection, and multi-teacher aggregation.

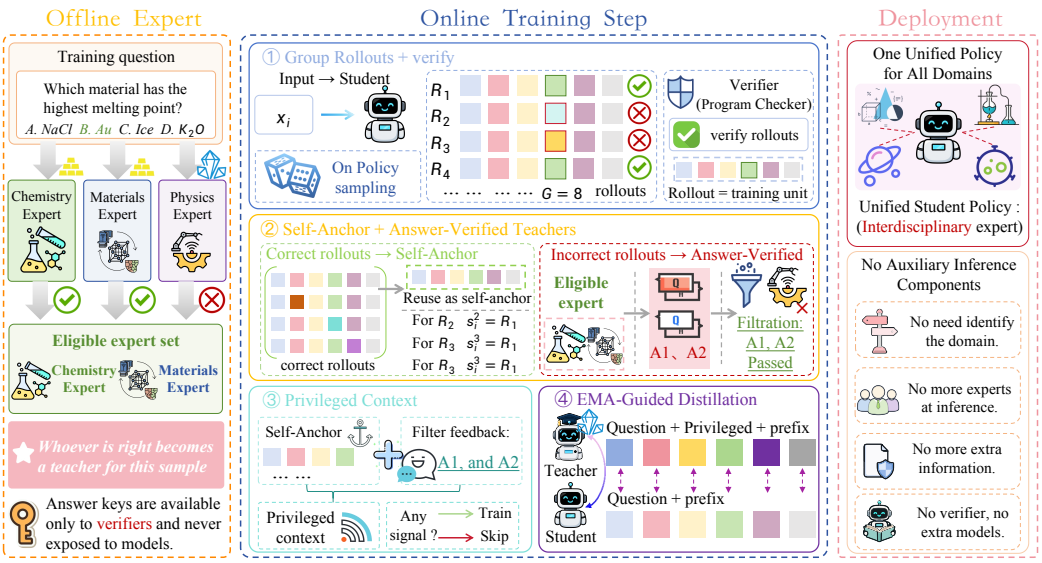


Figure 2: Overview of MT-SDPO. Offline answer verification builds a sample-level eligible teacher set. Online, self-anchors and sanitized teacher feedback condition an EMA self-teacher that distills into the student model, all that deployment retains.

2 Related Work

Turning separately trained domain teachers into one deployable policy raises three questions that organize prior work: how the teachers are consolidated into one model, how a student learns on-policy from a teacher, and how supervision is allocated across teachers on a shared sample.

(1) Capability Consolidation for Multi-Domain LLMs. Prior routes build or consolidate multi-domain capabilities by 1) joint or cascaded multi-domain RL (Yang et al., 2025; Wang et al., 2025), 2) off-policy finetuning on teacher completions (DeepSeek-AI, 2025), 3) weight-space merging (Wortsman et al., 2022; Ilharco et al., 2023), or 4) mixed-domain training with negative-transfer control (Yu et al., 2020; Cai et al., 2026; Yang et al., 2026b; Ye et al., 2025). Closest to us is MOPD (Ma et al., 2026), which routes the student's own rollouts to the domain-matched frozen teacher; we implement this rule as our *Domain Routing* baseline. It assumes the matched teacher is the right supervisor for every sample from its domain, whereas MT-SDPO keeps MOPD's two-stage split but selects the supervisor per sample by verification.

(2) On-Policy and Self-Distillation. Classical distillation matches a teacher on a fixed corpus and is off-policy, so the student can drift at inference (Hinton et al., 2015; Kim & Rush, 2016). Adaptive off-policy distillation instead reuses student-generated outputs (Ko et al., 2024), while on-policy distillation scores the student's current rollouts with a separate teacher (Gu et al., 2024; Agarwal et al., 2024), or with several teachers that debate over the student's state and are weighted by post-debate confidence (Wang et al., 2026). Self- and context-distillation let a model learn from signals produced by itself or by a context-privileged copy (Hübotter et al., 2026; Zhao et al., 2026; Kim et al., 2026; Song et al., 2026; Ye et al., 2026). Our baseline SDPO (Hübotter et al., 2026) anchors this self-teacher on the student's own correct rollout and, like group-relative RL under advantage collapse (He et al., 2026), gives no signal where that rollout is missing. MT-SDPO keeps the self-anchor but adds answer-verified external teachers, so those unanchored samples can still be supervised by whichever teacher is right.

(3) Multi-Teacher Distillation and Supervision Reliability. Distilling several teachers into one student (You et al., 2017) appears in recent LLMs as output mixing or per-sample routing (Xiaomi LLM-Core Team, 2026; Yang et al., 2026a), sometimes with adaptive per-instance teacher weights (Liu et al., 2020; Yuan et al., 2021). Learned routers instead pick one teacher per prompt,

trained on which teachers answer correctly (Jin et al., 2026) or on teacher response quality and student learnability (Zhang et al., 2026), and then distill off-policy from the selected rationales. Separately, verifiable-reward RL and generative verifiers turn correctness into a training signal that scores the student's own output (Lambert et al., 2024; Guo et al., 2025; Zhang et al., 2025); MT-SDPO draws on the same signal but spends it on deciding whom the student should trust. It differs from these allocation schemes in three ways: 1) a teacher may supervise a sample only if its own private answer on that sample passes a verifier, so eligibility is observed rather than predicted by a domain label, learned weight, or router; 2) eligibility is decided per sample and keeps every teacher that qualifies, so a cross-domain teacher helps whenever it is right; 3) all eligible teachers combine with the student's self-anchor into a single on-policy self-teacher.

3 PRELIMINARIES

Self-Distillation Policy Optimization. Given a prompt $x \sim \mathcal{D}$, SDPO (Hübotter et al., 2026) samples an on-policy rollout $y = (y_1, \ldots, y_T) \sim \pi_\theta(\cdot \mid x)$ and re-evaluates it with a self-teacher that additionally observes privileged information c. Write $p_{\theta,t} = \pi_\theta(\cdot \mid x, y_{<t})$ for the student distribution and $q_{\bar{\theta},t} = \mathrm{sg}[\pi_{\bar{\theta}}(\cdot \mid x, c, y_{<t})]$ for the self-teacher distribution, where $\bar{\theta}$ is a slowly updated copy of θ. The rollout-level objective is

$$\mathcal{L}_{\mathrm{SDPO}}(x, y, c) = \frac{1}{T} \sum_{t=1}^{T} D\big(p_{\theta,t} \big\| q_{\bar{\theta},t}\big)\,, \tag{1}$$

where D is a token-level divergence and sg blocks self-teacher gradients. The context c can contain environment feedback or another verified successful rollout; it changes the self-teacher distribution without changing the rollout on which the student is trained.

Multi-Teacher Setting. Each training instance consists of a prompt x_i, a verifiable reference answer a_i^*, and a domain label d_i. We train one student model π_θ with feedback from K frozen domain teachers $\{E_k\}_{k=1}^K$, each queried either for a private answer, written E_k^{ans}, or for a critique of a given rollout, written E_k^{fb}. A programmatic verifier $\textsc{Verify}(\cdot, a_i^*) \in \{0, 1\}$ extracts the answer from a completion and compares it with a_i^*; the reference answer reaches this verifier and the sanitizer only, and never enters a student or teacher prompt at any stage of MT-SDPO. Our problem is to decide, for each student rollout, which teachers may contribute to c and then consolidate their supervision into the single policy π_θ that training retains.

4 METHOD

We propose MT-SDPO, an answer-verified multi-teacher distillation framework that consolidates K frozen domain teachers into one deployable policy. A single allocation rule governs it: a teacher may supervise a sample only if its own private answer is verified correct. MT-SDPO therefore pairs a supervision source the student produces itself with a *sample-level eligible teacher set* that replaces domain routing.

4.1 FRAMEWORK OVERVIEW

Verification runs once, before training. As illustrated in Figure 2, MT-SDPO trains a single student model π_θ with three auxiliary parts that all disappear at inference: K frozen domain teachers, a programmatic verifier, and an EMA copy $\pi_{\bar{\theta}}$ of the student that serves as the self-teacher. Every teacher first answers each question privately, and the verifier caches the correct ones as that sample's eligible set, so training pays no recurring answering cost.

Each online step aligns supervision to individual rollouts. The student samples a group of rollouts per prompt and the verifier labels them. A correct rollout supervises the remaining rollouts of its group as their self-anchor, whereas an incorrect one is sent to the eligible teachers, whose sanitized feedback joins the anchor in a privileged context that only the self-teacher reads. Distilling the self-teacher into the student is the only parameter update.

4.2 ROLLOUT-SPECIFIC SELF-ANCHORS

The first source of privileged information is the student itself (Hübotter et al., 2026): a rollout group usually mixes correct and incorrect rollouts, and a correct one is a solution the current policy has already shown it can produce.

A programmatic check splits each rollout group. For prompt x_i, the student samples G rollouts $\{y_i^1, \ldots, y_i^G\} \sim \pi_\theta(\cdot \mid x_i)$, and

$$\mathcal{S}_i = \left\{ j : \textsc{Verify}(y_i^j, a_i^*) = 1 \right\} \tag{2}$$

indexes the correct ones.

Each rollout takes its self-anchor from a correct peer. The self-anchor of rollout j is

$$s_i^j = \begin{cases} \mathrm{Trunc}\left(y_i^{j^\star}\right), & \mathcal{S}_i \setminus \{j\} \neq \varnothing, \\ \varnothing, & \text{otherwise}, \end{cases} \tag{3}$$

where $j^\star = \min(\mathcal{S}_i \setminus \{j\})$ is the earliest correct rollout of the same group other than j, and Trunc keeps that rollout's solution segment within the anchor token budget.

Excluding the rollout itself prevents sequence leakage. If j were its own anchor, the self-teacher would observe a complete copy of the sequence whose prefixes it scores, and the supervision would degenerate into copying. An incorrect rollout can still reuse a correct peer, whereas a prompt that the student never solves has no self-anchor at all.

4.3 ANSWER-VERIFIED TEACHER ELIGIBILITY

Domain match indicates which teacher is stronger on average, not which is correct on a given question (Ma et al., 2026). MT-SDPO replaces this proxy with a per-sample check, and spends the resulting teacher queries only on rollouts the verifier marked incorrect, which is also where the self-anchor may be missing.

A private answer decides eligibility before training starts. Teacher k independently produces a private completion $\hat{y}_{i,k} = E_k^{\mathrm{ans}}(x_i)$ from the question, and the verifier keeps the teachers that are correct:

$$\mathcal{V}_i = \{k : \textsc{Verify}(\hat{y}_{i,k}, a_i^*) = 1\}\,. \tag{4}$$

The eligible set is sample-specific and ignores the domain label d_i: a teacher from another domain enters $\mathcal{V}_i$ whenever it answers correctly, and the nominal-domain teacher drops out whenever it does not.

Every eligible teacher is queried independently. Each $k \in \mathcal{V}_i$ receives the question and the incorrect rollout, but neither a_i^* nor $\hat{y}_{i,k}$. Only the messages that survive sanitization reach the context, concatenated in a fixed teacher order:

$$F_i^j = \begin{cases} \mathrm{Concat}_{k \in \mathcal{V}_i} \left[\Gamma\left(E_k^{\mathrm{fb}}(x_i, y_i^j) \right) \right], & j \notin \mathcal{S}_i, \\ \varnothing, & \text{otherwise}, \end{cases} \tag{5}$$

where the sanitizer Γ screens a message against x_i and a_i^* and returns either the accepted feedback or $\varnothing$, so that an empty $\mathcal{V}_i$ also leaves F_i^j empty. Eligibility fixes who is asked rather than who contributes: $|\mathcal{V}_i|$ eligible teachers supply at most $|\mathcal{V}_i|$ messages. Unlike conventional multi-teacher aggregation (You et al., 2017), MT-SDPO applies neither a majority vote nor domain-based weighting over those messages.

The sanitizer bounds label leakage and topical drift. Γ rejects an entire message that reveals the correct label, whether as a structured answer pattern, a targeted label expression, or the full text of the correct answer. It also requires a short concept anchor occurring verbatim in the question, so that the critique addresses the specific error. Both gates are lexical: they bound leakage and drift, but neither catch every paraphrase of the label nor certify that the retained feedback is semantically valid.

Verification is an allocation rule, not a correctness proof. Equation (4) tests one cached completion under a fixed prompt and decoding configuration; it neither tests the critique that the teacher writes later nor establishes general competence on the question. We therefore speak of feedback from answer-verified teachers rather than verified feedback, and justify the rule by its downstream effect, which Section 5.5 measures against domain routing and against unverified aggregation.

4.4 Privileged Distillation and Student Update

MT-SDPO merges the two sources into one training signal: it conditions the EMA self-teacher on both and transfers the resulting token distribution to the student, which observes neither.

The privileged context combines the two sources under one mask. For each rollout, MT-SDPO forms the context $c_i^j = [s_i^j; F_i^j]$, which the self-teacher reads after the original question under natural-language section headers, together with the eligibility mask

$$m_i^j = \mathbb{1}\left[c_i^j \neq \varnothing\right] = \mathbb{1}\left[s_i^j \neq \varnothing \ \vee \ F_i^j \neq \varnothing\right]. \tag{6}$$

The context has four cases: full conditioning when both sources exist, a fallback to self-anchor SDPO (Hübotter et al., 2026) when only s_i^j exists, and teacher-only SDPO when only F_i^j exists. The mask collapses the fourth, in which both are empty because the group holds no other correct rollout and no message survived, so that such a rollout contributes no distillation loss instead of an uninformative one. An empty source deletes its own block rather than leaving a placeholder, so the context stays a single natural-language prompt throughout. Appendix E reproduces the template.

The objective is a masked, importance-weighted token-level divergence. At token t of rollout y_i^j, the student scores the sampled prefix from the question alone, $p_{\theta,i,t}^j = \pi_\theta(\cdot \mid x_i, y_{i,<t}^j)$, while the self-teacher scores the same prefix with the context attached, $q_{\bar{\theta},i,t}^j = \mathrm{sg}[\pi_{\bar{\theta}}(\cdot \mid x_i, c_i^j, y_{i,<t}^j)]$. The teacher branch carries the EMA parameters $\bar{\theta}$ rather than the current θ, and sg only keeps the update out of that branch. Teacher forcing on a shared prefix keeps the two branches token-aligned. Extending Eq. (1) from one rollout to a masked, importance-weighted batch objective and instantiating D as the symmetric Jensen–Shannon divergence gives

$$\mathcal{L}_{\text{MT-SDPO}} = \frac{1}{Z} \sum_{i,j,t} m_i^j M_{i,t}^j w_{i,t}^j \, D_{\text{JSD}}\left(p_{\theta,i,t}^j \,\middle\|\, q_{\bar{\theta},i,t}^j\right), \tag{7}$$

where $M_{i,t}^j$ masks response tokens, $w_{i,t}^j$ is a detached and clipped importance weight (Schulman et al., 2017) that corrects policy drift on the sampled token, and Z counts the unmasked tokens. We adopt the symmetric divergence and the EMA self-teacher because both stabilize on-policy distillation (Agarwal et al., 2024; Hübotter et al., 2026). Since $\pi_{\bar{\theta}}$ is both an EMA copy and the only branch that reads c_i^j, the divergence transfers a context-conditioned and temporally smoothed shift rather than the isolated contribution of c_i^j. Appendix B gives the truncated support of the divergence, the two ratios behind $w_{i,t}^j$, and the per-micro-batch normalization that replaces Z.

Only the student model is updated and deployed. Gradients reach θ alone: every E_k stays frozen, and the EMA copy runs without gradients and follows $\bar{\theta} \leftarrow (1-\tau)\bar{\theta} + \tau\theta$ with EMA rate $\tau = 0.05$ once per collected rollout batch. A step samples rollouts from π_θ, computes every target in Eq. (7) from one fixed $\bar{\theta}$, takes the mini-batch updates on θ, and refreshes $\bar{\theta}$ last. Although the trainer reuses Group Relative Policy Optimization (GRPO) (Shao et al., 2024) infrastructure to collect rollouts, no scalar advantage enters Eq. (7): the binary check only marks success, builds self-anchors, and gates teacher queries. Algorithm 1 in Appendix B states one full training step. Once training finishes, MT-SDPO discards the EMA state and the teacher pool.

5 Experiments

We ask three questions. **Q1: Supervision allocation.** Does a prompt's domain identify a teacher correct on it? **Q2: Capability integration.** Can MT-SDPO produce one policy strong in aggregate and its weakest domain? **Q3: Component effects.** How much do answer verification, cross-domain reassignment, and aggregation each contribute?

Table 1: Capability integration across five students (avg@16 accuracy, %, step 100). Multi-Domain PT initializes the online rows below it. Bold: best online result per student. Suffix: change of MT-SDPO from Domain Routing, which differs only in the allocation rule.

Model / Method	Chemistry	Materials	Physics	Macro↑	Worst↑	Gap↓
Qwen3-1.7B (Base)	37.10	41.14	46.52	41.59	37.10	9.41
Multi-Domain PT	42.33	63.05	51.66	52.35	42.33	20.73
+ SDPO	40.74	55.93	50.40	49.02	40.74	15.19
+ Domain Routing	45.33	55.46	44.78	48.52	44.78	10.68
+ MT-SDPO (Ours)	46.52	56.65	46.20	**49.79** ↑1.27	**46.20** ↑1.42	**10.44** ↓0.24
Qwen3-4B (Base)	41.69	58.94	59.65	53.43	41.69	17.96
Multi-Domain PT	48.66	71.36	68.67	62.90	48.66	22.71
+ SDPO	51.98	70.17	64.40	**62.18**	**51.98**	18.20
+ Domain Routing	47.78	63.77	65.27	58.94	47.78	17.48
+ MT-SDPO (Ours)	50.63	62.26	63.92	58.94 ±0.00	50.63 ↑2.85	**13.29** ↓4.19
Qwen3-8B (Base)	44.38	58.47	66.22	56.36	44.38	21.84
Multi-Domain PT	49.29	70.25	67.88	62.47	49.29	20.96
+ SDPO	54.11	69.70	52.85	58.89	52.85	16.85
+ Domain Routing	62.03	67.33	63.29	64.21	62.03	**5.30**
+ MT-SDPO (Ours)	67.88	69.38	64.08	**67.11** ↑2.90	**64.08** ↑2.05	**5.30** ±0.00
Llama-3.1-8B-Instruct (Base)	30.38	43.04	41.53	38.32	30.38	12.66
Multi-Domain PT	56.33	59.57	51.42	55.78	51.42	8.15
+ SDPO	58.62	58.86	39.24	52.24	39.24	19.62
+ Domain Routing	57.20	53.96	43.43	51.53	43.43	**13.77**
+ MT-SDPO (Ours)	59.57	53.48	44.62	**52.56** ↑1.03	**44.62** ↑1.19	14.95 ↑1.18
OLMo-3-7B-Instruct (Base)	36.16	57.83	62.82	52.27	36.16	26.66
Multi-Domain PT	48.42	63.21	57.67	56.43	48.42	14.79
+ SDPO	45.41	62.66	54.91	54.33	45.41	17.25
+ Domain Routing	51.90	64.56	63.29	59.92	51.90	12.66
+ MT-SDPO (Ours)	55.69	65.82	64.55	**62.02** ↑2.10	**55.69** ↑3.79	**10.13** ↓2.53

5.1 Experimental Setup

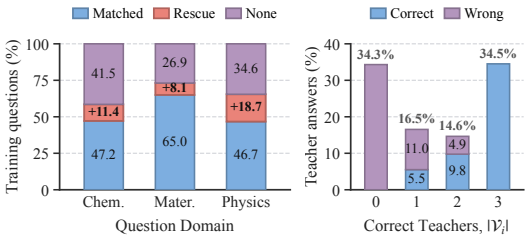
Datasets and Models. We use the Science Q&A subset of the L3 split of SciKnowEval (Feng et al., 2024), restricted to its chemistry, materials science, and physics domains, whose questions each carry a single reference answer that the verifier checks. Its training and test questions, balanced across the three domains, serve for online training and held-out evaluation. The students are Qwen3-1.7B, Qwen3-4B, Qwen3-8B (Yang et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and OLMo-3-7B-Instruct (Team Olmo, 2025), each supervised on the same prompts by a family-matched generator: Qwen3-32B, Llama-3.1-70B-Instruct, or OLMo-3.1-32B-Instruct. Domain-specific post-training gives every student three frozen teachers; multi-domain post-training on the union of the three training sets gives the shared initialization of all controlled online methods.

Protocol and Metrics. Every controlled online method starts from that initialization and trains for 100 steps with 8 on-policy rollouts per prompt on 8 H100 80GB GPUs. Evaluation draws 16 responses for each held-out question (temperature 0.6, top-$p = 0.95$, top-$k = 20$) and reports avg@16 domain accuracy A_d, the mean accuracy over the 16 responses to each held-out question of domain d, from which Macro, Worst, and Gap are the mean, minimum, and range over the three domains. We compare all methods at step 100 and, unless noted otherwise, measure point differences against the same-student Multi-Domain Post-Training policy.

Main Baselines. **SDPO** (Hübotter et al., 2026) spends the same verifier entirely on the student's own rollouts, promoting successful ones to privileged self-supervision, so it is the teacher-free control for what external feedback adds. **Domain Routing** implements MOPD's domain-assignment rule (Ma et al., 2026), sending every failed rollout to its nominal-domain teacher without checking whether it can solve the sample. It inherits the self-anchor, sanitizer, and privileged self-teacher, so only the routing rule differs.

5.2 Domain Match Misses Correct Teachers

Domain match holds on average, not per sample. Before training, we record the eligible set $\mathcal{V}_i$ of Eq. (4) on every training sample. Each teacher does improve its own held-out domain over the shared base, by 8.55, 9.57, and 2.77 points (Appendix D.1). Per sample they are far less separable: the matched teacher is correct on only 52.94% of samples, and on physics samples the chemistry teacher is the more accurate of the two. Teachers sharing one base model and one data pipeline overlap in competence more than their labels suggest.



(a) Teacher coverage (b) Composition

Figure 3: Teacher-pool diagnosis for the Qwen3-8B teachers on the training set. (a) Coverage from matched teachers, cross-domain rescue, and none correct. (b) Correct and wrong teacher answers per correct-teacher bin, as a share of all answers.

Per-sample selection turns that overlap into coverage. Admitting any teacher whose private answer is verified correct raises coverage to 65.69% and recovers 27.09% of what the matched teacher misses. Figure 3(a) breaks this rescue down by domain: largest in physics, at 18.74 points, where domain match is weakest. Figure 3(b) regroups the samples by number of correct teachers: roughly half admit two or three, making aggregation possible, while 34.31% admit none and bound any method built on this pool. That group is why MT-SDPO also keeps self-anchors.

5.3 Main Capability-Integration Results

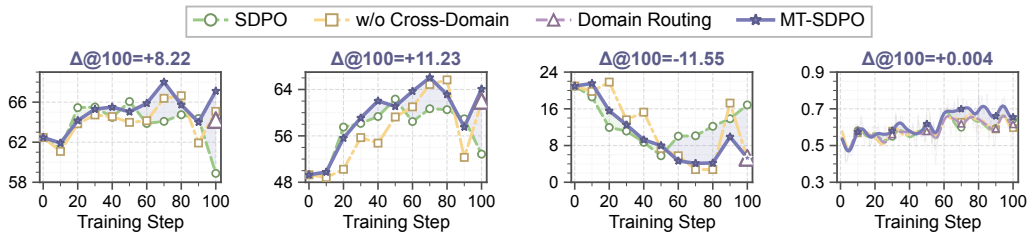
Answer-verified allocation outperforms domain routing on every student. Table 1 reports every method at step 100. Differing from Domain Routing only in the allocation rule, MT-SDPO reaches a higher Worst-domain accuracy on all five students, by 1.19 to 3.79 points ($p = 0.03$, one-sided sign test), and a higher or equal Macro accuracy on all five.

MT-SDPO recovers the weakest domain of Qwen3-8B. On Qwen3-8B, MT-SDPO improves Macro and Worst-domain accuracy by 4.64 and 14.79 points over Multi-Domain Post-Training, and narrows the gap from 20.96 to 5.30. Chemistry, the weakest domain, rises 18.59 points, while materials and physics together give back under five. MT-SDPO integrates capability instead of improving every domain, which is what one deployed policy needs. Appendix D.2 plots every Qwen3-8B variant in the Macro–Worst plane.

On Qwen3-8B, one integrated policy is more balanced than one matched teacher per domain. MT-SDPO improves on the three-model matched-teacher reference of Appendix D.1 by 3.79 Macro and 11.15 Worst-domain points with a single model at inference. Domain Routing reaches the same 5.30 gap but stays 2.90 and 2.05 points lower on the two accuracy metrics: routing alone recovers the balance, and verification and aggregation supply the accuracy.

Spending the verifier on allocation is what redistributes capability. Spent as a scalar reward instead, the signal varies the objective, not the allocation rule. On Qwen3-8B that GRPO reference widens the gap from 20.96 to 21.76, whereas MT-SDPO brings the same start to 5.30 and stands 9.17 points higher in its weakest domain. Appendix D.5 gives its setup and per-domain scores.

A flatter profile arises through two different mechanisms. All three Qwen3 scales cut the domain gap, by 49.6% at 1.7B, 41.5% at 4B, and 74.7% at 8B, but Macro improves at 8B alone. At 1.7B and 4B the flattening comes mostly from the strong domains falling: the weakest domain gains 4.19 and 1.97 points while the other two give back 11.86 and 13.85. At 8B and on OLMo-3-7B-Instruct, the weakest domain instead rises by 18.59 and 7.27 points and carries Macro with it, and OLMo gives the best Macro, Worst, and Gap of its block. Feedback from eligible teachers therefore levels these profiles, but lifts the aggregate only where the weakest domain itself improves.



(a) Macro accuracy ↑ (b) Worst-domain accuracy ↑ (c) Domain gap ↓ (d) Rollout accuracy ↑

Figure 4: Held-out dynamics (a)–(c) and rollout accuracy (d) on Qwen3-8B. Shading is the signed MT-SDPO–SDPO difference, not a confidence interval; Domain Routing appears only at step 100.

MT-SDPO still leads the controlled methods on Llama. Multi-domain post-training alone lifts Llama-3.1-8B-Instruct by 21.04 Worst-domain points and leaves it the flattest initialization of the five; no controlled method improves on it, but MT-SDPO keeps the best Macro and Worst among them.

5.4 Training Dynamics

Non-monotonic trajectories motivate the fixed step-100 comparison. Figure 4 places held-out metrics and training rollout accuracy side by side. SDPO peaks at step 50 and loses more than seven Macro points by step 100, whereas MT-SDPO peaks at step 70 and holds most of it. Because no method improves monotonically, we compare at step 100, not at per-method best checkpoints.

5.5 Ablation Studies

Cross-domain reassignment and aggregation play complementary roles. Table 2 expresses each variant as a removal from full MT-SDPO, with the self-anchor, sanitizer, EMA self-teacher, initialization, and update budget fixed. Restricting feedback to the verified domain-matched teacher (w/o Cross-Domain) starts the ladder; admitting one sample-eligible teacher instead (w/o Aggregation) adds 1.03 Macro and 0.47 Worst-domain points but widens the gap from 5.14 to 7.99; aggregating every eligible teacher in full MT-SDPO adds 1.00 and 1.82 points more and brings the gap to 5.30, nearly recovering the balance. The gain does not come from more frequent feedback: MT-SDPO uses retained teacher feedback on 5.64% of rollouts, against 9.63% for w/o Aggregation, as Appendix D.4 details.

Table 2: Ablation of supervision allocation on Qwen3-8B (%, step 100). Suffix: change from Multi-Domain PT. Bold and underline: best and second best.

Variant	Macro↑	Worst↑	Gap↓
w/o Cross-Domain	65.08 ↑2.61	61.79 ↑12.50	**5.14** ↓15.82
w/o Aggregation	<u>66.11</u> ↑3.64	<u>62.26</u> ↑12.97	7.99 ↓12.97
w/o Verification	60.55 ↓1.92	55.14 ↑5.85	10.92 ↓10.04
MT-SDPO (full)	**67.11** ↑4.64	**64.08** ↑14.79	<u>5.30</u> ↓15.66

Verification makes the larger teacher pool useful. Replacing the eligible set with all three unverified teachers (w/o Verification) loses 6.56 Macro and 8.94 Worst-domain points against MT-SDPO and falls 1.92 Macro points below initialization. The offline manifest shows why: the three teachers are wrong in 50.21% of teacher–sample pairs, against 47.06% under domain routing.

6 Conclusion

We showed that a teacher pool is better allocated per sample than per domain. Sample-level verification recovers a quarter of the samples domain routing misses; cross-domain reassignment then buys accuracy at the cost of balance, and aggregating every eligible teacher buys more of both. Extending MT-SDPO beyond exact-answer settings will require a reliable, task-appropriate alternative to exact-match eligibility.

AI USE STATEMENT

Generative AI tools, including ChatGPT and OpenAI Codex, were used to assist with literature retrieval and the drafting and polishing of manuscript text. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript, including its text, claims, analyses, and associated artifacts.

REPRODUCIBILITY STATEMENT

Section 4 and Algorithm 1 specify one full training step, and Appendix B.1 gives the implemented form of the objective. Appendices A–C list the post-training, online, and evaluation settings, including decoding parameters and the answer parser; Appendix B.2 describes how the private-answer manifest is built, and Appendix E reproduces every prompt verbatim. All experiments use publicly available models and the public SciKnowEval benchmark.

REFERENCES

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

Min Cai, Yu Liang, Longzheng Wang, Yan Wang, Yueyang Zhang, Long Xia, Zhiyuan Sun, Xi Ye, and Daiting Shi. Advancing general-purpose reasoning models with modular gradient surgery. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2602.02301.

DeepSeek-AI. DeepSeek-V3.2: Pushing the frontier of open large language models. Technical report, arXiv, 2025. URL https://arxiv.org/abs/2512.02556.

Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang, Yuqi Tang, Qiang Zhang, and Keyan Ding. SciKnowEval: Evaluating multi-level scientific knowledge of large language models. Technical report, arXiv, 2024. URL https://arxiv.org/abs/2406.09098.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. Technical report, arXiv, 2024. URL https://arxiv.org/abs/2407.21783.

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://openreview.net/forum?id=5h0qf7IBZZ.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Advantage collapse in group relative policy optimization: Diagnosis and mitigation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64547.

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

Jonas Hübotter, Frederike Lübeck, Lejs Behric, et al. Reinforcement learning via self-distillation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64121.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.04089.

Naman Jain, Jaskirat Singh, Manish Shetty, Liang Zheng, Koushik Sen, and Ion Stoica. R2E-Gym: Procedural environments and hybrid verifiers for scaling open-weights SWE agents. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2504.07164.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. Technical report, arXiv, 2025. URL https://arxiv.org/abs/2503.09516.

Ruihan Jin, Pengpeng Shao, Zhengqi Wen, Jinyang Wu, Mingkuan Feng, Shuo Yang, Chu Yuan Zhang, and Jianhua Tao. Exploring knowledge purification in multi-teacher knowledge distillation for LLMs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7pvJoB4aKO.

Jeonghye Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Rebellious student: Reversing teacher signals for reasoning exploration with self-distilled RLVR. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2605.10781.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 2024. URL https://arxiv.org/abs/2402.03898.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, et al. Tülu 3: Pushing frontiers in open language model post-training. Technical report, arXiv, 2024. URL https://arxiv.org/abs/2411.15124.

Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020. doi: 10.1016/j.neucom.2020.07.048. URL https://doi.org/10.1016/j.neucom.2020.07.048.

Wenhan Ma, Jianyu Wei, Liang Zhao, et al. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2606.30406.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, et al. WebGPT: Browser-assisted question-answering with human feedback. Technical report, arXiv, 2021. URL https://arxiv.org/abs/2112.09332.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. Technical report, arXiv, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. Technical report, arXiv, 2024. URL https://arxiv.org/abs/2402.03300.

Yuda Song, Lili Chen, Fahim Tajwar, Remi Munos, Deepak Pathak, J. Andrew Bagnell, Aarti Singh, and Andrea Zanette. Expanding the capabilities of reinforcement learning via text feedback. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2602.02482.

Team Olmo. Olmo 3. Technical report, arXiv, 2025. URL https://arxiv.org/abs/2512.13961.

Boxin Wang, Chankyu Lee, Nayeon Lee, et al. Nemotron-Cascade: Scaling cascaded reinforcement learning for general-purpose reasoning models. Technical report, arXiv, 2025. URL https://arxiv.org/abs/2512.13607.

Jianze Wang, Ying Liu, Jinlong Chen, Xuchun Hu, Qilong Zhang, Yu Cao, Jun Wang, Hua Yang, Yong Xie, and Qianglong Chen. MAD-OPD: Breaking the ceiling in on-policy distillation via multi-agent debate. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2605.01347.

Yuxiang Wei, Olivier Duchenne, Jade Copet, Quentin Carbonneaux, Lingming Zhang, Daniel Fried, Gabriel Synnaeve, Rishabh Singh, and Sida I. Wang. SWE-RL: Advancing LLM reasoning via reinforcement learning on open software evolution. In *Advances in Neural Information Processing Systems*, volume 38, pp. 78500–78525. Curran Associates, Inc., 2025. URL https://arxiv.org/abs/2502.18449.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, et al. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

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

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. Technical report, arXiv, 2025. URL https://arxiv.org/abs/2505.09388.

Zhuolin Yang, Zihan Liu, Yang Chen, et al. Nemotron-Cascade 2: Post-training LLMs with cascade RL and multi-domain on-policy distillation. Technical report, arXiv, 2026a. URL https://arxiv.org/abs/2603.19220.

Ziyu Yang, Guibin Chen, Yuxin Yang, Aoxiong Zeng, and Xiangquan Yang. Disentangling task conflicts in multi-task LoRA via orthogonal gradient projection. Technical report, arXiv, 2026b. URL https://arxiv.org/abs/2601.09684.

Hua Ye, Siyuan Chen, Haoliang Zhang, Weihao Luo, Yanbin Li, and Xuan Zhang. Synergy over discrepancy: A partition-based approach to multi-domain LLM fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 18893–18923. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1b4839ff1f843b6be059bd0e8437e975-Paper-Conference.pdf.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2602.12275.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1285–1294. Association for Computing Machinery, 2017. doi: 10.1145/3097983.3098135. URL https://doi.org/10.1145/3097983.3098135.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836. Curran Associates, Inc., 2020. URL https://arxiv.org/abs/2001.06782.

Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. Reinforced multi-teacher selection for knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14284–14291, 2021. doi: 10.1609/aaai.v35i16.17680. URL https://ojs.aaai.org/index.php/AAAI/article/view/17680.

Hengyuan Zhang, Shiping Yang, Xiao Liang, et al. Find your optimal teacher: Personalized data synthesis via router-guided multi-teacher distillation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14619–14637, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.666. URL https://aclanthology.org/2026.acl-long.666/.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.15240.

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. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2601.18734.

A Post-Training and Initialization

For five student backbones (Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Llama-3.1-8B-Instruct, and OLMo-3-7B-Instruct), we construct three domain teachers and one Multi-Domain Post-Training policy from the family-matched base model. The teachers use separate domain sets; the Multi-Domain policy uses their balanced union. Post-training, controlled online optimization, and evaluation follow the shared recipes in Tables 3–5. Identifiers are family-specific, and family-matched generators and validation/recovery pipelines produce the supervision. The teachers use their final three-epoch weights and remain frozen. Every controlled online method starts from the corresponding epoch-2 Multi-Domain checkpoint; no method-specific initialization is used.

Table 3: Post-training settings shared across all student backbones. Domain-specific and Multi-Domain Post-Training use the same recipe for Qwen3, Llama-3.1, and OLMo-3.

Parameter	Value
General	
Model update	Full-parameter post-training
Optimizer	Fused AdamW
Learning rate	1×10^{-5}
Schedule	Cosine, 5% warmup
Training epochs	3
Data	
Global batch size	4
Maximum sequence length	8,192 tokens
Response handling	Response tokens only; no packing
Training	
Weight decay / gradient clipping	0.1 / 1.0
Numeric precision	bfloat16

B Online Optimization and Teacher Feedback

Algorithm 1 gives the procedure described in Section 4. Table 4 lists the online optimization budget and distillation settings. Each controlled method processes 32 prompts per update, draws eight on-policy responses per prompt, and runs for 100 updates. The methods use the same prompt IDs, rollout-sampling configuration, and student-update budget. Each runs on eight H100 80GB GPUs. SDPO allocates all eight GPUs to the policy; methods with external teachers split four for policy training and four for serving the frozen teachers.

B.1 Implemented Form of the Distillation Objective

Equation (7) states the objective in its masked, importance-weighted form. Three details are fixed by the implementation. First, the Jensen–Shannon divergence $D_{\mathrm{JSD}}(p\|q) = \frac{1}{2}D_{\mathrm{KL}}(p\|u)+\frac{1}{2}D_{\mathrm{KL}}(q\|u)$, where $u = \frac{1}{2}(p+q)$, is evaluated on a truncated support rather than the full vocabulary. At each position, the operator $\mathcal{T}^j_{i,t}[\cdot]$ retains the 100 tokens with the highest student probability and collects the remaining mass in a tail symbol. The student defines the support, while each branch retains its own tail mass. Second, the weight $w^j_{i,t} = \min(w^{j,\mathrm{prox}}_{i,t}, \rho) \cdot \min(w^{j,\mathrm{TIS}}_{i,t}, \rho)$ combines the PPO proximal ratio $w^{j,\mathrm{prox}}_{i,t} = \pi_\theta(y^j_{i,t})/\pi_{\theta_{\mathrm{old}}}(y^j_{i,t})$ with the truncated-importance-sampling ratio $w^{j,\mathrm{TIS}}_{i,t} = \pi_{\theta_{\mathrm{old}}}(y^j_{i,t})/\pi_{\theta_{\mathrm{rollout}}}(y^j_{i,t})$. Both ratios are conditioned on $(x_i, y^j_{i,<t})$, detached, and clipped only from above at $\rho = 2$. Third, the normalizer is applied per micro-batch rather than globally. For dynamic micro-batches $\{\mathcal{M}_b\}_{b=1}^{B_\mu}$, let $n_b = |\mathcal{M}_b|$, $N = \sum_b n_b$, and $Z_b = \max\{1, \sum_{(i,j)\in\mathcal{M}_b,t} m^j_i M^j_{i,t}\}$.

Algorithm 1 Answer-Verified MT-SDPO

Input: data $\mathcal{D}$; student π_θ; EMA self-teacher $\pi_{\bar{\theta}}$; frozen teachers $\{E_k\}_{k=1}^{K}$; rollout count G; EMA rate τ
Output: updated student model π_θ

// Offline: cache sample-level teacher eligibility
$\mathcal{V}_i \leftarrow \{k : \textsc{Verify}(E_k^{\text{ans}}(x_i), a_i^*) = 1\},\ \forall (x_i, a_i^*) \in \mathcal{D}$
// Online: align supervision to each rollout
for each online training step **do**
  $\mathcal{B} \sim \mathcal{D}$; $y_i^{1:G} \sim \pi_\theta(\cdot \mid x_i)$ for each $x_i \in \mathcal{B}$
  $\mathcal{S}_i \leftarrow \{j : \textsc{Verify}(y_i^j, a_i^*) = 1\}$
  for each rollout y_i^j in the batch **do**
    $s_i^j \leftarrow \text{Trunc}(\text{earliest rollout in } \mathcal{S}_i \setminus \{j\}), \text{ or } \varnothing$ ▷ Eq. (3)
    $F_i^j \leftarrow \varnothing$
    if $j \notin \mathcal{S}_i$ **then**
      for each $k \in \mathcal{V}_i$ **do**
        $F_i^j \leftarrow \text{Concat}\left[F_i^j, \Gamma(E_k^{\text{fb}}(x_i, y_i^j))\right]$
      end for
    end if
    $c_i^j \leftarrow [s_i^j; F_i^j]$; $m_i^j \leftarrow \mathbb{1}[c_i^j \neq \varnothing]$
    Accumulate the (i, j) terms of $\mathcal{L}_{\text{MT-SDPO}}$ ▷ Eq. (7)
  end for
  Update θ over the collected batch
  $\bar{\theta} \leftarrow (1 - \tau)\bar{\theta} + \tau\theta$
end for
return π_θ; discard $\pi_{\bar{\theta}}$ and all teachers

The implemented loss is the sequence-count-weighted average of the micro-batch token means:

$$\mathcal{L}_{\text{MT-SDPO}} = \sum_{b=1}^{B_\mu} \frac{n_b}{N} \frac{1}{Z_b} \sum_{(i,j)\in\mathcal{M}_b} \sum_t m_i^j M_{i,t}^j w_{i,t}^j \cdot D_{\text{JSD}}\left(\mathcal{T}_{i,t}^j \left[p_{\theta,i,t}^j\right] \middle\| \mathcal{T}_{i,t}^j \left[q_{\bar{\theta},i,t}^j\right]\right). \tag{8}$$

B.2 Private-Answer Manifest

For every student family, each frozen teacher answers every training question once before online training. Decoding is deterministic, with temperature 0, top-$p = 1.0$, and a maximum of 6,144 tokens. Manifest construction uses the generic SciKnowEval L3 rollout prompt reproduced in Section E.2. This prompt is also used for online training and evaluation. The programmatic answer checker compares the parsed answer with the reference answer and records the sample-level eligible set. Neither the private-answer prompt nor the subsequent teacher-feedback prompt contains the reference answer. MT-SDPO queries every eligible teacher. The w/o Cross-Domain control (verified domain routing) restricts this set to the nominal-domain teacher. Domain Routing and w/o Verification (the unverified all-teacher control) do not use the manifest for eligibility.

B.3 Rollout-Specific Feedback

External feedback is requested only when the student rollout is incorrect. Eligible teachers are queried independently in the fixed order of chemistry, materials science, and physics. Decoding uses temperature 0.2, top-$p = 0.95$, and a 256-token output limit. A message is retained only if it has a valid concept anchor under the lexical sanitizer and passes the answer-leakage checks. Failed or rejected messages are dropped. The retained messages are concatenated with the available self-anchor under the token budgets in Table 4.

Table 4: Online settings shared across student families and controlled methods. The methods differ only in their available privileged context, as defined in Section 5.1.

Parameter	Value
Training	
Prompts per update / rollouts per prompt	32 / 8
Number of updates	100
Student learning rate	1×10^{-5}
Batching / Rollout	
Rollout sampling	Temperature 1.0, top-$p = 1.0$
Maximum prompt length	2,048 tokens
Maximum response length	8,192 tokens
Optimization mini-batch	32, dynamic micro-batching
Dynamic token cap	18,944 tokens per GPU
Distillation	
Distillation support	Student top-100 tokens plus tail mass
JSD mixture coefficient	Fixed at $1/2$
EMA update rate	$\tau = 0.05$ per rollout batch
Importance-ratio upper clip	2.0 for both ratios
Feedback processing and gating	
Anchor token budget	4,096
Feedback token budget	1,024
Maximum privileged re-prompt length	10,240 tokens

Table 5: Frozen evaluation settings shared across all student backbones.

Parameter	Value
Sampling	
Responses per question	16
Sampling	Temperature 0.6, top-$p = 0.95$, top-$k = 20$
Maximum new tokens	6,144
Evaluation	
Primary parser	Last explicit `answer` field
Missing, invalid, or truncated answer	Incorrect
Macro aggregation	Unweighted mean over three domains

C Evaluation Protocol

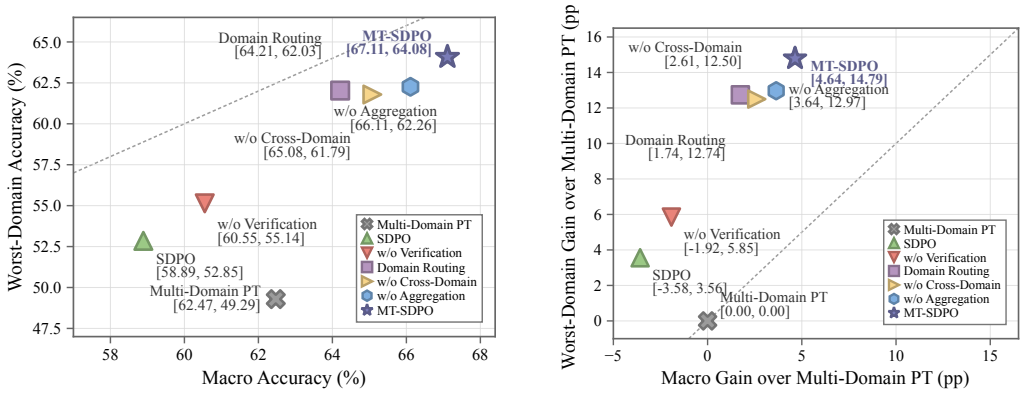
All five student backbones and all reported checkpoints use the same frozen evaluation protocol. Each held-out question receives 16 independently sampled responses. Table 5 records the settings that affect the reported avg@16 results.

C.1 Guarded Parsing Fallback

When the primary parser does not establish a correct answer, a fixed Qwen3-8B base model is queried as a deterministic binary judge (temperature 0, top-$p = 1.0$). The fallback cannot overturn an explicitly extracted non-gold answer, and a judge parse failure is counted as incorrect. This rule handles ambiguous formatting without allowing the judge to replace the answer key.

Table 6: Matched-teacher reference (avg@16 accuracy, %). Every domain is scored with its own frozen domain teacher of the same student family, so a row combines three models.

Student	Chemistry	Materials	Physics	Macro↑	Worst↑	Gap↓
Qwen3-1.7B	45.33	65.59	50.24	53.72	45.33	20.25
Qwen3-4B	51.27	69.38	66.06	62.24	51.27	18.12
Qwen3-8B	52.93	68.04	68.99	63.32	52.93	16.06
Llama-3.1-8B-Instruct	62.03	63.69	56.57	60.76	56.57	7.12
OLMo-3-7B-Instruct	48.02	61.55	53.24	54.27	48.02	13.53



(a) Absolute accuracy

(b) Accuracy gain

Figure 5: Step-100 capability integration on Qwen3-8B. (a) Macro against worst-domain accuracy. (b) Gains over Multi-Domain PT. Dashed diagonals mark parity.

D Additional Results

D.1 Matched-Teacher Reference

Table 6 scores each held-out domain with the student's frozen teacher for that domain. Each row therefore combines three models and is not a deployable policy. Section 5.2 compares these teachers with the base model, and Section 5.3 compares MT-SDPO against this reference on Qwen3-8B.

D.2 Step-100 Endpoints on Qwen3-8B

Figure 5 places every Qwen3-8B method and ablation variant of Tables 1 and 2 in the Macro–Worst plane at step 100, both in absolute terms and as gains over the shared Multi-Domain PT initialization.

D.3 Cross-Scale Training Dynamics

Figure 6 reports the held-out Macro and Worst-domain trajectories for every completed Qwen3 student scale. Intermediate checkpoints are diagnostic only: the main comparison continues to use the common, pre-specified step-100 checkpoint for every online method.

The trajectories differ across scales. At step 100, Qwen3-1.7B and Qwen3-4B primarily recover weakest-domain capability rather than improving Macro accuracy over their respective initializations. On Qwen3-8B, MT-SDPO finishes above its initialization and SDPO on both Macro and Worst-domain accuracy. The non-monotonic trajectories support the use of a common checkpoint rather than method-specific checkpoint selection.

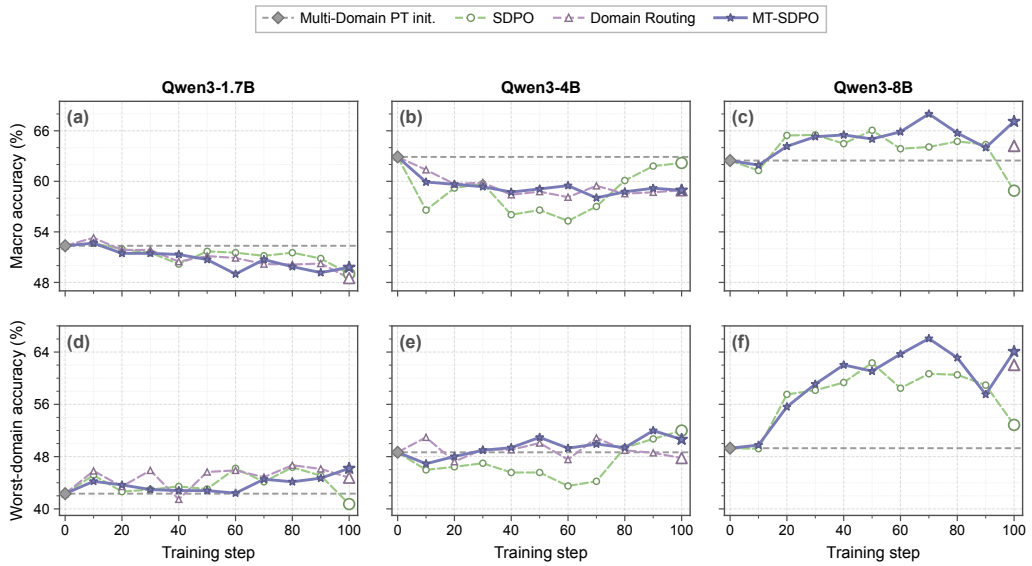
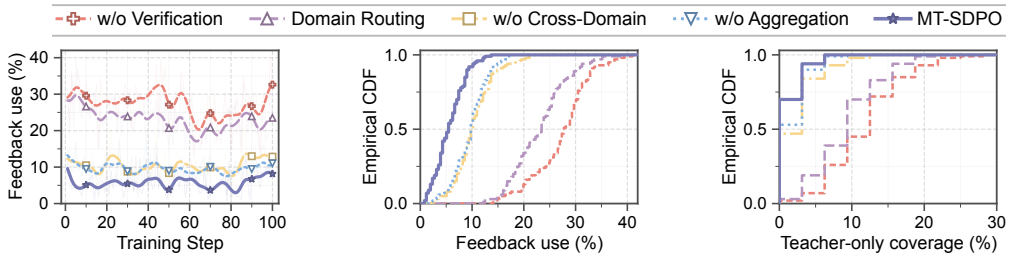


Figure 6: Cross-scale held-out dynamics, evaluated every 10 training steps. Columns are Qwen3-1.7B, Qwen3-4B, and Qwen3-8B, and the two rows report Macro and worst-domain accuracy. The horizontal reference marks each student's Multi-Domain PT initialization, and lines connect evaluated checkpoints without smoothing. Domain Routing has a held-out result only at step 100 on Qwen3-8B, so it appears as a single step-100 marker in (c) and (f).



(a) Feedback use trend (b) Feedback use distribution (c) Teacher-only coverage

Figure 7: Realized teacher-feedback exposure on Qwen3-8B over 100 training updates. (a) Smoothed share of rollouts that use retained teacher feedback. (b) Empirical distribution of the corresponding raw per-update shares. (c) Empirical distribution of the share of prompts with no successful student rollout but retained teacher feedback. Higher exposure alone implies neither more accurate feedback nor better held-out performance.

D.4 Additional Feedback Diagnostics

Figure 7 reports realized feedback exposure on Qwen3-8B. These diagnostics are distinct from the offline teacher-pool coverage and the held-out and rollout-accuracy dynamics in Section 5. Offline coverage asks whether a correct teacher exists for a training question. Here, feedback use is measured after the student-error condition, each method's eligibility rule, feedback generation, and sanitization. Panel (a) uses centered Gaussian smoothing ($\sigma = 2$) only for visualization. Panels (b) and (c) report empirical distributions over all 100 recorded updates.

Across the 100 updates, Domain Routing and w/o Verification use retained teacher feedback on 23.07% and 27.32% of rollouts on average. The corresponding averages are 10.39% for w/o Cross-Domain, 9.63% for w/o Aggregation, and 5.64% for MT-SDPO. Average teacher-only prompt coverage follows the same broad ordering. Domain Routing and w/o Verification reach 9.16% and 11.66%, while w/o Cross-Domain, w/o Aggregation, and MT-SDPO reach 2.44%, 1.81%, and 1.12%.

Table 7: Verifiable-reward RL reference from the shared initialization (step-100 avg@16 accuracy, %, under the protocol of Appendix C). Within each block, Multi-Domain PT is the shared initialization for GRPO and MT-SDPO. GRPO changes the training objective and lies outside the controlled online-distillation comparison.

Method	Chem.	Mat.	Phys.	Macro	Worst	Gap
Qwen3-1.7B						
Multi-Domain PT	42.33	63.05	51.66	52.35	42.33	20.73
GRPO	45.73	64.08	54.43	54.75	45.73	18.35
MT-SDPO	46.52	56.65	46.20	49.79	46.20	10.44
Qwen3-4B						
Multi-Domain PT	48.66	71.36	68.67	62.90	48.66	22.71
GRPO	52.53	74.53	66.22	64.43	52.53	21.99
MT-SDPO	50.63	62.26	63.92	58.94	50.63	13.29
Qwen3-8B						
Multi-Domain PT	49.29	70.25	67.88	62.47	49.29	20.96
GRPO	54.91	76.66	70.25	67.27	54.91	21.76
MT-SDPO	67.88	69.38	64.08	67.11	64.08	5.30
Llama-3.1-8B-Instruct						
Multi-Domain PT	56.33	59.57	51.42	55.78	51.42	8.15
GRPO	70.65	66.46	64.16	67.09	64.16	6.49
MT-SDPO	59.57	53.48	44.62	52.56	44.62	14.95
OLMo-3-7B-Instruct						
Multi-Domain PT	48.42	63.21	57.67	56.43	48.42	14.79
GRPO	47.47	55.85	36.87	46.73	36.87	18.98
MT-SDPO	55.69	65.82	64.55	62.02	55.69	10.13

This exposure ordering differs from the step-100 performance ordering in Tables 1 and 2. Feedback volume alone therefore does not explain integration quality. In these runs, answer verification is associated with less frequent, more selective external supervision.

The teacher-only metric uses all prompts as its denominator. It is the joint frequency of a missing successful student rollout and available retained teacher feedback. It is not the conditional probability of receiving feedback given a missing self-anchor. Neither metric tests whether the retained feedback is semantically correct or repairs the student's answer. Performance claims remain anchored to the step-100 checkpoint and exact held-out trajectories reported in Section 5.

D.5 Verifiable-Reward RL Reference

The ablation comparison of Section 5.5 holds the teacher pool, self-anchor, sanitizer, privileged self-teacher, initialization, and update budget fixed while varying the supervision-allocation rule. GRPO is outside that comparison because it changes the objective and uses no teacher pool. It replaces the token-level distillation objective of Eq. (7) with a policy gradient on a scalar reward. The programmatic verifier used by MT-SDPO as an eligibility gate supplies that reward. Table 7 reports GRPO only as a separate objective reference under the same update and rollout budget.

Each GRPO reference run starts from the corresponding epoch-2 Multi-Domain Post-Training checkpoint used by the controlled methods. It runs for 100 updates with 8 on-policy rollouts per prompt and uses eight H100 80GB GPUs. The verifier and evaluation protocol in Section C remain the same. No teacher, feedback, or self-teacher component is active, and the reward is the binary verifier outcome on the student's own rollouts.

The GRPO results are reported descriptively across all five backbones. Because GRPO changes the training objective, these results are separate from the controlled supervision-allocation comparison and do not establish a general ordering between the two objectives.

E PROMPT TEMPLATES

This section reports the exact message text used for SciKnowEval L3 data construction, post-training, evaluation, and online training. The SFT construction prompts in Section E.1 document the Qwen3 pipeline. Llama and OLMo use family-matched generators and their corresponding validation/retry pipelines rather than reusing Qwen3 responses. The student rollout, evaluation, online self-teacher reprompt, and teacher-feedback instructions are shared across families at the natural-language level. Tokenization and chat-template handling remain model-specific. Braced fields denote runtime substitutions rather than literal input. The domain-specific field {`domain_expert`} is `chemistry expert`, `materials science expert`, or `physics expert`. The field {`expert_domain`} is the corresponding display name.

E.1 SFT DATA CONSTRUCTION

INDEPENDENT GENERATOR SAMPLING

Two independent candidates are sampled for each training record. The reference answer is not included in this prompt.

Independent Generator Sampling

System:
You are a knowledgeable and rigorous `{domain_expert}`.

User:
Solve the following multiple-choice question. Give a concise, self-contained derivation using the relevant scientific principles, equations, calculations, or evidence. Include enough explicit reasoning for a reviewer to check, but avoid unnecessary background, repeated calculations, speculative alternatives, or restating the problem. Do not return only an answer.

Question:
`{question}`

Choices:
`{choices_block}`

End your response with exactly one final line in this format: "answer": "C". Replace C with your selected choice letter. Do not add Markdown styling, a code fence, a heading, or any text after that final line.

GOLD-CONDITIONED FALLBACK

This fallback is used only if neither independent candidate yields a parseable correct label. Failed independent completions are not inserted into the fallback prompt.

Gold-Conditioned Fallback Generation

System:
You are a knowledgeable and rigorous `{domain_expert}`. You may receive a private correctness constraint. Use it only internally, and never mention or allude to its existence in your response.

User:
Construct a concise, self-contained solution to the following multiple-choice question. Use the private correctness constraint internally and build a scientifically valid derivation that genuinely supports it. Include enough explicit reasoning for a reviewer to check, but avoid unnecessary background, repeated calculations, speculative alternatives, or restating the problem. Do not return only an answer.

Never mention or allude to the private constraint, an answer key, a hint, or any supplied answer. If the constraint cannot be scientifically justified, explain only the scientific inconsistency in the problem without disclosing that a constraint was supplied.

Question:
`{question}`

Choices:
`{choices_block}`

Private correctness constraint:
`{answer_key}. {answer_text}`

End your response with exactly one final line in this format: "answer": "C". Replace C with the constrained choice letter. Do not add Markdown styling, a code fence, a heading, or any text after that final line.

SELF-VERIFIER

Self-Verifier for Generated Reasoning

System:
You are a rigorous reviewer of scientific reasoning.

User:
Review the proposed solution to the following multiple-choice question.

Check whether its scientific facts, equations, calculations, units, and logical steps are correct. A correct final choice is not sufficient: reject the solution if it reaches the correct answer through incorrect, unsupported, circular, or fabricated reasoning.

Do not rewrite the solution. Briefly explain your judgment and finish with exactly one of the following plain-text lines. Do not apply Markdown styling to the verdict line.

Verdict: ACCEPT
Verdict: REJECT
Verdict: REVIEW

Question:
`{question}`

Choices:
`{choices_block}`

Reference answer:
`{answer_key}. {answer_text}`

Proposed solution:
`{cot_response}`

For the Qwen3 data, the generator and verifier are both Qwen3-32B, so this stage is self-verification rather than independent verification. The retained completion becomes the SFT assistant target without summarization, rewriting, or answer normalization.

E.2 Student SFT and Initial Rollout

Student SFT uses a domain-specific system message. Online rollout, offline private-answer manifest construction, and evaluation instead use the generic system message below. Their user message is identical.

System Messages for SFT and Rollout

Student SFT system:
You are a knowledgeable and rigorous `{domain_expert}`.

Online rollout, manifest, and evaluation system:
You are a knowledgeable and rigorous scientific reasoning assistant.

Shared Multiple-Choice User Message

Solve the following multiple-choice question. Give a concise, self-contained derivation using the relevant scientific principles, equations, calculations, or evidence. Include enough explicit reasoning for a reviewer to check, but avoid unnecessary background, repeated calculations, speculative alternatives, or restating the problem. Do not return only an answer.

Question:
`{question}`

Choices:
`{choices_block}`

Please show your choice in the answer field with only the choice letter, e.g., "answer": "C".

E.3 Evaluation Prompts

Candidate generation uses exactly the generic system message and shared user message in Section E.2. Any operational prefix inserted by a family's chat template is not part of the natural-language instruction or the generated completion.

The fallback judge is called only when the primary regex parser has not established correctness. It cannot overturn an explicitly extracted non-gold answer or accept a truncated response with no answer.

Guarded Fallback Judge

System:
You are a strict multiple-choice answer verifier. Ignore all reasoning, compare only the candidate's final selected option with the reference answer, and output only true or false.

User:
Candidate response:
<candidate_response>
`{candidate_response}`
</candidate_response>

Reference answer:
`{answer_key}`. `{answer_text}`

The reference answer label is `` `{answer_key}` ``. Ignore all reasoning. Output exactly `true` only if the candidate's last explicit `answer` field or `Final Answer` is also `` `{answer_key}` ``. Output exactly `false` for any other answer or no final answer. Do not output any explanation.

E.4 Online Self-Teacher Reprompts

The self-teacher reprompt is assembled from the original rollout prompt and independently available context blocks. An unavailable solution or feedback source contributes an empty string.

SDPO

SDPO Self-Teacher Reprompt

Main template:

```
{prompt}{solution}{feedback}
```

Correctly solve the original question.

Successful solution section:

Correct solution:

```
{successful_previous_attempt}
```

Environment feedback section:

The following is feedback from your unsuccessful earlier attempt:

```
{feedback_raw}
```

The formal SDPO baseline sets environment feedback to disabled, so the feedback section above is not inserted. The successful solution section is inserted only when a reward-verified successful attempt is available.

MT-SDPO AND ROUTED VARIANTS

MT-SDPO Self-Teacher Reprompt

Main template:

```
{prompt}{solution}{feedback}
```

Use the available information to reconsider the reasoning and correctly solve the original question. Follow the original response format and end with an answer field containing only the selected choice letter.

Successful solution section:

A reward-verified successful attempt previously generated by the current target policy is shown below:

```
{successful_previous_attempt}
```

Frozen-expert feedback section:

The following diagnostic feedback was produced by frozen domain specialists for the current attempt:

```
{feedback_raw}
```

MT-SDPO, Domain Routing, and the w/o Cross-Domain, w/o Aggregation, and w/o Verification variants use the same reprompt text. The three ablations correspond to verified domain routing, one eligible teacher, and unverified all-teacher feedback, respectively. They differ only in teacher selection and correctness filtering. The anchor, feedback, and total reprompt budgets are 4,096, 1,024, and 10,240 tokens, respectively.

E.5 FROZEN TEACHER DIAGNOSTIC PROMPT

Each selected teacher receives the question and the current student response. The concept tag is used for anchor filtering and is removed before retained paragraphs enter the self-teacher's privileged context. Retained outputs are concatenated in the fixed order of chemistry, materials science, and physics.

Frozen Teacher Diagnostic

System:
You are a knowledgeable and rigorous scientific reasoning assistant.

User:
You are acting as a frozen `{expert_domain}` specialist. Diagnose one response produced by the target model without revealing the final answer.

Complete two steps.

Step 1 — Problem anchor:
Identify one technical term or short concept, containing 1–3 words, that is central to the question. Copy the wording directly from the question. Output it first inside <concept>...</concept>. Do not use generic words such as problem, question, answer, value, calculation, formula, model, student, expert, or response.

Step 2 — Diagnostic feedback:
After the concept tag, write a short paragraph of 3–5 sentences and no more than approximately 80 words. Focus on the aspects relevant to your `{expert_domain}` expertise when useful. Identify the specific step, assumption, or concept in the target response that is correct, incorrect, or missing, and explain how the reasoning should be reconsidered.

Do not solve the entire problem from scratch.
Do not reveal the correct choice letter.
Do not quote the correct option text.
Do not write a final answer.
Do not emit an "answer": "X" field.
Do not mention that an answer key or verifier exists.

Question:
`{question}`

Choices:
`{choices_block}`

Target response:
`{response}`

Output:
<concept>...</concept>

Decoding: temperature $= 0.2$, top-$p = 0.95$, and maximum new tokens $= 256$.

F Feedback Effectiveness

We separate feedback exposure from feedback effectiveness. We compare the exposure statistics in Figure 7 with the held-out step-100 results in Section 5. Methods using more retained feedback do not necessarily score higher. This descriptive mismatch means that feedback volume is not itself evidence of utility. The held-out comparisons in Section 5 measure the end-to-end outcome of each training method. They do not isolate the causal contribution of an individual feedback message.

The fixed held-out protocol supports a method-level comparison, not a message-level causal repair rate. A direct estimate requires post-feedback completions or paired runs with and without feedback on the same erroneous rollouts.