RECOVERING GENERAL CAPABILITIES VIA UNCERTAINTY-CALIBRATED MULTI-TEACHER ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Specializing large language models to vertical domains improves domain-specific behavior but often degrades general capabilities. We study this trade-off in Multi-Teacher On-Policy Distillation (MOPD), where a specialized model learns from domain and general teachers on its own sampled trajectories. Standard MOPD faces two limitations: ordinary on-policy sampling rarely exposes tokens with large positive teacher–student advantages, and advantage sign alone does not establish whether the proposed update direction is reliable. We propose Uncertainty-Calibrated MOPD (UCMOPD), which addresses these limitations through two complementary mechanisms. Golden-Gain Enhancement combines higher-temperature exploration with a standard-temperature anchor and retains trajectories whose positive learning signal matches or exceeds the prompt-specific anchor. Teacher-Endorsement Filtering then uses centered log-likelihood (CLL) to estimate each retained token's plausibility relative to the teacher's uncertainty and probabilistically preserves updates whose directions are supported by that endorsement. Across role-playing and medical-domain specialization, UC-MOPD improves the general-capability average over standard MOPD by 4.48% and 7.86%, respectively, while maintaining vertical-domain performance. Component ablations and diagnostic analyses support the intended roles of the two mechanisms: exposing and selecting stronger positive signals at the trajectory level and validating update directions through teacher endorsement at the token level.

1 INTRODUCTION

Large language models have demonstrated strong general-purpose capabilities across reasoning, coding, instruction following, and open-ended generation, and are increasingly adapted into vertical-domain systems (Chen et al., 2024b; Lu et al., 2024; Yang et al., 2025b). In domains such as role-playing and medical consultation, specialization is essential: the model must internalize domain-specific knowledge and capabilities (Wang et al., 2026; Dou et al., 2026; Jin et al., 2021). At the same time, practical vertical-domain systems cannot specialize at the expense of general capabilities, because domain tasks often require combining domain knowledge with general reasoning, instruction following, and open-ended communication skills (Liu et al., 2024; Luo et al., 2023; Dong et al., 2024). After supervised fine-tuning or preference-based optimization on domain data, the model may become better at the target domain while losing part of its general ability, including reasoning, mathematical problem solving, coding, instruction following, and open-ended writing. This domain–general trade-off is especially problematic when the final system is expected to remain a capable general assistant while exhibiting strong domain behavior.

A straightforward remedy is to mix vertical-domain and general-domain data during specialization. Although technically simple, this strategy creates additional practical burdens. Beyond constructing a large, high-quality vertical-domain corpus, developers must invest further effort in collecting and filtering high-quality general data. Even after both sources are available, their mixture ratio must be tuned heuristically: too little general data may fail to preserve general capabilities, whereas too much may weaken domain specialization (Bethune et al., 2025). Maintaining general capability during specialization therefore adds substantial data-construction and mixture-tuning costs. This

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.

raises a natural question: can training first focus exclusively on domain capability and then use an open general model to integrate domain and general capabilities into a single model?

Multi-Teacher On-Policy Distillation (MOPD) (Xiao et al., 2026; Dou et al., 2026; Yang et al., 2026b; Xu et al., 2026a) offers a viable solution. Training can first focus exclusively on vertical-domain capability to obtain a domain model that serves as the domain-expert teacher, while an open general model serves as the general-expert teacher; MOPD then integrates the capabilities of both experts into a single model. Because vertical-domain models are typically obtained by continuing training from general models, the overall problem can be viewed as restoring general capability after vertical-domain training. A standard setup initializes the student from the vertical-domain model and uses a frozen copy of that model and the original general model as the domain and general teachers, respectively. The student generates responses on vertical-domain and general prompts matched to the two teachers' capability domains, and the corresponding teacher provides supervision, thereby preserving vertical-domain capability while restoring general capability (Chen et al., 2026).

Effective MOPD generally relies on prompts that adequately cover the capability distributions of the expert teachers (Chen et al., 2026), yet the original training prompts of open general models are typically unavailable. Publicly available post-training prompts can serve as practical surrogates, but assembling a sufficiently comprehensive set is costly and still unlikely to reproduce the original training distribution. We therefore seek to improve the data efficiency of MOPD, aiming to recover general capabilities effectively from a limited set of publicly available prompts. Given a prompt, the student autoregressively samples a complete rollout trajectory from its own conditional distribution, after which the teacher evaluates each sampled token under the same prefix. The difference between the conditional log-probabilities assigned to that token by the teacher and student defines the distillation advantage, $A_t = \log \pi_T(y_t \mid q, y_{<t}) - \log \pi_S(y_t \mid q, y_{<t})$. When $A_t > 0$, the corresponding update increases the student's conditional log-probability of the sampled token; when $A_t < 0$, it decreases that conditional log-probability. More importantly, a larger positive advantage induces a stronger positive update, allowing the student to move toward the teacher more rapidly (Gu et al., 2024).

Existing MOPD still faces two important problems. (1) Because the student samples each token according to its own conditional distribution, sampled tokens tend to have relatively high student probabilities. Positive advantage requires the teacher probability to exceed the student probability, so large positive advantages are difficult to observe under ordinary sampling, limiting learning from strong positive signals. (2) Using only the sign of advantage to determine whether to increase or decrease the sampled token's probability is not always reliable. As Figure 1(a) illustrates, several plausible continuations may divide the teacher's probability mass, while the student assigns slightly more probability to one of these teacher-endorsed tokens. If that token is sampled, it has $A_t < 0$ and is mistakenly suppressed. Conversely, Figure 1(b) shows that a weak student may sample a poorly learned token that the teacher endorses only weakly; if the teacher probability is nevertheless slightly higher than the student's, the token has $A_t > 0$ and is mistakenly reinforced. Thus, advantage sign provides only a candidate update direction; its reliability must be assessed independently using the teacher's own endorsement of the current token. Here, *teacher endorsement* refers to the teacher's judgment of the token's plausibility based solely on its own distribution.

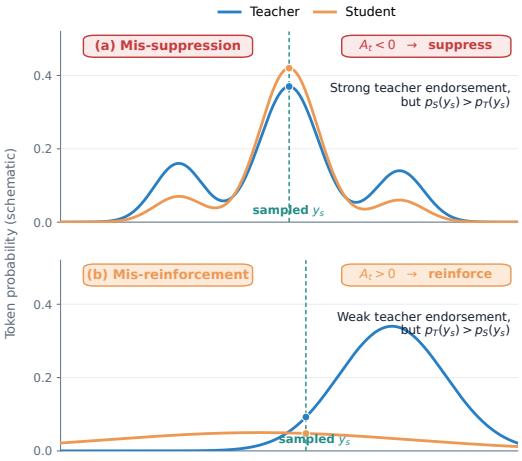

Figure 1: **Two failure cases of advantage-sign updates.** (a) A strongly teacher-endorsed token can be suppressed when $p_S(y_s) > p_T(y_s)$. (b) A weakly endorsed token can be reinforced when $p_T(y_s) > p_S(y_s)$.

The above analysis points to two complementary needs. First, we seek to increase the proportion of signals with large positive advantages in training trajectories; we call these *golden-gain* opportunities. Second, we seek to use teacher endorsement to validate the update direction proposed at each retained token. When the advantage indicates reinforcing a token, stronger teacher endorsement

makes that direction more reliable; when it indicates suppressing a token, weaker teacher endorsement makes that direction more reliable.

In this paper, we propose Uncertainty-Calibrated MOPD (UCMOPD) with two complementary mechanisms. Golden-Gain Enhancement uses dual-temperature sampling to generate a standard-temperature anchor and higher-temperature exploration responses, then retains exploration trajectories whose positive learning signal matches or exceeds the anchor. This increases exposure to strong positive advantages without accepting every exploratory trajectory. Inspired by Zhang et al. (2026b), Teacher-Endorsement Filtering introduces centered log-likelihood (CLL) to validate the reliability of token-level update directions. For each retained token, CLL assesses endorsement relative to the teacher's own uncertainty: tokens that the teacher considers more plausible provide stronger support for reinforcement, whereas less plausible tokens provide stronger support for suppression. It then probabilistically retains updates whose proposed directions are better supported by the teacher. Experiments on role-playing and medical specialization show that UCMOPD maintains vertical-domain performance. Relative to standard MOPD, it improves the average score across thirteen general-capability benchmarks by 4.48% in role-playing and 7.86% in medical specialization. Component ablations and diagnostic analyses further show that the proposed trajectory- and token-level mechanisms address their intended failure modes.

The contributions of this paper are as follows:

- We provide a comprehensive token-level analysis of MOPD signal utilization, showing that advantage sign alone is insufficient for deciding whether teacher feedback should be trusted.
- We design UCMOPD with two complementary mechanisms: Golden-Gain Enhancement for discovering and selecting stronger positive signals, and Teacher-Endorsement Filtering for validating token-level update directions.
- We validate the method in role-playing and medical-domain specialization, showing that it improves general-capability recovery over vanilla MOPD while preserving or improving vertical-domain performance.

2 Related Work

On-policy distillation. Knowledge distillation transfers predictive structure from teacher to student (Hinton et al., 2015), with sequence-level extensions for autoregressive generation (Kim & Rush, 2016). OPD instead queries the teacher on student-sampled trajectories to reduce exposure bias, sharing the motivation of interactive imitation learning (Ross et al., 2011). Representative objectives include generalized divergences (Agarwal et al., 2024), reverse-KL policy gradients (Gu et al., 2024), and skewed or contrastive losses (Ko et al., 2024; 2025); later work explores relaxed, selective, calibrated, prefix-based, curriculum-style, and related variants (Ko et al., 2026; Huang et al., 2025a; Zheng et al., 2026; Zhang et al., 2026a; Xu et al., 2026b; 2025; Xie et al., 2026; Kim & Baek, 2026; Jin et al., 2026; Yang et al., 2026a), as surveyed by Song & Zheng (2026). Unlike these studies, we ask whether an advantage-proposed update direction agrees with the teacher's own endorsement of the sampled token.

Multi-teacher distillation and capability recovery. Multi-teacher distillation aggregates complementary teachers or tasks (You et al., 2017; Fukuda et al., 2017). Recent LLM systems use on-policy or staged multi-teacher distillation for capability integration, medical adaptation, and reasoning recovery (Xiao et al., 2026; Dou et al., 2026; Yang et al., 2026b; Xu et al., 2026a). Most closely, Counteraction-Aware MOPD selects samples to reduce cross-domain gradient counteraction during capability recovery (Chen et al., 2026). We instead analyze MOPD at token level and select updates by consistency between their direction and entropy-calibrated teacher endorsement.

Vertical specialization and catastrophic forgetting. Vertical specialization can degrade broad abilities, closely relating to catastrophic forgetting (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Rolnick et al., 2019; Lesort, 2020). In LLMs, forgetting depends strongly on data composition, task overlap, and general-data injection (Luo et al., 2023; Dong et al., 2024; Bethune et al., 2025). Replay or self-synthesized rehearsal (Huang et al., 2024) and general-data injection or capability-integration recipes (Bethune et al., 2025; Liu et al., 2024; Dong et al.,

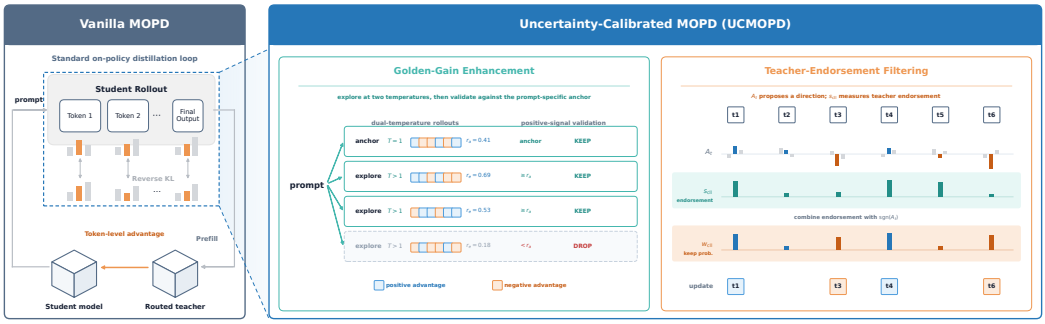

Figure 2: Overview of vanilla MOPD (left) and UCMOPD (right). Golden-Gain Enhancement combines dual-temperature rollouts with prompt-specific trajectory validation. Teacher-Endorsement Filtering distinguishes teacher endorsement s_{cll} from the update keep probability w_{cll}, which additionally incorporates the advantage sign.

2024) can mitigate it, but require suitable data or carefully curated mixtures. We instead assume that the original general training data are unavailable and recover capability from teacher supervision on student-sampled trajectories.

3 METHOD

In this section, we first review the MOPD objective and distinguish positive learning opportunities from the reliability of advantage-proposed update directions in Section 3.1. We then present Golden-Gain Enhancement, which combines high-temperature exploration with prompt-specific anchor validation to expose and retain stronger positive-advantage signals, in Section 3.2. Next, we introduce Teacher-Endorsement Filtering for validating token-level update directions in Section 3.3. Finally, we summarize the complete training procedure in Section 3.4.

3.1 MOPD BACKGROUND

We study Multi-Teacher On-Policy Distillation (MOPD) from frozen teacher policies to a student policy π_S. Each prompt is associated with a domain label d and is routed to the corresponding teacher $\pi_{T(d)}$; for example, role-playing prompts use the domain teacher, while general prompts use the general teacher. This gives a teacher distribution matched to the intended capability of each sampled trajectory without mixing teacher distributions at the token level. Following the on-policy distillation formulation in MOPD (Xiao et al., 2026), the student samples a response $y = (y_1, \ldots, y_T)$ from a rollout policy μ_S, while the training policy is denoted by π_S. MOPD uses reverse-KL distillation on student-sampled tokens and optimizes the surrogate

$$\mathcal{L}_{\text{MOPD}}(\pi_S) = -\mathbb{E}_{q,d,\, y\sim\mu_S}\left[\frac{1}{T}\sum_{t=1}^{T} A_t^{\text{distill}} \log \pi_S(y_t \mid q, y_{<t})\right],$$

where the stop-gradient MOPD distillation advantage is

$$A_t^{\text{distill}} = \text{sg}\left[\log \frac{\pi_{T(d)}(y_t \mid q, y_{<t})}{\pi_S(y_t \mid q, y_{<t})}\right].$$

Positive values indicate that the routed teacher assigns higher probability than the student to the sampled token, so reinforcing the token moves the student toward the teacher. Negative values indicate that the teacher assigns lower probability than the student, so a naive update suppresses the sampled token. For brevity, we write A_t for A_t^{distill} and π_T for the routed teacher $\pi_{T(d)}$ when the domain label is clear from context.

Token-level MOPD supervision raises two complementary questions. The first is how much the student can learn from a sampled token. For positive advantages, the magnitude of the teacher–student likelihood gap provides a continuous opportunity score: a large positive gap indicates that

the token is substantially under-modeled by the student relative to the teacher. We refer to such high-value positive signals as *golden-gain opportunities*. The second question is whether the update direction proposed by the advantage sign is reliable. This requires a teacher-internal endorsement estimate that does not depend on the student probability.

Advantage sign and teacher endorsement therefore play distinct roles. The sign proposes reinforcement or suppression, while endorsement determines whether that direction agrees with the teacher. Table 1 summarizes the resulting four conceptual cases. The strong/weak endorsement distinction is used only for exposition; Section 3.3 implements the same logic with a continuous keep probability.

Table 1: Direction–endorsement decomposition of token updates. CLL implements these retention tendencies continuously rather than assigning hard endorsement labels.

Advantage direction	Teacher endorsement	Retention tendency
$A_t > 0$	Strong	More likely to retain reinforcement
$A_t > 0$	Weak	More likely to filter reinforcement
$A_t < 0$	Strong	More likely to filter suppression
$A_t < 0$	Weak	More likely to retain suppression

3.2 Golden-Gain Enhancement

Standard on-policy sampling concentrates on continuations that the student already assigns high probability. This is appropriate for matching the deployment distribution, but it can repeatedly visit well-modeled regions and provide few opportunities with large positive teacher–student gaps. Increasing the sampling temperature is a natural way to explore less familiar continuations. However, a higher temperature only broadens the candidate pool; it does not guarantee that every resulting trajectory contains stronger positive-advantage signals. We therefore retain a standard-temperature response as a prompt-specific anchor and use it to determine whether each high-temperature exploration trajectory provides a genuine improvement in positive learning signal.

Concretely, for every prompt q, we sample one anchor response and m exploration responses,

$$y^a \sim \pi_S(\cdot \mid q; T_a), \qquad y_j^e \sim \pi_S(\cdot \mid q; T_e), \quad j = 1, \ldots, m,$$

where $T_a = 1.0$ and $T_e > T_a$. The anchor represents the student's ordinary sampling behavior. The exploration branch additionally uses top-p truncation, broadening token coverage while limiting the low-probability tail. To evaluate whether this exploration is effective, we compare each exploratory response against the anchor generated for the same prompt rather than applying a global threshold across prompts of different difficulty and advantage scales.

We score each response with its positive-advantage density,

$$r(y) = \frac{1}{\max(n_+, 1)} \sum_{t: A_t > 0} \left(\log \pi_T(y_t \mid q, y_{<t}) - \log \pi_S(y_t \mid q, y_{<t}) \right),$$

where n_+ is the number of positive-advantage tokens in the response; when $n_+ = 0$, the score is zero. This score measures the average strength of the response's positive teacher–student gaps. The anchor score $r(y^a)$ defines the prompt-specific threshold: the anchor is always retained, and an exploration response is retained only when

$$r(y_j^e) \geq r(y^a).$$

The comparison keeps exploratory responses whose positive learning signal is at least as strong as the student's ordinary response for the same prompt. In this way, Golden-Gain Enhancement uses higher-temperature sampling to search for golden-gain opportunities and the anchor-based criterion to ensure that only effective exploration contributes to training.

3.3 Teacher-Endorsement Filtering

Our goal is to use teacher endorsement to calibrate the update direction proposed at each retained token. An intuitive approach is to construct the teacher's top-k token set: a student-sampled token

inside this set is treated as endorsed, whereas one outside it is treated as unendorsed. However, this decision is sensitive to k. A small k may exclude plausible alternatives when the teacher distribution is broad, whereas a large k may include weakly supported tokens when it is sharp.

Inspired by Zhang et al. (2026b), we instead use centered log-likelihood (CLL) to measure teacher endorsement; its distribution-discrimination interpretation traces back to Cox's tests of separate families of hypotheses (Cox, 1961). For a sampled token $x = y_t$, CLL centers its teacher log-likelihood by the teacher entropy,

$$c_{\mathrm{cll}}(x) = \log p_T(x) + H[p_T] = \log p_T(x) - \mathbb{E}_{z\sim p_T}[\log p_T(z)].$$

Because $-H[p_T] = \mathbb{E}_{z\sim p_T}[\log p_T(z)]$, $c_{\mathrm{cll}}(x)$ compares the sampled token's log-probability with the teacher's probability-weighted mean log-probability over the vocabulary. In particular,

$$c_{\mathrm{cll}}(x) \geq 0 \quad \Longleftrightarrow \quad p_T(x) \geq \exp(-H[p_T])\,.$$

Thus the reference probability adapts to the shape of the teacher distribution rather than relying on a fixed likelihood or rank threshold. We convert CLL into a bounded teacher-endorsement score,

$$s_{\mathrm{cll}}(x) = \exp\left(\min\left(\log p_T(x) + H[p_T], 0\right)\right),$$

where $p_T(x) = \pi_T(x \mid q, y_{<t})$ and $H[p_T]$ is the teacher entropy estimate. A token satisfying $p_T(x) \geq \exp(-H[p_T])$ receives maximal endorsement, whereas a token below this entropy-induced typical-probability scale receives proportionally weaker endorsement. This score is independent of advantage sign.

Advantage sign enters only when endorsement is converted into update reliability. Let $d_t = \mathrm{sgn}(A_t) \in \{-1, +1\}$. For every nonzero-advantage token, we define a single keep probability

$$w_{\mathrm{cll}}(A_t, x) = \mathbf{1}[A_t > 0]s_{\mathrm{cll}}(x) + \mathbf{1}[A_t < 0]\big(1 - s_{\mathrm{cll}}(x)\big) = \frac{1 + d_t\big(2s_{\mathrm{cll}}(x) - 1\big)}{2}.$$

Equivalently, its two directional instances are

$$w^{+}_{\mathrm{cll}}(x) = s_{\mathrm{cll}}(x) = \exp(\min\left(\log p_T(x) + H[p_T], 0\right)),$$
$$w^{-}_{\mathrm{cll}}(x) = 1 - s_{\mathrm{cll}}(x) = 1 - \exp(\min\left(\log p_T(x) + H[p_T], 0\right)).$$

These are not separate endorsement criteria. CLL always computes the same $s_{\mathrm{cll}}(x)$; the sign merely specifies whether the proposed update increases or decreases the token probability. Strong endorsement agrees with reinforcement and conflicts with suppression, while weak endorsement induces the opposite retention preference.

During training, every nonzero-advantage token receives a Bernoulli mask sampled with probability $w_{\mathrm{cll}}(A_t, y_t)$. Updates whose direction agrees with teacher endorsement therefore remain more often, while endorsement-inconsistent updates are removed more often. No hard endorsement threshold is introduced into the executable rule. We discuss an alternative weighting variant, which multiplies rather than samples the same keep probability, in Appendix B.3.

The two mechanisms also admit simple local guarantees: Golden-Gain Enhancement cannot retain an exploration trajectory whose positive-advantage density is below the prompt-specific anchor, while Teacher-Endorsement Filtering implements entropy-adaptive direction–endorsement consistency. Appendix C provides the corresponding derivations.

3.4 Overall Training Procedure

The resulting training step applies the two mechanisms in sequence. Golden-Gain Enhancement samples anchor and exploratory trajectories and retains those whose positive-advantage density at least matches the prompt-specific anchor baseline. Teacher-Endorsement Filtering then applies CLL to every nonzero-advantage token in the retained trajectories, validating individual update directions against teacher-internal endorsement. This preserves the teacher-guided MOPD objective while making its supervision more selective at both sequence and token levels. Algorithm 1 summarizes the complete procedure.

Algorithm 1 UCMOPD

Require: Student policy π_S, teacher policies $\{\pi_T\}$, prompt batch $\mathcal{Q}$, anchor temperature $T_a = 1.0$, exploration temperature T_e
Ensure: Updated student policy π_S

for each prompt $q \in \mathcal{Q}$ **do**
  Sample anchor response $y^a \sim \pi_S(\cdot \mid q; T_a)$
  Sample exploration responses $\{y_j^e\}_{j=1}^m \sim \pi_S(\cdot \mid q; T_e)$
  $\mathcal{Y}_q \leftarrow \{y^a\} \cup \{y_j^e\}_{j=1}^m$
  for each response $y \in \mathcal{Y}_q$ **do**
    Compute student and teacher log probabilities for each token
    Compute token advantages $A_t = \log \pi_T(y_t \mid q, y_{<t}) - \log \pi_S(y_t \mid q, y_{<t})$
    Compute positive-advantage density $r(y)$ over tokens with $A_t > 0$
  end for
  Keep y^a and keep each exploration response y_j^e only if $r(y_j^e) \geq r(y^a)$
  for each retained response y **do**
    for each token position t with $A_t \neq 0$ **do**
      Estimate teacher entropy $H_T(t)$ at position t
      $s_t \leftarrow \exp(\min(\log \pi_T(y_t \mid q, y_{<t}) + H_T(t), 0))$
      $w_t \leftarrow \mathbf{1}[A_t > 0]s_t + \mathbf{1}[A_t < 0](1 - s_t)$
      Keep token update t with probability w_t
    end for
  end for
end for
Update π_S with the MOPD loss on retained responses and retained tokens

4 Experiments

4.1 Experimental Setup

We evaluate domain specialization recovery in two vertical settings: role-playing and medical-domain adaptation. In each setting, the student is initialized from the supervised domain-specialized model, and MOPD uses a domain teacher and a general teacher to provide token-level on-policy supervision. We compare against the original general model, the supervised fine-tuned (SFT) domain model, vanilla MOPD, and three additional distillation baselines: SelecTKD (Huang et al., 2025a), ReOPOLD (Ko et al., 2026), and CaMOPD (Chen et al., 2026). Our dual-temperature rollout uses one anchor response at temperature 1.0 in both domains. The role-playing setting uses seven exploration responses at temperature 1.5, while the medical setting uses three exploration responses at temperature 1.5. Appendix A provides the complete checkpoints, teacher–student scale relationship, routing rules, prompt sources, precision, and optimization configuration.

We report both general capabilities and vertical-domain performance. General capability is measured by MMLU-Redux (Gema et al., 2024), MMLU-Pro (Wang et al., 2024), GPQA-Diamond (Rein et al., 2023), AIME25 and HMMT25 (Balunovic et al., 2025), ZebraLogic (Lin et al., 2025), LiveCodeBench v5 and v6 (Jain et al., 2024), IF-Eval (Zhou et al., 2023), IFBench (Pyatkin et al., 2025), WritingBench (Wu et al., 2025b), Arena-Hard v2 (Li et al., 2024), and LiveBench 1125 (White et al., 2025). Role-playing ability is measured by storyline consistency, anthropomorphism, character fidelity, and storyline quality following the CoSER evaluation setting (Wang et al., 2026). Medical-domain ability is measured by MedQA-USMLE (Jin et al., 2021), MedXpertQA Text (Zuo et al., 2025), and PubMedQA (Jin et al., 2019). Unless otherwise stated, higher is better for all metrics.

4.2 Main Results

Table 2 and Figure 3 summarize the role-playing results. SFT substantially improves role-playing ability, raising the vertical average from 30.92 to 41.08, but it causes a large drop in the general average from 56.93 to 41.11. Vanilla MOPD recovers much of the general ability (52.09) but barely improves the role-playing average over SFT (41.22). UCMOPD obtains a strong trade-off among the compared methods: it reaches the best general average of 54.43, slightly above CaMOPD's 54.27, and the best role-playing average (45.00, tied with CaMOPD). Compared with vanilla MOPD, UCMOPD improves the general average by $+2.33$ points and the role-playing average by $+3.78$ points.

The medical-domain results show a complementary pattern. Medical SFT improves the medical average from 56.48 to 60.96, but reduces the general average from 59.48 to 49.97. Vanilla MOPD recovers only part of the general capability (51.81). In contrast, UCMOPD reaches a general average of 55.88, improving over vanilla MOPD by +4.07 points while maintaining a medical average of 60.95, essentially matching SFT (60.96) and exceeding CaMOPD (59.40). This indicates that the proposed selection mechanism is especially helpful when the domain-specialized model suffers a large general-capability drop.

Table 2: Main specialization results for role-playing (left) and medical-domain (right) specialization. Base and SFT denote the pre-specialization and supervised domain-specialized models; the remaining columns are general-capability recovery methods. Bold marks the best recovery-method result per metric. "Gen. Avg." averages general benchmarks, while "Vertical Avg." averages the listed vertical benchmarks.

Metric	*Role-playing specialization*							*Medical-domain specialization*						
	Base	SFT	MOPD	SelecTKD	ReOPOLD	CaMOPD	UCMOPD	Base	SFT	MOPD	SelecTKD	ReOPOLD	CaMOPD	UCMOPD
MMLU-Redux	82.39	75.05	78.86	78.60	78.65	79.07	**80.56**	84.67	85.77	86.21	**86.30**	86.21	86.21	86.11
MMLU-Pro	68.33	54.45	62.20	61.60	62.36	64.31	**65.72**	73.08	74.80	74.53	**74.67**	74.14	74.53	74.04
GPQA-Diamond	61.49	54.29	59.60	60.98	63.51	62.63	**63.64**	62.50	62.37	60.10	58.46	61.49	**61.99**	61.74
AIME25	45.21	40.62	**47.29**	45.00	43.12	45.21	43.54	65.00	41.67	54.37	53.75	52.08	54.37	**56.25**
ZebraLogic	79.50	31.50	71.10	70.60	71.40	**76.30**	76.20	84.80	71.60	72.60	72.10	73.20	72.60	**73.70**
HMMT25	33.12	26.67	31.04	31.04	31.25	31.04	**32.08**	45.00	28.33	33.96	33.54	32.92	34.58	**35.62**
LiveCodeBench v5	34.76	25.09	35.12	34.41	32.26	37.28	**39.78**	58.42	48.75	51.25	54.48	53.76	**56.63**	54.84
LiveCodeBench v6	20.57	16.00	**25.14**	24.00	22.86	24.00	**25.14**	31.43	27.43	24.00	**29.71**	29.14	**29.71**	27.43
IF-Eval	83.18	67.65	78.00	78.93	79.67	80.41	**80.96**	84.66	48.98	48.06	49.17	48.61	59.89	**71.53**
IFBench	30.95	23.81	26.53	26.53	26.19	26.87	**27.55**	21.09	22.45	**22.11**	20.41	21.09	**22.11**	21.09
WritingBench	83.80	69.58	80.30	80.53	80.39	81.64	**81.71**	79.87	79.42	79.62	79.78	**79.84**	79.62	79.73
Arena-Hard v2	51.15	6.95	22.60	22.00	23.70	**33.25**	29.35	30.65	23.65	23.90	25.40	25.15	24.10	**26.05**
LiveBench 1125	65.60	42.80	59.40	58.90	59.00	**63.50**	61.30	52.10	34.40	42.80	41.80	40.70	49.00	**58.30**
Storyline Consistency	34.74	42.27	43.57	43.87	45.42	**47.81**	47.39					–		
Anthropomorphism	23.09	33.42	34.91	32.28	33.35	33.96	**35.96**					–		
Character Fidelity	23.89	38.75	41.45	39.46	41.53	**44.65**	42.98					–		
Storyline Quality	41.98	49.89	51.50	51.54	52.84	53.47	**53.67**					–		
MedQA-USMLE				–				79.03	86.17	86.02	85.78	85.94	85.78	**86.10**
MedXpertQA Text				–				17.22	22.90	23.55	23.51	**24.12**	23.63	23.06
PubMedQA				–				73.19	73.81	72.77	73.41	54.07	68.79	**73.69**
Gen. Avg.	56.93	41.11	52.09	51.78	51.87	54.27	**54.43**	59.48	49.97	51.81	52.27	52.18	54.26	**55.88**
Vertical Avg.	30.92	41.08	41.22	43.36	43.28	**45.00**	**45.00**	56.48	60.96	60.78	60.90	54.71	59.40	**60.95**

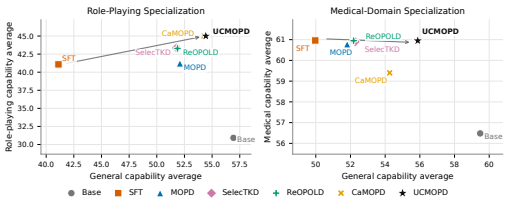

Figure 3: Domain–general trade-off in role-playing and medical specialization. UCMOPD improves general recovery while preserving vertical-domain ability.

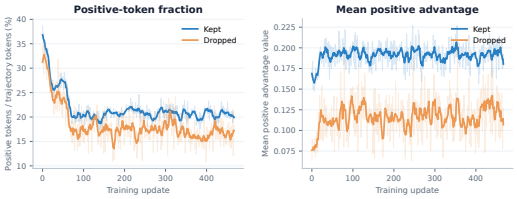

Figure 4: Training-update diagnostics of Golden-Gain Enhancement before Teacher-Endorsement Filtering. Retained trajectories exhibit a higher positive-token fraction and a higher mean positive advantage.

4.3 Ablations

Unless otherwise specified, ablations and diagnostics use the role-playing specialization setting. Table 3 starts from full UCMOPD and cumulatively removes positive-advantage-density filtering, dual-temperature sampling, and Teacher-Endorsement Filtering. Removing either Golden-Gain operation weakens the overall trade-off, while removing Teacher-Endorsement Filtering further lowers both averages.

Table 3: Cumulative component removal from UCMOPD on role-playing specialization.

Method	Gen. Avg.	Vertical Avg.	IF-Eval	GPQA	Zebra	Arena-Hard v2
UCMOPD	**54.43**	**45.00**	**80.96**	**63.64**	**76.20**	**29.35**
− Positive-advantage-density filtering	53.32	43.24	80.04	63.13	74.30	28.45
− Dual-temperature sampling	52.17	43.36	77.63	62.63	72.00	23.70
− Teacher-Endorsement Filtering = MOPD	52.09	41.22	78.00	59.60	71.10	22.60

Table 4: Token-endorsement criterion comparison. Both criteria retain updates whose direction agrees with estimated teacher endorsement; CLL replaces hard top-k membership with a continuous entropy-centered score.

Criterion	Gen. Avg.	Vertical Avg.	AIME25	GPQA	Zebra	Arena-Hard v2
Hard top-k endorsement ($k = 3$)	51.95	42.31	43.33	60.73	**74.40**	23.20
Hard top-k endorsement ($k = 1$)	52.01	41.32	42.71	**63.26**	71.90	22.65
CLL sample mask	**52.17**	**43.36**	**44.58**	62.63	72.00	**23.70**

To verify that the improvement is not simply due to more responses per prompt, we compare against a rollout-budget-matched vanilla MOPD baseline. Both methods use eight responses per prompt, and UCMOPD achieves higher general and role-playing averages; Appendix Table 7 reports the full results.

We use an exploration temperature of $T = 1.5$ because it yields the strongest role-playing average among the tested temperatures; the full sensitivity study is in Appendix B.6.

4.4 Analysis of the Proposed Mechanisms

Golden-Gain Enhancement. Ordinary on-policy sampling rarely exposes large positive teacher–student gaps. Golden-Gain Enhancement broadens the candidate pool through dual-temperature sampling and retains exploration trajectories only when their positive-advantage density matches or exceeds the prompt-specific anchor. In Table 3, removing trajectory filtering reduces the general average from 54.43 to 53.32 and the vertical average from 45.00 to 43.24; further removing dual-temperature sampling reduces the general average to 52.17. Appendix Table 7 rules out rollout count as the explanation. Figure 4 further shows that retained trajectories consistently contain stronger positive signals: their aggregate positive-token fraction and mean positive advantage are 0.2168 and 0.1911, compared with 0.1850 and 0.1158 for rejected trajectories (Appendix Table 8). Although only positive-advantage magnitude directly determines selection, the higher positive-token fraction indicates an empirical enrichment: trajectories with stronger positive signals also tend to contain such signals more frequently. Thus, Golden-Gain Enhancement exposes more candidates and concentrates training on stronger positive learning opportunities.

Teacher-Endorsement Filtering. Advantage sign compares teacher and student probabilities but does not measure the teacher's own endorsement, so a strongly endorsed token may be suppressed or a weakly endorsed token reinforced, as Figure 1 illustrates. Teacher-Endorsement Filtering therefore uses CLL to evaluate each retained token according to direction–endorsement consistency. Removing this mechanism from the CLL-only configuration reduces the general average from 52.17 to 52.09 and the vertical average from 43.36 to 41.22 (Table 3). Table 4 also shows that entropy-centered CLL gives the best general and vertical averages among the tested endorsement criteria, outperforming fixed top-k endorsement, which ignores distribution shape. Together, the results confirm the intended division of labor: Golden-Gain Enhancement discovers stronger positive opportunities, while Teacher-Endorsement Filtering validates the proposed token-level update directions. Additional direction–endorsement diagnostics and realized CLL keep rates are reported in Appendix B.4.

5 Conclusion

We study general-capability recovery after vertical-domain specialization as a signal-selection problem in Multi-Teacher On-Policy Distillation. UCMOPD uses Golden-Gain Enhancement to expose and select trajectories with stronger positive learning signals and Teacher-Endorsement Filtering to validate token-level update directions. Across role-playing and medical specialization, UCMOPD improves the general-capability average over standard MOPD by 4.48% and 7.86%, respectively, while maintaining vertical-domain performance. Ablations and diagnostic analyses further show that these gains do not merely arise from a larger rollout budget and that the two mechanisms address their intended failure modes. These results support uncertainty-aware signal discovery and validation as an effective approach to preserving general capabilities during domain specialization.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation, supporting qualitative and thematic data analysis, interpreting results. We have not used generative AI tools for generating synthetic data sets, helping to develop theoretical models or conceptual frameworks, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting dataset. Formulating mathematical claims, providing critical ingredients for proving mathematical claims are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, creating artifacts, sourcing/searching for information, editing a research paper to improve readability. We have reviewed all AI-assisted work. AI-assisted text and analyses were checked against the source materials and experimental results, and AI-assisted code was manually reviewed and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available language-model checkpoints and benchmark datasets. It does not involve human-subject experiments, collection of personal information, or interaction with identifiable individuals. The role-playing and medical-domain evaluations use existing benchmark prompts and follow their intended evaluation protocols. As with other language-model research, generated outputs may contain factual errors or undesirable content; they should not be treated as professional medical advice or as a substitute for human review.

REPRODUCIBILITY STATEMENT

We provide the algorithm, training objective, filtering rules, model roles, prompt-mixture construction, rollout temperatures, precision settings, and optimization configuration in the main paper and Appendix A. The benchmark definitions and citations, complete per-benchmark results, ablations, and diagnostic statistics are also reported in the paper or its appendices. These details are intended to make the reported comparisons and training procedure reproducible using the cited public checkpoints and datasets.

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.

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanovic, and Martin Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*, 2025. URL https://arxiv.org/abs/2505.23281.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, et al. Llama-Nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*, 2025. URL https://arxiv.org/abs/2505.00949.

Louis Bethune, David Grangier, Dan Busbridge, Eleonora Gualdoni, Marco Cuturi, and Pierre Ablin. Scaling laws for forgetting during finetuning with pretraining data injection. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.06042.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. HuatuoGPT-o1: Towards medical complex reasoning with LLMs. *arXiv preprint arXiv:2412.18925*, 2024a. URL https://arxiv.org/abs/2412.18925.

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*, 2026. URL https://arxiv.org/abs/2605.27115.

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*, 2024b. URL https://arxiv.org/abs/2405.01769.

David Roxbee Cox. Tests of separate families of hypotheses. 1961.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. How abilities in large language models are affected by supervised fine-tuning data composition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2310.05492.

Chengfeng Dou, Fan Yang, Fei Li, Jiyuan Jia, Qiang Ju, Shuai Wang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Hongda Zhang, et al. Baichuan-M3: Modeling clinical inquiry for reliable medical decision-making. *arXiv preprint arXiv:2602.06570*, 2026. URL https://arxiv.org/abs/2602.06570.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pp. 3697–3701, 2017.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. Are we done with MMLU? *arXiv preprint arXiv:2406.04127*, 2024. URL https://arxiv.org/abs/2406.04127.

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

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

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. SelecTKD: Selective token-weighted knowledge distillation for LLMs. *arXiv preprint arXiv:2510.24021*, 2025a. URL https://arxiv.org/abs/2510.24021.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1416–1428, 2024. URL https://aclanthology.org/2024.acl-long.77/.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *arXiv preprint arXiv:2504.00869*, 2025b. URL https://arxiv.org/abs/2504.00869.

Intelligent Internet. Ii-medical-7b-preview: Medical reasoning model, 2025. URL https://huggingface.co/Intelligent-Internet/II-Medical-7B-Preview.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024. URL https://arxiv.org/abs/2403.07974.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. URL https://www.mdpi.com/2076-3417/11/14/6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2567–2577, 2019. URL https://aclanthology.org/D19-1259/.

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

Minsang Kim and Seung Jun Baek. Explain in your own words: Improving reasoning via token-selective dual knowledge distillation. *arXiv preprint arXiv:2603.13260*, 2026. URL https://arxiv.org/abs/2603.13260.

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, 2016. URL https://arxiv.org/abs/1606.07947.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. URL https://arxiv.org/abs/1612.00796.

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*, 2024. URL https://arxiv.org/abs/2402.03898.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2503.07067.

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

Timothée Lesort. Continual learning: Tackling catastrophic forgetting in deep neural networks with replay processes. *arXiv preprint arXiv:2007.00487*, 2020.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024. URL https://arxiv.org/abs/2406.11939.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. ZebraLogic: On the scaling limits of LLMs for logical reasoning. *arXiv preprint arXiv:2502.01100*, 2025. URL https://arxiv.org/abs/2502.01100.

Chengyuan Liu, Yangyang Kang, Shihang Wang, Lizhi Qing, Fubang Zhao, Chao Wu, Changlong Sun, Kun Kuang, and Fei Wu. More than catastrophic forgetting: Integrating general capabilities for domain-specific llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7531–7548, 2024.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Wei Lu, Rachel K. Luu, and Markus J. Buehler. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities. *arXiv preprint arXiv:2409.03444*, 2024. URL https://arxiv.org/abs/2409.03444.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023. URL https://arxiv.org/abs/2308.08747.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174, pp. 248–260, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hanna Hajishirzi. Generalizing verifiable instruction following. In *Advances in Neural Information Processing Systems 38*, pp. 54870–54896, 2025. doi: 10.52202/085713-1645. URL https://arxiv.org/abs/2507.02833.

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. URL https://arxiv.org/abs/2311.12022.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011. URL https://arxiv.org/abs/1011.0686.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026. URL https://arxiv.org/abs/2604.00626.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. Coser: A comprehensive literary dataset and framework for training and evaluating llm role-playing and persona simulation, 2026. URL https://arxiv.org/abs/2502.09082.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuntian Ni, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024. URL https://arxiv.org/abs/2406.01574.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.19314.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. MedReason: Eliciting factual medical reasoning steps in LLMs via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025a. URL https://arxiv.org/abs/2504.00993.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. WritingBench: A comprehensive benchmark for generative writing. *arXiv preprint arXiv:2503.05244*, 2025b. URL https://arxiv.org/abs/2503.05244.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. MiMo-V2-Flash technical report. *arXiv preprint arXiv:2601.02780*, 2026. URL https://arxiv.org/abs/2601.02780.

Xurong Xie, Zhucun Xue, Jiafu Wu, Jian Li, Yabiao Wang, Xiaobin Hu, Yong Liu, and Jiangning Zhang. LLM-oriented token-adaptive knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2510.11615. Also available as arXiv:2510.11615.

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

Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.11325.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, and Zhipeng Wang. PACED: Distillation and on-policy self-distillation at the frontier of student competence. *arXiv preprint arXiv:2603.11178*, 2026b. URL https://arxiv.org/abs/2603.11178.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Chenghan Yang, Ruiyu Zhao, Yang Liu, and Ling Jiang. Survey of specialized large language model. *arXiv preprint arXiv:2508.19667*, 2025b. URL https://arxiv.org/abs/2508.19667.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026a. URL https://arxiv.org/abs/2602.12125.

Zhuolin Yang, Zihan Liu, Yang Chen, Wenliang Dai, Boxin Wang, Sheng-Chieh Lin, Chankyu Lee, Yangyi Chen, Dongfu Jiang, Jiafan He, et al. Nemotron-Cascade 2: Post-training LLMs with cascade RL and multi-domain on-policy distillation. *arXiv preprint arXiv:2603.19220*, 2026b. URL https://arxiv.org/abs/2603.19220.

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, 2017.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler, Qian Qian, Gregory D. Lyng, Sanjit Singh Batra, and Robert E. Tillman. Fast and effective on-policy distillation from reasoning prefixes. *arXiv preprint arXiv:2602.15260*, 2026a. URL https://arxiv.org/abs/2602.15260.

Miaosen Zhang, Yishan Liu, Shuxia Lin, Xu Yang, Qi Dai, Chong Luo, Weihao Jiang, Peng Hou, Anxiang Zeng, Xin Geng, and Baining Guo. Towards on-policy sft: Distribution discriminant theory and its applications in llm training. *arXiv preprint arXiv:2602.12222*, 2026b. URL https://arxiv.org/abs/2602.12222.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. SCOPE: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting. *arXiv preprint arXiv:2604.10688*, 2026. URL https://arxiv.org/abs/2604.10688.

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. URL https://arxiv.org/abs/2311.07911.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025. URL https://arxiv.org/abs/2501.18362.

A Detailed Experimental Setup

A.1 Models and Systems

For role-playing, a CoSER-fine-tuned Qwen3-4B-Instruct-2507 checkpoint (Yang et al., 2025a; Wang et al., 2026) initializes the student and serves as the frozen domain teacher; the original checkpoint serves as the general teacher. Medical adaptation uses II-Medical-7B-Preview (Internet, 2025) for student initialization and the domain teacher, with Qwen3-8B as the general teacher. Table 5 summarizes these model roles.

Both settings therefore use *homogeneous, same-scale* MOPD: the student and both teachers share the same Qwen3 architecture and parameter count. The teachers differ in specialization rather than scale, and only the student parameters are updated.

Table 5: Model checkpoints and precision. "Same as student initialization" means that the domain teacher is a frozen copy of the SFT checkpoint from which the trainable student is initialized.

Setting	Role	Checkpoint	Size	Training dtype	Inference dtype
Role-playing	Student	Our CoSER SFT of Qwen3-4B-Instruct-2507	4B	BF16	BF16
Role-playing	Domain teacher	Same as student initialization (frozen)	4B	–	BF16
Role-playing	General teacher	Qwen3-4B-Instruct-2507 (frozen)	4B	–	BF16
Medical	Student	II-Medical-7B-Preview	8B	BF16	BF16
Medical	Domain teacher	Same as student initialization (frozen)	8B	–	BF16
Medical	General teacher	Qwen3-8B (frozen)	8B	–	BF16

Student optimization uses Megatron-Core with BF16 parameters, while student rollout generation and both teachers' log-probability inference use asynchronous SGLang with BF16 weights; no model is quantized. Actor training uses pipeline parallelism of 2 for role-playing and 4 for the larger medical model. Student and teacher rollout inference use tensor parallelism of 1. Both settings run on 4 nodes with 8 NVIDIA H800 GPUs per node.

A.2 Data Construction and Routing

Each recovery corpus concatenates a vertical-domain prompt set with a general prompt set and is globally shuffled before training. The role-playing corpus contains 10,000 CoSER prompts and 10,000 general prompts. The medical corpus contains 9,926 medical prompts and the same 10,000 general prompts. The two mixtures are therefore exactly 1:1 and approximately 1:1, respectively.

Role-playing data. We rank the CoSER conversations (Wang et al., 2026) by their number of dialogue turns and reserve the 10,000 longest conversations as unseen role-playing prompts for MOPD. For each reserved conversation, the first $n-1$ turns form the prompt and context, while the final continuation is generated on-policy by the current student. These reserved conversations are never used for role-playing SFT. The remaining approximately 300,000 conversations are used to fine-tune Qwen3-4B-Instruct-2507 and obtain the role-playing initialization and domain teacher.

General recovery data. We draw the general prompts from the chat, code, math, and STEM subsets of the Nemotron Post-Training Dataset v1 (Bercovich et al., 2025). We sample 10,000 prompts while preserving the relative proportions of these four selected subsets. MOPD uses only the prompt and conversational context; reference responses from the source dataset are not included in the distillation loss.

Medical data. We construct the medical prompt pool from four public sources: 4,000 prompts from MedMCQA (Pal et al., 2022), 3,000 from MedReason (Wu et al., 2025a), 2,000 English prompts from Medical-R1-Distill-Data (Chen et al., 2024a), and 1,000 from m23k-tokenized

(Huang et al., 2025b). After preprocessing, 9,926 valid prompts remain. This stream is prompt-only: gold answers and explanations are not provided during MOPD training. The current student generates on-policy responses, and the medical teacher supplies token-level log-probability feedback on those responses.

Overlap control and teacher routing. We verify that none of the role-playing, general, or medical training prompts overlaps with any instance in the evaluation sets used in this paper. Teacher routing is deterministic and defined at the prompt level. Each example carries a source label that is preserved through rollout generation. Role-playing labels `coser` and `general` route to the CoSER SFT teacher and Qwen3-4B-Instruct-2507 teacher, respectively; medical labels `medical` and `general` route to the medical SFT teacher and Qwen3-8B teacher, respectively. All anchor and exploration responses sampled from one prompt inherit the same label and are scored by the same routed teacher at every token. We do not average, interpolate, or otherwise mix teacher distributions within a trajectory.

A.3 Training Configuration

Table 6 reports the optimization and sampling configuration. A "prompt batch" counts distinct prompts before rollout replication. The role-playing setting samples eight responses per prompt (one anchor and seven exploration responses), whereas the medical setting samples four (one anchor and three exploration responses); both therefore process 1,024 candidate responses per outer update before positive-advantage-density filtering. The anchor uses temperature 1.0 and top-$p = 1.0$, and exploration uses temperature 1.5 and top-$p = 0.9$. To reduce memory and communication overhead, the teacher entropy used by CLL is estimated from the teacher's top-50 log-probabilities, $\widehat{H}_T(t) = -\sum_{i=1}^{50} p_T^{(i)}(t) \log p_T^{(i)}(t)$, rather than from the full vocabulary; this top-k approximation avoids materializing and transferring the full distribution. We train for three passes over each shuffled prompt mixture. Because the dataloader drops the last incomplete batch, this gives $\lfloor 20000/128 \rfloor \times 3 = 468$ outer updates for role-playing and $\lfloor 19926/256 \rfloor \times 3 = 231$ for medical adaptation.

Table 6: Training and rollout configuration. Batch sizes are global prompt counts before response replication.

Configuration	Role-playing	Medical
Domain prompts	10,000	9,926
General prompts	10,000	10,000
Global prompt batch size	128	256
Responses per prompt	8 (1+7)	4 (1+3)
Candidate responses/update	1,024	1,024
Training epochs	3	3
Outer training updates	468	231
Actor updates per batch	1	1
Learning rate	2×10^{-6}	2×10^{-6}
Gradient clipping	1.0	1.0
Prompt/response limit	4,096/32,768	4,096/32,768

The two settings use different thinking configurations. Role-playing uses the non-thinking configuration of Qwen3-4B-Instruct-2507, so the student and teacher rollouts do not contain explicit `<think>` traces. Medical specialization uses the native thinking configuration of II-Medical-7B-Preview and Qwen3-8B; both are run with thinking enabled, and the resulting rollouts can contain explicit `<think>` reasoning. Thus, the two experiments evaluate UCMOPD without and with thinking traces, respectively. The consistent improvements across these settings suggest that UCMOPD is effective under both generation regimes, although we do not conduct a controlled comparison between thinking and no-thinking modes.

B Additional Experiments

B.1 Rollout-Budget Control

Table 7 compares vanilla MOPD and UCMOPD under the same eight-response rollout budget.

Table 7: Rollout-budget control with eight sampled responses per prompt.

Method	Gen. Avg.	Vertical Avg.	IF-Eval	GPQA	Zebra	Arena-Hard v2
MOPD, 8 rollouts	53.31	43.38	80.22	60.98	72.90	26.90
UCMOPD, 8 rollouts	**54.43**	**45.00**	**80.96**	**63.64**	**76.20**	**29.35**

B.2 Golden-Gain Enhancement: Aggregate Diagnostics

Table 8 summarizes the diagnostics visualized across training in Figure 4. The filter retains 88.6% of exploratory trajectories in this diagnostic run. Retained trajectories have both a larger positive-token fraction and a higher mean positive advantage before Teacher-Endorsement Filtering is applied.

Table 8: Aggregate diagnostics of Golden-Gain Enhancement before Teacher-Endorsement Filtering.

Subset	Positive-token fraction ↑	Mean positive advantage ↑
Kept trajectories	**0.2168**	**0.1911**
Dropped trajectories	0.1850	0.1158
Kept − dropped	0.0318	0.0752

B.3 Teacher-Endorsement Filtering: Sample Masking versus Weighting

The main method uses CLL in sample-mask mode: every nonzero-advantage token update is stochastically kept or dropped according to the direction-consistency probability $w_{\mathrm{cll}}(A_t, y_t)$. We also evaluate a weighting variant that keeps the full token set but multiplies each update by the same probability. Table 9 shows that both variants improve over vanilla MOPD on the general average, but sample masking gives a stronger role-playing average and substantially better GPQA than weighting. Weighting gives a slightly higher general average and slightly better scores on some individual benchmarks in this early CLL-only comparison, but continues to apply reduced updates where masking removes unreliable signals entirely. We therefore use sample masking as the default CLL mechanism in the main method.

Table 9: Implementation comparison for Teacher-Endorsement Filtering without Golden-Gain Enhancement.

Method	Gen. Avg.	Vertical Avg.	WritingBench	GPQA	Zebra	LCB v5
MOPD	52.09	41.22	80.30	59.60	71.10	**35.12**
CLL sample mask	52.17	**43.36**	80.62	**62.63**	72.00	34.76
CLL weighting	**52.34**	42.92	**80.71**	59.09	**72.30**	31.90

B.4 Teacher-Endorsement Diagnostics

We additionally quantify potential direction–endorsement conflict over valid nonzero-advantage tokens in the role-playing diagnostic data. Among negative-advantage tokens, 67.75% coincide with the teacher's top-1 prediction and 76.68% have $s_{\mathrm{cll}} \geq 0.5$, indicating substantial potential over-suppression; among positive-advantage tokens, 12.16% have $s_{\mathrm{cll}} < 0.5$, indicating potential over-reinforcement. The top-1 and CLL-threshold rates are descriptive proxies for direction–endorsement conflict rather than ground-truth error labels. The executable filter applies the continuous keep probability w_{cll} to both positive- and negative-advantage updates rather than using either binary criterion.

The score-based sampling statistics further confirm the expected behavior of the stochastic filter. Table 10 reports candidate-token-weighted expected keep rates and the corresponding realized Bernoulli keep rates. Their close agreement shows that the stochastic sample-mask implementation matches its CLL-weighted expectation at the observed token count.

Table 10: Expected and realized keep rates under signed-scope CLL sample masking. Expected rates are candidate-token-weighted means of w_{cll}; realized rates are computed from the sampled Bernoulli masks.

Candidate subset	Candidate tokens	Expected keep (%)	Realized keep (%)
Negative advantage	299,307,540	9.88	9.88
Positive advantage	115,874,361	81.88	81.88
All signed candidates	415,181,901	29.97	29.97

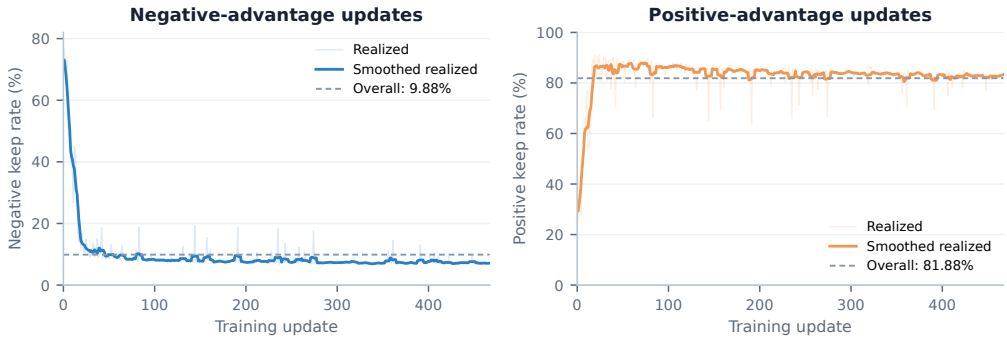

Figure 5: Per-update realized CLL retention rates under signed-scope sample masking. The left and right panels show negative- and positive-advantage updates, respectively. Faint traces show the raw rates, solid curves show rolling averages, and dashed lines mark the corresponding overall token-weighted rates reported in Table 10.

Figure 5 directly explains the apparently low negative-branch rate and the asymmetry between the two branches. The per-update realized keep rate is high in the first few updates, then falls and remains near 7–8% for most later updates, while the dashed line marks the token-weighted aggregate of 9.88%. Because $w^- = 1 - s_{\text{cll}}$, this decline is consistent with increasing teacher endorsement among sampled negative-advantage tokens: even when the student assigns them slightly higher probability, they are increasingly judged plausible by the teacher and therefore less often retained for suppression. Conversely, $w^+ = s_{\text{cll}}$ retains positive-advantage updates more often because their reinforcement directions are more frequently supported by teacher endorsement. This pattern indicates that local student–teacher alignment is improving.

B.5 Golden-Gain Enhancement: Trajectory-Filtering Criterion

Table 11 studies the trajectory-level filtering criterion under the same dual-temperature setting. Positive-advantage-density filtering outperforms trajectory-likelihood filtering under both strict and non-strict keep rules, matching our hypothesis that the useful signal is not generic teacher likelihood but the density of teacher-student improvement opportunities. We therefore use the non-strict positive-advantage-density rule in UCMOPD because it provides the strongest role-playing score among the positive-advantage-density variants while retaining a competitive general average.

B.6 Golden-Gain Enhancement: Exploration-Temperature Sensitivity

Table 12 varies the exploration temperature in Golden-Gain Enhancement. Moderate exploration works best among the tested settings: $T = 1.5$ gives the strongest role-playing average and is used as the default in our main configuration. Higher temperatures occasionally improve individual benchmarks such as LiveCodeBench v5 and LiveBench, but are less stable overall, showing that broader exploration does not necessarily yield stronger positive learning signals.

Table 11: Auxiliary comparison of trajectory filtering criteria under the same dual-temperature setting.

Filtering criterion	Gen. Avg.	Vertical Avg.	WritingBench	ZebraLogic
Positive-advantage density ($>$)	52.87	41.46	81.26	72.10
Likelihood ($>$)	51.31	41.32	79.48	71.40
Positive-advantage density ($\geq$)	**53.05**	**42.66**	**81.32**	72.80
Likelihood ($\geq$)	51.81	41.54	80.47	**74.20**

Table 12: Exploration-temperature ablation for positive-advantage-density filtering.

Temperature	Gen. Avg.	Vertical Avg.	IF-Eval	Zebra	LCB v5	LiveBench
1.2	52.30	42.09	79.11	**73.90**	31.54	60.30
1.5	**53.05**	**42.66**	**79.48**	72.80	34.76	61.30
1.8	52.52	41.10	78.19	72.20	**39.78**	61.00
2.0	52.62	42.30	79.11	73.30	33.69	**62.30**

C Properties of the Selection Rules

We give two simple properties that justify the two mechanisms in UCMOPD. These results are not intended as a global convergence guarantee for neural policy optimization. Instead, they show that Golden-Gain Enhancement provides positive-signal enrichment at the trajectory level, while Teacher-Endorsement Filtering provides entropy-adaptive direction-consistency control at the token level.

Trajectory-Level Positive-Signal Guarantee. For each prompt, let y^a be the anchor response and let $\mathcal{E}$ be the set of exploration responses. The filter retains

$$\mathcal{E}_{\text{keep}} = \{y \in \mathcal{E} : r(y) \geq r(y^a)\}.$$

By construction, every retained exploration response satisfies

$$r(y) - r(y^a) \geq 0, \qquad \forall y \in \mathcal{E}_{\text{keep}}.$$

Consequently, if $\mathcal{E}_{\text{keep}}$ is nonempty, its average positive-advantage-density score is lower bounded by the anchor score:

$$\frac{1}{|\mathcal{E}_{\text{keep}}|} \sum_{y \in \mathcal{E}_{\text{keep}}} r(y) \geq r(y^a).$$

Including the always-retained anchor preserves the same lower bound for the retained candidate set $\{y^a\} \cup \mathcal{E}_{\text{keep}}$. The trajectory filter therefore cannot admit an exploration response whose positive-advantage density is below the prompt-specific anchor threshold.

Entropy-Adaptive Direction–Endorsement Consistency. Let $b_T = \exp(-H[p_T])$ denote the entropy-induced typical probability scale of the teacher distribution at a token position. The CLL endorsement score can be rewritten as

$$s_{\text{cll}}(x) = \exp\left(\min\left(\log p_T(x) + H[p_T], 0\right)\right) = \min\left(\frac{p_T(x)}{b_T}, 1\right).$$

Accordingly, the unified keep probability becomes

$$w_{\text{cll}}(A_t, x) = \mathbf{1}[A_t > 0] \min\left(\frac{p_T(x)}{b_T}, 1\right) + \mathbf{1}[A_t < 0]\left[1 - \frac{p_T(x)}{b_T}\right]_+,$$

where $[z]_+ = \max(z, 0)$. Thus CLL compares the sampled-token probability against an entropy-adaptive baseline rather than a fixed top-k set, and then measures whether the proposed update direction agrees with the resulting endorsement. If $p_T(x) \geq b_T$, endorsement is maximal: increasing the token is fully consistent, while decreasing it is inconsistent. If $p_T(x) < b_T$, endorsement decreases smoothly as $p_T(x)/b_T$, continuously transferring retention probability from reinforcement to suppression.

This form also gives the desired entropy dependence. For fixed $p_T(x)$ below the typical scale, $s_{\mathrm{cll}}(x) = p_T(x)\exp(H[p_T])$ is non-decreasing in teacher entropy. A broader teacher distribution therefore recognizes more tokens as plausible, while a confident teacher assigns lower relative endorsement to the same low-probability token. The consistency rule automatically turns this single change in endorsement into opposite retention tendencies for probability-increasing and probability-decreasing updates.