SHARP AT THE PEAK, RICH IN THE PRIOR: PEAK-PRESERVING PRIOR DISTILLATION FOR RL-READY LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised fine-tuning (SFT) for large language models should acquire demonstrated behavior while retaining useful pretrained preferences and alternatives for subsequent reinforcement learning (RL). Conventional one-hot supervision promotes the demonstrated token without specifying which relationships among alternative continuations should be preserved. We introduce *Peak-Preserving Prior Distillation* (P^3D), a principled approach that specifies the desired supervisory distribution before optimization. Guided by *Greedy Guarantee* and *Minimal Intervention*, P^3D constructs a closed-form target that promotes demonstrated-token dominance while preserving relative probabilities among retained non-target alternatives from the current student. This directly embeds acquisition and retention into the supervision target, providing dense feedback without teacher probability vectors or an additional teacher forward pass. Experiments across three models from two families on mathematical and medical reasoning demonstrate stronger SFT and subsequent GRPO performance alongside strong general capabilities. Compared with the respective HardSFT baselines, domain-averaged greedy accuracy improves by up to 6.1 percentage points after SFT and 10.3 points after GRPO. Subsequent analysis further support effective task adaptation with limited predictive drift and retained prior preferences. These findings highlight principled target design as a foundation for learning demonstrated behavior while preserving the capacity for further reinforcement learning. Our code has been open-sourced at the anonymous link https://anonymous.4open.science/P3D-433F.

1 INTRODUCTION

The scaling of pretraining has reshaped the role of supervised fine-tuning (SFT) in reasoning models. Pretrained models already encode broad knowledge and a rich prior over plausible continuations (Gekhman et al., 2024; Zhou et al., 2023; Ren et al., 2024; Zhao et al., 2025). The success of RL directly from base models (Zeng et al., 2025; Guo et al., 2025) and their strong performance under extensive sampling (Chen et al., 2025d) suggest that these priors contain useful capabilities and diverse reasoning paths worth preserving. In the SFT-then-RL pipeline, SFT therefore serves as a bridge: it learns from expert demonstrations while preparing a reliable initialization for subsequent reinforcement learning (RL) (Chu et al., 2025; Chen et al., 2025c;b). Its quality depends on balancing *acquisition* of demonstrated behavior with *retention* of useful alternatives that RL can explore (Zhang et al., 2026).

Standard one-hot supervision does not explicitly balance these roles. It pushes the demonstrated token toward unit probability without specifying which preferences among alternative continuations should survive, potentially narrowing the exploration space available to subsequent RL (Zhao et al., 2026a;b; Kang et al., 2025; Li et al., 2025b). Existing remedies soften this pressure through label smoothing and self-distillation (Szegedy et al., 2016; Wang et al., 2026a), or adaptive gradient reweighting and reference-model regularization (Wang et al., 2026b; Wu et al., 2025; Zhu et al., 2025). Yet the core indicators in these methods like entropy alone does not identify which alternatives are useful: distributions with similar entropy can encode different preferences, and additional uncertainty need not support meaningful exploration (Cui et al., 2025; Çagatan & Akgün, 2025). More fine-grained methods such as GEM account for competition between target and non-target to-

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.

kens (Li et al., 2025c), but this does not explicitly protect the relative preferences *among non-target tokens*. These observations motivate a target-design question: *how should demonstrations reshape a student's distribution while retaining its useful structure?*

Our perspective is inspired by on-policy distillation (Agarwal et al., 2023), in which a teacher can provide dense distributional supervision at student-generated contexts. Such supervision conveys more than the identity of a single preferred token. Here, however, we work with teacher-generated demonstrations whose probability vectors are unavailable. We instead use the student's own predictive distribution as an informative reference for constructing soft targets at demonstration prefixes. The two sources play complementary roles: teacher demonstrations specify what to learn, while student predictions supply the relative preferences to preserve.

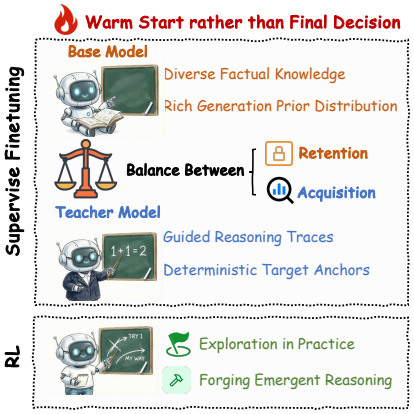


Figure 1: **SFT as a warm start for reinforcement learning.** SFT should balance acquiring demonstrated reasoning strategies with retaining pretrained priors, providing an initialization from which RL can explore and refine reasoning behavior.

We introduce *Peak-Preserving Prior Distillation* (P^3D) to pursue the acquisition–retention balance illustrated in Figure 1. P^3D directly specifies the desired supervisory distribution through two mathematical principles. *Greedy Guarantee* requires the demonstrated token to be the target's preferred continuation, while *Minimal Intervention* seeks the smallest change to the student's prior needed to acquire that preference (Jaynes, 1957). Together, these principles guide the construction of soft targets that support learning from demonstrations while preserving useful distributional structure for subsequent RL exploration.

To summarize, our contributions are threefold:

1. **Conceptually**, we identify preservation of non-target preferences as a target-design criterion and shift the focus of SFT from loss design and aggregate metrics to directly specifying the target distribution through guiding principles, making acquisition and retention explicit requirements on the supervision itself.
2. **Technically**, we propose P^3D, an efficient, mathematically derived soft-target construction that combines teacher demonstrations with the current student's distribution to promote demonstrated preferences while retaining useful prior structure.
3. **Empirically**, we evaluate three models from two families on mathematical and medical reasoning, covering domains with different prior strengths (Li et al., 2025a). P^3D improves greedy accuracy over standard SFT in all six model–domain settings and both greedy accuracy and pass@5 after GRPO. Token-level probes, general-capability evaluations, and training dynamics support these gains with evidence of prior retention.

2 Preliminaries

Setup and notation. Let π_θ be an autoregressive model over vocabulary $\mathcal{V}$, initialized from a pretrained model π_{base}. A demonstration corpus $\mathcal{D}$ contains prompt–response pairs (x, y). At response position t, the context is $s_t = (x, y_{<t})$ and the demonstrated token is $y^\star = y_t$. Standard supervised fine-tuning minimizes negative log-likelihood:

$$\mathcal{L}_{\text{NLL}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid s_t). \tag{1}$$

Distributional supervision. A soft target $Q_t \in \Delta^{|\mathcal{V}|}$ generalizes the one-hot supervision above through token-level cross-entropy,

$$\ell_t(\theta; Q_t) = -\sum_{v\in\mathcal{V}} Q_t(v) \log \pi_\theta(v \mid s_t). \tag{2}$$

The target is held fixed during differentiation. NLL is recovered when Q_t places all mass on $y^\star$. Our method constructs Q_t using both the demonstrated token and student's distribution. We review related work in Appendix A and provide a unified comparison of fine-tuning objectives in Appendix F.

3 A Distributional View of SFT Supervision

Setup. We examine what an SFT objective asks the model to change before optimization. Moving beyond the common focus on mathematics in SFT evaluation (Li et al., 2025a), we examine both mathematical and medical demonstrations to cover contrasting model-prior regimes (Appendix C.2). We probe the next-token prior $P_S = \pi_{\text{base}}(\cdot \mid s_t)$ of the initial checkpoint at teacher-forced response positions. We compare the demonstrated token $y^\star$ with the base's candidates and inspect the supervision imposed on these distributions. The details are described in the Appendix C.

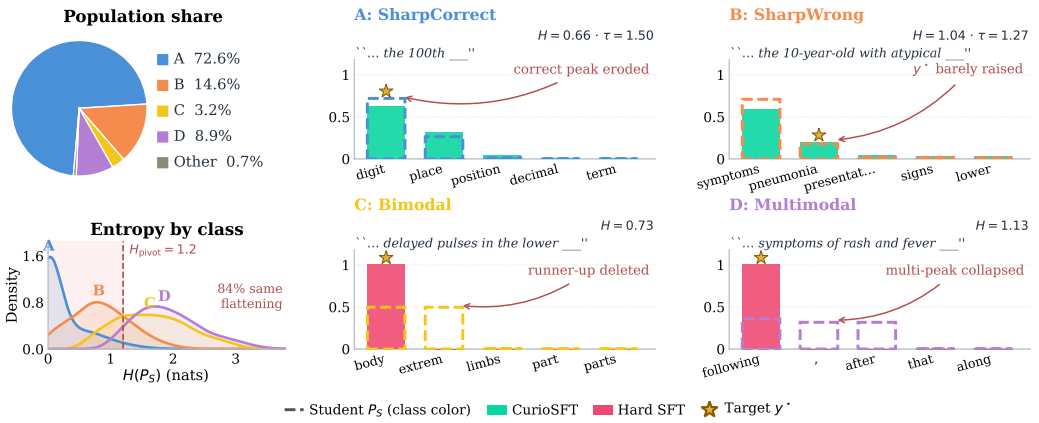


Figure 2: **Different positions require different distributional interventions.** Left: class shares and entropy distributions. Right: top-100 student model distributions (dashed) and one-hot targets or CurioSFT teachers (filled) for Cases A–D; stars mark demonstrated tokens.

3.1 Different Positions Require Different Interventions

Figure 2 organizes positions jointly by the shape of P_S and its agreement with $y^\star$. Classes A and B are both sharp, single-preference cases, but differ in whether the base model's top prediction matches the demonstration: Class A agrees, whereas Class B favors a different token. Classes C and D instead exhibit two dominant peaks or broader multi-peaked candidate structure. These are not merely different levels of uncertainty, but empirically distinct supervision situations.

Most pilot positions are already aligned with the demonstration: $y^\star$ is the base argmax at 77.4% of positions, with Class A alone accounting for 72.6%. For these positions, supervision is often more confirmatory than corrective: it need not overturn the model's current preference, and additional sharpening is not obviously necessary. Class B accounts for 14.6% of positions; at minimum, these cases call for increasing the probability of the demonstrated token, since it is at least a demonstrated acceptable alternative. In many cases, the base instead exhibits a strong mismatched prior, or even a strong potentially hallucinatory preference that needs correction. Classes C and D account for 3.2% and 8.9% of positions, respectively (12.1% combined), and exhibit broader or multi-peaked candidate structure. These cases often call for reinforcing the demonstrated token while, where possible, preserving credible alternative continuations. These patterns suggest that neither uniformly sharpening nor uniformly flattening the distribution suits all positions equally well.

3.2 Hard Labels and Entropy Leave Different Gaps

Hard labels discard the alternative structure. Hard SFT uses the one-hot target $Q = e_{y^\star}$, directing probability toward the demonstration. This provides a clear learning signal, but its independently fitted target retains no mass on other continuations. Classes C and D make the cost visible: two-peak

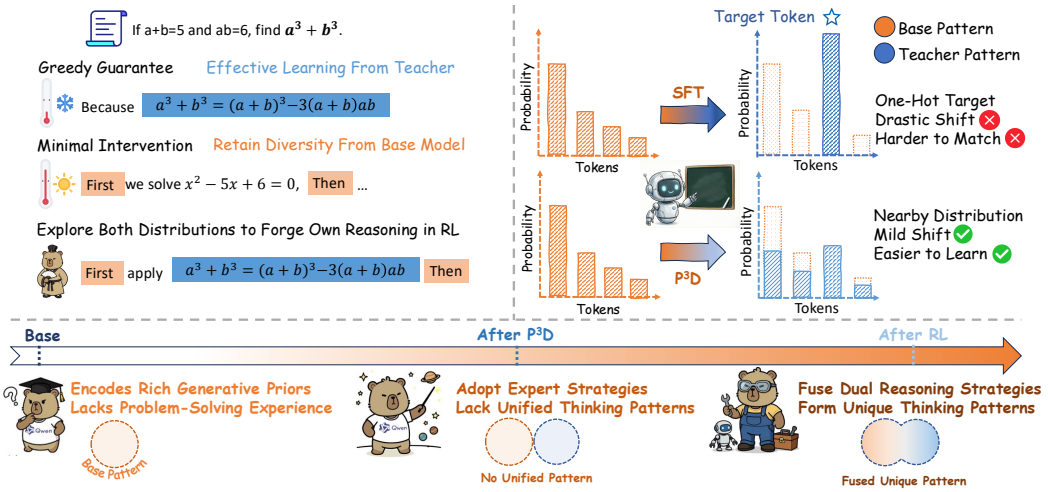


Figure 3: **Constructing distributional supervision from a hard demonstration.** P^{3}D adjusts demonstrated-token probability while retain possible alternatives from the current student.

and multi-peak priors are mapped to the same one-hot form, regardless of the relative plausibility of their alternatives. The issue is therefore the absence of a retention requirement in the target, not the absence of demonstration learning.

Entropy does not identify the required intervention. Smoothing and entropy regularization can counter over-concentration, but entropy describes only *how diffuse* a distribution is, not whether its preference already agrees with the demonstration. Our A/B examples make this distinction concrete: in A, the model already assigns the highest probability to the demonstrated token $y^\star$, so learning mainly requires preserving the preferred token; in B, the model favors a different token, so learning requires changing that preference. These two cases can have similar entropy, yet call for qualitatively different updates. For illustration, A/B overlay the self-distillation teacher induced by entropy-guided temperature scaling (Wang et al., 2026a). Because temperature scaling changes concentration while preserving candidate ordering, it cannot determine whether the preferred token should change. This is especially clear in case B, where the added entropy barely moves probability toward $y^\star$ but instead spreads mass over other implausible candidates in a form of ineffective "uninformative entropy". Together with the overlapping empirical entropy distributions in the lower-left panel, A/B show that confidence alone cannot capture such fine-grained model–demonstration agreement.

Implication. More broadly, many methods attempt to repair SFT through different surrogate constructions or auxiliary metrics, but still do not directly specify the central object of interest: *what distribution should training induce, so that the demonstrated choice is learned without unnecessarily disturbing the model's existing candidate structure?*

4 Peak-Preserving Prior Distillation

Guided by the implication in Section 3.2, we argue that a well-designed SFT target should operate directly on the model's distributional structure, rather than through a lossy aggregate such as entropy. Concretely, we require the target to satisfy two properties:

1. **Greedy Guarantee (Acquisition).** The demonstrated token should strictly dominate the supervisory distribution. This expresses acquisition of the teacher's demonstrated preference at the target level: greedy decoding from the target selects the teacher-provided token.
2. **Minimal Intervention (Retention).** The target should retain the student's existing distributional preferences wherever the demonstration does not require a change. This expresses retention of the student's prior through the smallest distributional adjustment needed to acquire the demonstrated preference.

At context s_t, let $P = \mathrm{stopgrad}[\pi_\theta(\cdot \mid s_t)]$, $p^\star = P(y^\star)$, and $m = \max_{y \neq y^\star} P(y)$. This reference carries the student's pretrained preferences together with information acquired through preceding SFT updates.

4.1 From Minimum KL to Capped Rescaling

We first formulate the target as a minimum-KL projection under demonstrated-token dominance:

$$Q^{\mathrm{cap}} = \arg\min_{Q \in \Delta^{|\mathcal{V}|}} D_{\mathrm{KL}}(Q\|P) \quad \text{s.t.} \quad Q(y^\star) \geq Q(y), \quad \forall y \neq y^\star. \tag{3}$$

The KKT conditions yield the capped-rescaling solution

$$Q^{\mathrm{cap}}(y^\star) = q, \qquad Q^{\mathrm{cap}}(y) = \min\{aP(y), q\}, \quad y \neq y^\star, \tag{4}$$

where a and q satisfy the optimality and normalization conditions. Large competitors are clipped to a common ceiling, while the remaining candidates are uniformly rescaled.

This exact solution exposes two limitations. A direct solver enumerates candidate clipped prefixes to determine the active set at each position, adding overhead to online target construction. Moreover, clipping can map distinct competitors to the same probability, erasing relative preferences that the demonstration does not instruct us to change. Thus, minimum aggregate KL does not by itself preserve non-target structure. Appendix B gives the full derivation and active-set computation.

4.2 Uniform Rescaling and the Supervisory Target

To avoid selective clipping, we extend the exact solution's unclipped rescaling form to all non-target tokens and allocate the remaining mass to the demonstration. Writing $c = Q(y^\star)$ gives

$$Q_c(y) = \begin{cases} c, & y = y^\star, \\ \dfrac{1-c}{1-p^\star} P(y), & y \neq y^\star. \end{cases} \tag{5}$$

This construction immediately preserves the student's relative preferences:

$$\frac{Q_c(y_1)}{Q_c(y_2)} = \frac{P(y_1)}{P(y_2)}, \qquad y_1, y_2 \neq y^\star. \tag{6}$$

Equivalently, the non-target conditional distribution remains unchanged. The construction therefore adjusts only the probability allocation between the demonstration and the alternatives, without introducing a new ordering among those alternatives.

We now select the minimum-KL member of this family satisfying dominance. Substitution into the KL objective gives

$$D_{\mathrm{KL}}(Q_c\|P) = c \log \frac{c}{p^\star} + (1-c) \log \frac{1-c}{1-p^\star}. \tag{7}$$

Dominance requires $c \geq b := m/(1-p^\star+m)$. Since the Bernoulli KL is minimized at $c = p^\star$, the unique solution is

$$c^\dagger = \max\{p^\star, b\}, \qquad Q^\dagger = Q_{c^\dagger}. \tag{8}$$

If the student already satisfies dominance, the target remains P; otherwise, it reaches the dominance boundary with the smallest KL change within the uniform-rescaling family. This is an exact optimum within that family and a structured approximation to Eq. equation 3. The appendix reports an empirical KL ratio of approximately 1.05 relative to the unrestricted optimum under its stated aggregation, quantifying the observed cost of retaining non-target structure.

Practical preference strength. The zero-margin construction can place the demonstration at a tie. In training, we use $c^\star = \max\{p^\star, b + \delta\}$, subject to numerical clipping, and set $Q_\delta = Q_{c^\star}$. The offset δ raises the target-probability floor above the tie boundary while retaining the same rescaling structure. It is a practical choice of supervision strength, selected empirically. Its precise meaning and numerical treatment are detailed in Appendix B.7.

4.3 Fitting the Constructed Distribution

The training objective fits the detached target with cross-entropy:

$$\mathcal{L}_{\mathrm{P^3D}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \sum_t \sum_{v\in\mathcal{V}} Q_\delta(v \mid s_t) \log \pi_\theta(v \mid s_t). \tag{9}$$

For an individual position, its logit gradient is

$$\frac{\partial \ell_t}{\partial z_\theta(v)} = \pi_\theta(v \mid s_t) - Q_\delta(v \mid s_t). \tag{10}$$

Thus, the demonstrated-token correction depends on the increase required by the constructed target. When $p^\star \geq b + \delta$, the ideal target equals P and the corresponding local logit gradient vanishes. To better understand the method from the perspective of optimization dynamics, we derive an **adaptive decomposition of the objective** based on the student's prior distribution and target token at each position. This provides a dynamical interpretation of demonstration learning and prior retention; the derivation is presented in Appendix B.8.

For efficiency, we apply the same top-K selection and target-inclusion simplifications used in our implementations of related works. Implementation details are provided in Appendix D.3.

5 Experiments

5.1 Setup

We evaluate P^{3}D on three base models from two families (Qwen3-4B/8B and Llama-3-8B) in mathematical and medical reasoning. We compare SFT objectives under a shared subsequent GRPO configuration, and include RFT and Prefix-RFT as dedicated RFT baselines. Table 1 reports greedy accuracy (T=0), pass@5 (T=0.6), and mean zero-shot accuracy over seven general benchmarks (Gen. Avg.). We additionally report mean@5, the average accuracy across five responses sampled at $T = 0.6$. Training data, hyperparameters, and evaluation details are in Appendices D and E.

5.2 Main Results

Table 1 shows strong SFT performance across model families and reasoning domains. **P^{3}D improves greedy accuracy over HardSFT in all six model–domain settings and achieves the highest greedy accuracy and pass@5 when averaged across these settings.** On Qwen3-8B Math, it reaches 65.5% greedy accuracy and 77.8% pass@5, improving on HardSFT by $\uparrow$ 6.1 and $\uparrow$ 8.3 respectively. The gains transfer to Llama3-8B, with greedy improvements of $\uparrow$ 5.8 on Math and $\uparrow$ 5.2 on Medical. On Medical, P^{3}D also leads both metrics when averaged across the three backbones.

The advantage extends to subsequent GRPO: P^{3}D outperforms HardSFT initialization on both task metrics in all six settings. Under the shared GRPO configuration, the Math greedy gain reaches $\uparrow$ 10.3 points on Qwen3-8B, while Qwen3-4B Math pass@5 rises from 64.9% to 77.4%. P^{3}D achieves the highest post-GRPO Math greedy accuracy and pass@5 on all three models among the reported methods, including RFT and Prefix-RFT. Its Medical results likewise lead both metrics when averaged across backbones, establishing strong overall performance in both domains.

These task gains are accompanied by strong general-capability performance. Across the six model–domain settings, P^{3}D attains the highest mean Gen. Avg. at both the SFT and post-GRPO stages. The joint results show that effective specialization can coexist with broad capabilities, making P^{3}D a strong foundation for reasoning post-training. Additional results on sampled performance are provided in Appendix J.

5.3 Acquisition with Prior Retention

We examine retention through both policy shift and general-benchmark performance of Qwen3-4B Base in the math domain.

In the distributional view (Figure 4(a)), LS and HardSFT achieve strong task performance, but both incur larger shifts than P^{3}D. ASFT remains closer to Base, yet moves toward a lower-performing

Table 1: **Main results (%).** (A) SFT; (B) SFT+GRPO, with one-stage RFT baselines in italics. **T=0**: greedy accuracy; **@5**: pass@5 at $T = 0.6$. **Gen. Avg.**: mean zero-shot accuracy over seven general benchmarks. **Bold**/underline: best/second best per column, model, and stage. Arrows: percentage-point changes from HardSFT (A) or HardSFT+GRPO (B). Dashes: incomplete coverage.

Method	Math													Medical								
	MATH500		GSM8K		AIME'24		AIME'25		Minerva		Avg		Gen.	MedQA		MedMCQA		PubMedQA		Avg		Gen.
	T=0	@5	T=0	@5	T=0	@5	T=0	@5	T=0	@5	T=0	@5	Avg.	T=0	@5	T=0	@5	T=0	@5	T=0	@5	Avg.
(A) SFT stage																						
Qwen3-4B																						
HardSFT	73.4	85.4	89.0	96.4	20.0	36.7	13.3	30.0	56.6	70.4	50.5	63.8	53.7	49.9	86.7	46.4	80.8	65.6	87.2	53.9	84.9	54.4
LS	71.4	86.4	89.4	97.0	**23.3**	**46.7**	23.3	33.3	55.8	69.9	52.6↑2.2	66.7↑2.9	53.8	53.7	86.5	47.1	81.4	66.6	**88.0**	**55.8**↑1.9	85.3↑0.4	54.6
EntropyReg	71.0	84.2	88.1	97.3	**23.3**	36.7	13.3	36.7	54.5	69.2	50.0↓0.4	64.8↑1.0	53.8	46.0	86.8	44.3	80.8	66.2	86.0	52.2↓1.8	84.5↓0.3	**54.8**
GEM	69.8	82.4	82.2	96.2	20.0	40.0	13.3	30.0	60.4	76.3	49.2↓1.3	65.0↑1.2	49.5	47.4	87.5	45.6	82.1	66.8	80.0	53.3↓0.7	83.2↓1.7	48.7
DFT	71.4	88.2	65.9	80.2	13.3	20.0	10.0	20.0	**61.3**	73.3	44.4↓6.1	56.3↓7.4	48.3	49.1	75.7	41.1	63.7	**71.0**	84.6	53.7↓0.2	74.7↓10.2	47.2
ASFT	68.4	79.4	86.7	80.1	16.7	23.3	13.3	16.7	30.7	75.5	43.1↓7.3	55.0↓8.8	52.4	**55.5**	84.8	48.4	73.9	61.0	83.6	55.0↑1.0	80.8↓4.1	52.4
EAFT	73.0	**88.4**	89.6	**97.9**	**23.3**	**46.7**	23.3	40.0	57.6	**78.4**	53.4↑2.9	**70.3**↑6.5	52.9	48.1	88.5	48.3	81.5	65.0	85.6	53.8↓0.1	85.2↑0.3	52.8
CurioSFT	69.6	72.6	86.8	89.5	16.7	10.0	13.3	6.7	51.3	54.3	47.5↓2.9	46.6↓17.2	47.7	48.6	40.1	**48.7**	76.8	68.2	52.2	55.2↑1.2	56.4↓28.5	42.9
P^{3}D (Ours)	**73.8**	88.2	**90.5**	97.2	20.0	**46.7**	**26.7**	**43.3**	59.0	74.3	**54.0**↑3.5	69.9↑6.1	**54.5**	48.0	**88.7**	47.9	**82.2**	68.2	87.2	54.7↑0.7	**86.0**↑1.1	52.9
Qwen3-8B																						
HardSFT	79.0	90.2	90.4	96.4	36.7	36.7	26.7	46.7	64.2	77.7	59.4	69.5	54.7	61.7	84.8	54.0	82.0	64.8	82.6	60.2	83.1	53.8
LS	76.2	91.2	84.3	94.5	30.0	**53.3**	30.0	60.0	63.2	77.1	56.7↓2.6	75.2↑5.7	53.9	56.9	87.3	56.1	81.4	62.6	84.4	58.5↓1.7	84.4↑1.3	54.0
EntropyReg	77.0	85.4	90.3	96.9	26.7	30.0	23.3	33.3	62.5	75.9	56.0↓3.4	64.3↓5.2	54.2	67.6	86.1	54.8	80.3	66.4	86.0	62.9↑2.7	84.1↑1.0	54.7
GEM	73.4	88.8	86.8	95.5	23.3	40.0	20.0	36.7	60.5	**86.5**	52.8↓6.6	69.5	51.1	64.5	86.2	55.8	**82.8**	65.4	80.8	61.9↑1.7	83.3↑0.2	50.8
DFT	74.6	87.0	86.6	95.7	30.0	46.7	30.0	43.3	57.0	70.3	55.6↓3.7	68.6↓0.9	54.6	65.2	85.9	48.7	73.0	67.0	83.2	60.3↑0.1	80.7↓2.4	52.5
ASFT	71.0	84.4	85.9	92.7	20.0	33.3	16.7	33.3	55.2	79.4	49.7↓9.6	64.6↓4.9	54.9	**69.7**	89.2	**57.2**	81.7	56.8	85.0	61.2↑1.0	85.3↑2.2	53.7
EAFT	79.6	**93.4**	89.9	97.5	26.7	46.7	26.7	43.3	62.7	76.1	57.1↓2.3	71.4↑1.9	54.3	64.5	87.1	54.8	81.5	65.0	84.2	61.4↑1.2	84.3↑1.2	53.6
CurioSFT	59.0	60.4	73.9	81.6	20.0	30.0	10.0	26.7	44.8	60.1	41.5↓17.8	51.8↓17.8	47.4	57.6	53.4	47.2	70.5	61.6	57.6	55.4↓4.7	60.5↓22.6	42.8
P^{3}D (Ours)	**80.8**	93.0	**91.9**	**97.8**	**50.0**	**53.3**	**36.7**	**63.3**	**68.3**	81.8	**65.5**↑6.1	**77.8**↑8.3	**55.2**	66.0	**89.9**	56.8	82.3	**70.6**	**86.8**	**64.4**↑4.2	**86.3**↑3.2	**54.8**
Llama3-8B																						
HardSFT	50.8	74.6	79.5	95.1	10.0	**30.0**	13.3	33.3	38.0	58.4	38.3	58.3	47.4	24.4	75.3	48.2	77.5	56.6	87.0	43.1	80.0	49.6
LS	51.6	77.0	79.5	**95.8**	6.7	26.7	16.7	30.0	38.5	60.3	38.6↑0.3	57.9↓0.4	47.0	25.9	75.4	45.8	76.1	61.4	85.0	44.4↑1.3	78.8↓1.1	48.5
EntropyReg	51.4	74.6	73.8	95.2	10.0	26.7	16.7	**40.0**	38.0	59.8	38.0↓0.4	**59.2**↑1.0	47.4	19.6	76.5	43.7	76.1	57.2	86.8	40.2↓2.9	79.8↓0.1	48.4
GEM	38.6	57.8	73.2	95.3	3.3	16.7	13.3	23.3	38.6	**60.4**	33.4↓4.9	50.7↓7.6	44.5	24.4	69.3	46.8	77.2	65.0	82.6	45.4↑2.3	76.3↓3.6	45.8
DFT	34.6	47.0	55.4	66.3	3.3	10.0	3.3	20.0	26.1	36.9	24.6↓13.8	36.0↓22.2	43.6	**39.1**	69.0	31.1	56.9	35.0	66.8	35.1↓8.0	64.3↓15.7	43.3
ASFT	29.8	56.0	1.4	40.7	3.3	10.0	3.3	33.3	26.2	46.9	12.8↓25.5	37.4↓20.9	46.2	28.4	57.3	33.6	65.7	33.2	72.0	31.7↓11.4	65.0↓15.0	46.3
EAFT	48.4	71.2	75.4	95.6	6.7	23.3	13.3	36.7	34.1	58.5	35.6↓2.7	57.1↓1.2	**47.8**	30.6	79.0	48.5	**78.7**	60.8	85.2	46.6↑3.6	81.0↑1.0	49.8
CurioSFT	25.8	40.6	62.2	64.5	6.7	6.7	3.3	13.3	16.9	30.1	23.0↓15.3	31.0↓27.2	37.0	21.9	31.6	37.5	70.7	58.8	61.0	39.4↓3.7	54.4↓25.6	37.8
P^{3}D (Ours)	**55.2**	**77.4**	**79.6**	94.6	**13.3**	26.7	**26.7**	30.0	**45.7**	60.3	**44.1**↑5.8	57.8↓0.5	47.7	29.2	**81.9**	**48.9**	77.7	**66.6**	**87.8**	**48.2**↑5.2	**82.5**↑2.5	**49.8**
(B) RL stage (post-GRPO)																						
Qwen3-4B																						
HardSFT+GRPO	78.8	86.2	92.1	95.2	16.7	26.7	13.3	33.3	73.5	83.0	54.9	64.9	53.3	68.3	87.7	60.3	80.8	69.2	85.0	65.9	84.5	52.9
LS+GRPO	76.0	87.6	90.4	96.4	13.3	33.3	16.7	33.3	69.8	84.0	53.2↓1.6	66.9↑2.1	53.0	71.9	88.5	58.9	79.0	70.6	83.0	67.1↑1.2	83.5↓1.0	53.0
EntropyReg+GRPO	73.8	88.0	90.4	95.6	23.3	53.3	13.3	33.3	70.5	83.6	54.3↓0.6	70.8↑5.9	53.0	54.2	86.5	43.5	82.0	63.6	**89.0**	53.8↓2.2	85.8↑1.4	52.9
GEM+GRPO	74.4	84.0	92.5	95.6	23.3	33.3	20.0	36.7	**79.4**	**87.5**	57.9↑3.0	67.4↑2.5	50.0	74.4	89.4	60.0	81.1	69.4	86.0	67.9↑2.0	85.5↑1.0	49.6
DFT+GRPO	74.6	81.4	91.5	93.8	10.0	23.3	10.0	26.7	70.4	78.5	51.3↓3.6	60.7↓4.1	51.3	52.8	72.3	51.1	73.6	67.0	84.6	56.9↓9.0	76.8↓7.7	46.5
ASFT+GRPO	78.6	87.8	87.6	87.9	20.0	26.7	13.3	36.7	73.6	81.7	54.6↓0.2	64.1↓0.7	52.7	70.2	86.9	58.3	77.9	68.8	82.0	65.8↓0.2	82.2↓2.2	52.6
EAFT+GRPO	76.8	85.4	92.3	96.5	16.7	30.0	20.0	33.3	74.3	83.4	56.0↑1.1	65.7↑0.9	**53.4**	73.1	90.7	60.6	79.8	65.0	85.0	66.2↑0.3	85.2↑0.7	52.6
CurioSFT+GRPO	42.0	66.8	66.1	91.7	13.3	43.3	16.7	30.0	49.4	65.5	37.5↓17.4	59.5↓5.4	41.8	57.5	76.9	50.3	81.7	67.8	74.4	58.5↓7.4	77.7↓6.8	43.5
P^{3}D (Ours)+GRPO	**82.0**	**92.2**	**94.2**	**97.7**	**30.0**	**60.0**	30.0	**50.0**	77.5	87.3	**62.7**↑7.8	**77.4**↑12.6	53.1	**74.5**	90.3	**62.9**	**82.9**	**72.4**	**89.0**	**70.0**↑4.0	**87.4**↑2.9	**53.4**
RFT	72.4	88.0	88.2	97.0	**30.0**	46.7	**36.7**	46.7	57.5	76.6	57.0↑2.1	71.0↑6.1	49.1	64.6	89.0	54.2	82.4	69.2	85.6	62.7↓3.3	85.7↑1.2	50.1
Prefix-RFT	76.2	81.2	88.9	92.3	13.3	20.0	16.7	26.7	61.5	78.0	51.3↓3.5	59.6↓5.2	52.7	67.2	**90.9**	55.9	81.4	67.8	88.8	63.7↓2.3	87.0↑2.5	52.6
Qwen3-8B																						
HardSFT+GRPO	68.0	86.2	93.2	**97.8**	26.7	40.0	20.0	43.3	72.0	84.3	56.0	70.3	54.5	79.3	92.2	65.1	81.1	71.0	84.8	71.8	86.0	53.5
LS+GRPO	82.4	90.8	89.9	94.5	30.0	50.0	16.7	33.3	67.7	78.2	57.3↑1.4	69.4↓1.0	53.4	81.2	92.2	66.7	82.6	74.0	84.6	74.0↑2.2	86.5↑0.4	54.4
EntropyReg+GRPO	70.4	91.8	85.8	95.8	23.3	36.7	20.0	46.7	62.1	82.9	52.3↓3.7	70.8↑0.5	54.0	68.0	86.2	54.8	80.5	66.4	86.8	63.1↓8.7	84.5↓1.5	**54.6**
GEM+GRPO	79.0	81.6	88.4	96.6	23.3	40.0	23.3	43.3	61.6	**87.8**	55.1↓0.9	69.9↓0.5	50.6	79.4	92.2	66.2	82.2	**75.2**	84.2	73.6↑1.8	86.2↑0.2	51.2
DFT+GRPO	77.4	83.8	90.4	92.6	20.0	30.0	13.3	26.7	70.4	76.6	54.3↓1.7	61.9↓8.4	53.9	71.3	89.5	60.2	79.6	69.8	83.8	67.1↓4.7	84.3↓1.7	51.7
ASFT+GRPO	80.0	87.2	90.9	95.5	16.7	33.3	13.3	26.7	**74.6**	84.1	55.1↓0.9	65.3↓5.0	54.3	78.2	91.9	64.8	82.7	67.4	80.8	70.2↓1.6	85.1↓0.9	53.2
EAFT+GRPO	**86.0**	92.0	94.4	97.3	**33.3**	60.0	36.7	46.7	74.4	83.3	65.0↑9.0	75.9↑5.5	54.8	81.8	92.2	62.6	81.2	67.4	83.8	70.6↓1.2	85.7↓0.3	53.3
CurioSFT+GRPO	64.2	76.2	66.1	91.7	13.3	36.7	20.0	33.3	59.8	69.8	44.7↓11.3	61.5↓8.8	49.3	57.7	60.4	47.3	80.6	62.4	70.8	55.8↓16.0	70.6↓15.4	43.2
P^{3}D (Ours)+GRPO	**86.0**	**92.8**	**94.5**	97.6	**33.3**	**66.7**	**43.3**	**56.7**	74.4	81.7	**66.3**↑10.3	**79.1**↑8.7	**54.8**	**82.2**	**93.6**	**68.1**	83.3	74.6	86.8	**75.0**↑3.2	**87.9**↑1.9	54.5
RFT	76.6	89.4	90.4	97.3	10.0	56.7	23.3	40.0	62.8	80.9	52.6↓3.4	72.9↑2.5	51.1	72.3	91.9	59.5	**83.7**	70.2	83.0	67.4↓4.4	86.2↑0.2	51.1
Prefix-RFT	68.0	86.4	82.6	96.2	6.7	30.0	16.7	26.7	65.3	81.0	47.9↓8.1	64.1↓6.3	54.4	79.3	92.7	63.9	83.2	70.4	**87.4**	71.2↓0.6	87.8↑1.7	53.5
Llama3-8B																						
HardSFT+GRPO	53.4	66.6	71.9	93.8	10.0	20.0	16.7	26.7	30.8	53.2	36.5	52.0	46.1	69.8	85.7	60.2	79.7	74.0	84.0	68.0	83.1	48.5
LS+GRPO	55.2	67.4	67.2	93.8	**16.7**	23.3	13.3	23.3	33.5	54.3	37.2↑0.6	52.4↑0.4	47.2	71.7	86.6	61.3	80.4	**77.4**	84.6	70.1↑2.2	83.9↑0.7	47.7
EntropyReg+GRPO	58.0	75.0	72.9	94.8	10.0	23.3	20.0	**40.0**	37.1	59.8	39.6↑3.1	58.6↑6.5	46.9	41.1	76.5	44.6	76.7	62.8	86.2	49.5↓18.5	79.8↓3.3	48.1
GEM+GRPO	46.2	70.0	68.4	93.4	13.3	20.0	13.3	23.3	36.0	**66.3**	35.5↓1.1	54.6↑2.6	45.2	71.6	88.7	60.6	80.6	72.2	83.4	68.1↑0.1	84.2↑1.1	46.3
DFT+GRPO	**59.4**	37.0	65.7	66.3	10.0	13.3	23.3	20.0	26.1	29.7	36.9↑0.4	33.3↓18.8	45.0	60.0	84.5	48.9	72.9	57.4	85.4	55.5↓12.5	80.9↓2.2	42.2
ASFT+GRPO	44.0	47.2	58.3	76.0	10.0	10.0	6.7	16.7	29.3	35.5	29.7↓6.9	37.1↓15.0	46.4	68.1	87.3	56.5	80.0	57.2	81.8	60.6↓7.4	83.0↓0.1	46.7
EAFT+GRPO	54.8	63.4	69.4	93.4	13.3	6.7	16.7	30.0	38.6	57.1	38.6↑2.0	50.1↓1.9	**48.1**	**74.7**	**90.7**	62.4	**82.1**	75.4	85.4	**70.8**↑2.9	86.1↑2.9	49.3
CurioSFT+GRPO	55.8	60.6	62.2	84.5	**16.7**	16.7	16.7	26.7	26.9	50.1	35.6↓0.9	47.7↓4.3	37.4	30.7	40.4	37.6	70.5	58.8	81.0	42.3↓25.6	64.0↓19.2	38.4
P^{3}D (Ours)+GRPO	56.0	**81.8**	76.7	**96.0**	**16.7**	**26.7**	**26.7**	36.7	40.5	66.3	**43.3**↑6.8	**61.5**↑9.4	47.9	71.0	87.8	**63.0**	80.6	75.8	86.6	69.9↑2.0	85.0↑1.9	**49.5**
RFT	30.2	37.8	44.4	79.6	10.0	10.0	6.7	23.3	11.9	30.0	20.6↓15.9	36.1↓15.9	45.7	63.0	89.5	51.8	81.2	71.4	**87.6**	62.1↓5.9	**86.1**↑3.0	45.2
Prefix-RFT	57.0	79.8	**79.9**	95.9	13.3	20.0	20.0	33.3	**44.2**	65.4	42.9↑6.4	58.9↑6.8	46.7	69.5	90.0	59.3	80.3	75.4	84.8	68.1↑0.1	85.0↑1.9	49.2

region. **Among the plotted checkpoints that outperform HardSFT, P^{3}D achieves the highest task accuracy with the smallest distributional shift.** This pattern reflects the minimal-intervention preference of our target construction: reaching a high-performing solution while staying close to the original predictive distribution.

The benchmark view complements this picture by examining whether the task gains extend to broader capabilities (Figure 4(b)). **P^{3}D combines the strongest task performance with the highest general-benchmark average among the plotted checkpoints.** Together, the two views support

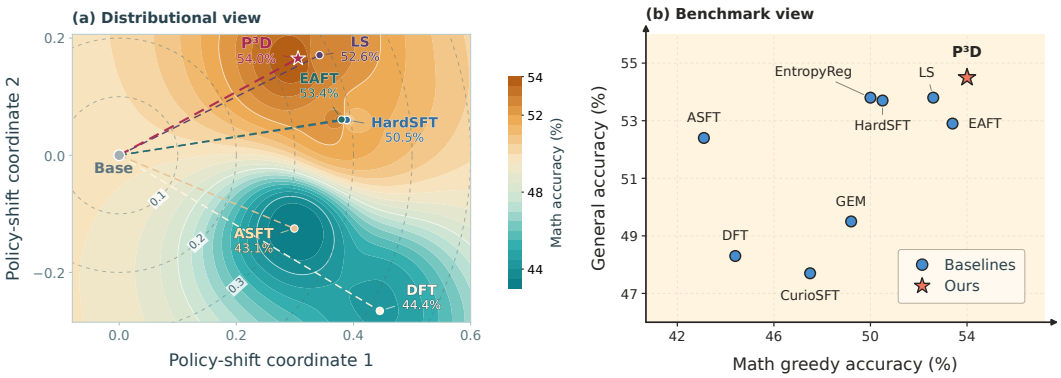
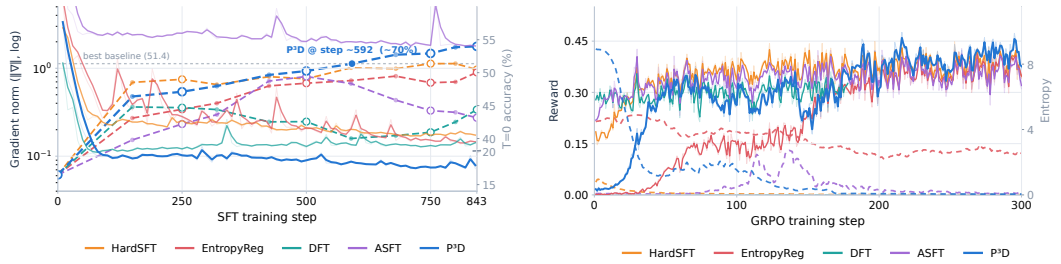


Figure 4: **Acquisition and retention after SFT.** (a) Projected distribution shifts from Base; rings indicate distance and background colors are reasonably interpolated/extrapolated performance. Lines connect Base to checkpoints, not training trajectories. (b) Task versus general accuracy; circles denote baselines and the star denotes P^{3}D.

effective domain adaptation with restrained distributional change and strong general performance. Detailed benchmark results, distributional probe details, and examples are provided in Appendix G.

5.4 Learning Efficiency and Subsequent Adaptation



(a) SFT learning efficiency

(b) Subsequent GRPO adaptation

Figure 5: **Learning efficiency and subsequent adaptation.** (a) SFT gradient norm (solid, left axis) and greedy accuracy (dashed, right); horizontal reference: best observed baseline score. The annotated crossing time is estimated from the accuracy curve. (b) GRPO reward (solid, left) and entropy (dashed, right), with $T = 1.0$. For clarity of presentation, both panels show a representative subset of SFT-loss variants spanning distinct loss-design families across the SFT and GRPO stages.

Figure 5(a) shows how the gradient signals and task accuracy evolve together during SFT. P^{3}D's gradient norm falls to a small magnitude early in training and remains low as its accuracy continues to improve. **Sustained performance gains accompany the smallest average gradient norm among the plotted methods.** Together with the limited distributional drift in Figure 4(a), these dynamics support the method's tendency toward effective adaptation with restrained changes to the pretrained model's predictions.

The accuracy curves also show efficient use of demonstration exposure. **P^{3}D is estimated reaches the best displayed baseline accuracy at approximately 70% of the training schedule.** With the shared batch size, this corresponds to roughly 70% of the scheduled training-example presentations. Accuracy continues to increase thereafter, showing that the method reaches a competitive level earlier and improves further with additional training.

At the start of GRPO in Figure 5(b), P^{3}D combines strong greedy performance with the highest initial policy entropy among the plotted methods by retaining alternatives in the prior distribution for sampling, even exceeding EntropyReg with an explicit entropy bonus. As entropy decreases, reward rises rapidly; following intermediate fluctuations, further improvement yields the highest late-stage reward among the displayed methods. HardSFT and ASFT start with higher reward but make more limited subsequent gains, while EntropyReg maintains higher entropy late in training

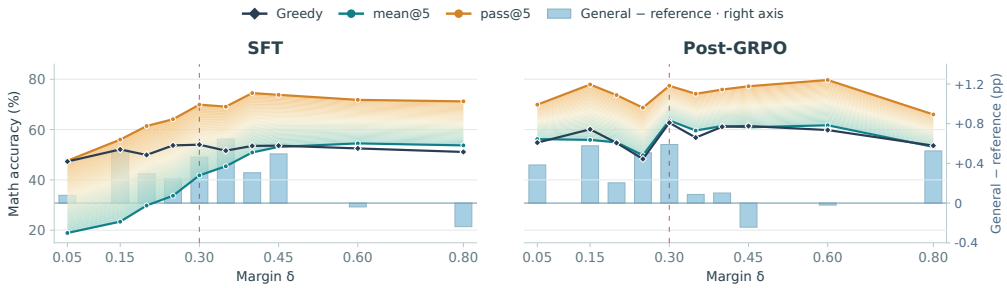


Figure 6: **Effect of δ.** The shaded gap between mean@5 and pass@5 highlights the sampling headroom. Bars show general-accuracy changes (right axis, pp) relative to Base for SFT and Base+GRPO for post-GRPO. The dashed line marks the default $\delta = 0.30$.

without attaining a reward advantage. **This contrast supports the usefulness of the alternatives retained by distributional supervision for subsequent RL: high initial entropy accompanies substantial reward learning and stronger final performance.** Additional training dynamics are provided in Appendix H.

5.5 Effect of Intervention Strength

During SFT, increasing δ strengthens target promotion. Mean@5 rises toward and eventually exceeds greedy accuracy before saturating (Figure 6, left). At larger margins, pass@5 gains level off, the shaded band narrows, and general-capability gains diminish. **Stronger intervention improves sampled correctness, but its benefits for coverage and general-capability retention diminish.**

After GRPO, mean@5 tracks greedy accuracy more closely across much of the sweep, while pass@5 remains higher. Intermediate margins sustain strong task performance; the largest margin reduces all scores, and stronger intervention brings no consistent general-accuracy benefit. Retaining alternatives during SFT leaves room for reward-based refinement. **Moderate supervision provides an effective RL starting point, combining reliable sampled performance with sampling headroom.**

All tested margins from 0.30 through 0.60 outperform HardSFT on greedy accuracy and pass@5 at both stages on Math, supporting the target construction across a broad operating range. **We use $\delta = 0.30$, the lower end of this range, as the default to retain sampling headroom with modest intervention.** It also maintains general accuracy above the respective reference at both stages. Medical results and the complete sweep are provided in Appendix I.

6 Discussion and Conclusion

We presented P^{3}D, which brings demonstrated-token dominance and minimal intervention into the design of distributional supervision. **The key design choice is to specify the desired supervisory distribution directly from acquisition and retention principles.** This shifts attention from indirect control through loss regularization or gradient modulation to the target that optimization is asked to fit. Across mathematical and medical reasoning, P^{3}D improves supervised adaptation and subsequent RL performance while maintaining strong general capabilities. The accompanying distributional and training analyses support a consistent picture: effective task learning can coexist with limited changes to the student's existing predictions. Selective adaptation offers a practical way to acquire new behavior while retaining useful foundations from pretraining.

More broadly, this work highlights the importance of the predictive distribution that SFT leaves for subsequent learning. Demonstrations identify the preferences to acquire, while the student's prior provides structure among alternative continuations. Preserving this structure makes target design relevant to both immediate performance and the opportunities available for later reward-based refinement. Our findings therefore motivate evaluating SFT by the behavior it teaches and the capacity for further improvement it retains. **Effective supervision should strengthen the model's preferred behavior while preserving its ability to learn further.**

AI Use Statement

We used generative AI tools solely for prose polishing and for assisting with related-work literature search. All other components of this work, including the P^3D formulation, the theoretical derivations, the figures, the code, and the experimental design, were completed by the human authors. We have reviewed all AI-assisted portions of the manuscript, and we take full responsibility for the final content of this work.

Ethics Statement

This work uses only widely adopted public benchmarks and open-source pretrained models that are permitted for academic research. It does not involve human subjects, private data, or safety-critical deployment. We therefore do not foresee any ethical concerns raised by this submission.

Reproducibility Statement

The full source code required to reproduce our experiments is released at an anonymous repository at https://anonymous.4open.science/P3D-433F. All theoretical derivations, including the KL-projection formulation of P^3D and its closed-form solution, are presented in full in the main text and the appendix. Experimental results are reported with evaluation in multiple random seeds and complete evaluation configurations, and training hyperparameters for every model family, domain, and baseline are documented in the appendix so that all experiments can boduced end-to-end.

References

Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. GKD: generalized knowledge distillation for auto-regressive sequence models. *CoRR*, abs/2306.13649, 2023. doi: 10.48550/ARXIV.2306.13649. URL https://doi.org/10.48550/arXiv.2306.13649.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Ömer Veysel Çagatan and Baris Akgün. Failure modes of maximum entropy RLHF. *CoRR*, abs/2509.20265, 2025. doi: 10.48550/ARXIV.2509.20265. URL https://doi.org/10.48550/arXiv.2509.20265.

Huayu Chen, Kaiwen Zheng, Qinsheng Zhang, Ganqu Cui, Yin Cui, Haotian Ye, Tsung-Yi Lin, Ming-Yu Liu, Jun Zhu, and Haoxiang Wang. Bridging supervised learning and reinforcement learning in math reasoning. *CoRR*, abs/2505.18116, 2025a. doi: 10.48550/ARXIV.2505.18116. URL https://doi.org/10.48550/arXiv.2505.18116.

Jack Chen, Fazhong Liu, Naruto Liu, Yuhan Luo, Erqu Qin, Harry Zheng, Tian Dong, Haojin Zhu, Yan Meng, and Xiao Wang. Step-wise adaptive integration of supervised fine-tuning and reinforcement learning for task-specific llms. *CoRR*, abs/2505.13026, 2025b. doi: 10.48550/ARXIV.2505.13026. URL https://doi.org/10.48550/arXiv.2505.13026.

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, 2024. URL https://arxiv.org/abs/2412.18925.

Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. Beyond two-stage training: Cooperative SFT and RL for LLM reasoning. *CoRR*, abs/2509.06948, 2025c. doi: 10.48550/ARXIV.2509.06948. URL https://doi.org/10.48550/arXiv.2509.06948.

Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025d. URL http://papers.nips.cc/paper_files/paper/2025/hash/537d5aa768c2d534016a4d06f87bc8fb-Abstract-Conference.html.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/chu25c.html.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. *CoRR*, abs/2505.22617, 2025. doi: 10.48550/ARXIV.2505.22617. URL https://doi.org/10.48550/arXiv.2505.22617.

Muxi Diao, Lele Yang, Wuxuan Gong, Yutong Zhang, Zhonghao Yan, Yufei Han, Kongming Liang, Weiran Xu, and Zhanyu Ma. Entropy-adaptive fine-tuning: Resolving confident conflicts to mitigate forgetting. *arXiv preprint arXiv:2601.02151*, 2026.

Yihong Dong, Xue Jiang, Yongding Tao, Huanyu Liu, Kechi Zhang, Lili Mou, Rongyu Cao, Yingwei Ma, Jue Chen, Binhua Li, Zhi Jin, Fei Huang, Yongbin Li, and Ge Li. RL-PLUS: countering capability boundary collapse of llms in reinforcement learning with hybrid-policy optimization. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 43029–43053. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1994. URL https://doi.org/10.18653/v1/2026.acl-long.1994.

Yuhao Du, Zhuo Li, Pengyu Cheng, Zhihong Chen, Yuejiao Xie, Xiang Wan, and Anningzhe Gao. Rlhf in an sft way: From optimal solution to reward-weighted alignment, 2026. URL https://arxiv.org/abs/2502.11026.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. SRFT: A single-stage method with supervised and reinforcement fine-tuning for reasoning. *CoRR*, abs/2506.19767, 2025. doi: 10.48550/ARXIV.2506.19767. URL https://doi.org/10.48550/arXiv.2506.19767.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *CoRR*, abs/2503.01307, 2025. doi: 10.48550/ARXIV.2503.01307. URL https://doi.org/10.48550/arXiv.2503.01307.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 7765–7784. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.444. URL https://doi.org/10.18653/v1/2024.emnlp-main.444.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel,

Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan

Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021b.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/ed873d79e7c268c020c4b4db13a2812a-Abstract-Conference.html.

Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo M. Ponti, and Ivan Titov. Blending supervised and reinforcement fine-tuning with prefix sampling. *CoRR*, abs/2507.01679, 2025. doi: 10.48550/ARXIV.2507.01679. URL https://doi.org/10.48550/arXiv.2507.01679.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William 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 (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Lou Jost. Entropy and diversity. *Oikos*, 113:363–375, 2006. URL https://api.semanticscholar.org/CorpusID:86256823.

Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. Quagmires in SFT-RL post-training: When high SFT scores mislead and what to use instead. *CoRR*, abs/2510.01624, 2025. doi: 10.48550/ARXIV.2510.01624. URL https://doi.org/10.48550/arXiv.2510.01624.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022. URL https://arxiv.org/abs/2206.14858.

Gaotang Li, Ruizhong Qiu, Xiusi Chen, Heng Ji, and Hanghang Tong. Beyond log likelihood: Probability-based objectives for supervised fine-tuning across the model capability continuum. *CoRR*, abs/2510.00526, 2025a. doi: 10.48550/ARXIV.2510.00526. URL https://doi.org/10.48550/arXiv.2510.00526.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://hf-mirror.com/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving diversity in supervised fine-tuning of large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=NQEe7B7bSw.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving diversity in supervised fine-tuning of large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025c. URL https://openreview.net/forum?id=NQEe7B7bSw.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.229. URL https://doi.org/10.18653/v1/2022.acl-long.229.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Lixuan He, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, Ning Ding, and Bowen Zhou. Towards a unified view of large language model post-training. *CoRR*, abs/2509.04419, 2025. doi: 10.48550/ARXIV.2509.04419. URL https://doi.org/10.48550/arXiv.2509.04419.

HF Mirror. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (eds.), *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 3505–3506. ACM, 2020. doi: 10.1145/3394486.3406703. URL https://doi.org/10.1145/3394486.3406703.

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.

Mengjie Ren, Boxi Cao, Hongyu Lin, Cao Liu, Xianpei Han, Ke Zeng, Guanglu Wan, Xunliang Cai, and Le Sun. Learning or self-aligning? rethinking instruction fine-tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 6090–6105. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.330. URL https://doi.org/10.18653/v1/2024.acl-long.330.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Mingyu Su, Jian Guan, Yuxian Gu, Minlie Huang, and Hongning Wang. Trust-region adaptive policy optimization. *CoRR*, abs/2512.17636, 2025. doi: 10.48550/ARXIV.2512.17636. URL https://doi.org/10.48550/arXiv.2512.17636.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

Luong Quoc Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 7601–7614. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.410. URL https://doi.org/10.18653/v1/2024.acl-long.410.

Bo Wang, Qinyuan Cheng, Runyu Peng, Rong Bao, Peiji Li, Qipeng Guo, Linyang Li, Zhiyuan Zeng, Yunhua Zhou, and Xipeng Qiu. Implicit reward as the bridge: A unified view of SFT and DPO connections. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025a. URL http://papers.nips.cc/paper_files/paper/2025/hash/6599417a0b34d6ab7836cf86f8dd138c-Abstract-Conference.html.

Hao Wang, Hao Gu, Hongming Piao, Kaixiong Gong, Yuxiao Ye, Xiangyu Yue, Sirui Han, Yike Guo, and Dapeng Wu. Learning while staying curious: Entropy-preserving supervised fine-tuning via adaptive self-distillation for large reasoning models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 13567–13581. Association for Computational Linguistics, 2026a. doi: 10.18653/V1/2026.ACL-LONG.617. URL https://doi.org/10.18653/v1/2026.acl-long.617.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025b. URL https://arxiv.org/abs/2506.01939.

Xinyu Wang, Changzhi Sun, Yuanbin Wu, and Xiaoling Wang. Stabilizing llm supervised fine-tuning via explicit distributional control, 2026b. URL https://arxiv.org/abs/2605.04468.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the generalization of SFT: A reinforcement learning perspective with reward rectification. *CoRR*, abs/2508.05629, 2025. doi: 10.48550/ARXIV.2508.05629. URL https://doi.org/10.48550/arXiv.2508.05629.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/a9d5c33e38442450f9781bd7bad3c81e-Abstract-Conference.html.

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, 2025. URL https://arxiv.org/abs/2505.09388.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *CoRR*, abs/2503.18892, 2025. doi: 10.48550/ARXIV.2503.18892. URL https://doi.org/10.48550/arXiv.2503.18892.

Dylan Zhang, Yufeng Xu, Haojin Wang, Qingzhi Chen, and Hao Peng. Good SFT optimizes for sft, better SFT prepares for reinforcement learning. *CoRR*, abs/2602.01058, 2026. doi: 10.48550/ARXIV.2602.01058. URL https://doi.org/10.48550/arXiv.2602.01058.

Shiyue Zhang, Shijie Wu, Ozan Irsoy, Steven Lu, Mohit Bansal, Mark Dredze, and David S. Rosenberg. Mixce: Training autoregressive language models by mixing forward and reverse cross-entropies. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 9027–9050. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.502. URL https://doi.org/10.18653/v1/2023.acl-long.502.

Eric Zhao, Pranjal Awasthi, and Nika Haghtalab. From style to facts: Mapping the boundaries of knowledge injection with finetuning. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/90043ebd68500f9efe84fedf860a64f3-Abstract-Conference.html.

Kai Zhao, Yanjun Zhao, Jiaming Song, Shien He, Lusheng Zhang, Qiang Zhang, and Tianjiao Li. SABER: switchable and balanced training for efficient LLM reasoning. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 34950–34958. AAAI Press, 2026a. doi: 10.1609/AAAI.V40I41.40799. URL https://doi.org/10.1609/aaai.v40i41.40799.

Zhengyang Zhao, Lu Ma, Yizhen Jiang, Xiaochen Ma, Zimo Meng, Chengyu Shen, Lexiang Tang, Haoze Sun, Peng Pei, and Wentao Zhang. Gift: Reconciling post-training objectives via finite-temperature gibbs initialization, 2026b. URL https://arxiv.org/abs/2601.09233.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL http://arxiv.org/abs/2403.13372.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: less is more for alignment. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html.

He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. Anchored supervised fine-tuning. *CoRR*, abs/2509.23753, 2025. doi: 10.48550/ARXIV.2509.23753. URL https://doi.org/10.48550/arXiv.2509.23753.

A Related Work

The Role of SFT in the SFT-then-RL Paradigm. Supervised fine-tuning followed by reinforcement learning is a standard post-training recipe for reasoning and instruction-following mod-

els (Ouyang et al., 2022; Guo et al., 2025; Chu et al., 2025). Studies emphasize SFT's role in eliciting capabilities acquired during pretraining (Zhou et al., 2023; Gekhman et al., 2024), while direct-RL approaches investigate reasoning improvement without a conventional supervised warm start (Hu et al., 2025; Yu et al., 2025). Work on accessible reasoning capabilities and post-training limitations (Chen et al., 2025d; Dong et al., 2026; Yan et al., 2025), together with analyses of entropy and exploration during RL (Cui et al., 2025), motivates evaluating SFT by both the demonstrated behavior it acquires and the useful alternatives it retains. Our method addresses this balance through supervisory-target construction.

Hybrid and Reward-Guided Training. RFT (Trung et al., 2024) learns from reward-selected generations. Hybrid approaches include LUFFY (Yan et al., 2025), which incorporates reweighted off-policy demonstrations; SRFT (Fu et al., 2025) and HPT (Lv et al., 2025), which integrate supervised and reward-based learning; Prefix-RFT (Huang et al., 2025), which guides generation with expert prefixes; and RL-PLUS (Dong et al., 2026), which introduces exploration-aware advantage functions. NFT (Chen et al., 2025a), MixCE (Zhang et al., 2023), and GOLD (Lu & Lab, 2025) also connect imitation with policy learning. These approaches modify training trajectories or the interaction between learning objectives. Our focus is complementary: specifying the distributional supervision at each demonstrated position.

Regularization and Target-Distribution Design in SFT. Standard one-hot SFT (HardSFT) (Ouyang et al., 2022) promotes demonstrations without explicitly retaining alternative-token mass, motivating remedies for excessive concentration (Bai et al., 2022; Kirk et al., 2024). *Concentration control*: Label Smoothing (LS) (Szegedy et al., 2016), entropy regularization (EntropyReg) (Jost, 2006), and CurioSFT (Wang et al., 2026a) soften supervision or retain uncertainty, but uncertainty alone does not specify which alternative preferences to preserve. *Position-level reweighting*: DFT (Wu et al., 2025) and EAFT (Diao et al., 2026) adjust CE strength using target confidence and entropy, respectively; these statistics do not directly specify a target satisfying demonstrated-token dominance and non-target structure preservation. Related implicit-RL interpretations also motivate per-position weighting (Wang et al., 2025a; Du et al., 2026). *Candidate competition*: GEM (Li et al., 2025c) uses a game-theoretic update with temperature β controlling competitor weights. It models target–competitor interactions, but its temperature-controlled weighting remains a heuristic with respect to jointly enforcing demonstrated-token dominance and preserving non-target structure: these requirements do not determine the competitor weights. *Reference retention*: ASFT (Zhu et al., 2025) combines confidence weighting with reference anchoring, while PSFT (Su et al., 2025) constrains policy drift. These encourage retention, but a reference constraint alone does not ensure demonstrated-token dominance. Appendix F provides detailed comparisons.

On-policy distillation (OPD) (Agarwal et al., 2023) offers a complementary setting in which a teacher supplies distributional supervision at student-generated contexts. When only hard demonstrations are available, however, the probabilities of alternative continuations remain unspecified. The gap highlighted by these comparisons is a principled mechanism that turns explicit acquisition and retention requirements into a computable supervisory distribution, specifying both how the demonstrated preference should be acquired and which relationships among alternatives should remain intact.

B Minimum-Information Projection for Hard-Label Distribution Construction

In black-box hard-label distillation, the teacher provides only a partial supervision signal: the target token $y^\star$ is required to be preferred over all alternatives, while no calibrated probabilities are specified for the remaining tokens. Consequently, the target distribution is not uniquely determined by the teacher signal alone. The problem is therefore to characterize a supervisory distribution Q that satisfies the hard preference constraint while preserving as much information from the student's original distribution P as possible.

This issue is particularly important in post-training large language models. A direct modification of the standard SFT objective, such as optimizing hard-label cross-entropy or augmenting it with

auxiliary regularizers, affects the student only indirectly through the loss function. Such objective-level interventions do not explicitly determine the target distribution being imposed, and thus provide limited control over what information from P is retained after optimization. In particular, target collapse toward a one-hot distribution centered at $y^\star$ guarantees top-1 correctness, but destroys the relative structure among non-target tokens, removes calibrated uncertainty, and suppresses the diversity inherited from pretraining. These components of the original distribution are precisely the ones that define a meaningful exploration space for subsequent reinforcement learning.

For this reason, our formulation does not proceed by altering the SFT objective itself. Instead, we directly construct an explicit target distribution Q and fit the student to that target. This perspective makes the supervision mechanism fully explicit and allows precise control over the deviation from P. The desired target distribution must satisfy two properties:

- It must place $y^\star$ among the highest-probability tokens; a positive margin, introduced after the structured projection, strengthens this weak dominance to strict preference in the constructed target.
- It should preserve the characteristic structure of P to the greatest extent possible, including the relative preference pattern over non-target tokens and the residual diversity of plausible alternatives.

In this way, the resulting supervision remains compatible with the teacher signal while retaining the distributional features of the pretrained model that are most relevant for effective downstream RL.

The analysis below first studies the information-theoretically minimal perturbation of P under the top-1 dominance constraint through a KL projection formulation. It then introduces an additional structure-preserving principle that retains the original geometry of P over non-target tokens, which yields the ratio-preserving rescaling adopted in our method.

B.1 Problem Setup

Let $P \in \Delta^{|\mathcal{V}|}$ denote the student's current next-token distribution over vocabulary $\mathcal{V}$, treated as fixed during target construction, and let $y^\star \in \mathcal{V}$ be the teacher-selected target token. We consider constructing a new distribution $Q \in \Delta^{|\mathcal{V}|}$. We assume $P(y) > 0$ for all tokens, as with a finite-logit softmax. In training, $P = \mathrm{stopgrad}(\pi_\theta(\cdot \mid s_t))$: it carries both pretrained preferences and information acquired through preceding SFT updates, without requiring a separate frozen reference model.

A fully unconstrained formulation is:

$$\begin{aligned} \min_{Q \in \Delta^{|\mathcal{V}|}} \quad & D_{\mathrm{KL}}(Q\|P) = \sum_{y \in \mathcal{V}} Q(y) \log \frac{Q(y)}{P(y)} \\ \text{s.t.} \quad & Q(y^\star) \geq Q(y), \qquad \forall y \neq y^\star. \end{aligned} \tag{11}$$

This objective corresponds to a minimum-information update from P to a new distribution Q that makes $y^\star$ a maximizer of the constructed target. Ties are allowed at this stage; this constraint does not by itself guarantee the trained model's decoding behavior.

B.2 The Unconstrained KL Projection Problem

We first characterize the optimal solution of equation 11. For convenience, rewrite the constraints as

$$Q(y) - Q(y^\star) \leq 0, \qquad \forall y \neq y^\star.$$

We also impose the normalization constraint $\sum_y Q(y) = 1$. Since the KL objective is strictly convex in Q and all constraints are linear, the problem is convex, and the KKT (Karush-Kuhn-Tucker) conditions are necessary and sufficient.

Lagrangian. Introduce Lagrange multipliers $\lambda_y \geq 0$ for each inequality constraint $Q(y) - Q(y^\star) \leq 0$ with $y \neq y^\star$, and $\nu \in \mathbb{R}$ for the normalization constraint. The Lagrangian is

$$\mathcal{L}(Q, \lambda, \nu) = \sum_{y \in \mathcal{V}} Q(y) \log \frac{Q(y)}{P(y)} + \sum_{y \neq y^\star} \lambda_y \big(Q(y) - Q(y^\star)\big) + \nu \left(\sum_{y \in \mathcal{V}} Q(y) - 1 \right). \tag{12}$$

Stationarity. For $y \neq y^\star$,

$$\frac{\partial \mathcal{L}}{\partial Q(y)} = \log \frac{Q(y)}{P(y)} + 1 + \lambda_y + \nu = 0,$$

which gives

$$Q(y) = c\, P(y) e^{-\lambda_y}, \qquad y \neq y^\star, \tag{13}$$

where $c := e^{-1-\nu} > 0$.

For $y = y^\star$,

$$\frac{\partial \mathcal{L}}{\partial Q(y^\star)} = \log \frac{Q(y^\star)}{P(y^\star)} + 1 - \sum_{y \neq y^\star} \lambda_y + \nu = 0,$$

which gives

$$Q(y^\star) = c\, P(y^\star) \exp\left(\sum_{y \neq y^\star} \lambda_y \right). \tag{14}$$

Complementary slackness. The KKT complementary slackness condition is

$$\lambda_y \big(Q(y) - Q(y^\star) \big) = 0, \qquad \forall y \neq y^\star.$$

Thus:

- if $Q(y) < Q(y^\star)$, then $\lambda_y = 0$, and by equation 13, $Q(y) = cP(y)$;
- if $Q(y) = Q(y^\star)$, then λ_y may be positive, corresponding to an active constraint.

Therefore, the optimal solution has the following capped-rescaling form.

Proposition 1 (Unconstrained KL projection structure)**.** *The unique solution of equation 11 has the form*

$$Q^\star(y^\star) = q^\star, \qquad Q^\star(y) = \min\{cP(y),\, q^\star\}, \quad \forall y \neq y^\star, \tag{15}$$

for some $q^\star \in (0, 1)$ and $c > 0$, together with the normalization condition

$$q^\star + \sum_{y \neq y^\star} \min\{cP(y),\, q^\star\} = 1. \tag{16}$$

Interpretation. Proposition 1 shows that the true unconstrained optimum does *not* in general correspond to uniform rescaling over all non-target tokens. Instead, large competitor tokens may be clipped at the boundary $Q(y) = Q(y^\star)$, while lower-probability tokens are uniformly rescaled.

B.3 Why We Do Not Use the Full Unconstrained Optimum

Cost of enumerating the clipped prefix. The capped-rescaling form does not by itself determine the scaling factor or the clipping threshold. For the zero-margin problem, sort the non-target probabilities as $r_1 \geq \cdots \geq r_n$, where $n = |\mathcal{V}| - 1$, and write $p^\star = P(y^\star)$. The clipped competitors form a sorted prefix. For a candidate prefix length k, the KKT conditions give

$$a_k = \left(p^\star \prod_{i=1}^{k} r_i \right)^{1/(k+1)}, \qquad c_k = \frac{1}{(k+1)a_k + \sum_{i=k+1}^{n} r_i}, \qquad q_k = c_k a_k.$$

The candidate must satisfy $r_k \geq a_k \geq r_{k+1}$, with the corresponding missing inequality omitted for $k = 0$ or $k = n$. A direct exact solver enumerates candidate prefix lengths k and checks these KKT consistency conditions, evaluating the geometric mean in log space. The appropriate k is position dependent, so this search must be resolved separately for each next-token distribution. Repeating the search across response tokens and training steps adds target-construction time to online SFT, making the exact projection less practical for large-scale training. Sorting and prefix computations can accelerate the search, but can not avoid determining a clipping set altogether.[1]

[1] This enumeration concerns sorted prefixes, not arbitrary token subsets. The number k of clipped competitors is distinct from the implementation's top-K support size.

Changes to competitor structure. Although the unconstrained optimum in Proposition 1 is mathematically the minimum-KL solution, its structural behavior is misaligned with the information content of black-box hard-label supervision. To see this concretely, consider a typical "confidently wrong" position where the student assigns a decreasing sequence of probabilities to the plausible tokens while the teacher's $y^\star$ lies in the tail:

$$P(y_1) = 0.60,\ P(y_2) = 0.10,\ P(y_3) = 0.05,\ \ldots,\ P(y^\star) \approx 0.001.$$

The general form clips the top-k non-target tokens down to a common plateau $q^\star$, replacing the ordered decreasing sequence with a flat plateau.

$$Q^\star(y_1) = Q^\star(y_2) = \cdots = Q^\star(y_k) = q^\star.$$

That is, the general form *flattens* the student's own probability judgments among the top competitors and treats them as indistinguishable.

This flattening has three concrete drawbacks:

1. **Erasure of student calibration.** The student's original relative probabilities encode meaningful information: y_1 at 0.60 says "I am confident in this", while y_2 at 0.10 says "this is a plausible alternative". Collapsing these to the same value $q^\star$ destroys the gradient of confidence, treating tokens the student regarded as very different as identical.
2. **Injection of unsupported structure.** A hard label only tells us that $y^\star$ should dominate. It carries *no* information about whether the top competitors should be equal-probability or preserve their original ordering. Enforcing equality among them injects structure that the supervision itself does not imply, which a scalar KL objective does not explicitly prevent.
3. **Degradation of downstream exploration.** When the SFT-supervised model is subsequently used as a warm-start for RL, the runner-up ordering is precisely what sampling exploits during exploration. A flat plateau of clipped tokens presents RL with an artificially indistinguishable set of alternatives, discarding useful prior information that the student had learned about their relative likelihoods.

Both drawbacks arise from selective clipping: it requires identifying which competitors reach the ceiling and collapses those competitors to a common probability. This suggests retaining the unclipped rescaling component of the exact solution and applying it to all non-target tokens, while assigning the remaining mass to the demonstration. We develop this construction next and then examine what structure it preserves.

B.4 From Capped Rescaling to Uniform Non-Target Rescaling

Removing selective clipping. In Proposition 1, every unclipped competitor is multiplied by the same factor. We extend this form to the entire non-target support:

$$Q(y) = \alpha P(y), \qquad y \neq y^\star.$$

This is motivated by the absence of a clipped prefix ($k = 0$) in the non-target profile. We retain this profile while allowing the demonstrated-token mass to adjust independently.

Writing $c := Q(y^\star)$ and $p^\star := P(y^\star)$, normalization gives $\alpha = (1 - c)/(1 - p^\star)$ and hence the family

$$Q(y) = \begin{cases} c, & y = y^\star, \\ \dfrac{1 - c}{1 - p^\star}\, P(y), & y \neq y^\star, \end{cases} \qquad c \in (0, 1). \tag{17}$$

Constructing a member of this family requires no search over clipped competitors; only its target mass c remains to be chosen.

Preservation follows from the construction. For any two non-target tokens, the shared factor cancels:

$$\frac{Q(y_1)}{Q(y_2)} = \frac{P(y_1)}{P(y_2)}. \tag{18}$$

Equivalently, the entire conditional distribution of the alternatives is unchanged:

$$Q(y \mid y \neq y^\star) = P(y \mid y \neq y^\star), \qquad \forall y \neq y^\star. \tag{19}$$

Thus the construction preserves a strong form of non-target structure: both relative preferences and conditional uncertainty remain intact. In particular, for Shannon entropy H (Jaynes, 1957),

$$H(Q(\cdot \mid y \neq y^\star)) = H(P(\cdot \mid y \neq y^\star)) .$$

No extra entropy is introduced within this conditional distribution which conform to *minimum-information augmentation principle* (Jaynes, 1957). The entropy of the full distribution can change as mass moves between the demonstration and the alternatives. These properties follow from uniform rescaling; they are not assumptions needed to obtain the preceding capped-rescaling analysis. They also match the information available from a hard label, which identifies the preferred token without specifying new relative probabilities among the alternatives.

B.5 Minimum-KL Target Within the Uniform-Rescaling Family

Having chosen the family equation 17 to avoid selective clipping, we select its least-KL member that satisfies demonstrated-token dominance:

$$\begin{aligned} \min_{c \in (0,1)} \quad & D_{\mathrm{KL}}(Q \| P) \\ \text{s.t.} \quad & Q(y^\star) \geq Q(y), \qquad \forall\, y \neq y^\star, \end{aligned} \tag{20}$$

where Q is parameterized by equation 17. Since uniform rescaling is equivalent to conditional-distribution preservation, this also establishes exact optimality under that structural requirement.

Solving the reduced problem. Substituting equation 17 into the KL objective yields

$$D_{\mathrm{KL}}(Q \| P) \;=\; c \log \frac{c}{p^\star} \;+\; (1-c) \log \frac{1-c}{1-p^\star},$$

the two-outcome KL between Bernoulli distributions $(c, 1-c)$ and $(p^\star, 1-p^\star)$. This is strictly convex in c on $(0,1)$, with unconstrained minimum at $c = p^\star$ (recovering $Q = P$ and $\mathrm{KL} = 0$).

Let $m := \max_{y \neq y^\star} P(y)$. The dominance constraint reduces to the single scalar inequality

$$c \;\geq\; \frac{1-c}{1-p^\star}\, m \quad \Longleftrightarrow \quad c \;\geq\; \frac{m}{1-p^\star+m}.$$

When $p^\star \geq m$ (i.e., $y^\star$ is already the argmax of P), the unconstrained minimum $c = p^\star$ is feasible, and the projection is trivially $Q = P$. When $p^\star < m$, the unconstrained minimum is infeasible and, by strict convexity, the KL-optimum sits at the boundary

$$c^\dagger = \frac{m}{1-p^\star+m}. \tag{21}$$

Combining both cases, we obtain the following result.

Proposition 2 (Structured minimum-information projection)**.** *Among all distributions satisfying the top-1 dominance condition and preserving the non-target conditional structure as in equation 19, the unique minimizer of $D_{\mathrm{KL}}(Q\|P)$ is*

$$Q^\dagger(y) = \begin{cases} c^\dagger, & y = y^\star, \\ P(y)\,\dfrac{1-c^\dagger}{1-p^\star}, & y \neq y^\star, \end{cases} \qquad c^\dagger = \max\left\{ p^\star, \frac{m}{1-p^\star+m} \right\}. \tag{22}$$

This is the exact minimum-KL solution under dominance and conditional-structure preservation. When $p^\star \geq m$, it returns $Q^\dagger = P$; otherwise, it reaches the weak-dominance boundary while preserving all non-target probability ratios.

B.6 Empirical KL Gap to the Exact Projection

The uniform-rescaling target is an exact optimum within its family and a structured approximation to the original projection. We now assess the KL cost of removing selective clipping. Both targets in this comparison use the original zero-margin dominance requirement, so the comparison isolates the cost of preserving non-target structure before any positive-margin extension.

Empirical KL gap. On our pilot corpus of 134,651 response-token positions across chat, medical, and math domains, we compute both the general-form optimum $Q^\star$ (Proposition 1) and our closed-form projection $Q^\dagger$ (Proposition 2) at each position. The pattern-share weighted mean of the pattern-internal KL ratios is

$$\overline{\frac{D_{\mathrm{KL}}(Q^\dagger \| P)}{D_{\mathrm{KL}}(Q^\star \| P)}} \approx 1.05,$$

i.e., about a 5% relative KL overhead under this aggregation. This is an empirical aggregate, not a pointwise approximation bound. Together with the exact conditional-preservation property, this result supports uniform rescaling as a practical alternative: it avoids clipping-set search and competitor flattening while remaining close to the minimum KL under the reported empirical aggregation. The additional KL is the cost of the chosen structural restriction. When the student already satisfies dominance, both projections equal P; equality can also occur where the exact solution belongs to the uniform-rescaling family.

B.7 The Peak-Preservation Margin δ

From weak dominance to a target-probability margin. The preceding comparison concerns the cost of uniform rescaling under weak dominance. We now strengthen the demonstrated-token preference for training. Define the ratio-preserving flip boundary

$$b := \frac{m}{1 - p^\star + m}, \qquad c^\dagger = \max\{p^\star, b\}.$$

When correction is needed, the zero-margin projection places the target exactly at a tie with the strongest competitor. To construct a target with a stronger demonstrated-token preference, we raise the target-probability floor above this boundary:

$$c^\star = \max\{p^\star, b + \delta\}, \qquad 0 \le \delta < 1 - b. \tag{23}$$

Here δ is an additive margin in target probability above the flip boundary. It is neither a fixed target-versus-competitor probability gap nor an increment added to $c^\dagger$ or to $p^\star$. The associated target Q_δ uses equation 17 with $c = c^\star$.

Minimum KL under the strengthened floor. The margin extension remains an exact structured projection:

$$\min_Q D_{\mathrm{KL}}(Q\|P) \quad \text{s.t.} \quad Q(\cdot \mid y \neq y^\star) = P(\cdot \mid y \neq y^\star), \qquad Q(y^\star) \ge b + \delta.$$

Substitution of equation 17 gives the same Bernoulli KL, now on $c \ge b + \delta$. Its unique minimizer is equation 23. Thus, positive δ deliberately strengthens the original weak-dominance requirement; it does not retain minimum KL under the weaker requirement alone. It preserves the non-target conditional distribution exactly while allowing additional target-versus-rest mass transfer.

What probability gap does the margin induce? For any member of the ratio-preserving family,

$$g(c) := Q(y^\star) - \max_{y \neq y^\star} Q(y) = c - \frac{1-c}{1-p^\star} m = \frac{1 - p^\star + m}{1 - p^\star}(c - b).$$

Consequently, before numerical clipping,

$$g(c^\star) \ge \delta \frac{1 - p^\star + m}{1 - p^\star},$$

with equality when $c^\star = b + \delta$. The induced probability-gap requirement is position dependent. For example, $p^\star = 0.1$ and $m = 0.6$ give $b = 0.4$; $\delta = 0.1$ sets $c^\star = 0.5$ and the strongest competitor to $1/3$, giving a gap of $1/6$, not 0.1. A positive target margin supplies a strict preference in Q_δ, but does not guarantee boundary crossing in one optimizer step or robustness of the trained model under inference-time perturbations.

Preservation and the role of the current student. If $p^\star \ge b + \delta$, the optimum remains $Q_\delta = P$. Otherwise, it increases the demonstrated-token mass and rescales all non-target probabilities by the same factor. Thus, increasing δ can also update positions where the demonstration is already top-1 but its advantage is small. Preservation refers to the student's current non-target conditional preferences, which include both pretrained information and preceding SFT updates; it does not fix the absolute non-target probabilities. Larger margins reduce the total mass available to alternatives even though their ratios remain unchanged.

Implementation: online top-K targets and numerical clipping. The implementation constructs P from the current student's detached logits on a top-K support, replacing the lowest selected token with $y^\star$ if necessary, and normalizes both the reference and the predicted distribution on this same support. All probabilities and maxima in the implemented construction refer to that normalized support. No additional teacher forward pass is required. With $\epsilon_{\text{num}} = 10^{-6}$, the implemented scalars are

$$\widehat{b} = \frac{m}{\max\{1 - p^\star + m, \epsilon_{\text{num}}\}},$$
$$\widehat{c} = \min\{1 - \epsilon_{\text{num}}, \max\{p^\star, \widehat{b} + \delta\}\},$$
$$\widehat{\alpha} = \frac{1 - \widehat{c}}{\max\{1 - p^\star, \epsilon_{\text{num}}\}}.$$

The target assigns $\widehat{c}$ to $y^\star$ and $\widehat{\alpha}P(y)$ to retained non-target tokens, then uses soft cross-entropy on this support. The upper clip allows a fixed experimental δ even when $b + \delta$ approaches or exceeds one; once it binds, the unclipped gap bound above no longer applies. Away from the numerical guards, this is exactly the structured construction. At extreme saturation, the upper clip can reduce a $p^\star > 1 - \epsilon_{\text{num}}$, and the denominator guard can make the target sum differ slightly from one; exact normalization and non-regression statements concern the ideal construction. Non-target ratios are preserved on the retained support, not asserted for the full vocabulary.

B.8 Learning Dynamics Through a Prior-Adaptive Decomposition

Having established the closed-form target $Q^\dagger$ in Proposition 2, we examine the same objective from the perspective of learning dynamics. Its decomposition exposes coefficients that adapt to the student's reference distribution at each position, making the contributions of demonstration learning and prior retention explicit. This is an equivalent analysis of the objective induced by the target construction. We first consider the zero-margin solution and then extend the analysis to the positive-margin target.

Mixture form of the target. Recall that for any position with student prior P and target token $y^\star$, the structure-preserving solution is

$$Q^\dagger(y^\star) = c^\dagger, \qquad Q^\dagger(y) = P(y)\,\frac{1 - c^\dagger}{1 - p^\star} \quad (y \neq y^\star).$$

Define the *non-target conditional prior*

$$\pi_{-y^\star}(y) := \frac{P(y)}{1 - p^\star}\,\mathbb{1}[y \neq y^\star], \qquad \sum_y \pi_{-y^\star}(y) = 1. \tag{24}$$

Then $Q^\dagger$ can be written as a *mixture at the distribution level*:

$$\boxed{Q^\dagger \;=\; c^\dagger \cdot \delta_{y^\star} \;+\; (1 - c^\dagger) \cdot \pi_{-y^\star},} \tag{25}$$

i.e., a convex combination of the Dirac at the target and the student's own conditional prior over non-target tokens.

Loss decomposition. Substituting equation 25 into $\mathrm{CE}(Q^\dagger, P_\theta) = -\sum_y Q^\dagger(y) \log P_\theta(y)$ and using the fact that $\delta_{y^\star}$ and $\pi_{-y^\star}$ have disjoint support, we obtain

$$\underbrace{\mathrm{CE}(Q^\dagger, P_\theta)}_{\mathcal{L}_{\mathrm{p^3_D}}(\theta)} \;=\; c^\dagger \cdot \underbrace{\big(-\log P_\theta(y^\star)\big)}_{\mathcal{L}_{\mathrm{SFT}}(\theta)} + (1 - c^\dagger) \cdot \underbrace{\mathrm{CE}\big(\pi_{-y^\star},\, P_\theta\big)}_{\mathcal{L}_{\mathrm{prior}}(\theta)}. \tag{26}$$

The right-hand side is a convex combination of two familiar objectives:

- $\mathcal{L}_{\mathrm{SFT}}$ is exactly the classical hard-label cross-entropy loss used in one-hot SFT—it drives $P_\theta(y^\star)$ upward and shrinks non-target mass uniformly.
- $\mathcal{L}_{\mathrm{prior}}$ is a *self-distillation* loss against the student's own non-target conditional distribution $\pi_{-y^\star}$. This reference is recomputed at each step and held fixed during differentiation, providing supervision from the current relative preferences among non-target alternatives.

Loss decomposition with a margin. The mixture identity equation 25 and loss decomposition equation 26 extend to the positive-margin target from Section B.7 with $c^\dagger$ replaced by $c^\star$:

$$\mathcal{L}_\delta(\theta) = c^\star \mathcal{L}_{\mathrm{SFT}}(\theta) + (1 - c^\star)\mathcal{L}_{\mathrm{prior}}(\theta). \tag{27}$$

At $\delta = 0$ this recovers the structured minimum-KL projection. As $b + \delta$ approaches one, the ideal target approaches one-hot supervision; the implementation stops just below one for numerical stability. For the detached current-student reference, positions with $Q_\delta = P$ have zero logit gradient under cross-entropy on the same normalized support, apart from numerical effects. Margin selection therefore controls the strength of demonstrated-token preference and the remaining mass for alternatives, while the conditional-structure constraint controls which relationships are preserved.

With a positive margin, the zero-gradient condition is $p^\star \geq b + \delta$; the weaker condition $p^\star \geq m$ only establishes that the demonstration is already a most-probable token. For example, at $\delta = 0.30$, $p^\star = 0.45$ and $m = 0.40$ give $c^\star \approx 0.721$, so the target-logit gradient remains nonzero. In contrast, $p^\star = 0.80$ and $m = 0.10$ give $b+\delta \approx 0.633 < p^\star$, hence $Q_\delta = P$ in the ideal construction. Floating-point computation can leave small residual gradients even in this latter case. These statements concern one position's contribution to the logit gradient; other positions and optimizer state can still change the shared model parameters.

Consequences for the zero-margin projection. The following three observations concern $c^\dagger$ at $\delta = 0$. With a positive margin, the same loss decomposition holds with $c^\star$, but zero perturbation requires $p^\star \geq b + \delta$, and the zero-margin bound $c^\dagger \leq m$ need not hold for $c^\star$.

1. **Zero perturbation on confident positions.** When $y^\star$ is already the student's argmax ($p^\star \geq m$), the zero-margin dominance constraint is satisfied, $c^\dagger$ collapses to $p^\star$, and $Q^\dagger = P$. The loss becomes $\mathcal{L}_{\mathrm{P^3D}} = -\sum_y P(y) \log P_\theta(y) = H(P) + D_{\mathrm{KL}}(P \| P_\theta)$, whose gradient vanishes at $P_\theta = P$. In this regime the objective performs no one-hot update which simply certifies the current distribution as consistent with the teacher signal. This is the mechanism by which P^{3}D avoids destroying the student's prior at positions where student and teacher already agree.
2. **Bounded target-token weight.** In the wrong-student regime ($p^\star < m$), the closed form satisfies
$$c^\dagger = \frac{m}{1 - p^\star + m} \in \left[\tfrac{m}{1+m}, m\right],$$
so the SFT-branch coefficient *never exceeds* m: even when the target is deeply implausible under the student prior, P^{3}D refuses to push a full one-hot signal. In the extremal case where $m \to 0$ (student's distribution is diffuse across many alternatives), we have $c^\dagger \to 0$ and the loss reduces to pure self-distillation against $\pi_{-y^\star}$. This mildness is a design feature—attempting to close a large probability gap in a single SFT step would require KL-costly transports that empirically manifest as gradient-norm spikes and post-SFT forgetting.
3. **Adaptive, self-referential weighting.** The required correction depends on the demonstrated token's probability relative to its strongest competitor. When $p^\star < m$, the increase in target probability is
$$c^\dagger - p^\star = \frac{(1 - p^\star)(m - p^\star)}{1 - p^\star + m}.$$
With the current student as the detached reference, the target-logit gradient is $p^\star - c^\dagger$, so this increase is also the magnitude of the per-token target-logit gradient. For fixed $p^\star$, a stronger competitor requires a larger correction; for fixed m, the correction decreases to zero as $p^\star$ approaches m from below. Once dominance is satisfied, $Q^\dagger = P$ and the local logit gradient vanishes. Thus, the zero-margin construction supplies a supervisory target with only the probability adjustment needed to reach dominance within the uniform-rescaling family, while preserving the relative probabilities of non-target alternatives. With a positive margin, the corresponding ideal correction is $c^\star - p^\star = \max\{0, b + \delta - p^\star\}$: it vanishes once the stronger target-probability floor is met, before numerical clipping.

B.9 Summary

Our construction separates two aspects of supervision: how strongly to prefer the demonstrated token, and how to distribute the remaining probability among alternatives. The demonstration guides

the former, while the current student supplies the latter. Uniform rescaling preserves these relative preferences and permits a closed-form minimum-KL target within the chosen family; the margin controls the strength of demonstrated-token preference. This makes the scope of intervention explicit: supervision changes the target-versus-rest allocation while retaining the student's conditional structure among alternatives.

C Additional Details of the Pilot Study

This appendix provides supporting details for the pilot study: the corpus and taxonomy definitions, the rationale for comparing mathematical and medical demonstrations, the lexical characteristics of the four classes, and the case identities and visualization settings.

C.1 Corpus and Distributional Taxonomy

The taxonomy analysis uses 134,651 teacher-forced response-token positions from Qwen3-8B base on 786 prompt–response pairs across chat, medical and mathematical domains.

The Math and Medical pilot examples are drawn from the respective SFT training corpora described in Appendix D.2. For Chat, we use the `train_prefs` split of `HuggingFaceH4/ultrafeedback_binarized` as a general-instruction reference for comparing prior strength across domains.

For each response-token position, let $p_1 \geq p_2 \geq p_3 \geq \cdots$ denote the top-K probabilities of P_S in descending order, and let $\mathrm{rank}(y^\star)$ denote the rank of the target token in this order. The four structural classes and the residual bucket are defined by the following thresholds.

Table 2: Class definitions and population shares in the pilot corpus.

Class	Definition	Population share
A (SharpCorrect)	$p_1 > 0.5$ and $\mathrm{rank}(y^\star) = 1$	72.6%
B (SharpWrong)	$p_1 > 0.5$ and $\mathrm{rank}(y^\star) \neq 1$	14.6%
C (Bimodal)	$p_1 \leq 0.5$, $p_1 > 0.20$, $p_2 > 0.15$, $p_3 < 0.10$	3.2%
D (Multimodal)	$p_1 \leq 0.5$, $p_3 \geq 0.08$, excluding Class C	8.9%
Others	positions outside Classes A–D	0.7%

The sharp–spread split at $p_1 = 0.5$ separates positions where a single token dominates from positions where mass is distributed. Within the sharp regime, rank of $y^\star$ separates confirmatory (A) from corrective (B) positions. Within the spread regime, the top-3 probability separates two-way ambiguity (C) from higher-order multi-modality (D). Class D excludes positions already assigned to C, making the classes mutually exclusive. The residual class collects positions satisfying none of these definitions; it is small enough (0.7%) to be discussed only as a residual, without a dedicated analysis.

C.2 Domain Selection and Prior Differences

Motivation. Beyond Log Likelihood (Li et al., 2025a) shows that the relative merits of SFT objectives depend on the strength of the model's existing task-relevant priors, and highlights the limitations of validating objectives in a single domain. Following its model-capability continuum, we examine mathematics as a *model-strong* setting and medicine as a *model-intermediate* setting. Mathematical pretraining and familiar symbolic patterns provide a strong foundation for predicting mathematical demonstrations; medical reasoning draws on existing world knowledge but also requires more specialized knowledge and context-sensitive judgments. This contrast lets us study supervision where the prior already supports much of the demonstration and where learning must more frequently revise its preferences.

Operational criterion. We characterize this strong/intermediate distinction empirically on the untrained Qwen3-8B base model, using three prior-side statistics computed at every teacher-forced response-token position. Let P_S denote the base model's next-token distribution and let $y^\star$ denote the teacher-forced target.

- **Token prediction accuracy:** the fraction of positions with $\arg\max P_S = y^\star$, i.e. how often the model correctly predicts the demonstrated token.
- **Prior concentration:** the distribution of $P_S(y^\star)$ across positions, summarized by its mean and median.
- **Prior spread:** the distribution of $H(P_S)$ across positions, also summarized by mean and median.

Following the use of mean demonstration probability as a proxy for prior strength in Beyond Log Likelihood (Li et al., 2025a), we interpret these statistics jointly as evidence of how readily the model can use its prior to predict the demonstrations. Higher token prediction accuracy suggests greater familiarity with the domain and a stronger starting point for learning, although it is not itself a measurement of final-answer accuracy. All three statistics are measured before task-specific fine-tuning.

Measurements. Table 3 reports the three statistics on our two target domains (math, medical) together with a general-chat domain (Chat) used as a control. Figure 7 visualizes the structural taxonomy shares of Section 3.1 broken down by domain and the empirical CDF of $P_S(y^\star)$ per domain.

Table 3: Pre-fine-tuning statistics for Math, Medical, and Chat on Qwen3-8B Base.

Statistic	Math	Medical	Chat
Token prediction accuracy			
$\Pr[\arg\max P_S = y^\star]$	86.3%	69.1%	69.8%
Prior concentration on $y^\star$			
mean $P_S(y^\star)$	0.827	0.647	0.637
median $P_S(y^\star)$	0.998	0.877	0.825
$\Pr[P_S(y^\star) < 0.1]$	7.0%	19.2%	17.9%
Prior spread			
mean $H(P_S)$ (nats)	0.32	0.60	0.75
median $H(P_S)$	0.02	0.35	0.49
Class shares (Sec. 3.1)			
Class A (SharpCorrect)	83.8%	62.6%	62.8%
Class B (SharpWrong)	9.4%	21.7%	17.8%
Class C (Bimodal)	1.8%	4.1%	4.6%
Class D (Multimodal)	4.9%	11.1%	13.3%
Number of positions (N)	63,307	25,768	45,576

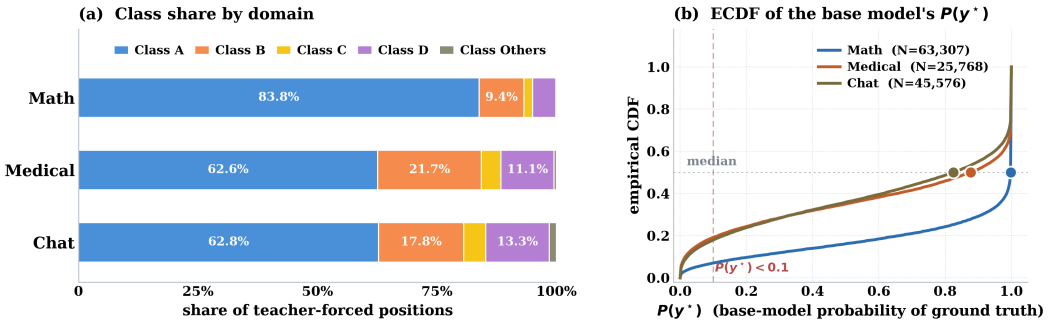


Figure 7: **Prior distributions by domain.** (a) Class shares. (b) Empirical CDFs of $P_S(y^\star)$; dots indicate medians.

The statistics suggest that the mathematical demonstrations are more familiar and locally easier for this model to predict. The model matches the demonstrated token at 86.3% of positions, and the

median $P_S(y^\star)$ is 0.998. Repeated symbolic constructions and familiar reasoning patterns plausibly contribute to this strong prior, making existing knowledge easier to exploit during fine-tuning. This interpretation is consistent with the model-strong mathematical settings studied in Beyond Log Likelihood (Li et al., 2025a). The prior is not uniformly reliable: 9.4% of positions belong to Class B, while 7.0% assign the demonstration less than 0.1 probability. Nevertheless, the predominance of aligned predictions suggests that much of the supervision builds on knowledge and patterns the model already possesses.

Medical reasoning presents a more mixed learning situation. The model still matches the demonstration at 69.1% of positions, indicating substantial usable prior knowledge, but Class B rises to 21.7% and Classes C/D together to 15.2%, compared with 9.4% and 6.7% in math. Medical terminology, contextual distinctions, and variation in explanatory phrasing plausibly make the next demonstrated token harder to predict. These observations support a model-intermediate interpretation: the model can draw on existing knowledge, yet encounters more positions where its preferred continuation needs revision or its alternatives deserve attention. Chat exhibits a similar profile, suggesting that the contrast is not restricted to medical terminology. Relative to math, these domains place greater demands on balancing demonstration learning with the retention of existing preferences.

Implication for the pilot. Math and Medical therefore expose different mixtures of the same supervision requirements. Math emphasizes learning from a comparatively reliable prior, while Medical more frequently combines usable prior knowledge with demonstration mismatch and distributed candidate mass. Both domains contain positions that call for preservation and positions that call for correction; neither should be assigned a single intervention wholesale. As Beyond Log Likelihood (Li et al., 2025a) demonstrates, an objective's advantage under strong priors need not persist when prior strength changes. **Examining both domains thus provides a more informative test of supervision than relying on the mathematical setting alone.**

C.3 Lexical Characteristics of the Four Classes

The taxonomy of Section 3.1 classifies each response-token position by the *shape* of P_S, but the shape itself carries no linguistic content. To close this gap, we look at the surface tokens that $y^\star$ actually takes in each class, restricting attention to the two domains studied in the main text (math and medical). For each (class, domain) pair we count target tokens, discard punctuation, stopwords and non-ASCII fragments, and assign every remaining surface token exclusively to the class whose *rate*, $\mathrm{count}(t, c) \,/\, |c|$, is largest which prevents dominant classes such as Class A on math from monopolizing every high-frequency token. We display the class-distinctive vocabulary in Figure 8. Word size reflects frequency in the assigned class and colors follow the class palette.

Class A: syntactic determinism versus lexical determinism. On math, the Class A cloud is dominated by LaTeX command tokens (`frac`, `sqrt`, `sin`, `cos`, `cdot`, `left`, `right`, `quad`, `triangle`, `quadratic`) and structural glue (`text`, `ify`, `ant`). These tokens suggest a contribution from familiar symbolic syntax: repeated command and delimiter patterns can make local continuations highly predictable. On medical, Class A is instead dominated by clinical nouns (`diagnosis`, `blood`, `disease`, `syndrome`, `injury`, `glucose`) and morphological suffixes (`itis`, `osis`, `emia`, `ology`, `ine`). The suffix positions are locally deterministic in the same sense (once "`pneumon`" has been emitted, "`ia`" is nearly forced), but the covering fraction of such deterministic constructions is far smaller than the fraction of LaTeX-constrained positions in math text. This provides a concrete linguistic explanation for the population gap of Section 3.1: Class A occupies 83.8% of math positions and only 62.6% of medical positions.

Class B: learning beyond confident predictions. Class B positions share the low-entropy region with Class A, but their preferred continuations disagree with the demonstration. In math, the cloud contains reasoning and output-structure tokens such as `First`, `equation`, `formula`, `Solution`, `boxed`, and `answer`, suggesting that these positions may carry learning signals for how solutions are organized and presented. In medicine, tokens such as `symptoms`, `condition`, and `test`, together with `associated` and `likely`, concern clinical concepts and how they are related or qualified. Such positions may therefore require learning domain knowledge and its expression, beyond changes in wording alone. These lexical patterns suggest that confident predictions may still contain errors in reasoning conventions, output formats, or domain-specific knowledge that

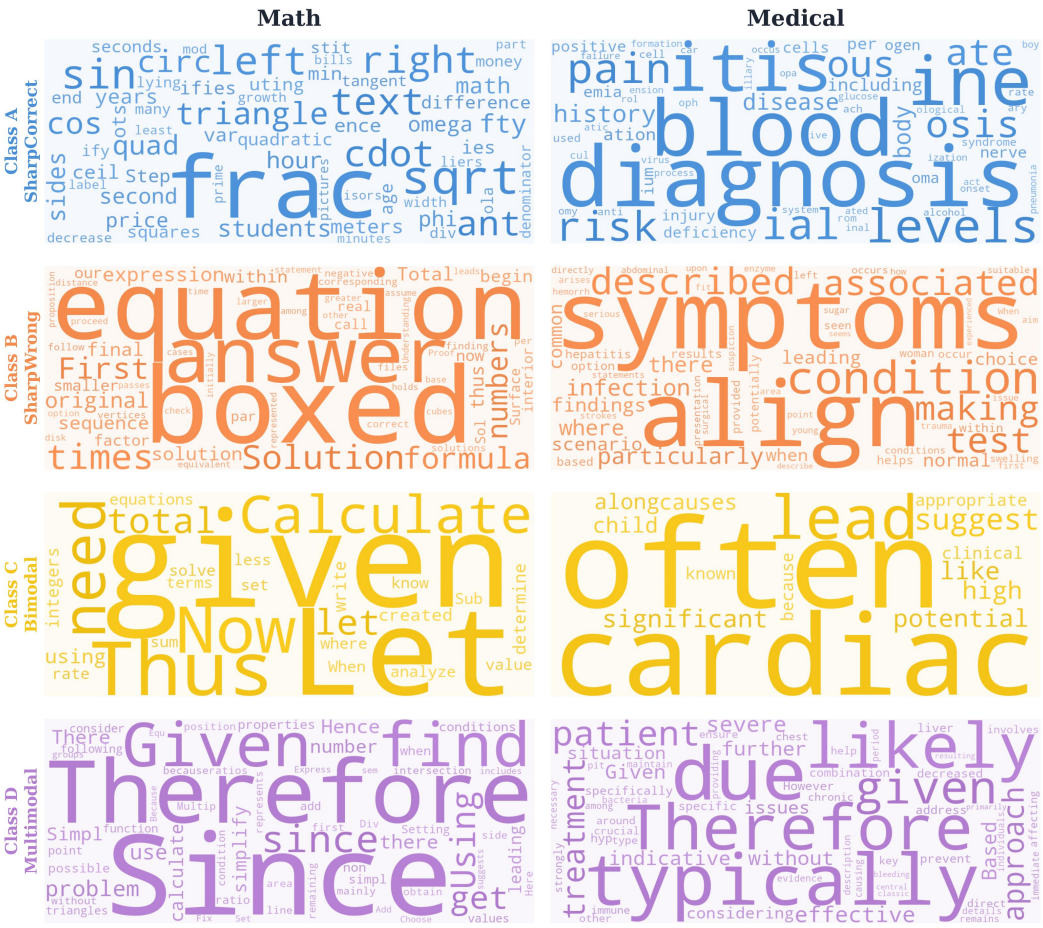


Figure 8: **Distinctive demonstration tokens.** Rows: Classes A–D. Columns: Math (left) and Medical (right).

require correction through demonstration learning. Low entropy alone therefore does not justify preserving the existing preference.

Classes C and D: interchangeable connectives and clinical hedges. Classes C and D contain positions where the base assigns non-negligible mass to several candidate tokens. The clouds show the types of demonstrated tokens occurring at these multi-candidate positions. On math, Class C is dominated by discourse markers that open a new derivation step (`Let`, `given`, `Now`, `Thus`, `Calculate`, `know`, `need`) and Class D by logical connectives that close one (`Since`, `Therefore`, `Given`, `Hence`, `Using`, `find`). Some of these expressions can serve similar discourse functions, suggesting room for variation in how a derivation is presented. On medical, Class C surfaces clinical hedges paired with specific nouns (`often`, `cardiac`, `lead`, `suggest`, `significant`, `potential`, `appropriate`, `high`) and Class D shows clinical reasoning tokens (`Therefore`, `typically`, `likely`, `due`, `severe`, `indicative`, `treatment`, `chronic`, `effective`). These lexical patterns suggest that Class C/D positions may admit synonymous expressions or alternative logical connectives, giving the model several ways to formulate and link reasoning steps. Such choices can provide an important source of diversity in reasoning expression. Preserving entropy at such connective-token positions has been motivated as a way to maintain exploration opportunities during subsequent RL. Learning the demonstrated continuation should therefore be accompanied by an effort to preserve the original range of plausible alternatives, rather than unnecessarily narrowing the available ways to continue (Wang et al., 2025b).

Table 4: Contexts and leading candidates for Cases A–D.

Class	Context before the token	Demonstration	Leading candidates
A	...the remainder is 0, indicating that the 100th ...	`digit`	`digit`, `place`, `position`
B	...the 10-year-old boy with atypical ...	`pneumonia`	`symptoms`, `pneumonia`, `presentation`
C	...weaker, delayed pulses in the lower ...	`body`	`body`, `extrem`, `limbs`
D	...the patient's new symptoms of rash and fever ...	`following`	`following`, `,`, `after`

C.4 Illustrative Cases in Context

The statistics and word clouds summarize patterns without the sentence context that gives individual choices their meaning. Table 4 places the four examples from Figure 2 in context, pairing each demonstrated token with the model's leading candidates. The figure shows their probability distributions; the table retains token fragments such as `extrem` as displayed there.

In A, the model already favors `digit`, agreeing with the demonstration's description of a decimal expansion. In B, it favors `symptoms`, while the demonstration specifies `pneumonia`. The latter choice names a disease concept rather than a generic description, illustrating how a confident continuation can still require adjustment toward the demonstration. Read together, A/B show that similar concentration can accompany different learning needs. The CurioSFT teacher overlays in the figure use its adaptive temperature scaling on top-100 distributions to illustrate concentration adjustment.

C and D make the range of continuation choices visible. In C, `body` and `extrem` carry the two dominant peaks, suggesting different ways to describe the location of the clinical finding. In D, `following` and `after` offer related ways to introduce temporal context, while the comma opens another syntactic continuation. These alternatives need not be interchangeable in every completed sentence, but they illustrate the expressive possibilities represented by the original distribution. Learning the demonstrated choice should therefore go together with retaining plausible alternatives that allow reasoning to be formulated and connected in different ways.

D Training Configuration Details

This section documents the setup used across all Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) experiments. To isolate the effect of the training objective itself, we hold every non-objective factor fixed within each domain: identical base models, identical data, identical optimizer schedule, and identical decoding evaluation. Method-specific hyperparameters (e.g., the entropy weight for entropy regularization or the teacher temperature for CurioSFT) follow the values reported in each method's original paper.

D.1 Base Models

We conduct experiments on three base models from two families: Qwen3-4B-Base, Qwen3-8B-Base (Yang et al., 2025), and Meta-Llama-3-8B (Grattafiori et al., 2024). We choose these models to cover different levels of pretrained prior strength. Prior comparisons suggest that Qwen3 base models are generally stronger and exhibit richer priors than comparable Llama3 models (e.g., Gandhi et al. (2025)). This setup therefore allows us to test whether the benefits of P^3D hold across models with different initial capabilities, rather than being tied to a single backbone. In addition, the two Qwen3 models allow us to examine the effect of scale within a stronger model family.

D.2 Training Datasets

We train and evaluate P^3D in two complementary domains, chosen to stress different aspects of the reasoning distribution. (Details in Appendix C.2)

Mathematics. For the SFT stage, we use a curated 143K-example chain-of-thought corpus assembled from OpenR1-Math-220k (Mirror, 2025) and NuminaMath-CoT (LI et al., 2024). Each example consists of a mathematical problem paired with a long-form solution that includes intermediate reasoning. For the subsequent RL stage, we use the OpenR1-Math prompt split distributed with the verl framework (Sheng et al., 2024), formatted for policy gradient training.

Medical. For the SFT stage, we use the medical-o1-reasoning-SFT corpus (Chen et al., 2024), which contains $\sim$ 43K USMLE-style clinical vignettes paired with detailed chain-of-thought answers distilled from a strong medical-reasoning model. For the subsequent RL stage, we use a rule-based reward defined over MedQA (Jin et al., 2021) train-set, where the ground-truth label is used to score each rollout.

D.3 Supervised Fine-Tuning (SFT)

Framework. We use LLaMA-Factory (Zheng et al., 2024) for all SFT experiments. Training is distributed across 8 NVIDIA H800 GPUs with DeepSpeed ZeRO Stage-3 (Rasley et al., 2020) for memory-efficient full-parameter fine-tuning.

Hyperparameters. Table 5 summarizes the SFT training hyperparameters. *All* methods share the same configuration within each domain; only the training objective differs.

Table 5: Shared SFT training hyperparameters across all methods and model families.

Hyperparameter	Value
Fine-tuning type	Full parameter
Parallelism	DeepSpeed ZeRO-3
Number of GPUs	8 (H800)
Effective batch size (math / medical)	512 / 128
Learning rate	1×10^{-5}
LR scheduler	Cosine
Warmup steps	60
Number of epochs	3
Max sequence length	4096
Precision	BF16
Gradient checkpointing	Enabled

For P^{3}D, target construction and fitting operate on a top-K candidate support with $K = 200$. We select the highest student logits and replace the lowest selected token with the demonstration if it is absent. Both the detached student reference and the prediction used for fitting are normalized on this same support, so non-target ratio preservation applies to the retained candidates. This reuses the student's forward pass without an additional teacher evaluation. Our EntropyReg, CurioSFT, and GEM implementations share the same candidate-selection routine, with support sizes configured per objective. Numerical guards and margin handling are detailed in Appendix B.7.

SFT Training Time. On 8 H800 GPUs, a single SFT run takes roughly 4 hours for medical (Qwen3-4B) and 11–15 hours for math (across the three families). Wall-clock differences among all methods are within $\pm 5\%$ of HardSFT, except ASFT, which is $\sim 30\%$ slower on math due to its additional base-model forward pass for the $D_{\mathrm{KL}}(P_{\mathrm{base}} \| \pi_\theta)$ term.

D.4 Reinforcement Learning (GRPO)

Framework. We use verl (Sheng et al., 2024) for all RL experiments with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which eliminates the need for a separate critic model by estimating advantages from group-level relative rewards.

Reward Function. We use a rule-based binary reward. For each prompt the model generates $n = 8$ rollouts at sampling temperature 1.0; each response is scored 1 (correct) or 0 (incorrect)

by comparing its extracted answer against the ground-truth reference. Answer extraction primarily parses the last `\boxed{...}` span, with a numeric-tolerance fallback for mathematical answers. The same reward configuration is applied to every SFT initialization, so that any downstream performance difference reflects the quality of the SFT warm start rather than RL-side tuning.

Hyperparameters. Table 6 summarizes the shared GRPO hyperparameters.

Table 6: Shared GRPO training hyperparameters.

Hyperparameter	Value
Framework	verl (GRPO)
Rollouts per prompt (n)	8
Rollout temperature	1.0
Max response length (math / medical)	8192 / 4096
Actor learning rate	1×10^{-6}
KL loss coefficient	0.001
Train batch size	128
PPO mini-batch size	32
Save / test frequency	every 40 steps
Training steps (math / medical)	300 / 157
Precision	BF16

Fair Comparison Across Methods. The same GRPO configuration is applied to *all* SFT initializations, including HardSFT, Label Smoothing, GEM, DFT, ASFT, CurioSFT, and P^{3}D. For math experiments we report the checkpoint at step 300; for medical experiments, where the training signal saturates earlier, we report the checkpoint at step 157. When comparing families and scales we align comparison points at the same training step to prevent any method from benefiting from extra optimization budget.

E Evaluation Setting

E.1 Evaluation Framework

We conduct all evaluations using the `lm-evaluation-harness` framework (Gao et al., 2024), which provides standardized task implementations, prompt templates, answer extraction, and metric aggregation. All experiments are executed via the `vLLM` (Kwon et al., 2023) backend to ensure efficient batched generation and consistent decoding behavior across models.

E.2 Benchmarks

To comprehensively evaluate both the reasoning abilities targeted by P^{3}D and its side effects on other capabilities, we select benchmarks spanning three complementary categories: *in-domain reasoning* in mathematics and medicine, and *pretrain-likelihood preservation*. Table 7 summarizes all benchmarks.

E.3 Metrics and Aggregation

Reported metrics. For each open-ended benchmark we report two quantities: (i) greedy accuracy at $T = 0$, which we denote *pass@1*; and (ii) two aggregated metrics computed from five independent samples at $T = 0.6$, namely *pass@5* (the probability that at least one of the five samples is correct) and *mean@5* (the arithmetic mean accuracy across the five samples). These three quantities together characterize both the deterministic-decoding quality of a model and the breadth of its sampled distribution, which is the property most directly relevant to downstream RL.

Answer extraction. All in-domain benchmarks parse the last `\boxed{...}` span produced by the model as the final answer, falling back to the harness's default filter chain if no boxed span is emitted. This unified extraction protocol removes prompt-format artifacts across methods.

Table 7: Benchmarks used in our evaluation, grouped into three categories. In-domain benchmarks are evaluated in a free-form CoT setting; pretrain-likelihood benchmarks are scored under the standard log-likelihood protocol.

Benchmark	HF Dataset	Dataset Summary	Primary Abilities Tested
In-Domain Reasoning: Mathematics			
MATH-500	MATH-500	500 competition-level problems curated from MATH Hendrycks et al. (2021b), spanning 7 subjects (algebra, geometry, number theory, etc.).	Formal math reasoning; symbolic manipulation.
GSM8K	gsm8k	Grade-school math word problems (1,319 test) requiring 2–8 steps of elementary arithmetic Cobbe et al. (2021).	Multi-step arithmetic; text-to-equation mapping.
AIME 2024	AIME_2024	30 problems from AIME 2024 (I+II) with integer answers, requiring creative multi-step reasoning in algebra, combinatorics, and geometry.	Competition-level problem solving.
AIME 2025	aime25	30 problems from AIME 2025, temporally disjoint from typical pretraining corpora.	Generalization to unseen contest problems.
Minerva Math	minervamath	Undergraduate-level quantitative-reasoning problems from Minerva Lewkowycz et al. (2022).	Applied and derivative-style reasoning.
In-Domain Reasoning: Medical			
MedQA	med_qa	1,273 USMLE-style multiple-choice questions from medical licensing exams Jin et al. (2021).	Clinical knowledge; multi-step medical reasoning.
MedMCQA	medmcqa	4,183 questions covering 2,400 topics across 21 medical subjects Pal et al. (2022).	Medical subject-matter knowledge.
PubMedQA	PubMedQA	500 biomedical yes/no/maybe questions grounded in scientific abstracts Jin et al. (2019).	Reading comprehension over biomedical text.
Pretrain-Likelihood Preservation			
ARC-Challenge	ai2_arc	Challenge split (1,172 questions) of the AI2 science exam benchmark Clark et al. (2018), filtered to defeat retrieval-based methods.	Scientific reasoning; commonsense knowledge.
HellaSwag	hellaswag	Commonsense completion problems Zellers et al. (2019).	Everyday commonsense inference.
MMLU	mmlu	57-subject knowledge benchmark spanning STEM, humanities, and professional exams Hendrycks et al. (2021a).	Broad-coverage academic knowledge.
MMLU-Pro	MMLU-Pro	Leaderboard configuration of MMLU-Pro, scored via log-likelihood.	Cross-domain knowledge (likelihood mode).
GPQA-Diamond	gpqa_diamond	The diamond subset of GPQA, the highest-quality expert-agreed slice Rein et al. (2023).	Expert-level scientific likelihood.
TruthfulQA-MC1 / MC2	truthful_qa	817 questions probing common misconceptions Lin et al. (2022). MC1 uses a single-correct-answer format; MC2 uses multi-truth aggregation.	Truthfulness under adversarial priors.

Table 8: **Core choices in SFT supervision.** Benefits and limitations refer to the objective's design.

Method	Core choice	Intended benefit	Remaining limitation
HardSFT	One-hot supervision	Direct imitation	Does not reserve mass for alternatives
LS	Uniform soft target	Reduce overconfidence	Treats different alternatives alike
EntropyReg	Reward uncertainty	Discourage concentration	Entropy does not specify candidate identity
GEM	Target–competitor game	Moderate competitive updates	Does not require non-target ratio preservation
CurioSFT	Entropy-guided self-distillation	Adapt uncertainty retention	Entropy alone does not determine label agreement
DFT	Weight by target confidence	Attenuate unlikely-label updates	Can also attenuate necessary correction
EAFT	Weight by entropy	Reduce updates at confident positions	Confidence can accompany a wrong preference
ASFT	Demonstration learning with reference anchoring	Limit reference drift	The compromise need not make the target dominant

`weight_by_size` aggregation. For benchmarks composed of multiple subsets (e.g., BBH, MMLU, MMLU-Pro), we follow the `lm-evaluation-harness` `weight_by_size` strategy and aggregate subset accuracies weighted by the number of examples per subset:

$$\mathrm{Acc}_{\mathrm{weighted}} = \sum_{s=1}^{S} \frac{N_s}{\sum_j N_j} \cdot \mathrm{Acc}_s. \tag{28}$$

E.4 Inference Configuration

For open-ended generation we use `top_p`$= 1.0$ and `top_k`$= -1$ (no truncation) for both $T = 0$ and $T = 0.6$ decoding, with a maximum generation length of $16{,}384$ tokens to give the model sufficient headroom to complete a full chain of thought without truncation. Pretrain-likelihood benchmarks are scored via log-likelihood without any sampling; we report the harness's native per-task metric.

F A Unified View of Fine-Tuning Objectives

This appendix compares the principles behind the objectives: what each method asks the model to learn, what it tries to retain, and where these choices can become limiting. We begin with their defining supervision mechanisms and then relate them to acquisition and retention. A limitation of an objective is not a claim that the method fails in every setting; the experimental results describe its behavior in our particular regime.

F.1 Targets, weights, and auxiliary objectives

At a fixed context, let $p = \pi_\theta(\cdot \mid s_t)$ denote the student prediction, $y^\star$ the demonstrated token, and $e_{y^\star}$ its one-hot target. Write P for a reference distribution when a method uses one. Many SFT objectives can be organized as

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \sum_t \left[w_t \, \mathrm{CE}(Q_t, p) + R_t \right], \qquad \mathrm{CE}(Q, p) = -\sum_v Q(v) \log p(v). \tag{29}$$

The three components answer different questions: Q_t specifies what to learn, w_t determines how strongly to learn at a position, and R_t expresses an additional preference such as avoiding overconfidence or reference drift. We use $\mathrm{sg}[\cdot]$ for stop-gradient, $H(p)$ for entropy, and U for a uniform distribution over the vocabulary. These categories overlap: a regularized objective can sometimes be rewritten using an effective target, and an online target construction can induce an equivalent weighting rule. The useful comparison is therefore which behavior a method explicitly controls.

F.2 Hard labels and distribution smoothing

HardSFT: concentrate on the demonstrated choice. Standard SFT (Ouyang et al., 2022) minimizes $\ell_{\mathrm{NLL}} = -\log p(y^\star)$, equivalently cross-entropy against $e_{y^\star}$. Its strength is a direct learning signal: every supervised position asks the model to assign more probability to the demonstration. For an independently fitted position, the optimum is one-hot. The same preference for concentration persists after the demonstrated token is already top-1, so the objective does not distinguish correcting a mismatch from further sharpening an already-aligned prediction. It supplies acquisition explicitly but leaves retention of alternatives unspecified.

Label Smoothing (LS): soften the target uniformly. LS (Szegedy et al., 2016) replaces the one-hot target with

$$Q_{\mathrm{LS}} = (1-\varepsilon)e_{y^\star} + \varepsilon U.$$

The demonstration remains preferred, but the model is no longer asked to assign it unit probability. This can reduce overconfidence. The limitation lies in where the remaining mass goes: all non-target tokens receive the same target probability, regardless of whether the student considered them plausible. LS specifies a soft distribution, but its uniform alternative component does not retain the student's relative preferences.

EntropyReg: discourage concentration in the prediction. Entropy regularization (Jost, 2006) adds an uncertainty reward,

$$\ell_{\mathrm{ent}} = -\log p(y^\star) - \alpha H(p).$$

Unlike LS, which changes the supervisory target, this objective rewards entropy in the student's own prediction. It can counter excessive concentration while the NLL term still promotes the demonstration. However, entropy measures the amount of uncertainty, not which candidates carry it. Distributions with the same entropy can favor different tokens; a confidently correct and a confidently wrong position can therefore look alike to the entropy term. This is the distinction illustrated in Section 3.1.

F.3 Candidate competition and adaptive self-distillation

GEM: regulate learning through candidate competition. GEM (Li et al., 2025c) motivates fine-tuning through a game between the demonstrated choice and competing continuations. Its contrastive supervision can be represented by a weighted comparison of their log-probabilities:

$$\ell_{\mathrm{GEM}} = \sum_j a_j \big[\log p(j) - \log p(y^\star)\big], \qquad q_j = \frac{\exp(z_j/\beta)}{\sum_k \exp(z_k/\beta)}. \tag{30}$$

Here a_j denotes the detached competitor weight; the temperature-shaped profile q and the choice of comparison function determine these weights. The important idea is that alternatives participate differently according to the student's competitive landscape. GEM is therefore more selective than uniformly rewarding entropy, and is not simply a scalar weight on ordinary CE. The temperature β controls how concentrated the competitor weighting is. These choices have a game-theoretic motivation, but are still overly heuristic. In particular, modeling target–competitor relations does not explicitly protect the relations among competitors themselves.

CurioSFT: learn the demonstration while retaining uncertainty. CurioSFT (Wang et al., 2026a) combines NLL with adaptive self-distillation:

$$\ell_{\mathrm{curio}} = -\log p(y^\star) + \alpha\, \mathcal{D}_{\mathrm{SD}}(p, \mathrm{sg}[Q_{\tau_t}])\,. \tag{31}$$

Here Q_{τ_t} is a temperature-adjusted self-teacher, with temperature chosen according to an entropy criterion, and $\mathcal{D}_{\mathrm{SD}}$ denotes the self-distillation penalty. The two terms have complementary roles: NLL acquires the demonstrated behavior, while the teacher signal counteracts excessive loss of uncertainty. This permits more adaptive softening than a fixed uniform target. Its limitation is that entropy alone does not determine whether the demonstration agrees with the student, or which non-target relations should remain intact. The NLL term still provides directed correction; the point is not that the complete objective ignores the demonstration, but that its uncertainty-control mechanism does not itself specify the desired acquisition–retention structure.

F.4 Reweighting and reference anchoring

DFT: attenuate supervision using target confidence. DFT (Wu et al., 2025) weights NLL by the student's detached probability of the demonstration:

$$\ell_{\mathrm{DFT}} = -\mathrm{sg}[p(y^\star)] \log p(y^\star).$$

Its defining choice is to reduce the influence of labels the student finds unlikely. This can avoid large, disruptive updates, but an unlikely label may also be precisely the correction the model needs. Here the gradient makes the trade-off explicit:

$$\frac{\partial \ell_{\mathrm{DFT}}}{\partial z_{y^\star}} = -p(y^\star)\big(1 - p(y^\star)\big).$$

It becomes small both when the demonstration is already highly probable and when it is extremely unlikely. Thus, attenuation can suppress redundant learning or needed correction, depending on the position.

EAFT: allocate learning according to uncertainty. EAFT (Diao et al., 2026) uses entropy to adapt the weight of hard-label learning. Its central mechanism is represented by

$$\ell_{\mathrm{EAFT}} = -\mathrm{sg}[f(H(p))] \log p(y^\star),$$

where f increases the learning weight with uncertainty. The intuition is to devote less effort to confident positions and more to uncertain ones. This can reduce unnecessary updates when confidence reflects agreement with the demonstration. But a confident mismatch also has low entropy: the same rule can downweight a position that needs correction. EAFT changes where learning is emphasized while retaining a one-hot target; entropy does not by itself determine the required change in preference.

ASFT: competing pressures in acquisition and retention. ASFT (Zhu et al., 2025) combines confidence-weighted demonstration learning with a KL anchor to a reference distribution P:

$$\ell_{\mathrm{ASFT}} = a_t \,\mathrm{CE}(e_{y^\star}, p) + \beta D_{\mathrm{KL}}(P \| p), \qquad a_t = \mathrm{sg}[p(y^\star)]. \tag{32}$$

ASFT explicitly represents both acquisition and retention. However, when the reference favors a competing token, acquiring the demonstration requires moving away from precisely the distribution that the KL term encourages the student to retain. As the student departs from the reference, the two terms can exert opposing optimization pressures: demonstration learning promotes the new preference, while anchoring resists that change. Their weighted sum leaves this tension to be resolved during optimization, so a needed correction can be weakened by the retention penalty. The resulting compromise is visible in the equivalent form

$$\ell_{\mathrm{ASFT}} = (a_t + \beta)\,\mathrm{CE}\left(\frac{a_t e_{y^\star} + \beta P}{a_t + \beta}, p\right) - \beta H(P), \tag{33}$$

for $a_t + \beta > 0$. If the reference disagrees strongly and the demonstration weight is small, this mixture need not make the demonstrated token top-1. Thus, explicitly including both goals in the loss does not ensure that the balance between their competing pressures meets the required learning outcome. This is a limitation of treating retention as an additive penalty: its strength controls how much correction is resisted, without specifying which changes are necessary and which relationships should be preserved. A constrained target construction addresses this tension before fitting the student by selecting a distribution that satisfies the stated acquisition and preservation requirements within its feasible family. The distinction is the explicit specification of that compromise, rather than the mere existence of a single effective target.

F.5 Methods combining SFT and RL

RFT and Prefix-RFT: obtain better training trajectories. RFT (Trung et al., 2024) selects model-generated responses using reward and learns from the selected outputs. With a binary verifier, it emphasizes successful rollouts. This can exploit solutions the model can already produce, but its coverage depends on finding enough useful samples. Prefix-RFT (Huang et al., 2025) uses expert prefixes to guide continuation generation, helping the model reach trajectories that are difficult to discover without guidance. Its benefit depends on the prefixes and the resulting continuations. Both methods primarily change where supervision comes from or which contexts it covers. They are complementary to designing the probability distribution imposed at each supervised position.

F.6 Empirical failure patterns of prior objectives

Reading the deltas of each baseline against HardSFT in our main table (Qwen3-4B/Qwen3-8B/Llama3-8B $\times$ math/medical $\times$ SFT/GRPO, 12 cells in total; Δ denotes gap to HardSFT on the domain-averaged metric), three failure typologies emerge. They line up with the objective-level gaps summarized in Table 8. All gaps are in percentage points, computed from the table's unrounded source scores and rounded only for reporting.

Attenuated correction on Llama3-8B Math (DFT, ASFT). Both methods scale the corrective signal by a confidence-like quantity that is small exactly where correction is most needed. DFT's target-logit gradient $-p(y^\star)\big(1 - p(y^\star)\big)$ vanishes as $p(y^\star) \to 0$, attenuating precisely the low-confidence positions the demonstration is supposed to fix. ASFT compounds this: a confidence-weighted CE term attenuates updates on the same low-$p(y^\star)$ positions where the base-anchoring $D_{\mathrm{KL}}(P_{\mathrm{base}} \| \pi_\theta)$ term simultaneously pulls the student back toward the uncertain base. At the SFT stage, both methods have their largest greedy and pass@5 regressions on Llama3-8B Math: DFT records $\Delta T{=}0 = -13.8$ / $\Delta\text{pass@5} = -22.2$, and ASFT records $\Delta T{=}0 = -25.5$ / $\Delta\text{pass@5} = -20.9$. ASFT's 25.5-point greedy decline is the largest SFT-stage greedy regression among the compared baselines. On Qwen3 Medical, the outcome depends on the metric and stage: both methods' SFT greedy scores remain close to HardSFT, but DFT loses 10.2 pass@5 points on Qwen3-4B. After GRPO, DFT's Medical greedy scores trail HardSFT by 9.0 points on Qwen3-4B and 4.7 points on Qwen3-8B.

Greedy regression with variable sampling gains (EntropyReg). EntropyReg's entropy bonus $-\alpha H(p)$ encourages uncertainty without distinguishing confident agreement from confident disagreement with the demonstration. Across the 12 settings, its mean gap to HardSFT is -3.9 points in greedy accuracy and $+0.6$ points in pass@5. Its largest greedy regression occurs on Llama3-8B Medical after GRPO (-18.5 points). Thus, the modest average gain in five-sample coverage coexists with lower average greedy accuracy and a substantial regression in this setting.

Unbounded self-distillation with a lagging teacher — universally weak on sampling (CurioSFT). The auxiliary term $\mathcal{D}_{\mathrm{SD}}(p, \mathrm{sg}[Q_\tau])$ is unbounded, and the self-teacher lags the student; small early divergences are amplified into a runaway suppression of non-target log-probabilities. The failure is base- and domain-agnostic: Δpass@5 is negative in all 12 (family, domain, stage) settings we evaluate (mean -16.6 points), while greedy accuracy also declines by 10.2 points on average. Both metrics therefore regress on average, with a larger loss in five-sample coverage. Sampling performance is particularly affected in these runs; greedy accuracy is not uniformly preserved.

Modest aggregate changes with setting-dependent outcomes (LS, EAFT, GEM$_{\text{game}}$). LS's uniform smoothing and EAFT's entropy-adaptive downweighting yield positive average gaps to HardSFT across the 12 settings: LS gains $+0.6$ / $+0.9$ points in $T{=}0$ / pass@5, and EAFT gains $+1.4$ / $+1.5$ points. These averages do not imply improvement in every setting. GEM has mean gaps within ± 1 percentage point of HardSFT across the 12 settings (-0.4 points in both greedy accuracy and pass@5), but its individual results also vary: Qwen3-8B Math SFT loses 6.6 greedy points, and Llama3-8B Math SFT loses 7.6 pass@5 points. Its aggregate proximity to HardSFT therefore does not imply that every result lies within one point of the baseline. None of these aggregate comparisons alone establishes preservation of non-target preferences.

Take-away. The observed patterns complement the objective-level gaps in Table 8: confidence-based attenuation can accompany weak demonstration learning, entropy regularization does not consistently improve greedy accuracy, and CurioSFT exhibits substantial losses in sampled coverage. The variation across models, domains, and stages motivates considering acquisition and retention jointly when specifying the supervisory target.

F.7 Summary: from separate controls to a specified target

Viewed through Eq. equation 29, the methods above control parts of the acquisition–retention problem. HardSFT specifies what to acquire but leaves alternative-token retention unspecified. LS,

EntropyReg, and CurioSFT soften supervision or regulate uncertainty, without making preservation of non-target preferences an explicit requirement. DFT and EAFT adjust learning strength, but confidence and entropy alone do not determine whether a position needs correction. GEM models target–competitor interactions without explicitly retaining relationships among the competitors. ASFT directly encourages reference retention, but can leave acquisition and retention as competing pressures whose weighted compromise does not ensure demonstrated-token dominance. Trajectory-selection methods address a complementary question: which examples and contexts provide supervision.

These distinctions motivate specifying the desired distribution before optimizing the student. A target should make the demonstrated preference explicit while identifying the alternative structure that should remain unchanged. We advocate resolving this balance through a constrained construction: state the acquisition and retention requirements, then compute the minimum-KL target within the resulting feasible family. In our case, retaining the non-target conditional distribution gives a closed-form correction of the target-versus-rest allocation. This makes the intended intervention explicit, rather than leaving it entirely to a proxy statistic or a compromise between loss terms. The target's properties are established by construction; how well training realizes them remains an empirical question.

G Acquisition and Prior Retention

This appendix supplements the main-text acquisition–retention analysis with distributional probes, local examples, and general-capability results. The probes use Qwen3-4B on Math and Medical; the evaluation tables also include Qwen3-8B and Llama3-8B. Each probe holds its own contexts fixed across methods. Teacher-forced demonstration positions, Base-generated trajectories, and model-generated responses answer different questions and are distinguished below. Learning dynamics are presented separately in Appendix H.

G.1 Distributional probes and prior retention

Distribution-shift projection settings in the main text. Figure 4(a) uses 19,859 shared Math teacher-forced demonstration positions. At each position, we retain token IDs recorded for all compared checkpoints, include the demonstrated token, and aggregate the remaining probability into one bin. A rank-two SVD of the Base-relative square-root probability vectors gives the displayed coordinates. The reported RMS Hellinger distances are computed before projection on these coarsened distributions. The background is a fitted visual guide, not an evaluation of models between the observed checkpoints.

Rank-shift and lexical probes. For the quantitative probe, Base generates trajectories from GSM8K and MedQA prompts, and every post-GRPO model is evaluated on those same prefixes. Let η be the trained model's rank of the Base-selected token. We classify positions as *unshifted* ($\eta = 1$), *marginal* ($\eta \in \{2, 3\}$), or *shifted* ($\eta > 3$). This measures preference retention, not whether the Base token is correct. The complementary lexical probe uses P^{3}D-generated MMLU responses and reverses the ranking query: it ranks each generated token under Base, using the same three categories. Each token type is assigned to its majority category, with word size proportional to the square root of its occurrence count. Thus, the two probes share category thresholds but differ in context source and ranking direction. Table 9 reports the quantitative results, and Figure 9 presents the lexical patterns for both domains.

Retention with selective changes. Table 9 shows that P^{3}D makes fewer large rank changes than HardSFT and the other non-anchored baselines in both domains, while retaining a substantial marginal group. In particular, when the preferred token changes, the original Base choice often remains among the leading alternatives rather than being displaced far down the ranking. This is consistent with our target construction: promoting the demonstrated token can move the previous favorite into a runner-up position while preserving relationships among non-target candidates. On Medical, P^{3}D has the largest marginal share among the compared methods; on Math, it combines increased top-1 agreement with fewer large shifts relative to HardSFT. Together with the task im-

Table 9: **Token-distribution shift after GRPO (Qwen3-4B).** Bold/underlined values rank first/second: descending for Unshifted and Marginal, ascending for Shifted. Marginal ranking does not imply quality.

Method	**Math** (gsm8k prompts)			**Medical** (medqa prompts)		
	Unshifted	Marginal	Shifted	Unshifted	Marginal	Shifted
HardSFT	80.2	**9.4**	10.4	89.1	6.7	4.3
LS	82.0	7.2	10.8	88.6	<u>7.3</u>	4.2
GEM	84.2	7.0	8.8	89.1	7.0	3.9
DFT	81.9	<u>8.4</u>	9.6	89.8	6.0	4.1
EAFT	81.8	5.4	12.8	89.9	6.2	3.9
ASFT	**90.1**	5.1	**4.8**	**93.6**	5.1	**1.3**
P^3D (Ours)	<u>85.7</u>	7.2	<u>7.1</u>	<u>90.0</u>	**7.5**	<u>2.5</u>

provements in Table 1, these patterns support learning new preferences while retaining the Base model's leading alternatives.

Retention must be paired with acquisition. ASFT is the exception to the rank-retention comparison: its reference anchor keeps it closer to Base than P^3D, but its task accuracy is lower in both reported Qwen3-4B post-GRPO domain comparisons. This comparison highlights the value of balancing acquisition with retention: P^3D learns the demonstrated behavior while keeping existing preferences available among the leading alternatives. The other baselines generally exhibit more large shifts without matching this joint task-performance and retention outcome.

Acquiring reasoning patterns while retaining prior preferences. The wordclouds in Figure 9 provide a lexical account of the strong acquisition–retention balance achieved by P^3D. Shifted expressions such as *try*, *recall*, and *Okay* are associated with attempting a reasoning step, invoking relevant knowledge, and organizing thought. Their prominence, together with the task improvements in Table 1, supports effective acquisition of reasoning-related discourse patterns. Meanwhile, the unshifted group contains task-framing words such as *question* and *options*, function words, digits, and content terms. Agreement with Base at these positions shows that the new behavior coexists with retained preferences in both task expression and content-bearing continuations. The marginal group further includes *Let* and *First*, which introduce or sequence reasoning steps: the post-GRPO model selects these expressions from candidates already ranked highly by Base. Subsequent adaptation therefore continues to draw on the existing repertoire of plausible continuations. Taken together, the lexical patterns and quantitative results support the central benefit of P^3D: strong acquisition of demonstrated behavior alongside retention of the prior preferences on which further learning can build.

G.2 Acquisition across base-prediction classes

We now examine whether the distributions realized after training reflect the acquisition–retention behavior intended by our two design principles, *Greedy Guarantee* and *Minimal Intervention*, and the resulting target construction (Section 4): learning the demonstrated preference while limiting unnecessary changes to already-aligned predictions and retaining probability mass on alternative continuations.

Using the A–D taxonomy introduced in Section 3.1, we examine how training changes demonstrated-token probabilities across base-prediction classes. We use teacher-forced probes on 200 demonstrations per domain, with at most 100 response positions per demonstration. Class assignments are determined once from the base distribution and held fixed across all trained models.[2] The probe contains 14,764/2,473/769/1,853 Math positions and 11,811/3,217/1,005/3,967 Medical positions in A/B/C/D, respectively.

[2]For this probe, A/B use a base top-1 probability above 0.5, with agreement determining A versus B. Among the remaining positions, C requires the top two probabilities to sum to more than 0.7 and the second to exceed 0.15; all others are assigned to D.



Figure 9: **Lexical patterns after GRPO.** Columns: Math and Medical; rows: the three rank-shift categories.

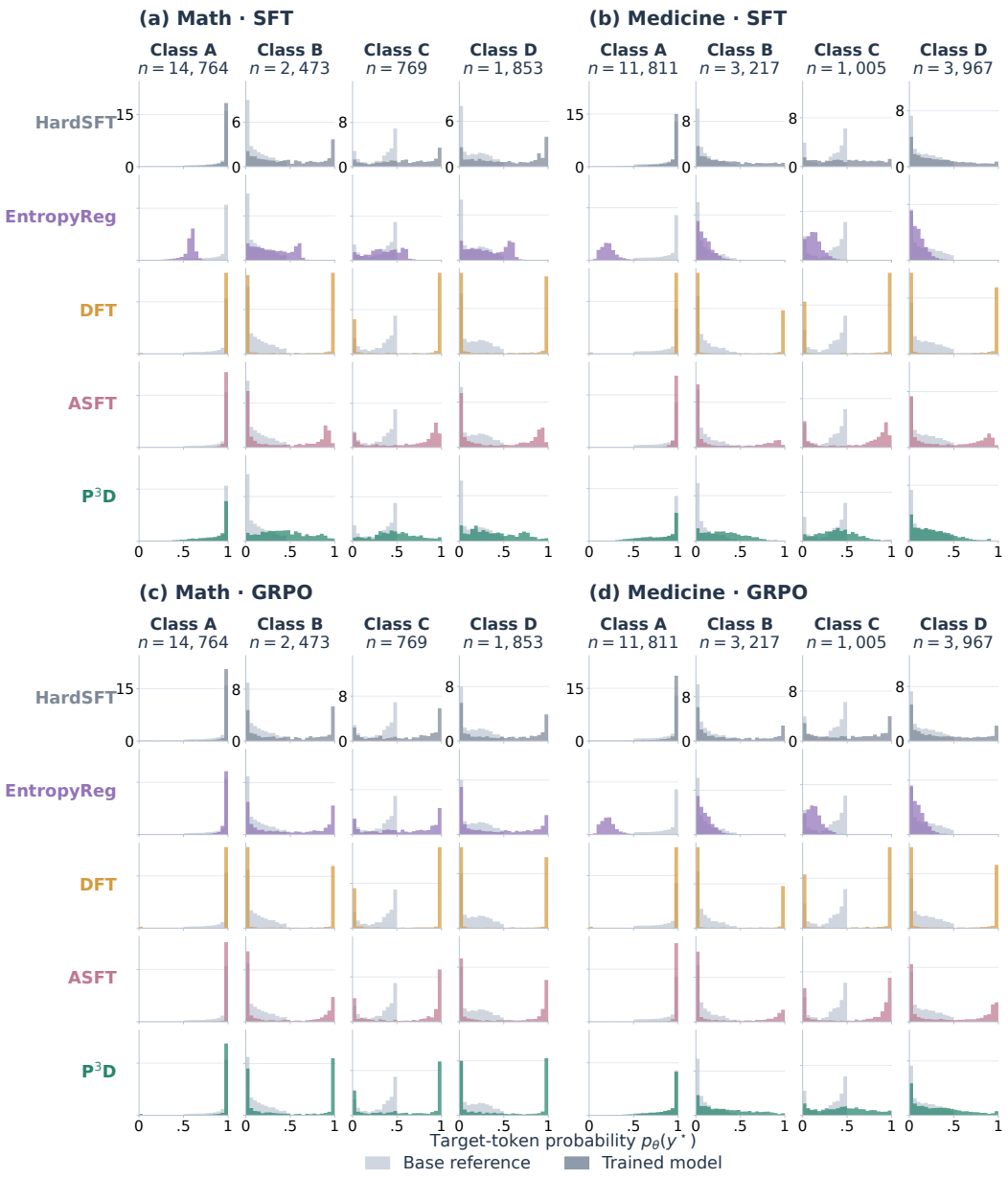


Figure 10: **Demonstrated-token probabilities by fixed Base class.** Top: SFT; bottom: GRPO. Grey: Base; color: trained model. Panels use matched positions and within-class density normalization.

Figure 10 compares HardSFT, EntropyReg, DFT, ASFT, and P^{3}D. Each panel overlays the trained distribution of $p_\theta(y^\star)$ with the base reference on the same positions. All panels use 24 equal-width bins on $[0, 1]$; densities integrate to one within a class, so histogram height does not indicate class size. The vertical scale is shared across methods within each class.

SFT: retaining confidence while learning mismatches. P^{3}D exhibits the selective acquisition intended by *Greedy Guarantee*: training strengthens the demonstrated choice where the base disagrees while retaining strong target confidence where it already agrees. In Class B, mean target probability increases from 0.129 to 0.448 on Math and from 0.106 to 0.327 on Medical, showing that the soft target supplies effective correction at initially mismatched positions. In Class A, the post-SFT medians remain high at 0.951 and 0.860, respectively, whereas EntropyReg produces much lower target probabilities in these runs. Together, these observations support the acquisition goal of our target design: learning the demonstrated preference without broadly weakening already-aligned predictions. The probability increases provide evidence of acquisition; token-wise dominance is a property of the constructed target and is not established by these histograms alone.

SFT: less concentration at the extremes. P^{3}D also exhibits the restrained concentration intended by *Minimal Intervention*: acquisition coexists with substantial probability mass remaining on non-target continuations. This is particularly visible in Class C, where the SFT median target probability is 0.476 on Math and 0.393 on Medical, compared with 0.999 and 0.992 for DFT; Class D likewise retains a broad range of intermediate probabilities. Class B shows that this retained mass accompanies meaningful correction: on Math, only 5.1% of positions remain at target probability at most 0.05, down from 40.2% under Base and compared with 44.5% under DFT, yet only 1.7% reach at least 0.95. The corresponding P^{3}D fractions on Medical are 10.6% and 0.1%. Thus, learning moves many mismatches away from near-zero target probability without pushing them to near-one-hot predictions, consistent with the target design's balance between correction and retention. These histograms establish the remaining non-target mass; the identities and relative preferences of those alternatives are examined through the complementary probes and cases in this appendix.

GRPO: further concentration depends on domain. The acquisition–retention balance established during SFT leaves an initialization that supports further reward-based adaptation: GRPO can increase concentration selectively while maintaining learned preferences at initially mismatched positions. On Math, the Class-C median rises from 0.476 to 0.755, close to HardSFT's 0.731, showing that the softer SFT distribution permits substantial subsequent concentration. On Medical, the Class-C median increases more moderately, from 0.393 to 0.452, while the Class-A median rises from 0.860 to 0.966. In Class B, the post-GRPO median remains above DFT's in both domains (0.438 vs. 0.009 on Math and 0.290 vs. below 0.001 on Medical), indicating that adaptation retains substantial target probability at positions the base initially disfavored. Together with the post-GRPO task gains in Table 1 and the reward dynamics in Appendix H.2, these patterns support the intended role of our two principles: acquiring demonstrated behavior while retaining alternatives during SFT provides a foundation for further learning, with the degree of subsequent concentration varying across domains.

G.3 Local examples of acquisition and retention

Figure 11 compares three selected positions in teacher-forced demonstrations (trained in SFT stage), rather than the generated trajectories above. We keep the same response prefix and demonstrated token across all available methods. Subsequent demonstration text is greyed out and does not enter the conditioning prefix. The same candidate token IDs are displayed for every method; candidates absent from saved top-K predictions are bounded by the smallest recorded probability rather than assigned zero.

A: retaining an existing preference. In the Math example, "the segment AC has a *point* B," the base assigns probabilities 0.561, 0.340, and 0.097 to *point*, *total*, and *length*. P^{3}D retains nearly the same probabilities, whereas HardSFT concentrates 0.842 on *point*. This position illustrates retention of the base's leading alternatives where it already agrees with the demonstration.

B: learning demonstrated content. In the pyrolysis example from the Medical corpus, the demonstrated continuation is "in the *absence* of oxygen." The base assigns *absence* probability

0.097, below *presence* (0.299) and *gas* (0.338). P^{3}D raises *absence* to 0.718, making it the preferred token. Retention therefore does not require leaving a mismatch with the demonstration uncorrected.

C: preserving plausible alternatives. For "The mesentery is the tissue that *attaches*," P^{3}D increases *attaches* from 0.225 to 0.558 while retaining 0.339 on *connects*. HardSFT instead assigns 0.894 and 0.094 to these tokens. This illustrates that learning the demonstrated choice can coexist with substantial probability on another plausible continuation. The example supports retention of alternatives, not exact realization of all target ratios by the trained model.

G.4 General-capability retention

Tables 10 pair SFT and post-GRPO zero-shot scores on ARC, HellaSwag, MMLU, MMLU-Pro, GPQA-D, and TruthfulQA-mc1/mc2. These are accuracy metrics obtained by likelihood-based answer scoring, rather than raw log-likelihood values (Appendix E). Reporting individual benchmarks makes retention and regression visible even when they offset one another in an aggregate. We additionally report the direct Base+GRPO control for Qwen3-4B Math, the setting for which this control is available. Changes from SFT to GRPO are assessed by comparing the two stage entries of each method.

Retention is benchmark-dependent. The benchmark-specific gains and regressions suggest that domain adaptation transfers unevenly: learning from reasoning demonstrations can benefit related reasoning tasks, while preservation of existing preferences does not ensure improvement on every form of knowledge or answer selection. For example, after Math SFT, Qwen3-8B P^{3}D improves MMLU-Pro from 46.4% under Base to 48.0% and GPQA-D from 36.9% to 40.9%, while HellaSwag rises from 78.7% to 79.6%. Yet ARC decreases from 56.9% to 55.1%, and MMLU changes from 74.8% to 74.3%. A plausible explanation is that demonstrated reasoning behavior transfers to some demanding question-answering tasks, whereas broad factual coverage and the relative likelihoods assigned to answer options receive less direct supervision. Even reasoning benchmarks need not benefit equally, as the ARC result illustrates. TruthfulQA introduces a further distinction: it tests resistance to common misconceptions, which mathematical or medical answer-correctness training does not explicitly target. Consistent with this distinction, Qwen3-4B Math SFT reduces TruthfulQA-mc1 from 34.8% to 33.8% despite its task gains. These patterns are consistent with task-dependent transfer and changes in answer preference; they motivate evaluating acquisition and retention jointly rather than expecting every general benchmark to improve.

GRPO changes the balance across capabilities. Comparing the same method before and after GRPO shows that further domain learning can both reduce and recover general-benchmark performance, depending on the capability being measured. For Qwen3-4B Math P^{3}D, ARC falls from 62.3% after SFT to 53.8% after GRPO, MMLU from 72.8% to 71.2%, and MMLU-Pro from 46.9% to 43.9%. In contrast, TruthfulQA-mc1 rises from 33.8% to 35.0% and mc2 from 51.8% to 54.7%, while HellaSwag remains similar (73.8% to 74.2%). The Medical run also improves GPQA-D (36.4% to 38.4%) and TruthfulQA-mc1 (36.8% to 38.2%). One plausible interpretation is that optimizing domain-specific rewards redirects answer preferences: this can weaken transfer to some tasks while improving likelihood-based answer selection on others, even without an explicit objective for those benchmarks. Thus, the in-domain gains after GRPO in Table 1 accompany a redistribution of general-benchmark performance rather than uniform improvement or forgetting. The available Qwen3-4B Math direct-RL control provides additional context: P^{3}D+GRPO exceeds Base+GRPO on GPQA-D (38.9% vs. 34.3%), but trails it on TruthfulQA-mc1 (35.0% vs. 37.7%). For the other models, the paired SFT/GRPO results describe stage-to-stage changes without attributing them to an advantage over direct RL.

H Learning Dynamics and Subsequent Adaptation

Our target design seeks to make demonstration learning efficient by combining acquisition with limited intervention in the student's existing distribution. This section connects three aspects of that design: smaller SFT gradient signals accompanied by limited policy drift and stronger task performance; efficient acquisition from demonstrations; and subsequent GRPO adaptation measured

Selective learning, retained alternatives

A Retain an existing preference

MATH

... First, the segment AC has a **point** B ...

	point	total	length	midpoint	Other
Base	**0.56**	0.34	0.10	<0.01	<0.01
HardSFT	**0.84**	0.15	0.01	<0.01	<0.01
LS	**0.80**	0.10	0.02	<0.01	0.08
EntropyReg	**0.48**	0.18	0.03	≤0.002	≈0.313
DFT	**1.00**	<0.01	<0.01	≤2.7e-14	≈0
ASFT	**0.98**	0.02	<0.01	<0.01	<0.01
EAFT	**0.72**	0.26	0.02	<0.01	<0.01
RFT	**0.75**	0.21	0.04	<0.01	<0.01
Prefix-RFT	**0.88**	0.11	0.01	0.00	<0.01
P^{3}D	**0.56**	**0.34**	**0.10**	**<0.01**	**<0.01**

B Learn the demonstrated content

MEDICAL

Plasma pyrolysis: ... break down organic materials in the **absence** of oxygen.

	absence	presence	gas	plasma	Other
Base	**0.10**	0.30	0.34	0.16	0.11
HardSFT	**0.95**	0.04	<0.01	0.01	<0.01
LS	**0.71**	0.12	<0.01	<0.01	0.17
EntropyReg	**0.33**	0.03	0.01	0.01	0.62
DFT	**1.00**	<0.01	<0.01	<0.01	<0.01
ASFT	**0.84**	0.06	0.05	0.03	0.01
EAFT	**0.86**	0.08	0.01	0.04	0.01
RFT	**0.64**	0.06	0.23	0.06	0.01
Prefix-RFT	**0.92**	0.08	0.00	0.00	<0.01
P^{3}D	**0.72**	**0.18**	**0.01**	**0.06**	**0.03**

C Keep plausible alternatives

MEDICAL

... The mesentery is the tissue that **attaches** ...

	attaches	connects	holds	supports	Other
Base	**0.23**	0.61	0.11	0.02	0.04
HardSFT	**0.89**	0.09	0.01	<0.01	<0.01
LS	**0.71**	0.14	0.01	0.01	0.13
EntropyReg	**0.27**	0.11	0.02	0.02	0.58
DFT	**0.99**	0.01	<0.01	<0.01	<0.01
ASFT	**0.88**	0.11	0.01	<0.01	<0.01
EAFT	**0.80**	0.18	0.02	<0.01	<0.01
RFT	**0.99**	0.01	<0.01	<0.01	<0.01
Prefix-RFT	**0.87**	0.10	0.00	0.03	<0.01
P^{3}D	**0.56**	**0.34**	**0.07**	**0.01**	**0.02**

Figure 11: **Selective learning at matched token positions.** A: retaining the base's existing preference; B: promoting demonstrated content; C: retaining plausible alternatives while learning the target. Rows compare measured Qwen3-4B probes with P^{3}D. Black text is the conditioning prefix, blue is the measured token, and grey is subsequent demonstration text, which is not part of the conditioning context. Blue, orange, and grey cells denote the demonstrated token, selected alternatives, and remaining vocabulary (Other), respectively. $\leq$ marks an upper bound for a candidate outside the saved top-K; $\approx$ marks a bounded residual estimate. These are selected qualitative examples.

Table 10: **General-capability retention.** Zero-shot accuracy (%). Bold: best adaptation method per model/domain/stage; Base and Base+GRPO are references. MMLU-P: MMLU-Pro; TQA1/2: TruthfulQA-mc1/mc2. n/a: inapplicable; —: unavailable.

Method	SFT							GRPO / RFT						
	ARC	HSwag	MMLU	MMLU-P	GPQA-D	TQA1	TQA2	ARC	HSwag	MMLU	MMLU-P	GPQA-D	TQA1	TQA2
Qwen3-4B														
Base	63.0	73.4	72.4	45.1	37.9	34.8	51.6	63.0	73.4	72.4	45.1	37.9	34.8	51.6
Math adaptation														
Base+GRPO				*n/a*				52.3	73.7	71.6	43.2	34.3	37.7	54.8
HardSFT	62.0	73.3	72.6	46.9	36.9	33.4	50.4	53.9	74.0	71.5	44.9	**39.4**	35.7	53.5
LS	**62.3**	72.2	72.3	45.1	**39.9**	33.3	51.2	**54.2**	72.9	71.1	43.1	38.4	**36.5**	54.5
EntropyReg	61.3	72.8	72.5	46.6	39.4	32.2	51.8	53.8	74.1	71.1	45.1	37.4	35.4	54.2
GEM	48.6	53.8	71.1	43.8	39.4	**36.4**	**53.7**	51.8	54.6	71.4	44.6	37.4	**36.5**	53.9
DFT	49.2	54.2	71.0	42.8	37.9	32.8	50.3	52.9	71.9	71.2	43.7	38.4	31.9	49.2
ASFT	53.4	72.9	71.1	43.5	37.9	35.4	52.6	53.2	72.8	71.2	45.0	38.9	35.0	53.1
EAFT	53.2	73.7	71.3	44.1	38.4	36.1	53.1	53.8	73.9	71.5	**45.2**	**39.4**	**36.5**	53.8
CurioSFT	47.7	56.6	72.3	46.5	34.9	25.7	50.0	33.2	42.5	69.4	39.4	35.9	23.5	48.8
P^{3}D	**62.3**	**73.8**	**72.8**	**46.9**	39.9	33.8	51.8	53.8	**74.2**	71.2	43.9	38.9	35.0	**54.7**
RFT				*n/a*				50.3	54.8	**71.6**	43.3	34.3	36.1	53.2
Prefix-RFT				*n/a*				51.5	73.7	71.3	44.1	38.9	36.0	53.5
Medical adaptation														
HardSFT	**63.9**	74.4	**72.5**	47.1	35.4	34.8	52.9	51.3	74.3	70.3	43.7	**38.4**	37.0	55.2
LS	63.6	73.6	72.4	46.4	**36.4**	35.7	54.4	51.7	74.2	70.3	44.0	37.4	37.6	56.0
EntropyReg	62.4	74.4	72.5	**47.3**	34.3	36.8	**56.0**	51.6	**75.3**	70.2	43.9	35.9	36.6	**56.9**
GEM	47.2	53.6	70.4	43.3	35.4	36.0	54.8	48.4	54.7	**71.2**	**46.0**	34.9	37.3	54.6
DFT	48.5	53.8	70.5	42.2	35.4	30.2	49.5	47.7	53.0	70.5	41.2	34.9	30.0	48.4
ASFT	53.8	73.1	71.1	43.1	35.9	35.6	53.9	**52.3**	73.2	71.0	43.5	37.4	36.2	54.7
EAFT	52.2	74.7	70.5	44.0	35.9	**37.0**	55.0	52.0	74.2	70.4	44.2	36.9	36.2	54.4
CurioSFT	36.0	49.0	68.4	39.1	35.9	23.6	48.6	37.5	48.6	67.4	40.7	35.1	25.3	49.9
P^{3}D	52.6	**75.3**	70.5	43.5	**36.4**	36.8	55.5	51.9	74.9	70.1	44.0	**38.4**	**38.2**	56.5
RFT				*n/a*				48.5	56.4	70.5	44.1	37.9	37.8	55.3
Prefix-RFT				*n/a*				51.4	74.3	70.3	43.6	36.9	37.0	54.7
Qwen3-8B														
Base	56.9	78.7	74.8	46.4	36.9	35.6	52.4	56.9	78.7	74.8	46.4	36.9	35.6	52.4
Math adaptation														
HardSFT	52.9	79.0	74.5	47.5	39.9	**35.9**	52.9	50.8	79.0	74.6	47.9	38.4	**36.7**	**54.0**
LS	51.1	78.2	74.2	47.7	37.9	35.4	53.2	50.6	78.8	74.5	**48.8**	34.3	35.5	51.6
EntropyReg	53.7	78.9	74.2	48.3	34.3	35.1	**54.8**	**54.9**	77.1	73.2	48.1	35.9	35.3	53.3
GEM	50.2	57.7	74.5	47.4	38.4	35.7	53.8	48.5	58.2	74.3	47.5	36.9	35.5	53.6
DFT	53.5	76.9	**74.8**	**48.5**	**41.9**	35.1	51.7	51.8	76.8	74.8	48.0	39.4	34.3	51.9
ASFT	**55.1**	78.1	74.4	48.1	41.4	35.5	51.9	51.2	78.2	74.8	48.4	37.9	36.2	53.1
EAFT	54.2	78.8	74.6	47.8	36.4	35.5	53.1	52.4	78.8	74.6	48.3	39.9	36.2	53.4
CurioSFT	34.0	71.7	73.8	44.4	35.4	23.9	48.4	38.4	72.7	72.2	46.4	35.8	29.3	50.6
P^{3}D	**55.1**	**79.6**	74.3	48.0	40.9	35.6	53.1	51.2	**79.1**	74.8	48.4	**40.4**	36.2	53.8
RFT				*n/a*				50.9	58.8	**75.0**	47.8	37.9	34.8	52.7
Prefix-RFT				*n/a*				51.2	78.9	74.3	47.4	39.9	36.1	53.2
Medical adaptation														
HardSFT	48.0	79.4	74.0	48.8	34.3	36.8	55.1	45.1	79.2	73.7	48.7	34.9	37.3	**56.0**
LS	51.5	79.0	73.9	48.1	33.8	36.6	55.3	**52.3**	79.3	74.0	49.3	34.9	36.0	55.2
EntropyReg	52.1	79.5	73.8	48.8	35.9	36.2	**56.7**	52.1	79.5	73.8	48.8	35.9	36.2	55.8
GEM	47.3	57.9	74.0	48.9	35.9	36.5	55.5	46.8	58.7	74.3	**49.9**	**37.9**	36.2	54.4
DFT	47.1	76.1	**74.1**	47.1	34.9	35.5	52.6	43.0	75.6	74.0	47.1	33.8	35.4	52.9
ASFT	**54.6**	77.9	74.1	47.3	**37.4**	33.8	50.9	51.1	77.7	74.1	47.3	36.9	33.9	51.6
EAFT	47.3	79.3	74.1	48.2	33.8	37.1	55.2	44.8	79.2	73.8	48.4	33.8	**37.8**	55.5
CurioSFT	32.1	50.1	72.2	40.2	33.3	23.5	48.4	31.3	51.3	73.3	40.0	33.6	23.2	49.9
P^{3}D	53.4	**79.6**	73.8	**49.0**	34.8	**37.4**	55.6	50.9	**79.6**	73.6	48.6	35.9	37.5	55.6
RFT				*n/a*				47.6	60.5	**74.4**	48.8	34.3	37.0	54.8
Prefix-RFT				*n/a*				46.6	79.3	74.0	48.9	34.3	36.6	54.8
Llama3-8B														
Base	53.4	79.1	62.1	30.9	31.8	27.2	44.0	53.4	79.1	62.1	30.9	31.8	27.2	44.0
Math adaptation														
HardSFT	53.0	80.0	61.0	30.2	31.3	29.6	46.3	53.2	79.2	60.6	29.4	28.3	27.8	44.1
LS	52.8	79.0	59.9	29.1	**31.8**	29.9	46.7	53.7	78.5	60.8	29.4	31.8	29.5	46.9
EntropyReg	53.3	79.1	**61.2**	30.2	31.3	27.2	**49.6**	51.1	78.8	61.0	28.7	32.3	27.7	48.6
GEM	51.1	59.9	60.7	30.1	**31.8**	30.6	47.2	53.1	60.5	59.8	29.5	**34.3**	31.2	47.6
DFT	48.5	63.2	60.8	29.3	29.8	27.7	45.6	50.3	68.9	60.0	29.1	32.8	28.1	45.5
ASFT	51.8	78.3	61.1	29.6	31.3	28.4	42.8	53.4	77.6	60.8	31.0	30.3	28.3	43.5
EAFT	**54.4**	80.2	61.0	**31.6**	29.3	**31.0**	47.4	**55.0**	79.5	60.6	31.7	31.3	30.6	48.0
CurioSFT	31.1	43.9	60.6	27.7	25.8	22.6	47.3	31.8	45.7	60.6	28.3	25.8	21.7	48.1
P^{3}D	54.1	**80.8**	60.8	29.8	30.3	30.7	47.5	54.7	**80.4**	60.6	29.8	30.8	30.7	48.0
RFT				*n/a*				51.7	60.6	**62.8**	**31.9**	28.8	**33.7**	**50.6**
Prefix-RFT				*n/a*				53.1	79.6	60.9	30.4	28.3	28.9	45.6
Medical adaptation														
HardSFT	53.0	80.8	61.2	34.0	**30.3**	34.5	53.6	47.7	79.7	61.2	33.5	29.3	33.9	54.4
LS	51.9	79.7	60.9	32.1	29.8	32.6	52.2	50.1	78.9	61.5	32.8	26.3	32.9	51.8
EntropyReg	52.8	80.2	61.1	33.5	27.3	30.0	54.1	**52.8**	80.3	60.8	32.7	25.8	30.2	54.1
GEM	50.9	61.3	60.7	33.3	28.3	33.5	52.4	49.6	61.2	61.0	32.8	32.8	33.8	53.0
DFT	46.2	64.4	58.5	29.2	29.3	29.5	46.2	44.9	63.2	58.7	28.1	26.3	30.0	44.5
ASFT	51.7	77.6	61.2	32.2	28.8	28.1	44.7	50.6	76.9	61.3	32.0	**35.4**	26.3	44.6
EAFT	53.3	81.1	**61.4**	33.6	29.3	**35.0**	**54.5**	48.9	79.1	**61.9**	33.6	28.3	**37.7**	**55.6**
CurioSFT	33.0	43.4	60.7	32.0	24.2	23.1	48.1	33.8	45.3	61.3	32.9	23.4	23.8	48.0
P^{3}D	**54.4**	**82.0**	61.3	**34.5**	29.2	34.6	52.3	**52.8**	80.5	61.8	33.5	29.3	34.3	54.5
RFT				*n/a*				49.9	61.4	59.7	30.3	32.8	31.9	50.1
Prefix-RFT				*n/a*				51.3	**80.5**	61.3	**34.3**	28.8	34.3	53.8

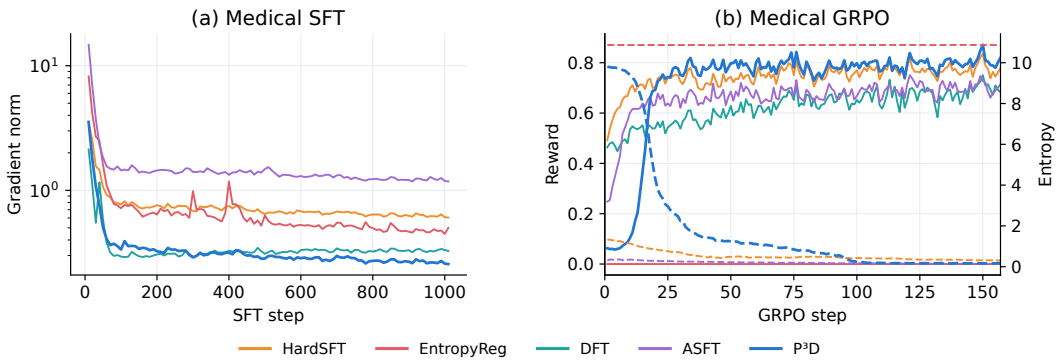


Figure 12: **Medical training dynamics.** (a) SFT gradient norms (log scale). (b) GRPO reward (solid, left axis) and entropy (dashed, right axis), at rollout temperature $T = 1.0$.

through reward and entropy. The main-text figures show Math, while the Medical traces below extend the analysis across domains. Training curves use Qwen3-4B.

H.1 Learning during SFT

Strong acquisition with limited policy drift. P^{3}D combines smaller SFT gradient signals with limited policy drift and stronger task performance, supporting the intended effect of *Minimal Intervention*. In Figure 5(a), its gradient norms settle below HardSFT's while its greedy accuracy continues to improve. The Medical gradients in Figure 12(a) show the same pattern of smaller late-stage gradient magnitudes relative to HardSFT. These optimization traces complement the measured distributional change in Figure 4(a): on Math, P^{3}D attains 54.0% greedy accuracy versus HardSFT's 50.5%, with a smaller RMS Hellinger distance from Base (0.264 vs. 0.287 on the shared, coarsened distributions). Together, the gradient and distributional views support learning with restrained intervention, while the accuracy improvement shows that this restraint preserves effective acquisition. ASFT's smaller drift but lower accuracy further highlights the importance of achieving both goals together.

Efficient learning from demonstration exposure. The training-progress view shows that this restrained adaptation also learns effectively with fewer SFT steps. In Figure 5(a), P^{3}D reaches the best displayed baseline accuracy before completing the training schedule and continues to improve thereafter. With a common batch size, reaching a given accuracy in fewer steps also requires fewer training-example presentations, supporting more efficient use of demonstration exposure. This behavior is consistent with the target construction: it directs correction toward the demonstrated preference and stops the local supervisory gradient once the chosen probability floor is met. The resulting training progress supports efficient acquisition alongside the limited intervention observed above.

H.2 Subsequent GRPO adaptation

Rapid reward adaptation with retained uncertainty. The P^{3}D initialization rapidly improves rollout reward while retaining higher policy entropy than HardSFT during early adaptation. On Medical, mean reward rises from 0.023 over steps 1–10 to 0.713 over steps 20–40, approaching HardSFT's 0.734 in the latter window. At the same time, P^{3}D retains entropy of 2.33, compared with HardSFT's 0.66. Math exhibits a similar combination: reward increases from 0.007 over steps 1–10 to 0.284 over steps 40–60, when entropy remains 1.30 versus HardSFT's 0.12. These paired observations show that the retained uncertainty is compatible with rapid acquisition of rewarded behavior. They support the role of the SFT target design in providing an initialization from which GRPO can learn while alternative continuations remain available.

Continued adaptation yields stronger late-stage reward. On Math, P^{3}D's early reward improvement develops into stronger late-stage performance: over steps 250–300, its mean reward reaches 0.397, compared with 0.378 for HardSFT and 0.370 for EntropyReg. On Medical, it main-

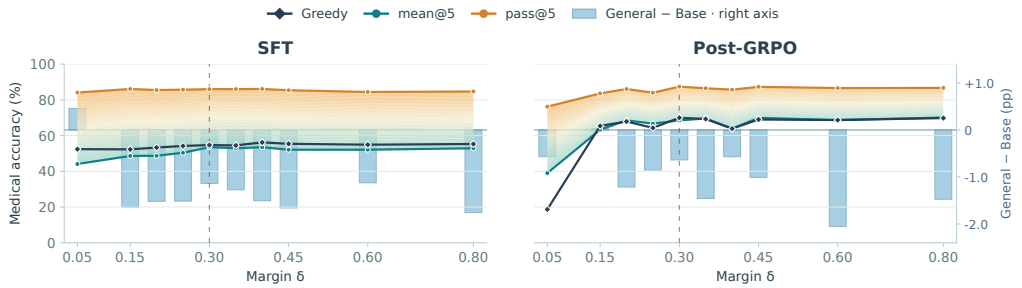


Figure 13: **Effect of intervention strength on Medical.** Left: SFT; right: post-GRPO. Lines show greedy accuracy, mean@5, and pass@5 from Table 12; shading spans mean@5 to pass@5. Bars show changes in mean accuracy on seven general benchmarks (right axis, pp), relative to the pretrained Base at both stages; zero denotes no change. The dashed vertical line marks our representative $\delta = 0.30$.

tains a high-reward regime after the initial rise, averaging 0.751 over steps 130–157. This progression links retention during SFT with the capacity for further reward-based learning. EntropyReg provides a useful contrast: it maintains high Medical entropy while reward stays near zero. The benefit of P^{3}D therefore lies in combining retained alternatives with effective adaptation to the reward signal.

I Full δ-Margin Sweep

The margin δ controls how strongly the supervisory target favors the demonstrated token. We examine the trade-off between sufficient supervision and retention of alternative continuations, and whether useful performance extends across a range of intervention strengths. The sweep therefore evaluates the operating range of the method; $\delta = 0.30$ is our representative configuration within that range.

I.1 Sweep setup and metrics

All sweeps use Qwen3-4B with $\delta \in \{0.05, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.60, 0.80\}$. Greedy accuracy uses $T = 0$; mean@5 and pass@5 use $T = 0.6$. Mean@5 measures average sampled correctness and pass@5 measures five-sample success coverage. The shaded bands in the overview figures span mean@5 to pass@5. General-capability bars report changes in the mean accuracy over seven likelihood-scored benchmarks. The Math overview uses Base for SFT and Base+GRPO for post-GRPO as stage-matched references. The Medical overview uses the pretrained Base at both stages to measure retention relative to the same starting model.

I.2 Task performance across margins

The Math overview in Figure 6 and the Medical overview in Figure 13 place the three task metrics alongside intervention strength. A larger margin increases the target's emphasis on the demonstration, while its effects on greedy accuracy, sampling reliability, and five-sample coverage can differ. Tables 11 and 12 retain the complete benchmark-level results, with $\delta = 0.30$ highlighted as the shared default.

I.3 A useful range of intervention strengths

The two extremes impose different limitations. Small margins provide little additional promotion beyond the dominance boundary. More positions can meet the chosen floor early and cease to receive a local supervisory gradient, limiting further acquisition under a fixed training budget. As the margin increases, the demonstrated-token mass approaches one and the constructed target approaches one-hot supervision, leaving less probability mass for alternatives. The margin therefore

Table 11: **Math margin sweep.** Per-benchmark and average greedy accuracy, pass@5, and mean@5 at SFT and post-GRPO stages. Shading marks $\delta = 0.30$.

δ	MATH500			GSM8K			AIME'24			AIME'25			Minerva			**Avg**		
	T=0	@5	m@5	T=0	@5	m@5	T=0	@5	m@5	T=0	@5	m@5	T=0	@5	m@5	T=0	@5	m@5
(A) SFT stage																		
0.05	69.0	57.6	24.8	83.9	80.8	40.1	10.0	20.0	4.7	23.3	36.7	8.7	50.1	43.0	16.2	47.3	47.6	18.9
0.15	71.4	69.2	32.0	87.3	92.0	45.9	20.0	23.3	5.3	26.7	33.3	8.7	55.2	62.1	25.1	52.1	56.0	23.4
0.20	70.2	78.0	43.6	87.6	93.6	54.8	13.3	30.0	8.0	23.3	36.7	11.3	54.9	68.6	31.3	49.9	61.4	29.8
0.25	71.2	82.6	51.7	85.0	93.9	57.1	**23.3**	36.7	8.0	**33.3**	33.3	12.0	55.6	73.8	39.8	53.7	64.1	33.7
0.30	73.8	88.2	61.9	**90.5**	97.2	69.6	20.0	46.7	12.0	26.7	43.3	17.3	**59.0**	74.3	48.1	**54.0**	69.9	41.8
0.35	71.6	88.2	68.1	84.1	95.8	73.1	20.0	43.3	14.0	26.7	40.0	16.7	55.7	**77.9**	55.2	51.6	69.1	45.4
0.40	70.2	88.2	72.8	88.3	97.0	79.3	20.0	53.3	20.0	**33.3**	**56.7**	22.7	55.4	77.3	59.8	53.5	**74.5**	50.9
0.45	72.6	**88.8**	**75.7**	85.9	96.8	81.3	**23.3**	**56.7**	**24.0**	30.0	50.0	22.7	56.2	76.9	**61.7**	53.6	73.8	53.1
0.60	72.8	85.8	74.4	89.8	**97.3**	**91.7**	20.0	53.3	20.7	23.3	50.0	**26.7**	56.8	72.6	59.0	52.5	71.8	**54.5**
0.80	**74.0**	86.4	75.0	88.2	96.9	90.2	13.3	50.0	19.3	23.3	50.0	25.3	56.6	72.9	58.9	51.1	71.2	53.7
(B) RL stage (post-GRPO)																		
0.05	77.4	89.4	80.1	92.4	96.4	92.2	13.3	43.3	18.0	16.7	36.7	16.7	74.1	83.9	74.0	54.8	69.9	56.2
0.15	77.8	89.4	77.0	90.8	96.7	91.4	26.7	46.7	14.0	36.7	**73.3**	28.7	68.6	83.4	68.5	60.1	77.9	55.9
0.20	67.6	84.0	67.4	91.8	97.2	92.9	16.7	53.3	15.3	30.0	56.7	30.7	67.4	77.1	68.1	54.7	73.7	54.9
0.25	71.0	85.2	69.7	88.9	96.0	88.6	3.3	40.0	12.7	13.3	40.0	14.7	65.2	82.1	63.4	48.3	68.7	49.8
0.30	82.0	92.2	82.7	**94.2**	**97.7**	93.3	30.0	60.0	30.7	30.0	50.0	**33.3**	**77.5**	**87.3**	**77.9**	**62.7**	77.4	**63.6**
0.35	81.0	91.2	82.0	92.0	96.7	93.0	23.3	60.0	27.3	20.0	46.7	26.7	67.7	76.5	69.0	56.8	74.2	59.6
0.40	**83.2**	**92.8**	**83.6**	93.5	97.2	93.2	**33.3**	63.3	32.0	26.7	50.0	28.7	69.1	76.0	69.8	61.1	75.9	61.4
0.45	**83.2**	91.6	83.0	93.6	97.3	93.4	20.0	63.3	30.7	**40.0**	56.7	27.3	70.2	77.1	70.3	61.4	77.2	60.9
0.60	80.6	92.0	81.9	93.6	96.9	**93.6**	23.3	**76.7**	**34.0**	33.3	56.7	30.0	68.3	76.0	68.8	59.8	**79.7**	61.7
0.80	77.0	85.2	75.8	86.6	95.4	90.3	6.7	33.3	10.7	23.3	33.3	15.3	74.5	82.9	73.3	53.6	66.0	53.1

Table 12: **Medical margin sweep.** Same metrics and stage organization as Table 11.

δ	MedQA			MedMCQA			PubMedQA			**Avg**		
	T=0	@5	m@5	T=0	@5	m@5	T=0	@5	m@5	T=0	@5	m@5
(A) SFT stage												
0.05	43.0	87.6	52.6	45.5	83.9	47.9	68.6	80.8	31.8	52.4	84.1	44.1
0.15	43.6	**89.2**	**54.5**	45.0	**84.5**	**49.5**	68.4	84.8	41.8	52.3	**86.1**	48.6
0.20	45.6	87.4	53.1	45.3	82.8	48.3	69.0	86.2	44.6	53.3	85.5	48.7
0.25	46.7	87.9	53.8	44.6	82.4	48.5	**71.2**	86.8	49.0	54.2	85.7	50.4
0.30	48.0	88.7	54.0	**47.9**	82.2	48.6	68.2	87.2	58.0	54.7	86.0	**53.5**
0.35	49.2	87.1	52.9	47.0	81.6	48.2	67.4	89.2	57.5	54.5	86.0	52.9
0.40	**54.4**	87.4	53.8	47.3	82.1	48.6	66.8	88.8	58.0	**56.2**	86.1	53.5
0.45	53.1	86.9	51.6	47.2	80.6	46.2	65.8	88.8	58.6	55.4	85.4	52.1
0.60	51.8	86.0	49.2	46.5	80.1	45.6	66.4	87.0	61.4	54.9	84.4	52.1
0.80	49.7	85.2	48.9	47.5	79.2	46.6	68.6	**89.8**	**63.1**	55.3	84.7	52.9
(B) RL stage (post-GRPO)												
0.05	0.9	80.1	38.2	54.8	79.9	54.7	0.4	68.4	24.1	18.7	76.2	39.0
0.15	68.1	86.9	70.8	59.0	79.9	60.4	69.0	84.0	58.8	65.4	83.6	63.3
0.20	70.5	87.7	72.8	58.9	79.7	59.9	**74.2**	**90.8**	**72.8**	67.9	86.1	68.5
0.25	69.8	87.0	71.5	56.2	77.7	58.6	67.0	87.2	70.0	64.3	84.0	66.7
0.30	74.5	90.3	74.5	62.9	**82.9**	62.9	72.4	89.0	68.0	**70.0**	**87.4**	68.5
0.35	76.0	90.3	76.1	62.8	81.2	63.8	69.2	88.0	68.8	69.3	86.5	69.6
0.40	**77.1**	90.4	76.2	62.9	82.0	63.2	51.6	84.6	52.4	63.9	85.7	63.9
0.45	75.9	90.7	76.6	62.3	81.3	62.1	68.8	90.0	70.9	69.0	87.3	69.9
0.60	75.6	90.8	75.7	**63.3**	80.9	**63.9**	67.4	88.0	66.4	68.7	86.6	68.7
0.80	75.7	**91.0**	**76.8**	62.2	80.5	62.3	71.6	88.4	71.1	69.8	86.7	**70.1**

balances the strength of demonstrated-token learning against the structure retained for subsequent adaptation.

Effective performance extends beyond a single margin. On Math, every tested margin from 0.30 through 0.60 improves both greedy accuracy and pass@5 over HardSFT at the SFT and post-GRPO stages. After GRPO, greedy accuracy spans 56.8–62.7% and pass@5 spans 74.2–79.7% within this range, compared with HardSFT's 54.9% and 64.9%. The gains therefore occur across several intervention strengths rather than at an isolated setting. The strongest tested margin, 0.80, reduces post-GRPO greedy accuracy to 53.6% and pass@5 to 66.0%, illustrating the cost of moving toward the one-hot end of the construction. We use $\delta = 0.30$ as a representative balance of acquisition, retained alternatives, and subsequent RL performance; the sweep supports a useful operating range without requiring a unique optimum for every metric.

J Additional Results on SFT and GRPO

We assess SFT and subsequent GRPO through three complementary measures: greedy accuracy captures the model's preferred output, mean@5 measures average sampled correctness, and pass@5 measures problem coverage within five samples. Reading these metrics together characterizes both immediate task performance after SFT and the gains realized through subsequent RL. The mean@5 results below complement the greedy and pass@5 scores in Table 1.

Strong greedy performance with useful sampling headroom. P^{3}D combines strong deterministic performance with successful exploration under sampling, reflecting the acquisition–retention balance intended for SFT. On Qwen3-4B Math, it achieves 54.0% greedy accuracy and 69.9% pass@5, compared with HardSFT's 50.5% and 63.8%, respectively (Table 1). Its domain-averaged mean@5 is lower than HardSFT's (41.8% vs. 53.7%; Table 13), but retains substantial sampled correctness. Read jointly, these metrics show that P^{3}D learns an effective greedy policy while retaining successful solutions that can be reached within five samples at $T = 0.6$. The lower mean@5 describes a trade-off in sampling reliability, while the higher pass@5 shows that this trade-off accompanies broader problem coverage under the same sampling budget. This combination supports useful sampling headroom after SFT and complements the distributional evidence for retaining alternatives.

GRPO converts the initialization's potential into reliable generation. P^{3}D achieves substantial gains in sampled correctness after GRPO, showing that its capacity for subsequent adaptation translates into held-out task performance. Table 13 reports Math mean@5 increases from 41.8% to 63.6% on Qwen3-4B, from 50.6% to 64.5% on Qwen3-8B, and from 36.6% to 45.9% on Llama3-8B. The corresponding HardSFT changes are 53.7% to 55.1%, 58.1% to 59.8%, and 41.5% to 35.7%. Thus, although P^{3}D starts with lower Math mean@5, it improves more during GRPO and finishes above HardSFT on all three models. These changes show how the sampling headroom retained during SFT accompanies more reliable correct generation after reward-based adaptation. Together with the greedy and pass@5 results in Table 1, these gains support evaluating SFT by both its immediate task learning and the foundation it provides for further RL. All mean@5 results use five samples at $T = 0.6$ under the evaluation protocol in Appendix E.

Table 13: **Mean@5 before and after GRPO (Math).** Each method pairs its SFT and post-GRPO scores at T=0.6. Avg is the arithmetic mean over all listed benchmarks; incomplete coverage is shown as a dash. The right block also includes the dedicated RFT methods RFT and Prefix-RFT. n/a marks their inapplicable standalone-SFT entries; dashes mark unavailable evaluations.

Method	SFT						Post-GRPO / RFT					
	MATH500	GSM8K	AIME'24	AIME'25	Minerva	Avg	MATH500	GSM8K	AIME'24	AIME'25	Minerva	Avg
Qwen3-4B												
HardSFT	73.4	90.7	18.0	27.3	59.3	53.7	78.4	92.1	15.3	16.0	73.8	55.1
LS	73.8	**91.8**	23.3	26.0	58.6	54.7	77.7	90.9	20.0	17.3	70.9	55.4
EntropyReg	72.8	91.0	19.3	28.0	57.4	53.7	77.2	90.4	21.3	17.3	72.5	55.8
GEM	64.2	74.0	16.0	18.0	**76.3**	49.7	71.7	91.7	18.0	10.7	77.7	54.0
DFT	72.8	63.1	7.3	9.3	61.9	42.9	74.0	91.8	14.0	10.7	68.8	51.9
ASFT	44.3	71.4	9.3	3.3	48.1	35.3	79.7	77.2	16.0	13.3	70.3	51.3
EAFT	**80.7**	91.6	**32.0**	**34.0**	60.2	**59.7**	76.6	92.4	8.0	14.7	74.2	53.2
CurioSFT	35.8	46.9	2.0	1.3	20.0	21.2	44.0	68.2	18.7	12.7	46.7	38.1
P^{3}D	61.9	69.6	12.0	17.3	48.1	41.8	**82.7**	**93.3**	**30.7**	**33.3**	**77.9**	**63.6**
RFT			*n/a*				64.9	80.1	24.0	24.7	56.8	50.1
Prefix-RFT			*n/a*				64.4	82.3	4.0	11.3	53.6	43.1
Qwen3-8B												
HardSFT	79.9	88.8	12.7	**39.3**	69.6	58.1	73.8	94.2	31.3	26.0	73.9	59.8
LS	79.0	83.0	**26.7**	37.3	68.9	59.0	82.0	90.0	34.0	18.7	67.9	58.5
EntropyReg	60.1	79.7	17.3	14.0	53.0	44.8	56.6	77.4	19.3	23.3	48.6	45.0
GEM	66.8	78.7	16.0	23.3	**76.0**	52.2	81.6	91.1	20.7	14.7	66.1	54.8
DFT	75.4	88.6	25.3	32.0	59.2	56.1	77.3	89.8	18.7	16.0	70.2	54.4
ASFT	70.5	66.3	8.7	12.7	55.8	42.8	78.9	91.7	19.3	17.3	**75.2**	56.5
EAFT	**82.6**	**90.9**	24.0	31.3	69.0	**59.6**	85.0	**94.7**	**36.7**	28.7	74.3	63.9
CurioSFT	48.4	66.4	14.7	8.0	47.6	37.0	59.1	68.2	14.7	16.0	53.7	42.4
P^{3}D	69.6	78.2	24.0	24.0	57.1	50.6	**85.9**	94.6	32.0	**37.3**	72.5	**64.5**
RFT			*n/a*				58.7	75.5	26.0	21.3	52.2	46.8
Prefix-RFT			*n/a*				70.0	84.3	8.7	7.3	67.4	47.5
Llama3-8B												
HardSFT	**55.9**	83.5	8.7	16.7	**42.5**	41.5	44.5	74.7	8.7	**16.7**	33.8	35.7
LS	54.8	**83.9**	**11.3**	16.0	41.8	**41.6**	44.1	79.5	5.3	12.0	32.6	34.7
EntropyReg	51.1	71.4	6.0	16.7	40.7	37.2	51.6	70.8	5.3	**16.7**	40.7	37.0
GEM	38.3	73.5	3.3	4.7	42.0	32.4	28.3	71.4	**10.0**	8.0	31.0	29.7
DFT	35.0	55.9	3.3	7.3	26.4	25.6	18.7	45.9	2.7	10.7	15.7	18.7
ASFT	30.8	14.6	2.0	10.0	25.0	16.5	25.5	50.9	2.7	5.3	20.0	20.9
EAFT	49.3	68.7	6.7	**18.0**	38.3	36.2	39.2	76.4	1.3	15.3	26.5	31.7
CurioSFT	22.7	34.5	1.3	2.7	13.1	14.9	47.0	62.6	6.0	11.3	28.3	31.1
P^{3}D	50.4	69.5	10.7	15.3	37.3	36.6	**65.2**	**88.2**	9.3	14.7	**52.1**	**45.9**
RFT			*n/a*				16.8	41.5	2.7	10.0	7.0	15.6
Prefix-RFT			*n/a*				61.2	86.1	8.0	**16.7**	50.1	44.4

Table 14: **Mean@5 before and after GRPO (Medical).** Each method pairs its SFT and post-GRPO scores at T=0.6. Avg is the arithmetic mean over all listed benchmarks; incomplete coverage is shown as a dash. The right block also includes the dedicated RFT methods RFT and Prefix-RFT. n/a marks their inapplicable standalone-SFT entries; dashes mark unavailable evaluations.

Method	SFT				Post-GRPO / RFT			
	MedQA	MedMCQA	PubMedQA	Avg	MedQA	MedMCQA	PubMedQA	Avg
Qwen3-4B								
HardSFT	54.5	49.5	63.3	55.8	73.7	61.5	68.4	67.9
LS	52.2	50.1	64.9	55.7	72.7	60.4	**70.5**	67.9
EntropyReg	49.0	47.1	58.8	51.6	49.5	47.7	61.8	53.0
GEM	50.4	47.5	57.6	51.9	74.5	61.7	68.5	68.2
DFT	50.8	41.2	**67.3**	53.1	41.0	51.8	67.4	53.4
ASFT	52.0	36.9	61.5	50.1	70.7	58.1	67.4	65.4
EAFT	**58.0**	**51.3**	64.7	**58.0**	**74.9**	61.4	67.6	68.0
CurioSFT	32.9	36.4	48.2	39.2	31.6	44.0	28.6	34.7
P^{3}D	54.0	48.6	58.0	53.5	74.5	**62.9**	68.0	**68.5**
RFT			*n/a*		65.3	55.5	65.9	62.3
Prefix-RFT			*n/a*		68.2	57.6	67.6	64.5
Qwen3-8B								
HardSFT	48.3	52.5	60.8	53.9	79.5	65.6	69.3	71.5
LS	52.1	56.6	63.0	57.3	80.8	67.0	71.9	73.2
EntropyReg	51.9	55.4	58.4	55.2	51.4	55.6	58.6	55.2
GEM	53.9	54.8	56.9	55.2	81.9	67.7	72.2	73.9
DFT	66.1	51.7	**68.0**	**61.9**	72.3	61.3	69.1	67.6
ASFT	**68.0**	**58.7**	57.1	61.3	78.4	65.3	64.4	69.4
EAFT	53.4	54.7	61.5	56.6	81.4	64.0	70.4	71.9
CurioSFT	44.0	40.1	48.1	44.1	47.8	50.1	55.5	51.1
P^{3}D	55.9	54.5	61.9	57.4	**82.4**	**68.4**	**74.1**	**75.0**
RFT			*n/a*		72.2	60.5	64.3	65.7
Prefix-RFT			*n/a*		77.9	64.6	68.9	70.4
Llama3-8B								
HardSFT	39.8	50.2	**63.2**	**51.1**	69.7	61.8	75.4	69.0
LS	35.5	47.4	62.6	48.5	75.1	61.5	**76.0**	70.9
EntropyReg	33.4	45.0	56.3	44.9	34.5	45.5	59.3	46.4
GEM	39.5	48.0	62.4	50.0	72.1	61.2	71.8	68.4
DFT	39.8	31.5	39.1	36.8	59.1	49.1	57.4	55.2
ASFT	19.3	28.4	30.0	25.9	68.1	57.0	58.4	61.2
EAFT	39.6	**50.9**	60.2	50.2	**76.2**	**63.3**	74.2	**71.2**
CurioSFT	16.1	30.1	21.8	22.7	24.8	40.1	59.2	41.4
P^{3}D	**42.1**	45.9	61.8	49.9	72.2	62.5	75.0	69.9
RFT			*n/a*		64.0	53.4	67.3	61.6
Prefix-RFT			*n/a*		68.8	58.7	72.4	66.7