

DIAG: Diagnostic Iterative Alignment and Generation for Data-Efficient Mathematical Preference Distillation

Guhan Chen^{1*}, Songtao Tian^{1*}, Bohan Li^{1,2}, Hejin Wang¹, Yexin Xie³, Zixiong Yu^{1,3,4†}

¹Tsinghua University ²Kyoto University ³Nanjing University

⁴Huawei Large Model Data Technology Lab

chen-gh23@mails.tsinghua.edu.cn xieye-xin@qq.com

{tiansongtao.2020, libh19, wanghj20, yuzx19}@tsinghua.org.cn

Abstract

Iterative preference optimization is essential for aligning Large Language Models on mathematical reasoning tasks, yet its efficiency is often throttled by signal scarcity: as the model improves, static problem sets become increasingly mismatched to the model’s evolving competence, producing rollouts that are either too easy or too hard and therefore non-informative, which leads to a scarcity of valid preference pairs. We propose DIAG, a *Diagnostic Iterative Alignment and Generation* framework that adaptively reshapes the practice distribution to increase informative supervision and focus training near the student’s current competence boundary. DIAG consists of two phases: (1) *diagnosing* valid preference-pair yield to calibrate the exploration–exploitation trade-off and allocate topic quotas via an Empirical Bayes shrinkage estimator, thereby prioritizing high-yield concepts; and (2) *generating* targeted practice, where a teacher synthesizes variants from the student’s failure traces. We further provide a theoretical view interpreting DIAG as a teacher-mediated approximation to KL-regularized reweighting of the practice distribution toward the student’s competence boundary, where valid preference-pair yield is maximized. Experiments show that DIAG boosts yield across iterations and delivers stronger reasoning performance under an iso-effective training budget, demonstrating that it can distill more informative preference supervision for mathematical reasoning.

1 Introduction

Artificial intelligence (AI) has advanced at an unprecedented pace in recent years, spanning natural language processing (Zhao et al., 2026b; Kang et al., 2026b), computer vision (Cui et al., 2021; Tian et al., 2026), multimodal understanding and generation (Yu et al., 2024; Wang et al., 2025;

Lu et al., 2026a,b; Dong et al., 2026), and intelligent agents (Wang et al., 2026a,b; Zhang et al., 2026). This progress has been driven by sustained advances in learning paradigms (Ouyang et al., 2022; Wu et al., 2026; Shi et al., 2026), model architectures (Vaswani et al., 2017; Yu et al., 2026b), data scaling (Yuan et al., 2023), and theoretical advances in understanding learning dynamics (Chen et al., 2024a; Yu et al., 2025) and generalization (Lai et al., 2023; Li et al., 2024b). Within this broader development, Large Language Models (LLMs) have emerged as a dominant paradigm in modern AI, achieving increasingly strong performance across a wide range of tasks. Their rapid advancement has spurred growing interest in improving their reasoning capabilities (Ye et al., 2025), instruction-following abilities (Zhou et al., 2023), and interactions with users (Qian et al., 2026) and environments (Zhu et al., 2026).

Against this backdrop, the LLM landscape has become increasingly heterogeneous, with models varying substantially in scale, architecture, and deployment settings, from massive frontier systems (GPT: OpenAI, 2024; Llama: Grattafiori et al., 2024) to specialized, lightweight variants (Phi-4-Mini: Microsoft, 2025; Gemini Nano: Gemini Team, 2025) optimized for edge deployment. This heterogeneity poses a challenge for model alignment: static training sets are increasingly unlikely to match the learning needs of models at different capability levels. Instances that are informative for one model may be too easy, too difficult, or even misleading for another, reducing the effectiveness of supervision and potentially leading to ineffective or even counterproductive learning.

Consequently, preference-based alignment has been increasingly adopted across diverse generative tasks (Rafailov et al., 2023; Lu et al., 2025), with recent reasoning methods further extending it to iterative optimization loops (Pang et al., 2024). By repeatedly generating candidate responses from

*Equal Contribution.

†Corresponding Author.

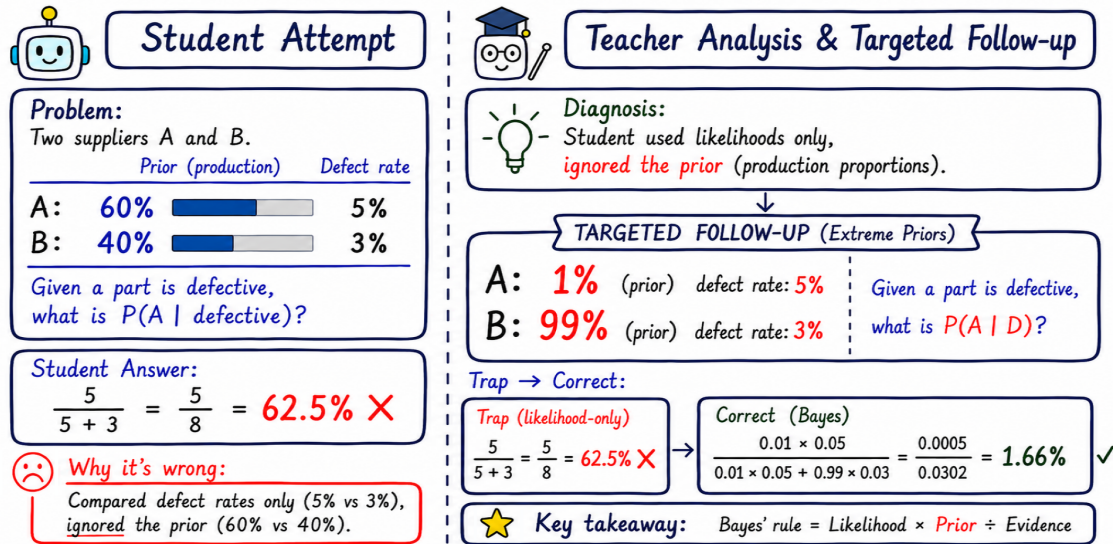


Figure 1: **An illustrative example of DIAG.** Left: The student directly compares likelihoods while ignoring prior probabilities, producing an incorrect posterior. Right: DIAG diagnoses the misconception from the student’s erroneous reasoning trace and generates a targeted follow-up with extreme priors that makes the error more salient, yielding more informative preference pairs for alignment.

the current policy, these iterative reasoning loops keep the learning signal calibrated to the evolving reasoning state of the student model. This capability is crucial for maintaining effective alignment across heterogeneous models. Under this paradigm, supervision typically arises from natural contrasts: the student samples multiple rollouts, and a preference pair is formed only when the set contains both correct and incorrect trajectories. However, the efficacy of online rollouts remains fundamentally capped by the nature of the underlying prompt pool. Even if rollouts are generated on-policy and updated in real time, a fixed practice distribution may still fail to sample problems that match the model’s current capability, thereby producing too few informative preference pairs and wasting computational resources. Existing remedies often rely on brute-force scaling: increasing rollout width, expanding the static prompt pool, or using large-scale synthetic datasets (Yu et al., 2023; Toshniwal et al., 2024; Chen et al., 2024c).

A seemingly more adaptive alternative is active learning, which selects high-uncertainty prompts from a candidate pool (Settles, 2009). Yet for iterative preference optimization, active learning remains practice-agnostic in a crucial sense: it typically selects among existing prompts rather than generating targeted follow-ups tailored to the student model’s current state. As a result, it can still produce redundant queries or abrupt difficulty shifts, worsening preference-pair scarcity rather

than alleviating it. This problem is conceptually a curriculum control problem. Classical curriculum learning and self-paced learning show that *how* examples are scheduled can be as important as *how many* examples are seen (Bengio et al., 2009; Kumar et al., 2010). However, applying curriculum ideas to iterative preference optimization introduces an additional challenge: the feedback signal is sparse, noisy at low sample counts, and non-stationary across iterations.

In this paper, we propose **DIAG**, a data-efficient framework for iterative preference distillation in mathematical reasoning. Our core insight is simple: *conditioning teacher generation on student mistakes can improve both the yield of preference pairs and the quality of preference supervision*. As illustrated in Figure 1, instead of treating the practice distribution as static, DIAG closes the loop between *student outcomes* and *next-iteration practice*. It follows a *Diagnose-and-Generate* paradigm: (i) diagnose how much usable preference signal the current curriculum yields; (ii) adaptively balance exploration and exploitation; (iii) allocate topic budgets using Empirical Bayes shrinkage to avoid low-count volatility (Efron and Morris, 1973); and (iv) generate targeted follow-up questions by conditioning a teacher on the student’s recent mistakes, focusing computation near the decision boundary through uncertainty-weighted resampling rather than static pool selection (Settles, 2009). This design increases the *quantity* of effective preference

pairs while improving the *quality* of supervision by making new questions explicitly target observed failure modes. Our contributions are as follows:

- We formalize iterative preference distillation in verifiable mathematical reasoning as a *signal-yield* problem, and empirically show that static practice distributions can lead to a rapid decline in natural contrastive supervision.
- We present **DIAG**, which couples a signal-aware pacing rule with Empirical Bayes topic scoring to robustly steer generation toward high-yield regions without over-reacting to low-count noise.
- We demonstrate that *mistake-conditioned generation* provides a simple and effective way to concentrate practice near the student’s decision boundary, yielding higher-quality preference pairs with greater information density per unit budget than practice-agnostic generation.

2 Related Work

Iterative Preference Optimization in Reasoning

Post-training for LLM alignment has expanded from online RLHF pipelines (Schulman et al., 2017; Ouyang et al., 2022) to simpler preference-based objectives such as DPO (Rafailov et al., 2023), which are easier to scale into iterative loops. Recent iterative variants for reasoning (e.g., IRPO, Pang et al. 2024; SVPO, Chen et al. 2024b) focus on improving the update rule or pair construction. Recent methods such as SAI-DPO dynamically select training data according to the model’s evolving reasoning abilities (Rao et al., 2026a). In contrast, DIAG diagnoses preference-pair yield and further generates targeted practice from the student’s erroneous reasoning traces. DIAG addresses an orthogonal bottleneck: even with a fixed verifier and rollout width, the practice distribution can quickly degenerate into all-correct or all-wrong regimes, starving the learner of usable preference pairs. In the spirit of curriculum learning and prioritized replay (Bengio et al., 2009; Kumar et al., 2010; Schaul et al., 2015; Jiang et al., 2021), our contribution is not a new objective but a signal-aware mechanism that allocates training budget where valid preference pairs are densest.

On-Policy Distillation (OPD) Recent on-policy distillation methods transfer teacher knowledge at the *signal level*: OPSD (Zhao et al., 2026a) and SDPO (Hübotter et al., 2026) let the same model play teacher and student under different contexts

and minimise a token-level KL divergence, while OPCD (Ye et al., 2026) enriches the teacher representation with auxiliary context. These approaches operate on fixed tasks or prompts and improve the supervisory distribution over the *same* problem. DIAG operates at an orthogonal, *data level*: the teacher analyses the student’s erroneous reasoning trace, diagnoses the underlying misconception, and synthesises a new question that forces the student to confront it, thereby reshaping the training distribution rather than the per-token supervision.

Reasoning Data Synthesis Throughout the life-cycle of large language models, data construction, curation, and feedback play central roles (Rao et al., 2025). Existing work has explored diverse data synthesis paradigms, including seed-based generation (Wang et al., 2023; Xu et al., 2025a), seedless generation (Xu et al., 2025b), and structured or design-guided synthesis (Yu et al., 2026c; Liu et al., 2026). Beyond general-purpose synthesis, domain-specific data construction has also been explored in multimodal and multilingual settings (Kang et al., 2026a; Gao et al., 2026). However, most existing approaches do not explicitly adapt generation to the evolving failure patterns of the student model. In contrast, DIAG uses the student’s erroneous reasoning traces to generate targeted follow-up questions, continually reshaping the practice distribution toward more informative preference supervision.

Mistake-Driven Augmentation Closest to our approach, LLM2LLM (Lee et al., 2024) prompts a teacher to generate new samples from student errors, but primarily conditions on the failed question rather than the student’s erroneous reasoning trace. Meanwhile, Rao et al. (2026b) improve exploration efficiency by mining and correcting failed reasoning trajectories, turning unsuccessful explorations into more informative training signals. Beyond reasoning-data augmentation, recent agent research has emphasized the value of execution traces, failure histories, and provenance for diagnosing erroneous states and guiding corrective actions (Yu et al., 2026a). In contrast, DIAG does not repair persistent agent memory; instead, it uses erroneous reasoning traces to diagnose misconceptions and generate targeted follow-up questions, integrating this mistake-driven augmentation with iterative preference distillation. Rather than correcting existing failures, it creates new questions near the decision boundary to improve the efficiency and informativeness of contrastive supervision.

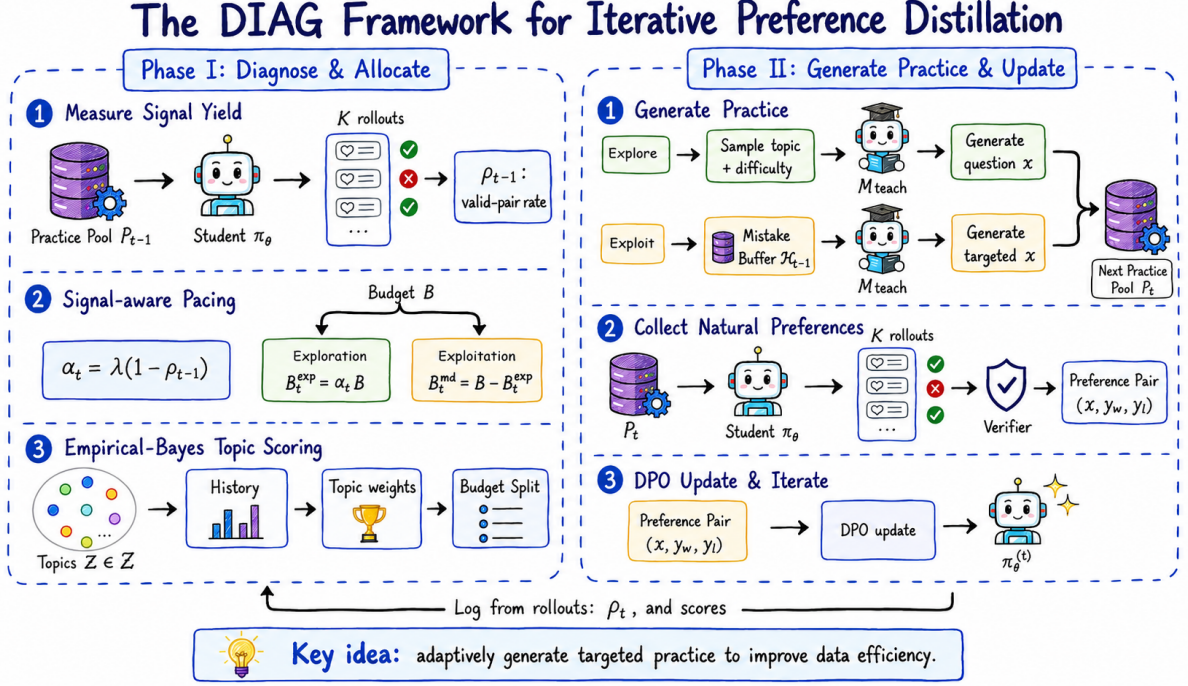


Figure 2: **The DIAG Framework** for Iterative Preference Distillation. DIAG identifies the student’s competence boundary and uncertainty, and generates targeted practice to sustain informative preference pairs, improving both supervision quality and data efficiency for math reasoning alignment.

3 Methods

3.1 Overview

As shown in Figure 2, we present **DIAG**, a framework that improves data efficiency in iterative preference distillation for mathematical reasoning. Rather than training on a fixed prompt pool (*Static Practice*), DIAG adopts a *Diagnose-and-Generate* paradigm: after each DPO iteration it uses the student’s failures to adaptively reshape the practice distribution for the next round.

Concretely, the method operates in two coupled phases (see Algorithm 1 for the overall procedure):

- (1) *Diagnose & Allocate*: estimate the effective pair yield from the previous iteration, set an exploration–exploitation budget split, and assign topic-level quotas via utility estimates smoothed by Empirical Bayes shrinkage to avoid noisy swings in the low-sample regime.
- (2) *Generate Practice*: perform weighted resampling over recent mistakes within each topic and query a teacher to generate new follow-up problems conditioned on the failed question and the student’s incorrect response.

3.2 Preliminaries: Natural Preference Signal

Let π_θ denote the student policy, and let $V(y; x) \in \{0, 1\}$ denote a task verifier. For each prompt x ,

we draw K stochastic rollouts

$$\{y^{(k)}\}_{k=1}^K \sim \pi_\theta(\cdot | x)$$

and obtain correctness labels $\{r^{(k)}\}_{k=1}^K$, where $r^{(k)} = V(y^{(k)}; x)$.

Natural Pair Existence A prompt x is *valid* for constructing preference pairs if its rollouts contain at least one correct and one incorrect response:

$$\mathbb{I}_{\text{valid}}(x) \triangleq \mathbb{I}(\exists i, j : r^{(i)} = 1 \wedge r^{(j)} = 0). \quad (1)$$

Define the signal yield of a practice pool P_t as

$$\rho_t \triangleq \frac{1}{|P_t|} \sum_{x \in P_t} \mathbb{I}_{\text{valid}}(x). \quad (2)$$

When ρ_t collapses (e.g., many prompts yield all-correct or all-wrong rollout sets), iterative preference learning experiences gradient starvation. DIAG explicitly adapts the practice generation process to maintain high ρ_t over iterations.

3.3 Phase I: Diagnose & Allocate

At each iteration, DIAG first diagnoses the amount of *usable preference signal* produced by the current practice distribution, then allocates the generation budget across topics accordingly. This phase combines signal-aware pacing with Empirical Bayes

Algorithm 1 DIAG Distillation Loop

```

1: Input: Topic set  $\mathcal{Z}$ , Difficulty levels  $\mathcal{D}$ , teacher  $\mathcal{M}_{\text{teach}}$ ,
   student  $\pi_\theta$ , rollout width  $K$ , per-iter budget  $B$ 
2: Init:  $N_z \leftarrow 0$ ,  $C_z \leftarrow 0 \forall z$ ;  $\mathcal{H}_0 \leftarrow \emptyset$ ;  $\alpha_1 \leftarrow 1$ 
3: for iteration  $t = 1$  to  $T$  do
4:   // Phase I: Diagnose & Allocate
5:    $B_t^{\text{rnd}} \leftarrow \lfloor \alpha_t B \rfloor$ ,  $B_t^{\text{exp}} \leftarrow B - B_t^{\text{rnd}}$ 
6:   Compute scores  $v_z$ ;  $\pi_{\text{sampler}}(z) \propto v_z$ 
7:    $\mathcal{D}_t \leftarrow \emptyset$ ,  $P_t \leftarrow \emptyset$ ,  $\mathcal{H}_t \leftarrow \emptyset$ 
8:   // Phase II: Generate Practice
9:   for  $j = 1$  to  $B$  do
10:    if  $j \leq B_t^{\text{rnd}}$  then
11:      Sample  $(z, d) \sim \mathcal{U}(\mathcal{Z} \times \mathcal{D})$   $\triangleright$  explore
12:       $x \sim \mathcal{M}_{\text{teach}}(\cdot | z, d)$ 
13:    else
14:      Sample  $z \sim \pi_{\text{sampler}}$ ;
15:      Get  $(x_0, d, y_{\text{err}}) \sim \mathcal{H}_{t-1}|z$   $\triangleright$  exploit
16:       $x \sim \mathcal{M}_{\text{teach}}(\cdot | z, d, x_0, y_{\text{err}})$ 
17:    end if
18:     $a^* \leftarrow \mathcal{M}_{\text{teach}}^{\text{ans}}(x)$   $\triangleright$  teacher solves
19:    Roll out  $\{y_k\} \sim \pi_\theta(\cdot | x)$ ;  $r_k \leftarrow \mathcal{V}(y_k; x, a^*)$ 
20:     $P_t \leftarrow P_t \cup \{x\}$ ;  $N_z \leftarrow N_z + 1$ 
21:    if  $\exists i, j : r_i = 1, r_j = 0$  then
22:       $y_w \leftarrow y_i$ ,  $y_l \leftarrow y_j$ 
23:       $\mathcal{D}_t \leftarrow \mathcal{D}_t \cup \{(x, y_w, y_l)\}$ 
24:       $C_z \leftarrow C_z + 1$ 
25:       $\mathcal{H}_t \leftarrow \mathcal{H}_t \cup \{(x, z, d, y_l)\}$ 
26:    end if
27:  end for
28:  Update  $\pi_\theta$  by minimizing  $\mathcal{L}_{\text{DPO}}$  on  $\mathcal{D}_t$ 
29:   $\rho_t \leftarrow |\{x \in P_t : \mathbb{I}_{\text{valid}}(x) = 1\}| / |P_t|$ 
30:   $\alpha_{t+1} \leftarrow \lambda(1 - \rho_t)$   $\triangleright$  next-iter explore rate
31: end for

```

topic scoring to produce an adaptive, targeted, and noise-resilient generation plan.

Signal-aware pacing At the beginning of iteration t , DIAG uses the yield measured on the previous practice pool P_{t-1} to set the exploration rate:

$$\alpha_t \triangleq \lambda(1 - \rho_{t-1}), \quad (3)$$

so that exploration increases when informative natural pairs are scarce and decreases when the current curriculum already produces abundant on-policy contrastive signal. For $t > 1$, by construction, $\alpha_t \in [0, \lambda] \subseteq [0, 1]$, where $\lambda \in (0, 1]$ is a fixed scaling factor controlling how aggressively the sampler shifts budget in response to the measured yield.

Budget split Given a fixed generation budget of B questions per iteration, we allocate

$$B_t^{\text{rnd}} \triangleq \lfloor \alpha_t B \rfloor, \quad B_t^{\text{exp}} \triangleq B - B_t^{\text{rnd}}, \quad (4)$$

where B_t^{rnd} is assigned to *uniform exploration* over the topic–difficulty space (z, d) , and B_t^{exp} is assigned to the *exploitation* stream. Intuitively, when ρ_{t-1} is low, we expand coverage to search for regions that can generate usable preference signal;

when ρ_{t-1} is high, we concentrate the exploitation budget on high-yield topics and recent mistakes to maximize sample efficiency. In practice, this acts as a damping mechanism: when the model encounters a “plateau” with low signal yield, the system automatically allocates more budget to broad exploration to discover new learnable frontiers. See Appendix E for details.

Empirical Bayes topic scoring We maintain per-topic statistics to estimate which topics are most likely to yield informative natural preference pairs, and use these estimates to guide the allocation of the exploitation budget B_t^{exp} . Let $z \in \mathcal{Z}$ denote a topic and let x be a question tagged with topic z . For each attempted question, we update

$$N_z \leftarrow N_z + 1, \quad C_z \leftarrow C_z + \mathbb{I}_{\text{valid}}(x), \quad (5)$$

where N_z counts how many questions of topic z have been attempted, while C_z counts how often topic z produced a usable natural contrastive signal.

We model the latent utility p_z (the probability that topic z yields a valid natural pair) as a Beta random variable with Empirical Bayes prior:

$$p_z \sim \text{Beta}(\alpha_0, \beta_0),$$

$$p_z | (C_z, N_z) \sim \text{Beta}(\alpha_0 + C_z, \beta_0 + N_z - C_z).$$

Here, α_0 and β_0 are prior pseudo-counts controlling the amount of shrinkage toward the global valid rate. Let $\bar{p}$ denote the global mean valid rate across topics, and let $\kappa > 0$ denote the total prior strength. We set $\alpha_0 = \kappa \bar{p}$ and $\beta_0 = \kappa(1 - \bar{p})$. The posterior mean gives a Stein-type shrinkage score:

$$v_z \triangleq \mathbb{E}[p_z | C_z, N_z] = \frac{C_z + \kappa \bar{p}}{N_z + \kappa}. \quad (6)$$

Topic sampling policy We sample topics for the exploitation stream proportionally to v_z :

$$\pi_{\text{sampler}}(z) = v_z / \sum_{z' \in \mathcal{Z}} v_{z'}. \quad (7)$$

This Empirical Bayes scheduler prevents the curriculum from over-reacting to low-count noise while focusing generation on topics that repeatedly yield informative on-policy supervision.

3.4 Phase II: Generate Practice

Given the budget split and topic sampler from Phase I, we next construct the practice pool for the following iteration. The exploration stream samples uniformly, while the exploitation stream conditions on the student’s recent failures to generate targeted follow-up questions.

Mistake buffer At iteration t , we collect a buffer of boundary mistake cases:

$$\mathcal{H}_t \triangleq \{(x, z, d, y_{\text{err}})\}, \quad (8)$$

where (z, d) denotes topic and difficulty, and y_{err} is an incorrect trace from a prompt that also admitted a correct rollout (i.e., a valid preference pair). Only such mixed-outcome prompts enter $\mathcal{H}_t$, as they lie near the student’s competence boundary. Degenerate episodes still update the signal-yield and topic statistics; when their mass grows, the pacing rule shifts budget toward exploration.

Trace-aware question synthesis For each exploitation slot, we first sample a topic $z \sim \pi_{\text{sampler}}(z)$, then retrieve a recent mistake case $(x, z, d, y_{\text{err}}) \in \mathcal{H}_{t-1}$ with the same topic (and matched difficulty), via uncertainty-weighted re-sampling (Settles, 2009), and query the teacher to generate a *targeted variant*:

$$x' \sim \mathcal{M}_{\text{teach}}(\cdot \mid z, d, x, y_{\text{err}}, \text{instruction}). \quad (9)$$

The instruction enforces: (i) preserve the underlying concept (same z), (ii) match difficulty d , and (iii) alter surface form (numbers/context) to avoid memorization while probing the same failure mode. For the exploration stream, questions are generated from metadata alone without mistake conditioning. The resulting questions together form the next practice pool P_t .

Theoretical interpretation Intuitively, generating from mistakes focuses computation near the student’s competence boundary. Appendix A.1 formalizes this: for rollout width K , the probability that a prompt yields a valid preference pair is $1 - p^K - (1 - p)^K$, which peaks at $p = 1/2$ and vanishes in the all-correct and all-wrong regimes. Appendix A.2 then derives a KL-regularized reweighting that exponentially up-weights high-yield prompts. DIAG approximates this ideal update via Empirical-Bayes topic allocation and teacher generation conditioned on recent boundary mistakes.

Policy update Once the practice pool P_t is constructed via Phases I–II, we collect preference pairs by labeling student rollouts based on correctness against the teacher-provided reference answer. The student policy π_θ is then updated by minimizing the standard DPO objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma(\beta[h_\theta(x, y_w) - h_\theta(x, y_l)]) \right], \quad (10)$$

where h_θ is the implicit reward difference and π_{ref} is a fixed reference policy for the current iteration.

4 Experiments

4.1 Setup

Benchmarks We evaluate on two standard math benchmarks GSM8K (Cobbe et al., 2021) and MATH500 (Hendrycks et al., 2021b), together with harder testbeds: MinervaMath (Lewkowycz et al., 2022), Gaokao2023En (Liao et al., 2024), College Math (Tang et al., 2024), OlympiadBench (He et al., 2024), AIME24 (AoPS, 2024), and AMC23 (AoPS, 2023). We use greedy decoding for all benchmarks except AMC23 & AIME24, whose small test sizes make single-run accuracy more sensitive to sampling noise. For these two benchmarks, we sample 8 solutions per problem with temperature $T = 0.1$ and top- $p = 0.95$, and report the mean pass@1 accuracy to reduce evaluation variance.

Student Models We evaluate three open-weight student models with approximately 7B–8B parameters: Qwen2.5-Math-7B (Yang et al., 2024), Qwen3-8B-Base (Yang et al., 2025), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024), spanning multiple model series and capability profiles.

Budget and fairness All methods share the same total student-sample budget. We run IDPO (Guo et al., 2024; Pang et al., 2024) for 7 iterations with 3×10^3 prompts per iteration, keeping the decoding protocol identical across methods. We employ Qwen3-235B (Int4) as the teacher model.

Practice pool construction (what differs) We compare three pool-construction strategies under the same student-sample budget:

- **Static Dataset Pool (Numina).** We use the open-source NuminaMath dataset (Li et al., 2024a) as a fixed practice pool, without online regeneration.
- **Static Gen (Teacher).** The teacher generates all $T \times B$ questions upfront using only metadata such as topic and difficulty. The resulting pool is fixed and partitioned equally across iterations.
- **Ours: Mistake-Conditioned.** At each iteration, the teacher regenerates the pool by conditioning on the student’s failures (Sec. 3).

4.2 Main Results

Table 1 reports the full results. Despite identical student-sample budgets, DIAG achieves the strongest average performance across all three student models. The gains are most pronounced on

Method	Standard			Exam		Competition (Hard)			Avg.
	GSM8K	MATH 500	Minerva Math	Gaokao 2023en	College Math	Olympiad Bench	AIME24 (Avg@8)	AMC23 (Avg@8)	
Qwen2.5-Math-7B									
BASE (Start)	84.2	70.2	26.5	58.7	35.0	33.8	24.2	54.7	48.4
NUMINA → IDPO	85.4	70.4	29.8	59.0	37.8	35.9	22.5	56.6	49.7
STATIC GEN → IDPO	86.5	69.6	33.8	60.8	38.7	35.7	22.1	59.7	50.9
DIAG (Ours)	87.1	70.8	34.9	61.0	41.9	38.4	25.0	63.1	52.8
Qwen3-8B-Base									
BASE (Start)	89.1	70.0	35.3	57.4	36.3	33.9	8.8	50.0	47.6
NUMINA → IDPO	90.9	73.4	34.9	62.3	38.1	40.1	8.3	56.2	50.5
STATIC GEN → IDPO	91.1	76.6	42.6	63.1	39.1	40.0	10.4	52.2	51.9
DIAG (Ours)	92.0	79.0	40.8	64.7	39.7	41.3	14.6	57.5	53.7
Llama-3.1-8B-Instruct									
BASE (Start)	84.8	47.6	25.7	41.6	28.8	14.5	7.1	21.2	33.9
NUMINA → IDPO	86.3	50.8	27.9	43.4	28.9	15.7	5.8	24.4	35.4
STATIC GEN → IDPO	86.9	50.2	29.4	42.1	30.0	15.6	6.7	26.6	35.9
DIAG (Ours)	88.4	48.2	31.6	45.2	32.1	17.6	7.5	28.4	37.4

Table 1: **Main Results on Mathematical Reasoning (Fixed Student-Sample Budget).** We compare four data sources under the *same total student-sample budget*: (BASE) the starting checkpoint, (NUMINA) IDPO on a static open-source pool, (STATIC GEN) IDPO on a fixed teacher-generated pool conditioned on (topic, difficulty), and **DIAG (Ours)** IDPO on online teacher generation conditioned on (failed question, wrong answer, topic, difficulty).

Task	Sampling Strategies		
	w/o Explore	w/o EB	Ours
GSM8K	87.6	86.4	87.1
MATH500	70.4	70.6	70.8
MinervaMath	35.7	35.3	34.9
Gaokao2023	59.5	59.7	61.0
Olympiad	37.5	38.7	38.4
CollegeMath	42.4	40.3	41.9
AIME24	24.2	24.2	25.0
AMC23	58.1	59.4	63.1
Avg.	51.9 (-0.9)	51.8 (-1.0)	52.8

Table 2: **Ablation on Sampling/Scheduling.** We ablate the uniform exploration stream (**w/o Explore**) and EB smoothing (**w/o EB**) under a fixed rollout budget.

harder testbeds, where generic practice pools are more likely to enter degenerate regimes, reducing the availability of natural preference supervision.

4.3 Ablation Study

Sampling & Scheduling We ablate the sampling & scheduling design on Qwen2.5-Math-7B under a fixed rollout budget: (i) **w/o Explore**: remove the uniform exploration stream, allocating the entire budget to score-based exploitation; (ii) **w/o Empirical Bayes (EB)**: replace EB shrinkage with raw per-topic rates; (iii) **Ours**: the full scheduler.

Table 2 shows that both components matter in

Method / Variant	Avg.	Δ
Full DIAG	52.8	–
<i>Impact of conditioning on failed question</i>		
w/o Wrong-QA (no failed context)	51.3	-1.5
<i>Impact of conditioning on wrong answer/trace</i>		
w/o Wrong-A (no wrong answer)	52.1	-0.7

Table 3: **Ablation on Teacher Conditioning Signals.** Δ denotes performance drop relative to the full method.

different ways. Removing exploration yields comparable scores on standard benchmarks but exhibits brittleness on competition-style evaluation (AMC23: 58.1 vs 63.1). Removing EB shrinkage leads to broad degradation on most benchmarks (Avg 51.8), confirming that naive frequency-based scoring is unstable in the low-sample regime.

Teacher Conditioning Signals We isolate which components of the teacher conditioning signal matter for targeted generation, all on Qwen2.5-Math-7B with identical budget and topic/difficulty controls: (i) **w/o Wrong-QA**: teacher conditioned only on topic/difficulty (no failed context); (ii) **w/o Wrong-A**: teacher receives the failed question but not the student’s wrong answer; (iii) **Ours**: full conditioning on topic + difficulty + failed question + wrong answer.

As shown in Table 3, performance increases

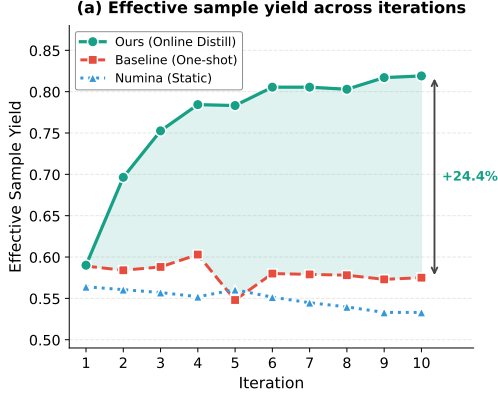


Figure 3: **Effective sample yield across iterations.** DIAG maintains a higher signal yield ρ_t than baselines and avoids degeneration into uninformative regimes.

monotonically as richer error context is provided. Removing the failed context (-1.5) weakens the teacher’s ability to anchor generation to the exact skill being tested. Removing the wrong answer (-0.7) confirms that error-aware generation is not equivalent to simple topic-conditioned rewriting: observing the student’s mistake helps the teacher probe the same misconception more reliably.

4.4 Cross-Task Generalization

To verify that DIAG does not incur an alignment tax, we evaluate Qwen3-8B-Base on BBH (Suzgun et al., 2023), HumanEval (Chen et al., 2021), MMLU (Hendrycks et al., 2021a), and TruthfulQA (Lin et al., 2022). Table 4 shows that DIAG maintains comparable performance on knowledge and truthfulness benchmarks while showing mild positive transfer on reasoning and code generation.

Method	BBH	Human Eval	MMLU	TruthfulQA	
				MC1	MC2
Qwen3-8B-Base	75.5	64.6	76.9	33.5	51.1
+ Static Gen	75.6 [↑]	64.6	76.9	33.6 [↑]	51.2[↑]
+ DIAG	75.6[↑]	64.7[↑]	76.9	33.9[↑]	51.0

Table 4: **Cross-Task Generalization and Alignment Tax Evaluation.** Bold denotes the best result; [↑] indicates improvement over the base model.

5 Analysis

5.1 Effective Sample Yield over Training

A central claim of our method is that optimizing the practice distribution improves sample efficiency under a fixed rollout budget. Following (2), we track the effective yield ρ_t , defined as the fraction

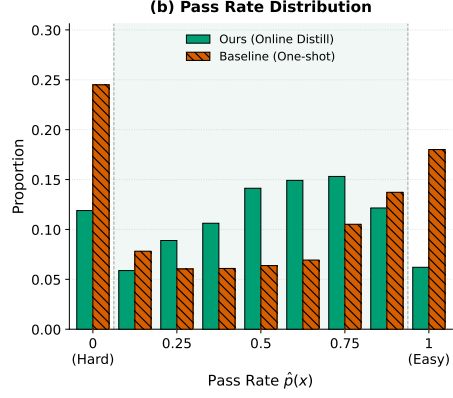


Figure 4: **Pass-rate distribution.** Our method concentrates more mass in the intermediate region, reducing both all-correct and all-wrong degeneracy.

of questions whose K rollouts contain both correct and incorrect outcomes, across iterations on Qwen2.5-Math-7B.

Figure 3 reports an extended 10-iteration diagnostic run: our method increases ρ_t from ~ 0.59 to ~ 0.83 (+24 pp) over 10 iterations, whereas both Static Gen and static Numina gradually decline to around 0.55, confirming that targeted generation keeps the training distribution near the student’s moving competence boundary.

5.2 Pass-Rate Distribution

The yield gap above suggests that the key advantage comes from *where* the model practices. We analyze the pass-rate distribution by estimating $\hat{p}(x) \triangleq \frac{1}{K} \sum_{k=1}^K r_k(x)$ for each generated question. Mass near $\hat{p}(x) \approx 1$ or $\hat{p}(x) \approx 0$ indicates wasted budget on trivial or intractable questions, respectively; both extremes cause gradient starvation. Figure 4 shows that our method concentrates more mass in the intermediate band, where the student is neither always correct nor always wrong. This suggests that mistake-conditioned generation acts as a calibration operator: it improves *difficulty matching* rather than simply making questions harder.

5.3 Iso-Effective Comparison

Our sampler produces more mixed episodes, which could confound improvements due purely to data quantity. To disentangle quantity from quality, we subsample each method to the same number ($\sim 8k$) of effective training instances and train the same Qwen2.5-Math-7B student. We additionally include two error-driven baselines under the same DPO framework: LLM2LLM-style (Lee et al., 2024), which generates new questions from stu-

Method	Eff. Samples	Avg.	Δ
Qwen2.5-Math-7B	—	48.4	—
+ SPIN (Chen et al., 2024d)	8k	48.9	+0.5
+ NuminaMath (Li et al., 2024a)	8k	49.0	+0.6
+ Static IDPO (Zhang et al., 2025)	8k	49.7	+1.3
+ LLM2LLM (Lee et al., 2024)	8k	50.1	+1.7
+ Ours	8k	50.7	+2.3

Table 5: **Iso-effective comparison.** All methods use $\sim 8k$ effective instances and the same gradient steps. LLM2LLM-style adopts error-driven generation but trains with DPO for fair comparison.

dent failures but without yield-aware scheduling (originally SFT; we use DPO for fair comparison); and SPIN (Chen et al., 2024d), which constructs self-play preference pairs on a fixed prompt pool.

As shown in Table 5, our method remains consistently better even when the effective sample count is held constant. SPIN (+0.5), operating on fixed prompts, confirms that prompt-level curriculum matters beyond response-level contrast. LLM2LLM-style (+1.7) benefits from error-driven generation but conditions the teacher only on the failed question; DIAG’s additional conditioning on the student’s wrong reasoning trace enables follow-ups that target the specific misconception, and its EB-guided scheduling further improves signal yield. This gap is consistent with the analysis in Appendix A.1 and A.5: boundary prompts produce more mixed rollouts, placing the model in a higher-curvature, more informative update region.

6 Conclusion

We showed that the dominant bottleneck in iterative preference distillation for math reasoning is signal scarcity: as the student improves, static practice pools degenerate into uninformative regimes. To address this, we proposed DIAG, a Diagnose-and-Generate framework that reshapes the practice distribution via signal-aware pacing, Empirical-Bayes topic prioritization, and mistake-conditioned generation. Across diverse students and challenging math evaluations, DIAG consistently improves both data efficiency and final reasoning performance, with gains persisting even under iso-effective training, indicating higher-quality supervision beyond increased signal yield.

Limitations

DIAG diagnoses signal scarcity in real time via Empirical Bayes shrinkage and uses mistake-

conditioned generation to produce targeted follow-up problems near the student model’s moving competence boundary. Two main limitations remain: (i) the Diagnose-and-Generate loop relies on a capable teacher model for question synthesis, which may not be available or affordable in all settings; (ii) our evaluation is restricted to mathematical reasoning with reliable verifiers, and it is unclear how well the framework transfers to domains where correctness cannot be automatically checked.

Ethics Statement

This work complies with the ACL Ethics Policy. All experiments are conducted using publicly available datasets and widely accessible models, with proper citations provided throughout. Our method does not involve human subjects, private data, or any form of data annotation beyond automated answer verification. We acknowledge that LLMs, including those trained with our framework, may still produce incorrect or misleading reasoning steps. While DIAG improves mathematical reasoning reliability, users should independently verify model outputs before relying on them in high-stakes or real-world decision-making scenarios.

Acknowledgements

We gratefully acknowledge the members of Huawei Large Model Data Technology Lab for their valuable advice and continued support throughout this work. We also appreciate the inspiring academic environment and scholarly exchanges at the Department of Statistics and Data Science and the Department of Mathematical Sciences, Tsinghua University, and thank the faculty members and fellow students for their guidance, perspectives, and constructive discussions. Finally, we thank the anonymous reviewers and area chairs for their thoughtful and constructive feedback, which contributed to improving the quality of this work.

References

- Art of Problem Solving AoPS. 2023. AMC Problems and Solutions. https://artofproblemsolving.com/wiki/index.php/AMC_12_Problems_and_Solutions. Accessed: 2026-08-20.
- Art of Problem Solving AoPS. 2024. AIME Problems and Solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions. Accessed: 2026-08-20.

- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. [Curriculum learning](#). In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.
- Guhan Chen, Yicheng Li, and Qian Lin. 2024a. [On the impacts of the random initialization in the neural tangent kernel theory](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 35909–35944. Curran Associates, Inc.
- Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024b. [Step-level value preference optimization for mathematical reasoning](#). *arXiv preprint arXiv:2406.10858*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. [Evaluating large language models trained on code](#). *Preprint*, arXiv:2107.03374.
- Nuo Chen, Ning Wu, Jianhui Chang, Linjun Shou, and Jia Li. 2024c. [ControlMath: Controllable data generation promotes math generalist models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12201–12217, Miami, Florida, USA. Association for Computational Linguistics.
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024d. [Self-play fine-tuning converts weak language models to strong language models](#). *arXiv preprint arXiv:2401.01335*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Quan Cui, Boyan Zhou, Yu Guo, Weidong Yin, Hao Wu, Osamu Yoshie, and Yubo Chen. 2021. [Contrastive vision-language pre-training with limited resources](#). *arXiv preprint arXiv:2112.09331*.
- Guangyuan Dong, Chuang Liu, Yangchen Zeng, Haoyu Wang, Xiaoyang Yu, Pinlong Zhao, Yuchao Hou, Ziwei Li, and Zheng Lin. 2026. [When generated images look right and retrieve wrong: Coverage-guided cross-scale re-indexing for knowledge-faithful generative perception](#). *Preprint*, arXiv:2608.20810.
- Bradley Efron and Carl Morris. 1973. [Stein’s estimation rule and its competitors—an empirical bayes approach](#). *Journal of the American Statistical Association*, 68(341):117–130.
- Jian Gao, Richeng Xuan, Zhaolu Kang, Dingshi Liao, Wenxin Huang, Zongmou Huang, Yangdi Xu, Bowen Qin, Zheqi He, Xi Yang, Changjinli, and Yonghua Lin. 2026. [LaoBench: A large-scale multidimensional Lao benchmark for large language models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23727–23743, San Diego, California, United States. Association for Computational Linguistics.
- Gemini Team. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- 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, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. 2024. [Direct language model alignment from online ai feedback](#). *Preprint*, arXiv:2402.04792.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. [OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. [Measuring mathematical problem solving with the math dataset](#). *Preprint*, arXiv:2103.03874.
- Jonas Hübötter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido

- Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. 2026. [Reinforcement learning via self-distillation](#). *arXiv preprint arXiv:2601.20802*.
- Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. 2021. [Prioritized level replay](#). In *International Conference on Machine Learning*, pages 4940–4950. PMLR.
- Zhaolu Kang, Junhao Gong, Jiaxu Yan, Wanke Xia, Yian Wang, Ziwen Wang, Huaxuan Ding, Zhuo Cheng, Wenhao Cao, Zhiyuan Feng, Siqi He, Shan-nan Yan, Junzhe Chen, Xiaomin He, Chaoya Jiang, Wei Ye, Kaidong Yu, and Xuelong Li. 2026a. [Hss-bench: Benchmarking humanities and social sciences ability for multimodal large language models](#). *Preprint*, arXiv:2506.03922.
- Zhaolu Kang, Yingjie He, Kehan Jiang, Leqi Zheng, Jiachen Qian, Qianyu Zhang, Chunlei Meng, Yujie Feng, Yuan Wang, Stephen Dou, Aming Wu, Pengxiang Zhao, Jiaxin Liu, Guansu Wang, Zeyu Zhang, Lei Wang, Qishi Zhan, Xiaomin He, Meisheng Zhang, and 2 others. 2026b. [How order-sensitive are llms? orderprobe for deterministic structural reconstruction](#). *Preprint*, arXiv:2601.08626.
- M. Kumar, Benjamin Packer, and Daphne Koller. 2010. [Self-paced learning for latent variable models](#). In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.
- Jianfa Lai, Zixiong Yu, Songtao Tian, and Qian Lin. 2023. [Generalization ability of wide residual networks](#). *Preprint*, arXiv:2305.18506.
- Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangu Peng, and Jiaya Jia. 2024. [Step-dpo: Step-wise preference optimization for long-chain reasoning of llms](#). *arXiv preprint arXiv:2406.18629*.
- Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael Mahoney, Kurt Keutzer, and Amir Gholami. 2024. [Llm2llm: Boosting llms with novel iterative data enhancement](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6498–6526.
- 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. 2022. [Solving quantitative reasoning problems with language models](#). *Preprint*, arXiv:2206.14858.
- 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. 2024a. [Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions](#). *Hugging Face repository*.
- Yicheng Li, Zixiong Yu, Guhan Chen, and Qian Lin. 2024b. [On the eigenvalue decay rates of a class of neural-network related kernel functions defined on general domains](#). *Journal of Machine Learning Research*, 25(82):1–47.
- Minpeng Liao, Chengxi Li, Wei Luo, Wu Jing, and Kai Fan. 2024. [MARIO: MATH reasoning with code interpreter output - a reproducible pipeline](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 905–924, Bangkok, Thailand. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Weize Liu, Yongchi Zhao, Yijia Luo, Mingyu Xu, Jiaheng Liu, Yanan Li, Xiguo Hu, Zhiqi Bai, Yuchi Xu, Wenbo Su, and Bo Zheng. 2026. [DESIGNER: Design-logic-guided multidisciplinary data synthesis for LLM reasoning](#). In *The Fourteenth International Conference on Learning Representations*.
- Shuo Lu, Yanyin Chen, Wei Feng, Jiahao Fan, Fengheng Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Ching Law, and Jian Liang. 2025. [Uni-layout: Integrating human feedback in unified layout generation and evaluation](#). *Preprint*, arXiv:2508.02374.
- Shuo Lu, Jianjie Cheng, Yinyu Xu, Yongcan Yu, Lijun Sheng, Peijie Wang, Siru Jiang, Yongguan Hu, Run Ling, Yihua Shao, Ao Ma, Wei Feng, Lingxiao He, Meng Wang, Qianlong Xie, Xingxing Wang, Nicu Sebe, Ran He, and Jian Liang. 2026a. [Do mllms really understand space? a mathematical reasoning evaluation](#). *Preprint*, arXiv:2602.11635.
- Shuo Lu, Haohan Wang, Wei Feng, Weizhen Wang, Shen Zhang, Yaoyu Li, Ao Ma, Zheng Zhang, Jingjing Lv, Junjie Shen, Ching Law, Bing Zhan, Yuan Xu, Huizai Yao, Yongcan Yu, Chenyang Si, and Jian Liang. 2026b. [One size, many fits: Aligning diverse group-wise click preferences in large-scale advertising image generation](#). *Preprint*, arXiv:2602.02033.
- Microsoft. 2025. [Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras](#). *Preprint*, arXiv:2503.01743.
- OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- 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. 2022. [Training language models to follow instructions with human feedback](#). *NeurIPS*, 35:27730–27744.

- Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. 2024. [Iterative reasoning preference optimization](#). *Advances in Neural Information Processing Systems*, 37:116617–116637.
- Pin Qian, Su Wang, Yihang Chen, Qiaolin Yu, Xiaoyuan Wang, Zhitong Guo, Zhicheng Wang, and Junxian You. 2026. [Toward user-conditioned evaluation of personal llm agents under temporal interventions](#). *Preprint*, arXiv:2607.21635.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). *Advances in neural information processing systems*, 36:53728–53741.
- Jun Rao, Xuebo Liu, Hexuan Deng, Zepeng Lin, Zixiong Yu, Jiansheng Wei, Xiaojun Meng, and Min Zhang. 2026a. [Dynamic sampling that adapts: Self-aware iterative data persistent optimization for mathematical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 28337–28350, San Diego, California, United States. Association for Computational Linguistics.
- Jun Rao, Xuebo Liu, Haotian Yan, Junjie Shen, Haosi Mo, Yanghaopeng Dong, Zihao Yan, Ziyi Wang, Zepeng Lin, Xiaojun Meng, Zixiong Yu, Liqun Deng, Jiansheng Wei, Yunhe Wang, and Min Zhang. 2025. [A data-centric perspective on the lifecycle of large language models](#). *TechRxiv*, 2025(1220).
- Jun Rao, Zixiong Yu, Xuebo Liu, Guhan Chen, Jing Li, Hejin Wang, Jiansheng Wei, Xiaojun Meng, and Min Zhang. 2026b. [Mining or synthesis? rethinking exploration efficiency in iterative alignment of mathematical reasoning](#). *Preprint*, arXiv:2602.05370.
- Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2015. [Prioritized experience replay](#). *arXiv preprint arXiv:1511.05952*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *arXiv preprint arXiv:1707.06347*.
- Burr Settles. 2009. [Active learning literature survey](#).
- Qiming Shi, Zhaolu Kang, Yunfan Zhou, Di Weng, and Yingcai Wu. 2026. [Spader: Step-wise peer advantage with diversity-aware exploration rewards for multi-answer question answering](#). *Preprint*, arXiv:2606.00593.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. [Challenging BIG-bench tasks and whether chain-of-thought can solve them](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.
- Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. 2024. [Mathscale: Scaling instruction tuning for mathematical reasoning](#). In *ICML*, pages 47885–47900.
- Songtao Tian, Guhan Chen, Bohan Li, Jingyi Ma, and Zixiong Yu. 2026. [Curvature-adaptive consistency flow matching: Autonomous trajectory optimization via reinforcement learning](#). *Preprint*, arXiv:2606.22394.
- Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. [Openmathinstruct-1: A 1.8 million math instruction tuning dataset](#). *Advances in Neural Information Processing Systems*, 37:34737–34774.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Xiaolong Wang, Zhaolu Kang, Wangyuxuan Zhai, Xinyue Lou, Yunghwei Lai, Ziyue Wang, Yawen Wang, Kaiyu Huang, Yile Wang, Peng Li, and Yang Liu. 2025. [MUCAR: Benchmarking multilingual cross-modal ambiguity resolution for multi-modal large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15026–15048, Suzhou, China. Association for Computational Linguistics.
- Yiqi Wang, Zihao Yan, Jiaqi Zhang, Zhangkai Wu, Mingkai Zheng, Zequn Sun, Yanming Zhu, and Taotao Cai. 2026a. [Map-graph: Provenance-aware shared memory for multi-agent workflows](#). *Preprint*, arXiv:2608.10509.
- Yiqi Wang, Jiaqi Zhang, Taotao Cai, Zirui Liu, Qingqiang Sun, Zequn Sun, Zhangkai Wu, Manqing Dong, Mingkai Zheng, Xuefei Yin, and Yanming Zhu. 2026b. [From agent traces to trust: A survey of evidence tracing and execution provenance in llm agents](#). *Preprint*, arXiv:2606.04990.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [Self-instruct: Aligning language models with self-generated instructions](#). In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.
- Wenxun Wu, Yuanyang Li, Guhan Chen, Linyue Wang, and Hongyang Chen. 2026. [Tool-augmented policy optimization: Synergizing reasoning and adaptive tool use with reinforcement learning](#). *Preprint*, arXiv:2510.07038.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2025a. [Wizardlm: Empowering large pre-trained language models to follow complex instructions](#). *Preprint*, arXiv:2304.12244.

- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2025b. [Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing](#). In *The Thirteenth International Conference on Learning Representations*.
- 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, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. [Qwen2.5-math technical report: Toward mathematical expert model via self-improvement](#). *Preprint*, arXiv:2409.12122.
- Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2026. [On-policy context distillation for language models](#). *Preprint*, arXiv:2602.12275.
- Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. [Limo: Less is more for reasoning](#). *Preprint*, arXiv:2502.03387.
- Bihui Yu, Dawei Liu, Huiyang Shi, GuiYong Chang, Jingxuan Wei, Linzhuang Sun, Songtao Tian, and Liping Bu. 2024. [Sam-wav2lip++: Enhancing behavioral realism in synthetic agents through audio-driven speech and action refinement](#). In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2999–3006.
- Caili Yu, Yiqi Wang, Jiaqi Zhang, Yiqun Duan, Mingkai Zheng, Zhangkai Wu, Kaize Shi, and Taotao Cai. 2026a. [From faulty memories to corrected actions: Dependency-guided rollback repair for memory-augmented agents](#). *Preprint*, arXiv:2608.10502.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhen-guo Li, Adrian Weller, and Weiyang Liu. 2023. [Metamath: Bootstrap your own mathematical questions for large language models](#). *arXiv preprint arXiv:2309.12284*.
- Zixiong Yu, Guhan Chen, Jianfa Lai, Bohan Li, and Songtao Tian. 2026b. [Branch scaling manifests as implicit architectural regularization for improving generalization in overparameterized resnets](#). In *Forty-third International Conference on Machine Learning*.
- Zixiong Yu, Jun Rao, Guhan Chen, Songtao Tian, Bohan Li, Jiansheng Wei, Min Zhang, and Xiaojun Meng. 2026c. [MathAgent: Adversarial evolution of constraint graphs for mathematical reasoning data synthesis](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 28298–28321, San Diego, California, United States. Association for Computational Linguistics.
- Zixiong Yu, Songtao Tian, and Guhan Chen. 2025. [Divergence of neural tangent kernel in classification problems](#). In *The Thirteenth International Conference on Learning Representations*.
- Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. [Scaling relationship on learning mathematical reasoning with large language models](#). *arXiv preprint arXiv:2308.01825*.
- Hanning Zhang, Jiarui Yao, Chenlu Ye, Wei Xiong, and Tong Zhang. 2025. [Online-dpo-r1: Unlocking effective reasoning without the ppo overhead](#). *GitHub repository*.
- Haobo Zhang, Xutao Mao, Guangyuan Dong, Ziwei Li, Xuanbo Su, Kaijie Chen, Jing Yang, and Zheng Lin. 2026. [Memmark: State-evolution attribution watermarking for agent long-term memory systems](#). *Preprint*, arXiv:2605.25002.
- Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026a. [Self-distilled reasoner: On-policy self-distillation for large language models](#). *Preprint*, arXiv:2601.18734.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, Xiaolei Wang, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, and 4 others. 2026b. [A survey of large language models](#). *Frontiers of Computer Science*, 20(12).
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. [LIMA: Less is more for alignment](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Kaijie Zhu, Yuzhou Nie, Yijiang Li, Yiming Huang, Jialian Wu, Jiang Liu, Ximeng Sun, Zhenfei Yin, Lun Wang, Zicheng Liu, Emad Barsoum, William Yang Wang, and Wenbo Guo. 2026. [Termigen: High-fidelity environment and robust trajectory synthesis for terminal agents](#). *Preprint*, arXiv:2602.07274.

A Theoretical Perspective: Signal Yield and Practice-Space Reweighting

This appendix provides a compact theoretical view of DIAG. The main idea is that useful natural preference pairs are most likely to arise near the student’s competence boundary, where the model sometimes succeeds and sometimes fails. DIAG can then be interpreted as a teacher-mediated approximation to an ideal practice-distribution reweighting that moves probability mass toward such high-yield regions.

A.1 Signal Yield Peaks at the Competence Boundary

Let $\pi_\theta(y \mid x)$ be the student policy and let $V(y; x) \in \{0, 1\}$ be the verifier. For a prompt x , define the student’s pass probability as

$$p_\theta(x) = \Pr_{y \sim \pi_\theta(\cdot \mid x)} [V(y; x) = 1]. \quad (11)$$

With rollout width $K \geq 2$, a valid natural preference pair exists when at least one rollout is correct and at least one rollout is incorrect. Thus the probability that x yields a valid pair is

$$u_{\theta,K}(x) = 1 - p_\theta(x)^K - (1 - p_\theta(x))^K. \quad (12)$$

Proposition A.1 For any $K \geq 2$, the function

$$u_K(p) = 1 - p^K - (1 - p)^K, \quad p \in [0, 1], \quad (13)$$

is uniquely maximized at $p = 1/2$, with value

$$u_K(1/2) = 1 - 2^{1-K}. \quad (14)$$

Moreover, $u_K(0) = u_K(1) = 0$.

Proof The derivative is

$$u'_K(p) = K[(1 - p)^{K-1} - p^{K-1}], \quad (15)$$

which vanishes only at $p = 1/2$. The second derivative is

$$u''_K(p) = -K(K - 1)[p^{K-2} + (1 - p)^{K-2}] < 0$$

for $p \in (0, 1)$ and $K \geq 2$, so u_K is strictly concave and has a unique maximum at $p = 1/2$. The boundary values follow directly. $\square$

This proposition formalizes why all-correct and all-wrong prompts are inefficient: both regimes have low probability of producing usable preference pairs. Prompts near the competence boundary, where $p_\theta(x)$ is intermediate, provide denser supervision under the same rollout budget.

For a practice distribution μ , define the expected valid-pair yield as

$$J_{\theta,K}(\mu) = \mathbb{E}_{x \sim \mu} [u_{\theta,K}(x)]. \quad (16)$$

If M prompts are sampled in an iteration, the expected number of effective DPO training instances is $MJ_{\theta,K}(\mu)$.

A.2 Ideal KL-Regularized Practice Reweighting

The previous result suggests that an ideal practice distribution should upweight prompts with large $u_{\theta,K}(x)$. Let μ_{ref} be the distribution induced by a metadata-only teacher generator. Consider the KL-regularized update ($\eta > 0$)

$$\mu_\eta \in \arg \max_{\mu \ll \mu_{\text{ref}}} \left\{ \mathbb{E}_{x \sim \mu} [u_{\theta,K}(x)] - \frac{1}{\eta} D_{\text{KL}}(\mu \parallel \mu_{\text{ref}}) \right\}. \quad (17)$$

Proposition A.2 The unique optimizer of (17) is

$$\mu_\eta(x) = \frac{\mu_{\text{ref}}(x) \exp(\eta u_{\theta,K}(x))}{Z_\eta}, \quad (18)$$

$$Z_\eta = \mathbb{E}_{x \sim \mu_{\text{ref}}} [\exp(\eta u_{\theta,K}(x))].$$

Proof By the calculus of variations, the stationarity condition of Eq. (17) gives

$$u_{\theta,K}(x) - \frac{1}{\eta} \left(\log \frac{\mu(x)}{\mu_{\text{ref}}(x)} + 1 \right) + c = 0, \quad (19)$$

where c enforces normalization. Rearranging yields

$$\mu(x) \propto \mu_{\text{ref}}(x) \exp(\eta u_{\theta,K}(x)), \quad (20)$$

which gives Eq. (18). $\square$

Eq. (18) is an idealized exponential reweighting of the practice distribution toward high-yield prompts. DIAG does not execute this update exactly, since estimating $u_{\theta,K}(x)$ for a large candidate space would be expensive. Instead, it approximates this reweighting using observable signals from the previous iteration: topic-level valid-pair statistics and local mistake contexts.

A.3 Mistake-Conditioned Generation as an Amortized Approximation

Let z denote prompt metadata such as topic and difficulty, and let $\nu_t(z)$ be the scheduler over metadata at iteration t . A metadata-only generator induces

a kernel $T_0(x | z)$, while the mistake-conditioned teacher induces

$$T_\phi(x | z, m), \quad (21)$$

where m contains a failed question and a wrong student trace. In DIAG, mistake contexts are drawn from recent prompts that produced valid natural preference pairs. These prompts are empirical boundary cases: the student has both successful and failed rollouts on them. Degenerate all-correct or all-wrong prompts are counted in yield statistics and affect pacing, but are not used as exploitation anchors.

Let $q_t(m | z)$ be the empirical distribution over mistake contexts for topic z . The metadata-only and DIAG-induced practice distributions are

$$\mu_0(x) = \mathbb{E}_{z \sim \nu_t} [T_0(x | z)], \quad (22)$$

$$\mu_{\text{DIAG}}(x) = \mathbb{E}_{z \sim \nu_t} \mathbb{E}_{m \sim q_t(\cdot | z)} [T_\phi(x | z, m)]. \quad (23)$$

For each metadata slice z , define the local yield gain

$$\Delta_z = \mathbb{E}_{m \sim q_t(\cdot | z)} \mathbb{E}_{x \sim T_\phi(\cdot | z, m)} [u_{\theta, K}(x)] - \mathbb{E}_{x \sim T_0(\cdot | z)} [u_{\theta, K}(x)]. \quad (24)$$

Then, by linearity of expectation,

$$J_{\theta, K}(\mu_{\text{DIAG}}) - J_{\theta, K}(\mu_0) = \mathbb{E}_{z \sim \nu_t} [\Delta_z]. \quad (25)$$

Thus, DIAG improves expected valid-pair yield whenever mistake-conditioned generation produces less degenerate follow-up prompts than metadata-only generation on average. This condition is not assumed as a theorem about all teachers; rather, it is the empirical property evaluated through the yield and pass-rate analyses in the main text.

A.4 Empirical-Bayes Stabilization

DIAG uses topic-level statistics to allocate the exploitation budget. For topic z , let N_z be the number of attempted prompts and C_z the number that yielded valid natural pairs. The raw rate C_z/N_z can be noisy when N_z is small. The Empirical-Bayes score is

$$v_z = \frac{C_z + \kappa \bar{p}}{N_z + \kappa} = \frac{N_z}{N_z + \kappa} \cdot \frac{C_z}{N_z} + \frac{\kappa}{N_z + \kappa} \cdot \bar{p}, \quad (26)$$

where $\bar{p}$ is the global valid-pair rate and κ is the prior strength. Thus v_z shrinks low-count topic estimates toward the global mean, preventing the scheduler from overreacting to a few lucky samples. As N_z grows, the topic-specific empirical rate dominates.

A.5 Connection to DPO Curvature

The signal-yield view is also consistent with the local geometry of the DPO loss. For a pair (x, y_w, y_l) , let

$$s_\theta = h_\theta(x, y_w) - h_\theta(x, y_l), \quad (27)$$

$$\ell(s_\theta) = -\log \sigma(\beta s_\theta). \quad (28)$$

With respect to $a = \beta s_\theta$, the curvature is

$$\frac{\partial^2 \ell}{\partial a^2} = \sigma(a)(1 - \sigma(a)), \quad (29)$$

which is maximized at $a = 0$. Hence, among valid preference pairs, updates are locally most sensitive when the model is uncertain between the preferred and dispreferred responses. This does not equate prompt-level pass probability with pairwise DPO uncertainty, but it supports the same design principle: practice near the student’s moving boundary tends to produce more informative supervision.

B Complexity Analysis

We analyze computational cost through the lens of *effective supervision*. Let C_{roll} be the unit cost of one student rollout and C_{teach} the unit cost of one teacher generation. At iteration t , we use a practice set P_t of size $|P_t| = M$ and sample K rollouts per question, so the student-side cost per iteration is

$$C_t^{\text{stu}} = MK C_{\text{roll}}. \quad (30)$$

We focus on verifiable preference learning where a question x yields a usable DPO pair iff it contains both correct and incorrect rollouts (Eq. (1)). Recall the effective (valid-pair) yield is defined as

$$\rho_t \triangleq \frac{1}{M} \sum_{x \in P_t} \mathbb{I}_{\text{valid}}(x), \quad (31)$$

following Eq. (2). The expected number of effective training pairs in iteration t is

$$\mathbb{E}[|D_t|] = M \rho_t. \quad (32)$$

Compute-matched teacher budget. In our experiments, STATIC GEN and DIAG are matched by the same total number of teacher-generated questions. Concretely, over T iterations both methods use TM teacher-generated questions, implying identical total teacher cost:

$$C_{\text{total}}^{\text{teach}} = TM C_{\text{teach}}. \quad (33)$$

The difference is *not* the amount of teacher compute, but *how* this fixed teacher budget is spent: STATIC GEN generates a fixed pool using only (topic, difficulty), while DIAG allocates the same number of generations online and conditions on the student’s recent failures to reshape the practice distribution.

Cost per effective pair (effective gradient cost).

Since the student rollout budget per iteration is fixed, the key efficiency question is: *how many of these rollouts produce gradient-bearing preference pairs?* Define the per-iteration compute cost per effective pair as

$$\begin{aligned}\mathcal{E}_t &\triangleq \frac{\mathcal{C}_t^{\text{stu}} + \mathcal{C}_t^{\text{teach}}}{\mathbb{E}[|\mathcal{D}_t|]} = \frac{MK C_{\text{roll}} + \mathcal{C}_t^{\text{teach}}}{M\rho_t} \\ &= \frac{K C_{\text{roll}}}{\rho_t} + \frac{\mathcal{C}_t^{\text{teach}}}{M\rho_t}.\end{aligned}\quad (34)$$

Under compute matching, $\mathcal{C}_t^{\text{teach}}$ is the same order for both methods (and the total teacher budget is identical), therefore the dominant factor is the yield ρ_t . When a static pool (or practice-agnostic generation) degenerates into all-correct or all-wrong regimes, ρ_t decreases and $\mathcal{E}_t$ increases, meaning a growing fraction of rollouts are wasted with no contrastive supervision.

DIAG improves effective sample efficiency. By explicitly optimizing the practice distribution to stay near the student’s moving decision boundary, DIAG sustains a substantially higher ρ_t across iterations (Sec. 5). Plugging this into Eq. (34) yields a strictly lower compute-per-effective-pair:

$$\begin{aligned}\rho_t^{\text{DIAG}} &> \rho_t^{\text{STATIC GEN}} \\ \implies \mathcal{E}_t^{\text{DIAG}} &< \mathcal{E}_t^{\text{STATIC GEN}}.\end{aligned}\quad (35)$$

Equivalently, for the same fixed rollout budget MK , DIAG produces more effective preference pairs $M\rho_t$ and therefore more gradient signal per unit compute. This explains why DIAG achieves higher final performance without increasing either student rollout compute or the total teacher-generation budget.

C Empirical Bayes Derivation and Implementation Details

This appendix provides a brief derivation of the Empirical Bayes (EB) topic score used in Phase I and summarizes the corresponding implementation choices.

Beta–Binomial model. For each topic z , let p_z denote the latent probability that a sampled question from topic z yields a valid *natural* preference pair (Eq. 1). Given per-topic counts (C_z, N_z) (Eq. 5), we model

$$\begin{aligned}p_z &\sim \text{Beta}(\alpha_0, \beta_0), \\ C_z \mid (p_z, N_z) &\sim \text{Binomial}(N_z, p_z).\end{aligned}\quad (36)$$

By conjugacy, the posterior is

$$p_z \mid (C_z, N_z) \sim \text{Beta}(\alpha_0 + C_z, \beta_0 + N_z - C_z), \quad (37)$$

and the posterior mean yields the EB score

$$v_z \triangleq \mathbb{E}[p_z \mid C_z, N_z] = \frac{C_z + \alpha_0}{N_z + \alpha_0 + \beta_0}. \quad (38)$$

Empirical Bayes prior from global statistics.

We set the Beta prior using the global valid-pair rate $\bar{p}$ and a smoothing strength κ (prior sample size):

$$\alpha_0 \triangleq \kappa \bar{p}, \quad \beta_0 \triangleq \kappa (1 - \bar{p}). \quad (39)$$

Substituting into Eq. 38 gives the shrinkage estimator used in the main text:

$$v_z = \frac{C_z + \kappa \bar{p}}{N_z + \kappa}. \quad (40)$$

When $N_z \ll \kappa$, the score shrinks towards $\bar{p}$, preventing over-reaction to low-count noise; as N_z grows, the empirical rate C_z/N_z dominates.

Implementation details. We compute $\bar{p}$ as the running global average of $\mathbb{I}_{\text{valid}}(x)$ over all attempted questions up to the current iteration (equivalently, $\bar{p} = \sum_z C_z / \sum_z N_z$ using the same statistics as Eq. 5). We fix κ as a small constant in all experiments, which controls the strength of shrinkage in the early, noisy regime.

D Prompts

Prompt Templates. We present the prompt templates employed for the Diagnose-and-Generate loop in Figure 6 and Figure 5. For the teacher model, the instruction explicitly incorporates the *Mistake Context*, comprising the failed question and the erroneous solution trace, instructing the model to synthesize a variant that probes the same underlying misconception while altering surface features. Regarding student rollouts, we employ

Prompt for Student Model Answer Generation

```
<|im_start|>system
Please reason step by step, and put your final answer within \boxed{ }.
<|im_end|>
<|im_start|>user
{problem} Let's think step by step and output the final answer within \boxed{ }.
<|im_end|>
<|im_start|>assistant
```

Prompt for Teacher Model Answer Generation

```
<|im_start|>system
You are a helpful assistant.
<|im_end|>
<|im_start|>user
{question} Let's think step by step and output the final answer within \boxed{ }.
<|im_end|>
<|im_start|>assistant
```

Figure 5: Prompt for Student Model and Teacher Model Answer Generation.

a standard format to elicit reasoning. To harmonize the reasoning structure across heterogeneous model architectures, we adopted an adaptive decoding strategy: for models requiring explicit structural guidance, the generation was primed with a sequential prefix (e.g., “Step 1:” (Lai et al., 2024)) to initiate the deduction process; for models with strong intrinsic instruction adherence, standard generation was utilized.

E Implementation Details

We provide the detailed hyperparameters and system configurations used in our experiments to ensure reproducibility.

E.1 Training Configuration

All student models (Qwen2.5-Math-7B, Qwen3-8B-Base, Llama-3.1-8B) were trained using the DPO objective with the AdamW optimizer, a cosine learning rate schedule with peak learning rate 5×10^{-7} , no weight decay, global batch size 128 (gradient accumulation 16), and 2 epochs per iteration. For DPO, we set the KL penalty $\beta = 0.1$, use sigmoid loss with label smoothing 0.1, and freeze the previous iteration’s checkpoint as the reference policy. Sequences are truncated at 4096 tokens (prompts at 1000), and training uses bfloat16 precision with gradient checkpointing enabled.

Regarding data attributes, we employ a hierarchical taxonomy annotated by Gemini-2.5 (Comanici et al., 2025). This includes five primary domains: *Algebra*, *Geometry*, *Calculus*, *Discrete Mathematics*, and *Probability*. Each contains several fine-grained sub-topics (e.g., *Polynomials*, *Modular Arithmetic*), along with four difficulty levels ranging from *Elementary* to *Expert*. For mistake-conditioned generation, failure traces are sampled from the mistake buffer $\mathcal{H}_{t-1}$ via an uncertainty-weighted probability proportional to $1 - \max(\hat{p}, 1 - \hat{p})$ to prioritize boundary cases, which is standard in (Settles, 2009). The pacing factor λ in (3) is set to 0.5 by default.

E.2 Generation & Rollout Configuration

The quality of the feedback loop relies heavily on the decoding strategies employed during the rollout phase.

- **Student Rollouts:** For collecting preference data, we sample $K = 8$ responses per prompt with temperature $T = 1.0$, top- $p = 0.99$, and top- $k = 50$ to ensure diversity in reasoning paths.
- **Teacher Generation:** For the targeted curriculum synthesis, the teacher model uses temperature $T = 1.0$ with top- $p = 0.99$ for question generation. For answer generation, we use greedy decoding ($T = 0$) to maximize correctness.

Prompt for Teacher Model Question Generation

You are an expert math curriculum designer and a master question creator. Your task is to generate high-quality, targeted training questions for a language model ("StudentLM"). Every question must be rigorous, original, aligned with the requested topic and difficulty, and directly target the student's weaknesses. You need to ensure that the answer to the generated question is a number or a formula, rather than a sentence or a proof.

1. Context & Goal

- General Topic: {topic}
- Difficulty Level: {difficulty}

2. Student Model Error Samples

{Error Sample: [Question (with Gold Solution), Student Failure]}

3. Additional Guidance

{guidance}

4. Output Format

First, analyze the student's error in a <guidance> block: identify the core misconception in the student's reasoning. Then, design a new problem that specifically forces the student to confront this misconception. Output in this exact structure:

<guidance>

...brief analysis of the student's misconception...

</guidance>

<question>

...problem statement...

</question>

Do NOT include any solutions, answers, or extra commentary outside these two blocks.

Figure 6: Prompt for Teacher Model Question Generation.

- **Verifier:** The verifier adopts rule-based evaluation (Zhang et al., 2025) to determine correctness.

E.3 Cross-Task Evaluation Implementation Details

To ensure reproducibility, we adopt standard evaluation protocols for all benchmarks. For **BBH** and **MMLU**, we utilize deterministic greedy decoding ($T = 0$) with a single rollout. For **TruthfulQA**, we follow the standard MC1 and MC2 validation setup, calculating scores based on the log-probabilities of candidate answers. For **HumanEval**, we employ a sampling-based approach: we generate $k = 20$ solutions per problem with a temperature of $T = 0.1$ and report the average pass rate.

F Case Studies

F.1 Ignoring Interior Singularities in Improper Integrals

A persistent failure mode among small models is mechanically applying the Fundamental Theorem of Calculus without checking whether the integrand is continuous on the integration interval. In the seed question below, the student is asked to evaluate $\int_{-1}^1 1/x^2 dx$. The integrand has a singularity at $x = 0$, making the integral divergent, but the student simply computes the antiderivative $-1/x$, substitutes the endpoints, and reports -2 , which is a negative value for a non-negative integrand, an inconsistency that goes entirely unnoticed.

The follow-up shifts the singularity from $x = 0$

to $x = 2 \in (1, 4)$ by asking the student to evaluate $\int_1^4 1/(x - 2)^2 dx$. This defeats the shallow heuristic “watch out when $x = 0$ is in the interval” while preserving the same underlying misconception. The naïve FTC evaluation now yields $-3/2$, again negative for a strictly positive integrand, so even a minimal sign-check would catch the error. Students who have internalized the continuity prerequisite immediately output 0 (divergent); those who have not will consistently report $-3/2$, producing a clean binary split in rollout outcomes.

F.2 Base-Rate Neglect in Bayesian Inference

The second case targets a different kind of oversight: ignoring prior probabilities when computing a posterior. Two machines have different production shares (60% vs. 40%) and defect rates (5% vs. 3%). The student is asked for $P(A \mid \text{defective})$ and normalizes the likelihoods directly, arriving at $5/(5 + 3) = 5/8$. The correct answer is $5/7$, obtained by weighting each likelihood by its prior. Notably, the student’s self-check ($5/8 + 3/8 = 1$) is tautological, since any normalization sums to one, yet it gives a false sense of correctness.

To expose this misconception more sharply, the follow-up keeps the same defect rates but pushes the priors to an extreme: Machine A now produces only 1% of items while Machine B produces 99%. The likelihood-only shortcut yields the same $5/8 \approx 62.5\%$, but the correct Bayesian posterior drops to $0.0005/0.0302 \approx 1.66\%$, which is off by a factor of nearly $40\times$. The dramatic gap ensures that any model still ignoring the prior will produce an answer that is obviously wrong, rather than merely slightly off.

Student Answer (Incorrect) – Improper Integral

Step 1: Find the antiderivative of $\frac{1}{x^2}$. We know that $\frac{1}{x^2} = x^{-2}$, so the antiderivative is $\frac{x^{-1}}{-1} = -\frac{1}{x}$.

Step 2: Apply the Fundamental Theorem of Calculus. That is $\int_{-1}^1 \frac{1}{x^2} dx = \left[-\frac{1}{x}\right]_{-1}^1$.

Step 3: Evaluate at the bounds. Upper bound: $-\frac{1}{1} = -1$. Lower bound: $-\frac{1}{-1} = 1$.

Step 4: Compute the result.

$$\left[-\frac{1}{x}\right]_{-1}^1 = (-1) - (1) = -2.$$

The answer is .

Teacher-Generated Follow-up (Targeted) – Improper Integral

Evaluate the integral $\int_1^4 \frac{1}{(x-2)^2} dx$. If the integral diverges, enter 0.

Correct Answer:

Trap answer via direct FTC: $\left[-\frac{1}{x-2}\right]_1^4 = -\frac{1}{2} - 1 = -\frac{3}{2}$.

Student Answer (Incorrect) – Bayesian Inference

Step 1: Identify the given information. Machine A: defect rate = 5%. Machine B: defect rate = 3%. We are told a defective item was found, and we need $P(A \mid \text{defective})$.

Step 2: Since the item is already known to be defective, we need to determine which machine is more likely to have produced it. We can compare the defect rates directly. The probability of being defective given Machine A is 5%, and given Machine B is 3%.

Step 3: Normalize to get the posterior.

$$P(A \mid \text{defective}) = \frac{P(\text{defective} \mid A)}{P(\text{defective} \mid A) + P(\text{defective} \mid B)} = \frac{0.05}{0.05 + 0.03} = \frac{5}{8}.$$

Step 4: Verify. $\frac{5}{8} + \frac{3}{8} = 1$. The probabilities sum to 1, so the answer is consistent. The answer is

.

Teacher-Generated Follow-up (Targeted) – Bayesian Inference

A factory has two machines. Machine A produces 1% of all items and has a 5% defect rate. Machine B produces 99% of all items and has a 3% defect rate. If a randomly selected item is found to be defective, what is the probability it was produced by Machine A?

Correct Answer: $\frac{0.0005}{0.0302} \approx 1.66\%$.

Trap answer via likelihood-only normalization: $\frac{5}{5+3} = \frac{5}{8} = 62.5\%$.

Figure 7: **Case Studies.** *Top pair (Improper Integral):* The student applies FTC across a singularity at $x = 0$, arriving at -2 ; the follow-up shifts the singularity to $x = 2$, where the trap value $-3/2$ is negative for a strictly positive integrand. *Bottom pair (Bayesian Inference):* The student normalizes likelihoods without prior weighting ($5/8$ instead of $5/7$); the follow-up pushes priors to 1% vs. 99%, where the same shortcut gives 62.5% but the correct posterior is $\approx 1.66\%$.