STRONG YET LEARNABLE: RETHINKING WHAT REASONING DISTILLATION NEEDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning distillation transfers the reasoning capabilities of large language models to smaller students by supervising their intermediate decisions. However, stronger teachers do not necessarily produce better students, as increasingly sophisticated reasoning can become difficult for weaker students to absorb. On-policy self-distillation reduces this mismatch by using the same student as a privileged teacher with access to the verified outcome, referred to as *hindsight*. Yet its decisions rely on information unavailable to the deployed student and may therefore be difficult to reproduce causally. This raises a more fundamental question: what makes supervision worth learning? Useful supervision must provide knowledge beyond the student's current behavior while remaining learnable for the student. Crucially, these two requirements need not be satisfied by the same source. Hindsight provides a way to assess learnability: it gives the same student additional task information while keeping the model and reasoning state fixed, revealing whether a teacher signal remains compatible with the student under richer information. We therefore introduce HIND, an on-policy distillation framework that decouples knowledge provision from learnability assessment: a strong external teacher supplies new reasoning, while the student's hindsight response assesses its learnability and weights the corresponding supervision accordingly. Across multiple student scales and model families, HIND consistently improves mathematical reasoning and transfers to coding and knowledge-intensive reasoning, with gains that persist under larger sampling budgets and grow with teacher strength.

1 INTRODUCTION

Reasoning distillation has become a practical route for transferring the capabilities of large language models into smaller and more efficient models (Hinton et al., 2015; Hsieh et al., 2023; Shridhar et al., 2023; Mukherjee et al., 2023). Beyond learning from final answers, students can imitate dense token-level predictions over reasoning, allowing a teacher to communicate its intermediate decisions throughout a solution. Recent work has increasingly adopted on-policy distillation (OPD): the student generates its own reasoning trajectories, and the teacher provides supervision along those student-generated partial solutions (Agarwal et al., 2024; Ye et al., 2026). OPD thereby keeps teacher feedback aligned with the student's evolving generation distribution and reduces the mismatch between training and inference.

A natural way to strengthen distillation is to use a more capable teacher. Empirical evidence, however, shows that this relationship is not monotonic: stronger teachers do not necessarily produce better students, and smaller students can benefit more from reasoning matched to their own capacity (Zhang et al., 2025; Li et al., 2025). Recent approaches similarly find that the utility of reasoning supervision depends on its relationship with the learner, rather than on teacher quality alone (Yang et al., 2026). This leaves a central question: *what makes teacher supervision actually suitable for the student?*

On-policy self-distillation (OPSD) offers an elegant answer by changing how the teacher is constructed (Zhao et al., 2026; Hübotter et al., 2026; Kaur et al., 2026). Instead of relying on a separate larger model, OPSD lets the same student model serve as the teacher by giving it privileged information, such as a verified answer or reference solution. This privileged view is commonly referred to as *hindsight*. Because the teacher and student share the same underlying model, the resulting supervision remains closely aligned with the student's own policy (Shenfeld et al., 2026).

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

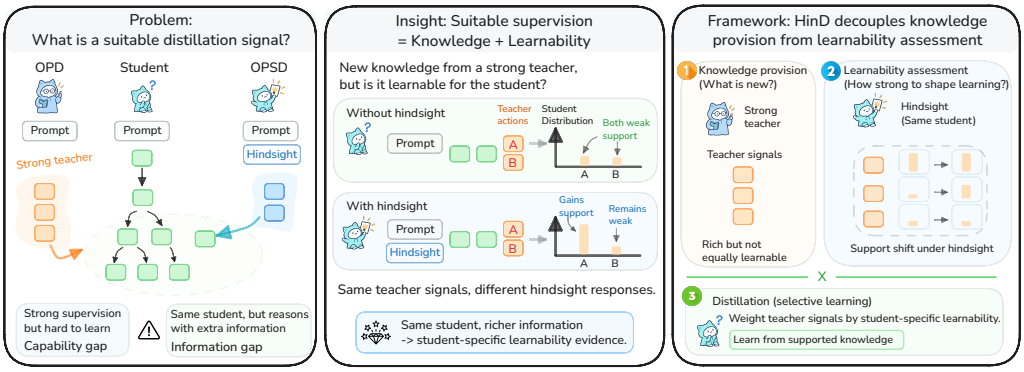

Figure 1: **From stronger supervision to suitable supervision. Problem:** A stronger teacher provides richer knowledge but may be difficult for a weaker student to learn from. OPSD reduces this capacity gap by using the same student as the teacher, but its privileged information creates an information gap that makes the resulting reasoning difficult to recover. **Insight:** Suitable supervision requires both new knowledge and learnability. Hindsight keeps the student and reasoning state fixed while changing only the available task information, providing student-specific evidence of whether a teacher signal is learnable. **Framework:** HIND decouples knowledge provision from learnability assessment: the teacher provides new reasoning, while the student's hindsight response assesses its learnability and determines how strongly each teacher signal shapes learning.

Self-distillation closes the capacity gap, but necessarily introduces an *information gap*. The privileged teacher makes decisions with information unavailable to the causal student, despite sharing the same underlying model. This distinction is especially consequential for autoregressive reasoning, where each decision is conditioned on a causal prefix that accumulates intermediate computations and information needed for later steps (Feng et al., 2023; Merrill & Sabharwal, 2024; Ton et al., 2025; Zhu et al., 2025). A privileged teacher enters each autoregressive decision with information that the deployed student would otherwise have to establish through its own reasoning. Its trajectories show how to proceed with this additional information, but not necessarily what the student can learn to recover without it. Consequently, matching teacher and student at the model level does not guarantee that the resulting supervision is suitable for the unprivileged student.

Yet this information gap also creates a unique opportunity. Hindsight changes the information available to the student while keeping the student model and reasoning state fixed, providing a controlled view of how the same student evaluates a teacher signal with more task information. This is particularly valuable because effective supervision must go beyond what the student already does, yet remain within what the student can plausibly absorb. If the same student assigns stronger support to teacher supervision under hindsight, that supervision is more consistent with the student's own reasoning. Hindsight therefore provides a student-specific assessment of whether externally supplied knowledge is learnable.

Building on this view, we introduce HIND, an on-policy distillation framework that decouples knowledge provision from learnability assessment. In this sense, the teacher determines *what is new*, while hindsight provides student-specific evidence of *what is learnable*. HIND operationalizes this distinction by measuring how teacher–student mismatch changes when the same student receives additional task information, and uses this signal to regulate the influence of teacher supervision at each on-policy state. This separation allows HIND to draw increasingly rich knowledge from stronger teachers while keeping transfer matched to what the student can learn, opening a path to exploit stronger teachers beyond what weaker students could directly imitate.

Our contributions are summarized as follows:

- **A new view of supervision suitability.** We characterize suitable supervision as requiring both new knowledge and student learnability, and show that hindsight provides a controlled way to assess the latter by changing only the information available to the same student.

- **A new OPD framework.** We propose HIND, which uses the teacher to provide new reasoning and the student's hindsight response to determine how strongly that reasoning should shape learning. This separation allows teacher capability to scale while learnability assessment remains student-specific, opening a path to exploit stronger teachers beyond the limits of direct imitation.
- **Comprehensive empirical validation.** Across student scales and model families, HIND consistently improves mathematical reasoning and transfers to coding and knowledge-intensive reasoning, and its gains persist under larger sampling budgets and grow with teacher strength.

2 Related Work

Reasoning distillation in LLMs. Reasoning distillation transfers teacher reasoning to smaller models through rationale supervision (Hsieh et al., 2023), problem decomposition (Shridhar et al., 2023), and progressive explanation learning (Mukherjee et al., 2023). Other methods enrich supervision with multiple reasoning chains (Ho et al., 2023; Li et al., 2023), while on-policy distillation reduces distribution mismatch through feedback on student-generated sequences (Agarwal et al., 2024). These advances improve where and how supervision is provided, but do not resolve which teacher signals are actually suitable for a given student.

Student-aware distillation. Student-aware distillation evaluates supervision relative to the learner. Capacity-gap studies show that larger teachers are not always better (Zhang et al., 2025), while mixing reasoning complexity or teacher sources can improve transfer (Li et al., 2025). RSR balances informativeness and alignment using token ranks and surprisal (Yang et al., 2026); Select to Think exploits local sufficiency to teach candidate selection (Ye et al., 2026); and LARK uses loss-reduction proxies to select trajectories balancing learnability and distributional coverage (Yu et al., 2026). The central challenge is therefore to identify supervision that introduces new knowledge while remaining learnable for the student.

On-policy self-distillation. OPSD constructs stronger supervision by giving the student privileged training-time information, including verified solutions (Zhao et al., 2026), textual feedback (Hübotter et al., 2026), demonstrations (Shenfeld et al., 2026), and privileged action information (Penaloza et al., 2026). This removes a large capacity gap but introduces an information gap: the teacher reasons with information unavailable to the deployed student, so its behavior may be difficult to recover from the student's causal view. Recent work shows that privileged self-distillation can impair long-horizon reasoning, verification, and self-correction, while privileged references do not always provide reliably useful information for learning (Kaur et al., 2026; Lim et al., 2026; Shrestha & Tessier, 2026). Thus, enriching what the teacher observes does not by itself ensure that the resulting supervision is suitable for a student that must reason without that information.

3 Method

Distillation faces a structural tension: supervision should extend beyond the student's current policy while remaining suitable for the student to absorb. OPD places teacher supervision on student-generated states, but stronger teacher signals can still be too unfamiliar for the student to learn effectively. OPSD reduces the capacity gap by using the student itself as a privileged teacher, yet its decisions rely on information unavailable at deployment and can therefore remain difficult for the student to recover from its reasoning prefix (Feng et al., 2023; Kim & Suzuki, 2025).

3.1 Observation: Hindsight reveals student-specific structure in external supervision

Strong teachers naturally introduce reasoning beyond the student's current policy, but such teacher–student disagreement can reflect very different relationships to the student itself. The key question is therefore not simply how far a teacher signal lies from the current student, but whether that signal remains connected to reasoning the same student can support. Can hindsight reveal this boundary from the student's own perspective?

For an on-policy prefix h_t, let $q_T(\cdot \mid h_t)$ denote the teacher distribution and $p_S(\cdot \mid h_t)$ the student distribution under the causal view, i.e., using only the information available in the reasoning prefix. We then evaluate the same student on the same state with the verified outcome y^* available, defining

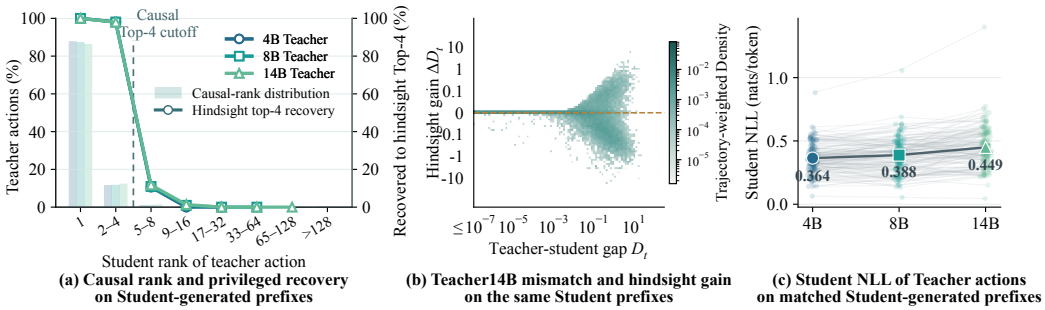

Figure 2: **Hindsight adds student-specific information beyond the causal student view.** On Qwen3-1.7B, (a) teacher actions outside the student's causal Top-4 can receive strong hindsight support; (b) teacher signals with similar causal mismatch can exhibit substantially different hindsight gains; and (c) student NLL of teacher actions on matched student-generated prefixes increases with teacher scale. Together, these results show that hindsight provides learnability information not captured by teacher–student mismatch alone, and that this information becomes increasingly relevant as teacher reasoning moves farther from the student distribution.

the corresponding hindsight view as $p_S(\cdot \mid h_t, y^*)$. Their teacher–student mismatches are

$$D_t = D_{\mathrm{KL}}(q_T(\cdot \mid h_t) \,\|\, p_S(\cdot \mid h_t)), \qquad D_t^* = D_{\mathrm{KL}}(q_T(\cdot \mid h_t) \,\|\, p_S(\cdot \mid h_t, y^*)). \tag{1}$$

Because the student, reasoning state, and teacher distribution remain fixed, their difference

$$\Delta D_t = D_t - D_t^* \tag{2}$$

directly measures how much more closely the same student supports the teacher distribution when additional task information is available.

We represent the teacher action at each prefix h_t by its greedy next-token prediction, $a_t^T = \arg\max_a q_T(a \mid h_t)$, and denote the rank of this action under the student's causal and hindsight views by

$$r_t = \mathrm{rank}_{p_S(\cdot|h_t)}(a_t^T), \qquad r_t^* = \mathrm{rank}_{p_S(\cdot|h_t, y^*)}(a_t^T), \tag{3}$$

where rank 1 is the most probable next token.

Hindsight reveals student support beyond current generation. Figure 2(a) groups teacher actions by their causal rank and reports how often each group enters the student's hindsight Top-4. A substantial fraction of actions outside the causal Top-4 move into the hindsight Top-4, showing that the same student can strongly support teacher reasoning that it would not ordinarily generate.

Hindsight distinguishes supervision with similar causal mismatch. Panel (b) places each reasoning state by its causal mismatch D_t and hindsight gain ΔD_t. Teacher actions with similar D_t can exhibit sharply different hindsight gains. Hindsight therefore provides additional student-specific information that distinguishes teacher supervision beyond its causal mismatch alone.

Learnability assessment becomes more important as teacher capability grows. Panel (c) reports the student's trajectory-level NLL of teacher actions evaluated on the same student-generated prefixes. The problem-balanced NLL rises from 0.364 to 0.388 and 0.449 nats/token from 4B to 8B and 14B. As teacher reasoning moves farther from the student's own distribution, deciding which parts remain learnable becomes increasingly important for effective distillation.

Together, these results establish hindsight as a student-specific source of learnability evidence for external supervision. This directly motivates HIND, which separates the two roles: the teacher provides new reasoning, while hindsight changes only the information available to the same student and thereby determines how strongly each teacher signal should shape learning.

3.2 HIND: STUDENT-AWARE DISTILLATION

When additional task information makes the same student more consistent with a teacher signal, that change provides student-specific evidence of its learnability. HIND uses this evidence to regulate

how strongly each teacher signal influences learning, while leaving the reasoning content entirely to the teacher. In this sense, the teacher determines *what is new*, and hindsight determines *what is learnable*.

Assessing learnability on student-generated states. HIND assesses learnability directly on states visited by the current student policy. At training round r, we freeze the current student S_{θ_r} and sample an on-policy trajectory $\tau \sim \pi_{\theta_r}(\cdot \mid x)$, where $h_t = (x, y_{<t})$ denotes the reasoning prefix at position t. We then evaluate the teacher, the causal student, and the hindsight view of the same frozen student on exactly this prefix, where the hindsight view differs only in receiving the verified final answer y^* as additional information.

Using the hindsight gain ΔD_t defined in Sec. 3.1, we quantify how much the same student's support for the teacher increases under hindsight. Equivalently,

$$\Delta D_t = \mathbb{E}_{a \sim q_T(\cdot \mid h_t)} \left[\log p_{\theta_r}(a \mid h_t, y^*) - \log p_{\theta_r}(a \mid h_t) \right], \tag{4}$$

which directly measures how the student's support for the teacher distribution changes under hindsight. A positive value therefore indicates that the teacher signal becomes more compatible with the same student when additional task information is available.

We convert this evidence into a bounded supervision weight,

$$g_t = [\Delta D_t]_+, \qquad w_t = \frac{g_t}{1 + g_t}. \tag{5}$$

The positive part assigns teacher influence only when hindsight provides affirmative support, while the bounded mapping prevents unusually large KL differences from dominating training. Importantly, we do not renormalize w_t: when few states receive such evidence, the total teacher influence decreases rather than being redistributed across unsupported states.

Distilling teacher knowledge with hindsight-assessed weights. For a batch of B on-policy trajectories $\{\tau_i\}_{i=1}^{B}$, let $\mathcal{R}_i$ denote the reasoning-token positions of trajectory i. The reasoning objective is

$$\mathcal{L}_R = \frac{1}{B} \sum_{i=1}^{B} \frac{1}{|\mathcal{R}_i|} \sum_{t \in \mathcal{R}_i} w_{it} D_{\mathrm{KL}} \left(q_T(\cdot \mid h_{it}) \,\|\, p_\theta(\cdot \mid h_{it}) \right). \tag{6}$$

where p_θ is the trainable student initialized from θ_r. The full teacher distribution remains the distillation target; hindsight changes only its position-wise influence.

To preserve supervision outside the reasoning span and stabilize policy drift, we include two auxiliary objectives. Let $\mathcal{G}_i$ denote all scored generated-token positions, $\mathcal{C}_i = \mathcal{G}_i \setminus \mathcal{R}_i$ the non-reasoning control positions, and $N_G = \sum_{i=1}^{B} |\mathcal{G}_i|$. We define

$$\mathcal{L}_C = \frac{1}{N_G} \sum_{i=1}^{B} \sum_{t \in \mathcal{C}_i} D_{\mathrm{KL}} \left(q_T(\cdot \mid h_{it}) \| p_\theta(\cdot \mid h_{it}) \right), \mathcal{L}_{\mathrm{ref}} = \frac{1}{N_G} \sum_{i=1}^{B} \sum_{t \in \mathcal{G}_i} D_{\mathrm{KL}} \left(p_\theta(\cdot \mid h_{it}) \| p_{\mathrm{ref}}(\cdot \mid h_{it}) \right). \tag{7}$$

The complete objective is

$$\mathcal{L} = \mathcal{L}_R + \lambda_C \mathcal{L}_C + \lambda_{\mathrm{ref}} \mathcal{L}_{\mathrm{ref}}, \tag{8}$$

where λ_C controls teacher supervision on non-reasoning positions and λ_{ref} controls regularization toward the frozen initial student. After one optimization pass, we refresh the student snapshot and generate new trajectories for the next round.

As training proceeds, both the visited states and their hindsight-based learnability signals are refreshed with the evolving student policy, keeping supervision grounded in the current learner throughout training.

4 EXPERIMENTS

To empirically evaluate HinD, we conduct a systematic analysis of its underlying premise and distillation effectiveness, guided by three research questions:

- **RQ1 (Overall effectiveness).** Does HIND improve on-policy distillation across student scales and reasoning domains?
- **RQ2 (Hindsight-based learnability).** Does hindsight identify more useful teacher supervision?
- **RQ3 (Training dynamics and efficiency).** How does HIND improve with additional budget?

4.1 SETUP

We outline the main experimental settings below, with data construction, optimization details and baseline implementations provided in Appendix B.

Models. We evaluate HIND across two model families and multiple student scales. For Qwen, we use Qwen3-1.7B and Qwen3-4B as students with Qwen3-32B as the teacher (Yang et al., 2025). For the Llama-based setting, we use NVIDIA Llama-3.1-Nemotron-Nano-4B-v1.1 as the student and NVIDIA Llama-3.3-Nemotron-Super-49B-v1 as the teacher (Bercovich et al., 2025).

Training data and evaluation. We use a fixed pool of $2{,}048$ unique problems selected from DAPO-Math-17K (Yu et al., 2025). For the main 200-update protocol, each update processes 16 prompts and samples one fresh student trajectory per prompt, yielding $3{,}200$ prompt exposures and $3{,}200$ on-policy trajectories in total. The pool is traversed with deterministic epoch-wise shuffling, so revisited problems receive newly generated trajectories from the current student policy. We evaluate mathematical reasoning on MATH-500 (Lightman et al., 2024), OlympiadBench (He et al., 2024), and AIME25 (Zhang & Math-AI, 2025), and cross-domain generalization on LiveCodeBench (Jain et al., 2025), MMLU-Pro (Wang et al., 2024), and GPQA Diamond (Rein et al., 2023).

On-policy training. At the beginning of each round, the current student snapshot generates the training trajectories. The teacher and the hindsight view of the same frozen student are then evaluated on exactly the same sampled prefixes. Hindsight receives only the verified final answer as additional information. After one optimization pass, the student snapshot is refreshed and new trajectories are generated, so both the training-state distribution and the HIND signal evolve with the policy. The teacher, hindsight view, and frozen reference are used only during training; inference requires only the student, with no auxiliary model calls.

Baselines. We compare against the pretrained BASE student, standard teacher-based on-policy distillation (OPD), privileged on-policy self-distillation (OPSD) (Zhao et al., 2026), and reward-based reinforcement learning with GRPO (Shao et al., 2024). Where budget matching is claimed, we match the number of sampled student trajectories and the maximum generation-token budget.

4.2 RQ1: OVERALL EFFECTIVENESS

Reasoning performance. Table 1 shows that HIND improves reasoning across student scales and model families. For both Qwen3 students, HIND achieves the strongest average mathematical performance, with substantial gains retained as the student grows from 1.7B to 4B. The improvement is especially pronounced on the more challenging AIME25 benchmark, where the 4B student rises from 58.8 to 70.0. On the Nemotron family, HIND likewise improves the Base student across several reasoning and generalization benchmarks, indicating that the benefit is not specific to a single architecture.

Generalization beyond the training domain. The gains extend beyond mathematical reasoning. Across LiveCodeBench, MMLU-Pro, and GPQA-D, HIND improves aggregate performance across the evaluated student settings, with gains spanning coding, broad knowledge, and knowledge-intensive reasoning. This indicates that the learned improvements transfer across heterogeneous reasoning domains.

Broader reasoning coverage under sampling. Table 2 evaluates whether the improvement extends beyond single-sample accuracy to a broader set of reasoning trajectories. Relative to the Base student, HIND retains gains from Pass@1 through Pass@16 on both OlympiadBench and GPQA-D, and achieves the strongest results at the larger Pass@k values. The persistence of the gain under larger sampling budgets suggests that HIND does more than promote a single preferred solution: it increases the availability of successful reasoning trajectories within the student's broader policy distribution.

Table 1: **Overall reasoning performance.** HIND consistently improves aggregate reasoning performance across student settings, with strong gains in mathematical reasoning and broad transfer beyond the training domain. Results are averaged over three decoding seeds.

Method	Math Reasoning				OOD Generalization		
	MATH-500	Olympiad	AIME25	Avg.	LiveCodeBench	MMLU-Pro	GPQA-D
Qwen3-1.7B							
Base	86.9	63.3	40.0	63.4	30.9	62.3	38.4
GRPO	88.2	65.6	40.0	64.6	29.7	**67.6**	40.4
OPD	87.9	65.2	36.7	63.3	31.8	63.7	**44.4**
OPSD	86.7	64.4	33.3	61.5	31.5	63.9	42.4
HinD	**92.2**$_{\pm 0.7}$	**67.1**$_{\pm 1.3}$	**43.3**$_{\pm 1.6}$	**67.5**	**32.0**$_{\pm 0.3}$	66.8$_{\pm 0.2}$	44.0$_{\pm 1.1}$
Qwen3-4B							
Base	93.5	74.3	58.8	75.5	46.3	70.8	53.6
GRPO	91.7	76.9	66.7	78.4	49.1	68.3	55.6
OPD	92.6	**77.5**	56.7	75.6	49.1	69.6	57.1
OPSD	**93.8**	75.2	60.0	76.3	48.0	72.1	53.5
HinD	**93.8**$_{\pm 0.7}$	77.1$_{\pm 1.0}$	**70.0**$_{\pm 0.0}$	**80.3**	**50.6**$_{\pm 1.2}$	**72.9**$_{\pm 0.0}$	**57.6**$_{\pm 0.8}$
Llama-3.1-Nemotron-Nano-4B							
Base	91.6	74.2	60.0	75.3	43.4	56.5	48.0
GRPO	92.8	74.6	56.7	74.7	**48.6**	54.3	**55.1**
OPD	91.2	74.6	53.3	73.0	42.9	55.5	51.5
OPSD	91.6	**74.8**	**63.3**	76.6	47.4	56.1	53.5
HinD	**93.2**$_{\pm 0.7}$	74.3$_{\pm 0.5}$	**63.3**$_{\pm 0.0}$	**76.9**	47.4$_{\pm 0.6}$	**59.8**$_{\pm 1.0}$	54.5$_{\pm 2.3}$

Table 2: **HinD improves the broader reasoning distribution, not only the top sampled response.** Shown on OlympiadBench and GPQA-D using Qwen3-1.7B

Method	OlympiadBench				GPQA-D			
	Pass@1	Pass@4	Pass@8	Pass@16	Pass@1	Pass@4	Pass@8	Pass@16
Base	63.3	76.3	80.2	83.4	38.4	58.3	66.2	72.7
OPD	65.2	77.0	80.6	83.8	**44.4**	59.3	67.0	72.7
OPSD	64.4	77.4	81.1	84.0	42.4	58.6	65.7	71.7
HinD	**67.1**	**78.6**	**81.5**	**84.9**	44.0	**61.0**	**69.2**	**75.8**

4.3 RQ2: HINDSIGHT-BASED LEARNABILITY

RQ2.1: Does the placement of HIND supervision matter? A defining feature of HIND is that teacher supervision is weighted differently across reasoning states. We isolate the value of this placement with matched controls that preserve the overall strength of HIND supervision while changing where it is applied. BASE receives no distillation update, while OPD applies the full teacher signal uniformly. UNIFORM-MATCHED preserves each trajectory's total HIND weight budget but distributes it uniformly; SHUFFLED-MATCHED preserves the exact weight multiset but permutes its positions; and CAUSAL-MATCHED assigns the same weights according to causal mismatch D_t. HIND retains the hindsight-based placement induced by ΔD_t. Together, these controls isolate supervision amount, weight distribution, and state-level placement.

Table 3 shows a clear hierarchy across these allocation rules. The five-benchmark macro averages MATH-500, OlympiadBench, AIME25, MMLU-Pro, and GPQA-D; the OOD average includes MMLU-Pro and GPQA-D. HIND performs best in all reported columns, showing that the gains depend not only on how much teacher supervision is applied, but on where it is emphasized. The matched controls rule out supervision budget, weight distribution, and causal mismatch alone as sufficient explanations, pointing instead to the state-level allocation induced by hindsight. Appendix C.2

Table 3: **Hindsight-based state assignment matters under matched supervision budgets.** Results on Qwen3-1.7B compare HIND with controls that progressively preserve its total weight mass and weight distribution while changing how those weights are assigned to reasoning states. *Weight Mass* denotes the total reasoning weight, *Weight Dist.* its within-trajectory distribution, and *State Assignment* the rule used to place those weights.

Method	Weight Mass	Weight Dist.	State Assignment	MATH	Olympiad	AIME	OOD Avg.
Base	–	–	–	86.9	63.3	40.0	50.4
OPD	Full	Uniform	All	87.9	65.2	36.7	54.1
Uniform-Matched	HinD	Uniform	All	88.0	64.5	36.7	50.1
Shuffled-Matched	HinD	HinD	Shuffled	87.4	65.6	30.0	48.1
Causal-Matched	HinD	HinD	D_t	88.0	64.8	36.7	50.5
HIND	HinD	HinD	HinD	92.2	67.1	43.3	55.4

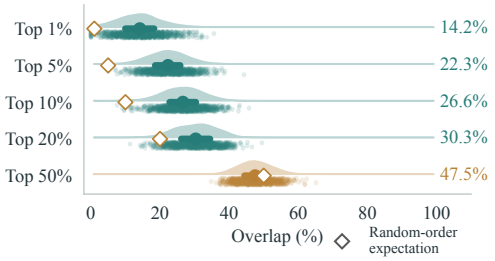

Figure 3: **Hindsight reorganizes teacher-signal priority beyond causal mismatch.** On Qwen3-1.7B, we compare the Top-α reasoning positions ranked by causal teacher–student mismatch D_t and by the HIND weight w_t. The overlap remains partial across cutoffs, showing that hindsight substantially changes which teacher signals receive the greatest influence.

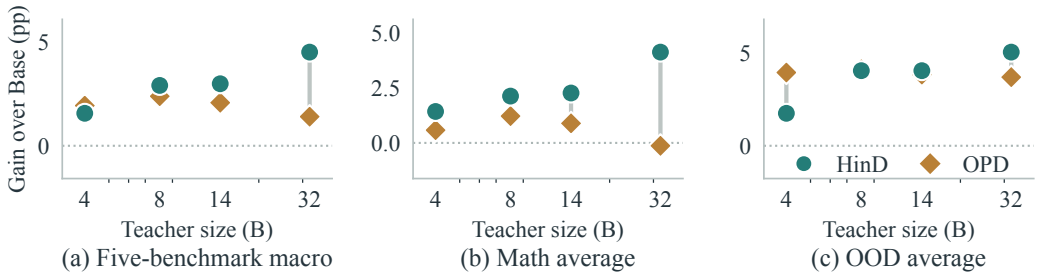

Figure 4: **HIND converts increasing teacher capability into larger student gains.** For a fixed Qwen3-1.7B student, we vary the Qwen3 teacher from 4B to 32B and report improvement over the Base student on the five-benchmark macro average, mathematical reasoning, and OOD evaluation. HIND shows a substantially stronger scaling trend than standard OPD.

provides two complementary views: Figure A1 visualizes how the allocation rules redistribute the same supervision budget along a fixed trajectory, while Figure A2 compares their induced priority rankings across all rollouts.

RQ2.2: Does hindsight add information beyond causal teacher–student mismatch? We compare HIND priorities with causal teacher–student mismatch on Qwen3-1.7B. Within each on-policy trajectory, we rank reasoning positions by D_t and by the HIND weight w_t, then compare the overlap between their Top-α positions. Figure 3 shows that the two signals are related but induce different priorities: at Top-10%, only 26.6% of positions are shared. Thus, hindsight contributes student-specific learnability information beyond causal mismatch and changes which teacher signals receive the greatest influence during distillation. Appendix C.3 provides a fuller characterization. Figure A3 examines hindsight gain across causal-mismatch quantiles, while Figure A4 compares the full within-trajectory rankings induced by D_t and w_t. Both views show that hindsight systematically reshapes teacher-signal prioritization beyond what causal mismatch alone captures.

RQ2.3: How does HIND behave as teacher strength increases? To evaluate how HIND translates increasing teacher capability into student improvement, we train the same Qwen3-1.7B student with Qwen3 teachers ranging from 4B to 32B. Figure 4 shows that HIND's gains increase overall with

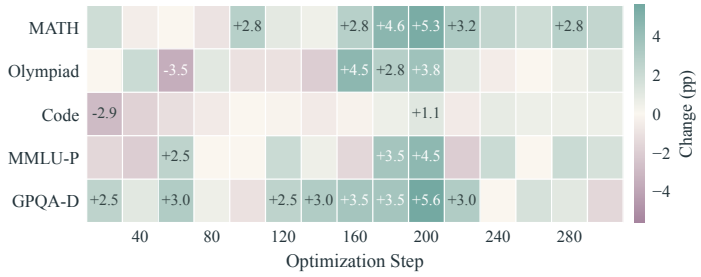

Figure 5: **HIND progressively expands its gains across tasks.** Cells report accuracy change relative to the Base model. Improvements become broadly positive across math and OOD tasks.

teacher strength across the five-benchmark average, with similar trends in mathematical reasoning and OOD evaluation. The 32B teacher yields the largest aggregate improvement, whereas standard OPD shows a substantially weaker scaling trend. In addition, Table A2 provides the full benchmark breakdown, while Figure A5 further analyzes the evolution of the training signal across teacher scales; see Appendix C.4 for details.

This behavior highlights a central benefit of HIND: teacher capability can scale while learnability remains student-specific, enabling increasingly strong teachers to contribute useful new reasoning.

4.4 RQ3: TRAINING DYNAMICS AND EFFICIENCY

RQ3.1: How does HIND scale with on-policy supervision budget? To study longer-horizon training dynamics, we extend training to 300 updates and track performance throughout the run. Figure 5 reports accuracy changes relative to the Base model across five evaluation benchmarks. As training progresses, HIND develops broader gains across both mathematical and out-of-domain reasoning tasks. Appendix C.5 analyzes this progression in more detail on GPQA-Diamond. Figure A6 shows how the solved set expands beyond the Base model, and Figure A7 relates accuracy to generation length throughout training. Together, these views show that the gains emerge by progressively solving additional questions while remaining in a comparable inference-length regime.

RQ3.2: Does HIND make better use of inference length? We examine how reasoning outcomes evolve as more inference tokens become available. For each sampled response, we rescore the same trajectory at 2k, 4k, 8k, 16k, and 32k tokens, recording whether reasoning is still ongoing, terminates without a valid answer, yields an incorrect answer, or reaches the correct solution. Figure A8 reports this progression on the mathematical evaluation set. At most budgets, and particularly at longer budgets, HIND converts a larger fraction of the available budget into completed solutions while maintaining higher overall accuracy. The gap becomes largest at longer budgets: accuracy reaches 91.3% and 92.2% at 16k and 32k, compared with 85.1% and 86.9% for the Base model. The response-state decomposition further shows a faster transition from ongoing reasoning to completed answers. Appendix C.6 extends the same analysis to GPQA-Diamond. There, Figure A9 shows an even stronger long-budget effect, with the advantage of HIND increasing to more than five percentage points at 16k and 32k. These results indicate that HIND improves how effectively additional inference length is used, rather than only shifting performance at a single decoding budget.

5 CONCLUSION

We study a central difficulty in reasoning distillation: supervision should extend beyond what the student already knows without becoming too difficult for it to learn. We realize that these two demands need not be satisfied by the same source. A capable teacher can contribute genuinely new reasoning, while hindsight provides a student-specific view of whether that reasoning remains learnable by evaluating the same student with additional task information. This distinction enables HIND to benefit increasingly from stronger teachers while preserving a learning signal tied to the student's own capabilities. Across model scales and families, the resulting gains extend from mathematical reasoning to coding and knowledge-intensive tasks, persist under larger sampling budgets, and strengthen as teacher capability grows. More broadly, our findings suggest that the effectiveness of distillation is governed not only by the quality of available knowledge, but by whether that knowledge can be translated into a learnable signal for the student.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce our experiments throughout the main paper and appendix. Section 3.2 specifies the HIND objective and training procedure, while Appendix B documents the model configurations, training data construction, optimization settings, baseline implementation and evaluation protocols. Additional appendices report the detailed analysis settings and ablation procedures used in the main experiments. We will provide the accompanying implementation and experiment configurations as supplementary material.

AI USE STATEMENT

Generative AI tools were used extensively throughout the development of this work, except for the research idea itself. They assisted with literature review, experimental planning, code generation and debugging, implementation of training and evaluation pipelines, data analysis, and manuscript editing. AI-generated code and analyses were reviewed, tested, and verified by the authors before being incorporated into the final work. All scientific decisions, experimental conclusions, and claims presented in the paper were ultimately made and validated by the authors, who take full responsibility for the content of this work.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*, 2025.

Phuong Tuan Dat, Qi Li, and Xinchao Wang. dopsd: On-policy self-distillation for diffusion language models. *arXiv preprint arXiv:2607.04428*, 2026.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in neural information processing systems*, 36:70757–70798, 2023.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 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)*, pp. 3828–3850, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 14852–14882, 2023.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the association for computational linguistics: ACL 2023*, pp. 8003–8017, 2023.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large

language models for code. In *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of llms? *arXiv preprint arXiv:2603.24472*, 2026.

Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. In *International Conference on Learning Representations*, volume 2025, pp. 44642–44668, 2025.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2665–2679, 2023.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25366–25394, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

Hyeongsoo Lim, Jinyoung Kim, Eun Seo Seo, Min Ho Jang, and Ji Won Yoon. Selfusion: Self-distillation for diffusion language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22077–22089, 2026.

Yifu Luo, Zeyu Chen, Haoyu Wang, Xinhao Hu, Yuxuan Zhang, Zhizhou Sha, and Shiwei Liu. Learning from the self-future: On-policy self-distillation for dllms. *arXiv preprint arXiv:2606.18195*, 2026.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *International Conference on Learning Representations*, volume 2024, pp. 7690–7706, 2024.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Samyak Shrestha and Alexander Tessier. Rethinking privileged information in on-policy self-distillation. *arXiv preprint arXiv:2608.18271*, 2026.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7059–7073, 2023.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in LLMs through information theory. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=IjOWms0hrf.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yuming Yang, Mingyoung Lai, Wanxu Zhao, Xiaoran Fan, Zhiheng Xi, Mingqi Wu, Chiyue Huang, Jun Zhao, Haijun Lv, Jian Tong, et al. Which reasoning trajectories teach students to reason better? a simple metric of informative alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42123–42150, 2026.

Wenxuan Ye, Yangyang Zhang, Xueli An, Georg Carle, and Yunpu Ma. Select to think: Unlocking slm potential with local sufficiency. *arXiv preprint arXiv:2604.26940*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2025.

Tianrun Yu, Kaixiang Zhao, Chih-Chun Chen, Amanda Hughes, Taylor W Killian, Fenglong Ma, Weitong Zhang, and Porter Jenkins. Lark: Learnability-grounded trajectory selection for efficient reasoning distillation. *arXiv preprint arXiv:2605.30651*, 2026.

Chen Zhang, Qiuchi Li, Dawei Song, Zheyu Ye, Yan Gao, and Yao Hu. Towards the law of capacity gap in distilling language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22504–22528, 2025.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Fangwei Zhu, Peiyi Wang, and Zhifang Sui. Chain-of-thought tokens are computer program variables. *arXiv preprint arXiv:2505.04955*, 2025.

A Comprehensive Related Works

Reasoning distillation in LLMs. Reasoning distillation transfers reasoning capabilities to smaller students through teacher-generated explanations and solution trajectories. Prior work has incorporated explanations as an additional source of supervision (Hsieh et al., 2023), organized intermediate reasoning through problem decomposition (Shridhar et al., 2023), and used rich explanation trajectories for progressive learning (Mukherjee et al., 2023). Other studies increase supervision diversity by sampling multiple teacher-generated reasoning chains for the same problem (Ho et al., 2023; Li et al., 2023). Building on these approaches, on-policy distillation provides teacher feedback directly on student-generated sequences, reducing the distribution mismatch between training and inference (Agarwal et al., 2024). However, placing supervision on the student's own trajectories does not by itself ensure that increasingly capable teacher reasoning remains learnable for the student.

Student-aware supervision. Student-aware distillation evaluates supervision relative to the learner rather than relying on teacher quality alone. Studies of teacher–student capacity gaps show that the largest teacher is not always the most effective one for a given student (Zhang et al., 2025). For reasoning distillation in smaller models, combining reasoning data of different complexity levels or from different teacher sources can improve transfer over using a single supervision source (Li et al., 2025). More recent methods make this dependence explicit: RSR combines student token rankings with surprisal to balance informativeness and alignment (Yang et al., 2026); Select to Think leverages local sufficiency within the student's candidate set to convert teacher guidance into candidate selection (Ye et al., 2026); and LARK selects trajectories using learning-loss reduction and proxy signals to balance learnability and distributional coverage (Yu et al., 2026). Together, these studies shift the focus from teacher quality alone to how supervision interacts with the learner.

On-policy self-distillation. On-policy self-distillation constructs a teacher from the student itself while providing additional training-time information. Prior work has used verified solutions to supervise the student's own trajectories (Zhao et al., 2026), converted textual environmental feedback into dense supervision (Hübotter et al., 2026), and conditioned teachers on demonstrations for continual learning (Shenfeld et al., 2026). Related work has also used privileged action information for capability transfer in agents (Penaloza et al., 2026). Similar ideas appear in diffusion language models, where self-generated answers serve as suffix conditions for step-level supervision (Luo et al., 2026), later denoising states provide more informative views of the same model (Dat et al., 2026), and correctness signals determine the direction of distillation across masking views (Lim et al., 2026).

These methods show that additional information can strengthen self-distillation, but they also expose an information gap: although teacher and student may share the same underlying model, the teacher can make decisions using information unavailable to the deployed student. This difference can make privileged reasoning difficult to recover from the student's causal context. Recent studies further report that richer privileged information does not always translate into better learning outcomes and, in some settings, can suppress uncertainty expression or degrade long-horizon reasoning and self-correction (Lim et al., 2026; Kim et al., 2026; Kaur et al., 2026). Existing OPSD methods therefore primarily use privileged information to construct stronger supervision, while the question of whether that supervision is learnable for the student remains separate.

HInD builds on this distinction by separating knowledge provision from learnability assessment: the teacher provides new reasoning, while the student's hindsight response determines how strongly that reasoning should shape learning.

B Experimental Setup

B.1 Models, training data, and budget

Models. We evaluate HInD with thinking-enabled language models from two model families. For the Qwen setting, we use Qwen3-1.7B and Qwen3-4B as students and Qwen3-32B as the teacher. For the Llama-based setting, we use NVIDIA Llama-3.1-Nemotron-Nano-4B-v1.1 as the student and NVIDIA Llama-3.3-Nemotron-Super-49B-v1 as the teacher. All runs start from the original student parameters and use full-parameter adaptation with a fresh optimizer. The teacher remains frozen throughout training. We additionally keep a frozen copy of the initial student for policy regularization,

Table A1: Training configuration for the three main student settings.

	Qwen3-1.7B	Qwen3-4B	Nemotron-4B
Teacher	Qwen3-32B	Qwen3-32B	Nemotron-49B
Max. rollout tokens	8,192	16,384	16,384
Prompts / update	16	16	16
Rollouts / prompt	1	1	1
Learning rate	10^{-6}	10^{-6}	10^{-6}

while the hindsight view always uses the student snapshot at the beginning of the current training round. The hindsight model is therefore refreshed after every update and is not a separately trained teacher.

All models are run in their native thinking mode, producing an explicit reasoning process before the final answer. For Qwen models, we enable this behavior through the model's chat template by setting `enable_thinking=True`, which allows the model to generate its native `<think>...</think>` reasoning block. For Nemotron, we enable detailed reasoning through the system message `detailed thinking on`. These settings follow each model's built-in prompting convention and do not introduce any additional training mechanism.

Training data. We construct a fixed training pool of $2{,}048$ unique problems from DAPO-Math-17K using a deterministic seed-42 selection procedure. The pool is disjoint from the held-out DAPO split by question ID and normalized question text. Before pool construction, exact normalized-text overlaps identified against the evaluation datasets are removed. This filtering targets exact overlap and should not be interpreted as semantic or near-duplicate decontamination. The selected training pool remains fixed throughout training.

Training budget. Our main reporting protocol uses 200 optimizer updates. Each HIND update processes 16 prompts and samples one fresh student trajectory per prompt, yielding $3{,}200$ prompt exposures and $3{,}200$ on-policy trajectories. Prompts are traversed using deterministic epoch-wise shuffling and reshuffled after the pool is exhausted. Consequently, the 200-update budget covers all $2{,}048$ unique training problems and revisits $1{,}152$ of them. Each revisit receives a newly generated trajectory from the current student policy. The resulting exposure corresponds to 1.5625 passes over the fixed problem pool.

The maximum number of newly generated training tokens is $8{,}192$ for Qwen3-1.7B and $16{,}384$ for Qwen3-4B and Nemotron-Nano-4B. Student rollouts use temperature 0.6, top-$p = 0.95$, and top-$k = 20$. These sampling parameters affect trajectory generation only; all distillation KL terms are computed over the full vocabulary with softmax temperature 1.

B.2 HIND TRAINING AND IMPLEMENTATION

On-policy round construction. At the beginning of training round r, let S_r denote the current student snapshot. For each training problem x_i, we sample a reasoning trajectory y_i from S_r without providing the verified answer. The sampled token sequence is held fixed throughout the round.

At each eligible reasoning position t, we evaluate three distributions on exactly the same sampled prefix:

$$p_{C,it} = S_r(\cdot \mid x_i, y_{i,<t}), \tag{9}$$

$$p^*_{C,it} = S_r(\cdot \mid x_i, a^*_i, y_{i,<t}), \tag{10}$$

$$q_{T,it} = T(\cdot \mid x_i, y_{i,<t}), \tag{11}$$

where a^*_i denotes the verified final answer. The hindsight view uses exactly the same student parameters and the same sampled prefix as the causal view, but receives a^*_i as additional context. It is teacher-forced on the original student trajectory and does not generate a separate privileged trajectory. Thus, the causal and hindsight views differ only in the task information available to the same frozen student.

Hindsight prompt. The hindsight view receives only the verified final answer rather than a gold reasoning trace. We append the following information to the final user message:

> For this reasoning task, the verified final answer is provided below as additional context. Continue solving the original problem step by step.
> <verified_final_answer>
> {answer}
> </verified_final_answer>

Neither the causal student nor the teacher receives this suffix.

Hindsight-weighted reasoning objective. For each reasoning position, we compute

$$D_{it} = D_{\mathrm{KL}}\left(q_{T,it} \| p_{C,it}\right), \qquad D^*_{it} = D_{\mathrm{KL}}\left(q_{T,it} \| p^*_{C,it}\right), \tag{12}$$

$$\Delta D_{it} = D_{it} - D^*_{it}, \qquad w_{it} = \frac{[\Delta D_{it}]_+}{1 + [\Delta D_{it}]_+}. \tag{13}$$

The resulting reasoning objective is

$$\mathcal{L}_R = \frac{1}{B} \sum_{i=1}^{B} \frac{1}{|\mathcal{R}_i|} \sum_{t \in \mathcal{R}_i} w_{it}\, D_{\mathrm{KL}}\left(q_{T,it} \| p_{\theta,it}\right). \tag{14}$$

where $\mathcal{R}_i$ contains the reasoning-token positions of rollout i and $p_{\theta,it}$ denotes the trainable student. The loss is first averaged over reasoning positions within each rollout and then over prompts. We do not normalize by the sum or empirical mean of the hindsight weights. Consequently, when learnability evidence is weak or sparse, the overall reasoning-update magnitude decreases rather than being redistributed across other positions.

Control and reference objectives. We further preserve teacher supervision on non-reasoning generated tokens and regularize the updated student toward its initialization. Let $\mathcal{G}_i$ denote all scored generated-token positions, $\mathcal{C}_i = \mathcal{G}_i \setminus \mathcal{R}_i$ the non-reasoning control positions, and

$$N_G = \sum_{i=1}^{B} |\mathcal{G}_i|. \tag{15}$$

The two objectives are

$$\mathcal{L}_C = \frac{1}{N_G} \sum_{i=1}^{B} \sum_{t \in \mathcal{C}_i} D_{\mathrm{KL}}\left(q_{T,it} \| p_{\theta,it}\right), \qquad \mathcal{L}_{\mathrm{ref}} = \frac{1}{N_G} \sum_{i=1}^{B} \sum_{t \in \mathcal{G}_i} D_{\mathrm{KL}}\left(p_{\theta,it} \| p_{\mathrm{ref},it}\right). \tag{16}$$

The complete objective is

$$\mathcal{L} = \mathcal{L}_R + 0.5\, \mathcal{L}_C + 0.1\, \mathcal{L}_{\mathrm{ref}}. \tag{17}$$

Here, $\mathcal{L}_C$ is a teacher forward-KL objective over the non-reasoning generated region, including control, final-answer, and end-of-sequence positions, while $\mathcal{L}_{\mathrm{ref}}$ is a reverse-KL regularizer to the frozen initial student. Both terms use the total number of scored generated tokens as their denominator.

Optimization. We use AdamW with a constant learning rate of 10^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, zero weight decay, and global gradient clipping at 1.0. In Eq. 8, we set $\lambda_C = 0.5$ and $\lambda_{\mathrm{ref}} = 0.1$. All student parameters are updated; we do not use LoRA. Training uses FP32 master parameters with BF16 computation, activation checkpointing, and chunked vocabulary projection. The latter reduces memory use while preserving exact full-vocabulary KL computation. After one optimization pass over the current rollout batch, the student snapshot is refreshed and new trajectories are sampled for the next round.

B.3 Baseline implementation

OPD and OPSD. For the main comparisons, OPD and OPSD use the same student initialization, training pool, 200-update budget, prompts per update, rollout count, generation limits, optimizer settings, and evaluation protocol as HIND. They differ only in how reasoning supervision is constructed. OPD applies the teacher distribution directly, while OPSD uses the privileged student as the distillation teacher. All other training components, including the control and reference regularization terms, are kept unchanged.

Allocation controls. The matched allocation controls preserve different components of the HIND weighting scheme. UNIFORM-MATCHED preserves each rollout's total hindsight weight but distributes it uniformly across reasoning positions. SHUFFLED-MATCHED preserves the exact multiset of hindsight weights while permuting their positions within the rollout. CAUSAL-MATCHED preserves the same weight multiset but assigns larger weights to positions with larger causal teacher–student KL. These controls isolate the effects of supervision amount, weight distribution, and state-level placement.

GRPO implementation. We implement GRPO using TRL 0.23.1 and DeepSpeed ZeRO-2. Each update samples two prompts with eight completions per prompt. The reward is binary verified-final-answer correctness, without a format bonus or correctness-based filtering. Advantages are normalized within each eight-completion group using the sample standard deviation plus 10^{-4}. We use a sequence-averaged GRPO objective, a clipping threshold of 0.2, one optimization iteration per rollout batch, and a sampled-reference-KL coefficient of 0.04 against the frozen initial student. Length-capped completions are retained.

Budget matching. The main HIND, OPD, and OPSD comparisons use identical training budgets and optimization settings. For GRPO, we match the number of optimizer updates, sampled student trajectories, and maximum generation-token budget. Over 200 updates, both GRPO and HIND sample $3{,}200$ student trajectories with the same token per-trajectory generation limit.

B.4 EVALUATION PROTOCOL

Generation protocol. Standard accuracy evaluation uses one completion per question with batched vLLM inference, BF16 weights, temperature 0.6, top-$p = 0.95$, and top-$k = 20$. The maximum generation budget is $32{,}768$ tokens. Given a context limit of $40{,}960$ and a 128-token safety margin, the generation limit for question i is

$$M_i = \min\left(32{,}768,\ 40{,}960 - L_{\mathrm{prompt},i} - 128\right). \tag{18}$$

Input questions are never truncated. Math and multiple-choice benchmarks are scored using a frozen final-answer parser, and unfinished reasoning without a valid final answer is counted as incorrect. LiveCodeBench uses its frozen public/private tests and evaluation pipeline.

Pass@k. For selected multiple-sample evaluations, we draw $n = 16$ independently seeded completions for each question and estimate

$$\widehat{\mathrm{pass@}k} = \frac{1}{N}\sum_{i=1}^{N}\left[1 - \frac{\binom{n-c_i}{k}}{\binom{n}{k}}\right], \qquad k \in \{4, 8, 16\}, \tag{19}$$

where c_i denotes the number of correct completions for question i. Thus, Pass@4 and Pass@8 are estimated from all 16 samples rather than from only the first 4 or 8 generations.

Inference-budget evaluation. For inference-length analyses, we evaluate prefixes of the same long sampled trajectory at budgets of $2{,}048$, $4{,}096$, $8{,}192$, $16{,}384$, and $32{,}768$ tokens. These results therefore measure how solution availability evolves along a common trajectory and should not be interpreted as five independently generated responses under separately imposed decoding limits.

Inference requirements. The teacher, hindsight view, and frozen reference are used only during training. At inference, HIND requires only the trained student, with no auxiliary model calls, retrieval, or reranking. This removes auxiliary inference overhead at a fixed decoding budget, although realized token usage may still vary because training can change response length.

C COMPREHENSIVE EXPERIMENT

C.1 ADDITIONAL DETAILS ON THE HINDSIGHT OBSERVATION

All analyses in Figure 2 are performed on on-policy reasoning states generated by Qwen3-1.7B. For each problem, the student first generates a reasoning trajectory without access to the verified answer.

The sampled trajectory is then held fixed, and the causal student, hindsight student, and teachers are evaluated on exactly the same prefixes. Thus, the three panels compare different supervision signals without changing the underlying student-generated state distribution.

Panel (a): Student support for teacher actions under hindsight. At an on-policy prefix

$$h_t = (x, y_{<t}), \tag{20}$$

let a_t^T denote the teacher action evaluated at that state. We measure its rank under the same student in the causal and hindsight views,

$$r_t = \mathrm{rank}_{p_S(\cdot|h_t)}(a_t^T), \qquad r_t^* = \mathrm{rank}_{p_S(\cdot|h_t, y^*)}(a_t^T), \tag{21}$$

where rank 1 denotes the most probable next token.

We group teacher actions by their causal student rank into 1, 2–4, 5–8, 9–16, 17–32, 33–64, 65–128, 129+. For each teacher and rank interval $\mathcal{B}_k$, the bars report $P(r_t \in \mathcal{B}_k)$, while the overlaid curves report $P(r_t^* \le 4 \mid r_t \in \mathcal{B}_k)$. The dashed line marks the causal Top-4 boundary. Actions outside this boundary can still move into the same student's hindsight Top-4, showing that teacher reasoning not favored under the current prefix can become strongly supported once additional task information is available.

Panel (b): Causal mismatch and hindsight gain. Panel (b) analyzes the full predictive distributions on the same on-policy states. The horizontal axis reports causal teacher–student mismatch D_t, while the vertical axis reports ΔD_t. Color denotes equal-trajectory-weighted density, so each rollout contributes the same total mass regardless of length. The horizontal axis is logarithmic and the vertical axis uses a monotonic asinh transform to display positive and negative gains over a wide dynamic range.

States with similar D_t exhibit substantially different hindsight gains. Hence, even at a fixed level of causal mismatch, hindsight provides additional student-specific information about how teacher supervision relates to the learner.

Panel (c): Teacher scale and distance from the on-policy student. Panel (c) asks how teacher reasoning changes relative to the student as teacher capability increases. All teacher scales are evaluated on the same student-generated trajectories and prefixes, so the underlying on-policy state distribution is identical across comparisons.

For teacher T, we measure the student's average negative log-likelihood of the corresponding teacher actions along each rollout,

$$L_i^T = \frac{1}{|\mathcal{R}_i|} \sum_{t \in \mathcal{R}_i} -\log p_S(a_{i,t}^T \mid h_{i,t}), \tag{22}$$

where $\mathcal{R}_i$ denotes the reasoning positions of the student-generated trajectory. The large markers report the problem-balanced mean

$$\bar{L}^T = \frac{1}{N} \sum_{i=1}^{N} L_i^T, \tag{23}$$

while faint lines connect the same problems across teacher scales.

The mean NLL increases from 0.364 to 0.388 and 0.449 nats/token for the 4B, 8B, and 14B teachers, respectively. Because teacher scale is varied while the student-generated states remain fixed, this trend shows that increasingly capable teachers place more probability mass on reasoning that is less likely under the current student.

Summary. The three panels expose complementary aspects of hindsight on the same on-policy state distribution. Panel (a) shows that teacher actions outside the student's current high-probability region can gain substantial support under hindsight; Panel (b) shows that this response varies even among states with similar causal mismatch; and Panel (c) shows that the need for such student-specific learnability information grows as teacher capability increases. Together, these results support hindsight as a learnability assessment for teacher supervision rather than merely another source of supervision.



Figure A1: **Different allocation rules assign the same supervision mass to different reasoning states.** Shown on a fixed Qwen3-1.7B trajectory, the horizontal axis indexes reasoning positions and color indicates the applied supervision weight. Uniform-Matched spreads supervision evenly, Shuffled-Matched permutes the HIND weights, Causal-Matched reorders them by causal mismatch, and HIND retains its hindsight-based state assignment.

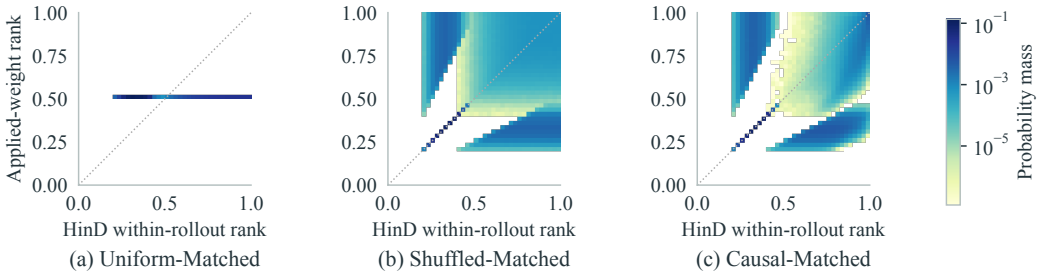

Figure A2: **Matched controls induce distinct transformations of HIND priority.** Within each rollout, the horizontal axis gives the percentile rank under the original HIND weights and the vertical axis the percentile rank under the applied control weights; color indicates equal-rollout probability mass. Uniform-Matched flattens priority, Shuffled-Matched redistributes it, and Causal-Matched imposes a structured but distinct ordering based on causal teacher–student mismatch.

C.2 ANALYSIS OF STATE-SPECIFIC SUPERVISION ALLOCATION

This section isolates the role of *where* reasoning supervision is applied while controlling its overall strength. For a trajectory with reasoning positions $t = 1, \ldots, T$ and HIND weights $\{w_t\}_{t=1}^{T}$, we consider three matched allocation rules.

Allocation rules *Uniform-Matched:* Each reasoning position receives the trajectory's mean HIND weight,

$$w_t^{\text{uni}} = \frac{1}{T} \sum_{j=1}^{T} w_j. \tag{24}$$

This preserves the total reasoning-weight mass while removing all within-trajectory prioritization.

Shuffled-Matched. The exact multiset of HIND weights is preserved, but reassigned by a random permutation π of the reasoning positions:

$$w_t^{\text{shuf}} = w_{\pi(t)}. \tag{25}$$

This preserves the full weight distribution while breaking its correspondence with individual states.

Causal-Matched. The exact HIND weight multiset is again preserved, but reassigned according to causal teacher–student mismatch D_t. Let

$$D_{t_1} \geq D_{t_2} \geq \cdots \geq D_{t_T}, \qquad w_{(1)} \geq w_{(2)} \geq \cdots \geq w_{(T)} \tag{26}$$

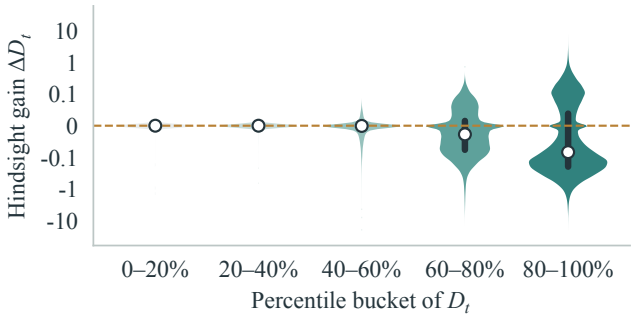

Figure A3: **Hindsight responses become more heterogeneous as causal mismatch increases.** Reasoning positions are divided into equal-trajectory-weighted quintiles of causal mismatch D_t, and each violin shows the conditional distribution of hindsight gain ΔD_t. White dots denote weighted medians and dark segments weighted interquartile ranges. The widening distributions at larger D_t show that similar levels of teacher–student mismatch can correspond to substantially different hindsight responses.

denote the sorted causal mismatches and sorted HIND weights, respectively. Causal-Matched then assigns

$$w_{t_k}^{\text{causal}} = w_{(k)}, \qquad k = 1, \dots, T. \tag{27}$$

Thus, Causal-Matched and HIND share the same total weight mass, weight histogram, sparsity, and magnitude distribution; only the state assignment differs. This makes it the most direct comparison between causal-mismatch-based and hindsight-based allocation.

Allocation along a fixed reasoning trajectory. Figure A1 visualizes the four allocation rules on a single reasoning trajectory. Uniform-Matched spreads supervision evenly across positions, while Shuffled-Matched preserves the original HIND weight values but redistributes them across states. Causal-Matched concentrates the same weights according to causal mismatch, producing a structured pattern that differs visibly from the hindsight-based placement of HIND. This comparison shows that HIND is characterized not merely by sparse or uneven supervision, but by a specific state-level organization of that supervision.

Population-level priority geometry. Figure A2 extends this comparison to the full rollout population. Within each rollout, reasoning positions are converted to percentile ranks under both the original HIND weights and the weights applied by each matched control. The resulting joint rank density shows how each rule transforms the priority structure induced by HIND.

Uniform-Matched collapses the applied ranking to a nearly constant level, reflecting the removal of within-rollout prioritization. Shuffled-Matched preserves the weight multiset but redistributes its ranking across states. Causal-Matched induces a structured geometry, reflecting the relation between causal mismatch and hindsight, while still producing substantial reprioritization relative to HIND. Across the rollout population, these patterns show that the effectiveness of supervision depends not only on its amount or marginal distribution, but also on how it is assigned to states.

C.3 Additional analysis of hindsight-based prioritization

This section further characterizes the hindsight signal used by HIND. Unless otherwise specified, all distributions are equal-trajectory-weighted so that each trajectory contributes equally.

For each reasoning position t, we use the causal and hindsight teacher–student mismatches defined in Sec. 3.1,

$$D_t = D_{\mathrm{KL}}(q_T \| p_S), \qquad D_t^* = D_{\mathrm{KL}}(q_T \| p_S^*), \tag{28}$$

with hindsight gain

$$\Delta D_t = D_t - D_t^*. \tag{29}$$

We additionally analyze the bounded HIND weight w_t, which determines the position-wise influence of teacher supervision during training.

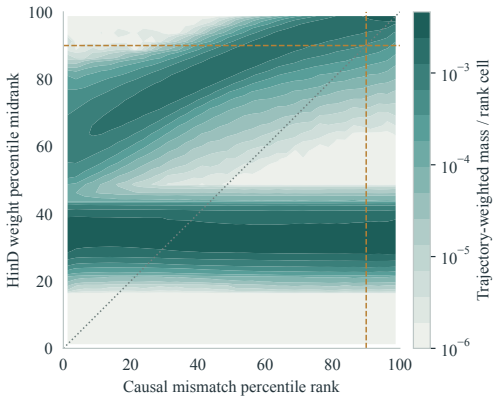

Figure A4: **Hindsight reshapes supervision priority across the full ranking space.** Within each trajectory, reasoning positions are converted to percentile midranks under causal mismatch D_t (horizontal axis) and the HIND weight w_t (vertical axis); color shows equal-trajectory-weighted mass in each rank cell. The dotted diagonal denotes identical rankings and the dashed lines mark the Top-10% regions. The broad off-diagonal mass shows that hindsight induces substantial reprioritization beyond any single Top-α cutoff.

Table A2: **Per-benchmark performance across external teacher scales.** Qwen3-1.7B is trained with HIND using Qwen3 teachers from 4B to 32B. Best results are **bolded**.

	Math Reasoning			OOD Generalization		
Teacher	**MATH-500**	**Olympiad**	**AIME25**	**LiveCodeBench**	**MMLU-Pro**	**GPQA-D**
-	86.9	63.3	40.0	30.9	62.3	38.4
4B	88.7	65.8	40.0	30.9	63.8	40.4
8B	88.3	**68.3**	40.0	**32.6**	63.3	**45.5**
14B	88.7	**68.3**	40.0	32.0	65.3	43.5
32B	**92.2**	67.1	**43.3**	32.0	**66.8**	44.0

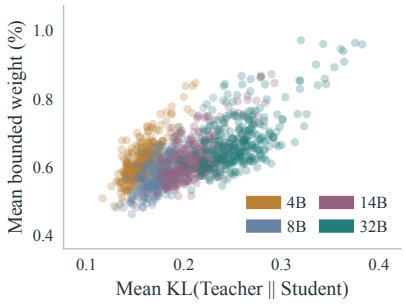

Figure A5: **HIND keeps supervision strength controlled as teacher–student mismatch grows.** Each point represents one training step across teacher scales, relating mean causal mismatch to the mean bounded HIND weight. Stronger teachers shift the training signal toward larger mismatch, while the corresponding increase in supervision weight remains comparatively moderate.

Conditional hindsight response. Figure A3 conditions ΔD_t on equal-trajectory-weighted quintiles of causal mismatch D_t. The conditional distributions broaden markedly as causal mismatch increases, showing that larger teacher–student mismatch is associated with a wider range of hindsight responses rather than a single monotonic pattern. Accordingly, the hindsight signal does not merely track the magnitude of mismatch; it differentiates how that mismatch is organized from the student's perspective.

Global ranking structure. Figure A4 examines the same relationship across the full within-trajectory ranking space. Rather than selecting a particular Top-α cutoff, we compare the percentile rankings induced by causal mismatch D_t and the HIND weight w_t. The broad off-diagonal density shows that hindsight substantially reorganizes supervision priority throughout the ranking space. This complements the Top-α overlap result in the main text by showing that the reprioritization induced by hindsight is a global property of the supervision signal rather than a phenomenon tied to one operating point.



Figure A6: **Training changes the balance between newly solved questions and regressions.** For matched GPQA-Diamond questions, the curves report the fraction changing from Base-incorrect to HIND-correct and from Base-correct to HIND-incorrect across checkpoints. The strongest checkpoints exhibit a favorable balance in which newly solved questions exceed regressions.

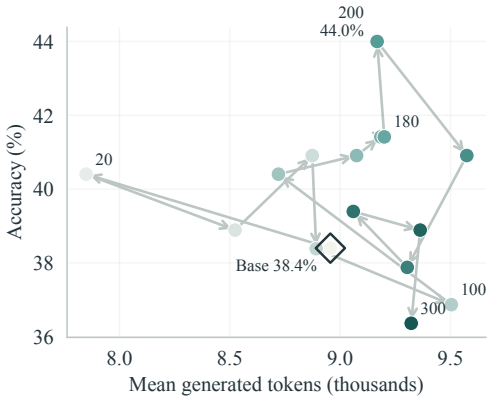

Figure A7: **Accuracy gains are not explained by longer generations alone.** Each checkpoint is positioned by its mean generated tokens and GPQA-Diamond accuracy. Substantial accuracy variation occurs within a comparable generation-length range, with the peak GPQA-Diamond checkpoint at 200 updates remaining close in length to neighboring checkpoints.

C.4 ADDITIONAL ANALYSIS OF TEACHER-STRENGTH SCALING

Per-benchmark results. Table A2 reports the full benchmark breakdown across teacher scales. Performance improves broadly as teacher capability increases, with the 32B teacher achieving the highest six-benchmark average and the best results on MATH-500, AIME25, and MMLU-Pro. Other benchmarks peak at intermediate teacher scales, but remain substantially above the Base student, showing that the aggregate scaling trend reflects broad improvements rather than gains from a single task.

Training-signal geometry. Figure A5 examines how the HIND signal changes with teacher strength. Stronger teachers shift training toward larger teacher–student mismatch, while the mean bounded weight increases much more moderately. Because the weight is determined by hindsight gain rather than mismatch magnitude itself, larger teacher–student gaps do not translate directly into proportionally stronger updates. This separation allows increasingly capable teachers to contribute richer reasoning while keeping supervision strength tied to the student's hindsight response.

C.5 DETAILED DYNAMICS ACROSS SUPERVISION BUDGETS

The cross-benchmark heatmap in Figure 5 summarizes how performance evolves across an extended 300-update training run. We provide two complementary analyses on GPQA-Diamond to examine the corresponding question-level transfer and generation behavior.

Question-level transfer. Figure A6 tracks, relative to the Base student, the fraction of questions newly solved by each checkpoint and the fraction that become incorrect. As training progresses, these two sets evolve non-monotonically, but the strongest checkpoints exhibit a clear favorable balance in which newly solved questions substantially outnumber regressions. This shows that the aggregate gains reflect broader successful coverage rather than uniform improvement on every question.

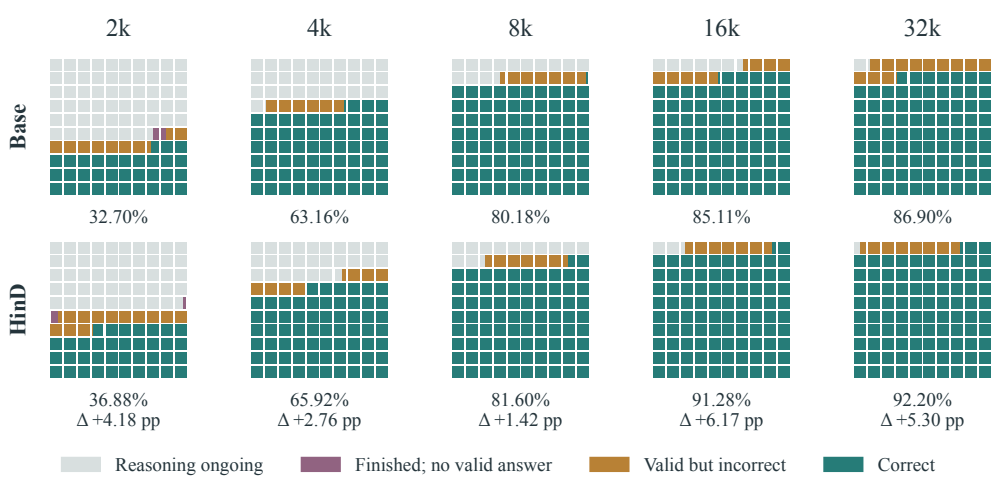

Figure A8: **HIND uses additional inference length more effectively on MATH-500.** Columns increase the token budget from 2k to 32k, while colors indicate whether the response is still reasoning, finished without a valid answer, valid but incorrect, or correct. HIND produces a larger completed fraction at matched budgets and achieves its clearest accuracy gains at longer reasoning lengths.

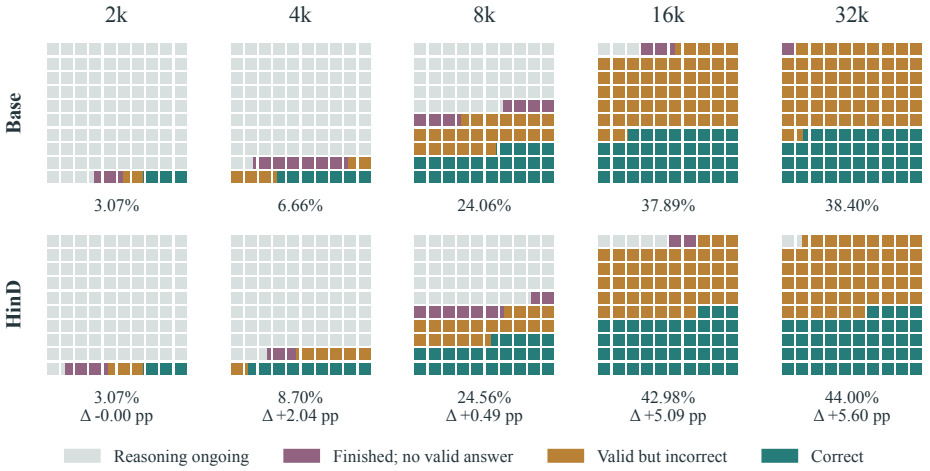

Figure A9: **HIND's inference-length advantage extends beyond the training domain.** Shown on GPQA-Diamond, each column corresponds to a larger token cutoff applied to the same generated responses, and colors indicate whether reasoning is still ongoing, ends without a valid answer, yields an incorrect answer, or reaches the correct solution. The advantage grows at longer budgets, showing that HIND converts additional inference length into correct completed answers more effectively.

Accuracy and generation length. Figure A7 places each checkpoint by its mean generated tokens and GPQA-Diamond accuracy. Large changes in accuracy occur within a relatively narrow range of generation lengths, with the peak GPQA-Diamond checkpoint at 200 updates remaining in the same general length regime as neighboring checkpoints. Thus, the training gains are not explained simply by progressively longer generations, but by changes in how effectively the student uses its reasoning budget.

Together, these analyses complement the cross-benchmark dynamics in Figure 5: additional training changes which questions the student can solve while improving performance without a corresponding monotonic increase in generation length.

C.6 CROSS-DOMAIN INFERENCE-LENGTH ANALYSIS

We extend the inference-length analysis of RQ3.2 to GPQA-Diamond, which lies outside the mathematical training domain. Using the same protocol as in the main text, we rescore a single generated response at 2k, 4k, 8k, 16k, and 32k token cutoffs, and categorize each prefix by its response state.

Figure A9 shows that the advantage of HIND becomes more pronounced as additional reasoning length is available. Accuracy improves from 37.89% to 42.98% at 16k and from 38.40% to 44.00% at 32k, while the response-state composition shifts toward correct completed answers. This mirrors the mathematical-domain trend in Figure A8 and shows that HIND converts additional inference budget into successful reasoning even beyond the domain used for training.