MASKED SELF-DISTILLATION: LETTING REASONING MODELS KEEP THINKING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) has emerged as a post-training paradigm for LLMs. However, we show that naive OPSD degrades reasoning models on tasks that demand exploration and reflection, such as competition mathematics. We trace this degradation to a context asymmetry: exploration and reflection are redundant to a solution-aware teacher, so uniform distillation causes the student to lose both abilities. Specifically, on identical student prefixes, the privileged teacher assigns far lower probability than the student to a sparse set of *epistemic markers* linked to exploration and reflection, such as "Wait" and "perhaps". We propose *masked self-distillation*, which withholds the distillation loss at marker positions and retains it everywhere else. By letting reasoning models keep thinking, the masked objective reverses the degradation in every tested configuration and outperforms SFT and GRPO. More broadly, hindsight is a poor guide to exploration: privileged self-distillation should transfer its preferences selectively, not uniformly.

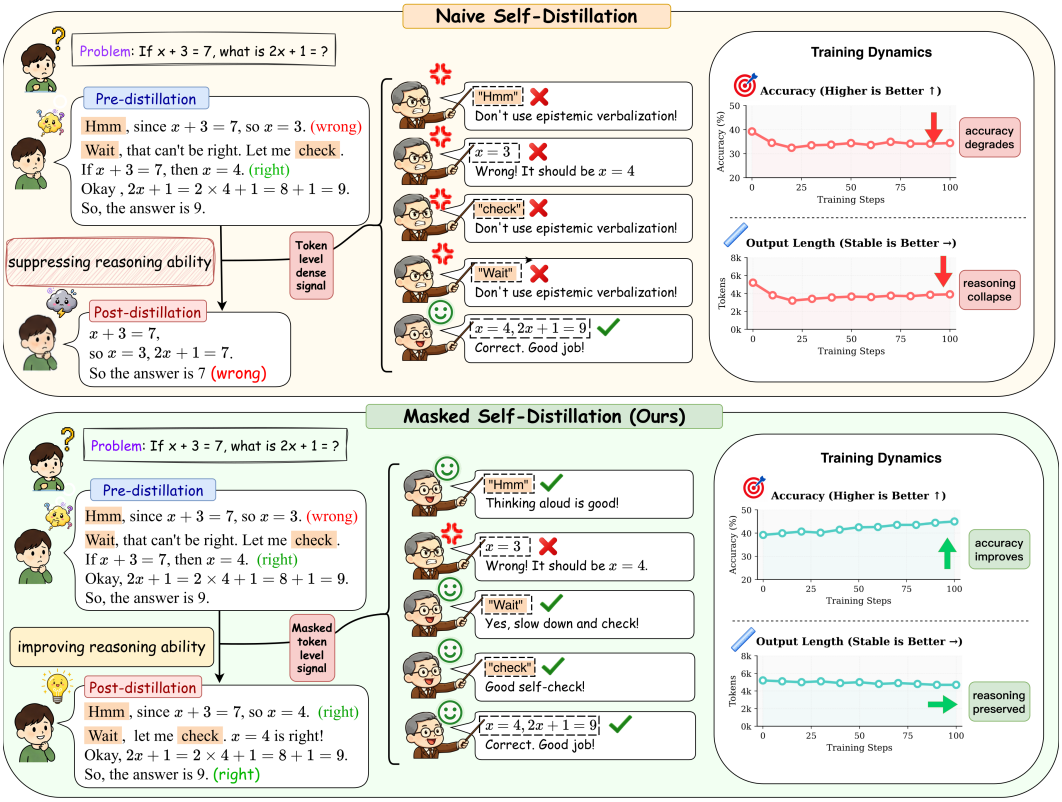


Figure 1: **Conceptual overview.** Under naive OPSD, the privileged teacher pushes the student away from its own epistemic markers, and marker use, response length, and accuracy decline together. Masked self-distillation withholds the loss at those positions and retains it elsewhere. Examples and curves are schematic.

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.

1 INTRODUCTION

On-policy distillation (OPD) trains a student on its own sampled responses using dense token-level teacher supervision, reducing the trajectory mismatch of imitating teacher-written sequences (Agarwal et al., 2024). On-policy *self*-distillation (OPSD) removes the separate teacher (Zhao et al., 2026): both branches start from the same checkpoint, but the student reasons from the problem alone while the fixed teacher additionally sees a verified solution (Figure 1, top). Only the student is updated during training; both branches score the same student-generated prefixes. OPSD variants report gains across reasoning, science QA, tool use, and code (Hübotter et al., 2026; Zhang et al., 2026b; Jin et al., 2026).

Privileged context, however, raises a question: *should every local preference of a teacher that has seen the solution be transferred to a student that will not have it?* An unassisted student may need *exploration* (considering alternative paths) and *reflection* (checking and revising intermediate steps), even when these behaviors are redundant to a solution-aware teacher. Access to a verified derivation can make a direct continuation preferable to checking an alternative. Transferring that preference uniformly may therefore discourage the student from revisiting an uncertain step, even though it must still find and verify a solution without the teacher's information. Competition mathematics exposes this tension: naive OPSD reduces accuracy below the starting checkpoint on AIME24, AIME25, and MATH-500 for DeepSeek-R1-Distill-Qwen (1.5B and 7B) and OpenMath2-Llama3.1-8B (Table 2). Kim et al. (2026) report related suppression of epistemic verbalization; we ask where the damage originates in the objective and whether it can be removed selectively.

We trace this mismatch through *epistemic markers*, lexical expressions signaling uncertainty, evaluation, or revision. "Perhaps" (uncertainty) and "Wait" (reconsideration) are illustrative examples of a broader class identified by our fixed token-level lexicon (Section 4.1). Internal states preceding "Wait" can modulate subsequent reasoning (Troitskii et al., 2025). On identical student-generated prefixes, the privileged teacher assigns sampled markers far lower probability than the student does, while mean probabilities for sampled non-marker tokens are similar (Table 1). The respective teacher and student probabilities are 0.127 versus 0.417 at marker positions, compared with 0.751 versus 0.761 elsewhere. Because both branches score identical prefixes, this comparison isolates the effect of privileged context rather than differences between independently generated reasoning trajectories. Marker positions occupy only 3% of response tokens but carry 20% of the position-clipped teacher–student divergence (Figure 3). During naive OPSD training, marker frequency, response length, and accuracy fall together (Figures 4 and 5); Appendix G illustrates the resulting behavior.

This diagnosis suggests a simple but effective remedy. *Masked self-distillation* withholds the distillation term where the student has already generated a marker and retains supervision elsewhere (Figures 1 and 2). It requires no additional model or reward, never inserts or rewards markers, and leaves 97% of positions untouched. Masking reverses degradation on all three mathematical benchmarks; on DeepSeek-R1-Distill-Qwen, it exceeds naive OPSD by 5.0–12.1 points and outperforms SFT and GRPO (Table 2). A non-marker mask matched on count, frequency, position, entropy, and divergence stays within 1.2 points of naive OPSD, while banning markers during decoding hurts the untrained model (Table 2). These controls support the importance of where supervision is withheld, beyond simply reducing its amount. Masking also improves scientific reasoning on SciKnowEval (Table 3).

Our contributions are threefold:

1. **Diagnosing naive OPSD's failure.** We identify a context asymmetry: the solution-aware teacher downweights epistemic markers linked to exploration and reflection that the unassisted student may still need. Same-prefix probes localize this mismatch: 3% of positions carry 20% of the clipped diagnostic divergence, motivating selective transfer.
2. **Masked self-distillation.** We introduce a selective objective that withholds distillation at student-generated marker positions and retains supervision elsewhere. It requires no additional model or reward and never inserts or rewards markers.
3. **Empirical validation.** Across three checkpoints and three mathematical benchmarks, masking reverses naive OPSD's degradation in all nine configurations. On DeepSeek-R1-Distill-Qwen, it improves accuracy by 5.0–12.1 points over naive OPSD and outperforms SFT and GRPO. Cross-task controls further characterize its scope.

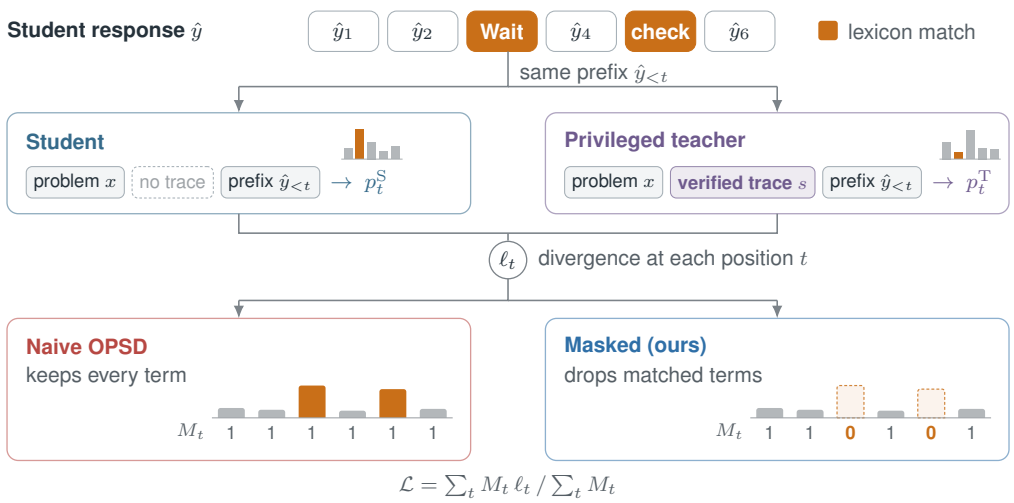


Figure 2: **Masked self-distillation.** Student and teacher score the same student-generated prefix; only the teacher sees the verified trace s. Naive OPSD applies the local term ℓ_t at every position. The masked objective zeroes it at lexicon-matched positions and keeps it elsewhere; it removes a term and adds no reward. Bar heights are schematic.

2 Self-Distillation with Privileged Context

2.1 Objective

Let x be a problem, s privileged information such as a verified solution, and $\hat{y} \sim \pi_\theta^{\mathrm{S}}(\cdot \mid x)$ a response sampled by the student. At response position t, student and teacher score the same sampled prefix under different context:

$$p_t^{\mathrm{S}} = \pi_\theta(\cdot \mid x, \hat{y}_{<t}), \tag{1}$$

$$p_t^{\mathrm{T}} = \pi_{\theta_0}(\cdot \mid x, s, \hat{y}_{<t}). \tag{2}$$

Naive OPSD minimizes the teacher-to-student divergence uniformly over response positions,

$$\mathcal{L}_{\mathrm{SD}}(\theta) = \frac{1}{L} \sum_{t=1}^{L} D_{\mathrm{KL}}\left(\mathrm{sg}[p_t^{\mathrm{T}}] \,\|\, p_t^{\mathrm{S}}\right), \tag{3}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient: the teacher branch provides the target distribution, and no gradient flows through it, so only the student branch is updated. Here θ_0 denotes the initial policy, and the teacher is fixed to it rather than following the student (Zhao et al., 2026). With identical conditioning and forward-pass settings, the two branches coincide before any update, so $p_t^{\mathrm{T}} = p_t^{\mathrm{S}}$ and the distillation gradient vanishes; privileged context creates the nonzero learning signal studied here. Because both branches score the same prefix, comparing p_t^{T} with p_t^{S} isolates what the privileged context changes in the next-token distribution, rather than comparing two independently generated trajectories. This comparison is the basis of our diagnosis.

2.2 Related Work

Context and on-policy distillation. Context distillation moves information available in a teacher prompt into the parameters (Snell et al., 2022), and on-policy distillation reduces the train–test trajectory mismatch by supervising student-generated responses (Agarwal et al., 2024). Zhao et al. (2026) combine the two into OPSD, with one model under privileged teacher and unprivileged student contexts. Related approaches use environment feedback, multi-task training, long context, and other forms of privileged information (Hübotter et al., 2026; Jin et al., 2026; Zhang et al., 2026b; Ye et al., 2026; Peñaloza et al., 2026). Separately, Zhang et al. (2026a) improve code generation by applying supervised fine-tuning to a model's own sampled outputs. We study a failure mode of privileged-context OPSD and its remedy.

Failure modes and reasoning compression. Distillation often shortens outputs, which connects it to work on concise reasoning (Shrivastava et al., 2026; Sang et al., 2026; Deng et al., 2026; Yang et al., 2026; Luo et al., 2026). Closest to us, Kim et al. (2026) show that richer teacher context can suppress epistemic verbalization and that the cost of this compression depends on model behavior and task coverage. Our study is complementary and intervention-focused: we measure the local objective on student prefixes and test whether a structured subset of teacher preferences should be withheld.

Epistemic markers and self-correction. Chain-of-thought methods elicit intermediate reasoning and exploration (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023; Yao et al., 2023). Language models can verbalize uncertainty, although verbalized and internal uncertainty align only imperfectly (Lin et al., 2022; Kadavath et al., 2022; Yona et al., 2024). Work on self-correction and process supervision asks when models can detect and repair their own errors (Madaan et al., 2023; Huang et al., 2023; Lightman et al., 2023; Uesato et al., 2022; Cobbe et al., 2021). Internal states preceding "Wait" modulate subsequent reasoning (Troitskii et al., 2025), whereas NoWait can suppress explicit thinking tokens without reducing utility in its tested settings (Wang et al., 2025). We therefore treat markers as contextual cues, not universally necessary triggers; our intervention acts on their training loss.

3 Diagnosing the Failure Mode at the Token Level

We investigate the signal behind the accuracy decline, shorter responses, and reduced epistemic-marker use observed under naive OPSD (Section 5). Whether such words themselves trigger reflection has been studied elsewhere (Wang et al., 2025; Troitskii et al., 2025); our question is narrower: does privileged context make the diagnostic divergence unusually concentrated at these observable positions?

3.1 Same-Prefix Probability Shift and Signal Concentration

Setup. With DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), we draw 12 student samples for each of the 30 AIME24 and 30 AIME25 problems (AIME I and II from each year; Mathematical Association of America, n.d.), giving 720 trajectories and 3.06M response positions. The student sees only the problem; the stopped-gradient teacher additionally sees the full reference reasoning. Both branches score the identical student-generated prefixes, and a position is *marker-matched* if the sampled token belongs to the tokenizer-level marker set used in training (Section 4). We report two quantities. The *probability probe* records the probability that each branch assigns to the token the student actually sampled, averaged over the positions in a group. The *signal share* measures concentration. We compute the full-vocabulary divergence at each position and clip it at $\tau = 5$ so that a few extreme positions cannot dominate,

$$d_t^{\mathrm{diag}} = \min\big(D_{\mathrm{KL}}(p_t^{\mathrm{T}} \| p_t^{\mathrm{S}}), \tau\big)\,, \tag{4}$$

and define a group's share as its sum of d_t^{diag} divided by the sum over all positions, that is, the fraction of the total clipped divergence that the group contributes. This position-level clipping is a diagnostic convention; the training loss bounds each vocabulary component instead (Section 4).

Table 1: **Aggregate same-prefix diagnostic.** Probabilities are means for the sampled student token. Signal is the position-clipped diagnostic divergence of Equation 4. "Clipped pos." is the fraction of positions in a row that reach $d_t^{\mathrm{diag}} = \tau$. Counts are rounded, so group counts need not sum exactly to the total.

Group	Positions	Pos. %	p^{T}	p^{S}	Signal %	Clipped pos. %
Marker-matched	92.1K	3.01	0.127	0.417	20.3	3.8%
Other	2.97M	96.99	0.751	0.761	79.7	–
All	3.06M	100.00	–	–	100.0	0.62%

Results. Table 1 and Figure 3 summarize the probe. At marker-matched positions the teacher assigns the sampled token far lower probability than the student does (0.127 versus 0.417); for sampled non-marker tokens the two branches assign similar mean probabilities (0.751 versus 0.761). Under the unclipped KL objective with a stopped-gradient teacher, this gap implies a local gradient that lowers the sampled marker's student logit (Appendix K); component-wise clipping changes that gradient.

The diagnostic divergence is concentrated. Marker-matched positions are 3.01% of all positions but carry 20.3% of the clipped divergence, and 3.8% of them reach the clip τ, compared with 0.62% of positions overall. In the single trajectory at the top of Figure 3, marker positions receive a mean signal comparable to that of the annotated error positions (0.123 versus 0.147): marker and annotated error positions show comparable diagnostic mismatch, rather than necessarily comparable training gradients.

These measurements establish a strong, localized mismatch in the objective; Section 5 tests whether withholding it matters.

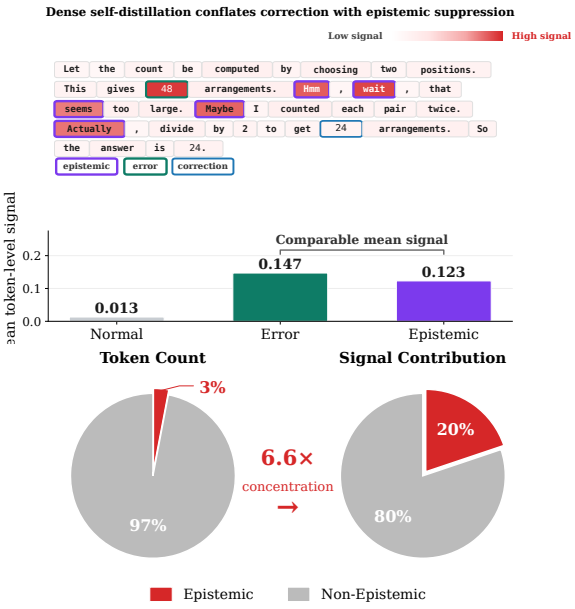


Figure 3: Position-level concentration of the diagnostic signal. Top: one illustrative trajectory ("epistemic" in the figure denotes marker-matched positions). Bottom: summary over 720 trajectories; percentages are rounded.

3.2 Why the Privileged Teacher Prefers Direct Continuations

The conditioning asymmetry offers a simple account of this concentration. Privileged context can reduce the teacher's need for exploration and reflection, while the student must still reason without that context. After a prefix ending in "Wait" or "let me check", the student may explore an alternative path or reflect on a previous step. The solution-aware teacher may instead prefer the direct derivation that the reference confirms. Uniform distillation transfers this context-dependent preference to the unprivileged student, who may still need to explore, check, and revise. Appendix I illustrates the two generation styles on one problem.

A free-generation probe supports this account. Holding the checkpoint fixed and varying the context from none to a final answer, a concise solution, or a full reasoning trace, marker frequency falls from 3.1% to 2.6%, 1.8%, and 1.3%, respectively, while responses become shorter and more accurate (Appendix A.1). Richer privileged context makes the teacher more direct, and it is this directness that uniform distillation transfers. Length and accuracy change along with marker use, so the probe corroborates the account without isolating the markers themselves.

4 Masked Self-Distillation

The diagnosis motivates selective transfer at lexical cues associated with exploration and reflection. We target these observable cues rather than attempting to annotate the behaviors directly. Masked self-distillation withholds the objective at marker-matched positions, where the privileged teacher's preferences differ sharply from the student's, and retains it elsewhere.

4.1 Marker Inventory and Tokenizer-Level Matcher

We use *epistemic markers* as a lexical proxy for exploration and reflection, including verbalized uncertainty, reconsideration, self-checking, and error recognition. Our marker categories draw on

Algorithm 1 Masked Self-Distillation

Require: Pairs $\mathcal{S} = \{(x_i, s_i)\}$; initial policy π_{θ_0}; marker-ID set $\mathcal{I}$; component bound τ.
Initialize student parameters $\theta \leftarrow \theta_0$; keep teacher parameters θ_0 fixed.
while not converged **do**
  Sample minibatch $\mathcal{B} \subset \mathcal{S}$
  for all $(x, s) \in \mathcal{B}$ **do**
    Sample $\hat{y} \sim \pi_\theta^{\mathrm{S}}(\cdot \mid x)$
    $M_t \leftarrow \mathbf{1}[\hat{y}_t \notin \mathcal{I}]$ for $t = 1, \ldots, |\hat{y}|$
    for $t = 1, \ldots, |\hat{y}|$ **do**
      $p_t^{\mathrm{T}} \leftarrow \mathrm{sg}[\pi_{\theta_0}(\cdot \mid x, s, \hat{y}_{<t})]$
      $p_t^{\mathrm{S}} \leftarrow \pi_\theta^{\mathrm{S}}(\cdot \mid x, \hat{y}_{<t})$
      $\ell_t \leftarrow \sum_{v \in \mathcal{V}} \min\left(p_{t,v}^{\mathrm{T}} \log \frac{p_{t,v}^{\mathrm{T}}}{p_{t,v}^{\mathrm{S}}}, \tau\right)$
    end for
    $\mathcal{L}(x; \theta) \leftarrow \frac{\sum_t M_t \ell_t}{\max(\sum_t M_t, 1)}$
  end for
  Update θ using the minibatch mean of $\mathcal{L}(x; \theta)$
end while

prior work on uncertainty and self-reflection (Yona et al., 2024; Wang et al., 2025; Troitskii et al., 2025). We use Qwen3-32B to identify candidate marker expressions in 1,000 student-generated reasoning trajectories, then manually review and filter the candidates (Appendix F). A matched occurrence identifies a lexical cue; its role in the surrounding reasoning can vary with context.

To construct the training matcher, we take the content words of each expression, enumerate capitalization, punctuation, and whitespace variants, and retain only variants encoded as a single non-special token. The resulting token-ID set $\mathcal{I}$ is fixed before training. A response position is *marker-matched* exactly when its sampled token belongs to $\mathcal{I}$; matching uses token identity without contextual semantic annotation. We rebuild the set for each model and task because token IDs are tokenizer-specific.

The main DeepSeek-R1-Distill-Qwen results use the full inventory. Separately, we test sensitivity to a smaller, audited *core* list, selected by occurrence count, same-prefix probability shift, and student entropy. Appendix A gives the selection criteria, compares the full and core lists, and reports candidate audits across models and tasks; an audited list on a new tokenizer is termed an *adapted* list.

4.2 Masked Objective

For a student response $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_L)$, define

$$M_t = \begin{cases} 0 & \text{if } \hat{y}_t \in \mathcal{I}, \\ 1 & \text{otherwise,} \end{cases} \tag{5}$$

so that $M_t = 0$ exactly when the student has already generated a matched marker. The mask acts only on the loss: it never forces, inserts, or rewards markers. Naive and masked runs share the same per-position loss term, in which each vocabulary component of the divergence is bounded before summation,

$$\ell_t = \sum_{v \in \mathcal{V}} \min\left(p_{t,v}^{\mathrm{T}} \log \frac{p_{t,v}^{\mathrm{T}}}{p_{t,v}^{\mathrm{S}}},\ \tau \right), \tag{6}$$

which matches the training implementation and differs from the position-level clipping of Equation 4. The masked loss is

$$\mathcal{L}_{\text{masked}}(\theta) = \frac{\sum_{t=1}^{L} M_t \ell_t}{\max(\sum_{t=1}^{L} M_t,\ 1)}. \tag{7}$$

Normalizing by the number of retained positions keeps the loss on the same scale as naive OPSD, which is recovered by setting $M_t = 1$ everywhere. We compute ℓ_t in chunks of response positions to avoid materializing the full $L \times |\mathcal{V}|$ tensor; Algorithm 1 summarizes the training procedure.

5 EXPERIMENTS

We train two representative reasoning models with different backbones and training recipes, DeepSeek-R1-Distill-Qwen (1.5B and 7B) and OpenMath2-Llama3.1-8B, and compare masked self-distillation with naive OPSD, the untrained base checkpoint, supervised fine-tuning (SFT), and group relative policy optimization (GRPO) (Shao et al., 2024). We then test whether the gain is specific to marker positions and how it depends on the task.

5.1 SETUP

Models and data. We train DeepSeek-R1-Distill-Qwen at 1.5B and 7B (Guo et al., 2025), Qwen models distilled from an RL-trained reasoner, and OpenMath2-Llama3.1-8B (Toshniwal et al., 2024), a Llama model fine-tuned on step-by-step mathematical solutions, each with a marker list built for its tokenizer (Section 4.1). Training problems and reference traces come from up to 30K OpenThoughts mathematics examples (Guha et al., 2025); the teacher receives the trace as privileged context. The SFT and GRPO baselines, the controls of Section 5.3, and the cross-task experiments use DeepSeek-R1-Distill-Qwen; the cross-task controls train the 7B model on SciKnowEval L3, ToolAlpaca, and MASSIVE (Feng et al., 2024; Tang et al., 2023; FitzGerald et al., 2023); Appendix A.4 lists their output formats and metrics.

Baselines. Naive OPSD uses the same rollouts, teacher context, component-wise bound, and normalization as the masked objective and sets $M_t = 1$ everywhere. SFT trains on the same reference traces with token-level cross-entropy (Ouyang et al., 2022); GRPO optimizes a sparse outcome reward on the same problems. All methods run for 100 updates with the hyperparameters of Appendix C.

Evaluation. We evaluate on AIME 2024, AIME 2025, and MATH-500 (Guo et al., 2025; Hendrycks et al., 2021). AIME reports avg@12, the mean correctness over 12 sampled responses per problem, averaged over problems; MATH-500 reports pass@1. Answers are extracted from `\boxed{}` and checked with Math-Verify (Kydlíček, n.d.). Cross-task and prompt-robustness conditions report mean $\pm$ standard deviation over three runs; the remaining tables report point estimates without run-level uncertainty, for which we make no significance claims.

Implementation. Our implementation builds on veRL (Sheng et al., 2024); each run uses 8 NVIDIA A800-80GB GPUs.

5.2 MAIN RESULTS

Table 2 reports the main comparison; checkpoint curves are in Appendix J. Naive OPSD ends below the base model in all nine model–benchmark configurations, by 1.1–8.3 points. Masking the marker positions reverses this in every configuration: masked self-distillation ends above the base model throughout. On DeepSeek-R1-Distill-Qwen it is 5.0–12.1 points above naive OPSD and above SFT and GRPO on every benchmark; on OpenMath2-Llama3.1-8B, with a marker list rebuilt for its tokenizer, it is 2.0–4.4 points above naive OPSD. Both the failure and the remedy thus appear across backbones and training recipes. SFT gives small, consistent gains from the reference traces, whereas GRPO does not improve consistently under the fixed 100-update protocol, which is unsurprising for checkpoints already distilled from an extensively RL-trained model (Guo et al., 2025). Over training, marker frequency, response length, and accuracy fall together under naive OPSD and remain higher under masking (Appendix B); because accuracy falls with length, the shortening accompanies a loss of accuracy rather than preserving performance. The qualitative examples illustrate fewer explicit checks.

5.3 IS THE GAIN SPECIFIC TO MARKER POSITIONS?

Matched non-marker mask. Masking removes supervision, so a gain could reflect supervising fewer positions, or positions that happen to be noisy, rather than marker positions in particular. The matched control in Table 2 addresses this. Within each response it masks the same number of non-marker positions, chosen to match the marker positions on token-frequency bin, response-position decile, student-entropy decile, and clipped-divergence decile (standardized mean differences below 0.05 on all four). It ends within -0.3 to $+1.2$ points of naive OPSD, whereas the marker mask exceeds it by 4.4–11.3 points. Equal-count random masking likewise tracks naive OPSD, and masking the

Table 2: **Training and intervention results.** AIME: avg@12 (%); MATH-500: pass@1 (%). Average is the unweighted mean of the three scores. Bold marks the highest point estimate in each training block; no significance claim is intended. The final block changes decoding only.

Method	AIME24	AIME25	MATH-500	Average
DeepSeek-R1-Distill-Qwen · 1.5B				
Base	27.2	20.8	69.6	39.2
SFT	28.8	21.8	73.5	41.4
GRPO	26.0	20.2	71.3	39.2
Naive OPSD	21.2	16.5	65.5	34.4
Matched non-marker mask (control)	22.4	16.2	66.3	35.0
Masked self-distillation (ours)	**32.5**	**24.5**	**77.6**	**44.9**
DeepSeek-R1-Distill-Qwen · 7B				
Base	47.8	38.1	86.8	57.6
SFT	48.2	38.5	87.2	58.0
GRPO	47.0	38.2	85.6	56.9
Naive OPSD	39.5	30.8	84.0	51.4
Matched non-marker mask (control)	40.3	30.5	84.6	51.8
Masked self-distillation (ours)	**49.0**	**39.8**	**89.0**	**59.3**
OpenMath2-Llama3.1 · 8B				
Base	11.7	4.7	69.4	28.6
Naive OPSD	9.2	3.6	66.8	26.5
Masked self-distillation (ours)	**13.3**	**5.6**	**71.2**	**30.0**
Decoding only: DeepSeek-R1-Distill-Qwen-7B *(no training)*				
Standard decoding	47.8	38.1	86.8	57.6
Core-marker ban	43.1	34.2	84.7	54.0

highest-student-entropy positions recovers only part of the gain (Appendix B). The benefit is therefore not explained by mask size or by the marginal statistics of the masked positions; what the masked positions mean is what matters.

Decoding-time marker ban. If marker positions matter for reasoning, removing them from an untrained model should hurt. We set the logits of the single-token core markers to $-\infty$ while decoding from the base 7B checkpoint, renormalize, and keep every other decoding setting; multi-token expressions are left untouched. Accuracy falls on all three benchmarks, most on AIME24 and AIME25 (Table 2, last block). Because the ban alters every subsequent token, it shows that access to these markers matters for the model's behavior; the surface forms and the continuations they trigger are removed together.

5.4 When Does Masking Help? Cross-Task Controls

Our account predicts that masking matters only where the distillation signal concentrates at marker positions, which should happen on tasks that call for exploration and reflection. Table 3 examines this prediction in four settings with the same objective, budget, and model. The benefit of masking follows the concentration rather than a math versus non-math split. AIME and SciKnowEval show substantial concentration, degradation under naive OPSD, and recovery under masking. ToolAlpaca shows weak concentration, no degradation, and no difference between the objectives. MASSIVE, whose constrained label decoding contains no marker positions, makes the two objectives identical. Four settings cannot establish concentration as a causal moderator, but the graded association is what the account predicts.

All four controls use the same 7B checkpoint and 100-update training budget; the trained scores in Table 3 are means and standard deviations over three runs (Appendix A.4). Masking improves AIME24 and SciKnowEval by 9.5 and 3.5 points over naive OPSD, respectively. The 0.2-point ToolAlpaca difference is small relative to the reported run-to-run variation; on MASSIVE the objectives coincide because there are no matched positions.

Table 3: **Cross-task controls.** Marker and signal columns are shares of response positions and of position-clipped diagnostic divergence. Trained conditions are mean ± standard deviation over three runs. AIME position statistics pool AIME24/25; the displayed score is AIME24 avg@12.

Task / metric	Marker %	Signal %	Base	Naive	Masked
AIME24 / avg@12	3.01	20.3	47.8	39.5±0.7	49.0±0.5
SciKnowEval L3 / norm. aggregate	2.34	13.1	41.8	39.2±0.6	42.7±0.5
ToolAlpaca / task success	0.62	2.1	54.2	55.4±0.7	55.6±0.6
MASSIVE / intent accuracy	0.00	0.00	83.9	84.5±0.3	84.5±0.3

6 Discussion and Limitations

What the evidence shows. Privileged context lowers the teacher's probability of the student's own marker tokens and concentrates the divergence at those positions; withholding the loss there reverses the degradation while random, entropy-based, and statistically matched masks do not; and removing the markers at decoding time hurts an untrained model. Together these results establish that marker positions are where the privileged teacher's preferences conflict with what the student needs, and that withholding the loss there is what restores accuracy. They leave open how much of the benefit flows through the words themselves and how much through response length, exploration and reflection, or internal states; evidence that pre-marker internal states modulate reasoning (Troitskii et al., 2025) leaves this distinction unresolved. Masked self-distillation is best understood as removing a measured local pressure at lexical cues associated with exploration and reflection. We do not separately quantify either behavior, and our qualitative examples primarily illustrate checking and revision.

Scope. The study covers English outputs, three checkpoints from two model families, four task settings, and 100-update protocols. The cross-task controls support a graded, task-dependent pattern but not transfer to arbitrary models, languages, prompts, or longer training, and prompt robustness is shown only for the three formulations in Appendix A. Most tables report point estimates without run-level uncertainty.

Matcher. The hand-specified lexicon depends on the language, tokenizer, model, and task. Single-token matching misses multi-token expressions and implicit reflection, and words such as "actually" can be false positives. The audit must be repeated for new settings (Appendix A). Soft masks and contextual or internal-state selectors remain untested.

Future work. Future work can extend selective self-distillation to multilingual reasoning and broader task settings, using task-appropriate contextual markers to identify which teacher preferences to transfer.

Broader implication. A uniform token-level objective applies the same loss rule to positions that can play different roles in reasoning. In the OPSD settings tested here, it fails to distinguish a step the teacher can skip from a check the student needs. We take this as a reason to **account for token function when designing post-training objectives**, although we have not tested the idea beyond self-distillation.

7 Conclusion

Privileged-context self-distillation can transfer preferences that suit a solution-aware teacher but conflict with an unassisted student's need for exploration and reflection. Our diagnosis locates this mismatch at sparse epistemic-marker positions, which account for a disproportionate share of teacher–student divergence. Masked self-distillation withholds supervision at these positions and retains it elsewhere, reversing the accuracy decline under naive OPSD across the tested mathematical reasoning configurations. The intervention preserves dense teacher supervision at the remaining 97% of response positions and requires neither an additional model nor a reward signal. Matched-mask controls support the importance of where supervision is withheld, while cross-task results link the benefit to diagnostic signal concentration. Together, these findings motivate selective transfer of privileged teacher preferences, leaving room for the student's own exploration and reflection.

AI USE STATEMENT

Large Language Models (LLMs) were used to assist with manuscript drafting and revision, including sentence rephrasing, grammar checking, and improvements to readability and clarity, as well as literature search. The authors take full responsibility for the manuscript's scientific content, accuracy, and originality.

REPRODUCIBILITY STATEMENT

Section 2 defines the teacher–student conditioning and the loss, Section 4 the matcher and masked objective, and Section 5 the data, metrics, and compute. The appendix provides prompts, the marker inventory and matching rules, baseline objectives, optimization and decoding hyperparameters, robustness protocols, training curves, and qualitative examples, and every table states whether it reports three-run means or point estimates. All checkpoints, datasets, Math-Verify, and veRL are publicly available.

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 *The Twelfth International Conference on Learning Representations*, 2024.

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

Jie Deng, Shining Liang, Jun Li, Hongzhi Li, and Yutao Xie. Conpress: Learning efficient reasoning from multi-question contextual pressure. *arXiv preprint arXiv:2602.01472*, 2026.

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

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4277–4302. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.235. URL https://aclanthology.org/2023.acl-long.235/.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. OpenThoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

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

Yiqiao Jin, Yiyang Wang, Lucheng Fu, Yijia Xiao, Yinyi Luo, Haoxin Liu, B. Aditya Prakash, Josiah Hester, Jindong Wang, and Srijan Kumar. Unisd: Towards a unified self-distillation framework for large language models. *arXiv preprint arXiv:2605.06597*, 2026.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

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. URL https://arxiv.org/abs/2603.24472.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Hynek Kydlíček. Math-Verify: Math verification library. https://github.com/huggingface/Math-Verify, n.d. Software; official citation metadata accessed 2026-09-13.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Qin-Wen Luo, Sheng Ren, Xiang Chen, Rui Liu, Jun Fang, Naiqiang Tan, and Sheng-Jun Huang. Compress the easy, explore the hard: Difficulty-aware entropy regularization for efficient llm reasoning. *arXiv preprint arXiv:2602.22642*, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2023.

Mathematical Association of America. MAA invitational competitions. https://maa.org/maa-invitational-competitions/, n.d. Historical problem sets reproduced with MAA permission at LIVE by Po-Shen Loh: 2024 AIME I, 2024 AIME II, 2025 AIME I, and 2025 AIME II. Accessed: 2026-09-16.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 2022.

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

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. CRISP: Compressed reasoning via iterative self-policy distillation. *arXiv preprint arXiv:2603.05433*, 2026.

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

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

Vaishnavi Shrivastava, Ahmed H. Awadallah, Vidhisha Balachandran, Shivam Garg, Harkirat Behl, and Dimitris Papailiopoulos. Sample more to think less: Group filtered policy optimization for concise reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a00ef870c140f09263ba2288855be50b-Abstract-Conference.html.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023. URL https://arxiv.org/abs/2306.05301.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024. URL https://arxiv.org/abs/2410.01560.

Dmitrii Troitskii, Koyena Pal, Chris Wendler, and Callum Stuart McDougall. Internal states before wait modulate reasoning patterns. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18640–18649, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1012. URL https://aclanthology.org/2025.findings-emnlp.1012/.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Chenlong Wang, Yuanning Feng, Dongping Chen, Zhaoyang Chu, Ranjay Krishna, and Tianyi Zhou. Wait, we don't need to "wait"! removing thinking tokens improves reasoning efficiency. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 7459–7482, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.394. URL https://aclanthology.org/2025.findings-emnlp.394/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Yongxin Wang, Wenlei Shi, Yiwei Wang, Xiaodan Liang, and Jing Tang. Accordion-thinking: Self-regulated step summaries for efficient and readable llm reasoning. *arXiv preprint arXiv:2602.03249*, 2026.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7752–7764, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.443. URL https://aclanthology.org/2024.emnlp-main.443/.

Ruixiang Zhang, Richard He Bai, Huangjie Zheng, Navdeep Jaitly, Ronan Collobert, and Yizhe Zhang. Embarrassingly simple self-distillation improves code generation. *arXiv preprint arXiv:2604.01193*, 2026a.

Xinsen Zhang, Zhenkai Ding, Tianjun Pan, Run Yang, Chun Kang, Xue Xiong, and Jingnan Gu. Opsdl: On-policy self-distillation for long-context language models. *arXiv preprint arXiv:2604.17535*, 2026b.

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. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64784.

A Additional Diagnostic and Robustness Results

A.1 Reference-Context Generation Probe

We vary the context used for free generation while holding the model checkpoint fixed. This probe compares generation modes; it is not a set of training ablations. "Student" means no reference, while the other rows are teacher-mode generations with progressively richer privileged information.

Table 4: Free-generation statistics under four conditioning contexts. Output length and marker-position columns are reported per-generation averages. Length and accuracy vary jointly with marker frequency, so the table is corroborative rather than a controlled lexical intervention.

Condition	Avg. length	Avg. marker positions	Marker %	Acc.
None (student)	4,300	133	3.1%	33.3%
Final answer only	4,706	122	2.6%	38.3%
Concise solution	3,005	54	1.8%	74.7%
Full reasoning trace	1,279	17	1.3%	90.0%

A.2 Prompt-Formulation Robustness

We test three semantically equivalent student/teacher prompt styles. Diagnosis uses fixed student trajectories, while masked-OPSD training uses identical data and hyperparameters across styles. Each training condition has three runs. Let $T_{\mathcal{E}}$ denote matched positions and define $\Delta p = \mathbb{E}_{t \in T_{\mathcal{E}}}[p_t^{\mathrm{S}}(\hat{y}_t) - p_t^{\mathrm{T}}(\hat{y}_t)]$.

Table 5: Prompt robustness for DeepSeek-R1-Distill-Qwen-7B masked self-distillation. Scores are mean $\pm$ standard deviation over three runs. Signal uses the position-clipped diagnostic in Equation 4.

Prompt	Δp	Signal %	AIME24	AIME25	MATH-500
Original	0.290	20.3	49.0±0.5	39.8±0.4	89.0±0.2
Concise paraphrase	0.284	19.9	48.7±0.6	39.9±0.5	88.8±0.2
Minimal formatting	0.279	19.5	48.9±0.5	39.4±0.6	88.6±0.3

The maximum downstream difference is 0.5 points and lies within the displayed run-to-run variation. This supports robustness to the three tested formulations, not arbitrary prompt formats.

A.3 Candidate Audit, Core-List Sensitivity, and Transfer

This audit tests whether the gains depend on the full breadth of the source inventory. We select a smaller diagnostic core and compare training with this subset against the full-list matcher used for the main DeepSeek-R1-Distill-Qwen results. We also examine how candidate statistics transfer across models and tasks. For a candidate expression e, the supplementary robustness audit records its occurrence count, same-prefix probability shift, and student uncertainty. Let T_e denote the positions matched to e. We define

$$\Delta p(e) = \mathbb{E}_{t \in T_e}\left[p_t^{\mathrm{S}}(\hat{y}_t) - p_t^{\mathrm{T}}(\hat{y}_t)\right], \tag{8}$$

and

$$U^{\mathrm{S}}(e) = \mathbb{E}_{t \in T_e}\left[H(p_t^{\mathrm{S}})\right]. \tag{9}$$

The diagnostic core requires at least 50 occurrences, $\Delta p(e) > 0$, and $U^{\mathrm{S}}(e)$ above the median of frequency- and position-matched non-candidates. The audit uses held-out trajectories that do not overlap with model training; it is a robustness analysis of the manually reviewed inventory described in Appendix F. Spearman ρ is computed over candidates occurring at least 50 times in both settings; overlap is measured against the 41-expression DeepSeek-R1-Distill-Qwen-7B core.

Table 6: Transfer of the candidate ranking relative to the DeepSeek-R1-Distill-Qwen-7B mathematical diagnostic. The lower cross-task overlap argues for rebuilding the list rather than assuming a universal vocabulary.

Diagnostic setting	ρ	**Core overlap**	**Core count**
DeepSeek-R1-Distill-Qwen-1.5B (math)	0.86	0.80	39
DeepSeek-R1-Distill-Qwen-7B (math)	1.00	1.00	41
DeepSeek-R1-Distill-Llama-8B (math)	0.78	0.71	38
OpenMath2-Llama3.1-8B (math)	0.71	0.63	36
DeepSeek-R1-Distill-Qwen-7B (SciKnowEval L3)	0.41	0.34	19

Table 7: Full-list versus core-list sensitivity for DeepSeek-R1-Distill-Qwen-7B. The audited core retains most of the downstream score of the full list used for the main results.

Mask list	**AIME24**	**AIME25**	**MATH-500**
Full list (primary matcher)	49.0	39.8	89.0
Diagnostic core list	48.7	39.4	88.6

For the supplementary candidate audit and transfer diagnostics, a normalized multi-token expression is represented by its complete tokenizer-ID sequence and is counted only when that sequence occurs contiguously; component IDs do not independently trigger these diagnostic matches. This sequence-level diagnostic rule is distinct from the single-token content-word matcher used for the training results in Table 2, described in Appendix F. For the full-list versus core-list training comparison, we convert the selected core expressions into a token-ID set using the same content-word extraction and single-token variant rules as for the full inventory. The comparison therefore varies the expression inventory while keeping the tokenizer conversion procedure fixed.

A.4 Cross-Task Protocol and Reporting

Table 8: Output and evaluation formats for the cross-task controls.

Setting	**Output format**	**Primary metric**
AIME24/25	Open-ended mathematical reasoning	avg@12
SciKnowEval L3	Open-ended scientific reasoning	normalized aggregate
ToolAlpaca	Full tool-use trajectory	overall task success
MASSIVE (English)	Constrained intent label	intent accuracy

All cross-task controls use DeepSeek-R1-Distill-Qwen-7B, the same reference-conditioned OPSD objective, and the same 100-update budget as the main experiments. Naive and masked conditions use three runs and are reported as mean $\pm$ standard deviation. In contrast, Table 2 reports point estimates without run-level uncertainty. AIME avg@12 already averages 12 sampled responses per problem and should not be confused with independent training runs.

B Training Dynamics and Equal-Budget Masks

The top panel of Figure 4 tracks the 1.5B model. Under naive OPSD, marker frequency and response length fall early and only partly recover; under masked self-distillation, both remain higher. Accuracy declines with length in the naive run, so this is not successful reasoning compression in the present setup.

The bottom panel compares equal-count random masking and a highest-student-entropy mask under the same budget. Random masking stays close to naive OPSD, arguing against the explanation that the gain comes from supervising fewer positions alone. Entropy masking stays near the base

checkpoint but is weaker than the marker mask. One possible explanation is that entropy selects both marker positions and decision points where teacher information is useful; the experiment does not isolate that interpretation.

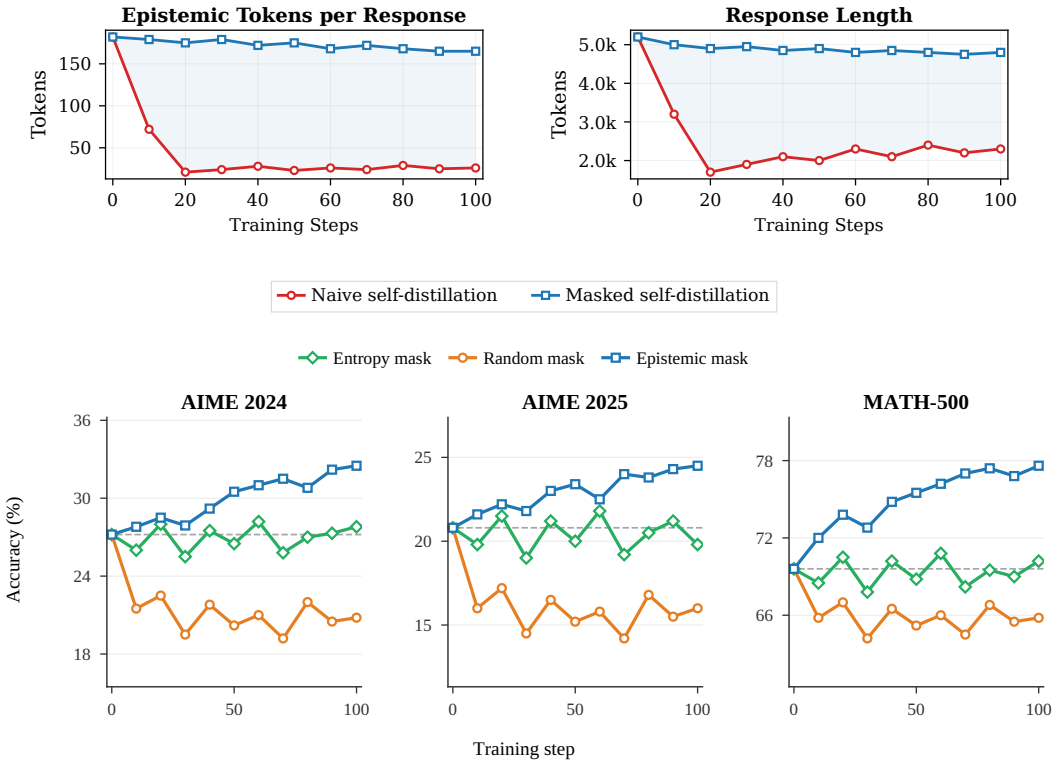


Figure 4: Training behavior and mask selectivity for the 1.5B model. Top: marker-frequency and response-length trajectories. Bottom: equal-budget entropy, random, and marker masks on all three benchmarks.

C Training Details

Table 9 summarizes the main optimization and generation settings used in our experiments. The checkpoints are `deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B`, `deepseek-ai/DeepSeek-R1-Distill-Qwen-7B`, and `nvidia/OpenMath2-Llama3.1-8B`. The self-distillation column applies to both naive OPSD and masked self-distillation. The two variants differ only in whether marker-matched positions are masked from the distillation loss.

Table 9: Training hyperparameters for all methods. "–" indicates that a setting is not applicable.

Hyperparameter	Self-Distillation	SFT	GRPO
Optimization			
Optimizer	AdamW	AdamW	AdamW
Learning rate	5×10^{-6}	5×10^{-6}	5×10^{-6}
LR schedule	constant	constant	constant
Warmup ratio	0.03	0.03	0.03
Weight decay	0.01	0.01	0.01
Gradient clip norm	1.0	1.0	1.0
Training			
Training steps	100	100	100
Global batch size	32	32	32
Mini batch per GPU	8	8	8
Gradient accumulation	1	1	1
Rollouts per problem	1	1	8
Self-Distillation Specific			
Per-vocabulary bound τ	5.0	–	–
GRPO Specific			
Clip ε	–	–	0.2
KL penalty coefficient	–	–	0.04
Advantage normalization	–	–	group
Reward normalization	–	–	enabled
Generation (training)			
Temperature	1.0	–	1.0
Top-p	0.9	–	0.9
Repetition penalty	1.0	–	1.0
Max prompt length	8,192	8,192	8,192
Max completion length	16,384	16,384	16,384
Generation (evaluation)			
Temperature	0.6	0.6	0.6
Top-p	0.95	0.95	0.95
Repetition penalty	1.0	1.0	1.0
Max prompt length	8,192	8,192	8,192
Max completion length	16,384	16,384	16,384

D PROMPT TEMPLATES

Both student and teacher share the same system message:

System message

```
You are a helpful assistant.
```

Student prompt

```
Problem: {question}
Please reason step by step, and put your final answer within \boxed{}.
```

Teacher prompt

```
Problem: {question}
Below is a verified reference solution. Use it to reason about the problem.
Your own response wording may differ.
{reference_solution}
Please reason step by step, and put your final answer within \boxed{}.
```

Teacher and student are initialized from the same checkpoint. The teacher remains fixed during training and receives the reference solution as privileged context, while the student is updated and sees only the problem. During training, the teacher does not generate a new trajectory; its stopped-gradient next-token distribution is computed on the student response through a forward pass with the extra context. The student minimizes the component-bounded loss in Equation 6, except at positions removed by Equation 5. The teacher conditions on the `COT_Reason` field of OpenThoughts (Guha et al., 2025).

The teacher prompt includes "Your own response wording may differ" to reduce pressure to copy the reference verbatim. Removing this phrase had negligible effect in a diagnostic probe, suggesting that the observed marker-probability gap is not explained by this wording alone.

E SUPERVISED FINE-TUNING (SFT)

Supervised Fine-Tuning (SFT) trains the model to predict a target response given a problem prompt using a standard next-token cross-entropy loss. Given a dataset of pairs $\{(x_i, y_i)\}$, where x_i is a math problem and y_i is the corresponding full reasoning trace, SFT minimizes:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\log \pi_\theta(y_{i,t} \mid x_i, y_{i,<t}).$$

The loss is computed only on response tokens. Prompt tokens are masked. We use the same training data as self-distillation (the `COT_Reason` field of OpenThoughts dataset), with 100 steps, learning rate 5×10^{-6}, and batch size 32. Training hyperparameters are listed in Table 9.

F MARKER EXPRESSION INVENTORY AND TOKENIZER MATCHING

Inventory construction. We select 500 OpenThoughts mathematics problems as a separate inventory-construction subset and check them for overlap with the final evaluation problems. The initial DeepSeek-R1-Distill-Qwen-7B checkpoint generates two responses per problem, yielding 1,000 reasoning trajectories. Both correct and incorrect responses are retained. These trajectories are separate from those used for the subsequent candidate audit in Appendix A.

We use Qwen3-32B to extract candidate expressions of uncertainty, reconsideration, self-checking, and error recognition from these trajectories, followed by manual review and filtering. The resulting

expressions form the source inventory listed below. This construction procedure is distinct from the statistical selection of the smaller diagnostic core in Appendix A; the main DeepSeek-R1-Distill-Qwen experiments use the full inventory.

Tokenizer-level matching. The inventory below is the semantic source list used to construct the matcher; it is not a list of phrase-level patterns applied directly during training. We lowercase each expression, split it into words, remove common function words, and enumerate capitalization, leading/trailing whitespace, comma, and period variants of each remaining content word. A variant enters $\mathcal{I}$ only if the model tokenizer encodes the complete variant as exactly one non-special token. Multi-token encodings and their component IDs are not added. Training then uses exact token-ID membership, $M_t = \mathbf{1}[\hat{y}_t \notin \mathcal{I}]$. This description follows the implementation used for the reported results.

Source Expression Inventory ($\mathcal{E}$)

Uncertainty / Hedging

wait, hmm, hmmm, perhaps, maybe, seems, appears, might, likely, possibly, probably, roughly, approximately, presumably, I think, I believe, I suppose, I assume, assuming, not sure, not entirely sure, not necessarily, unclear, uncertain, indeed, I doubt, suspicious.

Reconsideration / Revision

actually, actually no, or rather, or maybe, so actually, instead, well, on second thought, upon reflection, that said, having said that, but then, although, however, what if, could it be that, in other words, more precisely, which means, go back, try again.

Self-Checking / Verification

check, double-check, cross-check, let me check, let me verify, let me confirm, let me see, let me try, let me compute, let's see, verify, confirm, validate, make sure, ensure, did I, is that right, is this correct, does this make sense, careful, I need to be careful, just to be sure.

Error Recognition / Correction

hold on, but wait, no wait, oh, oh no, not quite, not exactly, not right, that's not right, that can't be, something's off, doesn't work, mistake, I made a mistake, I just realized, incorrect, error, redo, recalculate, let me correct, let me redo, scratch that, adjust, I mean, actually I think.

G Qualitative Case Study: Contrasting Self-Correction Behaviors

We compare outputs from three training conditions based on DeepSeek-R1-Distill-Qwen-1.5B: the pretrained model (step 0), naive OPSD (100 steps), and masked self-distillation (100 steps). Outputs are lightly condensed for presentation. This example illustrates the behavioral contrast but is not used as aggregate or causal evidence. Epistemic markers are highlighted in orange; not every highlighted phrase is necessarily a member of the single-token training matcher. Errors are shown in red.

Evaluation prompt (MATH-500)

Let $f(x) = (\arccos x)^2 + (\arcsin x)^2$. Find the range of $f(x)$. All functions are in radians.

Base model: reflective reasoning with self-correction

Let's solve this. I remember that $\arccos x + \arcsin x = \pi/2$ for any $x \in [-1, 1]$. Let $\theta = \arccos x$. Then $\arcsin x = \pi/2 - \theta$, so $f(x) = \theta^2 + (\pi/2 - \theta)^2 = 2\theta^2 - \pi\theta + (\pi/2)^2$.
Since $\theta = \arccos x$ and $x \in [-1, 1]$, θ ranges over $[0, \pi]$. This is a quadratic opening upwards; the vertex at $\theta = \pi/4$ gives the minimum.
Wait, let me verify the identity: at $x = 0$, $\arccos 0 = \pi/2$, $\arcsin 0 = 0$, sum $\pi/2$. At $x = 1$, $\arccos 1 = 0$, $\arcsin 1 = \pi/2$. OK, correct.
Minimum at $\theta = \pi/4$: $f = 2(\pi^2/16) - \pi^2/4 + \pi^2/4 = \pi^2/8$.
Endpoints: $\theta = 0$ (i.e., $x = 1$): $f = 0 + (\pi/2)^2 = \pi^2/4$. $\theta = \pi$ (i.e., $x = -1$): $f = 2\pi^2 - \pi^2 + \pi^2/4 = \pi^2 + \pi^2/4 = 3\pi^2/4$.
Wait, that's not right – π^2 is $4\pi^2/4$, so adding $\pi^2/4$ gives $5\pi^2/4$, not $3\pi^2/4$. I mistakenly combined the numerators without a common denominator. Correcting: $f(\pi) = 5\pi^2/4$.
Let me double-check by computing directly: $f(-1) = (\arccos(-1))^2 + (\arcsin(-1))^2 = \pi^2 + (-\pi/2)^2 = \pi^2 + \pi^2/4 = 5\pi^2/4$. Good, that matches.
Let me also find critical points via derivative: $f'(x) = \frac{2}{\sqrt{1-x^2}}[\pi/2 - 2\arccos x]$.
Setting to zero gives $\arccos x = \pi/4$, i.e., $x = \sqrt{2}/2$–the same vertex.
Hmm, maybe I should verify with the Extreme Value Theorem: f is continuous on $[-1, 1]$, so it attains max and min at critical points or endpoints. The only critical point is at $\sqrt{2}/2$ (minimum), and the maximum at $x = -1$ gives $5\pi^2/4$.
Thus range: $\boxed{\left[\frac{\pi^2}{8}, \frac{5\pi^2}{4}\right]}$.

Naive OPSD: confident error without verification

I need the range of $f(x) = (\arccos x)^2 + (\arcsin x)^2$. Hmm, let me think.
First, $\arccos x + \arcsin x = \pi/2$. Let $a = \arccos x$, $b = \arcsin x$. Then $b = \pi/2 - a$. $f(x) = a^2 + (\pi/2 - a)^2 = 2a^2 - \pi a + \pi^2/4$.
Since $a = \arccos x$ with $x \in [-1, 1]$, $a \in [0, \pi/2]$.
Vertex at $a = \pi/4$: $f = \pi^2/8$ (minimum). Endpoints: $a = 0$ gives $\pi^2/4$; $a = \pi/2$ also gives $\pi^2/4$.
Therefore, the range of $f(x)$ is $\boxed{\left[\frac{\pi^2}{8}, \frac{\pi^2}{4}\right]}$.

Analysis
The naive-OPSD model proceeds directly from setup to conclusion with no reflective pauses. The domain of $\arccos x$ is incorrectly truncated to $[0, \pi/2]$ (the correct range is $[0, \pi]$). Only two endpoints are evaluated—both yielding $\pi^2/4$—and the model immediately concludes the maximum is $\pi^2/4$, never examining $x = -1$ where the true maximum $5\pi^2/4$ occurs. The sole "Hmm" at the start is the only highlighted marker in the entire response: no "Wait", no "let me check", no self-questioning of the domain assumption. In contrast to the base model—which pauses to verify the identity, catches an arithmetic slip, differentiates to confirm the critical point, and invokes the Extreme Value Theorem—the naive-OPSD model simply executes the derivation and stops.

Masked self-distillation: verification and self-correction

I need the range of $f(x) = (\arccos x)^2 + (\arcsin x)^2$. Hmm, let me think.
First, I remember that for any $x \in [-1, 1]$, $\arccos x + \arcsin x = \pi/2$. That seems important. Let $a = \arccos x$ and $b = \arcsin x$. Then $a + b = \pi/2$, so $b = \pi/2 - a$.
Substituting: $f(x) = a^2 + (\pi/2 - a)^2 = a^2 + (\pi^2/4 - \pi a + a^2) = 2a^2 - \pi a + \pi^2/4$.
This is a quadratic in a. Since $a = \arccos x$ and $x \in [-1, 1]$, when $x = 1$ we have $a = 0$; when $x = -1$ we have $a = \pi$. So $a \in [0, \pi]$.
The coefficient of a^2 is positive, so the parabola opens upward—minimum at the vertex, maximum at an endpoint. Vertex: $a = \pi/4$, which is inside $[0, \pi]$.
Now evaluate at the three candidate points:

- $a = 0$: $f = 0 - 0 + \pi^2/4 = \pi^2/4$.
- $a = \pi/4$: $f = 2(\pi^2/16) - \pi^2/4 + \pi^2/4 = \pi^2/8$.
- $a = \pi$: $f = 2\pi^2 - \pi^2 = \pi^2$, so $f = \pi^2/4$.

Wait, hold on – I dropped the $+\pi^2/4$ term! $f(\pi) = 2\pi^2 - \pi^2 + \pi^2/4 = \pi^2 + \pi^2/4 = 5\pi^2/4$, not $\pi^2/4$. That was a careless omission.
Let me verify with the original function directly: at $x = 1$, $f = 0 + (\pi/2)^2 = \pi^2/4$. At $x = -1$, $f = \pi^2 + (-\pi/2)^2 = \pi^2 + \pi^2/4 = 5\pi^2/4$. All consistent.
I considered the identity, transformed the function, checked the vertex and all three boundary candidates, caught a slip, and verified against the original. I'm confident the range is $[\pi^2/8, 5\pi^2/4]$.
Thus: $\boxed{\left[\frac{\pi^2}{8}, \frac{5\pi^2}{4}\right]}$.

Analysis
In this example, the masked-distilled model uses the correct domain $[0, \pi]$, checks all three candidate points, catches an arithmetic omission, and verifies the result directly. The contrast is consistent with the aggregate marker, length, and accuracy trends, but one selected trace cannot attribute the behavior to a particular surface expression or objective component by itself.

H GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is a reinforcement learning algorithm for fine-tuning policy LLMs π_θ without a separately learned value function. Its key idea is to normalize rewards within a group of responses sampled from the same prompt, thereby stabilizing optimization.

Given a query q, the old policy π_θ^{old} generates G candidate responses $\{o_1, \ldots, o_G\}$. Each response o_i is assigned a scalar reward $R(q, o_i)$. Group-wise z-score normalization is then applied to obtain an advantage shared across the tokens of the response:

$$\hat{A}_{i,t} = \frac{R(q, o_i) - \text{mean}(\{R(q, o_j)\}_{j=1}^G)}{\text{std}(\{R(q, o_j)\}_{j=1}^G) + \varepsilon}.$$

Let $r_\theta(o_{i,t}) = \pi_\theta(o_{i,t} \mid q, o_{i,<t}) \,/\, \pi_\theta^{\text{old}}(o_{i,t} \mid q, o_{i,<t})$. The policy is updated with a clipped surrogate objective, similar to PPO, combined with a KL-divergence penalty to constrain policy drift:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_\theta(o_{i,t})\,\hat{A}_{i,t},\ \text{clip}\big(r_\theta(o_{i,t}),\,1{-}\varepsilon,\,1{+}\varepsilon\big)\,\hat{A}_{i,t}\Big){+}\beta\,\text{KL}(\pi_\theta \,\|\, \pi_{\text{ref}}).$$

Unlike self-distillation, GRPO uses no teacher-provided token distributions; in our setup, it derives the policy update from a scalar reward per response. Training hyperparameters are listed in Table 9.

I Qualitative Teacher Generation under a Full Reasoning Trace

To illustrate how conditioning can accompany a change in generation style, we compare generations of DeepSeek-R1-Distill-Qwen-1.5B on the same MATH-500 problem used in Appendix G: (1) **student** mode with no privileged information and (2) **teacher** mode conditioned on a full reasoning trace. This example is descriptive; the same-prefix aggregate probe in Section 3 supplies the quantitative evidence. The full reasoning trace provided to the teacher is the base model's reflective response from Appendix G (shown below as the reference for comparison). For the teacher prompt, this trace is presented via the standard teacher template (Appendix D) as a sample solution.

Evaluation prompt (MATH-500)

Let $f(x) = (\arccos x)^2 + (\arcsin x)^2$. Find the range of $f(x)$. All functions are in radians.

Reference: full reasoning trace (fed to teacher)

Let's solve this. I remember that $\arccos x + \arcsin x = \pi/2$ for any $x \in [-1, 1]$. Let $\theta = \arccos x$. Then $\arcsin x = \pi/2 - \theta$, so $f(x) = \theta^2 + (\pi/2 - \theta)^2 = 2\theta^2 - \pi\theta + (\pi/2)^2$.
Since $\theta = \arccos x$ and $x \in [-1, 1]$, θ ranges over $[0, \pi]$. This is a quadratic opening upwards; the vertex at $\theta = \pi/4$ gives the minimum.
Wait, let me verify the identity: at $x = 0$, $\arccos 0 = \pi/2$, $\arcsin 0 = 0$, sum $\pi/2$. At $x = 1$, $\arccos 1 = 0$, $\arcsin 1 = \pi/2$. OK, correct.
Endpoints: $\theta = 0$ (i.e., $x = 1$): $f = 0 + (\pi/2)^2 = \pi^2/4$. $\theta = \pi$ (i.e., $x = -1$): $f = 2\pi^2 - \pi^2 + \pi^2/4 = \pi^2 + \pi^2/4 = 3\pi^2/4$.
Wait, that's not right – π^2 is $4\pi^2/4$, so adding $\pi^2/4$ gives $5\pi^2/4$, not $3\pi^2/4$. I mistakenly combined the numerators without a common denominator. Correcting: $f(\pi) = 5\pi^2/4$.
Let me double-check by computing directly: $f(-1) = (\arccos(-1))^2 + (\arcsin(-1))^2 = \pi^2 + (-\pi/2)^2 = \pi^2 + \pi^2/4 = 5\pi^2/4$. Good, that matches.
Let me also find critical points via derivative: $f'(x) = \frac{2}{\sqrt{1-x^2}}[\pi/2 - 2\arccos x]$.
Setting to zero gives $\arccos x = \pi/4$, i.e., $x = \sqrt{2}/2$–the same vertex.
Hmm, maybe I should verify with the Extreme Value Theorem: f is continuous on $[-1, 1]$, so it attains max and min at critical points or endpoints. The only critical point is at $\sqrt{2}/2$ (minimum), and the maximum at $x = -1$ gives $5\pi^2/4$.
Thus range: $\boxed{\left[\frac{\pi^2}{8}, \frac{5\pi^2}{4}\right]}$.

Student generation (no privileged information)

Let's find the range of $f(x) = (\arccos x)^2 + (\arcsin x)^2$. Hmm, I remember that $\arccos x + \arcsin x = \pi/2$ for any $x \in [-1, 1]$.
Let me verify the identity: at $x = 0$, $\arccos 0 = \pi/2$ and $\arcsin 0 = 0$, sum is $\pi/2$. At $x = 1$, $\arccos 1 = 0$ and $\arcsin 1 = \pi/2$, sum is $\pi/2$. At $x = -1$, $\arccos(-1) = \pi$ and $\arcsin(-1) = -\pi/2$, sum is $\pi/2$. Okay, that identity holds.
Let $\theta = \arccos x$, then $\arcsin x = \pi/2 - \theta$. Substituting: $f(x) = \theta^2 + (\pi/2 - \theta)^2 = 2\theta^2 - \pi\theta + \pi^2/4$.
Since $\theta = \arccos x$ with $x \in [-1, 1]$, θ ranges over $[0, \pi]$. This is a quadratic with positive leading coefficient, so the vertex at $\theta = \pi/4$ is a minimum, giving $f_{\min} = \pi^2/8$.
Now, for the maximum I need to check the endpoints of $[0, \pi]$. At $\theta = 0$: $f = \pi^2/4$. At $\theta = \pi$: $f = 2\pi^2 - \pi^2 + \pi^2/4 = 5\pi^2/4$. Hmm, the function is continuous on a closed interval, so by the Extreme Value Theorem it attains all values between min and max.
Let me also verify with the original function directly: $f(1) = (\arccos 1)^2 + (\arcsin 1)^2 = 0 + (\pi/2)^2 = \pi^2/4$. $f(-1) = (\arccos(-1))^2 + (\arcsin(-1))^2 = \pi^2 + (-\pi/2)^2 = 5\pi^2/4$.
Wait, I should also check the derivative to confirm there are no other critical points: $f'(x) = \frac{2}{\sqrt{1-x^2}}[\pi/2 - 2\arccos x]$. Setting to zero gives $\arccos x = \pi/4$, which is the same vertex. Good, only one critical point.
Thus range: $\boxed{\left[\frac{\pi^2}{8}, \frac{5\pi^2}{4}\right]}$.

Teacher generation (conditioned on the full reasoning trace)

I need to find the range of $f(x) = (\arccos x)^2 + (\arcsin x)^2$. I remember that $\arccos x + \arcsin x = \pi/2$ for any $x \in [-1, 1]$. That seems useful.
Substituting $\theta = \arccos x$, we have $\arcsin x = \pi/2 - \theta$, so: $f(x) = \theta^2 + (\pi/2 - \theta)^2 = 2\theta^2 - \pi\theta + \pi^2/4$.
Now $\theta = \arccos x$ ranges over $[0, \pi]$. This is a quadratic with positive leading coefficient, so the vertex at $\theta = \pi/4$ gives the minimum: $f_{\min} = \pi^2/8$.
For the maximum, since the parabola opens upwards, it occurs at an endpoint. At $\theta = 0$: $f = \pi^2/4$. At $\theta = \pi$: $f = 2\pi^2 - \pi^2 + \pi^2/4 = 5\pi^2/4$.
Wait, let me double-check the endpoints. At $x = 1$, $\arccos 1 = 0$ and $\arcsin 1 = \pi/2$, so $f(1) = \pi^2/4$. At $x = -1$, $\arccos(-1) = \pi$ and $\arcsin(-1) = -\pi/2$, so $f(-1) = 5\pi^2/4$. That checks out.
Thus range: $\boxed{\left[\frac{\pi^2}{8}, \frac{5\pi^2}{4}\right]}$.

Analysis
The unprivileged generation contains more highlighted discourse markers and uses identity verification, derivative checking, and direct evaluation. The reference-conditioned generation is shorter and uses only one explicit "Wait" before checking the endpoints. This selected contrast is consistent with Table 4: access to a full trace is accompanied by more direct output and fewer explicit markers.

J Full Training Curves

Figure 5 shows the checkpoint trajectories underlying the final values in Table 2.

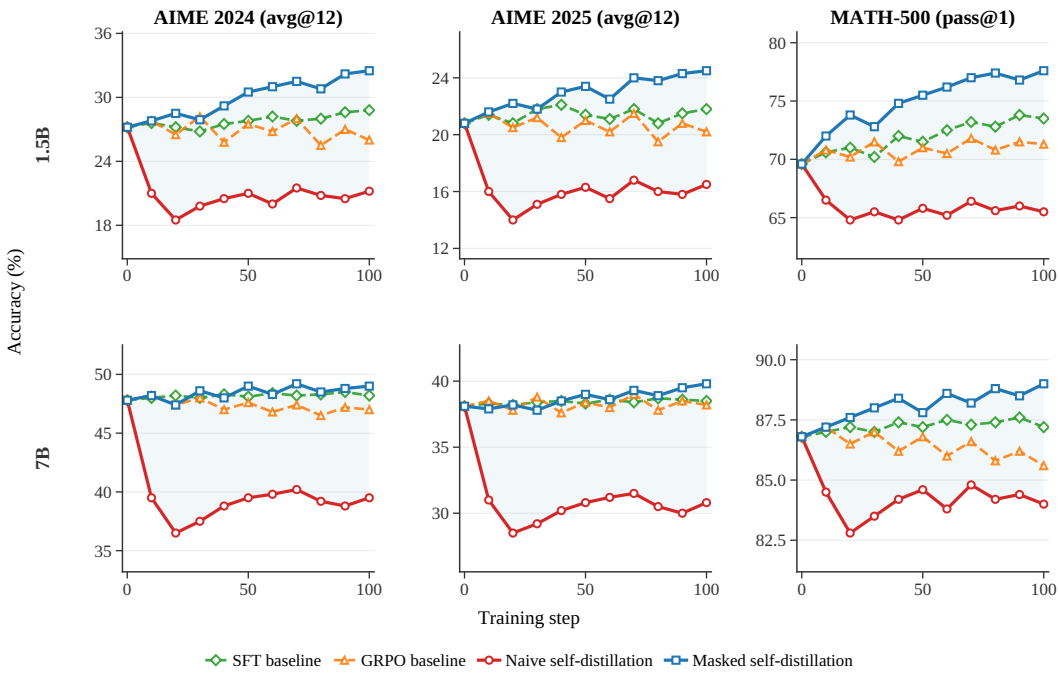


Figure 5: Checkpoint trajectories for the 1.5B (top) and 7B (bottom) models. Columns show AIME24, AIME25, and MATH-500; all panels share one legend.

K Gradient Intuition for the Unclipped Objective

The implementation uses the component-wise bound in Equation 6; its gradient is piecewise because bounded components stop contributing above τ. The standard, unclipped KL gives a useful local intuition without making claims about full parameter-gradient norms. Let $z^{\mathrm{S}}_{t,w}$ be the student logit for vocabulary item w. Dropping the teacher-entropy constant, the position-level cross-entropy is

$$\tilde{\ell}_t = -\sum_{v \in \mathcal{V}} p^T_{t,v} \log p^S_{t,v} \tag{10}$$

and its logit gradient is

$$\frac{\partial \tilde{\ell}_t}{\partial z^S_{t,w}} = p^S_{t,w} - p^T_{t,w} \tag{11}$$

For the sampled token $\hat{y}_t$, if $p^{\mathrm{S}}_{t,\hat{y}_t} > p^{\mathrm{T}}_{t,\hat{y}_t}$, gradient descent reduces its student logit relative to teacher-favored alternatives. The same-prefix values in Table 1 show that this gap is much larger on average at marker-matched positions. Masking sets the entire position-level contribution to zero. This argument explains the direction of the local pressure; it does not imply that the sampled token dominates the parameter gradient, that marker positions have a particular gradient norm, or that their surface forms uniquely determine subsequent reasoning.