

A Privileged Teacher That Only Scores Cannot Search: Critique-Conditioned On-Policy Self-Distillation for Reasoning

Mayug Maniparambil^{1*} Ayush Rai² Ekansh Verma¹ Noel O'Connor² Fergal Reid¹
¹Fin.ai ²Dublin City University

Abstract

On-policy self-distillation promises to replace the sparse outcome signal of verifiable-reward RL with dense token-level credit: a teacher conditioned on the reference solution scores the student’s own rollout, and the resulting distributional shift is distilled back into the policy. We investigate whether this scoring signal actually contains corrective credit. Using gold error spans in student rollouts, we test the two signals consumed by self-distillation objectives, the full-distribution KL and the token log-ratio, for two properties: whether they localize errors and whether they push probability away from erroneous student tokens. They exhibit neither of these properties nor can they reliably distinguish correct from incorrect rollouts within the same problem. We argue that this failure is structural: locating an error in a reasoning trace is itself a search problem, while a scoring teacher is never explicitly asked to perform that search. The same model, with the same privileged information, localizes errors substantially better when allowed to generate a critique. This forms the motivation for *critique-conditioned self-distillation*: first generate a critique of the current rollout, then condition the scoring teacher on that critique. The resulting signal localizes errors, is corrective within error spans, and shifts probability toward tokens that initiate backtracking such as “Wait”, “But”, and “Actually”. We further show that the critique is useful both at test time, where resuming a wrong rollout with the critique in context converts math problems the model cannot solve by sampling, and in cold-start GRPO, where it provides an advantage signal on prompts for which verifiable rewards alone provide none.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) provides a reliable but sparse learning signal: a reasoning rollout is correct or incorrect, but the verifier does not identify which intermediate decisions caused the outcome. In group-based objectives, the limitation is sharper when all sampled rollouts for a problem fail, since the relative advantage can vanish entirely. These all-wrong groups are precisely where additional supervision is most useful, yet outcome-based RL provides little guidance about how the reasoning should change.

On-policy self-distillation (OPSD) offers a natural way to obtain denser supervision [Zhao et al., 2026a, Hübotter et al., 2026, Yang et al., 2026]. The model generates a rollout as a student and then scores the same rollout as a teacher with privileged information, such as a reference solution. The discrepancy between the student and teacher distributions provides a signal at every token, even when the final answer is wrong. But density alone does not imply credit assignment. For the signal to be corrective, it should be larger where the reasoning goes wrong and should push probability away from the student’s erroneous choices.

*Correspondence: mayugmaniparambil@gmail.com

We test these properties directly on annotated error spans in logic and competition-mathematics rollouts. Across tasks and teacher scales, reference-conditioned scoring fails both tests: its distributional shift is only weakly concentrated at errors, its token-level signal does not reliably suppress erroneous student tokens, and even when aggregated over the full trajectory it does not distinguish correct from incorrect rollouts of the same problem. The limitation is therefore not in how a particular loss consumes the signal. Scoring itself exposes little information about where the reasoning failed.

We argue that locating an error is itself a reasoning problem. A scoring teacher is given the reference but is never explicitly asked to compare it against the rollout and search for the mistake. When the same model is instead allowed to generate a critique, it can localize and explain the error. This motivates *critique-conditioned self-distillation*: first generate a localized critique of the student’s rollout, then condition the scoring teacher on that critique. Generation performs the search, and scoring converts the error it finds into a dense token-level signal.

Our contributions are threefold. We show that the privileged teacher’s signal is neither localized nor corrective at gold error spans, and is not predictive of rollout correctness even at the whole-trace level (§5). We show that the failure lies in scoring, not in the information. The same model, given the same reference, localizes errors when it generates a critique (§6). Finally, we introduce *critique-conditioned self-distillation*, which uses generated critiques as privileged context for dense token-level supervision, and provide initial results using this signal for test-time inference and training (§8, Appendix A).

2 Related Work

Privileged self-distillation. Our work sits between privileged self-distillation and error-localization methods. On-policy self-distillation uses a teacher conditioned on privileged information to provide dense token-level supervision over the student’s own rollout [Zhao et al., 2026a, Hübötter et al., 2026, Yang et al., 2026, Pan et al., 2026]. Recent work has identified several failure modes of this setup, including context-induced bias and weak or stylistic distributional shifts [Harne et al., 2026, Kaur et al., 2026, Pan et al., 2026]. We differ by measuring the resulting signal directly at gold error spans and asking whether it carries corrective credit.

Locating errors is a reasoning problem. A separate line of work shows that locating reasoning errors can itself require generation or explicit reasoning, motivating generative verifiers, process supervision, and critique-based reward models [Tyen et al., 2024, Zhang et al., 2025a, Ankner et al., 2024, Khalifa et al., 2025]. Extended related work appears in Appendix C.

3 Preliminaries

3.1 Group-relative RL and the all-wrong group

For a prompt x the policy π_θ samples a group of G rollouts $y^1, \dots, y^G$, a verifier scores each one $r_i \in \{0, 1\}$, and every token of rollout i receives the group-normalised advantage $A_i = (r_i - \bar{r})/\sigma_r$ [Shao et al., 2024]. The signal is one bit per rollout, and it vanishes entirely when the group is unanimous: if all G rollouts are wrong, $A_i = 0$ for every i and the group contributes no gradient. These *all-wrong* groups are the problems at the frontier of the model’s competence—the ones it cannot yet solve once in G tries—and they are the reason verifiable-reward RL tends to sharpen what the base model can already do rather than extend it [Yue et al., 2025].

3.2 On-policy self-distillation

Let $P_S(\cdot \mid x, y_{<t})$ denote the student’s next-token distribution at a prefix of its own rollout, and let $P_T(\cdot \mid x, r, y_{<t})$ denote the same network evaluated at that prefix with privileged information r —here, the reference solution—in context. Because the teacher evaluates a rollout it did not generate, we call this *scoring*. Scoring yields a dense signal: at each position, the discrepancy between P_T and P_S is interpreted as token-level credit.

Two forms of that discrepancy are in use. OPSD distills the full distribution,

$$\mathcal{L}_{\text{KL}} = \sum_t \text{KL}(P_T(\cdot \mid x, r, y_{<t}) \parallel P_S(\cdot \mid x, y_{<t})), \quad (1)$$

so its per-token signal is the KL divergence at each position [Zhao et al., 2026a, Hübotter et al., 2026, Kaur et al., 2026]. RLSD [Yang et al., 2026] instead scores the token the student wrote,

$$\Delta_t = \log P_T(y_t \mid x, r, y_{<t}) - \log P_S(y_t \mid x, y_{<t}), \quad (2)$$

and uses Δ_t to modulate the group advantage token by token. The verifier therefore determines the update direction, while the privileged teacher modulates its magnitude. RLCSO [Pan et al., 2026] keeps this form but replaces the reference by sibling rollouts from the same group, contrasting a correct sibling y^+ against a wrong one y^- , $e_t = \log P_T(y_t \mid x, y^+, y_{<t}) - \log P_T(y_t \mid x, y^-, y_{<t})$, thereby removing shifts induced by privileged context alone. Neither construction resolves the all-wrong regime: there is no correct sibling, and group-relative advantage vanishes. We therefore focus on reference-conditioned scoring and ask a more basic question: does $P_T(\cdot \mid x, r, y_{<t})$ expose useful credit at the positions where the student’s reasoning goes wrong?

3.3 A general form of the probe

Fix a rollout y and let b denote privileged context prepended to the teacher prompt. Scoring the same rollout with and without b gives two per-token signals:

$$\text{KL}_t(b) = \text{KL}(P_T(\cdot \mid x, b, y_{<t}) \parallel P_T(\cdot \mid x, y_{<t})), \quad (3)$$

$$\Delta_t(b) = \log P_T(y_t \mid x, b, y_{<t}) - \log P_T(y_t \mid x, y_{<t}). \quad (4)$$

When $b = r$, these recover the two OPSD readouts of Eqs. 1 and 2, since the unprivileged teacher is the student.

The two signals probe different properties of the same scoring pass. $\text{KL}_t(b)$ measures how strongly the privileged context changes the teacher’s distribution at position t . If that context identifies where the rollout fails, this shift should concentrate at the error spans, so we call it the *localization* signal. $\Delta_t(b)$ measures how the context changes the probability of the token the student actually wrote. At an erroneous token, a corrective signal should reduce that probability, so we call Δ_t the *correction* signal and read negative values as corrective. We evaluate both signals at gold error spans in §4.

To isolate the information carried by privileged context, we pair each context b with a matched placebo $\tilde{b}$ and report the content signals $\text{KL}_t(b) - \text{KL}_t(\tilde{b})$ and $\Delta_t(b) - \Delta_t(\tilde{b})$. This controls for distributional shifts caused by the presence of a solution-shaped preamble alone. For the reference condition, $b = r$ and $\tilde{b} = r'$ is a matched-length reference from another problem. For the critique condition, $b = c(y)$ and $\tilde{b} = c(y')$ is a critique of another problem’s rollout.

The critique is generated from the problem, reference, and student rollout as $c(y) = \{(q_j, \rho_j)\}_j$, where q_j is a verbatim quote of an erroneous sentence and ρ_j explains the error. *Critique-conditioned self-distillation* then replaces the reference in OPSD with $c(y)$: generation performs the search for the error, and scoring converts the resulting critique into a dense signal over the rollout. We ablate the quote and explanation separately in §7.

4 Setup

4.1 Models, data, and labels

Models and tasks. We conduct the token credit signal analysis on the Qwen3-1.7B [Yang et al., 2025], Olmo-3-7B-Think and Olmo-3-32B-Think [Olmo et al., 2026] rollouts. In each case the teacher is the same model with privileged context, the on-policy self-distillation setting. For Qwen3-1.7B we also run Qwen3-8B as a scale-control teacher. The tasks are Knights & Knaves logic puzzles [Xie et al., 2025a,b] (K&K, 4–8 people) and DeepMath [He et al., 2025b] at two difficulty bands (L5-7, L7-10).

Rollouts, references, and gold error spans. We study wrong rollouts: verifier-incorrect, not truncated, and ending in a boxed answer. We keep the verifier-correct rollouts of the same problems for the correctness test in §5. The privileged reference is a verified step-by-step derivation for K&K and the dataset’s reference solution plus answer for DeepMath. We obtain gold error spans from a grader model that reads the problem, the rollout and the reference and flags each error as a verbatim quote of one sentence plus a reason. The grader is a DeepSeek model [DeepSeek-AI et al., 2026a,b]

for the Qwen rollouts and claude-sonnet-5 [Anthropic, 2026] for the Olmo-3 rollouts. We then have gpt-5.2 [OpenAI, 2025a] and claude-sonnet-5 judge every flag independently, and we count a flag as gold when both accept it. To measure how lenient the judges are, we replace a quarter of the flags sent to each judge with position-matched control sentences carrying the same reason and report their acceptance rate with each study (Appendix D). We also perform the same signal analysis on human-annotated error spans from ProcessBench (Appendix J).

Critics. We use three critics to write the critiques $c(y)$ of §3.3. All use the same prompt: the rollout is presented sentence-numbered, each error is a verbatim quote plus a reason, and recall is favoured over precision (Appendix D.1). All receive the same reference r that the teacher receives in the reference condition. We use gpt-5.2 as the oracle critic. We also test whether the student can identify its own errors by having Qwen3-1.7B critique its own rollouts. We call this the self-critique. To assess whether a larger critic is more effective, we have Qwen3-8B perform the critiques as well. Appendix B shows an oracle critique and a self-critique of the same rollout on each task. Because critic and teacher see the same information, the comparison in §6 is between generating and scoring, not between amounts of information. The Olmo-3 models each critique their own rollouts (Table 1, Appendix I).

4.2 The error-span Z readout

Per-token signals vary substantially across positions and rollouts, so we normalise each rollout against its own clean regions, following baseline-normalisation conventions in event-related neurophysiology [Mercier et al., 2022]. For each rollout, we measure the signal in a short window immediately after the first gold error span and compare it with length-matched windows after clean spans in the prefix. We use the first error only because the reasoning after it is often wrong in ways the gold spans do not mark. If $\bar{s}_W$ denotes the mean signal in window W , we define

$$Z_r = \frac{\bar{s}_{W_{\text{err}}} - \text{mean}_k \bar{s}_{W_k}}{\text{sd}_k \bar{s}_{W_k}},$$

which we call the *error-span Z*. It measures how far the signal at the error departs from the rollout’s own baseline, in units comparable across signals, teachers, and tasks.

A positive localization score means that the teacher’s distribution changes more at the error than at clean regions while a negative correction score means that privileged context lowers the probability of the student’s erroneous token. For localization, we log-transform the window-mean KL before computing Z_r to reduce the effect of its heavy tail. We report the mean Z_r across rollouts, calibrated by subtracting the score obtained when a random clean sentence from the same rollout is treated as the error. Thus, a signal with no error-specific content is centered at zero. Window sizes and other implementation details are given in Appendix D.

5 The Privileged Signal Is Empty Where It Matters

We report the localization and correction signals of §3.3 at the positions of the student’s rollouts known to be wrong. Figure 1 shows both signals around the gold error sentences (blue), aligned to sentence onset and to sentence end and averaged over all error sentences of all rollouts, against the same average over the clean sentences that precede the first error (orange). Table 1 reports the error-span Z of §4.2. A privileged teacher provides usable token-level credit at an error if its error-span Z is positive for localization and negative for correction. We apply the same criterion to the reference here and to the critique in §7.

Not localized at the errors. On K&K, error and control traces are nearly indistinguishable for both localization and correction. On DeepMath, the reference induces a small error-specific shift, most visible near the sentence boundary, but the effect is much weaker than under critique conditioning (Figure 3). Thus, the reference does affect the teacher’s distribution, but not selectively at the student’s errors.

Table 1 shows the same comparison at the rollout level. Across students, tasks, and both signals, reference-conditioned scores remain close to the rollout’s clean baseline, and no setting satisfies the corrective criterion. When the reference signal does depart from zero, it can even point in the wrong

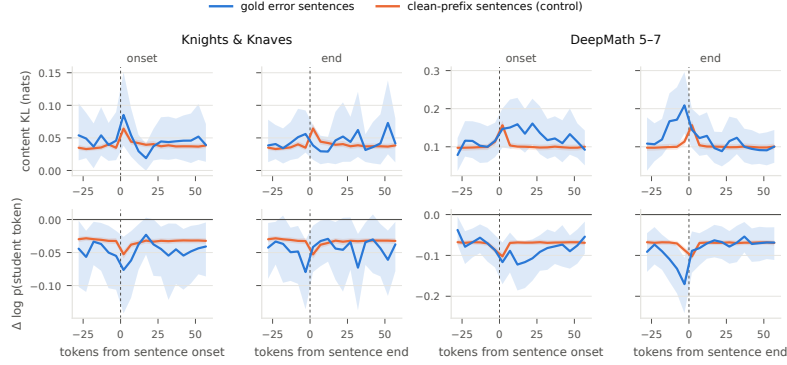


Figure 1: **The reference-conditioned teacher barely reacts at the student’s errors.** Content KL (top) and $\Delta \log p$ of the student’s token (bottom) around gold error sentences (blue) and clean sentences (orange), aligned to sentence onset and end; Qwen3-1.7B scoring its own rollouts. Bands: 95% CI.

Table 1: **Error-span Z at the first gold error, for three students each scoring its own rollouts.** Localization: log content KL (positive = reacts at the error); correction: $\Delta \log p$ of the student’s token (negative = corrective). Bold: 95% CI excludes 0.

conditioning	K&K		L5-7		L7-10	
	loc	corr	loc	corr	loc	corr
<i>Qwen3-1.7B</i> (355 / 516 / 794 problems)						
reference	-0.17	+0.11	+0.14	-0.14	+0.07	+0.17
critique, oracle	+0.75	-0.62	+0.99	-0.90	+0.78	-0.54
critique, self	+0.16	-0.07	+0.25	-0.14	+0.18	-0.07
<i>Olmo-3-7B-Think</i> (197 / 398 / 631 problems)						
reference	+0.20	-0.00	+0.11	-0.08	-0.01	-0.03
critique, oracle	+0.63	-0.29	+0.89	-0.59	+0.88	-0.49
critique, self	+0.13	-0.14	+0.41	-0.18	+0.33	-0.14
<i>Olmo-3-32B-Think</i> (114 / 111 / 133 problems)						
reference	+0.09	-0.41	+0.21	-0.09	+0.15	-0.06
critique, oracle	+0.76	-0.43	+0.96	-0.50	+1.20	-0.62
critique, self	+0.47	-0.01	+0.54	-0.39	+0.40	-0.42

direction: on L7-10, the Qwen teacher assigns *more* probability to the student’s erroneous token when conditioned on the reference. Increasing teacher capacity does not rescue the signal. Re-scoring the same Qwen rollouts with the 8B teacher leaves the reference-conditioned scores similarly small, while critique-conditioned signals become stronger across settings (Appendix E).

Not even informative about correctness. A diffuse signal could still be useful even if it does not localize errors, but at minimum it should recover the verifier’s one bit: whether the rollout is correct. We therefore test whether the localization or correction signal averaged over all tokens of a rollout ranks correct rollouts above incorrect siblings of the same problem. Across all three tasks and both teacher scales, neither reference-conditioned signal does so reliably: all twelve within-problem AUC confidence intervals include chance. Pooled across problems, some cells move away from chance (Figure 2), but the same deviations appear under content-free controls, indicating that they largely reflect problem-level factors such as length and difficulty. A trained probe over richer trace-level summaries does no better than the content-free controls (Appendix G). The reference-conditioned signal is therefore not only poorly localized. Even aggregated over the full trajectory, it carries little information about whether the reasoning is correct.

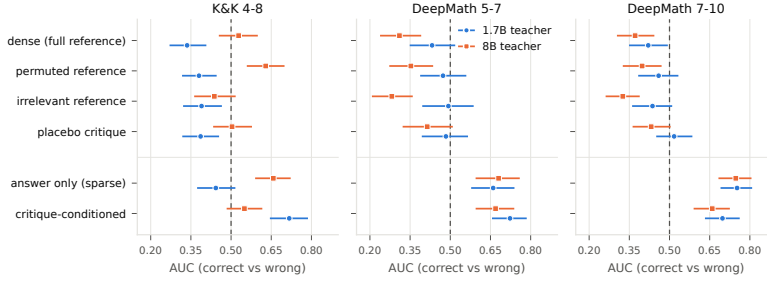


Figure 2: **The reference-conditioned signal cannot tell correct rollouts from wrong ones.** Pooled AUC of trace-mean $\Delta \log p$, correct vs. wrong rollouts, by privileged context and teacher.

Table 2: **Generating finds the errors that scoring cannot.** Percent of wrong rollouts with a critique quote within five sentences of a gold error (localized ± 5), quotes found verbatim in the rollout (quote-match), and critiques in the required format (parse).

critic	localized ± 5			quote-match			parse		
	K&K	L5-7	L7-10	K&K	L5-7	L7-10	K&K	L5-7	L7-10
gpt-5.2 (oracle)	57	62	67	98	96	95	100	100	100
Qwen3-8B	26	22	27	96	82	83	99	86	85
Qwen3-1.7B (self)	20	18	21	77	64	68	90	83	83

6 Dense Feedback Is a Search Problem

Why does the reference as privileged context induce so little signal at the error? We posit that locating where a rollout goes wrong, and why, is itself a reasoning problem: the reference and rollout must be compared until the discrepancy is found. A scoring pass is never asked to perform that search. It only produces a next-token distribution with the reference in context. If the failure is one of scoring rather than of information, the same model should recover the errors when allowed to generate. We test this directly by asking critics of different scales to read the problem, reference, and rollout and produce a list of errors, each with a verbatim quote and an explanation (§3.3). The quote makes the prediction directly localizable against the gold error spans and later provides an anchor for critique-conditioned teacher scoring.

Table 2 confirms this prediction. The strongest critic localizes a gold error on well over half of the rollouts across tasks, while smaller critics succeed less often but preserve the same ordering with scale. Their main failures are practical rather than conceptual: they often paraphrase instead of quoting the rollout or violate the required output format. The key result is that the same reference that yields little error-specific signal under scoring produces explicit error locations under generation. This suggests that search is the missing computation. The next section asks whether these generated critiques can in turn be converted back into a dense signal by scoring.

7 Critique-Conditioned Self-Distillation Delivers

We replace the reference in the teacher’s context with a generated critique and score the student’s rollout exactly as before, using a critique from another problem as the placebo. Figure 3 repeats Figure 1 with the critique in place of the reference. The teacher, alignment, and estimand are the same. Under critique conditioning, the previously overlapping error and control traces separate. The teacher’s distribution shifts after the flagged error sentence, and its probability on the student’s continuation decreases there rather than elsewhere. This pattern holds across tasks.

Table 1 confirms the same result at the rollout level. With oracle critiques, every setting satisfies the corrective criterion of §5 on both signals: errors are localized relative to clean regions, and probability is pulled away from the student’s erroneous continuation, at several times the magnitude of the reference-conditioned signal. The critic sees only the reference that the scoring teacher already has. The critique therefore carries the same information in a form the scoring pass can use. Self-generated

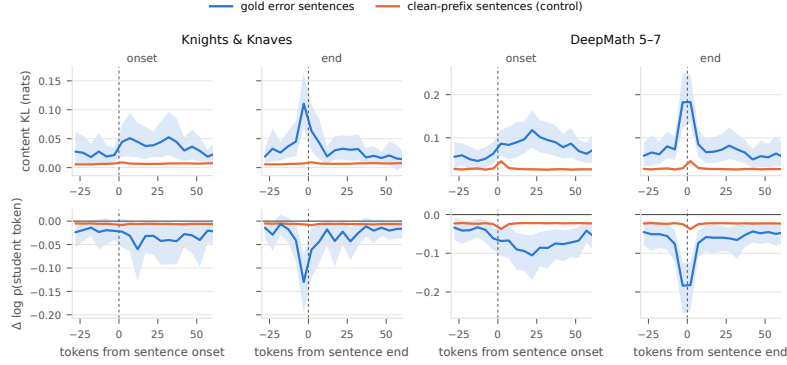


Figure 3: **With a critique in context, the same teacher reacts at the errors.** Same layout as Figure 1, with gpt-5.2 critiques in place of the reference; Qwen3-1.7B scoring its own rollouts. The error curve leaves its control on every panel.

Table 3: **Critique conditioning redirects probability toward self-correction markers; the reference does not.** Share of post-error tokens whose argmax becomes a self-correction marker, at errors (err) and controls (ctl).

task	reference		critique	
	err	ctl	err	ctl
K&K	4.9	4.9	23.9	7.7
L5-7	3.7	2.9	9.0	5.2
L7-10	2.6	2.6	10.6	4.8

Table 4: **The quote carries localization, the reason carries correction.** Error-span Z with the oracle’s critiques ablated to quote only or reason only; Qwen3-1.7B teacher unless stated.

critique	K&K		L5-7		L7-10	
	loc	corr	loc	corr	loc	corr
quote + reason	+0.75	-0.62	+0.99	-0.90	+0.78	-0.54
reason only	+0.48	-0.47	+0.75	-0.58	+0.61	-0.30
quote only	+0.71	-0.16	+0.94	-0.39	+0.69	-0.27
quote + reason, 8B teacher	+1.20	-0.79	+1.09	-0.96	+0.91	-0.62
quote only, 8B teacher	+0.82	+0.21	+1.17	-0.22	+0.90	+0.03

critiques preserve the same effect at reduced magnitude: with Qwen3-1.7B critiquing its own rollouts, localization is positive and correction is negative on all three tasks, and the effect strengthens with model scale. The weaker self-critique signal still trains. In the training runs of Appendix A, the self-critique arm recovers most of the oracle-critique arm’s gain over GRPO (Table 6). Larger scoring teachers likewise amplify the critique-conditioned signal while leaving the reference-conditioned signal near baseline (Appendix E). Critique conditioning also recovers the trace-level correctness signal absent under the reference (Figure 2). The same pattern holds on human-labelled error steps: on ProcessBench [Zheng et al., 2025], with each solution scored by the Qwen2.5 model that wrote it, the reference is neither localized nor corrective, while a critique localizes the labelled step for both generators and corrects once the tokens it quotes are excluded (Appendix J).

Correction resolves into self-correction language. The correction signal in Table 1 shows that critique conditioning moves probability mass away from the student’s continuation, but not where that mass goes. We therefore inspect high-content tokens in the post-error window and bucket the teacher’s argmax prediction by destination. Under reference conditioning, error windows show no excess of self-correction markers such as *Wait*, *But*, or *However* while under critique conditioning, the reconsideration bucket rises to two to three times the control rate on every task (Table 3). Thus, the probability mass released from the student’s erroneous continuation is redirected toward self-correction language only when the teacher is told what is wrong, not merely shown the correct solution.

Delivery decomposes into pointing and correcting. A critique contains two separable components: a quote that points to the erroneous step, and a reason that explains why it is wrong. Table 4 ablates each in turn. Localization is carried primarily by the quote: removing the reason leaves it largely unchanged, while removing the quote weakens it across tasks. Correction depends on the reason:

Table 5: **At test time the critique’s content converts math problems the model cannot solve by sampling.** Olmo-3-32B-Think resumes its own wrong L7-10 rollouts from a cut, eight samples per rollout; sample accuracy in percent on all problems and on the problems unsolved in eight samples. Bold = 95% CI of the paired difference to the control excludes 0.

cut	context at the cut	all (39)	unsolved by sampling (19)
fixed depth	cue	10.9	4.3
critique’s flagged sentence	cue + critique item	13.4	9.3

without it, the corrective shift shrinks substantially and, with the larger teacher, can disappear or even reverse. The two components therefore play distinct roles in delivery: the quote tells the scoring teacher *where* to react, while the reason tells it *how* to change the student’s distribution.

8 Using the Signal at Test Time

We test whether the critique is useful at inference. We take wrong rollouts of Olmo-3-32B-Think on L7-10, 55 rollouts from 39 problems, 19 of which the model does not solve in eight samples. The model critiques each rollout with the reference solution in context, the self-critique setting of Table 1. We cut the rollout at the first sentence the critique flags and let the model continue from the cut, eight samples per rollout. The critique item is placed at the cut together with the cue “Wait, I should correct this reasoning.” The control cuts at a fixed depth and places the cue alone. Table 5 reports sample accuracy. With the critique in context the model solves the problems it never solved by sampling about twice as often as the control. The critic had the reference in context, so the gain measures what a privileged critique delivers at inference, not unaided self-correction. We also train with the signal (Appendix A). On four Knights & Knaves puzzles Olmo-3-7B-Think never solves by sampling, the critique-conditioned arm solves two puzzles that the GRPO control never solves (Appendix A.1). On a 320-puzzle training set, the critique-conditioned term raises out-of-distribution accuracy over GRPO and over reference-conditioned self-distillation, and the student as its own critic recovers most of the oracle critic’s gain (Appendix A.2).

9 Conclusion and Limitations

We show that reference-conditioned self-distillation provides dense supervision but not reliable corrective credit: its signal does not localize student errors and is not informative about rollout correctness. The same teacher can recover those errors when allowed to generate a critique, suggesting that credit assignment requires an explicit search over the reasoning trace. Critique-conditioned self-distillation separates that search from scoring, using generation to identify the error and a scoring pass to turn it into dense token-level supervision. **Limitations.** Our study has three main limitations. First, our training evidence is small in scale, a 320-puzzle set and 100 steps (Appendix A.2), and we have not run critique-conditioned self-distillation at full RL scale. Second, critique conditioning adds a generation pass over the wrong rollouts of each batch. In our runs this costs 6% of step time with the student as critic and about a fifth with an API critic (Appendix A.3). Third, every critique in this paper, in the signal analysis, at test time and in training, is written with the reference solution in context. Our results show that a critique conveys privileged information in a form the scoring teacher can use, not that the model corrects itself unaided.

References

- Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D. Chang, and Prithviraj Ammanabrolu. Critique-out-loud reward models, 2024. URL <https://arxiv.org/abs/2408.11791>.
- Anthropic. Introducing claude sonnet 5, jun 2026. URL <https://www.anthropic.com/news/claude-sonnet-5>. Accessed 2026-09-12. Announcement dated June 30, 2026; model ID claude-sonnet-5.
- Xiushi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. Rm-r1: Reward modeling as reasoning, 2026a. URL <https://arxiv.org/abs/2505.02387>.

Yuxin Chen, Yu Wang, Yi Zhang, Ziang Ye, Zhengzhou Cai, Yaorui Shi, Qi Gu, Hui Su, Xunliang Cai, Xiang Wang, An Zhang, and Tat-Seng Chua. Learning to self-verify makes language models better reasoners, 2026b. URL <https://arxiv.org/abs/2602.07594>.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanbiao Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-rl: Incentivizing reasoning capability in llms via reinforcement learning, 2026a. URL <https://arxiv.org/abs/2501.12948>.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Donghao Li, Dongjie Ji, Erhang Li, Fang Wei, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanting Chen, Guoai Cao, Guolai Meng, Guowei Li, Han Yu, Han Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoling Zhang, Haoming Luo, Haoran Wei, Haotian Yuan, Haowei Zhang, Haowen Luo, Haoyu Chen, Haozhe Ji, Hengqing Zhang, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J Yang, JQ Zhu, Jia Luo, Jia Song, Jia Yu, Jialiang Huang, Jialu Cai, Jian Liang, Jiangting Zhou, Jiasheng Ye, Jiashi Li, Jiabin Xu, Jiewen Hu, Jieyu Yang, Jin Chen, Jin Yan, Jingchang Chen, Jingli Zhou, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jingzi Zhou, Jinhua Zhu, Jiping Yu, Joseph Sun, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junmin Zheng, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Leyi Xia, Li Zhang, Liang Zhao, Lihua Guo, Lingxiao Luo, Linwang Ma, Linyan Zhu, Litong Wang, Liyu Cai, Liyue Zhang, Longhao Chen, MS Di, MY Xu, Max Mei, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Mingxu Zhou, Minmin Han, Ning Wang, Panpan Huang, Panpan Wang, Peixin Cong, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qiwei Jiang, Rui Tian, Ruifan Xu, Ruijie Lu, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqian Chen, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, Ruyi Chen, SH Liu, Shanghao Lu, Shangmian Sun, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoheng Nie, Shaoqing Wu, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Shuying Yu, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian Pei, Tian Ye, Tianle Lin, Tianran Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tun Wang, W Zhang, WL Xiao, Wangding Zeng, Wei An, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjing Yao, Wenjun Gao, Wenkai Yang, Wenlve Huang, Wenqing Hou, Wentao Zhang, Wenting Ma, Xi Gao,

- Xiang He, Xiangwen Wang, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyu Zhang, Xu Chen, Xuanyu Wang, Xuecheng Su, Xueyin Chen, Xuheng Lin, Xuwei Fu, YC Yan, YQ Wang, YW Ma, Yanfeng Luo, Yang Zhang, Yanhong Xu, Yanru Ma, Yanwen Huang, Yao Li, Yao Li, Yao Xu, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Shao, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yijia Wu, Yiliang Xiong, Yiling Ma, Ying He, Ying Tang, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, YuKun Li, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuanhao Li, Yudian Wang, Yuehan Yang, Yuer Xu, Yuhan Wu, Yuhao Meng, Yuheng Zou, Yukun Zha, Yunfan Xiong, Yupeng Chen, Yuping Lin, Yuqian Cao, Yuqian Wang, Yushun Zhang, Yuting Yan, Yutong Lin, Yuxian Gu, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhen Huang, ZF Wu, Zehao Wang, Zehua Zhao, Zehui Ren, Zekai Zhang, Zhangli Sha, Zhe Fu, Zhe Ju, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zheren Gao, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhixuan Chen, Zhiyu Wu, Zhizhou Ren, Zhongyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihua Qu, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Ziyi Wan, Zizheng Pan, and Zongqing Yao. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026b. URL <https://arxiv.org/abs/2606.19348>.
- Sarthak Harne, Chinmay Karkar, Yash Pandya, Ahmed Awadallah, and Akshay Nambi. Privileged, but biased: How pi-conditioned teachers break self-distillation, 2026. URL <https://arxiv.org/abs/2608.04794>.
- Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zhongyuan Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, and Bo Zheng. Can large language models detect errors in long chain-of-thought reasoning?, 2025a. URL <https://arxiv.org/abs/2502.19361>.
- Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning, 2025b. URL <https://arxiv.org/abs/2504.11456>.
- 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, 2024. URL <https://arxiv.org/abs/2310.01798>.
- Jonas Hübner, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Büning, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation, 2026. URL <https://arxiv.org/abs/2601.20802>.
- Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.
- Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models, 2026. URL <https://arxiv.org/abs/2607.05184>.
- Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process reward models that think, 2025. URL <https://arxiv.org/abs/2504.16828>.
- Jeonghye Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Rebellious student: Reversing teacher signals for reasoning exploration with self-distilled rlvr, 2026. URL <https://arxiv.org/abs/2605.10781>.
- Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe, 2026a. URL <https://arxiv.org/abs/2604.13016>.

- Yu Li, Shu Hong, and Tian Lan. Localizing credit at the divergence: Path-conditioned self-distillation for llm reasoning, 2026b. URL <https://arxiv.org/abs/2606.15576>.
- 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, 2023. URL <https://arxiv.org/abs/2305.20050>.
- Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning, 2026. URL <https://arxiv.org/abs/2605.21606>.
- Manuel R Mercier, Anne-Sophie Dubarry, François Tadel, Pietro Avanzini, Nikolai Axmacher, Dillan Cellier, Maria Del Vecchio, Liberty S Hamilton, Dora Hermes, Michael J Kahana, et al. Advances in human intracranial electroencephalography research, guidelines and good practices. *NeuroImage*, 260:119438, 2022.
- Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznan-ski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL <https://arxiv.org/abs/2512.13961>.
- OpenAI. Gpt-5.2 model. OpenAI API model documentation, 2025a. URL <https://developers.openai.com/api/docs/models/gpt-5.2>. Accessed 2026-09-12. Model snapshot listed on the page: gpt-5.2-2025-12-11.
- OpenAI. Gpt-5 system card, aug 2025b. URL <https://cdn.openai.com/gpt-5-system-card.pdf>. Accessed 2026-09-12. Title page dated August 13, 2025.
- Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation, 2026. URL <https://arxiv.org/abs/2606.11709>.
- 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, 2024. URL <https://arxiv.org/abs/2402.03300>.
- Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. Gpt-4 doesn’t know it’s wrong: An analysis of iterative prompting for reasoning problems, 2023. URL <https://arxiv.org/abs/2310.12397>.
- Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. Llms cannot find reasoning errors, but can correct them given the error location, 2024. URL <https://arxiv.org/abs/2311.08516>.
- 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, 2022. URL <https://arxiv.org/abs/2211.14275>.
- Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations, 2024. URL <https://arxiv.org/abs/2312.08935>.
- Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation, 2026. URL <https://arxiv.org/abs/2605.26844>.

- Yubo Wang, Xiang Yue, and Wenhua Chen. Critique fine-tuning: Learning to critique is more effective than learning to imitate, 2025. URL <https://arxiv.org/abs/2501.17703>.
- Jinwei Xiao, Zhuowen Han, Yueqing Sun, Zhengxi Lu, Yuxin Liu, Zhiyuan Yao, Wentao Chen, Qi Gu, and Xunliang Cai. Finding the evidence: Discovering decision-supporting tokens for on-policy reasoning distillation, 2026. URL <https://arxiv.org/abs/2606.22830>.
- Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. On memorization of large language models in logical reasoning, 2025a. URL <https://arxiv.org/abs/2410.23123>.
- Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning, 2025b. URL <https://arxiv.org/abs/2502.14768>.
- Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation, 2026. URL <https://arxiv.org/abs/2604.14084>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr, 2026. URL <https://arxiv.org/abs/2604.03128>.
- Zichao Yu, Chengzhi Yu, Shengze Xu, Yujin Han, Bingqing Jiang, Xu Wang, and Difan Zou. Mismatch matters: On-policy distillation beyond token agreement, 2026. URL <https://arxiv.org/abs/2608.09836>.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL <https://arxiv.org/abs/2504.13837>.
- Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction, 2025a. URL <https://arxiv.org/abs/2408.15240>.
- Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning, 2025b. URL <https://arxiv.org/abs/2501.07301>.
- Siyao 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, 2026a. URL <https://arxiv.org/abs/2601.18734>.
- Xuyang Zhao, Liting Zhang, Zichen Xu, Zhihu Wang, Xu Caiyue, Shiwan Zhao, and Qicheng Li. Is more privileged information better? from solution traces to problem-solving structure in self-distilled reasoning, 2026b. URL <https://arxiv.org/abs/2608.01589>.
- Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning, 2025. URL <https://arxiv.org/abs/2412.06559>.

Appendix

A Training with the Critique-Conditioned Signal

We report two training experiments. The first is a cold-start pilot on four puzzles with Olmo-3-7B-Think, the pilot referred to in §8. The second compares the critique-conditioned term against GRPO and reference-conditioned self-distillation over 100 steps on Olmo-3-7B-Instruct, with the oracle critic and with the student as its own critic. We then report the compute cost of the critique pass.

A.1 Cold-start pilot on four puzzles

We train Olmo-3-7B-Think on four 16-person Knights & Knaves puzzles it never solves in 64 samples. The student writes the rollouts, critiques its own wrong rollouts with the verified derivation as reference, and scores them as its own teacher. On all-wrong groups, where GRPO has no gradient, the critique-conditioned correction signal in the critic-located error span becomes the advantage, with a zero-sum positive term on the prefix before the error. The control is GRPO alone. Both arms run 30 steps of 32 rollouts with one seed. Figure 4 shows the cumulative number of solved training rollouts. The control solves one of the four puzzles by chance and never the other three. The critique-conditioned arm solves two of those three puzzles.

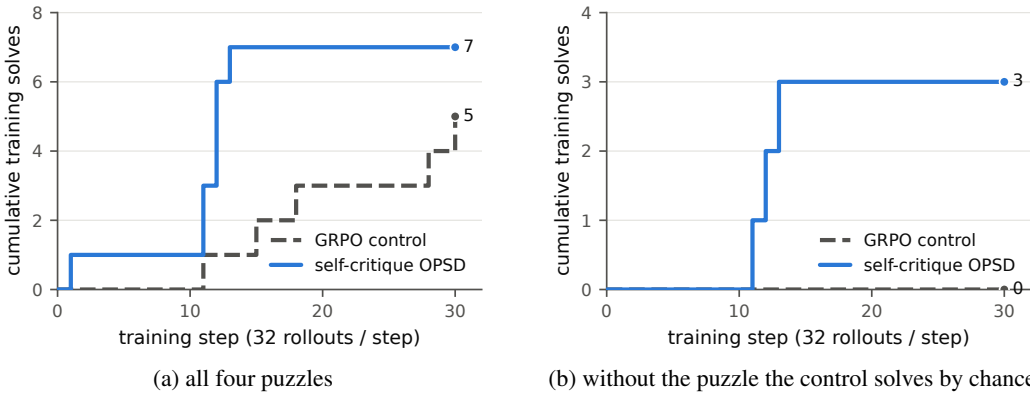


Figure 4: Cumulative solved training rollouts on four certified-cold Knights & Knaves puzzles, critique-conditioned advantage against the GRPO control. Olmo-3-7B-Think as student, critic, and teacher.

A.2 Training against GRPO and reference-conditioned self-distillation

We train Olmo-3-7B-Instruct [Olmo et al., 2026] on 320 Knights & Knaves puzzles with 4–8 people for 100 steps, with 16 puzzles and 8 rollouts per step, learning rate 5×10^{-6} and a generation cap of 8,192 tokens. We evaluate the final checkpoint at a 32,768-token cap on 500 held-out puzzles with 4–8 people and on 100 puzzles each with 9, 10 and 11 people. OOD is the mean over the three larger sizes. The baselines are GRPO (§3.1) and reference-conditioned self-distillation with the verified derivation in the teacher’s context (RLCSD, Pan et al., 2026). The critique-conditioned arms add a term on the wrong rollouts. The critic reads the problem, the rollout and the derivation and writes the critique of §3.3. The teacher scores the rollout with the critique in context, and the corrective part of its log-ratio against the student is applied with weight 3 on the sentence the critique quotes and the 50 tokens after it. The critic is gpt-5-mini (oracle) or the student itself (self-critique).

The critique term on top of GRPO is worth as much as the reference-conditioned teacher term, +5.5 against +5.8 OOD points over GRPO, and on top of RLCSD it adds a further 3.9 (Table 6). The two gains add. The self-critique is the weaker critic, its quotes match a sentence of the rollout 82% of the time against 96% for the oracle, and it still recovers most of the oracle’s gain, 0.593 against 0.606 with RLCSD at 0.567.

Table 6: Accuracy after 100 steps on Olmo-3-7B-Instruct. In-domain is Knights & Knaves with 4–8 people, OOD the mean over 9, 10 and 11 people. The critique term is confined to the quoted sentence and the 50 tokens after it.

method	critic	in-domain	9	10	11	OOD
base model	–	0.666	0.517	0.415	0.275	0.403
GRPO	–	0.700	0.595	0.532	0.400	0.509
GRPO + critique	gpt-5-mini	0.744	0.680	0.568	0.445	0.564
RLCSD	–	0.745	0.688	0.583	0.430	0.567
RLCSD + critique	gpt-5-mini	0.754	0.698	0.623	0.497	0.606
RLCSD + critique	self	0.735	0.693	0.595	0.490	0.593

A.3 Compute cost of the critique pass

Critique conditioning adds one generation per critiqued rollout. Only wrong rollouts in groups with at least one correct rollout are critiqued, 16 to 18 of the 128 rollouts per step in the runs of Table 6. Table 7 gives the cost per step averaged over the 100 steps. The self-critique runs in the same inference engine as the rollouts. We push the frozen base weights, generate the critiques and push the current policy back. Olmo-3-7B-Instruct writes short critiques, 175 tokens on average, so the critique tokens are 0.4% of the rollout tokens generated per step and the whole pass is 6% of step time, one point of which is the weight swap. The API critic is a reasoning model and returns about 3,000 tokens per critique including its reasoning. Its share of step time is time spent waiting on the API, not GPU time, and could overlap with rollout generation.

Table 7: Cost of the critique pass per training step, mean over 100 steps, in the regime of Table 6. Tokens are per critique.

critic	arm	critiqued rollouts	prompt tokens	completion tokens	step time	share
gpt-5-mini (API)	RLCSD + critique	16	5,600	2,970	255 s	21%
gpt-5-mini (API)	GRPO + critique	18	4,960	2,890	256 s	22%
self (in-engine)	RLCSD + critique	18	6,770	175	219 s	6%

B Critique Examples

We show one wrong Qwen3-1.7B rollout per task with its gold error sentence, the gpt-5.2 critique and the student’s own critique. Both critics received the same prompt (Appendix D.1) and the same reference. The gold error sentence is in italics in the rollout excerpt.

Knights & Knaves. Problem. A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie. You meet 6 inhabitants on this island: Yara, Liam, Peter, Sophia, Noah, and Bob. Yara says: “Peter is a knave if and only if Liam is a knight.” Liam says: “It is not the case that Yara is a knave if and only if Yara is a knight.” Peter says: “Bob is a knight.” Sophia says: “Liam is a knight.” Noah says: “It is not the case that Liam is a knave.” Bob says: “Noah is a knave.”

Reference answer. Yara is a knight, Liam is a knight, Peter is a knave, Sophia is a knight, Noah is a knight, Bob is a knave.

Student’s answer. Yara is a knight, Liam is a knave, Peter is a knight, Sophia is a knave, Noah is a knave, Bob is a knight.

Rollout excerpt. [...] if Yara is a knight (Y is true), then the statement “Yara is a knave if and only if Yara is a knight” would be “False $\leftrightarrow$ True”, which is False. Similarly, if Yara is a knave (Y is false), then the statement is “True $\leftrightarrow$ False”, which is also False. Therefore, the statement is always false. *Therefore, Liam’s statement is false.* Therefore, if Liam is a knight, he would be making a false statement, which is impossible. Therefore, Liam must be a knave. Wait, that seems important. [...]

Gold error. Liam’s statement is the negation of the inner biconditional, and since that inner biconditional is always false, Liam’s statement is actually true.

Oracle critique (gpt-5.2). ERROR 1: “Therefore, the statement is always false.” – The biconditional “Yara is a knave iff Yara is a knight” is always false, so its negation (Liam’s statement) is always true, not false.

ERROR 2: “Therefore, Liam must be a knave.” – Since Liam’s statement is necessarily true, Liam must be a knight, not a knave.

ERROR 3: “Since we just determined that Liam is a knave,” – This propagates the earlier incorrect classification of Liam, making subsequent deductions unreliable.

Self-critique (Qwen3-1.7B). ERROR 1: “Liam is a knave” – Liam’s statement is a contradiction (always false), so he must be a knave. This is correct, but the final answer contradicts this conclusion.

ERROR 2: “Noah is a knave” – Noah’s statement is “Liam is a knight,” which is false because Liam is a knave. Therefore, Noah is a knave. This is correct.

ERROR 3: “Bob is a knight” – Bob’s statement “Noah is a knave” is true, so Bob is a knight. This is correct.

ERROR 4: “Peter is a knight” – Peter’s statement “Bob is a knight” is true, so Peter is a knight. This is correct.

ERROR 5: “Yara is a knight” – Yara’s statement is true, so she is a knight. This is correct.

Final Answer (boxed): Yara is a knight, Liam is a knight, Peter is a knave, Sophia is a knight, Noah is a knight, Bob is a knave.

The oracle quotes the sentence before the gold error and the conclusion drawn from it, and each reason says which way to correct it. The self-critique quotes fragments of the final answer, endorses the student’s reasoning, and then writes the reference answer as its own final answer. Its quotes carry no location and its reasons carry no correction before that last line.

DeepMath L5-7. Problem. Find the residue of the function $\sin\left(\frac{1}{z}\right)$ at its poles.

Reference answer. 1. **Student’s answer.** 0.

Rollout excerpt. [...] So, this is an infinite series, and there are infinitely many terms with positive powers of $1/z$. Therefore, the Laurent series has an infinite number of terms with negative powers of z , which means that $z = 0$ is an essential singularity, not a pole. But wait, the question says “at its poles”. *But if there are no poles, then the residue would be...* Well, if there are no poles, then the residue is zero? But maybe I need to check if there are any poles. But since $\sin(1/z)$ has an essential singularity at $z = 0$, there are no poles. So, the residue at the pole... but there are no poles. Therefore, the residue is zero? [...]

Gold errors. Three sentences are gold. The first is the italic one above (a residue can exist at an essential singularity and is the coefficient of the z^{-1} term, which is nonzero here). The other two are “Since there are no poles, the Laurent expansion does not have any $(z - a)^{-1}$ term, so the coefficient is zero” and “if a function has an essential singularity at a point, then the residue is not defined in the usual sense.”

Oracle critique (gpt-5.2). ERROR 1: “then the residue would be... Well, if there are no poles, then the residue is zero?” – Lack of poles does not imply the residue is zero; residue is defined at isolated singularities (including essential ones) via the z^{-1} coefficient.

ERROR 2: “if a function has an essential singularity at a point, then the residue is not defined in the usual sense.” – False: the residue is still defined for essential singularities as the coefficient of z^{-1} in the Laurent series.

ERROR 3: “Since there are no poles, the Laurent expansion does not have any $(z - a)^{-1}$ term” – Incorrect: $\sin(1/z) = 1/z - 1/(6z^3) + \dots$ does contain a z^{-1} term, so residue at 0 is 1.

Self-critique (Qwen3-1.7B). ERROR 1: “the residue of the function $\sin\left(\frac{1}{z}\right)$ at its poles is $\boxed{0}$ ” – The student incorrectly concludes that the residue is zero because there are no poles, but the residue is defined only for isolated singularities, not for the absence of poles. The function has an essential singularity at $z = 0$, and the residue at that singularity is 1, not zero.

ERROR 2: “the student’s final answer is known to be WRONG: the solution below contains at least one mistake” – The student’s final answer is wrong because the residue of $\sin\left(\frac{1}{z}\right)$ at its essential singularity (not poles) is 1, not 0.

[ERROR 3 to ERROR 5 repeat these two items.]

The oracle’s three quotes are the three gold error sentences, and each reason states the correct fact. The self-critique quotes the final answer and a line of its own prompt, neither of which is a step of the derivation, so it does not localize. Its reason is right: the residue at the essential singularity is 1. This is the division of labour of Table 4. The quote carries the location and the reason carries the direction, and here the self-critique has the second without the first.

C Extended Related Work

Self-distillation with a privileged teacher. On-policy self-distillation [Zhao et al., 2026a, Hübötter et al., 2026] produces a per-token signal in place of the single bit that verifiable-reward RL provides: the same model scores the student’s own rollout with the reference solution in its context. The line

has grown by varying pieces of that recipe. RLSD keeps the reference fixed and turns the log-ratio into a token-level advantage weight [Yang et al., 2026]. RLCSO contrasts a correct hint against a wrong one to suppress the style drift a hint alone induces [Pan et al., 2026], and Zhao et al. [2026b] replace the full solution with a more structured trace. Carried over to thinking models, the recipe degrades performance rather than helping [Kaur et al., 2026], and Rebellious Student reverses the signal on rollouts the student already gets right, to encourage exploration rather than correction [Kim et al., 2026]. Closest to our premise, Li et al. [2026b] show that a teacher conditioned only on the final answer produces no signal at the intermediate positions where guidance is needed, and condition it on a successful peer rollout instead. All of these variants take the scored signal as given and change how it is consumed or what the teacher is conditioned on. We ask whether the scored signal carries any credit at the error, and find that a full reference solution does not. Cross-model distillation selects the teacher’s tokens by mismatch, importance, learnability, position, or evidence [Yu et al., 2026, Xu et al., 2026, Wang et al., 2026, Liu et al., 2026, Xiao et al., 2026], and Li et al. [2026a] find instead that the high-overlap tokens carry the signal. Those methods assume a signal to select. A cross-model teacher’s gap reflects a real capability difference; a privileged teacher differs from the student only by context, and at the errors we find no signal to select.

Failure modes of the privileged teacher. Three papers report that the privileged teacher misbehaves. Harne et al. [2026] show that the teacher is biased by the privileged information it is given, and ask whether self-distillation alone teaches anything. Kaur et al. [2026] show that the same information degrades performance on thinking models. Pan et al. [2026] find that the distributional gap concentrates on style tokens rather than task-bearing ones. None of them anchors the diagnosis to the place where the student’s reasoning goes wrong. We measure the signal itself against gold error spans on the student’s own rollouts, and find that scoring fails to localize the error while generating does.

Locating an error is a reasoning problem. Tyen et al. [2024] show that language models cannot find their reasoning errors but can correct them given the error location, and Huang et al. [2024] and Kamoi et al. [2024] show that intrinsic self-correction without an external signal does not improve reasoning. Verifiers do better when they reason before they judge. Generative verifiers [Zhang et al., 2025a], critique-out-loud reward models [Ankner et al., 2024], and reward models trained as reasoners [Chen et al., 2026a] all take this route, and process reward models that think locate erroneous steps better than discriminative scorers with a fraction of the labels [Khalifa et al., 2025]. ProcessBench reports that trained step scorers underperform prompted critics [Zheng et al., 2025], and Chen et al. [2026b] find that verification is a capability separate from generation. Generation’s advantage is not universal. Tyen et al. [2024] also show that a trained classifier can localize mistakes without generating, and Stechly et al. [2023] find a critique’s content irrelevant to the benefit of iterative prompting in planning. Our claim is correspondingly narrow. The token-level likelihoods of a teacher scoring a forced trace do not localize the error, while the same model generating a critique does.

Process supervision and critics. Process supervision identifies the error when the step-level labels are created. Labels come from humans [Uesato et al., 2022, Lightman et al., 2023], from automatic rollouts [Wang et al., 2024], or from an LLM judge, and the automatically estimated labels generalize worse than the judged or human ones [Zhang et al., 2025b]. The generative verifiers and critique-out-loud models above, together with critics trained to critique rather than imitate [Wang et al., 2025], put the search into generation, which is the readout we use. Such critics are scored against human-annotated step-level error benchmarks [Tyen et al., 2024, Zheng et al., 2025, He et al., 2025a], which we revisit in Appendix J. Critique-conditioned self-distillation combines both ideas: a generated critique, delivered through the scoring teacher, becomes a per-token signal for on-policy self-distillation, with the critic and the teacher the same model.

D Setup Details

Rollouts and targets. Every rollout in the paper is sampled from the student itself, on the training distribution. A rollout is a *target* if the verifier scores it wrong, it is not truncated, and it ends in a boxed answer; verifier-correct rollouts of the same problems are kept as a control stratum for §5. K&K targets are drawn only from puzzles whose English rendering is decidable as worded, about half of the pool, because on the remainder a “wrong” rollout is often a legitimate alternative reading. Rollouts are sampled with the training parameters, temperature 1.0, top-p 0.95, top-k 20,

eight rollouts per prompt, with a 16,384-token budget for Qwen3-1.7B and a 32,768-token budget for the Olmo-3 models, whose thinking traces are longer.

Gold error spans. A grader model reads the problem, the rollout, and the reference, and lists the places where the reasoning goes wrong, earliest first, each as a verbatim quote of one sentence with a reason, a severity, and a confidence (§D.2). The grader is deepseek-reasoner for the original Qwen sample, deepseek-v4-flash for the problems added in the scale-up, and claude-sonnet-5 for the two Olmo-3 studies. Each flagged sentence is then judged independently by gpt-5.2 and by claude-sonnet-5, which see the problem, the reference, and the rollout with the sentence marked and the grader’s reason attached (§D.3). A quarter of the flags sent to each jury are replaced by position-matched control sentences carrying the same reason, which measures the jury’s leniency; the control acceptance rate is reported with each study. A flag is *strict* gold if both juries accept it and *lenient* gold if the claude jury accepts it. Error-span Z uses the first strict gold sentence of each rollout; the curves of Figs. 1 and 3 use every strict gold sentence; the critique gate of Table 2 uses the lenient tier. In the Olmo-3 studies the claude jury shares its model with the grader, so gpt-5.2 is the only fully independent judge there.

Placebos. Every conditioning block is scored against a placebo of the same shape, and all reported signals are block minus placebo. The placebo for the reference is a different problem’s reference of matched length, assigned by one fixed derangement of the problems that every arm shares. The placebo for a critique is the critique written for a different problem’s rollout, assigned by the same derangement. The undifferentiated arms are reported as negative controls in Table 9.

Error-span Z windows and statistics. The error window is the 20 tokens immediately after the first gold error sentence ends. Control windows have the same shape and sit after the ends of clean sentences before the first error. We use the 12 nearest control windows, non-overlapping, with a 60-token guard band around the error sentence. On the localization signal we take the log of each window mean before Z-scoring; the correction signal is already a signed log-probability difference and is used as is. The dataset-level statistic is the mean of Z_r over rollouts, winsorized at the 5th and 95th percentiles, with a 95% bootstrap interval clustered by problem (2,000 resamples). The null subtracted from every cell relabels a control sentence of the same rollout as the error and recomputes Z_r ; the reported value is $Z - Z_{\text{null}}$. Bold in Table 1 marks cells whose 95% interval excludes zero.

Which reference each model sees. The scoring teacher’s reference is a step-by-step derivation verified against the ground truth for K&K and the dataset’s reference solution with its final answer for DeepMath. The oracle critic receives the same reference as the teacher. The open-weights critics on the scale-up sample receive the K&K derivations and, for DeepMath, a clean derivation generated and verified against the ground-truth answer, because the dataset solutions are raw chains of thought that weaker critics mistake for a second student solution; the 60 rollouts without a verified derivation are dropped from those arms.

D.1 Critique prompt

The critic receives the problem, the student’s rollout with sentence tags, the reference solution, and the instructions below. The rollout is presented before the reference to prevent the critic from quoting the reference. The same prompt is used for every critic.

Problem: {problem}

You are grading a student’s solution to this problem.

The student’s final answer is known to be WRONG: the solution below contains at least one mistake.

```
=== Student Solution Begin ===
[S0] <sentence 0> [S1] <sentence 1> ...
=== Student Solution End ===
```

For reference, here is a correct solution:

```
=== Reference Solution Begin ===
{reference}
=== Reference Solution End ===
```

Coverage matters more than precision here: if you are unsure where the reasoning FIRST goes wrong, list EACH plausible location as its own ERROR line (up to 5), earliest first -- it is better to list several candidate locations than to miss the true one.

Write a CONCISE critique of the student's reasoning, at most 120 words total. Use EXACTLY this format, one line per error, earliest error first, at most 5 errors, and nothing else:

```
ERROR 1: "<short quote, at most 10 words, copied word-for-word from the
student's solution at the point the error occurs>" -- <one short
sentence: what is wrong there>
ERROR 2: ...
```

If the reasoning contains no errors, output the single line: NO ERRORS

Rules: each quote must be copied exactly, character for character, from the Student Solution above (never from the reference, never a paraphrase), and must be at most 10 words. Point at the sentence where the reasoning first goes wrong, not merely the final answer line. Only report NEW errors: once the reasoning is broken, list a later step only if it introduces a NEW, independent error. Report only errors you are CERTAIN about.

IMPORTANT REMINDER: every quote must be copied word-for-word from between "=== Student Solution Begin ===" and "=== Student Solution End ===" above -- NEVER from the Reference Solution, which is correct and is not being graded.

The student's sentences are tagged [S0], [S1], ... to help you keep track of position. Do NOT include any [S#] tag inside your quotes -- quote only the sentence text itself.

D.2 Gold-span grader prompt

The grader sees the problem, the rollout, and the reference (arm *ref*), with the same known-wrong line as the critic, followed by:

List the places where the student's REASONING goes wrong, in order of occurrence, starting with the FIRST error. Report ONLY errors you are CERTAIN about -- one certain error is worth more than several doubtful ones. List at most 8. Your final answer must use EXACTLY this format and nothing else, beginning with the line "ERROR 1:" (or the single line "NO ERRORS" if you find none):

```
ERROR 1:
QUOTE: "<one sentence copied word-for-word from the student's solution,
at the point where the error occurs>"
REASON: <one short sentence explaining what is wrong there>
SEVERITY: <integer 1-5: how much this error contributes to a wrong final
answer; 5 = directly causes it, 1 = minor slip>
CONFIDENCE: <integer 0-100: how certain you are that this is a genuine
error>
```

```
ERROR 2:
...
```

Rules: each QUOTE must be copied exactly, character for character, from the Student Solution above -- not from the reference, and not your own paraphrase. Point to the sentence in the student's reasoning where things go wrong, not merely the final answer line. ERROR 1 must be the EARLIEST point where the reasoning actually goes wrong. Only report NEW

errors: once the reasoning is broken, do not list every later step that merely inherits an earlier mistake -- list a later step only if it introduces a NEW, independent error. Do not include errors you are not sure about.

D.3 Jury prompt

Each grader flag is judged independently by two models. The judge sees the problem, the reference, and the rollout with the flagged sentence marked between «< and »>; a quarter of the flags are replaced by position-matched control sentences with the grader’s reason attached unchanged.

Above is a problem, a correct reference solution, and a student’s solution in which one sentence has been marked between <<< and >>>. A grader claims the student’s reasoning goes wrong AT THAT SENTENCE, for this reason:

GRADER’S CLAIM: {reason}

Judge the claim on the student’s actual reasoning. Answer YES only if a genuine reasoning error is committed at (or within a sentence or two of) the marked sentence -- i.e. the marked step is wrong given what the student derived so far, not merely downstream of an earlier mistake, and not a legitimate alternative reading. Answer NO otherwise.

Reply in EXACTLY this format:

VERDICT: <YES or NO>

CONFIDENCE: <integer 0-100>

WHY: <one short sentence>

E Error-Span Z: Full Results

This section reports error-span Z with 95% confidence intervals for every conditioning and teacher on the Qwen3-1.7B rollouts, including the 8B model as a scale-control teacher and as a critic, together with the undifferentiated hint arms as negative controls and the per-rollout sign consistency of the oracle-critique signal. Intervals are cluster bootstraps by problem. The corresponding tables for the Olmo-3 students are in Appendix I.

Table 8: Error-span Z at the first gold error with 95% confidence intervals, Qwen3-1.7B rollouts, scale-up sample, every conditioning and teacher. Rows: the reference solution or a critique placed in the scoring teacher’s context; teacher Qwen3-1.7B (the student itself) or Qwen3-8B; critiques by gpt-5.2 (oracle), by the 1.7B student on its own rollouts, or by Qwen3-8B. Localization on log content KL (positive = the distribution reacts at the error); correction on $\Delta \log p$ of the student’s token (negative = corrective). n = rollouts entering the cell.

conditioning	teacher	task	loc	95% CI	corr	95% CI	n
reference	1.7B	K&K	-0.17	[-0.33, -0.01]	+0.11	[-0.04, +0.26]	317
		L5-7	+0.14	[-0.01, +0.29]	-0.14	[-0.30, +0.02]	430
		L7-10	+0.07	[-0.08, +0.19]	+0.17	[+0.07, +0.28]	690
reference	8B	K&K	+0.13	[-0.03, +0.29]	-0.16	[-0.35, +0.02]	317
		L5-7	+0.05	[-0.11, +0.20]	-0.06	[-0.23, +0.10]	430
		L7-10	-0.05	[-0.18, +0.08]	+0.06	[-0.06, +0.19]	690
critique, oracle	1.7B	K&K	+0.75	[+0.54, +0.97]	-0.62	[-0.86, -0.40]	317
		L5-7	+0.99	[+0.82, +1.16]	-0.90	[-1.14, -0.65]	448
		L7-10	+0.78	[+0.63, +0.92]	-0.54	[-0.71, -0.35]	711
critique, oracle	8B	K&K	+1.20	[+0.95, +1.45]	-0.79	[-1.10, -0.49]	317
		L5-7	+1.09	[+0.91, +1.27]	-0.96	[-1.25, -0.68]	448
		L7-10	+0.91	[+0.77, +1.05]	-0.62	[-0.79, -0.45]	711
critique, 1.7B self	1.7B	K&K	+0.16	[-0.01, +0.32]	-0.07	[-0.23, +0.11]	316
		L5-7	+0.25	[+0.10, +0.40]	-0.14	[-0.29, +0.01]	438
		L7-10	+0.18	[+0.06, +0.30]	-0.07	[-0.19, +0.04]	648
critique, 8B critic	1.7B	K&K	+0.22	[+0.03, +0.39]	-0.30	[-0.50, -0.10]	317
		L5-7	+0.20	[+0.05, +0.34]	-0.19	[-0.34, -0.05]	427
		L7-10	+0.25	[+0.12, +0.37]	+0.03	[-0.08, +0.14]	634

Table 9: Negative controls and reliability, Qwen3-1.7B rollouts, scale-up sample. Left: error-span Z of each privileged context alone, without subtracting its placebo (reference alone; a donor critique alone), which should sit at zero. Right: per-rollout sign consistency of the oracle-critique signal, the share of rollouts whose first error window departs from its own control windows in the expected direction, on each channel. *: 95% CI excludes 0.

task	teacher	reference alone	critique placebo alone	sign% loc	sign% corr
K&K	1.7B	-0.13*	-0.06	68.1	55.5
K&K	8B	-0.15*	-0.08	72.2	55.8
L5-7	1.7B	-0.15	-0.07	74.8	55.6
L5-7	8B	-0.12	-0.14	74.8	56.7
L7-10	1.7B	-0.27*	-0.29*	67.1	46.0
L7-10	8B	-0.25*	-0.27*	68.5	49.6

F Full Figures

This section reports the figures of the main text for the remaining settings: all three tasks including DeepMath 7-10 at both alignments, the Qwen3-8B model as the scoring teacher, and the student’s own critiques in place of the oracle’s. Across every setting the pattern of the main text holds: with the reference in context the error and control curves move together, and with a critique in context the signal rises at the error sentences and nowhere else, more strongly for the 8B teacher and more weakly for the student’s own critiques.

G Correctness Prediction: Full Results

This section gives the design and the full numbers behind §5. The primary test is within-problem: for every problem that has at least one wrong and at least one correct rollout, 34 / 28 / 49 problems on K&K / DeepMath 5-7 / DeepMath 7-10, each with exactly two correct siblings, we ask whether the trace-mean of a signal orders the correct rollouts above the wrong ones, and report the pairwise AUC with a 95% interval from 2,000 bootstrap draws clustered by problem (Table 10). The pooled view, in which all correct rollouts of a task are scored against all wrong ones, is Fig. 2; it is shown because the content-free placebos reproduce its deviations from chance, which is what identifies them as length and difficulty effects. The trained probe (Table 11) is an L2 logistic regression on about sixty per-trace summaries of each signal, moments, quantiles, spike rates, and positional bins, fitted with grouped cross-validation by problem and scored out of fold; the length baseline uses the response length and its logarithm only, and the length of the critique is excluded from the critique arms because it encodes the critique’s own verdict. K&K is the weak task for this design, since wrong rollouts run much longer there and length alone reaches a pooled AUC of 0.71; the DeepMath splits are the clean case. One comparison in Table 11 is worth isolating: the full reference, which contains the answer, predicts correctness worse than the bare answer at both scales. All numbers in this section are on the original sample.

Table 10: Within-problem pairwise AUC of the reference-conditioned trace-mean signal, correct vs. wrong sibling rollouts of the same problem, with 95% CI. 0.5 = chance; no cell excludes it.

task	teacher	trace-mean KL	trace-mean $\Delta \log p$
K&K	1.7B	0.598 [0.451, 0.740]	0.396 [0.260, 0.536]
K&K	8B	0.435 [0.295, 0.582]	0.630 [0.490, 0.759]
L5-7	1.7B	0.609 [0.448, 0.780]	0.487 [0.329, 0.634]
L5-7	8B	0.495 [0.341, 0.668]	0.529 [0.364, 0.686]
L7-10	1.7B	0.602 [0.490, 0.709]	0.425 [0.317, 0.527]
L7-10	8B	0.566 [0.456, 0.679]	0.481 [0.368, 0.596]

H Critique Ablations: Full Results

The ablation result of §7 is reported here with confidence intervals (Table 12), as paired per-rollout drops in which each rollout’s complete critique is compared with its own ablated critique (Table 13), and with the same paired test on the Olmo-3-7B-Think rollouts. The pattern is the same in every

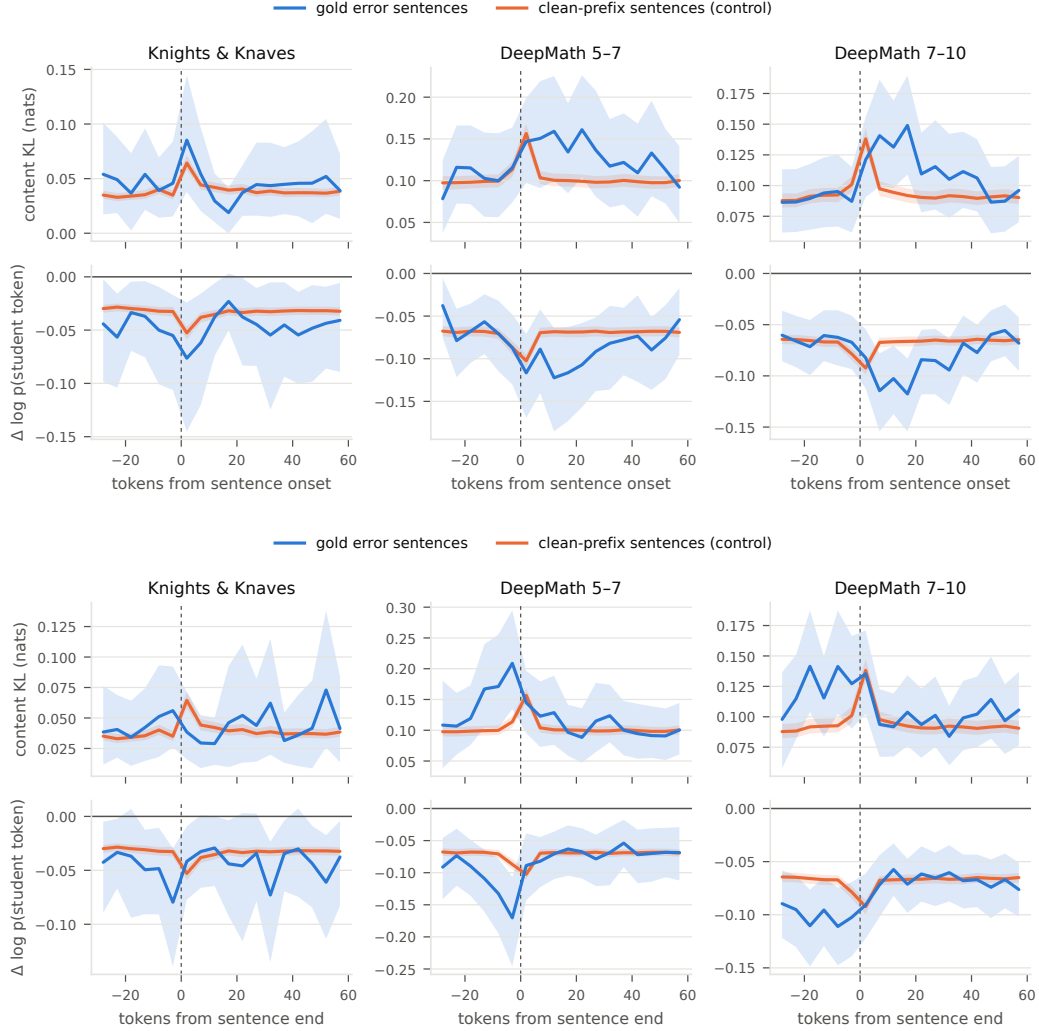


Figure 5: Reference-conditioned signal on all three tasks. Content KL (top row of each block) and $\Delta \log p$ of the student’s token (bottom row) around gold error sentences (blue) and clean-prefix sentences (orange), aligned to the sentence onset (upper block) and to the sentence end (lower block). Qwen3-1.7B scoring its own rollouts with the reference solution in context, minus the shuffled-reference placebo; scale-up sample.

form: removing the quotes lowers localization, removing the reasons lowers correction, and the 7B Olmo-3 student reproduces it.

I Replication on Olmo-3-7B-Think and Olmo-3-32B-Think

The error-span measurements of the main text are repeated on a second model family, with the student, the scoring teacher, and the self-critic all the same model: Olmo-3-7B-Think on a set comparable in size to the Qwen scale-up, and Olmo-3-32B-Think on a smaller set. Table 14 lists what differs from the Qwen pipeline. The Olmo-3 models reason at greater length, so their rollouts are sampled with a 32,768-token budget; the gold-span grader is claude-sonnet-5 rather than a DeepSeek model, which leaves gpt-5.2 as the only jury independent of the grader; part of the K&K references were derived by gpt-5.2 and verified against the ground truth, and the 32B study draws its K&K puzzles from the 632 that have a verified derivation. Everything else, the prompts, the placebo constructions, the probe, and the error-span Z readout, is shared.

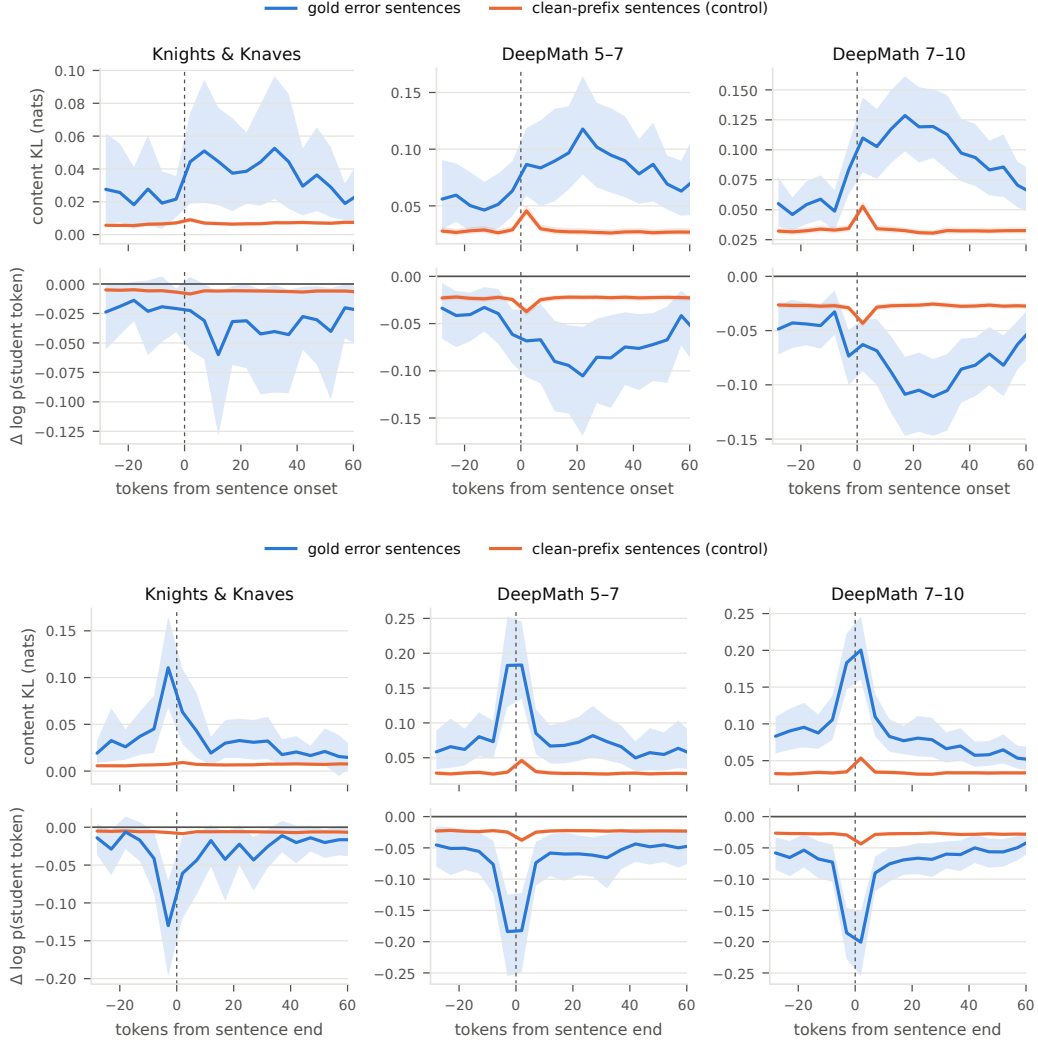


Figure 6: Critique-conditioned signal on all three tasks. Same quantities, alignments, and rollouts as the figure above, with gpt-5.2’s critique of the rollout in the teacher’s context in place of the reference, minus the donor-critique placebo. Teacher Qwen3-1.7B.

The result replicates (Table 15): on both Olmo-3 models the reference-conditioned teacher does not react at the errors, and the critique-conditioned teacher localizes and corrects on every task. The 32B model is the better self-critic: its own critiques localize on all three tasks and correct on DeepMath 7-10, where the 7B’s localize on the two math tasks only. Table 16 gives the critique localization gate for both models next to the full Qwen sweep; the Olmo-3 rollouts are two to three times longer in sentences, so their absolute rates are comparable only within a model.

J Validation on Human-Annotated Error Spans

Our main analysis uses error spans labelled by an LLM jury. To validate the result on human annotations, we repeat the analysis on ProcessBench [Zheng et al., 2025], which provides human-labelled first-error steps for solutions generated by Qwen2.5-7B-Instruct and Qwen2.5-Math-7B-Instruct. We score each solution with the model that generated it and compare reference conditioning against an oracle critique from gpt-5 [OpenAI, 2025b], generated without access to the human label. Table 17 reproduces the main pattern of Table 1: the reference is neither localized nor corrective, while critique conditioning localizes the error for both models. Correction is clear for Qwen2.5-

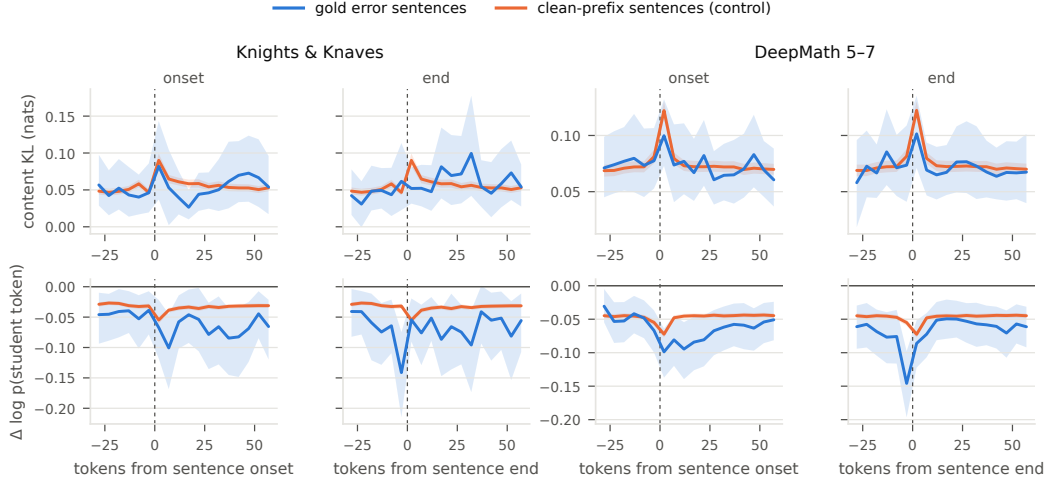


Figure 7: Reference-conditioned signal with Qwen3-8B as the scoring teacher on the Qwen3-1.7B rollouts. Layout as Fig. 1: content KL (top) and $\Delta \log p$ of the student’s token (bottom), Knights & Knaves and DeepMath 5-7, onset- and end-aligned; reference minus shuffled-reference placebo; scale-up sample.

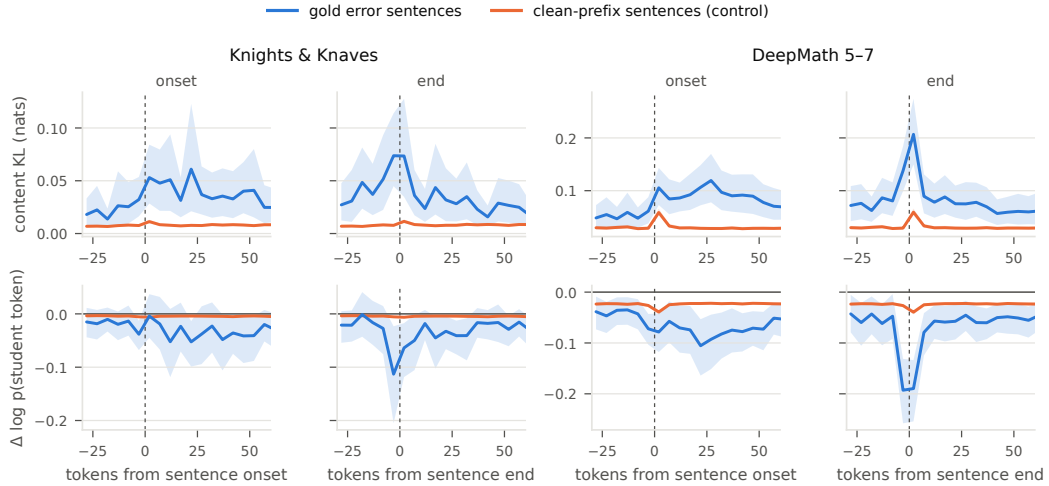


Figure 8: Critique-conditioned signal with Qwen3-8B as the scoring teacher on the Qwen3-1.7B rollouts, gpt-5.2 critiques. Layout as Fig. 3: content KL (top) and $\Delta \log p$ of the student’s token (bottom), Knights & Knaves and DeepMath 5-7, onset- and end-aligned; critique minus donor-critique placebo; scale-up sample.

7B-Instruct and initially weak for Qwen2.5-Math-7B-Instruct; consistent with §7, removing tokens quoted by the critique makes the signal corrective for both. More generally, correction appears when the critique points inside the human-labelled error span and disappears when it points elsewhere. Thus, human annotations reproduce the same picture: reference-conditioned scoring remains flat, while critique-conditioned scoring reacts where the critique identifies the error.

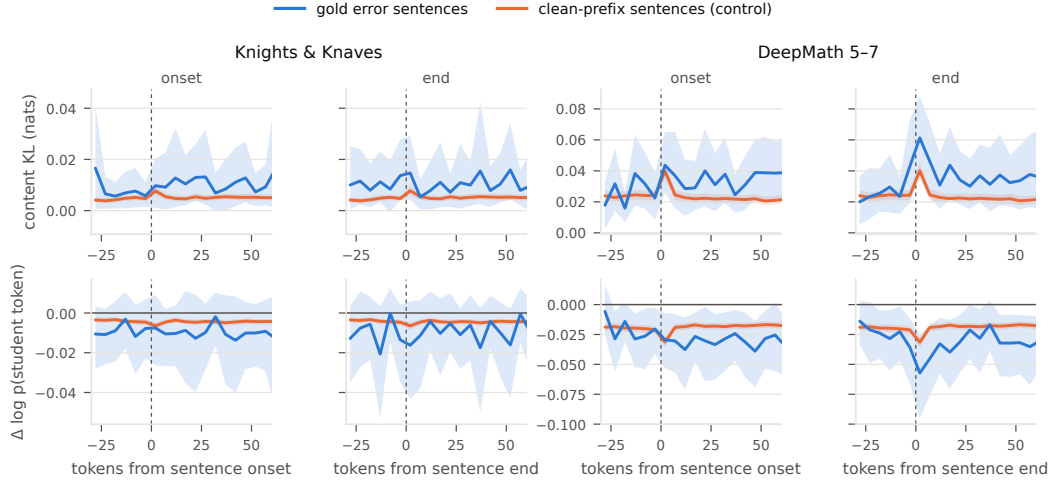


Figure 9: Critique-conditioned signal with the student’s own critiques: Qwen3-1.7B writes the critique, then scores its own rollout with that critique in context. Layout as Fig. 3: content KL (top) and $\Delta \log p$ of the student’s token (bottom), Knights & Knaves and DeepMath 5-7, onset- and end-aligned; critique minus donor-critique placebo; scale-up sample. Note the smaller vertical scale.

Table 11: Trained probe: out-of-fold AUC of an L2 logistic regression on per-trace summaries of each privileged context’s signal, DeepMath L5-7 / L7-10, both teachers. The reference block never beats its content-free placebos.

context	1.7B teacher	8B teacher
length only	.54 / .56	.54 / .56
reference (dense)	.57 / .58	.65 / .64
permuted reference	.60 / .60	.66 / .71
irrelevant reference (placebo)	.57 / .58	.70 / .58
answer only (sparse)	.69 / .75	.78 / .85
critique	.71 / .66	.68 / .69
placebo critique	.63 / .56	.61 / .54

Table 12: Ablating the critique, with 95% confidence intervals: error-span Z when the gpt-5.2 critique in the teacher’s context is complete (quote + reason), stripped of its quotes (reason only), or stripped of its reasons (quote only). Qwen3-1.7B rollouts, scale-up sample; scoring teacher Qwen3-1.7B in the upper block and Qwen3-8B in the lower. Critique minus donor-critique throughout.

critique	task	loc	95% CI	corr	95% CI
quote + reason, 1.7B teacher	K&K	+0.75	[+0.54, +0.97]	−0.62	[−0.86, −0.40]
	L5-7	+0.99	[+0.82, +1.16]	−0.90	[−1.14, −0.65]
	L7-10	+0.78	[+0.63, +0.92]	−0.54	[−0.71, −0.35]
reason only, 1.7B teacher	K&K	+0.48	[+0.31, +0.67]	−0.47	[−0.67, −0.26]
	L5-7	+0.75	[+0.60, +0.89]	−0.58	[−0.81, −0.36]
	L7-10	+0.61	[+0.48, +0.75]	−0.30	[−0.45, −0.16]
quote only, 1.7B teacher	K&K	+0.71	[+0.52, +0.90]	−0.16	[−0.39, +0.06]
	L5-7	+0.94	[+0.75, +1.14]	−0.39	[−0.58, −0.20]
	L7-10	+0.69	[+0.55, +0.85]	−0.27	[−0.43, −0.14]
quote + reason, 8B teacher	K&K	+1.20	[+0.95, +1.45]	−0.79	[−1.10, −0.49]
	L5-7	+1.09	[+0.91, +1.27]	−0.96	[−1.25, −0.68]
	L7-10	+0.91	[+0.77, +1.05]	−0.62	[−0.79, −0.45]
quote only, 8B teacher	K&K	+0.82	[+0.63, +1.02]	+0.21	[−0.03, +0.45]
	L5-7	+1.17	[+0.98, +1.36]	−0.22	[−0.45, −0.01]
	L7-10	+0.90	[+0.75, +1.05]	+0.03	[−0.10, +0.17]

Table 13: Paired per-rollout ablation drops with 95% CI: error-span Z of the complete critique minus that of the ablated critique on the same rollouts, gpt-5.2 critiques. Upper block: Qwen3-1.7B rollouts and teacher, scale-up sample. Lower block: Olmo-3-7B-Think rollouts and teacher, big set. A positive drop on localization or a negative drop on correction means the ablation removed signal; *: CI excludes 0.

student	signal	ablation	K&K	L5-7 / L7-10
Qwen3-1.7B	localization	remove quote	+0.30*	+0.35* / +0.19*
		remove reason	-0.00	+0.09 / +0.02
	correction	remove quote	-0.14	-0.39* / -0.27*
		remove reason	-0.35*	-0.44* / -0.32*
Olmo-3-7B-Think	localization	remove quote	+0.28* [+0.15, +0.42]	+0.27* [+0.19, +0.34] / +0.22* [+0.16, +0.28]
		remove reason	+0.09 [-0.06, +0.25]	+0.05 [-0.04, +0.14] / +0.01
	correction	remove quote	-0.05	-0.22* [-0.32, -0.12] / -0.15* [-0.23, -0.08]
		remove reason	-0.13	-0.13* [-0.25, -0.02] / -0.20*

Table 14: The three error-span studies side by side: models, rollout sampling, gold-span pipeline, reference solutions, and the number of problems and rollouts that enter error-span Z on each task (Knights & Knaves / DeepMath 5-7 / DeepMath 7-10).

	Qwen3-1.7B (scale-up)	Olmo-3-7B-Think (big set)	Olmo-3-32B-Think (minimal set)
student = teacher = self-critic	yes (plus 8B teacher / critic)	yes	yes
rollout cap	16,384	32,768	32,768
sampling	temp 1.0, top-p 0.95, top-k 20, $n=8$	same	same
gold-span grader	deepseek-reasoner / v4-flash	claude-sonnet-5	claude-sonnet-5
juries	gpt-5.2 + claude-sonnet-5	same (gpt-5.2 independent)	same
oracle critiques	gpt-5.2, same prompt	gpt-5.2, same prompt	gpt-5.2, same prompt
K&K references	verified deepseek derivations	452 deepseek + 180 gpt-5.2, verified	632 puzzles with a verified derivation
problems in error-span Z (K&K / L5-7 / L7-10)	355 / 516 / 794	197 / 398 / 631	114 / 111 / 133
rollouts in error-span Z	317 / 448 / 711	271 / 584 / 958	160 / 152 / 208

Table 15: Error-span Z at the first gold error with 95% confidence intervals for the two Olmo-3 students, each scoring its own rollouts: Olmo-3-7B-Think on the big set and Olmo-3-32B-Think on the minimal set. Rows: reference solution in context; gpt-5.2 critique in context; the student’s own critique in context. Localization on log content KL, correction on $\Delta \log p$ of the student’s token; n = rollouts entering the cell.

student	conditioning	task	loc	95% CI	corr	95% CI	n
Olmo-3-7B	reference	K&K	+0.20	[+0.01, +0.39]	-0.00	[-0.18, +0.18]	271
		L5-7	+0.11	[-0.03, +0.25]	-0.08	[-0.23, +0.08]	584
		L7-10	-0.01	[-0.11, +0.09]	-0.03	[-0.14, +0.08]	958
	critique, oracle	K&K	+0.63	[+0.39, +0.87]	-0.29	[-0.55, -0.03]	271
		L5-7	+0.89	[+0.72, +1.07]	-0.59	[-0.78, -0.40]	584
		L7-10	+0.88	[+0.75, +1.01]	-0.49	[-0.63, -0.33]	958
	critique, self	K&K	+0.13	[-0.08, +0.34]	-0.14	[-0.32, +0.04]	271
		L5-7	+0.41	[+0.27, +0.56]	-0.18	[-0.32, -0.04]	581
		L7-10	+0.33	[+0.21, +0.44]	-0.14	[-0.25, -0.03]	953
Olmo-3-32B	reference	K&K	+0.09	[-0.17, +0.35]	-0.41	[-0.72, -0.10]	160
		L5-7	+0.21	[-0.06, +0.49]	-0.09	[-0.33, +0.14]	152
		L7-10	+0.15	[-0.09, +0.37]	-0.06	[-0.32, +0.19]	208
	critique, oracle	K&K	+0.76	[+0.46, +1.07]	-0.43	[-0.79, -0.07]	160
		L5-7	+0.96	[+0.66, +1.27]	-0.50	[-0.83, -0.19]	152
		L7-10	+1.20	[+0.89, +1.48]	-0.62	[-0.99, -0.26]	208
	critique, self	K&K	+0.47	[+0.21, +0.72]	-0.01	[-0.30, +0.25]	160
		L5-7	+0.54	[+0.26, +0.83]	-0.39	[-0.85, +0.05]	150
		L7-10	+0.40	[+0.13, +0.65]	-0.42	[-0.73, -0.12]	206

Table 16: Critique localization gate for the two Olmo-3 students (oracle and self-critic) and the full window sweep for the Qwen3-1.7B rollouts. Percent of wrong rollouts with a gold span for which a quote lands within w sentences of a gold error sentence; lenient tier; relaxed parse for self-critics. Olmo-3 rollouts run 800–1,500 sentences against 450–550 for Qwen, so absolute rates are comparable only within a student.

student	critic	task	± 2	± 5	± 10	med. dist	quote-match	parse
Qwen3-1.7B	gpt-5.2 (oracle)	K&K / L5-7 / L7-10	48 / 54 / 59	57 / 62 / 67	65 / 69 / 73	3 / 2 / 1	98 / 96 / 95	100
	Qwen3-8B		15 / 13 / 16	26 / 22 / 27	40 / 31 / 37	17 / 18 / 13	96 / 82 / 83	99 / 86 / 85
	Qwen3-1.7B self, relaxed		9 / 12 / 13	20 / 18 / 21	33 / 26 / 31	16 / 23 / 16	77 / 64 / 68	90 / 83 / 83
	Qwen3-1.7B self, strict		4 / 8 / 9	9 / 13 / 15	16 / 19 / 22	12 / 16 / 13	33 / 42 / 45	43 / 54 / 53
Olmo-3-7B	gpt-5.2 (oracle)		31 / 37 / 46	39 / 45 / 55	43 / 51 / 64	16 / 8 / 3	97 / 89 / 94	100
	Olmo-3-7B self, relaxed		3 / 6 / 7	7 / 9 / 12	10 / 12 / 17	71 / 61 / 42	74 / 58 / 61	92 / 84 / 87
Olmo-3-32B	gpt-5.2 (oracle)		26 / 32 / 47	33 / 38 / 56	42 / 46 / 62	22 / 13 / 3	97 / 86 / 95	100
	Olmo-3-32B self, relaxed		4 / 9 / 13	8 / 11 / 21	12 / 17 / 24	110 / 76 / 51	87 / 80 / 87	98 / 93 / 97

Table 17: **On human-labelled error steps, the reference is flat and the critique reacts.** Process-Bench, each solution teacher-forced through the Qwen2.5 model that wrote it. Error-step Z as in Table 1: loc = log content KL (positive = reacts at the labelled step), corr = $\Delta \log p$ of the student’s token (negative = corrective), hint minus placebo. Bold = 95% CI excludes 0. The critique is written by gpt-5 without the label.

conditioning	Qwen2.5-7B-Instruct		Qwen2.5-Math-7B	
	loc	corr	loc	corr
reference	+0.18	−0.34	+0.27	−0.26
critique, oracle	+1.25	−0.55	+0.80	+0.13