DISCERN: DECOUPLING REASONING AND VERIFICATION FOR SELECTIVE ABSTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

An external verifier can help a language model abstain from incorrect answers, but its training responses determine which errors it encounters. Can the verifier itself guide the generation of more useful training responses? We propose DISCERN, an answer-supervised chunk-scoring verifier trained through a sparring curriculum. A temporary solver learns to obtain high scores from the current verifier; final-answer labels then distinguish correct responses from errors that received high scores, providing supervision for the next verifier update. Deployment requires no generator weights or logits. On 100 greedy BeyondAIME responses, the final 4B verifier achieves mean-score AUC 0.913. In the continuation comparison on frozen-base BeyondAIME responses, sparring raises mean-score AUC from 0.802 to 0.869, compared with 0.706 for static refresh with an unchanged solver, and lowers expected risk relative to round 1 by 16.5 percentage points at 20% coverage. Same-trace ProcessBench comparisons extend evaluation to GSM8K, MATH, OlympiadBench, and Omni-MATH. Tests with a 30B verifier and external generators broaden the evaluation, while SimpleQA remains difficult. The clearest gains are in mathematical response selection, particularly when accepting only a small fraction of responses.

1 INTRODUCTION

Language models can produce fluent multi-step solutions while misjudging their correctness. Self-assessed confidence is sensitive to prompts and distribution shifts (Kadavath et al., 2022), and training refusal is difficult when examples of intermediate difficulty are sparse (Wu et al., 2025). An external verifier separates answering from deciding whether to answer: given a question and a completed response, it assigns a score used to accept or withhold that response. This makes selective abstention available without modifying the deployed generator.

This separation leaves a training problem: which responses should the verifier learn to assess as its capabilities change? Sampling again from the same solver refreshes the examples, but leaves the data-generating policy unchanged. We instead use the verifier's current scores to update a temporary solver before generating the next training corpus, so the response distribution can adapt along with the verifier. The resulting responses are checked against final answers, so a high verifier score does not by itself make a response a positive training example. This lets verifier feedback guide where to look for training examples while answer correctness determines how to learn from them.

We implement this *sparring curriculum* in **DISCERN**, a verifier that annotates responses with knowledge and inference chunks and scores each chunk. The mean score supports default response selection; the minimum provides a conservative weakest-link reference. Because final-answer labels do not specify how scores should vary across chunks, we combine minimum-score error detection, difficulty-aware mean-score targets, and trajectory preferences. A bucketed training corpus increases the representation of intermediate-difficulty questions. Semantic tags expose the annotation for inspection, without requiring gold step-error labels. Only the verifier is retained for deployment.

Our contributions are: (1) a verifier-guided response-generation curriculum, tested against fresh-data continuation with the original solver held fixed; (2) an answer-supervised reward recipe for learning chunk scores; and (3) evidence that the resulting verifier improves mathematical response selection, with the clearest matched-coverage gains over round 1 at low coverage. Reward removals

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.

examine how the objectives affect ranking, and externally supplied traces test transfer beyond the initial solver. A separate audit finds 83.0% human agreement with the semantic type annotations.

2 Related Work

Uncertainty estimation includes verbalized confidence, latent statistics, multi-sample consistency, and semantic entropy (Xia et al., 2025; Kadavath et al., 2022; Malinin & Gales, 2020; Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). Fine-grained confidence methods (Wu et al., 2025) provide a related weakest-link interface. DISCERN learns chunk scores with a multi-term reward and an external-verifier curriculum.

Refusal and factuality alignment use question- or response-level supervision (Zhang et al., 2024; Tjandra et al., 2024; Wei et al., 2025; An & Xu, 2025; Xue et al., 2025). Process reward models instead score intermediate steps, often for reasoning search (Lightman et al., 2023; Zhang et al., 2025). We compare with an outcome-only trained verifier and an off-the-shelf PRM. Extended discussion and baseline protocols appear in Appendices L and M.

3 Method

3.1 Data and verifier interface

To expose the verifier to questions that the solver answers inconsistently, we construct a difficulty-bucketed corpus from DAPO-Math-17k. For each question q_i, the base Qwen3-4B-Instruct solver generates ten responses; their empirical final-answer accuracy p_i serves as a solver-specific difficulty proxy. We retain all questions with $p_i < 1$ and subsample the perfectly solved bucket to the mean size of the ten non-perfect buckets. This reduces the dominance of perfectly solved questions while retaining the other difficulty levels. After removing incomplete traces, the corpus contains 10,248 problems and 101,585 traces (Appendix K).

An independently initialized Qwen3-4B-Instruct verifier receives a question q and the solver's response a, including its proposed answer, and emits tagged chunks with scores $c_k \in [0, 1]$. Knowledge chunks identify facts, formulas, or theorems; inference chunks identify derivations and calculations. This vocabulary makes the scored response easier to inspect. The verifier may lightly paraphrase intermediate reasoning and is instructed to preserve the boxed final answer. Training uses final-answer correctness as a target, not as a verifier input, and requires no human step-error labels. The content reward supervises discrimination and score trajectories independently of tag identity.

Our default policy accepts a response if $\bar{c} = K^{-1} \sum_{k=1}^{K} c_k \geq t$ and abstains otherwise. We choose mean confidence for its stronger observed ranking (Table 1) and finer score resolution. The minimum $c_{\min} = \min_k c_k$ remains a training statistic and a conservative reference: at the same threshold and boundary convention, it accepts a subset of the mean policy's responses. Both aggregations provide learned decision scores.

3.2 Answer-supervised reward

A final-answer label tells us whether a response ends with the right answer, but does not assign a score to each chunk. We therefore give the reward three roles: identify incorrect responses, set difficulty-aware mean scores, and shape score changes along a response. The verifier is trained with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), using the sequence-level reward

$$R = R_{\text{format}}\big(0.4R_{\text{detect}} + 0.4R_{\text{accuracy}} + 0.2R_{\text{topo}}\big).$$

The binary format gate checks complete valid tagging and boxed-answer consistency. The round-1 gate tolerates missing boxed answers and rejects a mismatch when both answers are extractable (Appendix K.3). Let $y \in \{0, 1\}$ denote final-answer correctness. The detection term encourages a low minimum on incorrect responses and a high minimum on correct ones:

$$R_{\text{detect}} = y\, c_{\min} + (1 - y)(1 - c_{\min}).$$

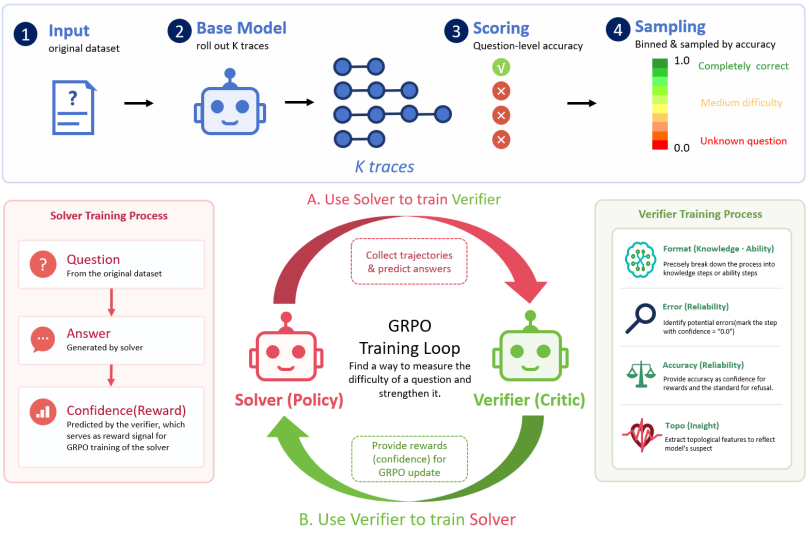

Figure 1: Sparring curriculum. Verifier scores train a temporary solver; regenerated, answer-labeled responses train the continuing verifier. Deployment uses the verifier on a frozen generator.

With $\mathrm{Tri}(u, v) = 1 - |u - v|$, the difficulty-aware term is

$$R_{\text{accuracy}} = y\ \mathrm{Tri}(\bar{c}, 1) + (1 - y)\,\mathrm{Tri}(\bar{c}, p_{\text{prior}}),$$

where p_{prior} is the empirical solver success rate for the question. Detection alone permits every chunk of an incorrect response to receive zero; when $p_{\text{prior}} > 0$, the mean-score target discourages that uniform suppression. The difficulty proxy thus sets a training preference for the score profile, rather than supplying intermediate-step labels.

The trajectory term R_{topo} adds an ordering preference that the mean and minimum alone cannot express. It rewards high-confidence plateaus for correct answers and local or head-to-tail drops, or low mean confidence, for incorrect answers. Its statistics and fixed cutoffs appear in Appendix K. Together, the terms train the minimum for error detection and the mean for response selection, while shaping the trajectory of chunk scores.

For each supplied question–response pair, GRPO samples tagged outputs and normalizes their sequence rewards to obtain relative advantages. These advantages train both the annotation and its discrete confidence outputs. The format gate suppresses rewards for invalid tagging or detected boxed-answer mismatches; correctness labels remain attached to the supplied solver response.

3.3 Sparring Curriculum

Starting from verifier $V^{(t)}$, we update a temporary solver with GRPO to maximize

$$\mathbb{E}_{q\sim\mathcal{D},\ a\sim\pi_{\theta_{\text{solver}}}(\cdot|q)}[c_{\min}^{V^{(t)}}(q, a)].$$

Optimizing verifier reward can produce either correct responses or errors that the verifier misses. Final-answer checking separates these cases: after regeneration, we assign correctness labels and continue training from $V^{(t)}$ on the refreshed corpus. In this way, a response that receives high verifier confidence but has an incorrect answer becomes a negative example for the next update. Deployment retains the second-round verifier selected at the validation plateau.

The natural comparison is fresh sampling from an unchanged solver: does updating the response generator help beyond replacing old examples with new ones? We continue the same round-1 verifier on responses from either the unchanged solver (static refresh) or the verifier-updated solver (sparring). Both branches receive fresh responses. Their evaluation on aligned solver responses tests the usefulness of the resulting training corpora.

4 EXPERIMENTS

4.1 SETUP AND COMPARISON SCOPE

We first ask whether DISCERN selects better answers from fixed responses, then examine the reward and curriculum choices behind those gains. External mathematical traces and factual questions test where the learned scores transfer.

The base solver and initial verifiers use Qwen3-4B-Instruct. DISCERN is trained on DAPO-Math-17k (Yu et al., 2025) with GRPO and continues verifier training across curriculum rounds. Round-1 and sparring verifiers use 240 updates per round, eight rollouts per prompt, training batch size 256, and actor learning rate 10^{-6}. BeyondAIME (ByteDance Seed et al., 2025) (100 problems) and AIME-2024/2025 (30 each) evaluate mathematical response selection; SimpleQA (Wei et al., 2024) evaluates factual transfer. Full configurations and control-specific settings appear in Appendix M.

We evaluate the final system under greedy decoding and use five-seed inference evaluations for the reward and continuation controls. Within each control comparison, the verifiers assess the same supplied solver responses under shared final-answer correctness labels. Mean confidence is the default response-selection score, with minimum confidence reported as a conservative reference. The comparison-specific settings and five-seed summaries are in Appendices M and A.

Controls. The untrained verifier measures the effect of verifier training. The trained *single-reward control* compares outcome-only and full training recipes with the same chunk interface and initial training corpus; component removals assess the individual content-reward terms. Static refresh and sparring compare continuation with responses from an unchanged or verifier-adapted solver, respectively. Table 1 places DISCERN alongside trained confidence and PPO baselines from Wu et al. (2025) and reproduced UALIGN and a response-probability adaptation of CISC (Xue et al., 2025; Taubenfeld et al., 2025); same-trace PRM comparisons additionally assess a pretrained process scorer. Baseline definitions and reproduction protocols appear in Appendix M.

Metrics. We distinguish ordering responses from deciding how many to accept. Conf AUC measures ranking, while selective risk is the error rate among accepted responses and AURC integrates that risk over coverage. SNR-Gain (Wu et al., 2025) measures a complementary quantity: threshold-integrated accepted-correct versus accepted-incorrect mass, relative to always answering. Under continuous integration it simplifies to

$$\text{SNR-Gain} = \log \frac{\mathbb{E}[s \mid y = 1]}{\mathbb{E}[s \mid y = 0]}.$$

This threshold-integrated metric measures score-weighted enrichment and depends on the numerical score scale. AUC and AURC assess selection quality; SNR-Gain supplies complementary enrichment evidence. Each aggregation is reported explicitly. Metric definitions, numerical integration, and secondary calibration diagnostics are in Appendix C.1.

4.2 ENDPOINT RANKING AND SELECTIVE ABSTENTION

Comparison with confidence and value-based baselines. Trained confidence and value-based methods provide a useful benchmark for an external verifier (Table 1). On BeyondAIME, DISCERN's mean-score AUC is 0.913, compared with 0.902 for Confidence-Brier and 0.881 for PPO-Value; its mean-score SNR-Gain is also higher than the reported trained confidence and PPO baselines. This places DISCERN competitively in the benchmark comparison. The table includes published results from Wu et al. (2025); we next turn to fixed-response controls to examine the effect of DISCERN training directly.

Same-trace untrained control. Training improves ranking on the same 100 greedy BeyondAIME responses: mean-score AUC rises from 0.784 to 0.913. The aggregation matters. Minimum-score AUC rises more modestly, from 0.762 to 0.791, despite a large increase in minimum SNR-Gain from 0.107 to 1.746. The mean therefore captures a stronger ranking gain on these responses, motivating its use for default selection. AIME results appear in Table 36.

Table 1: Main benchmark comparison on BeyondAIME and SimpleQA. DISCERN uses greedy decoding; mean is the default abstention score, minimum a conservative reference. The untrained 2507 verifier uses the same 100 BeyondAIME traces and prompt as DISCERN (33 correct answers). Published confidence/PPO results follow Wu et al. (2025); baseline provenance and evaluation protocols are specified in Appendix M. †UALIGN is reproduced on the same 100 BeyondAIME problems. smECE and Brier are secondary calibration diagnostics.

	BeyondAIME				SimpleQA			
Model	SNR-Gain ↑	Conf AUC ↑	smECE ↓	Brier ↓	SNR-Gain ↑	Conf AUC ↑	smECE ↓	Brier ↓
Qwen3-next-80b-a3b-thinking	0.006	0.496	0.408	0.425	–	–	–	–
GLM-4.5	0.016	0.519	0.440	0.449	–	–	–	–
Qwen3-max	0.043	0.678	0.453	0.413	–	–	–	–
Gemini-2.5-Pro	0.071	0.699	0.431	0.387	0.017	0.556	0.451	0.436
Qwen3-4B-Instruct	0.084	0.715	0.632	0.599	0.066	0.561	0.821	0.762
Kimi-K2-0905	0.087	0.678	0.575	0.529	–	–	–	–
Qwen3-4B-Instruct-ppo	0.101	0.786	0.560	0.518	–	–	–	–
GPT-oss-120b	0.154	0.703	0.419	0.402	–	–	–	–
GPT-5	0.207	0.706	0.160	0.208	0.498	**0.838**	**0.098**	**0.178**
o4-mini	0.515	0.738	0.164	0.241	–	–	–	–
Qwen3-4B-Instruct-confidence-min	0.683	0.863	0.220	0.194	–	–	–	–
Qwen3-4B-Instruct-confidence-brier	0.802	0.902	0.177	0.161	0.411	0.704	0.459	0.341
Qwen3-4B-Instruct-confidence-prod	0.806	0.883	0.166	0.166	–	–	–	–
Qwen3-4B-Instruct-confidence-ce	1.097	0.854	0.174	0.179	–	–	–	–
Qwen3-4B-Instruct-ppo-value	1.202	0.881	**0.061**	0.141	0.211	0.734	0.665	0.525
UALIGN†	1.243	0.637	0.224	0.230	–	–	–	–
CISC (response-probability adaptation)	0.436	0.779	0.299	0.273	–	–	–	–
Grok-4	–	–	–	–	0.147	0.664	0.307	0.327
Claude-sonnet-4.5-nothinking	–	–	–	–	0.282	0.748	0.318	0.284
GLM-4.6	–	–	–	–	0.358	0.739	0.377	0.310
Claude-sonnet-4.5-thinking	–	–	–	–	0.456	0.783	0.200	0.214
Gemini-2.5-Flash	0.020	0.630	0.569	0.522	–	–	–	–
Gemini-2.5-Flash (Two-pass Reflection)	0.212	0.661	0.489	0.448	–	–	–	–
Gemini-2.5-Flash + Our Verifier	0.500	0.724	0.304	0.308	–	–	–	–
Qwen3-4B-Instruct-2507 (mean)	0.045	0.784	0.705	0.650	–	–	–	–
Qwen3-4B-Instruct-2507 ($c_{\min}$)	0.107	0.762	0.637	0.567	–	–	–	–
DISCERN (mean)	1.456	**0.913**	0.092	**0.123**	0.441	0.674	0.291	0.321
DISCERN ($c_{\min}$)	**1.746**	0.791	0.160	0.160	**0.696**	0.644	0.225	0.251

Table 2: Default mean-score selection on BeyondAIME: paired expected-risk differences at matched coverage. Negative differences favor the first method. Panels identify separate evaluation cohorts; bracketed ranges are pointwise paired intervals.

Contrast	Coverage	$\Delta\mathbb{E}[R]$	Paired interval
Final system − Single-reward	20%	-0.231	$[-0.316, -0.145]$
Sparring − Round 1 (frozen-base)	20%	-0.165	$[-0.236, -0.095]$
	30%	-0.135	$[-0.183, -0.032]$
	50%	-0.032	$[-0.076, +0.014]$

Default-policy selective prediction. Better ranking is useful if it lowers the error rate among the answers we choose to keep. At 20% coverage, the final system has expected risk 0.282 versus 0.513 for the single-reward control, a paired difference of -0.231 with interval $[-0.316, -0.145]$. Table 2 places this complete-training-recipe comparison alongside sparring versus its round-1 starting checkpoint. Each contrast uses the same supplied responses within its own evaluation cohort. The latter comparison also lowers mean-score AURC from 0.556 to 0.483; Section 4.3 examines where the risk reduction occurs.

Appendix A.1 reports multi-dataset risk–coverage curves and operating points for the conservative minimum reference.

4.3 Reward and curriculum controls

Reward recipe. The full recipe improves score-weighted enrichment over outcome-only training: within-round minimum-score SNR-Gain rises from 0.862 ± 0.094 to 1.148 ± 0.147, with paired hierarchical-bootstrap difference $+0.286$ and 95% interval $[+0.120, +0.479]$ (Table 9). Component

Table 3: One-round reward-removal diagnostics on 100 BeyondAIME traces (33 correct; five inference seeds, mean ± SD). SNR-Gain follows the protocol in Table 39; mean and minimum AUC are shown separately. Table 9 reports the outcome-only reward control.

Reward	SNR-Gain	Mean AUC	Minimum AUC
Full	1.235 ± 0.045	0.781 ± 0.003	0.724 ± 0.004
Without topology	1.118 ± 0.045	0.760 ± 0.012	0.753 ± 0.006
Without accuracy	1.087 ± 0.092	0.768 ± 0.019	0.735 ± 0.011
Without detection	1.005 ± 0.000	0.729 ± 0.000	0.729 ± 0.000

Table 4: BeyondAIME continuation comparison (five inference seeds; mean ± SD). Static refresh continues verifier training with new responses from the unchanged solver; sparring uses responses from the updated solver. Columns identify the evaluation generator. Confidence metrics use parse-valid outputs.

Verifier	Base evaluation responses	Updated-solver responses
	SNR-Gain (mean confidence)	
Round 1	1.065 ± 0.108	0.653 ± 0.172
Static refresh	0.780 ± 0.259	0.790 ± 0.181
Sparring continuation	$\mathbf{1.324 \pm 0.173}$	$\mathbf{1.119 \pm 0.121}$
	Conf AUC	
Round 1	0.802 ± 0.029	0.716 ± 0.064
Static refresh	0.706 ± 0.061	0.724 ± 0.050
Sparring continuation	$\mathbf{0.869 \pm 0.046}$	$\mathbf{0.836 \pm 0.026}$

removals ask which objectives help the jointly trained scores (Table 3). The full reward gives the highest mean-score AUC and lowest mean-score AURC (0.542), whereas the minimum-score AUC ordering favors the removal variants. Thus, the observed ranking benefit is specific to the mean aggregation. Paired fixed-coverage risk intervals mostly include zero (Appendix E.2).

Producing valid annotations alone is not enough to obtain useful scores. A separate verifier trained with reference-aware, Qwen3-4B-judged binary rationale rewards produces parseable outputs in 98.25% of evaluations, yet its mean-score AUC is 0.483–0.527 across the three mathematical sets, versus 0.801–0.861 for round 1 (Appendix E.5). This provides a complementary comparison of supervision strategies.

Curriculum. Does verifier-guided solver adaptation help more than drawing fresh responses from the original solver? Table 4 compares both continuation branches with their round-1 starting checkpoint on the same supplied responses. On frozen-base BeyondAIME, sparring raises mean-score AUC from 0.802 to 0.869 and SNR-Gain from 1.065 to 1.324; static refresh reaches 0.706 and 0.780, respectively. The gain over round 1 shows that sparring improves the starting verifier, while the refresh comparison shows that new responses alone do not yield the same benefit in this experiment. Validation reward also declined in both static-continuation configurations, motivating adaptation of the training distribution as the verifier learns (Appendix E.4). Sparring exceeds static refresh in AUC by 0.163 on frozen-base and 0.112 on updated-solver responses. Appendix E.4 gives the three-dataset results and identifies the separate reference checkpoint used in additional diagnostics.

Mean-score risk at matched coverage. The risk reduction is concentrated among the highest-scoring responses. Relative to round 1, sparring lowers expected risk by 16.5 percentage points at 20% coverage and 13.5 points at 30% coverage (Table 2). Both paired intervals favor sparring; at 50% coverage the interval includes zero. The clearest practical gain is therefore in selecting a smaller, more reliable subset. Appendix E.2 reports the full grid and a separate static-reference comparison.

Continuation checkpoints score the same input responses, with confidence metrics evaluated on each checkpoint's parse-valid outputs. Risk–coverage evaluation retains parser failures below

valid scores. Reference-checkpoint diagnostics and primary static-refresh results appear in Appendices E.1 and E.4, respectively.

Which errors does the next verifier repair? To examine how the verifier handles responses from its training opponent, we cross-evaluate both verifier rounds on frozen-base and updated-solver responses. Table 5 pools the three mathematical datasets over five inference seeds and measures false acceptance among incorrect answers at mean confidence ≥ 0.5. Round 1 accepts 33.2% of updated-solver errors, compared with 24.47% of base-solver errors. Round 2 reduces these rates to 9.4% and 15.92%, respectively: improved rejection of the opponent's errors accompanies improvement on the original solver's errors.

Table 5: Pooled mathematics false-acceptance rate among incorrect responses at mean confidence ≥ 0.5. Each cell reports accepted incorrect responses divided by all incorrect responses from that solver. The columns contain frozen-base errors ($N = 515$) and updated-solver errors ($N = 500$, 15 fewer incorrect responses). Table 13 decomposes these pooled rates into item-level transitions.

Verifier	Frozen base solver ($N = 515$)	Updated sparring solver ($N = 500$)
Round-1 DISCERN verifier	126/515 (24.47%)	166/500 (33.20%)
Round-2 DISCERN verifier	82/515 (15.92%)	47/500 (9.40%)

Tracking individual errors reveals that round 2 repairs 129 of the 166 updated-solver errors accepted by round 1 (77.7%), versus 55 of 126 on the base population (43.7%). Newly missed errors account for similar fractions of the two populations: 10/500 (2.00%) and 11/515 (2.14%). A problem-clustered bootstrap over 139 unique problems, retaining all five seeds, gives a 95% interval of $[+23.1, +44.9]$ percentage points for the updated-minus-base conditional-repair difference (Appendix A.2). These transitions describe recovery on the two error populations; the continuation comparison separately tests adaptation against static refresh. The opponent supplies curriculum examples, with its own answer accuracy reported in Table 40.

Format-gate ablation. A format check also determines which outputs can receive a learning signal. Holding the content-reward terms and weights unchanged, tolerating a missing boxed answer yields higher mean-score AUC and lower AURC than suppressing the reward whenever either extraction is missing (Table 6). The advantage appears on all three mathematical datasets; BeyondAIME AUC is 0.802 versus 0.708. The strict gate has higher minimum-score SNR-Gain on AIME-2024, while the tolerant gate consistently improves mean-score ranking. This supports the original gate's distinction between a detected answer mismatch and a missing extraction. Operating-point diagnostics appear in Appendix E.3.

Table 6: Format-gate ablation on frozen-base responses (mean $\pm$ SD over five inference seeds). Tolerant is the original round-1 checkpoint; strict is the boxed-answer gate ablation. Content-reward terms and weights are unchanged. SNR-Gain uses the indicated aggregation; mean-score AUC is higher-is-better and AURC lower-is-better.

Dataset	Gate	SNR (mean)	SNR ($c_{\min}$)	Conf AUC	AURC
AIME-2024	Tolerant	0.986 ± 0.328	1.022 ± 0.339	0.801 ± 0.060	0.198 ± 0.075
	Strict	0.558 ± 0.109	1.718 ± 0.883	0.733 ± 0.051	0.241 ± 0.080
AIME-2025	Tolerant	1.296 ± 0.440	1.388 ± 0.558	0.861 ± 0.056	0.291 ± 0.063
	Strict	0.663 ± 0.260	1.075 ± 0.425	0.762 ± 0.092	0.384 ± 0.039
BeyondAIME	Tolerant	1.065 ± 0.108	1.126 ± 0.105	0.802 ± 0.029	0.556 ± 0.052
	Strict	0.527 ± 0.145	0.705 ± 0.296	0.708 ± 0.057	0.593 ± 0.062

4.4 Evaluation on externally supplied mathematical traces

ProcessBench (Zheng et al., 2024) tests whether the verifier can detect errors in externally supplied mathematical solutions. Unlike final-answer selection, this task asks whether any step in a trace is erroneous, so we use the minimum reference to flag a weak link. Official GSM8K and MATH sets

contain 400 and 1,000 supplied responses, respectively, and compare DISCERN with the UALIGN judge. OlympiadBench and Omni-MATH compare against Qwen2.5-Math-PRM-7B. Both methods assess the same traces within each comparison.

Table 7: Threshold-specific whole-trace error rejection on ProcessBench using the minimum-reference policy. Error recall measures rejection of erroneous traces; correct pass measures acceptance of fully correct traces. F1 is their harmonic mean, following this process-level protocol. The methods score the same fixed traces at the specified operating points; PRM and UALIGN are external comparison models.

Dataset	Method	Error recall	Correct pass	F1
GSM8K	DISCERN	56.0	94.3	70.3
	UALIGN judge	53.6	96.9	69.0
MATH	DISCERN	76.3	86.9	81.3
	UALIGN judge	45.1	92.1	60.6
OlympiadBench	DISCERN	83.1	74.0	78.3
	PRM-7B	95.9	63.4	76.4
Omni-MATH	DISCERN	89.7	67.6	77.1
	PRM-7B	98.2	56.0	71.3

Under this minimum-reference protocol, DISCERN rejects on GSM8K/MATH when $c_{\min} \leq 0.5$; UALIGN uses binary step judgments. For the hard-set PRM comparison, thresholds maximize F1 on GSM8K and transfer to OlympiadBench/Omni-MATH: 0.20 for DISCERN and 0.85 for the PRM. Table 7 shows the resulting trade-offs. On the hard sets, the PRM detects more erroneous traces, while DISCERN preserves more fully correct traces and obtains higher whole-trace F1. Related-step localization results appear in Appendix F.

Process correctness and final-answer correctness are evaluated separately. Using shared final-answer labels on MATH, DISCERN has 75.9% accepted-answer precision at 49.4% coverage, compared with 60.4% precision at 70.0% coverage for UALIGN. On GSM8K, the respective precision/coverage pairs are 68.1%/68.25% and 67.5%/70.75%. These results describe the reported threshold-specific precision–coverage trade-offs. The full confusion counts are retained in Table 30.

4.5 Scale, transfer, and trace diagnostics

Table 8: 30B verifier scale-up on the frozen Qwen3-4B-Instruct-2507 solver with temperature 0.2 and seeds 42–46. SNR-Gain values are mean $\pm$ sample SD across seeds; the remaining columns report summary metrics.

Dataset	SNR (mean)	SNR ($c_{\min}$)	AUC (mean)	Acc.	Abstain	Parse fail
AIME-2024	0.873 ± 0.218	1.060 ± 0.350	0.823	56.7%	37.3%	4.7%
AIME-2025	1.073 ± 0.225	1.297 ± 0.412	0.864	41.3%	42.7%	4.0%
BeyondAIME	1.223 ± 0.154	1.495 ± 0.221	0.862	27.6%	67.8%	4.4%

We test model scale and generator transfer separately. A 30B verifier reaches mean-score AUC 0.823–0.864 across the three mathematical datasets over five inference seeds, extending the tested scale within the Qwen family (Table 8); parse-failure rates are 4.0–4.7%. On a separate 687-item Gemini refusal set, the $t = 0.5$ rule raises accepted-answer accuracy from the 62.2% always-answer baseline to 78.8% at 62.5% coverage. The latter result tests use with a generator outside the verifier's model family.

Factual transfer gives a different picture. On SimpleQA, even accepting very few responses leaves low precision: the 30B mean policy reaches 22.2% at 5.2% coverage, and its minimum reference reaches 26.2% at 1.8% coverage. The 4B minimum reference has 89.4% risk at 26.0% coverage (Table 37). NQ-open, PopQA, and TriviaQA add 11,610 examples with positive SNR-Gain under inference seed 42 (Appendix G). These results distinguish score-weighted enrichment from the accepted-answer precision needed for reliable factual answering.

Inspectability and error pointing answer different questions about the annotation. In a 1,300-span audit, human type labels agree with the verifier's knowledge/inference tags in 83.0% of cases. Identifying the core error is harder: strict pointing accuracy is 49.3% on 67 incorrect BeyondAIME traces (Appendices H and I). Thus, readable semantic types and precise error localization should be assessed separately.

Figure 2 illustrates what a reader can inspect in the output. In Case 2, low confidence marks a restrictive assumption despite locally correct algebra; in Case 3, confidence recovers on later sub-calculations after an early warning. These selected examples show how scores vary within a response. Appendix J discusses the cases in detail.

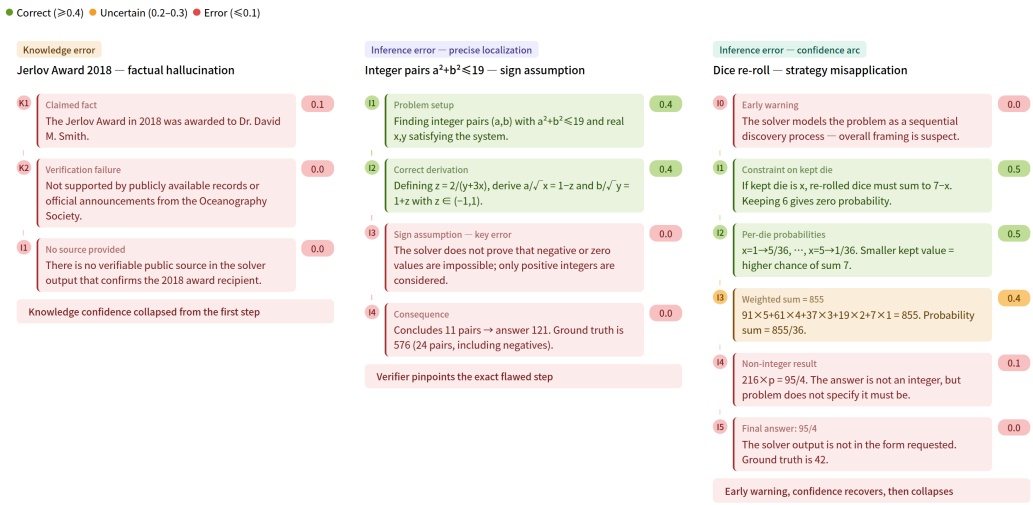

Figure 2: **Selected DISCERN outputs.** The panel shows: (Case 1) low confidence near a factual error; (Case 2) low confidence near an unproven restrictive assumption; and (Case 3) an early confidence decrease followed by recovery on later sub-calculations.

5 Conclusion

DISCERN makes the verifier part of training-data generation: its scores guide a temporary solver, while final-answer correctness supervises learning from the regenerated responses. On BeyondAIME, this curriculum improves mean-score ranking over both round 1 and frozen-solver static refresh, with the clearest matched-coverage risk reductions among the highest-scoring responses. The full reward favors mean-score ranking, and ProcessBench extends error-rejection evaluation to external mathematical traces. Together, these findings support improving a verifier through the responses it helps generate, with practical gains in low-coverage mathematical answer selection.

Scope and future work. The evaluation covers outcome-supervised response selection and semantic inspection. Study limitations and future comparisons are consolidated in Appendix O.

AI use statement

Generative AI assisted language editing, restructuring, utility code, and the BeyondAIME error-block review. A human inspected all audited traces and reconciled disagreements. The authors are responsible for verifying reported results and analysis artifacts.

Reproducibility statement

The appendices describe reward definitions, training and evaluation settings, control differences, and five-seed results (Appendices K, M, and A).

REFERENCES

Hao An and Yang Xu. Teaching llms to abstain via fine-grained semantic confidence reward. *arXiv preprint arXiv:2510.24020*, 2025.

Jaroslaw Blasiok and Preetum Nakkiran. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=XwiA1nDahv.

ByteDance Seed et al. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. *arXiv preprint*, abs/2504.13914, 2025. URL https://arxiv.org/abs/2504.13914.

Lang Cao. Learn to refuse: Making large language models more controllable and reliable through knowledge scope limitation and refusal mechanism. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3628–3646, 2024.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*, 2024.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5186–5200, 2024.

Jeremy Cole, Michael Zhang, Dan Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 530–543, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. Spuq: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2336–2346, 2024.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*, 2023.

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.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

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.

Wen Luo, Guangyue Peng, Wei Li, Shaohang Wei, Feifan Song, Liang Wang, Nan Yang, Xingxing Zhang, Jing Jin, Furu Wei, et al. Two pathways to truthfulness: On the intrinsic encoding of llm hallucinations. *arXiv preprint arXiv:2601.07422*, 2026.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Cheng Qian, Emre Can Acikgoz, Hongru Wang, Xiusi Chen, Avirup Sil, Dilek Hakkani-Tur, Gokhan Tur, and Heng Ji. Smart: Self-aware agent for tool overuse mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4604–4621, 2025.

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

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1030. URL https://aclanthology.org/2025.findings-acl.1030/.

Benedict Aaron Tjandra, Muhammed Razzak, Jannik Kossen, Kunal Handa, and Yarin Gal. Fine-tuning large language models to appropriately abstain with semantic entropy. *arXiv preprint arXiv:2410.17234*, 2024.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Sean Chen, Mohammad Kachuee, Teja Gollapudi, Tony Liao, Nicolas Scheffer, et al. Truthrl: Incentivizing truthful llms via reinforcement learning. *arXiv preprint arXiv:2509.25760*, 2025.

Jiayun Wu, Jiashuo Liu, Zhiyuan Zeng, Tianyang Zhan, Tianle Cai, and Wenhao Huang. Mitigating llm hallucination via behaviorally calibrated reinforcement learning. *arXiv preprint arXiv:2512.19920*, 2025.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21381–21396, 2025.

Boyang Xue, Fei Mi, Qi Zhu, Hongru Wang, Rui Wang, Sheng Wang, Erxin Yu, Xuming Hu, and Kam-Fai Wong. Ualign: Leveraging uncertainty estimations for factuality alignment on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6002–6024, 2025.

Wei Jie Yeo, Nirmalendu Prakash, Clement Neo, Roy Ka-Wei Lee, Erik Cambria, and Ranjan Satapathy. Understanding refusal in language models with sparse autoencoders. *arXiv preprint arXiv:2505.23556*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say "i don't know". In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7113–7139, 2024.

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. *arXiv preprint arXiv:2501.07301*, 2025.

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. *arXiv preprint arXiv:2412.06559*, 2024.

A FIVE-SEED EVALUATION AND UNCERTAINTY

We complement the greedy BeyondAIME comparison in Table 1 with five-seed evaluation. We ran the frozen base solver with inference seeds 42–46 (temperature 0.2) and passed the same generated responses to the final DISCERN and single-reward verifiers; the same BeyondAIME responses were also evaluated by the round-1 DISCERN verifier. Solver accuracy is therefore identical within each matched comparison. In this five-seed reference protocol, response confidence is $c_{\min} = \min_k c_k$ for SNR-Gain and abstention, while Conf AUC, smECE, and Brier use mean chunk confidence. SNR-Gain, Conf AUC, smECE, and Brier are computed on parse-valid confidence outputs; for risk–coverage, parser failures are retained and placed below valid confidence scores. Standard deviations summarize variation across inference seeds; the resampling procedures below estimate uncertainty for fixed trained checkpoints. A separate format-gate ablation is reported in Appendix E.3.

Table 9: Five-seed mean ± sample SD, complementary to the greedy main table. SNR-Gain uses $c_{\min}$, whereas Conf AUC, smECE, and Brier use mean chunk confidence; all confidence metrics use parse-valid outputs. The final column counts parse failures across the five seeds. The BeyondAIME round-1 row is the matched 240-step full-reward control. Pooled-mathematics Conf AUC is reported separately in Table 14.

Dataset	Method	SNR-Gain ($c_{\min}$)	Conf AUC (mean)	smECE (mean)	Brier (mean)	Parse fail ↓
BeyondAIME	DISCERN R2 (final)	1.851±0.294	0.855±0.040	0.114±0.015	0.147±0.020	0/500 (0%)
	DISCERN R1 (full reward)	1.148±0.147	–	–	–	1/500 (0.2%)
	Single-reward	0.862±0.094	0.772±0.030	0.293±0.027	0.306±0.032	2/500 (0.4%)
AIME-2024	DISCERN R2 (final)	1.352±0.478	0.844±0.058	0.173±0.036	0.179±0.044	0/150 (0%)
	Single-reward	0.735±0.514	0.715±0.107	0.257±0.072	0.261±0.079	10/150 (6.7%)
AIME-2025	DISCERN R2 (final)	1.872±0.364	0.917±0.039	0.164±0.056	0.143±0.064	0/150 (0%)
	Single-reward	1.204±0.457	0.837±0.065	0.223±0.076	0.202±0.079	7/150 (4.7%)
SimpleQA	DISCERN R2 (final)	0.725±0.039	0.687±0.008	0.290±0.004	0.317±0.003	1/21,630 (0.005%)
	Single-reward	0.218±0.023	0.582±0.006	0.385±0.004	0.438±0.003	2/21,630 (0.009%)

For the within-round BeyondAIME reward comparison, we performed 10,000 paired hierarchical bootstrap resamples: seeds are resampled first and aligned items are then resampled within each selected seed. The round-1-full-minus-single-reward SNR-Gain difference is $+0.286$ with a 95% interval of $[+0.120, +0.479]$. The corresponding per-seed item-bootstrap intervals exclude zero for seeds 42, 43, and 46, and cross zero for seeds 44 and 45; the hierarchical interval summarizes the five-seed experiment. For the five-seed appendix comparison of final DISCERN against the single-reward verifier, the SNR-Gain differences are $+0.836$, $+0.260$, $+0.411$, and $+0.507$ on BeyondAIME, AIME-2024, AIME-2025, and SimpleQA, respectively. All five seed intervals exclude zero on BeyondAIME and SimpleQA; the AIME intervals are wider because each set contains only 30 problems. A problem-by-seed hierarchical bootstrap gives a 95% interval for AIME-2024 SNR-Gain that crosses zero, reflecting greater uncertainty on this set.

Table 10: Reconciliation of BeyondAIME round-2 values. The greedy 1.456 is Table 1 (and the reprint in Table 39). The five-seed mean 1.457 ± 0.156 uses temperature 0.2 in the frozen-base evaluation setting for stochastic decoding. Table 12 reports the minimum aggregation.

Value	Verifier	Solver outputs	Aggregation	Seeds	Role
1.456	round-2 DISCERN	frozen base, greedy	mean	1	Table 1 (reprinted in Table 39)
1.746	round-2 DISCERN	frozen base, greedy	$c_{\min}$	1	Table 1
1.457±0.156	round-2 DISCERN	frozen base, T=0.2	mean	5	five-seed mean; not Table 12
1.851±0.294	round-2 DISCERN	frozen base, T=0.2	$c_{\min}$	5	Tables 9, 12
1.148±0.147	round-1 DISCERN	frozen base, T=0.2	$c_{\min}$	5	Tables 9, 12; not the ablation 1.235 ± 0.045
1.483±0.227	round-2 DISCERN	updated solver, T=0.2	$c_{\min}$	5	Table 12

The main BeyondAIME numbers are greedy decoding of frozen base-solver outputs, with mean aggregation as the default abstention score (SNR-Gain 1.456, Conf AUC 0.913) and $c_{\min}$ as a conservative reference (SNR-Gain 1.746). The five-seed mean in the frozen-base evaluation setting under temperature 0.2 is 1.457 ± 0.156 (Table 10), distinct from the greedy point estimate 1.456. Five-seed uncertainty for the reference $c_{\min}$ score is 1.851 ± 0.294 (Table 9 and Table 12). The updated-solver row reports results on the updated solver's response population.

Mean and minimum aggregation give different profiles on the same BeyondAIME responses (SNR-Gain/Conf AUC $1.456/0.913$ versus $1.746/0.791$). Mean aggregation provides the default response-

Table 11: Five-seed minimum-reference selective-prediction results (mean ± sample SD; lower AURC, nAURC, and risk are better). Coverage and risk in the last two columns use $c_{\min} \geq 0.5$.

Dataset	Method	AURC	nAURC	Coverage	Risk
BeyondAIME	DISCERN	0.506±0.039	0.379±0.073	0.278±0.019	0.287±0.058
	Single-reward	0.576±0.038	0.579±0.053	0.518±0.011	0.513±0.047
AIME-2024	DISCERN	0.261±0.058	0.448±0.110	0.620±0.045	0.216±0.057
	Single-reward	0.286±0.095	0.524±0.220	0.633±0.053	0.248±0.120
AIME-2025	DISCERN	0.334±0.087	0.284±0.145	0.507±0.064	0.220±0.110
	Single-reward	0.368±0.089	0.380±0.152	0.573±0.080	0.283±0.116
SimpleQA	DISCERN	0.921±0.003	0.816±0.013	0.260±0.005	0.894±0.004
	Single-reward	0.940±0.002	0.941±0.007	0.419±0.003	0.937±0.002

selection score, combining stronger observed ranking with finer score resolution; minimum aggregation supplies a conservative reference. We report the ranking and score-weighted enrichment of each aggregation separately.

A.1 Selective Prediction and Risk–Coverage

For this minimum-reference analysis at coverage q, we select the $\lceil qn \rceil$ highest-scoring responses under $c_{\min}$ and define selective risk as their error rate. A boundary tie is resolved by uniform random selection independent of correctness, and the reported risk is its expectation. Deterministic thresholds accept complete tied-score groups. Parser failures remain in the evaluation and rank below valid scores. AURC averages selective risk over coverages. We additionally report normalized AURC,

$$\mathrm{nAURC} = \frac{\mathrm{AURC} - \mathrm{AURC}_{\mathrm{oracle}}}{r - \mathrm{AURC}_{\mathrm{oracle}}},$$

where r is the always-answer error rate; zero is oracle ranking and one is random ranking.

Under the conservative minimum-reference policy, the final verifier has five-seed AURC 0.506 ± 0.039, compared with 0.576 ± 0.038 for the single-reward control (Table 11). At $c_{\min} \geq 0.5$, risk decreases from 0.513 ± 0.047 to 0.287 ± 0.058, while coverage decreases from 0.518 ± 0.011 to 0.278 ± 0.019. These operating points describe the minimum policy's coverage–risk trade-off. The minimum-score distribution is characterized in Appendix A.2.

Figure 3 and Table 11 place BeyondAIME alongside AIME and SimpleQA under the same minimum-reference evaluation. On AIME-2024/2025, the final verifier's AURCs are 0.261 ± 0.058 and 0.334 ± 0.087, compared with 0.286 ± 0.095 and 0.368 ± 0.089 for the control. The figure and table report inference-seed variation and the corresponding SimpleQA operating points.

Risk–coverage curves average risk over uniform selection within tied-score groups. These interior points describe randomized selection; deterministic thresholds operate at tied-group boundaries. Parser failures rank below valid scores. Appendix A.1 gives the AURC definition and evaluation protocol.

A.2 Matched Solver–Verifier Cross-Evaluation

We evaluate both verifier checkpoints on frozen base-solver and updated-sparring-solver traces to characterize the challenge–recovery loop. The round-1/base cell is the base-generator reference; round-1/updated measures the challenge, round-2/updated measures recovery, and round-2/base checks retention on the original population. Table 12 reports five-seed BeyondAIME SNR-Gain using $c_{\min}$, following Table 9. Tables 14 and 5 report pooled-mathematics mean-score Conf AUC and false acceptance among incorrect responses at mean confidence ≥ 0.5. This separates the default mean-score diagnostics from the minimum-score reference.

Table 13 decomposes the false-acceptance rates in Table 5 into item-level transitions. The columns contain $N = 515$ frozen-base errors and $N = 500$ updated-solver errors. Conditional repair rates use the 126 and 166 errors accepted by round 1 in the respective populations. On the frozen base-solver population, the round-2 verifier fixes 55 of the 126 responses that round-1 accepted ($b = 55$,

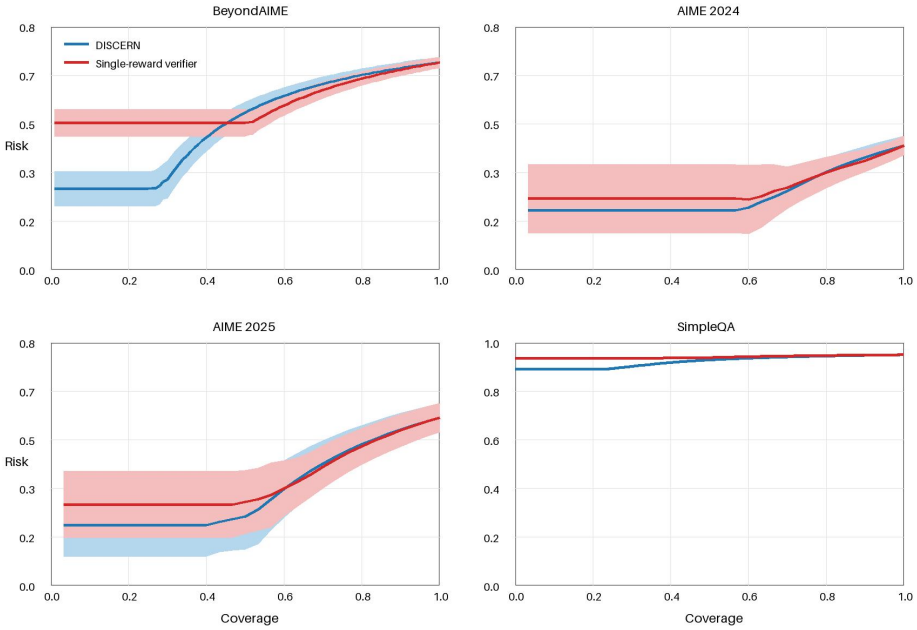

Figure 3: Conservative minimum-reference risk–coverage curves for the final round-2 verifier and the single-reward control (five-seed mean ± SD). Tied-score blocks use expected risk under uniform tie-breaking; the tie convention is defined in the evaluation protocol.

$55/515 = 10.68\%$ of incorrect responses) and newly misses 11 ($c = 11$, $11/515 = 2.14\%$). On the updated-solver population, it fixes 129 of the 166 responses that round-1 accepted ($b = 129$, $129/500 = 25.80\%$) and newly misses 10 ($c = 10$, $10/500 = 2.00\%$). The updated-solver population thus exhibits a much larger repair mass b/N (25.80% vs. 10.68%) with essentially unchanged new misses (2.00% vs. 2.14%). For the primary between-population comparison, we condition on round-1 acceptance: the repair rate is $55/126 = 43.7\%$ for the base population and $129/166 = 77.7\%$ for the updated population (observed difference $+34.1$ pp). To account for repeated problems across inference seeds while preserving the two solver populations, we resample the 139 unique problems with replacement, keeping all five seeds of each sampled problem, jointly for both solver populations (10,000 replicates). The resulting 95% percentile interval for the updated-minus-base conditional-repair difference is $[+23.1, +44.9]$ pp and excludes zero; none of the replicates had a non-positive difference. A seed-clustered bootstrap remains the uncertainty method for the unconditional masses: the 95% interval for the between-population difference in b/N is $[+11.5, +18.3]$ pp excluding zero, while the corresponding interval for c/N is $[-1.1, +0.8]$ pp and includes zero. Under this mean-score operating-point protocol, the observed repair fraction is larger on updated-solver errors and the new-miss fractions are similar. The continuation controls separately evaluate the benefit of updating the data-generating solver.

Table 12: Five-seed BeyondAIME SNR-Gain in the asymmetric solver–verifier cross-evaluation (mean ± sample SD), under $c_{\min}$ aggregation. The round-1/base cell is the in-distribution reference; round-1/updated is the challenge diagnostic; round-2/updated is the recovery cell; and round-2/base is the regression check. The protocol reconciliation is in Table 10.

Verifier	Frozen base solver	Updated sparring solver
Round-1 DISCERN verifier	1.148±0.147	1.000±0.160
Round-2 DISCERN verifier	1.851±0.294	1.483±0.227

Figure 4 describes the observed score quantization. Round 1 emits only $\{0.0, 0.1, 1.0\}$ over 4,366 valid chunks. Every response with a 0.1-valued chunk also has a zero-valued chunk, making $c_{\min}$ binary for all 767 valid round-1 responses. In this pooled updated-solver diagnostic, round 2 emits

Table 13: Item-level transitions between the round-1 and round-2 verifiers among incorrect responses at mean confidence ≥ 0.5. Cell b counts responses accepted by round-1 and rejected by round-2 (fixed); cell c counts responses rejected by round-1 and accepted by round-2 (newly missed). Absolute counts and both b/N and the conditional repair rate $b/(a+b)$ are reported. For the frozen-base and updated-solver error sets ($N = 515$ and $N = 500$), the conditional-repair denominators are 126 and 166 round-1 false acceptances, respectively. The net change $(b - c)/N$ equals the pooled FA-rate difference in Table 5: $+8.54$ pp for the frozen base solver and $+23.80$ pp for the updated solver. Unclustered McNemar tests describe the item-level transitions ($p = 3.6 \times 10^{-8}$ and $p = 1.7 \times 10^{-27}$, respectively); primary inference uses the problem-clustered between-population analysis in the surrounding text.

	Frozen base solver ($N = 515$)		Updated solver ($N = 500$)	
	R2 accept	R2 reject	R2 accept	R2 reject
R1 accept	71	**55**	37	**129**
R1 reject	11	378	10	324
R1-accept base $a+b$	126		166	
conditional repair $b/(a+b)$	55/126 = 43.7%		129/166 = 77.7%	
b/N	55/515 = 10.68%		129/500 = 25.80%	
c/N	11/515 = 2.14%		10/500 = 2.00%	

Table 14: Five-seed pooled mathematics Conf AUC (BeyondAIME+AIME-2024+AIME-2025) in the asymmetric solver–verifier cross-evaluation (mean aggregation). The round-1/updated cell is a challenge diagnostic, round-2/updated is the recovery cell, and round-2/base is a regression check. Dataset-specific results appear in Table 9.

Verifier	Frozen base solver	Updated sparring solver
Round-1 DISCERN verifier	0.823±0.030	0.791±0.020
Round-2 DISCERN verifier	0.878±0.037	0.849±0.020

10 values over 13,563 chunks, but its minimum remains binary for 97.9% of responses. These statistics characterize the pooled updated-solver diagnostic. The mean aggregates the more diverse chunk scores, whereas the minimum concentrates response scores at the endpoints.

The false-acceptance table describes challenge and recovery at mean confidence ≥ 0.5. Round-1 false acceptance is 24.47% on base-solver errors and 33.20% on updated-solver errors, an 8.73-point pooled difference. Round 2 reduces these rates to 15.92% and 9.40%, respectively. The remaining 6.52-point gap is *lower* false acceptance on updated-solver errors. These rates describe the two solver error populations at the specified mean-score threshold.

The SNR-Gain table shows the same qualitative pattern under $c_{\min}$. Replacing frozen-base-solver responses with updated-sparring-solver responses lowers the round-1 verifier from 1.148 to 1.000. Replacing the round-1 verifier with the round-2 verifier on that same updated-solver population then raises SNR-Gain from 1.000 to 1.483. The round-2 verifier also improves on frozen base-solver responses from 1.148 to 1.851, providing no indication of a trade-off against the original response population in this metric.

These observations characterize the solver's role as a curriculum generator. False acceptance characterizes threshold-specific rejection, Conf AUC characterizes ranking, and SNR-Gain also reflects score scale. The separate continuation comparison in Table 4 includes newly generated training responses in both branches and supports updating the solver relative to the tested frozen-solver refresh recipe.

B Adaptation and Modification of Reference Experimental Results

We adapt CISC to confidence-based rejection by using response probability as the confidence signal. The original CISC procedure uses P(True)-based weighted voting; our implementation evaluates its response-probability adaptation for selective abstention.

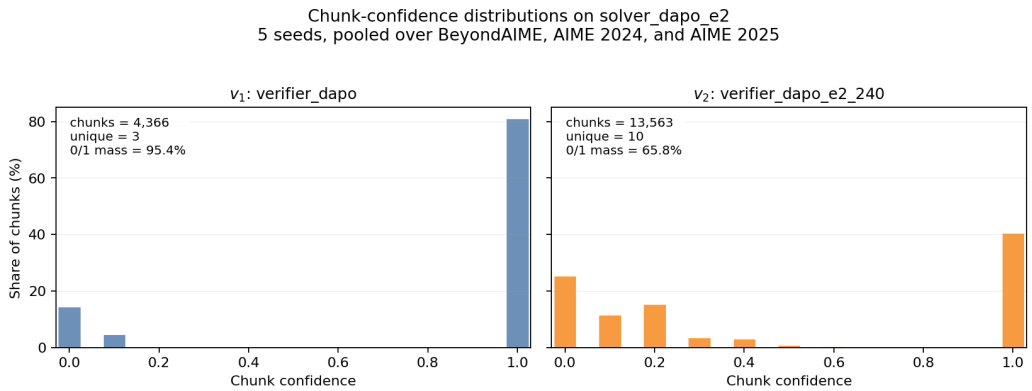

Figure 4: Chunk-level confidence distributions on the same updated-solver evaluation setting, pooled over BeyondAIME, AIME-2024, and AIME-2025 across five seeds. The valid parsed outputs contain 4,366 round-1 chunks and 13,563 round-2 chunks. Round 1 emits three distinct values with 95.4% of mass at 0 or 1; round 2 emits ten values and reduces endpoint mass to 65.8%.

C Definitions and Interpretation of Evaluation Metrics

C.1 Definition of SNR-Gain

Let a model, for a given prompt x, decide either to **Answer** ($a = \mathrm{ANS}$) with response y or to **Abstain** ($a = \mathrm{ABS}$). Define $\mathrm{valid}(y) \in \{0, 1\}$ as the binary correctness of y. For a score threshold $t \in [0, 1]$, the model's policy induces a joint distribution over answers and correctness. We define

$$\mathrm{Acc}(t) = \mathbb{E}\big[\mathrm{valid}(y) \wedge a(t) = \mathrm{ANS}\big]$$

$$\mathrm{Hal}(t) = \mathbb{E}\big[\neg\mathrm{valid}(y) \wedge a(t) = \mathrm{ANS}\big]$$

$$\mathrm{SNR}(t) = \frac{\mathrm{Acc}(t)}{\mathrm{Hal}(t)}$$

When t varies, the average behaviour over the whole spectrum $t \in [0, 1]$ is captured by the *integrated* SNR

$$\mathrm{SNR}([0, 1]) = \frac{\int_0^1 \mathrm{Acc}(t)\,\mathrm{d}t}{\int_0^1 \mathrm{Hal}(t)\,\mathrm{d}t}.$$

The baseline accuracy–hallucination trade-off is evaluated at $t = 0$ (model answers everything). The **SNR-Gain** quantifies the relative improvement brought by the model's ability to abstain according to its score:

$$\text{SNR-Gain} = \log\left(\frac{\mathrm{SNR}([0, 1])}{\mathrm{SNR}(0)}\right)$$

A positive SNR-Gain indicates that the abstention policy raises the overall "signal-to-noise" ratio of the generated answers; the larger the gain, the greater this integrated ratio.

Evaluation scope of SNR-Gain. SNR-Gain is the answer–abstain metric used by Wu et al. (2025). It integrates over the numerical score scale using the reference protocol, complementing ranking metrics and fixed-threshold operating points. Fixed-threshold SNR is infinite if accepted-correct mass is positive and accepted-error mass is zero, and undefined if both are zero; integrated SNR uses the accepted masses accumulated over the full threshold range.

Dependence on score scale. Let $Y \in \{0, 1\}$ denote correctness and $s \in [0, 1]$ a response score, with acceptance when $s \geq t$. Under uniform continuous threshold integration,

$$\int_0^1 \mathrm{Acc}(t)\,\mathrm{d}t = \mathbb{E}[Ys], \qquad \int_0^1 \mathrm{Hal}(t)\,\mathrm{d}t = \mathbb{E}[(1 - Y)s].$$

For a nondegenerate correct/incorrect population and positive conditional score means, normalization by the always-answer odds gives

$$\text{SNR-Gain} = \log \frac{\mathbb{E}[s \mid Y = 1]}{\mathbb{E}[s \mid Y = 0]}.$$

The calculation integrates accepted-correct and accepted-incorrect masses separately before forming the normalized log ratio. A strictly increasing nonlinear transformation $g(s)$ preserves rankings and threshold-selected subsets but can change the conditional-mean ratio. For example, if every correct response scores 0.8 and every incorrect response scores 0.4, SNR-Gain is $\log 2$; replacing s by s^2 gives $\log 4$ with the same ordering and attainable risk–coverage trade-off. Numerical quadrature approximates the integral and can introduce additional discretization effects. AUC and tie-aware risk–coverage quantify ranking, while SNR-Gain quantifies score-weighted enrichment.

C.2 Definition of Actuality Alignment Benchmarks

This section defines **Precision**, **Truthfulness**, **Specificity**, and **Unknown Recall** under the actuality-alignment protocol. The protocol evaluates answering and refusal using operational known/unknown question categories and the model's answering behavior. Its **Precision** uses all known questions as the denominator; accepted-answer precision in the QA and ProcessBench answer-level tables uses all accepted answers, $KC/(KC + UI)$. Table 15 defines the categories used in this subsection.

Table 15: Notations for different answer types.

Notation	Description
KC	Known & answered correctly
KI	Known & answered incorrectly
KR	Known & refused to answer
UC	Unknown & answered correctly
UI	Unknown & answered incorrectly
UR	Unknown & refused to answer

Precision is the fraction of known questions answered correctly:

$$\text{Precision} = \frac{\text{KC}}{\text{KI} + \text{KC} + \text{KR}}. \tag{1}$$

Truthfulness counts correct answers on known questions and refusals on unknown questions, normalized by all evaluated questions:

$$\text{Truthfulness} = \frac{\text{UR} + \text{KC}}{\text{KC} + \text{KI} + \text{KR} + \text{UC} + \text{UI} + \text{UR}}. \tag{2}$$

In Equation (2), the numerator counts correct answers on known questions and refusals on unknown questions under the benchmark's categorization. The denominator counts all evaluated samples. Always refusing can obtain high Truthfulness when unknown questions dominate. We therefore include an always-abstain reference and report answering behavior alongside these metrics.

Specificity is the refusal rate over all operationally unknown questions:

$$\text{Specificity} = \frac{\text{UR}}{\text{UR} + \text{UI} + \text{UC}}. \tag{3}$$

A high Specificity indicates frequent refusal on the operationally unknown subset; always abstaining maximizes it.

Unknown Recall is the refusal fraction among unknown questions that receive a refusal or an incorrect answer:

$$\text{Unknown Recall} = \frac{\text{UR}}{\text{UR} + \text{UI}}. \tag{4}$$

A high Unknown Recall indicates more refusals relative to incorrect answers on the operationally unknown subset. Specificity additionally includes correct answers on that subset (UC) in its denominator, whereas Unknown Recall excludes them. We therefore report them alongside Truthfulness and Precision.

D Separate MATH-Trained AIME Stress Test

We further trained the verifier exclusively on the simpler MATH dataset (which contains problems of generally lower difficulty than AIME benchmarks) using a fixed, relatively high sampling temperature of $T = 0.8$. The high temperature is used to increase generation diversity and provides a separate stress-test setting. The greedy BeyondAIME evaluation appears in Table 1.

We adopt the **SNR-Gain** metric proposed in (Wu et al., 2025), which measures the log-scale improvement in the signal-to-noise ratio (ratio of correctly accepted answers to incorrectly accepted answers) across the risk-threshold spectrum $t \in [0, 1]$, relative to the always-answer baseline ($t = 0$). A formal definition is provided in Appendix C.1.

Table 16: Verifier rejection performance on AIME 2025 ($T = 0.8$, verifier trained on MATH). No correct answer was rejected.

Metric	**AIME 2025**
Problems	30
Rejected	14
Correct rejections	14
Incorrectly accepted	5
SNR-Gain	0.847

Under this protocol, the verifier rejects 14 of 19 incorrect AIME 2025 answers and rejects no correct answer, yielding an SNR-Gain of 0.847. This 30-problem stress test is reported separately from the greedy BeyondAIME evaluation.

Note on AIME 2024. At the reported operating point in this separate stress test, the verifier rejected all 9 incorrect AIME 2024 responses and rejected no correct response. The operating-point SNR is therefore infinite because accepted-correct mass is positive and accepted-error mass is zero; SNR-Gain instead uses accepted masses integrated over the threshold range. The main BeyondAIME results in Table 1 use the greedy frozen-solver protocol and report mean and $c_{\min}$ aggregation separately.

E In-depth Analysis of Reward Components

To examine how reward components affect training dynamics, we compare the full DISCERN reward with runs excluding individual components. Figure 5 reports Accuracy (mean@1) on the validation set across 220 training steps. These single-run curves complement the endpoint ablation in Table 39.

Diagnostic Analysis of Reward Components. The training curves show distinct optimization behavior under component removal:

- $R_{\textbf{topo}}$**:** The variant without R_{topo} (Blue) eventually reaches high accuracy but shows more stagnation and fluctuation between steps 100 and 140, whereas the full reward (Red) follows a steadier trajectory.
- $R_{\textbf{detect}}$**:** Excluding R_{detect} (Green) produces the most erratic curve, including a drop near step 180.

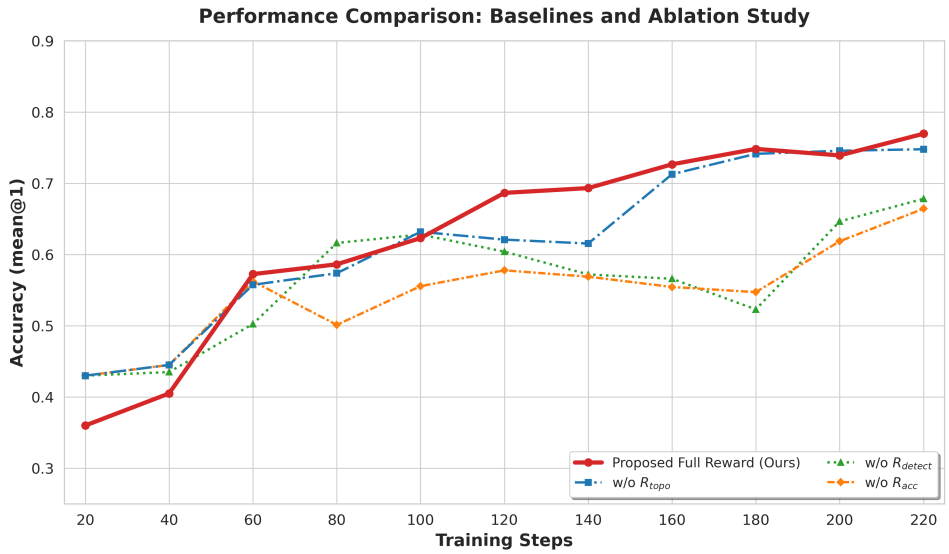

Figure 5: **Reward-Ablation Learning Curves.** Verifier validation Accuracy (mean@1) during training. The full reward has higher observed accuracy than variants lacking R_{topo}, R_{detect}, or R_{acc}.

- R_{acc}**:** The version without R_{acc} (Orange) consistently lags behind the others, supporting the practical value of the difficulty-aware reliability term within this training recipe.

In summary, Figure 5 motivates retaining all three reward components, while the one-round endpoint ablation in Table 39 provides the SNR-Gain comparison. The ranking diagnostics below report the effects on mean and minimum aggregation separately.

E.1 Ranking diagnostics for reward and continuation controls

Do changes in score-weighted enrichment also improve response ordering? Tables 17 and 18 answer this question with Conf AUC for both mean and minimum aggregation. The summaries use inference seeds 42–46. Each checkpoint receives the same supplied traces within a comparison; Conf AUC is computed on that checkpoint's parse-valid outputs.

Table 17: One-round BeyondAIME reward controls: Conf AUC, mean ± sample SD over five inference seeds. All four checkpoints receive the same 100 supplied traces (33 correct answers), corresponding to the reward-removal evaluation in Table 39.

Reward	Mean-score AUC	Minimum-score AUC
Full reward	0.781 ± 0.003	0.724 ± 0.004
Without R_{topo}	0.760 ± 0.012	0.753 ± 0.006
Without R_{accuracy}	0.768 ± 0.019	0.735 ± 0.011
Without R_{detect}	0.729 ± 0.000	0.729 ± 0.000

The full reward has the highest mean-score AUC point estimate among these four checkpoints. Under minimum aggregation, the component-removal variants have higher AUC point estimates. The matched-coverage analysis below reports mean-score AURC and paired risk intervals for the same reward comparison.

Sparring has higher AUC point estimates than both comparison checkpoints on all three datasets under both aggregations. This table uses the additional static-reference checkpoint; the primary static-refresh comparison is in Appendix E.4. BeyondAIME mean-score paired intervals for this reference comparison appear in Table 25.

Table 18: Additional continuation-reference diagnostics on frozen-base solver responses: Conf AUC, mean ± sample SD over five inference seeds. Checkpoints receive the same supplied evaluation responses within each dataset. This table uses the static-reference checkpoint from Table 25; the primary static-refresh checkpoint is evaluated in Table 4. Both static refresh and sparring use newly generated training responses, from unchanged and updated solver weights, respectively.

Dataset	Checkpoint	Mean-score AUC	Minimum-score AUC
AIME-2024	Round 1	0.801 ± 0.060	0.772 ± 0.057
	Frozen-solver fresh-data	0.691 ± 0.029	0.645 ± 0.063
	Sparring	0.830 ± 0.050	0.792 ± 0.039
AIME-2025	Round 1	0.861 ± 0.056	0.847 ± 0.060
	Frozen-solver fresh-data	0.723 ± 0.064	0.687 ± 0.049
	Sparring	0.897 ± 0.056	0.876 ± 0.063
BeyondAIME	Round 1	0.802 ± 0.029	0.770 ± 0.018
	Frozen-solver fresh-data	0.675 ± 0.034	0.616 ± 0.028
	Sparring	0.869 ± 0.046	0.814 ± 0.040

E.2 Mean-score expected risk at matched coverage

We supplement the ranking controls with mean-score AURC and expected selective risk, $\mathbb{E}[R]$, on the same supplied responses within each comparison. We report the continuation, one-round reward-removal, and final-system evaluation cohorts separately. Negative paired risk differences favor sparring, full reward, or the final system, respectively. Risk within tied-score groups is averaged over random selection, as defined in the selective-prediction protocol. Bracketed ranges are pointwise paired intervals. Fresh-data entries in this subsection use the additional static-reference checkpoint from Table 25.

Table 19: Additional mean-score selective-prediction summaries on BeyondAIME. Lower AURC and expected risk are better. The two panels use their respective evaluation cohorts. Dashes mark unreported metrics.

Comparison	Checkpoint	AURC	$\mathbb{E}[R]$ at 50%
Continuation, frozen-base	Round 1	0.556	–
	Static reference	0.646	–
	Sparring	0.483	–
One-round reward removal	Full reward	0.542	0.477
	Without topology	0.565	0.483
	Without accuracy	0.557	0.478
	Without detection	0.578	0.504

Continuation relative to round 1. Table 20 reports the full coverage grid. On BeyondAIME, paired intervals exclude zero at 20% and 30% coverage; the intervals include zero at 50%, 70%, and 90%. The strongest risk reduction relative to round 1 occurs at low coverage. The AIME point estimates are mostly smaller; the supplied 20% intervals include zero for both AIME sets.

Table 20: Mean-score expected-risk differences, sparring minus round 1, on frozen-base responses. Negative values favor sparring. The grid includes all five nominal coverage points. Paired intervals for BeyondAIME at 20%, 30%, and 50% appear in Table 21.

Dataset	20%	30%	50%	70%	90%
BeyondAIME	-0.165	-0.135	-0.032	$+0.001$	-0.002
AIME-2024	-0.005	-0.005	-0.005	-0.044	-0.004
AIME-2025	-0.046	-0.046	-0.028	-0.007	0.000

Continuation relative to fresh-data refresh. Against the additional static-reference checkpoint, the BeyondAIME paired intervals favor sparring at every reported coverage point. The 50% difference is -0.116 with interval $[-0.167, -0.059]$. The round-1 comparison independently shows a low-coverage benefit; primary static-refresh ranking and operating-point diagnostics appear in Appendix E.4.

Table 21: Reported paired expected-risk differences on BeyondAIME using mean scores. Each row compares methods within its own evaluation cohort. Negative differences favor the first method. Intervals are pointwise; no simultaneous-coverage or independent-training-run claim is made.

Contrast	Coverage	$\Delta\mathbb{E}[R]$	Paired interval
Sparring − Round 1	20%	-0.165	$[-0.236, -0.095]$
	30%	-0.135	$[-0.183, -0.032]$
	50%	-0.032	$[-0.076, +0.014]$
Sparring − Static reference	50%	-0.116	$[-0.167, -0.059]$
Full reward − Without detection	20%	-0.064	$[-0.159, +0.016]$
Final system − Single-reward	20%	-0.231	$[-0.316, -0.145]$

Reward removal and final-system comparisons. Full reward has the lowest mean-score AURC point estimate among the one-round variants (Table 19). At 50% coverage, expected risk is 0.477, versus 0.483, 0.478, and 0.504 without topology, accuracy, and detection. Paired fixed-coverage risk intervals mostly include zero; for example, full minus no-detection at 20% is -0.064 with interval $[-0.159, +0.016]$. In the separate final-system comparison, expected risk at 20% coverage is 0.282 for final DISCERN and 0.513 for the single-reward control; the paired difference is -0.231 with interval $[-0.316, -0.145]$. Differences diminish at higher coverage, with intervals often including zero from 50% onward. This final-system comparison shows lower expected risk at low coverage.

E.3 Format-gate ablation

A boxed-answer check can reject a detected answer change or suppress learning whenever extraction fails. We compare these choices in round-1 training. The original checkpoint uses a *missing-box-tolerant* gate: the boxed-answer check sets R_{format} to zero only when answers are extracted from both the supplied response and the verifier output and disagree. The *strict* variant additionally sets the gate to zero whenever either boxed answer is missing. Both use the same detection, accuracy, and topology reward logic with weights $0.4/0.4/0.2$ (Appendix K.3).

Both checkpoints score the same frozen-base solver responses on AIME-2024, AIME-2025, and BeyondAIME, with inference seeds 42–46, temperature 0.2, top-$p = 0.9$, and maximum output length 16,384. The operating-point threshold is 0.5. Table 6 in the main text reports ranking and enrichment; Table 22 below reports operating-point diagnostics.

The missing-box-tolerant gate has higher mean-score AUC by 0.068, 0.099, and 0.094 on AIME-2024, AIME-2025, and BeyondAIME, respectively, and lower AURC on all three. It also yields lower false-acceptance rates at the reported threshold (Table 22). The strict variant retains positive mean-score SNR-Gain and above-chance AUC on every reported inference seed. Its minimum-score SNR-Gain is higher on AIME-2024 but lower on the other two sets, distinguishing score-weighted enrichment from the consistent mean-ranking advantage of the tolerant gate.

Unlike the strict gate, the tolerant rule keeps otherwise valid annotations eligible for content reward when a boxed answer cannot be extracted. The observed ranking advantage supports this less restrictive gating choice for learning useful verification scores.

E.4 Frozen-solver static-refresh evaluation

Static refresh continues verifier 1 using newly generated responses from the unchanged original solver. We distinguish the primary static-refresh checkpoint in Table 4 from the additional static-reference checkpoint in Table 25. The latter supplies the additional ranking and expected-risk diagnostics; results and intervals retain their checkpoint-specific correspondence.

Table 22: Operating-point diagnostics for the tolerant and strict format-gate configurations at the reported threshold $t = 0.5$. FA is the fraction of incorrect responses accepted (mean $\pm$ SD across inference seeds); coverage is the average of the rounded per-seed coverage values. Parser failures are pooled counts over five seeds, with 150 responses per AIME set and 500 for BeyondAIME.

Dataset	Gate	FA	Coverage	Parse failures
AIME-2024	Tolerant	0.289 ± 0.086	60.7%	2/150
	Strict	0.427 ± 0.073	59.3%	5/150
AIME-2025	Tolerant	0.233 ± 0.122	52.7%	4/150
	Strict	0.425 ± 0.087	58.0%	4/150
BeyondAIME	Tolerant	0.237 ± 0.034	38.8%	1/500
	Strict	0.389 ± 0.066	47.6%	9/500

Evaluation protocol. Both static checkpoints receive the same frozen-base and updated-sparring solver responses. Evaluation uses inference seeds 42–46, verifier temperature 0.2, top-p 0.9, a 16,384-token limit, and threshold 0.5. Table 23 reports both SNR-Gain aggregations, Conf AUC, and AURC; Table 24 records threshold diagnostics and parser failures. Under the continuation evaluation protocol, confidence metrics use each checkpoint's parse-valid outputs.

Table 23: Static-refresh evaluation: reported mean $\pm$ SD over five inference seeds on each solver population. SNR-Gain uses the indicated aggregation; Conf AUC uses mean confidence. AURC follows this continuation protocol; the tie-averaged expected-risk protocol is specified in Appendix E.2.

Dataset	SNR-Gain (mean)	SNR-Gain ($c_{\min}$)	Conf AUC	AURC
Frozen-base evaluation responses				
AIME-2024	0.743 ± 0.248	0.876 ± 0.343	0.738 ± 0.067	0.251 ± 0.078
AIME-2025	1.085 ± 0.228	1.528 ± 0.596	0.824 ± 0.029	0.323 ± 0.039
BeyondAIME	0.780 ± 0.259	0.960 ± 0.424	0.706 ± 0.061	0.588 ± 0.042
Updated-solver evaluation responses				
AIME-2024	0.757 ± 0.306	1.162 ± 1.069	0.750 ± 0.090	0.222 ± 0.049
AIME-2025	0.863 ± 0.098	1.076 ± 0.321	0.791 ± 0.026	0.385 ± 0.064
BeyondAIME	0.790 ± 0.181	0.968 ± 0.389	0.724 ± 0.050	0.608 ± 0.033

Table 24: Static-refresh threshold diagnostics at $t = 0.5$. FA is the fraction of incorrect responses accepted. FA is the reported mean $\pm$ SD of seed-level rates; coverage averages the displayed seed-level values. Parse-failure counts sum evaluations over the five inference seeds.

Dataset	False acceptance	Coverage	Parse failures
Frozen-base evaluation responses			
AIME-2024	0.425 ± 0.122	66.0%	0/150
AIME-2025	0.353 ± 0.096	58.0%	2/150
BeyondAIME	0.307 ± 0.070	41.0%	12/500
Updated-solver evaluation responses			
AIME-2024	0.422 ± 0.114	69.3%	5/150
AIME-2025	0.392 ± 0.074	58.0%	10/150
BeyondAIME	0.329 ± 0.030	43.0%	18/500

Comparison and scope. On frozen-base BeyondAIME, static refresh attains mean-score SNR-Gain 0.780 and AUC 0.706. On updated-solver BeyondAIME, it reaches SNR-Gain 0.790 and AUC 0.724, compared with round 1's 0.653 and 0.716. Performance depends on the evaluation population: frozen-base AUC and mean-score SNR-Gain remain below round 1 on all three datasets. Sparring has higher BeyondAIME AUC point estimates than static refresh by 0.163 on frozen-base and 0.112 on updated-solver responses.

Table 25: Additional BeyondAIME continuation-reference comparison (five inference seeds; mean ± SD), using the additional static-reference checkpoint. This comparison matches verifier-training budgets and evaluates confidence metrics on parse-valid outputs, with paired problem-bootstrap intervals for the listed checkpoints. Table 4 reports the primary static-refresh comparison.

Verifier	Base evaluation responses	Updated-solver responses
SNR-Gain (mean confidence)		
Round 1	1.065 ± 0.108	0.653 ± 0.172
Static reference	0.310 ± 0.075	0.213 ± 0.055
Sparring continuation	$\mathbf{1.324 \pm 0.173}$	$\mathbf{1.119 \pm 0.121}$
Sparring − Fresh-data	$+1.014$	$+0.906$
95% problem-bootstrap CI	$[+0.781, +1.291]$	$[+0.743, +1.092]$
Conf AUC		
Round 1	0.802 ± 0.029	0.716 ± 0.064
Static reference	0.675 ± 0.034	0.625 ± 0.043
Sparring continuation	$\mathbf{0.869 \pm 0.046}$	$\mathbf{0.836 \pm 0.026}$
95% CI, sparring − fresh-data	$[+0.139, +0.255]$	$[+0.148, +0.272]$

Table 25 and the static-reference expected-risk entries in Appendix E.2 report the additional reference checkpoint; Table 23 reports the primary static-refresh checkpoint.

During training, validation reward declined in both the primary static-refresh run and the additional static-reference run. We suspect overfitting contributed to the performance loss. This observation motivates a curriculum that adapts the training-response distribution to the evolving verifier, rather than relying only on fresh samples from an unchanged solver. Sparring implements this idea by using the current verifier's feedback to update the response generator before continuing verifier training.

E.5 Reference-aware rationale-judged reward

As an additional alternative-supervision comparison, we evaluate a verifier trained with a *reference-aware rationale-judged reward*. A Qwen3-4B meta-evaluator has access to the reference answer and assesses whether the rationale given by the trainable verifier is correct, assigning a binary training reward of 1 or 0. The trainable verifier retains the chunk-scoring interface and receives this rationale judgment as its training reward. This rationale-judged verifier is distinct from the answer-label-based single-reward control (Appendix M.1).

Evaluation. Both checkpoints receive the same frozen-base solver responses on AIME-2024, AIME-2025, and BeyondAIME, with inference seeds 42–46, verifier temperature 0.2, top-p 0.9, and a 16,384-token limit. All scores here use mean chunk confidence. AUC, AURC, and top-k risk use each checkpoint's parse-valid outputs. At nominal coverage c, top-k risk selects the top $\mathrm{round}(cn)$ responses among n parse-valid outputs; we report these as parse-valid top-k diagnostics. The tie-averaged expected-risk analysis is reported separately in Appendix E.2.

Results and interpretation. The rationale-judged checkpoint produces parseable outputs in $786/800 = 98.25\%$ of evaluations, but its mean AUCs are 0.483, 0.527, and 0.526, compared with 0.801, 0.861, and 0.802 for round-1 DISCERN (Table 26). DISCERN has higher AUC and lower AURC in all 15 dataset–inference-seed comparisons. On BeyondAIME, mean top-k risk at nominal 20% coverage is 0.867 versus 0.440. These observations illustrate that parseable chunk-confidence outputs and reference-aware rationale supervision need not yield effective response-level ranking. This alternative-supervision comparison complements the single-reward and component-removal experiments.

F Off-the-shelf PRM and ProcessBench-style Diagnostics

We score Qwen2.5-Math-PRM-7B (Zhang et al., 2025) as an off-the-shelf process scorer. BeyondAIME SNR-Gain uses the same 100 traces as Table 39 (33 correct). Refusal fires if any DIS-

Table 26: Reference-aware rationale-judged reward versus round-1 DISCERN on frozen-base responses. Entries are reported mean ± SD across five inference seeds. AUC is higher-is-better; AURC and risk are lower-is-better. Parser failures are summed over the five inference seeds. Risk is measured within each checkpoint's parse-valid subset, at the indicated nominal coverage.

Dataset	Verifier	AUC	AURC	Parse failures
AIME-2024	Rationale-judged	0.483 ± 0.050	0.372 ± 0.088	0/150
	DISCERN R1	0.801 ± 0.060	0.198 ± 0.075	2/150
AIME-2025	Rationale-judged	0.527 ± 0.061	0.460 ± 0.036	3/150
	DISCERN R1	0.861 ± 0.056	0.291 ± 0.063	4/150
BeyondAIME	Rationale-judged	0.526 ± 0.025	0.762 ± 0.038	11/500
	DISCERN R1	0.802 ± 0.029	0.556 ± 0.052	1/500
Dataset	Verifier	Risk@20%	Risk@30%	Risk@50%
AIME-2024	Rationale-judged	0.467 ± 0.139	0.422 ± 0.122	0.347 ± 0.110
	DISCERN R1	0.133 ± 0.139	0.139 ± 0.127	0.204 ± 0.109
AIME-2025	Rationale-judged	0.333 ± 0.000	0.475 ± 0.079	0.519 ± 0.062
	DISCERN R1	0.200 ± 0.075	0.183 ± 0.067	0.249 ± 0.102
BeyondAIME	Rationale-judged	0.867 ± 0.057	0.791 ± 0.014	0.744 ± 0.029
	DISCERN R1	0.440 ± 0.102	0.440 ± 0.080	0.540 ± 0.042

CERN chunk or PRM step is at most t. PRM inputs use the official `[Step i]` tags, one tag per step.

On those traces, min-aggregated PRM SNR-Gain is 0.223, against DISCERN R1 1.235 ± 0.045. At t=0.5 on seed 42, precision is 0.67 for DISCERN and 0.41 for the PRM. The greedy mean endpoint appears separately in Table 1.

For the hard-set diagnostics below we use ProcessBench (Zheng et al., 2024) OlympiadBench and Omni-MATH. Thresholds are selected by maximizing F1 on GSM8K and then transferred: t=0.20 for the round-2 DISCERN checkpoint and t=0.85 for the PRM. The localization diagnostic below evaluates related-step detection. A separate official-trace evaluation on GSM8K and MATH uses the $c_{\min} \leq 0.5$ rule and is reported below with UALIGN judge comparisons.

Whole-problem refusal. A trace is flagged if any step is $\leq t$. Table 27 reports error detection (recall on traces with an error), pass rate on fully correct traces, and F1. DISCERN has higher F1 on both hard sets (mean 77.7 vs. 73.9). The PRM detects more errors but rejects more fully correct traces.

Table 27: Minimum-reference whole-problem refusal on OlympiadBench and Omni-MATH. A trace is refused if any step/chunk is $\leq t$. Thresholds are GSM8K-tuned and transferred.

Setting	Error recall	Correct-trace pass	F1
OlympiadBench, DISCERN e2_240	83.1	74.0	**78.3**
OlympiadBench, Qwen2.5-Math-PRM-7B	95.9	63.4	76.4
Omni-MATH, DISCERN e2_240	89.7	67.6	**77.1**
Omni-MATH, Qwen2.5-Math-PRM-7B	98.2	56.0	71.3

Related-step localization. A hit requires that a gold error step is $\leq t$, or that the first low-score step lies within ± 1 of a gold error. Fully correct traces must remain $> t$ throughout. Table 28 shows that mean F1 is essentially tied (71.9 vs. 72.2). The PRM records more hits on erroneous traces and a lower pass rate on correct traces. The strongest gains over the PRM occur in whole-problem refusal F1 and BeyondAIME SNR-Gain.

Official ProcessBench GSM8K/MATH evaluation. To complement the hard-set diagnostics above, we evaluate the same round-2 DISCERN checkpoint on externally generated official ProcessBench traces. GSM8K contains 400 traces and MATH contains 1,000 traces with standard first-error

Table 28: Related-step localization on ProcessBench hard sets. A hit is a gold error step $\leq t$ or a first low-score step within gold ± 1.

Setting	Hit rate	Correct-trace pass	F1
OlympiadBench, DISCERN e2_240	68.5	74.0	71.2
OlympiadBench, Qwen2.5-Math-PRM-7B	90.5	63.4	**74.6**
Omni-MATH, DISCERN e2_240	78.3	67.6	**72.6**
Omni-MATH, Qwen2.5-Math-PRM-7B	92.2	56.0	69.7

and final-answer labels. We compare against UALIGN using its judge protocol (binary step judgments); the UALIGN native refusal protocol is reported only as a degenerate reference because it almost never refuses. Threshold $t = 0.5$ is used for DISCERN's conservative-reference $c_{\min}$ score, and the reported process metrics follow the ProcessBench definitions.

Table 29: Whole-trace error rejection on official ProcessBench GSM8K and MATH generator traces. Error recall is measured on traces with a gold first error, correct-trace pass is measured on fully correct traces, and F1 is their harmonic mean.

Split	Method	n	Error recall	Correct pass	F1
GSM8K	DISCERN $c_{\min} \leq 0.5$	400	56.0	94.3	70.3
GSM8K	UALIGN judge	400	53.6	96.9	69.0
MATH	DISCERN $c_{\min} \leq 0.5$	1000	76.3	86.9	81.3
MATH	UALIGN judge	1000	45.1	92.1	60.6

For answer-level selective prediction, both methods use the same final-answer correctness labels on the fixed supplied responses. The whole-trace process diagnostic uses the annotation `label == -1` for traces marked error-free. In Table 30, KC and KR count correct final answers that are accepted and rejected, respectively; UI and UR count incorrect final answers that are accepted and rejected. Thus, $KC + KR$ is invariant across methods: 200 of 400 responses on GSM8K and 500 of 1,000 on MATH.

Under this common answer-level labeling, DISCERN reaches GSM8K coverage/precision of 68.25%/0.681 and MATH coverage/precision of 49.40%/0.759, compared with 70.75%/0.675 and 70.00%/0.604 for UALIGN judge, respectively. Thus, on MATH the verifier trades coverage for higher accepted-answer precision, while on GSM8K the two methods are closely matched. The corresponding false-acceptance rates are 43.5% and 23.8% for DISCERN on GSM8K and MATH, versus 46.0% and 55.4% for UALIGN judge.

Table 30: Answer-level selective prediction on fixed official ProcessBench GSM8K and MATH traces, using the same final-answer correctness labels for both methods. Coverage is $(KC+UI)/n$, precision is $KC/(KC + UI)$, and false acceptance is $UI/(UI + UR)$. Coverage and FA are percentages. The correct-answer total $KC + KR$ is invariant across methods.

Split	Method	Coverage	Precision	FA	(KC, KR, UI, UR)	$KC + KR$
GSM8K	DISCERN $c_{\min} \leq 0.5$	68.25	0.681	43.5	(186,14,87,113)	200
GSM8K	UALIGN judge	70.75	0.675	46.0	(191,9,92,108)	200
MATH	DISCERN $c_{\min} \leq 0.5$	49.40	0.759	23.8	(375,125,119,381)	500
MATH	UALIGN judge	70.00	0.604	55.4	(423,77,277,223)	500

G Extended Closed-book QA Transfer

Protocol. We evaluate NQ-open (3,610 examples), PopQA (4,000), and TriviaQA (4,000) using the closed-book QA evaluation pipeline: a frozen 4B solver generates answers with temperature 0.2 and a maximum of 512 output tokens, and the mathematics-trained round-2 verifier scores the responses with temperature 0.2 and top-$p = 0.9$. Evaluation uses inference seed 42. Correctness is determined by normalized matching against gold answer aliases, an automatic reference-based criterion. This short-answer, gold-alias protocol covers NQ-open, PopQA, and TriviaQA; TruthfulQA is outside its evaluation scope.

Table 31: Single-seed closed-book QA transfer with the round-2 verifier. SNR-Gain integrates the threshold spectrum; solver accuracy is measured by normalized gold-alias matching. Sample counts identify the evaluated subsets.

Dataset	n	Solver acc.	SNR-Gain (mean)	SNR-Gain ($c_{\min}$)
NQ-open	3,610	25.7%	0.299	0.437
PopQA	4,000	18.6%	0.554	0.776
TriviaQA	4,000	52.8%	0.618	0.836

Table 32: False acceptance at $t = 0.5$ on the same closed-book QA responses. Each entry is an accepted-incorrect count divided by the total incorrect count, $UI/(UI + UR)$, followed by its percentage.

Dataset	FA (mean)	FA ($c_{\min}$)
NQ-open	1694/2681 (63.2%)	1315/2681 (49.0%)
PopQA	1624/3255 (49.9%)	1277/3255 (39.2%)
TriviaQA	841/1888 (44.5%)	664/1888 (35.2%)

Interpretation and scope. Both aggregations yield positive SNR-Gain on all three evaluated sets, supporting factual-transfer benefits beyond SimpleQA under the reference-matching protocol. At $t = 0.5$, minimum aggregation accepts fewer incorrect responses than mean aggregation, consistent with its conservative decision rule. These experiments report score-weighted enrichment and false acceptance for inference seed 42 on a frozen solver's responses.

The tables report SNR-Gain and false acceptance separately for mean and minimum aggregation.

H Human Validation of Chunk-Type Semantics

We perform a post-hoc audit of the semantic meaning of the verifier's two chunk types. Human labeling assigns individual spans to *knowledge* or *inference* while the original DISCERN tag is hidden; we then compare the human label with the verifier tag. This audit is independent of training and contains 1,300 parsed spans with no parse failures. It evaluates semantic type agreement on the generated spans.

Table 33: Human validation of DISCERN's knowledge/inference chunk types. The original verifier tag was hidden during human labeling.

Slice	Spans	Tag accuracy
All	1,300	83.0%
BeyondAIME	500	87.4%
SimpleQA	800	80.2%

The pooled confusion matrix (rows are model tags and columns are human labels) is:

model tag\human label	knowledge	inference
knowledge	415	85
inference	136	664

The observed agreement supports semantic interpretability on the audited spans. The table reports the dataset and model-tag composition of the audit.

I Error-Block Localization Diagnostic

We separately score whether, on incorrect solver traces, the verifier points at a block that is actually mathematically problematic. This error-block task complements the semantic type audit. A downstream or nearby invalid block counts as correct under the relaxed criterion if it is genuinely relevant to the error.

The labeled set is the same 100 BeyondAIME traces scored by the same round-2 verifier as Table 1. These same traces contain 33 correct and 67 incorrect solver answers, consistent with Table 1. This appendix evaluates error-block pointing on that fixed response population. Every trace was inspected by a human annotator and by an AI assistant; disagreements were reconciled into a consensus label.

Table 34: Dual-reviewed error-block diagnostic on the same 100 BeyondAIME traces and round-2 verifier as Table 1. The localization population consists of incorrect solver traces. Relaxed correctness includes core and genuinely related/downstream error blocks; strict correctness counts core blocks. Solver correctness is 33/100, consistent with the main table.

Quantity	Count	Rate
Solver correct / incorrect	33 / 67	33% / 67%
Correct accepted / correct rejected	16 / 17	48.5% / 51.5% of correct
Wrong accepted (localization missing)	3	4.5% of incorrect
Wrong rejected	64	95.5% of incorrect
Core error block (among 67 incorrect)	33	49.3%
Related/downstream error block	21	31.3% of incorrect
Off-target reason among wrong-rejected	10	15.6% of 64
Relaxed localization among incorrect	54/67	80.6%
Relaxed localization among wrong-rejected	54/64	84.4%
Trace-level semantic correctness	70/100	70.0%

Under the relaxed criterion, 54 of 67 incorrect traces point at a core or related error block (80.6%); conditioning on rejection raises this to 54/64 (84.4%) because the three false acceptances are counted as missing localization in the 67-trace denominator. Strict core-block pointing is 33/67 (49.3%). Trace-level semantic correctness is 70/100, pulled down by 17 rejections of correct solutions. On these BeyondAIME traces, the relaxed diagnostic shows that rejected-wrong traces usually identify a mathematically problematic block; the strict criterion separately measures core-block pointing.

J Qualitative Examples of Reasoning and Confidence Patterns

Response-level metrics do not show what a low chunk score refers to. Figure 2 presents three selected outputs: a factual error, an unproven restrictive assumption, and a confidence drop followed by recovery on later calculations. These examples show how the annotation can be inspected alongside the reasoning; Appendix I quantifies error-block pointing.

J.1 Reasoning under Restrictive Assumptions

An answer can go wrong because it restricts the solution space too early, even when the subsequent algebra is locally correct.

In **Case 2** (the integer pairs problem), the solver produces a correct local algebraic derivation ($z \in (-1, 1)$) but fails to consider the full solution space (e.g., negative integers). It implicitly assumes positivity and continues without revisiting that assumption. The final-answer label identifies the unsuccessful response; inspecting the trace reveals the restrictive assumption behind it.

DISCERN assigns low confidence at Step I3, directing attention to the restrictive assumption in this example. The dual-reviewed rates appear in Appendix I.

J.2 Dynamic Confidence Recovery

In **Case 3** (the dice re-roll problem), confidence decreases at Step I0 and then temporarily recovers on sub-calculations I1–I2. The later chunks therefore receive distinct scores despite the earlier warning, giving the reader a more differentiated annotation than a single response-level score.

K Detailed Method and Training Protocol

K.1 Cold-start Dataset Construction

The cold-start corpus uses repeated solver attempts to estimate which questions are answered consistently and which produce mixed outcomes. For each DAPO-Math-17k question q_i, **Qwen3-4B-Instruct** generates $N = 10$ distinct reasoning paths and answers $\{a_{i,1}, \ldots, a_{i,N}\}$. Their empirical accuracy is

$$p_i = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}[\text{answer } a_{i,j} \text{ is correct}],$$

which serves as a solver-specific proxy for question difficulty: low p_i indicates a low observed final-answer success rate.

A key challenge is the concentration of solver accuracies near the two extremes. We therefore use a reproducible bucketed sampling rule. Let S_k be the set of questions with $\mathrm{round}(p_i \times 10) = k$ for $k = 0, \ldots, 9$. We retain all questions in these ten non-perfect buckets, compute $N_{\text{avg}} = \lfloor (\sum_{k=0}^{9} |S_k|)/10 \rfloor$, and uniformly sample N_{avg} questions from the perfectly solved bucket S_{10}. Over 17,398 cold-start problems, the resulting bucket counts are $1,709, 606, 547, 522, 531, 563, 729, 917, 1,250, 1,944, 931$ (for $p_i = 0, 0.1, \ldots, 1.0$); after removing incomplete traces the final set contains 10,248 problems and 101,585 traces. This increases coverage of intermediate difficulty without synthesizing or reweighting examples.

K.2 Verifier Design and Training

The verifier is a single Qwen3-4B-Instruct generative Transformer, initialized independently of the solver. It takes a problem q and a reasoning path a and, in one pass, outputs a structured segmentation with localized confidence scores. We decompose the answer into alternating semantic chunks of two types:

- **Knowledge chunks:** statements of objective facts, formulas, or theorems, tagged with `<knowledge>`.
- **Inference chunks:** logical derivations, calculations, or intermediate conclusions, tagged with `<inference>`.

Each chunk is wrapped with an HTML-style tag that includes a predicted confidence attribute, e.g., `<knowledge confidence="0.8">...<inference confidence="0.4">...`. The verifier is instructed to preserve the **final answer** (`\boxed{}`) and annotates intermediate reasoning through tag insertion and light paraphrasing. The reward's boxed-answer check is defined below.

The semantic vocabulary is specified in the prompt, while training uses answer-level outcomes and score trajectories without human step-error labels. We assess whether readers interpret the resulting types consistently in a post-hoc audit with the original tags hidden. Agreement with human type labels is 83.0% (87.4% on BeyondAIME and 80.2% on SimpleQA; Appendix H). Error-block pointing asks the separate question of whether a low score identifies the core error (Appendix I).

Because tag generation and confidence prediction are non-differentiable, we train the verifier with **GRPO** (Group Relative Policy Optimization). Let π_θ be the verifier's policy. For each (q, a), it autoregressively produces a tagged version $\hat{a}$ and updates parameters via

$$\nabla_\theta J(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D}} \Big[\hat{A} \sum_t \nabla_\theta \log \pi_\theta(\hat{a}_t \mid \hat{a}_{<t}, q, a) \Big],$$

where $\hat{A}$ is the sequence-level advantage obtained by normalizing the cumulative rewards of a group of outputs for the same (q, a). Over training, the verifier learns to decompose mathematical answers and assign bounded confidence values $c_k \in [0, 1]$ to each semantic chunk. The model emits discrete confidence values; their empirical distributions are reported in Appendix A.2.

K.3 Reward Function Design

We combine an output-format gate with a multi-term sequence-level reward for answer-level discrimination and score structure:

$$R = R_{\text{format}} \times R_{\text{content}},$$

where $R_{\text{content}} = 0.4 \times R_{\text{detect}} + 0.4 \times R_{\text{accuracy}} + 0.2 \times R_{\text{topo}}$. The content terms supervise answer-level discrimination and score trajectories independently of the knowledge/inference tag identity. For continuous confidence values, we use a **triangular kernel** $\mathrm{Tri}(p,t) = 1 - |p - t|$, whose gradient magnitude is piecewise constant away from $p = t$.

Format Reward (R_{format}). Let $T(\hat{a})$ indicate valid, complete tagging: after stripping valid tag blocks from the verifier output $\hat{a}$, the remaining text is empty modulo whitespace. Let $b(a)$ and $b(\hat{a})$ be the extracted boxed answers, with $\perp$ denoting a missing extraction. The round-1 missing-box-tolerant gate is

$$R^{\text{tol}}_{\text{format}} = T(\hat{a})\, G_{\text{tol}}, \qquad G_{\text{tol}} = \begin{cases} 0, & b(a) \neq \perp,\ b(\hat{a}) \neq \perp,\ b(a) \neq b(\hat{a}), \\ 1, & \text{otherwise.} \end{cases}$$

Thus, a detected disagreement suppresses the content reward; a missing boxed answer alone does not. The strict ablation instead uses

$$R^{\text{strict}}_{\text{format}} = T(\hat{a})\, \mathbb{1}[b(a) \neq \perp,\ b(\hat{a}) \neq \perp,\ b(a) = b(\hat{a})]\,.$$

Both gates multiply the same content reward with weights $0.4/0.4/0.2$ and identical detection, accuracy, and topology logic. Appendix E.3 compares the resulting checkpoints.

Error Detection Reward (R_{detect}). We use a weakest-link surrogate for answer-level error detection. An incorrect final answer motivates assigning a low score to at least one chunk. Let $c_{\min} = \min_k c_k$. We define

$$R_{\text{detect}} = \begin{cases} c_{\min}, & \text{if answer is correct,} \\ 1 - c_{\min}, & \text{if answer is incorrect.} \end{cases}$$

This rewards low minimum scores for incorrect final answers and high minimum scores for correct final answers. For a correctness probability p, the standalone expected reward is $(1-p) + (2p-1)c_{\min}$, favoring an endpoint when $p \neq 0.5$. This classification-oriented objective trains $c_{\min}$ as a decision score.

Difficulty-aware Reliability Reward (R_{accuracy}). To set a difficulty-aware target for the mean chunk score $\bar{c} = \frac{1}{K}\sum_{k=1}^{K} c_k$, we use final-answer correctness and the solver-specific difficulty proxy. The target is 1 for correct answers and p_{prior} for incorrect answers. The latter target discourages assigning uniformly low scores to every incorrect response.

$$R_{\text{accuracy}} = \begin{cases} \mathrm{Tri}(\bar{c}, 1), & \text{if answer is correct,} \\ \mathrm{Tri}(\bar{c}, p_{\text{prior}}), & \text{if answer is incorrect.} \end{cases}$$

When $p_{\text{prior}} > 0$, this term discourages assigning zero to every chunk of an incorrect response. The term acts on the mean chunk score.

Score interpretation and aggregation roles. The mean chunk score $\bar{c}$ is a learned response-level score shaped by the outcome, difficulty, and trajectory objectives. We use **mean confidence** for default ranking and abstention (accept if $\bar{c} \geq t$), and **minimum confidence** as a conservative weakest-link reference. Since $c_{\min} \leq \bar{c}$, minimum aggregation accepts a subset of the mean policy's responses at the same threshold and boundary convention. Conf AUC and AURC assess the ranking of each aggregation; SNR-Gain measures score-weighted enrichment (Appendix C.1). Calibration metrics are secondary diagnostics.

Topological Reward (R_{topo}). This term encodes preferred score-trajectory shapes by treating $\mathbf{c} = (c_1, \ldots, c_K)$ as an ordered sequence. We define head-tail degradation $\Delta = c_1 - c_K$, maximum step drop $D_{\max} = \max_{2 \le k \le K}(c_{k-1} - c_k)$, and volatility σ (standard deviation).

For *correct final answers*, $R_{\text{topo}} = 1.0$ is assigned for high-confidence plateaus ($\sigma < 0.15, c_{\min} > 0.6$), and 0.8 for trajectories satisfying $\Delta < 0.2$ and $D_{\max} < 0.4$. For *incorrect final answers*, we assign 1.0 for a large local drop ($D_{\max} > 0.6$) or a large head-to-tail decrease ($\Delta > 0.4$), and a baseline reward of 0.8 for low mean confidence ($\bar{c} < 0.5$).

We use R_{topo} as a training-stabilization term, retaining its form and thresholds unchanged from earlier Qwen2.5-3B experiments; at that scale we observed that without it the verifier tended to assign high confidence to clearly erroneous steps. At 4B, removing it lowers one-round five-seed SNR-Gain from 1.235 ± 0.045 to 1.118 ± 0.045 (Table 39). Figure 8 shows the associated training-trajectory effect: confidence on incorrect answers rises with step count without R_{topo} but decreases toward zero under the full reward.

The cutoffs above define training-reward preferences; evaluation thresholds determine which responses are accepted. SNR-Gain integrates over evaluation thresholds, whereas coverage, risk, and precision at a specified threshold describe an operating point. The component-removal ablation evaluates the contribution of the trajectory term as a whole.

L Extended Related Work and Design Context

Research on uncertainty estimation for hallucination detection encompasses verbalizing, latent information, consistency-based, and semantic clustering approaches (Xia et al., 2025). Verbalizing methods directly elicit confidence scores but suffer from overconfidence and brittleness to prompts (Kadavath et al., 2022); latent methods compute token-level or hidden-state statistics such as perplexity or predictive entropy, yet they produce a single response-level score that cannot separate factual errors from reasoning flaws (Manakul et al., 2023; Malinin & Gales, 2020; Chen et al., 2024). Consistency-based approaches measure agreement across multiple sampled responses (Cole et al., 2023; Huang et al., 2023; Gao et al., 2024; Chen & Mueller, 2024) and semantic clustering methods group equivalent meanings to compute semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024), but both remain at the question level. Our verifier instead emits structured `<knowledge>` and `<inference>` segments with per-chunk scores, which we aggregate for response-level abstention.

Existing refusal and factuality-alignment methods use different supervision granularities. R-Tuning and SE-Tuning provide question-level signals (Zhang et al., 2024; Tjandra et al., 2024); TruthRL and FiSCoRE use response-level rewards (Wei et al., 2025; An & Xu, 2025); other self-awareness frameworks use external knowledge bases or manual annotation (Qian et al., 2025; Cao, 2024). Evidence of differences between factual and reasoning processes (Luo et al., 2026; Yeo et al., 2025) motivates inspecting both kinds of statements. Like these response-supervised approaches, DISCERN trains from answer-level labels. It differs in emitting typed chunk scores under a multi-term reward and refining an external verifier through sparring. Process Reward Models (PRMs) also assign step-level scores (Lightman et al., 2023; Zhang et al., 2025), typically for process correctness and Best-of-N search. We evaluate Qwen2.5-Math-PRM-7B (Zhang et al., 2025) as an off-the-shelf process scorer on the same traces with a min-based reference rule; this comparison evaluates a pretrained process scorer for response rejection.

M Detailed Evaluation Protocols and Full Reference Results

Training configuration. The base solver, round-1 verifier, and outcome-only control start from Qwen3-4B-Instruct; later verifier rounds continue from the preceding checkpoint. We train on DAPO-Math-17k using the veRL GRPO implementation. Round-1 and sparring verifier checkpoints use 240 updates per round, eight rollouts per prompt, training batch size 256, and actor learning rate 10^{-6}. The outcome-only control's sequence limits, micro-batching, and reward configuration are described below.

Evaluation map. Table 1 supplies benchmark context and a same-trace untrained control. Table 2 summarizes default-policy matched-coverage risk; Table 4 compares continuation checkpoints

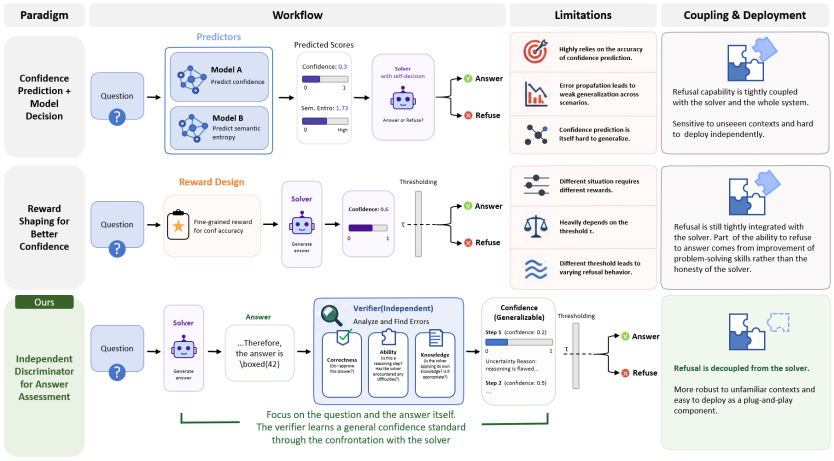

Figure 6: Schematic comparison of solver-coupled refusal and a decoupled verifier. DISCERN assigns chunk scores to supplied solver responses and aggregates them for selective abstention. Decoupling permits evaluation on responses from different generators. This schematic was created manually by the authors.

on fixed response populations. Appendix E.2 provides the corresponding mean-score risk grids, while Appendix A.1 collects the minimum-reference curves and operating points. ProcessBench evaluates whole-trace error rejection separately from answer selection. The primary static-refresh checkpoint and the additional static-reference checkpoint retain separate identities and result tables in Appendix E.4.

Evaluation cohorts. The greedy endpoint uses one response per problem. The outcome-only, component-removal, and continuation comparisons each align supplied responses within their evaluation cohort; inference repetitions use seeds 42–46 for fixed trained checkpoints. Each table identifies its score aggregation. For example, round-1 SNR-Gain 1.148 in the outcome-only comparison uses minimum scores, whereas 1.065 in the continuation comparison uses mean scores.

M.1 Internal Single-Reward Verifier

The **single-reward verifier** uses the same Qwen3-4B model family, GRPO implementation, and structured output interface as DISCERN: it segments the response into `<knowledge>` and `<inference>` chunks and assigns a confidence to each chunk. Its only training reward is the outcome-based quadratic score

$$R_{\text{single}}(y, p) = 2py - p^2, \qquad p = c_{\min},$$

where $y \in \{0, 1\}$ is answer correctness. The control optimizes this quadratic outcome reward directly. DISCERN multiplies a format/preservation gate by its weighted detection, difficulty-aware reliability, and confidence-trajectory rewards.

The single-reward verifier serves primarily as a *within-round reward control* for the round-1 DISCERN verifier. Both start from Qwen3-4B-Instruct, train for 240 steps on the same first-round DAPO verifier set, and share the model family, structured output interface, GRPO implementation, rollout count, batch size, and learning rate. Their five-seed step-240 comparison in Table 9 compares the outcome-only and full-reward training recipes on the same frozen solver responses; a paired hierarchical bootstrap quantifies the difference. The single-reward control uses maximum prompt/response lengths of 10,240/8,192 tokens and an actor micro-batch of 8 per GPU; round-1 DISCERN uses 12,288/12,288 tokens and an actor micro-batch of 2 per GPU. The finer removals of individual reward components are five-seed SNR-Gain diagnostics in Table 39, supplemented by mean- and minimum-score AUC in Appendix E.1.

The main BeyondAIME table reports the completed DISCERN verifier after the second-round curriculum, under greedy decoding, with mean and $c_{\min}$ aggregation shown separately. The within-round table compares the reward recipes; the endpoint table evaluates the final deployed verifier.

M.2 Evaluation Setup

We assess ranking with **Confidence AUC** and **AURC**, and retain **SNR-Gain** from Wu et al. (2025) as a complementary threshold-integrated, score-scale-dependent metric (Appendix C.1). We report **smECE** (Blasiok & Nakkiran, 2024) and **Brier Score** as secondary diagnostics. Factuality alignment uses **Truthfulness**, **Precision**, **Specificity**, and **Unknown Recall** (Appendix C.2). SNR-Gain separately integrates accepted-correct and accepted-incorrect masses over $t \in [0, 1]$ and takes their ratio relative to the always-answer baseline. Appendix C.1 gives the derivation and an example of its score-scale dependence.

We evaluate on **BeyondAIME** (ByteDance Seed et al., 2025) (100 ultra-hard problems, primary testbed), **AIME-2024/2025** (competition-level mathematics), and **SimpleQA** (Wei et al., 2024) (zero-shot cross-domain transfer). The main BeyondAIME, AIME, and SimpleQA comparisons use greedy decoding (one sample per problem; BeyondAIME solver accuracy 33%). We report both mean chunk confidence and $c_{\min} = \min_k c_k$. SNR-Gain is reported for each dataset and solver population; SimpleQA evaluates cross-domain transfer. A separate MATH-trained, high-temperature AIME stress test is reported in Appendix D, with its checkpoint and evaluation protocol specified separately.

To assess verifier scale, we also evaluate a separate 30B checkpoint on the frozen Qwen3-4B-Instruct-2507 solver with temperature 0.2 and inference seeds 42–46. We report mean and $c_{\min}$ SNR-Gain, mean-confidence Conf AUC, solver accuracy, abstention rate at the reported operating point, and parse-failure rate. The same checkpoint is evaluated on SimpleQA with temperature 0.2, top-$p = 0.9$, maximum generation length 8192, and maximum model length 16384; SNR-Gain and AUC use parse-valid outputs. Together with the Gemini generator-transfer evaluation below, these experiments provide scale and generator diversity beyond the main 4B configuration.

We compare against frontier models (GPT-5, GPT-OSS-120b, o4-mini, Claude-Sonnet-4.5, Grok-4, Gemini-2.5-Pro, GLM-4.5/4.6, Qwen3-max, Kimi-K2-0905) and the confidence baselines (Confidence-Brier/CE/min/prod, PPO-Value, PPO Only) from Wu et al. (2025), where frontier model confidence scores are elicited via prompting. We additionally reproduce **UALIGN** (Xue et al., 2025) and evaluate a **response-probability adaptation of CISC** (Taubenfeld et al., 2025) on the same benchmark with Qwen3-4B-Instruct (Appendix B). Our reproduced Qwen3-4B-Instruct baseline matches the reported reference value as a sanity check. All SNR-Gain values use the reference definition and threshold-integration protocol; published confidence/PPO values are attributed to their source, and the reproduced methods use the protocols described here. The single-reward, component-removal, continuation, and PRM comparisons evaluate their checkpoints on aligned supplied responses.

Among the response-level baselines in Table 1, `confidence-min` and `confidence-prod` aggregate token-level scores and obtain SNR-Gain 0.683 and 0.806. The table also reports calibration diagnostics: `ppo-value` has smECE 0.061, compared with DISCERN's 0.092.

Wu et al. also report claim-level confidence baselines under a weakest-link refusal rule, rejecting the response if any claim is below threshold. This is a whole-response minimum-aggregation policy. Their Qwen3-4B claim-level confidence-min and confidence-prod obtain SNR-Gain 0.301 and 0.183 (Conf AUC 0.788 and 0.741). DISCERN obtains SNR-Gain 1.456 with mean aggregation and 1.746 with $c_{\min}$, and mean-confidence Conf AUC 0.913. These published claim-level results provide fine-grained confidence references alongside the response-level baselines.

We additionally evaluate an untrained Qwen3-4B-Instruct-2507 verifier on the same 100 greedy BeyondAIME solver traces, using the same prompt and temperature 0, without GRPO. Separately, we score an off-the-shelf process reward model, Qwen2.5-Math-PRM-7B (Zhang et al., 2025), on the same BeyondAIME traces used for the one-round ablation (Appendix F). This same-trace comparison evaluates an off-the-shelf 7B process scorer alongside the 4B verifier.

Table 35: Greedy 4B evaluation. BeyondAIME contains 100 frozen-solver traces (33 correct); the untrained verifier receives the same traces and prompt. AIME sets contain 30 problems each. The complete baseline and calibration results are in Tables 1 and 36.

Dataset / verifier	Aggregation	Conf AUC	SNR-Gain
BeyondAIME / untrained	Mean	0.784	0.045
	Minimum	0.762	0.107
BeyondAIME / DISCERN	Mean	0.913	1.456
	Minimum	0.791	1.746
AIME-2024 / DISCERN	Mean	0.857	1.121
	Minimum	0.826	1.264
AIME-2025 / DISCERN	Mean	0.819	0.768
	Minimum	0.819	0.974

For an additional refusal-oriented transfer check, we evaluate the verifier at threshold $t = 0.5$ on a separate 687-item Gemini refusal set with standard answers. We report coverage and accepted-answer accuracy relative to the always-answer baseline on this set.

M.3 Mathematical and generator-transfer results

Evaluation on BeyondAIME. Table 1 reports BeyondAIME under greedy decoding (one sample per problem, 100 problems, solver accuracy 33%), with mean and $c_{\min}$ aggregation shown as separate DISCERN rows. We use mean aggregation as the default ranking and abstention score: it attains Conf AUC 0.913 and SNR-Gain 1.456. smECE 0.092 and Brier 0.123 are secondary diagnostics (Section 3.2). The conservative-reference $c_{\min}$ aggregation attains SNR-Gain 1.746 and Conf AUC 0.791. The headline AUC 0.913 thus characterizes the default mean score; the minimum reference has different discrimination and operating points. An untrained Qwen3-4B-Instruct-2507 verifier, scored on the same 100 solver traces with the same prompt at temperature 0, has lower SNR-Gain (0.045 under mean aggregation and 0.107 under $c_{\min}$). Seed-level intervals, paired bootstrap results, and risk–coverage curves for the five-seed protocol are reported in Appendix A. The within-round five-seed reward comparison is included in Table 9: on the same frozen-solver traces, the round-1 full-reward verifier raises SNR-Gain from 0.862 ± 0.094 to 1.148 ± 0.147 relative to the single-reward control. Table 11 additionally reports $c_{\min}$-based AURC for the completed round-2 verifier versus that same single-reward control (0.506 ± 0.039 vs. 0.576 ± 0.038; nAURC 0.379 ± 0.073 vs. 0.579 ± 0.053). This AURC contrast compares the completed two-round system with the single-reward control. AURC summarizes selective risk across the full coverage range. At the reference $c_{\min} \geq 0.5$ operating point, risk falls from 0.513 ± 0.047 to 0.287 ± 0.058 while coverage falls from 0.518 ± 0.011 to 0.278 ± 0.019; these operating points describe the minimum policy's coverage–risk trade-off. The Gemini comparison further shows that an external 4B verifier raises SNR-Gain from 0.020 to 0.500 while the generator is held fixed. On a separate five-seed evaluation of 100 frozen Gemini-2.5-Flash traces (45 correct solver responses), the round-2 verifier attains mean SNR-Gain 0.925 ± 0.114 under mean aggregation and 0.959 ± 0.414 under $c_{\min}$; the corresponding coverages are $29.6\% \pm 4.2$ percentage points and $10.4\% \pm 2.3$ percentage points. This provides a multi-seed cross-generator evaluation with both aggregation policies.

Performance on AIME. Table 36 reports greedy decoding on AIME-2024 and AIME-2025. Conf AUC describes ranking and SNR-Gain describes score-weighted selection; smECE and Brier in the same lists are secondary diagnostics. Mean aggregation attains SNR-Gain 1.121 / Conf AUC 0.857 / smECE 0.097 / Brier 0.137 on AIME-2024 and 0.768 / 0.819 / 0.246 / 0.236 on AIME-2025; $c_{\min}$ attains 1.264 / 0.826 / 0.167 / 0.167 and 0.974 / 0.819 / 0.248 / 0.227. Five-seed intervals are reported in Appendix A.

M.4 Verifier scale and factual transfer

The 30B scale-up preserves useful discrimination under the same frozen-solver protocol: mean-confidence Conf AUC is 0.823, 0.864, and 0.862 on AIME-2024, AIME-2025, and BeyondAIME,

Table 36: Greedy evaluation on AIME-2024 and AIME-2025 (one sample per problem). DISCERN reports mean and $c_{\min}$ aggregation separately. Baseline rows report SNR-Gain from their original protocols.

	AIME-2024				AIME-2025			
Model	SNR-Gain ↑	Conf AUC ↑	smECE ↓	Brier ↓	SNR-Gain ↑	Conf AUC ↑	smECE ↓	Brier ↓
Confidence-Brier	1.153	–	–	–	1.097	–	–	–
PPO-Value	1.051	–	–	–	1.105	–	–	–
Confidence-CE	**1.320**	–	–	–	**1.352**	–	–	–
PPO Only	0.205	–	–	–	0.213	–	–	–
Qwen3-4B-Instruct	0.177	–	–	–	0.153	–	–	–
DISCERN (mean)	1.121	**0.857**	**0.097**	**0.137**	0.768	**0.819**	**0.246**	0.236
DISCERN ($c_{\min}$)	1.264	0.826	0.167	0.167	0.974	**0.819**	0.248	**0.227**

respectively. Its mathematics parse-failure rate is 4.0–4.7%. This experiment extends evaluation to a larger verifier under the frozen-solver protocol.

Table 37: 30B verifier on SimpleQA over five seeds (42–46; $n = 4326$). SNR-Gain values are mean $\pm$ sample SD. Coverage, precision, and false-acceptance (FA) are reported for mean and $c_{\min}$ aggregation; parse failures are 0.0%.

Aggregation	SNR-Gain	Coverage	Precision	FA	Parse fail
Mean	0.876 ± 0.034	$5.2\% \pm 0.1$	$22.2\% \pm 1.5$	$4.2\% \pm 0.2$	0.0%
$c_{\min}$	1.169 ± 0.079	$1.8\% \pm 0.1$	$26.2\% \pm 3.7$	$1.4\% \pm 0.1$	0.0%

On SimpleQA, the 30B checkpoint reaches SNR-Gain 0.876 ± 0.034 (mean) and 1.169 ± 0.079 ($c_{\min}$), alongside the reported 4B five-seed $c_{\min}$ value 0.725 ± 0.039. The 30B checkpoint attains accepted-answer precision of 22.2% (mean) or 26.2% ($c_{\min}$), at coverage of 5.2% and 1.8%, respectively.

Zero-Shot Generalization to SimpleQA. On SimpleQA under greedy decoding, mean aggregation attains SNR-Gain 0.441 and Conf AUC 0.674 (secondary smECE 0.291, Brier 0.321); $c_{\min}$ attains SNR-Gain 0.696 and Conf AUC 0.644. In the separate five-seed protocol, $c_{\min}$ SNR-Gain is 0.725 ± 0.039, while nAURC is 0.816 ± 0.013. Table 11 reports selective risk 0.894 ± 0.004 at $c_{\min} \geq 0.5$ and coverage 0.260 ± 0.005, corresponding to 10.6% accepted-answer precision. This is above the frozen solver's roughly 5% always-answer accuracy, at the reported coverage. Accepted-answer precision measures selection at this operating point; SNR-Gain summarizes score-weighted enrichment across thresholds.

On the separate 687-item Gemini refusal set with standard answers, the $t = 0.5$ policy accepts 62.5% of items and achieves 78.8% accuracy among accepted answers, compared with 62.2% for the always-answer baseline (+16.6 percentage points). Because the generator is Gemini rather than Qwen3-4B-Instruct, this result provides a generator-transfer check complementary to the 30B verifier scale-up.

Extended Closed-book QA Transfer. To assess factual transfer beyond SimpleQA, we additionally evaluate the mathematics-trained round-2 verifier on NQ-open, PopQA, and TriviaQA, totaling 11,610 examples with a frozen 4B solver and inference seed 42. Mean/$c_{\min}$ SNR-Gain is positive on all three evaluated sets: $0.299/0.437$, $0.554/0.776$, and $0.618/0.836$, respectively (Appendix G). These results extend the evaluation to three additional factual QA datasets under the same frozen-solver, reference-matching pipeline.

Actuality Alignment. Table 38 and Figure 7 report four actuality-alignment metrics on BeyondAIME following the UALIGN protocol (Xue et al., 2025), with all learned methods reproduced on Qwen3-4B-Instruct. We include an always-abstain reference because Specificity and Unknown Recall can be maximized trivially by refusing every question. DISCERN attains 84.0% Truthfulness, 12 points above the 72.0% always-abstain reference, while retaining 67.9% Precision on known questions. Always abstention reaches 100% Specificity and Unknown Recall, compared with 90.3% for DISCERN. For DISCERN, the 72 unknown questions comprise 65 refusals, 7 incorrect answers,

Table 38: Actuality-alignment metrics on BeyondAIME. CISC uses the response-probability adaptation; always-abstain supplies a refusal-only reference.

Method	Truth.	Prec.	Spec.	U-Rec.
Always abstain	72.0	0.0	100.0	100.0
DISCERN	84.0	67.9	90.3	90.3
CISC (adapted)	69.0	75.0	66.7	69.6
UAlign	35.0	64.5	21.7	24.6

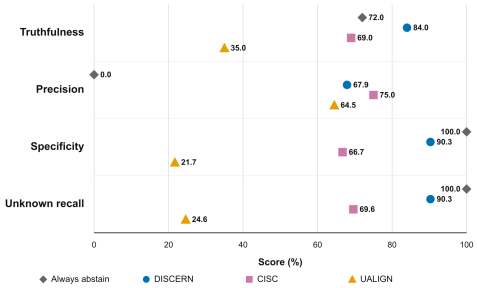

Figure 7: Actuality-alignment comparison with the always-abstain reference. DISCERN improves Truthfulness over trivial refusal by correctly answering known questions, with the refusal-only metrics shown alongside.

and no correct answers ($UR = 65$, $UI = 7$, $UC = 0$); consequently, Specificity and Unknown Recall are numerically identical. The result characterizes the trade-off between answering known questions and refusing unknown ones.

N Additional Reward and Curriculum Diagnostics

Table 39: BeyondAIME one-round reward ablation and an off-the-shelf PRM, five inference seeds 42–46 where a SD is shown (mean $\pm$ sample SD). Full, no-topo, no-detect, and no-accuracy checkpoints share the first-round corpus. The evaluation set contains 100 problems with 33 correct answers; reference-policy refusal fires if any chunk or PRM step is $\leq t$. The within-round outcome-only comparison and greedy endpoint are reported separately in Tables 9 and 1. Qwen2.5-Math-PRM-7B is a single point estimate on the same traces.

Model	SNR-Gain ↑
Qwen2.5-Math-PRM-7B	0.223
Ours (no accuracy)	1.087 ± 0.092
Ours (no detect)	1.005 ± 0.000
Ours (no topo)	1.118 ± 0.045
DISCERN R1 (full)	$\mathbf{1.235 \pm 0.045}$
DISCERN R2 (greedy; Table 1)	$\mathbf{1.456}$

N.1 Reward and Training-Round Diagnostics

Reward Component Ablation All reward ablations are evaluated after one round of training, before cumulative supervision and regenerated second-round data enter the pipeline. The single-reward and full-reward round-1 verifiers share the base initialization, first-round corpus, and 240-step budget. Table 39 reports five-seed SNR-Gain: full reward 1.235 ± 0.045, no R_{topo} 1.118 ± 0.045, no R_{accuracy} 1.087 ± 0.092, and no R_{detect} 1.005 ± 0.000 (identical on all five seeds). On the same traces, Qwen2.5-Math-PRM-7B (Zhang et al., 2025) attains SNR-Gain 0.223. At t=0.5 on seed 42, DISCERN answers 30 problems with precision 0.67 (20 correct / 10 incorrect), while the PRM answers 34 with precision 0.41 $(14/20)$. The comparison evaluates the 7B PRM and the 4B verifier on the same supplied traces. The matched $c_{\min}$ comparison of the archived outcome-only control against round-1 full reward remains in Table 9: SNR-Gain rises from 0.862 ± 0.094 to 1.148 ± 0.147, mean paired difference $+0.286$, seed-and-item hierarchical-bootstrap 95% interval $[+0.120, +0.479]$. Figure 8 further shows that confidence on incorrect answers rises with step count without R_{topo} but decreases toward zero under the full reward. Whole-problem refusal and related-step localization against ProcessBench-style gold steps are in Appendix F; the same appendix now also reports official GSM8K/MATH traces with standard first-error labels, where DISCERN's whole-trace F1 is 70.3/81.3 versus 69.0/60.6 for the UALIGN judge.

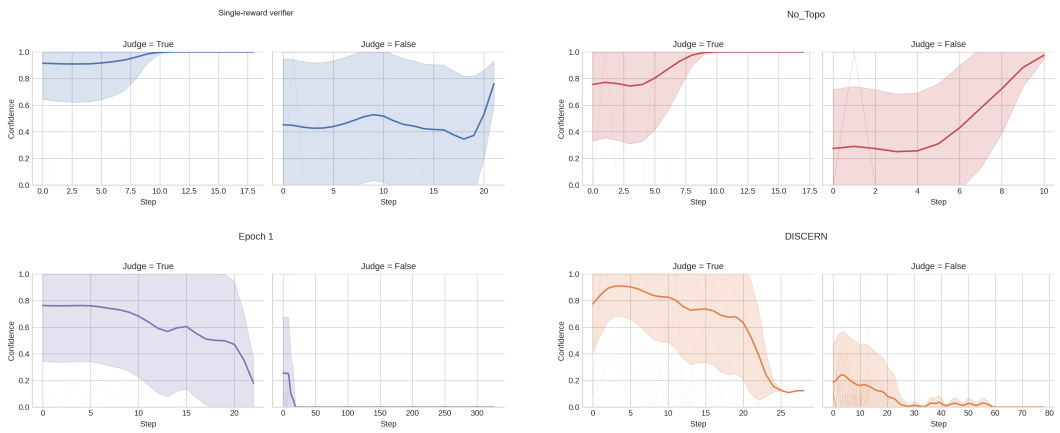

Figure 8: Mean confidence trajectories (with standard deviation bands) across four verifier variants on BeyondAIME.

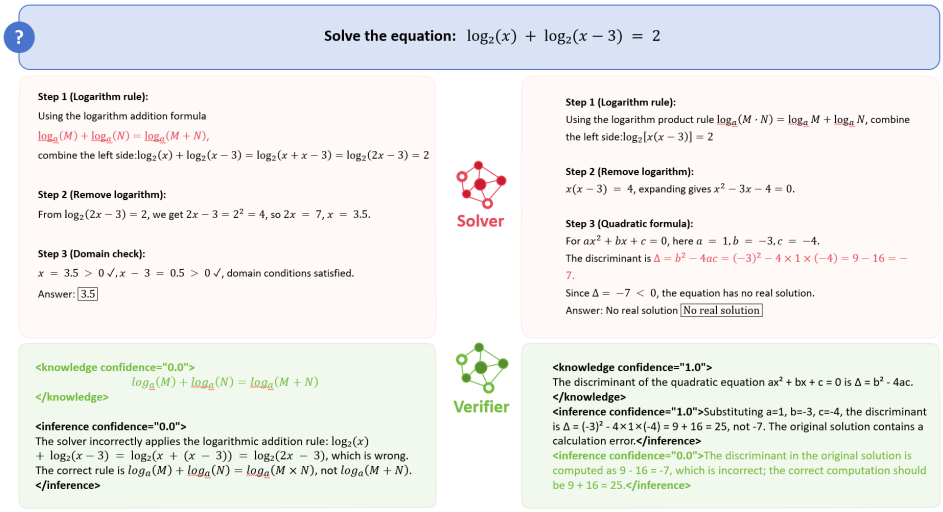

Figure 9: Selected excerpts of verifier annotations for two solver responses. The displayed chunks highlight a logarithmic-rule error (left) and a discriminant-calculation error (right). Remaining annotations and the preserved final answer are omitted from the verifier excerpts for readability.

Training Epochs We further examine performance across sparring rounds. The first round supplies the within-round reward comparison; the completed second round is the deployed checkpoint. The fixed cross-evaluation reports more round-1 false acceptance on updated-solver errors and subsequent recovery by round 2. Chunk-score endpoint mass falls from 95.4% to 65.8% and the number of distinct values rises from 3 to 10 (Figure 4), characterizing the change in chunk-score diversity. A third round reaches 1.754 ± 0.232 SNR-Gain, with other changes within inference-seed variation, so we retain the two-round checkpoint selected at the validation plateau. Table 4 compares the complete curriculum against frozen-solver fresh-data continuation. Appendix E.1 extends the ranking comparison to both aggregations on AIME-2024, AIME-2025, and BeyondAIME.

N.2 Output and Solver Diagnostics

Comparison with Verbalized Confidence. Our method differs from verbalized confidence approaches by emitting structured chunk scores rather than only a scalar response score. DISCERN attains higher SNR-Gain in our evaluation.

Table 40: Solver accuracy on the matched five-seed traces (temperature 0.2). Base is the frozen Qwen3-4B-Instruct-2507 solver; Trained solver is the sparring-updated opponent used to regenerate verifier training data, prompted as a generator. Entries are five-seed mean ± sample SD (%). Δ is the pooled difference in percentage points; brackets are a problem-clustered bootstrap 95% CI (unique problems resampled, all seeds kept, 10,000 replicates).

Dataset	Base solver	Trained solver	Δ (pp)
BeyondAIME	27.6 ± 1.9	29.2 ± 4.1	$+1.6\ [-2.2, +5.4]$
AIME-2024	56.7 ± 3.3	57.3 ± 4.9	$+0.7\ [-4.7, +6.7]$
AIME-2025	41.3 ± 5.1	45.3 ± 5.1	$+4.0\ [-2.0, +10.0]$
Pooled	35.6 ± 1.5	37.5 ± 2.3	$+1.9\ [-1.0, +4.8]$

Case Study. Figure 9 shows selected verifier annotations for two incorrect solutions to a logarithmic equation. The excerpts distinguish a misapplied logarithmic identity from an arithmetic error in the discriminant of $x^2 - 3x - 4 = 0$, illustrating how chunk scores expose different types of reasoning issues. Appendix I reports a dual-reviewed, relaxed error-block diagnostic on the same 100 BeyondAIME traces and the same round-2 verifier as the main table.

Sparring Challenge–Recovery Diagnostics. Across five seeds, the $c_{\min}$-trained sparring solver changes which problems it solves (Table 40): BeyondAIME from $27.6 \pm 1.9\%$ to $29.2 \pm 4.1\%$, AIME-2024 from $56.7 \pm 3.3\%$ to $57.3 \pm 4.9\%$, AIME-2025 from $41.3 \pm 5.1\%$ to $45.3 \pm 5.1\%$, and pooled from $35.6 \pm 1.5\%$ to $37.5 \pm 2.3\%$. The problem-clustered intervals for Δ are reported in the table. Its role is to search for responses that challenge the current verifier's confidence judgments and thereby expose hard examples for the next verifier update.

All four problem-clustered 95% intervals for the solver-accuracy difference include zero. The opponent's training objective is to expose hard responses for verifier training.

Solver adaptation versus frozen-solver data refresh. For the additional continuation-reference evaluation, inference seeds are 42–46, solver and verifier temperature is 0.2, and verifier top-p is 0.9. Its confidence intervals use 10,000 paired problem-bootstrap replicates, keeping the five inference seeds fixed and averaging the per-seed metric within each replicate. Confidence metrics use each checkpoint's parse-valid outputs; the separate risk–coverage analysis in Appendix A.1 retains failures below valid scores. The primary static-refresh evaluation appears in Appendix E.4. Table 25 reports the additional static-reference comparison. Its SNR-Gain differences are $+1.014$ on frozen-base responses and $+0.906$ on updated-solver responses, with AUC rising from 0.675 to 0.869 and from 0.625 to 0.836. The paired intervals quantify this static-reference comparison. In the primary static-refresh comparison, the respective SNR-Gain gaps are $+0.544$ and $+0.329$, and AUC gaps are $+0.163$ and $+0.112$ (Table 4).

The primary static-refresh results on frozen-base and updated-solver populations are summarized in Appendix E.4.

O Limitations and Future Directions

Evaluation scope and training variability. BeyondAIME contains 100 problems and each AIME set contains 30; five inference seeds measure variation on these fixed problem sets and trained checkpoints. The format-gate ablation compares distinct reward configurations (Appendix E.3). Repeating both continuation branches from several shared Round-1 initializations remains necessary to assess the cross-run robustness of sparring gains. Both verifier scales belong to the Qwen family; Gemini evaluates generator transfer. Broader verifier-family evaluation and alternative static-refresh schedules are future comparisons.

Experimental isolation and score interpretation. The same-response controls evaluate prediction quality on aligned supplied answers. The outcome-only control differs in sequence-length limits, micro-batching, and format gating; the rationale-judged checkpoint is an additional supervision comparison. The off-the-shelf PRM also differs in training objective and model size. These

comparisons complement the component removals, but do not jointly isolate every design choice. Published reference rows without aligned item-level predictions support benchmark-level, rather than paired, comparisons. The full reward has the strongest mean-score AUC among the reported component variants, whereas the minimum-score ordering differs and paired fixed-coverage risk intervals mostly include zero. The scores are learned decision signals: the standalone linear detection objective is not a proper probability scoring rule, and SNR-Gain depends on the numerical score scale. Mean aggregation is motivated by observed ranking and score resolution; its operating-point trade-offs with minimum aggregation should be assessed at attainable matched coverages. In the pooled updated-solver diagnostic, 97.9% of round-2 minima are binary. Expected risks within tied groups describe randomized selection, whereas deterministic thresholds accept or reject a complete tied group.

Outcome supervision and interpretation. Answer labels and difficulty-aware targets do not certify the validity or location of intermediate steps. The trajectory cutoffs encode hand-specified preferences; component removals test the terms as a whole, while sensitivity to the individual cutoffs remains untested. Semantic type agreement assesses readability on the audited spans, rather than the independent effect of typing on abstention. The error-block audit uses human–AI consensus, with strict core-block pointing of 49.3%; inter-annotator reliability is not estimated. Related-step localization is approximately tied with the PRM, and the qualitative examples illustrate selected patterns rather than their prevalence.

Training and compute scope. Static refresh regenerates responses with the unchanged solver, whereas sparring updates the solver before regeneration. Matching verifier-training budgets does not match total pipeline cost, which also includes opponent training and response generation. Previously logged inference evaluation totals approximately 8.546 GPU-hours and training about 185.1 wall-clock hours on eight A800s (approximately 1,481 A800 GPU-hours). These logged totals exclude the strict-format ablation and additional static-refresh runs; the comparisons do not establish end-to-end compute efficiency.

Factual transfer and protocol coverage. SimpleQA remains difficult in absolute terms: the 4B minimum policy has 89.4% selective risk at 26.0% coverage, and the 30B policies reach 22.2% precision at 5.2% coverage (mean) and 26.2% at 1.8% (minimum). The additional 11,610 NQ-open, PopQA, and TriviaQA examples broaden evaluation, but use one inference seed and automatic gold-alias matching; valid paraphrases outside the alias set may be marked incorrect. These QA summaries lack matched baselines, coverage, accepted-answer precision, and documented parser-failure counts and handling. Their AUC/AURC aggregation was unspecified in the source summary, so those entries are omitted. A unified QA comparison should report these quantities together on identical responses and labels. More generally, confidence metrics that exclude parsing failures can use different valid subsets of the same supplied answers. The rationale-judged comparison's source top-k analysis does not specify its within-tie selection rule, so it is kept separate from the tie-averaged expected-risk analysis. These protocol boundaries and the observed QA operating points limit claims of broad reliable answering.

Future work: semantic types as a verification interface. The knowledge/inference representation combines a design idea with an empirical finding: distinct semantic tags make the verifier output inspectable, and their agreement with human span labels reaches 83.0% although the content reward is independent of tag identity. Existing claim-level confidence methods, including Wu et al. (2025), provide useful fine-grained reference comparisons. A matched typed-versus-untyped experiment holding data, initialization, reward, and training budget fixed would test the independent performance contribution of typing. A complementary trained scalar verifier would test decomposition, with differences in applicable reward terms explicitly accounted for. Answer-only and fixed-segmentation controls would distinguish answer checking from the contribution of reasoning annotation; type-conditioned supervision is another direction.

Ethics. This work uses existing language models and public benchmarks, together with post-hoc annotation of model outputs. Verifier scores can create false reassurance and should not substitute for domain-expert review in high-stakes settings.