Hard2Verify: A Step-Level Verification Benchmark for Open-Ended Frontier Math

Shrey Pandit*, **Austin Xu***, **Xuan-Phi Nguyen**, **Yifei Ming**,
Caiming Xiong, **Shafiq Joty**
Salesforce AI Research

Data: https://huggingface.co/datasets/Salesforce/Hard2Verify

Code: https://github.com/SalesforceAIResearch/Hard2Verify

Abstract

Large language model (LLM)-based reasoning systems have recently achieved gold medal-level performance in the IMO 2025 competition, writing mathematical proofs where, to receive full credit, each step must be not only correct but also sufficiently supported. To train LLM-based reasoners in such challenging, open-ended settings, strong verifiers capable of catching step-level mistakes are necessary prerequisites. We introduce *Hard2Verify*, a human-annotated, step-level verification benchmark produced with over 500 hours of human labor. Hard2Verify is designed to rigorously assess step-level verifiers at the frontier: Verifiers must provide step-level annotations or identify the first error in responses generated by frontier LLMs for very recent, challenging, and open-ended math questions. We evaluate 29 generative critics and process reward models, demonstrating that, beyond a few standouts, open-source verifiers lag closed source models. We subsequently analyze what drives poor performance in step-level verification, the impacts of scaling verifier compute, as well as fundamental questions such as self-verification and verification-generation dynamics.

1 Introduction

Mathematical reasoning serves as a gold-standard evaluation setting for benchmarking reasoning progress in large language models (LLMs). Over the past half-decade, benchmarks have been introduced to assess LLMs at the grade-school (Cobbe et al., 2021), high-school (Hendrycks et al., 2021), university (Zhang et al., 2023), and competition math level (MMA, 2025; He et al., 2024a; Gao et al., 2024). However, the progress of mathematical reasoning ability of LLMs has outpaced benchmark creation, with every subsequent release of a frontier LLM saturating new benchmarks, most recently with GPT-5 Pro achieving 96.5%+ on AIME 2024. As a result, recent efforts (Glazer et al., 2024; Phan et al., 2025) have written novel, unseen mathematical questions to test LLMs.

While the training approaches of closed frontier models remain a secret, open-source progress in mathematical reasoning has been driven by scaling *reinforcement learning from verifiable rewards* (RLVR) (Lambert et al., 2024), with the breakthrough of DeepSeek-R1 (Guo et al., 2025) leading to an explosion of interest. This paradigm requires training data with solutions that are easily *verifiable*, i.e., have solutions that can be easily checked against a known ground-truth by string matching or symbolic checkers. Math benchmarks, for the most part, also adopt the verifiable setup, where a model response is considered correct if its final answer matches the ground-truth. Answer correctness, while a necessary condition for overall solution correctness, is not sufficient: LLMs can produce incorrect intermediate reasoning but conclude with correct final answers (Lightman et al., 2023; Zheng et al., 2024a; Setlur et al., 2025).

The next frontier for LLMs is solving problems that are *hard to verify*. A grand example of such a problem is proving the *Riemann hypothesis*, where the expected solution is not a short phrase, but a multi-step proof. To verify correctness, each step must be rigorously checked. Hints of open-ended problem solving abilities already exist: advanced reasoning systems (OpenAI, 2025a; Google, 2025a; Huang and Yang, 2025) have achieved gold-level performance in the 2025 IMO. Here, LLM outputs were judged at the step-level by human experts who determined if steps are both correct and sufficiently supported, with supporting lemmas and claims all appropriately stated and applied.

Training reasoning LLMs capable of open-ended problem solving requires scalable *automatic evaluation*: Not every LLM rollout during RLVR train-

*Equal contribution. Correspondence:shrey.pandit@salesforce.com

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22502–22517
July 2-7, 2026

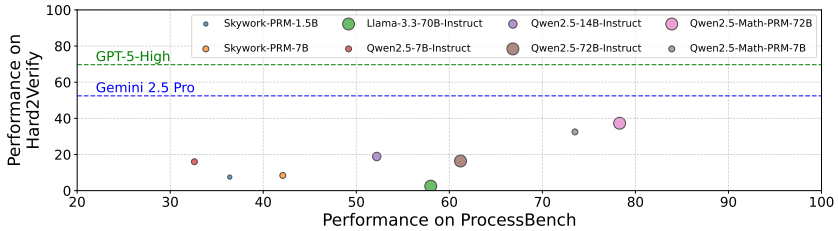

Figure 1: Comparison of models evaluated on both ProcessBench (Zheng et al., 2024a) and our Hard2Verify benchmark. Past benchmarks do not sufficiently evaluate in the frontier-level math settings that Hard2Verify does; On the same error identification task, Qwen2.5-Math-PRM-72B performance drops from ProcessBench state-of-the-art at 78.3 to 37.3 on Hard2Verify.

ing can be audited by human experts. Rather, evaluation in open-ended settings requires *step-level verifiers*, typically process reward models (PRMs) or generative critic models. Such verifiers have already been used to provide dense process rewards (Lightman et al., 2023; Shao et al., 2024; Zha et al., 2025). Furthermore, step-level verifiers are also used in many test-time scaling methods, selecting the most promising candidate from multiple solutions or steps (Snell et al., 2024; Yu et al., 2025; Lifshitz et al., 2025; Zhou et al., 2025b). However, are these step-level verifiers sufficient for pushing the frontier of mathematical reasoning?

This work introduces *Hard2Verify*, which benchmarks verifiers in assessing *frontier* LLM responses to difficult, recent, and open-ended math problems. We curate challenging problems from recent math competitions like IMO and Putnam, sample responses from three strong LLMs, GPT-5 (high) (OpenAI, 2025), Gemini 2.5 Pro (Google, 2025b), and Claude Sonnet 4 (thinking) (Anthropic, 2025), and employ PhD-level math experts to annotate each model-generated step. The resulting benchmark is the culmination of 500+ hours of human effort encompassing three rounds of independent agreement checks, yielding 1860 rigorously graded steps across 200 unique model responses.

Beyond operating at the frontier, *Hard2Verify* distinguishes itself from existing benchmarks for step-level annotation (Table 1). First, we emphasize collecting open-ended questions, with 78.5% of our samples being open-ended. This way, verifiers cannot "cheat" if they have seen the question or ground-truth answer during training; rather verifiers must substantively assess step correctness. Second, step correctness is judged not only on correctness, but also based on whether all invoked results, such as supporting lemmas or claims, are correctly stated and applied; saying "X follows from Y" receives no credit if Y is not sufficiently justified or properly invoked. Third, *Hard2Verify* focuses on benchmarking verifiers in naturally occurring application settings: Verifiers must assess *model-written* responses, which often differ dramatically from human-written reference answers.

We benchmark 29 models spanning proprietary models to open-weight models to PRMs. Compared to past work, *Hard2Verify* represents a step up in difficulty, as shown in Fig. 1; Models capable of scoring 60%+ on ProcessBench (Zheng et al., 2024a) are unable to crack 20% on *Hard2Verify*. Our analysis reveals that this degraded performance is because weaker verifiers cannot identify mistakes, marking nearly *every* step as correct. We additionally analyze several fundamental questions in step-level verification: How should one to scale verifier compute? What are the impacts of self-verification? How much easier is generation than verification for frontier models?

2 Background and Related Work

LLM-based verification. To meet demands for scalable evaluation, LLM-based evaluators were first used in chat settings (Zheng et al., 2023). However, as LLMs are deployed in challenging reasoning settings (Ke et al., 2025), recent work has shown the need for more capable reasoning evaluators (Frick et al., 2024; Tan et al., 2024; Zhou et al., 2025b). To get denser signals, focus quickly shifted to PRMs (Lightman et al., 2023) and synthetic ways to curate step-level training data (Wang et al., 2023; Luo et al., 2024). However, Shao et al. (2024) showed that dense reward signals for policy optimization brings limited improvement over outcome-level baselines. This observation stems from the fact that PRMs only measure if a step *could* lead to a correct final answer, not whether the step is correct in any absolute sense. As a result, *generative verifiers* (Mahan et al., 2024; Zhang et al., 2025a; Liu et al., 2025) have been

Table 1: Comparison between Hard2Verify and existing step-level math benchmarks.

	Question Difficulty	Open-Ended Responses?	Natural Responses?	Generator Strength	Annotator	Step-Level Labels?
MR-GSM8K (Zeng et al., 2023)	Easy	✗	✓	Weak	Human	✓
MR-MATH (Xia et al., 2025)	Easy	✗	✓	Weak	Human	✓
MR-Ben (Zeng et al., 2023)	Easy	✗	✓	Weak	Human	✓
ProcessBench (Zheng et al., 2024a)	Easy-Hard	10.3%	✓	Weak-Medium	Human	✗
PRMBench (Song et al., 2025)	Easy	✗	✗	Weak	Synth. + Human Check	✓
Hard2Verify (Ours)	Hard	78.5%	✓	Strong	Human	✓

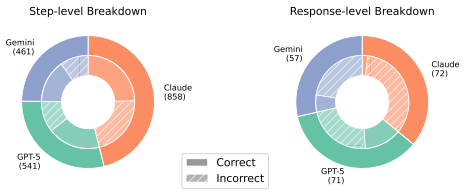

Figure 2: Breakdown of correct vs. incorrect steps (left) and responses (right) by model. We consider a response incorrect if *any* step in the response is labeled incorrect.

deployed. This allows for more precise description of evaluation criteria and increased inference-time compute. Generative approaches are either *reference-based* (Luong et al., 2025; Ma et al., 2025) or *reference-free* (Shao et al., 2025; Rahman et al., 2025; Xu et al., 2025). The former requires detailed grading rubrics; As a result, *Hard2Verify* operates in the latter, more scalable setting.

Benchmarking step-level math verifiers. Table 1 contrasts *Hard2Verify* with related benchmarks. MR-GSM8K (Zeng et al., 2023) annotate model responses to GSM8K (Cobbe et al., 2021) questions on a per-step basis to evaluate generative models as evaluators. MR-MATH (Xia et al., 2025) and MR-Ben (Zeng et al., 2024) are similar, using slightly harder sources like MATH (Hendrycks et al., 2021) and MMLU (Hendrycks et al., 2020). The two most relevant works to *Hard2Verify* are ProcessBench (Zheng et al., 2024a) and PRM-Bench (Song et al., 2025). ProcessBench uses a mix of easy (GSM8K and MATH) and hard (OlympiadBench and Omni-MATH) questions, but mostly contains samples that are not open-ended[1]. Further, ProcessBench only evaluates first error identification ability of verifiers. PRMBench obtains step-level annotations by taking fully correct *human-written* and model-generated solutions from the now easy PRM800K dataset and injecting errors with an LLM, yielding responses that are *not naturally occurring*: Human-and model-written text may have large differences in style and substance and injected errors may not represent naturally occurring errors in generation. Hard2Verify, in contrast, operates at the current frontier, tasking verifiers to evaluate responses from frontier-level LLMs to difficult, largely open-ended questions.

[1]The fraction of open-ended questions in ProcessBench in Table 1 is derived by counting the number questions from the Omni-MATH split that are not in the rule-based Omni-MATH. All other splits are not open-ended.

3 The Hard2Verify Benchmark

3.1 Design philosophy

Hard2Verify is designed to test verifiers at the frontier of LLM-based math reasoning. At the question, response, and annotation level, Hard2Verify is curated based on the following philosophy:

- **Questions.** To measure progress in step-level verification, we must characterize how verifiers perform on *extremely difficult, open-ended* math questions. Open-ended problems represent the next frontier of mathematical reasoning, one where verifiers become increasingly important in lieu of available ground-truth answers. We focus our data collection on very recent mathematical Olympiads, prioritizing open-ended questions.
- **Model responses.** The responses that verifiers evaluate must be from *highly capable, frontier-level models.* To push the frontier of math reasoning, verifiers must be able to tell when the most powerful models make potentially subtle mistakes. Moreover, such mistakes should be *naturally occurring*, i.e., arise from the model generation process. We do not inject or edit existing correct model-or human-written solutions. This is meant to closely approximate the response distribution that verifiers will see "in the wild", as they are applied in frontier math settings.
- **Annotation process.** We employ a *strict view* of response grading: Any step that contains a mistakes or is derived from a previous mistake is considered incorrect, i.e., we do not employ "Error Carried Forward" grading. This is inspired by competitive math settings, the entire solution must be correct to receive full points.

Based on this philosophy, we create Hard2Verify.

3.2 Curating hard questions

We construct our benchmark by collecting problem statements and official solutions (Q, A_{official}) from leading math competitions including the IMO, Putnam, and INMO; We provide a full list of sources in App. A. We focus question curation on recent (2024 and beyond) Olympiad-level math competitions. For each Olympiad, we parse the official PDFs using MathPix and extract all content in LaTeX to preserve mathematical typography and ensure stable equation rendering. We exclude image-dependent problems and only keep questions that could be solved using textual information. The resulting question set comprises 80 frontier-level problems from 10 distinct Olympiads.

3.3 Response generation

Using our curated question pool, we sample responses from three frontier LLMs: GPT-5 (with high reasoning), Gemini 2.5 Pro, and Claude Sonnet 4 (Thinking). We employ a standardized prompt (App. D), instructing models to produce exam-style, stepwise proofs that mirror how an Olympiad participant would structure a solution. We use the same prompt and decoding settings across models and disable access to external tools, like web search or code interpreters. Each model produces a single solution per problem, which we record for annotation. These samples are challenging; for example, Gemini 2.5 Pro takes up to 15 minutes to return a solution via API access. After curating all model responses to all questions, we filter out responses with undesirable qualities, such as a small number of long, dense steps or responses with degenerate outputs. This leaves us with a compact but high quality set of 200 responses.

3.4 Ensuring high-quality annotations

After sampling responses to our curated questions, human annotators meticulously annotate each model solution step-by-step. We partnered with Turing, a research accelerator. Turing employs mathematical experts, with a super-majority of our annotators having an advanced graduate level education in mathematics. To ensure consistent and high quality evaluations, we provided comprehensive annotation instructions as well official solutions A_{official} as references. Annotation began with a multi-round pilot study, where we hand-annotated three model responses, then worked together with annotators to review samples, solicited feedback from annotators, and finetuned evaluation instructions accordingly. We then performed annotations in batches of samples, performing spot-checks of samples as they became available. This is in addition to internal processes at Turing, which include initial human annotation and three rounds of human review, where annotations were reviewed for correctness and guideline alignment. Overall, *this process represents over 500 hours of manual human labor*. See App. E for more annotation details.

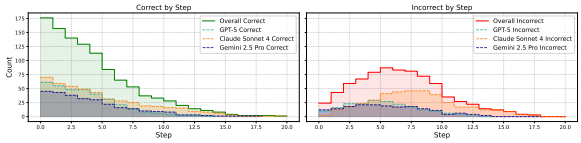

Figure 3: Count of correct (left) and incorrect (right) labels by model solution step. Models tend to begin solutions correctly, but get derailed after a few steps.

3.5 Overall dataset statistics.

Our annotation process yields 1,860 unique model steps annotated across 200 model solutions. 58% (1,080/1,860) steps are labeled correct, while the remaining 780 are labeled incorrect. Fig. 2 shows how models perform on a step-level and problem level. We consider a model response correct if *all* steps in the solution are graded correct by humans. Claude Sonnet 4 takes the most steps but gets the least percentage of steps correct, whereas GPT-5 and Gemini 2.5 Pro perform similarity in terms of step-level accuracy. However, at the response level, GPT-5 outperforms Gemini 2.5 Pro by larger margins. Claude Sonnet 4, while achieving over 50% step-level accuracy, fails to string correct steps together, only producing 4 entirely correct solutions out of 72. Fig. 3 visualizes how errors appear as a function of steps, with all three models following similar trends: Errors tend to occur in the middle of solutions, appearing after a few steps.

3.6 Evaluation tasks

Our step-level annotations allow us to construct three distinct tasks: (1) Step-level correctness (`Step-Level`), (2) Response-level correctness (`Response-Level`), and (3) First error identification (`ErrorID`). The `Step-Level` task corresponds to the setup in Song et al. (2025), whereas the `ErrorID` tasks corresponds to that of Zheng et al. (2024a). As we show in § 4, both tasks are challenging settings for current verifiers. We provide our evaluation prompts in App. D.

`Step-Level`. Here, the verifier is tasked with de-

termining the correctness of each step. Generative verifiers are prompted to output a binary yes/no label for each step, whereas PRM step-level scores are converted to binary labels via a fixed threshold.

`Response-Level`. Here, we derive an outcome-based task from `Step-Level` labels which reflects strict grading of open-ended math problems: For a question to be correct, *all* steps in the solution must be deemed correct. Therefore, if *any* step in the solution is incorrect, the solution is wrong[2]. From human labels, we create an overall response-level correctness label. Likewise, we create a response-level prediction from step-level verifier predictions. This task is more forgiving than `Step-Level`: Exact step labels need not match exactly for a verifier to agree with a human at the response level.

`ErrorID`. Here, the verifier is prompted to output the first step that contains a mistake, if present, or step -1, corresponding to "No error". For generative verifiers, the first error step can also be derived from `Step-Level` labels, similar to the `Response-Level` setting. Following ProcessBench, we prompt the verifier to output the step index directly, allowing us to more directly compare across benchmarks; we quantify differences between the direct prompting approach and converting from step-level labels in § 4.3. For PRMs, we select the first step below the correctness threshold.

4 Experiments

4.1 Evaluation Metrics

Let TPR and TNR denote the True Positive Rate and True Negative Rate, i.e., verifier accuracy on correct and incorrect samples, respectively. We define *Balanced Accuracy* as the mean and *Balanced F1 Score* as the harmonic mean of TPR and TNR[3]:

$$\text{Balanced F1 Score} = \frac{2 \text{ TPR} \cdot \text{TNR}}{\text{TPR} + \text{TNR}}, \quad (1)$$

We report Balanced Accuracy and Balanced F1 Score for all tasks. The ground-truth labels and model predictions vary based on task. For `Step-Level`, we aggregate all steps and all verifier predictions across all responses, whereas for `Response-Level` and `ErrorID`, we compute metrics at the response level. Balanced Accuracy and Balanced F1 both serve as aggregate measures: the former reflects average performance across both modes, while the latter penalizes imbalanced performance. An ideal verifier scores highly on both.

[2]Because we are concerned with ensuring completely correct responses, we apply this procedure to *all* responses/questions in *Hard2Verify*, including non-open-ended questions.

[3]This is equivalent to the ProcessBench "F1 Score", which differs from the typical F1 Score by using TNR instead of precision. To avoid confusion, we use "Balanced F1 Score".

4.2 Evaluated models

We select a variety of PRMs and generative models prompted as step-level critics. For prompted critics, we test a closed-source models as well as large ($\geq$ 70B) and small-medium ($<$70B) open-weight models. We evaluate all reasoning models at the maximum reasoning level (e.g., "high" for GPT-5), using suggested sampling parameters for various baselines. All Qwen3 models are evaluated with "thinking" on. For instruction-tuned models, we use greedy decoding. For all models, we set the maximum number of output tokens to be 32K. The full set of models is enumerated in App. F. For PRMs, we select Qwen2.5-Math-PRM (Zhang et al., 2025b), Skywork-PRM (He et al., 2024b), ReasonFlux-PRM (Zou et al., 2025), and Universal-PRM (Tan et al., 2025). We tune PRM thresholds following Zheng et al. (2024a); See App. F.1.

4.3 Main evaluation results

Table 2 presents our main results, with detailed results presented in App. C. Among proprietary models, GPT-5 stands out in its overall ability across all three tasks. Gemini 2.5 Pro follows closely for step-level identification, but lags in error identification. Finally, Claude Sonnet 4 with Thinking lags OpenAI models and Gemini 2.5 Pro, failing to match reasoning models from previous generations, like o3 and o4-mini. Among larger open-weight models, the gpt-oss series are clear standouts, with gpt-oss-120B roughly matching the performance of o3. Larger Qwen3 models and DeepSeek-R1 challenge for second place on `Step-Level`, but all lag on `ErrorID`. Notably, Llama-3.3-70B, which achieves 58.0 on ProcessBench (Fig. 1) achieves only 2.50 on `ErrorID`. Among smaller models, gpt-oss-20B performs extremely well on step-level and response-level tasks, but falters in identifying errors. ByteDance Seed-OSS-36B and Qwen3-30B-A3B are the next best performers, but only ByteDance Seed-OSS-36B is able to outperform random guessing performance on `ErrorID`. Finally, even strong PRMs perform below random guess performance on `ErrorID`. For example, Qwen2.5-Math-PRM-72B achieves only 37.28 Balanced F1.

Table 2: Main evaluation results on Hard2Verify across our three evaluation tasks (§ 3.6). We report Balanced Accuracy and Balanced F1 Score. **Best** and second-best scores in each category marked.

	Step-Level		Response-Level		ErrorID	
	Bal. Accuracy	Bal. F1	Bal. Accuracy	Bal. F1	Bal. Accuracy	Bal. F1
***Generative Critics**, proprietary models*						
GPT-5	**86.53**	**85.83**	**89.69**	**89.52**	**70.61**	**69.72**
Gemini 2.5 Pro	83.37	83.09	85.73	85.46	52.46	52.46
Claude Sonnet 4	70.61	60.37	78.24	73.44	53.45	39.30
GPT-5-Mini	81.06	78.73	81.93	81.92	65.96	60.04
o3	78.70	75.29	83.21	82.58	60.32	57.31
o4-Mini	74.90	68.09	83.94	81.71	67.31	57.62
GPT-4.1	56.17	24.66	58.94	33.55	52.44	21.29
***Generative Critics**, large ($\geq$ 70B) models*						
Kimi K2	61.79	42.83	65.34	51.66	49.10	31.40
DeepSeek-R1	68.92	62.30	73.95	72.75	54.23	45.35
Qwen3-235B-A22B	72.55	64.03	79.42	77.87	60.90	50.78
Qwen3-Next-80B-A3B	67.91	54.69	75.08	68.31	58.29	43.05
Qwen2.5-72B-Instruct	56.01	26.36	61.06	46.89	26.49	16.38
GLM-4.5-Air	57.40	29.40	61.78	41.00	41.97	17.81
gpt-oss-120B	**78.10**	**74.64**	**83.92**	**83.71**	**63.97**	**60.64**
Llama-3.3-70B-Instruct	54.28	18.37	57.04	28.16	49.44	2.50
***Generative Critics**, small/medium ($<$ 70B) models*						
Qwen3-32B	63.99	51.77	67.86	63.16	51.96	26.83
Qwen3-30B-A3B	70.71	61.91	73.79	71.02	58.83	**50.51**
ByteDance Seed-OSS-36B	66.79	53.09	72.54	63.88	**59.24**	45.18
gpt-oss-20B	**75.18**	**70.93**	**83.85**	**83.32**	46.13	45.28
Qwen3-14B	65.48	52.91	74.59	70.12	53.69	37.33
Qwen3-8B	65.26	53.51	77.61	72.45	45.92	34.26
Qwen2.5-14B-Instruct	60.45	47.59	63.40	63.23	43.47	18.86
Qwen2.5-7B-Instruct	48.82	22.84	55.67	44.18	29.75	15.96
***Process Reward Models**, open-source models*						
Qwen2.5-Math-PRM-72B	55.82	35.50	**66.80**	**64.91**	41.80	**37.28**
Qwen2.5-Math-PRM-7B	57.56	42.37	63.08	57.57	35.03	32.50
Skywork-PRM-7B	38.52	34.12	56.77	29.81	11.56	8.36
Skywork-PRM-1.5B	40.81	12.94	52.46	20.89	8.62	7.48
ReasonFlux-PRM-7B	53.09	22.40	55.89	53.82	**42.48**	28.71
UniversalPRM-7B	**64.17**	**60.27**	54.74	41.46	26.08	25.97

Table 3: `ErrorID` performance using two prompting approaches, with Δ = `Step-Level` − `ErrorID`. The `ErrorID` prompt tasks verifier to directly identify the first step with an error, as in ProcessBench (Zheng et al., 2024a). The `Step-Level` prompt tasks the verifier to produce step-level labels, from which the first error step is derived. Balanced Accuracy tends to improve with the `Step-Level` prompt, but Balanced F1 changes are mixed.

	`ErrorID` Prompt Bal. Accuracy	`Step-Level` Prompt Bal. Accuracy	$\Delta_{\text{Bal. Acc}}$	`ErrorID` Prompt Bal. F1	`Step-Level` Prompt Bal. F1	$\Delta_{\text{Bal. F1}}$
GPT-5	70.61	76.72	+6.11	69.72	75.66	+5.94
gpt-oss-120B	63.97	69.68	+5.71	60.64	64.81	+4.17
GPT-5-Mini	65.96	66.43	+0.47	60.04	63.25	+3.21
o4-Mini	67.31	67.16	-0.15	57.62	53.35	-4.27
o3	60.32	68.02	+7.70	57.31	60.61	+3.30
Gemini 2.5 Pro	52.46	66.11	+13.65	52.46	62.78	+10.32
Qwen3-235B-A22B	60.90	65.17	+4.27	50.78	55.35	+4.57
Qwen3-30B-A3B	58.83	60.19	+1.36	50.51	47.25	-3.26
DeepSeek-R1	54.23	61.53	+7.30	45.35	52.02	+6.67
gpt-oss-20B	46.13	66.44	+20.31	45.28	57.75	+12.47
ByteDance Seed-OSS-36B	59.24	58.94	-0.30	45.18	33.55	-11.63
Qwen3-Next-80B-A3B	58.29	63.37	+5.08	43.05	44.85	+1.80
Claude Sonnet 4	53.45	60.83	+7.38	39.30	38.59	-0.71
Qwen3-14B	53.69	56.56	+2.87	37.33	33.25	-4.08
Qwen3-8B	45.92	57.35	+11.43	34.26	29.09	-5.17
Kimi K2	49.10	54.26	+5.16	31.40	23.33	-8.07
Qwen3-32B	51.96	52.35	+0.39	26.83	31.09	+4.26
GPT-4.1	52.44	51.97	-0.47	21.29	11.89	-9.40
Qwen2.5-14B-Instruct	43.47	40.30	-3.17	18.86	23.04	+4.18
GLM-4.5-Air	41.97	53.24	+11.27	17.81	16.25	-1.56
Qwen2.5-72B-Instruct	26.49	48.09	+21.60	16.38	10.72	-5.66
Qwen2.5-7B-Instruct	29.75	43.01	+13.26	15.96	9.53	-6.43
Llama-3.3-70B-Instruct	49.44	50.71	+1.27	2.50	7.31	+4.81

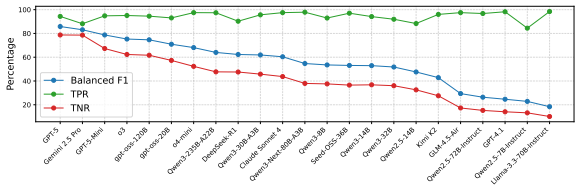

Figure 4: Weaker models are unable to find mistakes, eventually considering *all* steps correct: TNR tends toward 0 while TPR tends towards 1.

What separates strong and weak verifiers? To provide additional insights into variations across different verifiers, Fig. 4 plots the TPR and TNR for all generative critics models, sorted in performance from strongest (left) to weakest (right) in terms of Balanced F1 Score. A clear trend emerges: Verifier performance degrades because TNR drops quickly to near 0, while TPR rises gradually to almost 1. This indicates that almost all steps are labeled as correct, revealing that *weaker verifiers cannot catch errors*. Notably, the order of models from left to right approximately correlates with mathematical *generation* ability, i.e., the ability to solve extremely difficult math problems. As such, this may indicate that a baseline level of solving ability is a necessary prerequisite for verification. App. C shows this trend holds similarly for other tasks.

To identify errors, how should verifiers be prompted? Our `ErrorID` task adopts the setup of ProcessBench (Zheng et al., 2024a), which prompts the verifier to output the index of the first step with an error. However, the first error index can also be derived from step-level labels, like those produced in the `Step-Level` task. In Table 3, we compare the performance under the `ErrorID` Prompt and `Step-Level` Prompt. Surprisingly, directly prompting for the given task may not yield the best performance: In terms of Balanced F1, performance across models is mixed, with some models exhibiting very small performance changes and others exhibiting significant changes. For example, step-level prompting significantly degrades performance for ByteDance Seed-OSS-36B from 45.18 to 33.55, while boosting performance for Gemini 2.5 Pro from 52.46 to 62.78. Overall, we find that more capable models, like GPT-5 and Gemini 2.5 Pro, benefit the most from switching to deriving first identified error from `Step-Level` outputs. We hypothesize that requiring step-by-step annotations requires models inspect each step more carefully, allowing for better error identification. Models capable of performing step-level verification tend to benefit the most, while insufficiently capable models fare worse. As the change in performance is mixed across models, we advise practitioners optimize prompts on a per-verifier basis.

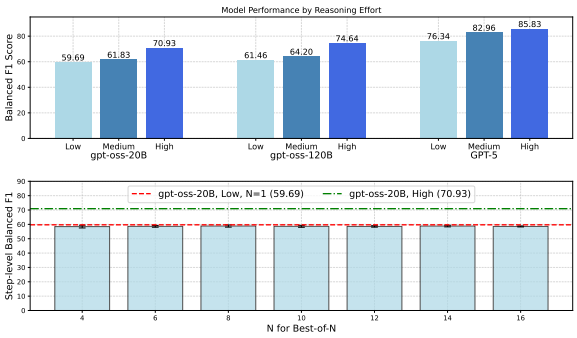

Figure 5: **Top**: Scaling inference-time compute sequentially leads to higher performance in GPT-5 and gpt-oss models **Bottom**: Parallel decoding has little effect on step-level F1 performance for gpt-oss-20B, failing to bridge the gap vs. gpt-oss-20B at high-reasoning effort.

5 Additional Analysis

5.1 How should we scale verifier inference-time compute?

Here, we scale verifier inference-time compute sequentially and in parallel. We find sequential scaling brings substantive gains unlike parallel scaling.

Sequential inference-time compute scaling. Here we explore scaling inference-time compute sequentially by letting the verifier generate more output tokens, focusing on the `Step-Level` task. We use gpt-oss-20B, gpt-oss-120B, and GPT-5, which all have low, medium, and high reasoning levels. In Fig. 5 (top), we plot Balanced F1. Affording the verifier to generate more "thinking" tokens at inference time generally improves performance, with gpt-oss-120B improving the most from low (61.46) to high (74.64) and gpt-oss-20B likewise improving significantly. Gains for GPT-5 are smaller compared to gpt-oss models, but still significant, with 12.3% relative improvement from low to high.

Parallel inference-time compute scaling. Here, we attempt to match the performance of gpt-oss-20B at high reasoning effort by sampling parallel outputs from gpt-oss-20B at low reasoning effort. We sample 32 responses per sample from gpt-oss-20B and simulate best of N from $N = 4, \ldots, 16$ via bootstrap sampling. Concretely, for each N, we sample N responses from the 32 without replacement, and aggregate predicted step-level labels via majority vote, breaking ties arbitrarily. To reduce variance, we repeat this process for 10 trials for each N, and report mean and standard deviation

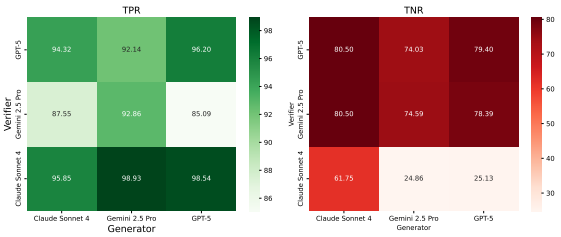

Figure 6: Verifier TPR and TNR based on generator model. For strong verifiers (GPT-5, Gemini 2.5 Pro), TPR varies based on generator, with GPT-5 being the most stable. Claude Sonnet 4 generates the easiest to catch mistakes, whereas Gemini 2.5 Pro produces the hardest to catch mistakes, as measured by TNR.

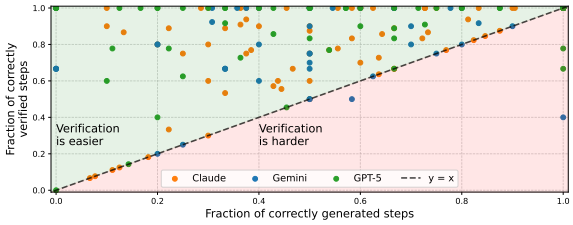

Figure 7: Each generator evaluates self-produced responses, and the fraction of steps correctly solved vs. fraction of steps correctly verified for a given question is plotted. In general, models are more successful in catching mistakes than generating error-free responses.

across trials Fig. 5 (bottom). We also plot the baseline gpt-oss-20B performance at low and high reasoning efforts. Surprisingly, Best-of-N does not meaningfully improve over sampling 1 response as N increases. An intuitive explanation for this phenomenon is that step-level verification is inherently a sequential task: Each step must be processed one-after-another. As such, affording the verifier more time to "think" about each step is more effective than sampling multiple "rushed" judgments.

5.2 How do verifiers verify their own responses?

We investigate the dynamics of self-verification, focusing on GPT-5, Gemini 2.5 Pro, and Claude Sonnet 4 as verifiers. Fig. 6 plots the step-level TPR and TNR performance based on response generator. The results notably depend on verifier strength: Table 2 shows that GPT-5 and Gemini 2.5 Pro are the top two performers, whereas Claude Sonnet 4 is a relatively weak proprietary verifier. We find that GPT-5 and Gemini 2.5 Pro as verifiers are more likely to consider a correct self-generated response as correct, as measured by TPR. Of the two, GPT-5 exhibits the least variation in TPR across models, while Gemini 2.5 Pro performance drops from 92.86 TPR on own-generated responses to as low as 85.09 TPR for GPT-5-generated responses. Claude Sonnet 4, on the other hand, overwhelmingly assigns "Correct" as a label, leading to high TPRs regardless of generator. Across all threee models, it is easier to identify errors from the weaker model (Claude Sonnet 4) than it is to identify errors from the stronger models. This result is consistent with recent work (Zhou et al., 2025a) studying verification, which finds weaker generators produce easier to catch errors. Interestingly, both GPT-5 and Gemini 2.5 Pro struggle have the lowest TNR on responses from Gemini 2.5 Pro, showing that GPT-5 is more reliable in self-critique than Gemini 2.5 Pro is. The fact that Gemini 2.5 Pro has the lowest TNR on self-generated responses is consistent with recent work analyzing self-reflection (Stechly et al., 2023, 2024; Huang et al., 2023), where LLMs were shown to have difficulties correcting their own mistakes in challenging reasoning settings. In contrast, Claude Sonnet 4 as a relatively weaker verifier cannot identify errors in stronger model responses.

5.3 Is verifying easier than solving?

Here, we examine if generating a solution is easier than verifying the same solution. We split Hard2Verify into three subsets corresponding to each of the three generator models and have the generators verify their own responses. For each response, we record the fraction of correctly generated steps ("solve rate"), as deemed by human annotators, and the fraction of correctly verified steps ("verification rate"), as deemed by agreement with human annotators. In Fig. 7, we plot the verification rate against the solve rate. We observe that the verification rate is consistently higher than the solve rate across all models; Only on a few problems does the verifier have a more difficult time verifying a problem than generating the problem. This result offers some optimism for future work in verification: Because verifying a solution tends to be "easier" than generating the solution, verifiers may not necessarily need to be as powerful as frontier generators to reliably identify errors.

6 Conclusion

We introduce *Hard2Verify*, a human-annotated, step-level benchmark aimed to assess how step-level verifiers operate in frontier settings. We focus our data curation on recent open-ended math problems, sampling responses from frontier LLMs. The

end result of over 500 hours of human annotation effort is a benchmark that challenges many current open-source verifiers, which are unable to match the performance of larger, proprietary models.

7 Limitation

Our work has some limitations. First, Hard2Verify is modest in scale (200 model-generated solutions). This is a consequence of careful filtering and the use of a strict cutoff date during data collection, which may lead to some variance in measured performance. Second, the benchmark is currently limited to English, Olympiad-style problems and text-only inputs. In future work, we plan to expand the dataset and analysis to a larger scale, include problems in additional languages, and incorporate diagram-based reasoning.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Anthropic. 2025. Introducing claude 4.

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

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. How to evaluate reward models for rlhf. *arXiv preprint arXiv:2410.14872*.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, and 1 others. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, and 1 others. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*.

Google. 2025a. Advanced version of gemini with deep think officially achieves gold-medal standard at the international mathematical olympiad. *https://deepmind.google/discover/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/*.

Google. 2025b. Gemini 2.5: Our most intelligent ai model. *https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024a. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Jujie He, Tianwen Wei, Rui Yan, Jiacai Liu, Chaojie Wang, Yimeng Gan, Shiwen Tu, Chris Yuhao Liu, Liang Zeng, Xiaokun Wang, Boyang Wang, Yongcong Li, Fuxiang Zhang, Jiacheng Xu, Bo An, Yang Liu, and Yahui Zhou. 2024b. Skywork-o1 open series.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

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

Yichen Huang and Lin F Yang. 2025. Gemini 2.5 pro capable of winning gold at imo 2025. *arXiv preprint arXiv:2507.15855*.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, and 1 others. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Shalev Lifshitz, Sheila A McIlraith, and Yilun Du. 2025. Multi-agent verification: Scaling test-time compute with multiple verifiers. *arXiv preprint arXiv:2502.20379*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and 1 others. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*.

Minh-Thang Luong, Dawsen Hwang, Hoang H Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, and 1 others. 2025. Towards robust mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35406–35430.

Wenjie Ma, Andrei Cojocaru, Neel Kolhe, Bradley Louie, Robin Said Sharif, Haihan Zhang, Vincent Zhuang, Matei Zaharia, and Sewon Min. 2025. Reliable fine-grained evaluation of natural language math proofs. *arXiv preprint arXiv:2510.13888*.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative reward models. *arXiv preprint arXiv:2410.12832*.

MMA. 2025. (american invitational mathematics examination). *https://maa.org*.

OpenAI. 2025. Gpt-5 system card.

OpenAI. 2025. Introducing GPT-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2025-09-25.

OpenAI. 2025a. Openai imo 2025 proofs. *https://github.com/aw31/openai-imo-2025-proofs*.

OpenAI. 2025b. Openai o3 and o4-mini system card.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Salman Rahman, Sruthi Gorantla, Arpit Gupta, Swastik Roy, Nanyun Peng, and Yang Liu. 2025. Spark: Stepwise process-aware rewards for reference-free reinforcement learning. *arXiv preprint arXiv:2512.03244*.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2025. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *The Thirteenth International Conference on Learning Representations*.

Zhihong Shao, Yuxiang Luo, Chengda Lu, ZZ Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, and Xiaokang Zhang. 2025. Deepseekmath-v2: Towards self-verifiable mathematical reasoning. *arXiv preprint arXiv:2511.22570*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. *arXiv preprint arXiv:2310.12397*.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2024. On the self-verification limitations of large language models on reasoning and planning tasks. *arXiv preprint arXiv:2402.08115*.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Xiaoyu Tan, Tianchu Yao, Chao Qu, Bin Li, Minghao Yang, Dakuan Lu, Haozhe Wang, Xihe Qiu, Wei Chu, Yinghui Xu, and 1 others. 2025. Aurora: Automated

training framework of universal process reward models via ensemble prompting and reverse verification. *arXiv preprint arXiv:2502.11520*.

ByteDance Seed Team. 2025. Seed-oss open-source models. https://github.com/ByteDance-Seed/seed-oss.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2023. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27723–27730.

Austin Xu, Xuan-Phi Nguyen, Yilun Zhou, Chien-Sheng Wu, Caiming Xiong, and Shafiq Joty. 2025. Foundational automatic evaluators: Scaling multi-task generative evaluator training for reasoning-centric domains. *arXiv preprint arXiv:2510.17793*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Fei Yu, Yingru Li, and Benyou Wang. 2025. Scaling flaws of verifier-guided search in mathematical reasoning. *arXiv preprint arXiv:2502.00271*.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. 2023. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. *arXiv preprint arXiv:2312.17080*.

Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, and 1 others. 2024. Mr-ben: A comprehensive meta-reasoning benchmark for large language models. *arXiv e-prints*, pages arXiv–2406.

Kaiwen Zha, Zhengqi Gao, Maohao Shen, Zhang-Wei Hong, Duane S Boning, and Dina Katabi. 2025. Rl tango: Reinforcing generator and verifier together for language reasoning. *arXiv preprint arXiv:2505.15034*.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025a. Generative verifiers: Reward modeling as next-token prediction. In *The Thirteenth International Conference on Learning Representations*.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025b. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024a. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Livia Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, and 1 others. 2024b. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583.

Yefan Zhou, Austin Xu, Yilun Zhou, Janvijay Singh, Jiang Gui, and Shafiq Joty. 2025a. Variation in verification: Understanding verification dynamics in large language models. *arXiv preprint arXiv:2509.17995*.

Yilun Zhou, Austin Xu, Peifeng Wang, Caiming Xiong, and Shafiq Joty. 2025b. Evaluating judges as evaluators: The jetts benchmark of llm-as-judges as test-time scaling evaluators. *arXiv preprint arXiv:2504.15253*.

Jiaru Zou, Ling Yang, Jingwen Gu, Jiahao Qiu, Ke Shen, Jingrui He, and Mengdi Wang. 2025. Reasonflux-prm: Trajectory-aware prms for long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2506.18896*.

A Detailed Dataset Sources

In table App. A we provide the distribution of the 80 problems we sourced from different Olympiads along with the date the Olympiads were conducted. For the IMO-shortlist, we report the earliest date that the shortlist questions were made publicly available, typically the calendar year *after* the Olympiad was conducted.

Contest	Date of Olympiad	# Questions
IMO - Shortlist 2023	21 July 2024	10
IMO - Shortlist 2024	23 July 2025	29
Putnam	7 Dec 2024	12
EGMO (European Girls' Mathematical Olympiad)	17 April 2025	6
IMO (International Mathematical Olympiad)	20 July 2025	6
BMO (British Mathematical Olympiad)	22 Jan 2025	4
CMO (Canadian Mathematical Olympiad)	6 March 2025	4
USA-JMO (Junior Mathematical Olympiad)	20 March 2025	4
INMO (Indian National Mathematical Olympiad)	19 Jan 2025	3
USAMO (United States of America Mathematical Olympiad)	20 March 2025	2
Total		**80**

Table 4: Distribution of questions from various Olympiads with Year-wise Splits

B Case study: Where do models and humans disagree?

We inspect outputs from a relatively strong open-source verifier, ByteDance Seed-OSS-36B (Team, 2025) on multiple IMO-level problems and found a recurring theme: The verifier incorrectly accepts *partial or under-justified claims as correct*. We provide two concrete examples below. These mismatches reflect larger *systematic* behavior in verifiers, revealed in § 4: Current verifiers are too *generous*, with TPR rate tending towards 1 and TNR tending toward 0, indicating that a vast majority steps are considered correct.

On IMO 2023 Shortlist, question A6, Gemini 2.5 Pro makes a generalized claim, but only proves the claim for a single input. Human annotators catch this mistake, noting "*The equality holds only at one point ... not a polynomial identity, so coefficients need not match.*" Seed-OSS-36B considers this step correct without mentioning the unfounded generalization. Similarly, on IMO 2024 Shortlist, question A1, Claude Sonnet 4 as generator constructs a proof by cases by invoking Weyl's equidistribution theorem, but considers only a single case: "*if α is not an even integer, then $\alpha = m + \beta$ with m odd and $2/3 \leq \beta < 1$...*". Seed-OSS-36B greenlights this step as correct, whereas human annotators find it incomplete: "*The case analysis ignores the branch where m is even and $0 < \beta < 1/3$,...*". Further, the theorem invocation itself is deemed under-specified: "*justification [for invoking Weyl's equidistribution theorem] should explicitly specify the estimate and the choice of n*".

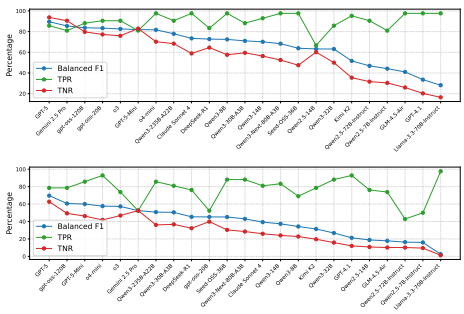

Figure 8: `Response-Level` and `ErrorID` tasks follow similar trends in TPR and TNR, with weaker verifiers unable to identify mistakes.

C Additional Experimental Results

We report TPR and TNR for all evaluated models in Table 5, alongside our aggregate metrics presented in § 4. We also visualize TPR and TNR trends for the `Response-Level` and `Step-Level` tasks, similar to Fig. 4. As shown in Fig. 8, TNR is the primary driver in poor Balanced F1 performance: The weaker the verifier, the more it struggles in identifying mistakes, opting to mark nearly every step as correct.

D Prompts for Generation and Evaluation

In this section we provide prompts used for generating responses to Olympiad-level math questions. We also provide the prompts used for the `Step-Level` and `ErrorID` tasks.

Table 5: Complete metrics for our three evaluation tasks, reporting Balanced Accuracy, Balanced F1, TPR, and TNR.

	Step-Level				Response-Level				ErrorID			
	TPR	TNR	Bal. Accuracy	Bal. F1	TPR	TNR	Bal. Accuracy	Bal. F1	TPR	TNR	Bal. Accuracy	Bal. F1
***Generative Critics**, proprietary models*												
GPT-5	94.35	78.72	86.53	85.83	85.71	93.67	89.69	89.52	78.57	62.66	70.61	69.72
Gemini 2.5 Pro	88.15	78.59	83.37	83.09	80.95	90.51	85.73	85.46	52.38	52.53	52.46	52.46
Claude Sonnet 4	97.50	43.72	70.61	60.37	97.62	58.86	78.24	73.44	80.95	25.95	53.45	39.30
GPT-5-Mini	94.81	67.31	81.06	78.73	80.95	82.91	81.93	81.92	85.71	46.20	65.96	60.04
o3	95.09	62.31	78.70	75.29	90.48	75.95	83.21	82.58	73.81	46.84	60.32	57.31
o4-Mini	97.50	52.31	74.90	68.09	97.62	70.25	83.94	81.71	92.86	41.77	67.31	57.62
GPT-4.1	98.24	14.10	56.17	24.66	97.62	20.25	58.94	33.55	92.86	12.03	52.44	21.29
***Generative Critics**, large ($\geq$ 70B) models*												
Kimi K2	96.02	27.56	61.79	42.83	95.24	35.44	65.34	51.66	78.57	19.62	49.10	31.40
DeepSeek-R1	90.28	47.56	68.92	62.30	83.33	64.56	73.95	72.75	76.19	32.28	54.23	45.35
Qwen3-235B-A22B	97.41	47.69	72.55	64.03	90.48	68.35	79.42	77.87	85.71	36.08	60.90	50.78
Qwen3-Next-80B-A3B	97.87	37.95	67.91	54.69	97.62	52.53	75.08	68.31	88.10	28.48	58.29	43.05
Qwen2.5-72B-Instruct	96.76	15.26	56.01	26.36	90.48	31.65	61.06	46.89	42.86	10.13	26.49	16.38
GLM-4.5-Air	97.50	17.31	57.40	29.40	97.62	25.95	61.78	41.00	73.81	10.13	41.97	17.81
gpt-oss-120B	94.54	61.67	78.10	74.64	88.10	79.75	83.92	83.71	78.57	49.37	63.97	60.64
Llama-3.3-70B-Instruct	98.43	10.13	54.28	18.37	97.62	16.46	57.04	28.16	97.62	1.27	49.44	2.50
***Generative Critics**, small/medium (< 70B) models*												
Qwen3-32B	91.94	36.03	63.99	51.77	85.71	50.00	67.86	63.16	88.10	15.82	51.96	26.83
Qwen3-30B-A3B	95.65	45.77	70.71	61.91	88.10	59.49	73.79	71.02	80.95	36.71	58.83	50.51
ByteDance Seed-OSS-36B	97.04	36.54	66.79	53.09	97.62	47.47	72.54	63.88	88.10	30.38	59.24	45.18
gpt-oss-20B	93.06	57.31	75.18	70.93	90.48	77.22	83.85	83.32	52.38	39.87	46.13	45.28
Qwen3-14B	94.17	36.79	65.48	52.91	92.86	56.33	74.59	70.12	83.33	24.05	53.69	37.33
Qwen3-8B	92.96	37.56	65.26	53.51	97.62	57.59	77.61	72.45	69.05	22.78	45.92	34.26
Qwen2.5-14B-Instruct	88.33	32.56	60.45	47.59	66.67	60.13	63.40	63.23	76.19	10.76	43.47	18.86
Qwen2.5-7B-Instruct	84.44	13.21	48.82	22.84	80.95	30.38	55.67	44.18	50.00	9.49	29.75	15.96
***Process Reward Models**, open-source models*												
Qwen2.5-Math-PRM-72B	89.50	22.14	55.82	35.50	55.56	78.05	66.80	64.91	55.56	28.05	41.80	37.28
Qwen2.5-Math-PRM-7B	87.13	27.99	57.56	42.37	44.44	81.71	63.08	57.57	44.44	25.61	35.03	32.50
Skywork-PRM-7B	51.55	25.50	38.52	34.12	17.65	95.89	56.77	29.81	17.65	5.48	11.56	8.36
Skywork-PRM-1.5B	74.53	7.08	40.81	12.94	11.76	93.15	52.46	20.89	11.76	5.48	8.62	7.48
ReasonFlux-PRM-7B	93.47	12.72	53.09	22.40	66.67	45.12	55.89	53.82	66.67	18.29	42.48	28.71
UniversalPRM-7B	80.00	48.35	64.17	60.27	27.78	81.71	54.74	41.46	27.78	24.39	26.08	25.97

Prompt used to generate responses to Olympiad questions

You are a careful, rigorous math proof assistant. Provide correct, detailed, and complete proofs.

Solve the following math problem formally. Return a detailed and formal solution that can be verified by a grader.

Use start the proof with <start> followed by each step with <step>...</step>, and end with <end>.

Only return the solution, in detailed steps, no headers, no explanations, no other text, only the <start> <step>...</step> <step>...</step> ... <end> tags.

Prompt used for the Step-Level task

The following is a math problem and a solution (split into steps, enclosed with tags and indexed from 0):

[Math Problem]

{problem}

[Solution]

{steps}

Your task is to review and critique the solution step-by-step.

For each step, determine if it is correct or incorrect. A correct step is one where all of the content is correct, and is logically consistent with all previous steps and information given in the problem.

An incorrect step is one where the content is incorrect, or is not logically consistent with all previous steps and information given in the problem, or is based on an error in a previous step.

Important: Any step that contains or is based on an error is considered incorrect. That is, if the error is carried forward from a previous step or is based on an error in the previous step, consider the step incorrect.

Provide reasoning for your correctness determinations. Your final verdict should be a comma-separated list of yes and no's, where each yes or no corresponds to a step's correctness, with yes meaning correct and no meaning incorrect.

Please use the following format to return your answer:

Reasoning: <your reasoning for each step>

Verdict: <your comma-separated list of yes and no's>

Do not use any other formatting, including markdown, bold text, code blocks, or any other formatting. If your formatting is incorrect, your evaluation will be affected.

Prompt used for the `ErrorID` task

The following is a math problem and a solution (split into steps, enclosed with tags and indexed from 0):

[Math Problem]

{problem}

[Solution]

{steps}

Your task is to identify the first incorrect step in the solution.

Instructions:

- Review each step carefully for mathematical correctness and logical consistency
- A step is incorrect if it contains mathematical errors, logical inconsistencies, or is based on errors from previous steps
- Find the FIRST step that is incorrect (0-indexed)
- If ALL steps are correct, return -1

Provide your reasoning and then give your final answer as a single number in the specified format.

Please use the following format to return your answer:

Reasoning: <your detailed reasoning explaining which steps are correct/incorrect and why>

Verdict: <the step number of the first incorrect step or -1 if all steps are correct>

Examples:

- If step 0 is the first incorrect step: 0
- If step 3 is the first incorrect step: 3
- If all steps are correct: -1

Do not use any other formatting, including markdown, bold text, code blocks, or any other formatting. If your formatting is incorrect, your evaluation will be affected.

E Annotation details

Each sample was annotated over four rounds: An initial annotation round and three rounds of reviews to resolve disagreements. A total of 52 annotators were employed for grading, with 35 having at least a graduate degree in mathematics or related fields. On average, a model response took 90 minutes to grade and 63 minutes to review, with the longest response taking up to 4 hours. Annotators were given access to external tools, such as the internet, python, Wolfram Mathematica, and LLMs strictly as assistive aids.

We present the detailed annotation guidelines provided to the math experts for step-by-step evaluation of each model solution below.

Annotation instructions to human annotators

When annotating, refer to the reference answer(s) as possible solution(s)/proof(s). Each question may have multiple valid approaches, as these are open-ended questions. The provided reference answer(s) is an example of a valid approach; it may not be the only such valid approach.

Base your correctness decision off of the following criteria:

Correct: A step is considered correct if it is:

Computationally valid: There are no mistakes in rote mathematical operations, such as addition or computing values of known functions (e.g., sin(pi/2))

Logically valid: The step follows logically from previous steps and information present in the original question. There are no intermediate mistakes in the reasoning. Any and all conclusions in the step must be logically deducible from previous correct steps.

If a step invokes any third-party mathematical results, such as known theorems / lemmas (e.g., fundamental theorem of calculus) or intermediate results from previous steps, then annotators must verify that the result is used in a valid way:

(1) all assumptions of the result (theorem) are met

(2) the consequence of the result (theorem) is correctly described and applied to the specific problem

Important: Do not apply "Error carried forward" grading.

If a current step is derived from a previous step that is incorrect, consider the current step incorrect, even if the logic/computation of the step is correct.

Example:

Step 1: 1 + 1 = 3 [Incorrect]

Step 2: We now must add 5 to Step 1's result, which gives us 8 [Incorrect, even though the computation in the step is correct; It is based on an incorrect Step 1]

Extra note:

"Hand-waviness": If a model produces a "hand-wavy" argument, wherein they say that a new result follows by similar logic/computation as a previously established result, then annotators must verify that the hand-wavy argument in-fact holds. This means verifying (1) The previously established result's assumptions are met by the new result scenario (2) The previously established computation/logic is applicable to the new

Example:

Step N: A valid proof of Case 1, yielding Result 1

Step N+1: Case 2 follows by a similar argument to Case 1, yielding Result 2.

[This is "hand-wavy", as the exact computation is omitted by appealing to previously computed Steps]

Incorrect: A step is considered incorrect if it is:

Based in any way on an incorrect past step.

Logically invalid: The model's output contains a reasoning error or mistake. Examples: Unfounded logical leap

Incorrectly invoking a mathematical result or past result when assumptions/conditions are not satisfied

Incorrect application of a mathematical result when conditions are met, i.e., mis-applying a theorem.

Failing to consider/cover a scenario or case within a proof, i.e., the proof concludes without covering all scenarios and is incomplete.

If the top-level proof misses a case/scenario: As this case involves text not in the model output, there is no concrete step to mark as incorrect. As a result, mark the conclusion of the proof (i.e., last step) as incorrect and provide corresponding justification.

If an intermediate result is stated, but the derivation of the intermediate result misses a case/scenario: Mark the step that states the intermediate result as incorrect (as well as any subsequent steps that depend on the intermediate result). As a concrete toy example Say a model is doing Proof by Cases for all real numbers.

It splits its analysis into 2 cases, Case 1 (positives) and Case 2 (negatives). For Case 1, it proves the claim for all positive integers, but does not consider non-integer reals.

Mark the step that contains the conclusion of Case 1 incorrect, as well as any subsequent steps that depend on Case 1.

Computationally invalid: Makes an operation / value computation mistake. This should be relatively easy to spot, but please verify all complex expressions, such as integrals, trigonometric functions, etc.

Note: This is not an exhaustive list of errors. Verify all computations, and document any error that occurs, no matter how minor.

F Evaluated baselines

Here we provide a comprehensive list of models that were evaluated on our benchmark.

- OpenAI: GPT-5, GPT-5-Mini (OpenAI, 2025), o3, o4-Mini (OpenAI, 2025b), GPT-4.1 (OpenAI, 2025), gpt-oss-120b, gpt-oss-20b (Agarwal et al., 2025)
- Google: Gemini 2.5 Pro (Google, 2025b)
- Anthropic: Claude Sonnet 4 (Anthropic, 2025)
- Moonshot (Kimi): Kimi-K2-Instruct-0905 (Team et al., 2025)
- DeepSeek: DeepSeek-R1 (Guo et al., 2025)
- Alibaba Qwen: Qwen3-235-A22B, Qwen3-Next-80B-A3B, Qwen3-32B, Qwen3-30B-A3B, Qwen3-14B, Qwen3-8B (Yang et al., 2025), Qwen2.5-72B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-7B-Instruct (Team, 2024), Qwen2.5-Math-PRM-72B, Qwen2.5-Math-PRM-7B (Zhang et al., 2025b)
- Zhipu GLM: GLM-4.5-Air (Zeng et al., 2025)
- Meta: Llama-3.3-70B-Instruct (Grattafiori et al., 2024)
- ByteDance: ByteDance Seed-OSS-36B (Team, 2025)
- Skywork: Skywork-PRM-7B, Skywork-PRM-1.5B (He et al., 2024b)
- ReasonFlux-PRM-7B (Zou et al., 2025)
- UniversalPRM-7B (Tan et al., 2025)

For Kimi K2, DeepSeek-R1, and GLM-4.5-Air, we used `together.ai` for inference. All other open-weight baselines were run locally, hosted via vLLM (Kwon et al., 2023) or SGLang (Zheng et al., 2024b).

F.1 PRM Threshold Tuning

To decide the cutoff threshold for evaluated PRMs, we select 100 responses at random from our benchmark and tune PRM performance against this subset, following (Zheng et al., 2024a). The same 100 responses are kept fixed across all baselines, and we sweep the threshold from 0.1 to 0.9 in increments of 0.05. To select the threshold, we compute the harmonic mean of the three task-specific Balanced F1 Scores, prioritizing selecting a threshold that yields strong yet balanced performance. We find that PRM performance can vary considerably based on chosen threshold.

Error Type	Description	%
Propagated Error	Locally valid logic depending on a prior incorrect step	39.3
Unjustified Leap	Claims without sufficient justification, or conclusions that do not follow from premises	31.2
Incorrect Math	Wrong computations, misapplied theorems, or invalid transformations	20.0
Incomplete Cases	Missing cases in case analysis, or overgeneralization	5.5
Invalid Setup	Wrong WLOG, invalid structural assumptions, or circular reasoning	4.0

Table 6: Taxonomy of incorrect steps across 455 erroneous steps in GPT-5-generated responses.

G Error Taxonomy of Incorrect Steps

To better characterize the kinds of mistakes frontier models make, we categorize the 455 incorrect steps from GPT-5-generated responses into five classes. *Propagated Error* dominates at 39.3%, consistent with the error-cascade patterns in Fig. 3, indicating that a substantial share of step-level errors are downstream consequences of a single earlier mistake rather than independent failures.

H GPT-5 Verifier Failure Modes

To understand where even the strongest verifier falls short, we manually categorize the 284 step-level disagreements between GPT-5 and human annotators into five failure modes. *Error Propagation Confusion* accounts for half of all disagreements, suggesting that improving verifiers' ability to separate local step validity from global proof correctness is the single most impactful direction for reducing verification failures.

Failure Mode	Description	%
Error Propagation Confusion	Conflating local vs. global correctness (e.g., accepting a locally flawed step because the proof direction seems right, or rejecting a locally correct step due to a prior flagged step)	50.0
Rigor Misassessment	Accepting heuristic arguments as proof, or rejecting standard techniques as insufficient	18.0
Surface-Level Evaluation	Judging based on superficial cues rather than mathematical substance	13.7
Mathematical Misunderstanding	The verifier's own reasoning contains errors (e.g., wrong counterexamples or misinterpretations)	13.4
Case Boundary Blindness	Missing incomplete case analysis, or flagging irrelevant edge cases	4.9

Table 7: Failure modes observed across 284 step-level disagreements between GPT-5 and human annotators.

Computational Budget and Infrastructure Details
Computation Time: 12 hours
Human Annotation time: 500 hours
GPU Hardware: 8 x NVIDIA H200 (143,771 MiB RAM each)

Table 8: Infrastructure Details while generating Hard2Verify Dataset.

I Computational Cost

This section reports the computational and human effort required to construct the Hard2Verify dataset, including model inference for solution generation, data processing, and expert annotation.

Table 8 summarizes the overall budget and infrastructure used in this work. The compute time reported corresponds to the end-to-end pipeline for dataset generation and preparation (including solution sampling and formatting), executed on dedicated GPU resources. Human annotation time reflects the cumulative effort spent by expert annotators and reviewers on step-level correctness labeling and quality control.