

DEEPTHEOREM: ADVANCING LLM REASONING FOR THEOREM PROVING THROUGH NATURAL LANGUAGE AND REINFORCEMENT LEARNING

Anonymous authors

Paper under double-blind review

ABSTRACT

Theorem proving serves as a major testbed for evaluating complex reasoning abilities in large language models (LLMs). However, traditional automated theorem proving (ATP) approaches rely heavily on *formal* proof systems that poorly align with LLMs' strength derived from *informal*, natural language knowledge acquired during pre-training. To fully leverage the theorem-proving knowledge acquired from pre-training, in this work, we present DeepTheorem, a comprehensive informal theorem-proving suite exploiting natural language to enhance LLM mathematical reasoning. DeepTheorem includes 1) **a large-scale dataset of 121K high-quality IMO-level informal theorems and proofs** spanning diverse mathematical domains, rigorously annotated for correctness, difficulty, and topic categories, accompanied by systematically constructed verifiable theorem variants; 2) **adaptation of RL-Zero explicitly to informal theorem proving**, leveraging the verified theorem variants to incentivize robust mathematical inference; 3) **comprehensive outcome and process evaluation metrics** examining proof correctness and the quality of reasoning steps; and 4) **a novel informal theorem proving benchmark** consolidated from three established math competitions, formatted for automatic evaluation. Extensive experimental analyses demonstrate DeepTheorem significantly improves LLM theorem-proving performance compared to existing datasets and supervised fine-tuning protocols, achieving state-of-the-art accuracy and reasoning quality. Our findings highlight DeepTheorem's potential to fundamentally advance automated informal theorem proving and mathematical exploration.

1 INTRODUCTION

Theorem proving is widely regarded as a pinnacle challenge for evaluating advanced reasoning capabilities of both human and artificial intelligence. It requires integrating diverse cognitive facets such as abstraction, strategic inference, pattern recognition, and meticulous logical deduction. Recent advancements in deep learning, especially in large language models (LLMs), have significantly reshaped the landscape of automated theorem proving (ATP). Much prior work attempts ATP by integrating LLMs with either formal proof engines such as Lean, Coq, and Isabelle (Zheng et al., 2022; Liu et al., 2023; Tsoukalas et al., 2024) or domain-specific languages from ProofWiki (Welleck et al., 2022). However, these proof methods impose a significant barrier for LLMs whose primary strength derives from the vast corpus of natural language and LaTeX-based mathematical texts used during pre-training. This inherent misalignment limits LLMs' capability in theorem proving, leaving a considerable gap between their potential and actual performance.

In this paper, we present **DeepTheorem**, a novel, comprehensive suite expressly designed to leverage natural language to unleash the latent mathematical reasoning ability of LLMs for theorem proving. Instead of relying on formal proof assistants, DeepTheorem offers a scalable, intuitive, and flexible alternative, enabling LLMs to generate informal mathematical proofs aligning closely with human mathematicians' heuristic-driven thinking. Central to our approach is the construction of a large-scale benchmark comprising 121K IMO-level informal mathematical theorems with precise annotations such as correctness labels, difficulty levels, diverse mathematical domains, and verifiable theorem variants amenable to advanced reinforcement learning paradigms. Additionally,

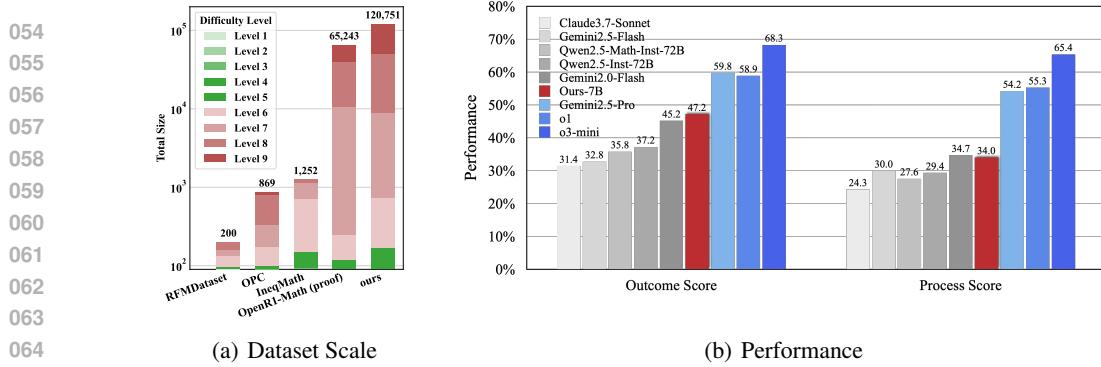


Figure 1: (a): Our dataset surpasses other informal theorem proving datasets in both scale and difficulty; (b): RL-Zero training with our DeepTheorem datasets on 7B model achieves strong results.

we adapt reinforcement learning algorithms to the problem of informal theorem proving for the first time, significantly enhancing LLM’s ability to reason mathematically beyond the constraints of supervised fine-tuning (SFT). To evaluate existing and newly trained models on informal theorem proving, we also construct a new benchmark sourced from established mathematics competitions, and propose comprehensive evaluation metrics that rigorously assess the correctness of generated proofs and the processes underlying the proofs themselves.

Through extensive experiments, we show that leading LLMs still exhibit significant limitations in theorem proving. However, when trained with DeepTheorem, they achieve substantial performance improvements over models trained using existing datasets, showcasing the effectiveness of our natural-language-focused approach. Our results underscore the promise of DeepTheorem to redefine LLM-driven mathematical reasoning, offering a robust platform for continued progress in automated and scalable informal theorem proving.

To sum up, our key contributions are:

- We introduce the *DeepTheorem* framework, a comprehensive informal theorem-proving suite exploiting natural language to enhance LLM mathematical reasoning;
- We open-source a large-scale natural-language theorem collection of 121K informal mathematical theorems and corresponding high-quality proofs at IMO-level difficulty, suitable for both SFT and RL;
- We innovatively adapt the RL-Zero training method explicitly to informal theorem proving, significantly enhancing LLM’s reasoning capacity beyond traditional SFT methods;
- We introduce a new benchmark for evaluating informal theorem proving, and develop a comprehensive evaluation framework assessing both the correctness of theorem proofs (outcome evaluation) and the completeness, logical validity, and correctness of generated reasoning processes (process evaluation);
- Through extensive experiments, we establish the superiority of our DeepTheorem training paradigm, achieving state-of-the-art performance and surpassing existing informal theorem datasets and training methods;

2 DATASET

Overview The *DeepTheorem* dataset¹ is a novel, large-scale resource designed to advance LLMs in informal mathematical theorem reasoning. Mined from a diverse web corpus, it addresses the need for challenging, decontaminated, and diverse topics to push LLMs toward frontier theorem proving. As illustrated in Figure 2, each entry in the dataset offers distinct features tailored to support diverse research objectives, including: 1) a mathematical theorem in standardized LaTeX format; 2) a True-or-False correctness label, where a False label indicates that the theorem can be mathematically disproved; 3) a fine-grained difficulty score ranging from level 5~10; 4) mathematical topics for targeted analysis; and 5) detailed step-by-step proof solutions generated by o3-mini.

¹In the rest of this section, *DeepTheorem* dataset refers specifically to the DeepTheorem training dataset.

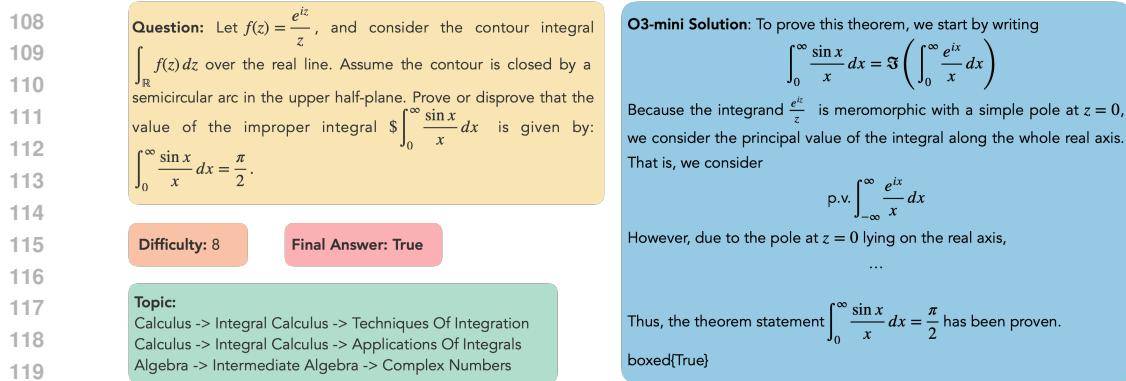
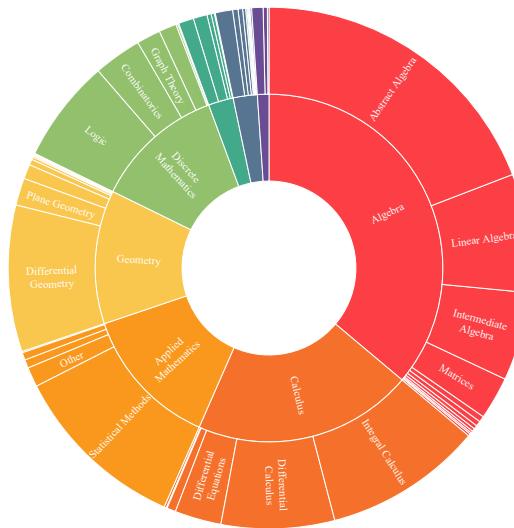


Figure 2: A data sample from the DeepTheorem dataset.

Figure 3: Statistics of *DeepTheorem* dataset hierarchical topics. See Appendix A for details.

in Figure 4, *DeepTheorem* dataset emphasizes theorems at high difficulty levels (6–9), surpassing existing corpora in complexity and challenge, presenting significant challenges for state-of-the-art LLMs while aligning with frontier, IMO-level benchmarks such as FIMO (Liu et al., 2023).

• **Strict decontamination:** To preserve evaluation integrity, *DeepTheorem* dataset employs rigorous decontamination processes to avoid overlap with widely used benchmarks. The targets of

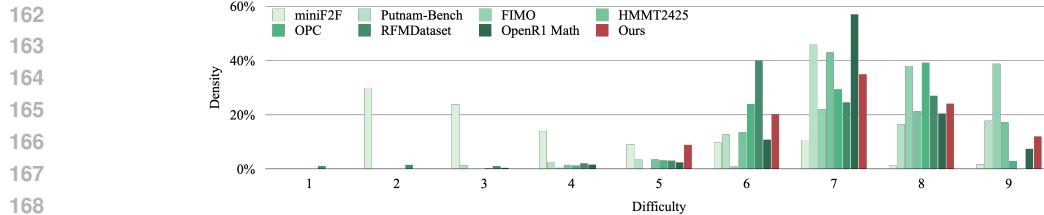


Figure 4: Difficulty density histogram comparison across widely used benchmarks (i.e. miniF2F, Putnam-Bench, FIMO, HMMT2425) and other recent theorem proving datasets (i.e. OPC, RFM-Dataset, theorem subset of OpenR1-Math).

our decontamination includes general math reasoning benchmarks - MATH (Hendrycks et al., 2021b), AIME (MAA, a), AMC (MAA, b), Minerva Math (Lewkowycz et al., 2022), Olympiad-Bench (He et al., 2024), Omni-MATH (Gao et al., 2025), MathOdyssey (Fang et al., 2024), GAOKAO (Zhong et al., 2024), JEEBench (Arora et al., 2023), MMLU-STEM (Hendrycks et al., 2021a), CMATH (Wei et al., 2023), OlympicArena (Huang et al., 2024), GSM8K (Cobbe et al., 2021), GPQA (Rein et al., 2024) - and theorem proving benchmarks: miniF2F (Zheng et al., 2022), PutnamBench (Tsoukalas et al., 2024), FIMO (Liu et al., 2023), and HMMT (Harvard-MIT Mathematics Tournament, 2024, 2025).

- **Proofs from advanced LLMs:** *DeepTheorem* dataset includes concise, high-quality proof solutions generated by o3-mini, tailored for supervised fine-tuning (SFT). These proofs provide a compact yet complete outline of the logical steps required to prove (or disprove) each theorem, optimized for clarity and brevity. Unlike verbose or overly formal proofs, these proofs, expressed in LaTeX, align with the informal nature of LLMs, making them an effective learning signal. By incorporating these proofs, the dataset enables models to internalize structured reasoning patterns, improving their ability to generate coherent and logically sound mathematical arguments.

2.1 CONSTRUCTING *DeepTheorem* DATASET

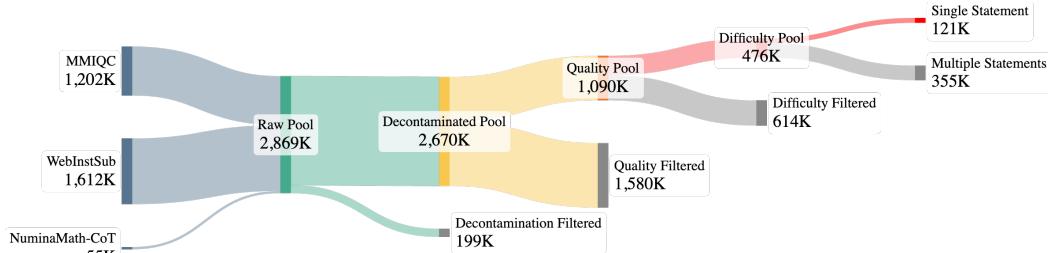


Figure 5: The process pipeline of constructing *DeepTheorem* dataset.

As shown in Figure 5, the construction pipeline of *DeepTheorem* dataset starts by aggregating raw data from multiple sources, including MMICQ (Liu et al., 2025), WebInstruct (Yue et al., 2024), and NuminaMath-CoT (LI et al., 2024).

Decontamination To ensure the integrity of *DeepTheorem* dataset, we implement a rigorous decontamination process to eliminate overlap between training and testing datasets. The process employs a recall-and-justify pipeline to identify and remove potential duplicates, consisting of three key steps:

1. *Embedding Generation*: We use an embedding model² (Reimers & Gurevych, 2019; Toshniwal et al., 2025) to generate sentence embeddings for all theorem statements in the training and testing datasets.

² paraphrase-multilingual-MiniLM-L12-v2

216 2. *Similarity Recall*: For each training sample, we compute its embedding cosine similarity to all
 217 test samples, and recall the top five testing samples exceeding a similarity threshold (set to 0.7).
 218 3. *Contamination Justification*: An LLM³ evaluates whether the recalled test samples are contami-
 219 nated within the current training sample (Section F.1).
 220

221 This process removed approximately 199K contaminated samples, effectively identifying identical
 222 cases, generalized questions, and converse theorems. Removed contaminated examples are shown
 223 in Appendix B. Approximately 2.6M samples remain for the next processing step.

224 **Quality control and proof generation** We also implement a rigorous quality control pipeline for
 225 generating and validating theorem statements and their proofs. The process involves four key steps:

226 1. *Theorem Justification*: An LLM verifies that the question is complete, and indeed a theorem-
 227 proving question (Section F.2).
 228 2. *Rationale Summarization*: An LLM summarizes the original question and generates a formatted,
 229 concise, self-contained theorem (Section F.3).
 230 3. *Proof Generation*: o3-mini (high effort) generates the proof solution with True-or-False conclu-
 231 sion about the theorem (Section F.4).
 232 4. *Logical Validation*: The LLM performs an extra justification step to check that the theorem-proof
 233 pair is logically coherent. (Section F.5)

234 This systematic approach yielded 1.08M high-quality, mathematically sound theorem-proof pairs.
 235

236 **Difficulty and single statement annotation** We annotate the difficulty levels of *DeepTheo-*
 237 *rem* dataset, and remove questions with multiple statements to prove.

238 1. *Difficulty Annotation*: An LLM analyzes each theorem statement following the strategy of Gao
 239 et al. (2025), considering factors such as logical complexity, mathematical prerequisites, and
 240 proof length, to assign a difficulty score on a scale of 1 to 9. Only questions with a difficulty
 241 score of at least 5 are retained. (Examples in Section B, prompt in Section F.6)
 242 2. *Single-Statement Filtering*: We filter out samples that query for proving multiple statements,
 243 retaining only those with a single, well-defined theorem to ensure clarity and consistency with
 244 evaluation.
 245 3. *Topic Annotation*: Finally, we annotate the topic domain of the mathematical theorems with
 246 LLMs. (Section F.7)

247 The difficulty and single statements filtering results in 121K challenging theorems, yielding the final
 248 *DeepTheorem* dataset.

253 3 THEOREM PROVING VIA REINFORCEMENT LEARNING

254 **Motivation** Conventionally, informal theorem-proving datasets are utilized through supervised
 255 fine-tuning (SFT), where models learn to generate proofs by imitating dataset examples. However,
 256 recent studies on RL-Zero demonstrate its superior performance over SFT by leveraging a base
 257 model’s pretrained knowledge and exploratory capabilities (Jaech et al., 2024; DeepSeek-AI et al.,
 258 2025). This raises a natural question: *Can we harness the base model’s exploration ability for infor-*
 259 *mal theorem proving?* In this section, we explore the possibility of utilizing RL-Zero for informal
 260 theorem proving. The process involves three key steps: 1) data augmentation to generate contradic-
 261 tory theorem variants for binary rewards; 2) RL-Zero training with GRPO (Shao et al., 2024); and
 262 3) Evaluation of the theorem-proof generation.

263 3.1 THEOREMS WITH VERIFIABLE REWARDS

264 **Theorems can be disproved** To construct a theorem with rewards for RL-Zero, we make the key
 265 observation:

266 267 268 269 ³ GPT-4o is used for annotation in this section unless otherwise specified.

270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323

Insight

A statement need not be correct but can be also proven incorrect, enabling a binary reward structure compatible with RL-Zero.

This observation allows us to transform *DeepTheorem*'s theorems into true-or-false variants, facilitating RL training that incentivizes robust reasoning.

To construct such training data, we use an LLM to expand the original theorems into contradictory variants that can be *disproved*. Specifically, we strictly limit the transformation made to the original theorem, so that the resulting variant is either entailed by or contradictory from the original theorem. Consider the example in Table 1 (omitting the hypotheses for simplicity): if the original theorem can be proved, Variant #1 is also correct and can be mathematically proved in the same manner as the original one, while Variant #2 must be incorrect and can be disproved.

With such logically entailing or contradictory transformations, we are able to construct variants of a theorem that are guaranteed to be correct or incorrect by only accessing the theorem itself but not the proof process, which makes this transformation task much easier than annotating new math statements, and thus allowing a relatively weaker LLM (e.g. Qwen2.5-72B-Instruct, Yang et al., 2024) to perform it. After this expansion phase, we further annotate the completeness of the resulting theorem pool and finally acquire a training set of 242K mathematical theorems that can either be proved or disproved, each with a complete proof trajectory (see Appendix F.8 for more details).

3.2 BINARY REWARDS ACTIVATE THEOREM PROOF GENERATION

With the aforementioned theorem variants, we can now apply reinforcement learning to natural language theorem proving. Specifically, we adopt the GRPO algorithm (Shao et al., 2024).

Proof generation with RL Inspired by the success of reasoning-specialized models such as R1 and its open-source reproductions (DeepSeek-AI et al., 2025; Hu et al., 2025), we encourage the model to enclose its reasoning process in `<think> </think>` tags in the system prompt to incentivize more detailed reasoning behaviours (see Appendix F.9), and then ask the model to end each proof with either “`\boxed{proved}`” or “`\boxed{disproved}`”. In the reward function, we extract this answer and compare it against the ground truth, giving a reward of 1 if the answer matches, and 0 otherwise. We also enforce several sanity checks to prevent model collapse: if the ratio of white spaces in a model’s solution is less than 0.05 or the average character repetition count is greater than 300, then a reward of 0 is issued regardless of the answer.

3.3 EVALUATION

The theorem-proving questions used for evaluation are drawn from two challenging benchmarks — FIMO (Liu et al., 2023) and Putnam (Tsoukalas et al., 2024) — and a newly constructed theorem-proving subset of HMMT (Harvard-MIT Mathematics Tournament, 2024, 2025).

Outcome evaluation Evaluating the correctness of natural language (NL) proofs poses a significant challenge, as it mirrors the complexity faced by humans in assessing the logical coherence and mathematical validity of informal reasoning. Unlike formal theorem-proving systems that rely on structured logic, NL proofs lack a standardized format, making their evaluation inherently subjective and difficult to automate. To address this, we propose a novel evaluation framework that leverages multiple en-

Table 1: An example of theorem variants given an original theorem.

Theorems	Example
Original	$x > 1$
Variant 1	$x > 0$
Variant 2	$x < 1$

Table 2: Test data statistics. Each original theorem is manually expanded into multiple entailing or contradictory variants. *Random accuracy* indicates the expected score of random guessing following the outcome criteria described below.

Bench	Scale	Variants (Avg.)	Random Acc.
FIMO	172	2.7	17.4
HMMT	205	3.5	11.2
Putnam	281	2.9	15.4

tailing and contradictory variants derived from each theorem. By assessing the model’s ability to consistently assign correct truth values across these variants, we indirectly estimate its theorem justification ability. When the number of variants is sufficiently large, this approach provides a robust proxy for evaluating the correctness of NL proof generation.

Thus, we manually expand each question in the three data sources into multiple entailing or contradictory variants following the same variant generation protocol in Section 3.1, and the resulting benchmarks are shown in Table 2. When evaluating a model, we ask it to either prove or disprove each theorem and corresponding variants, and evaluate the results with the criteria below:

Outcome Criteria

A test case in a theorem testing set is passed if and only if:

1. The model explicitly produces a truth value (true or false) for theorems and variants;
2. The predicted truth value for the original theorem is correct;
3. The predicted truth values for all entailing variants are the same as the original theorem;
4. The predicted truth values for all contradictory variants are the inverse of the original theorem.

Process evaluation Since theorem proving requires generating logically validated proofs for each reasoning step, we also develop a process evaluation framework that evaluates the quality of proof along four dimensions:

- **Logical Validity:** Check if each step follows logically from the previous one. Flag any logical errors.;
- **Completeness:** Verify if all necessary cases and steps are included to prove the theorem;
- **Correctness:** Confirm if the final conclusion is correct;
- **Clarity:** Assess if the proof is clear, unambiguous, and well-explained.

We use GPT-4o as the LLM judge and ask it to score the proof using a weighted sum of the four dimensions (prompt given in Appendix F.10). In Appendix D, we also present the results using o3-mini and Ling-1T (Team & AI, 2025) as the judge as well as human evaluation.

4 EXPERIMENTS

4.1 SETTINGS

We train two sets of models, using supervised fine-tuning (SFT) and zero reinforcement learning (RL-Zero) respectively, starting from Qwen2.5-Base (Yang et al., 2024) (additional results with Qwen3-Base (Yang et al., 2025) are provided in Appendix C). For SFT, we train the models for 3 epochs on the complete proof solutions in the dataset, using one machine for training each model. For RL-Zero, we adopt GRPO with batch size 128, group size 64, and maximum rollout length 8192. We train the models for 1000 steps, and distribute each model across two machines during training. Following the settings of Hu et al. (2025), we do not apply any KL regularization or entropy loss, as we find that KL regularization has a negligible impact on model performance, while entropy loss leads to model collapse.

As a baseline, we select the theorem-proving subset of OpenR1-Math (Face, 2025), the highest-quality existing theorem-proving dataset with complete questions and responses. We apply the same processing pipelines to it as detailed in Section 2.1, which yields 66K original theorems and 130K variants in total. We dub this processed dataset *OpenR1-Math-Proof*.

4.2 MAIN RESULTS

DeepTheorem with RL-Zero achieves the best performance The main results are presented in Table 3. *DeepTheorem* demonstrates superior performance over OpenR1-Math-Proof, especially

378
379 Table 3: Outcome (out.) and Process (proc.) evaluation of models trained on OpenR1-Math-Proof
and DeepTheorem.

380 381 Model	382 Strategy	383 Data	384 FIMO		385 HMMT		386 Putnam		387 Avg.	
			388 out.	389 proc.	390 out.	391 proc.	392 out.	393 proc.	394 out.	395 proc.
396 1.5B	397 SFT	398 OpenR1-Proof	399 20.63	400 8.66	401 11.86	402 4.80	403 35.42	404 18.98	405 22.64	406 10.81
		407 DeepTheorem	408 31.75	409 18.86	410 15.25	411 9.41	412 36.46	413 21.43	414 27.82	415 16.57
	416 RL	417 OpenR1-Proof	418 34.92	419 8.54	420 16.95	421 5.10	422 55.21	423 17.92	424 35.69	425 10.52
		426 DeepTheorem	427 31.75	428 15.23	429 23.73	430 10.15	431 52.08	432 22.79	433 35.85	434 16.06
397 3B	398 SFT	399 OpenR1-Proof	400 23.81	401 12.85	402 15.25	403 6.90	404 43.75	405 27.96	406 27.60	407 15.90
		408 DeepTheorem	409 33.33	410 20.38	411 20.34	412 12.15	413 36.46	414 25.43	415 30.04	416 19.32
	417 RL	418 OpenR1-Proof	419 34.92	420 14.33	421 23.73	422 11.72	423 57.29	424 35.11	425 38.65	426 20.39
		427 DeepTheorem	428 38.10	429 23.39	430 25.42	431 13.56	432 52.08	433 33.84	434 38.53	435 23.60
398 7B	399 SFT	400 OpenR1-Proof	401 30.16	402 18.23	403 15.25	404 8.63	405 48.96	406 32.95	407 31.46	408 19.94
		409 DeepTheorem	410 34.92	411 26.69	412 22.03	413 15.41	414 41.67	415 33.50	416 32.87	417 25.20
	418 RL	419 OpenR1-Proof	420 42.86	421 22.79	422 25.42	423 13.15	424 60.42	425 38.94	426 42.90	427 24.96
		428 DeepTheorem	429 55.56	430 39.07	431 28.81	432 20.85	433 57.29	434 42.20	435 47.22	436 34.04

393
394 for the 7B backbone and in terms of process evaluation. On the other hand, our RL-Zero training paradigm consistently outperforms SFT, validating the effectiveness of RL-Zero in pushing the models' reasoning capabilities beyond the limit of SFT.

400
401 **DeepTheorem achieves strong parameter efficiency** We demonstrate that our
402 DeepTheorem-RL strategy achieves strong
403 parameter efficiency in Figure 6. Compared
404 to the Qwen2.5 series, training DeepTheorem
405 on 1.5 to 7B models significantly improves
406 the informal theorem proving boundary at
407 parameter-performance space. Moreover, when
408 extrapolated DeepTheorem parameter efficiency
409 also surpasses SOTA commercial models such as
410 o1 and o3-mini.

412
413 **SOTA performance at equal model scale** In
414 Table 4, we also provide the evaluation results
415 of SOTA LLMs on the three benchmarks. These
416 results suggest that theorem proving, especially
417 our newly constructed HMMT benchmark, is still
418 quite challenging for LLMs. On the other hand,
419 our 7B model, trained with RL-Zero on *DeepTheorem*,
420 outperforms SOTA models of much larger
421 sizes, including those specialized in math and reasoning,
422 demonstrating the superior quality of
423 *DeepTheorem* and our innovative outcome-supervised RL training approach for theorem proving.

424
425 **Reasoning with theorem proving skills** In
426 Figure 7, we visualize the techniques used by
427 our 7B model trained with RL on *DeepTheorem*,
428 where direct proof is most commonly used,
429 followed by proof by exhaustion and construction.
430 In Appendix E, we provide a non-cherry-picked
431 example generation, finding the model to deliver
432 a clear and correct disproof, highlighting its effi-
433 cacy in tackling advanced mathematical problems
434 with precision and clarity.

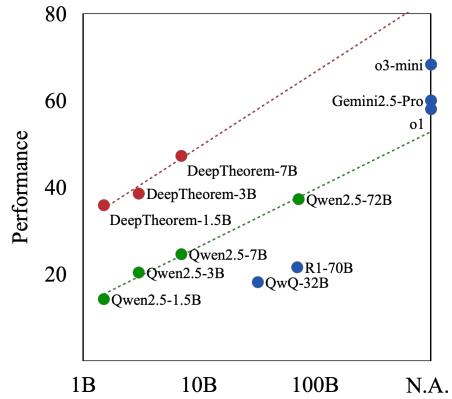


Figure 6: Performance improvement of models trained with DeepTheorem over baselines on theorem proving benchmarks.

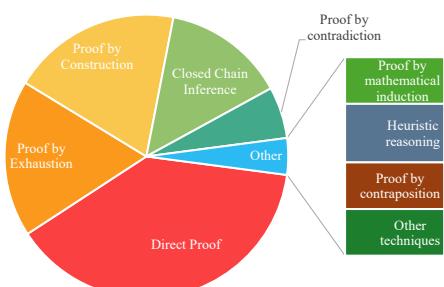


Figure 7: Distribution of proof techniques used by DeepTheorem-7B.

432
 433 Table 4: Evaluation comparison of our model trained with RL on DeepTheorem with SOTA LLMs
 434 for both commercial models and open source models. *: denotes our method. Inst: Instruct; DS:
 435 Deepseek.

436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485	436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485		436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485		436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485		436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485		436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485		
		FIMO		HMMT		Putnam		Avg.		#Rank	
		<i>out.</i>	<i>proc.</i>	<i>out.</i>	<i>proc.</i>	<i>out.</i>	<i>proc.</i>	<i>out.</i>	<i>proc.</i>	<i>out.</i>	<i>proc.</i>
<i>Commercial Models</i>											
Claude3.7-Sonnet	34.92	26.28	13.56	8.29	45.83	38.33	31.44	24.30	12	12	
Gemini2.5-Pro	57.14	54.06	57.63	49.82	64.58	58.75	59.78	54.21	4	4	
Gemini2.5-Flash	30.16	28.95	25.42	22.02	42.71	38.98	32.76	29.98	11	8	
GPT-4o	34.92	30.70	16.95	14.59	22.92	18.88	24.93	21.39	14	13	
o1-mini	60.32	55.23	35.59	30.90	61.46	52.88	52.46	46.34	6	6	
o1	66.67	61.00	47.46	47.30	62.50	57.55	58.88	55.28	5	3	
o3-mini	80.95	77.61	45.76	43.47	78.12	75.12	68.28	65.40	2	1	
<i>Open Source Models</i>											
Qwen2.5-Inst-7B	30.16	21.13	10.17	6.83	33.33	25.39	24.55	17.78	15	15	
Qwen2.5-Inst-72B	49.21	37.35	13.56	9.78	48.96	41.00	37.24	29.38	8	9	
Qwen2.5-Math-Inst-7B	28.57	18.86	3.39	1.61	25.00	18.79	18.99	13.09	17	18	
Qwen2.5-Math-Inst-72B	47.62	36.02	11.86	8.61	47.92	38.04	35.80	27.56	10	11	
R1-Distill-7B	6.35	4.27	0.00	0.00	4.17	2.58	3.51	2.28	19	19	
R1-Distill-70B	17.46	14.05	16.95	13.52	30.21	23.10	21.54	16.89	16	16	
QwQ-32B	17.46	15.41	11.86	10.10	25.00	18.19	18.11	14.57	18	17	
Llama3.3-Inst-70B	41.27	27.33	10.17	4.12	36.46	25.30	29.30	18.92	13	14	
Qwen3-32B	73.02	55.04	52.54	36.88	79.17	57.81	68.24	49.91	3	5	
Kimi-K2-Thinking-1T	25.40	19.49	32.20	25.77	51.04	40.40	36.21	28.56	9	10	
Ling-1T	85.71	64.27	55.93	37.09	83.33	64.79	74.99	55.38	1	2	
*DeepTheorem-RL-7B	55.56	39.07	28.81	20.85	57.29	42.20	47.22	34.04	7	7	

5 RELATED WORK

Theorem proving Theorem proving is a non-trivial task for natural language processing. In the context of LLMs, most works on theorem proving focus on formal languages such as Lean, Coq, and Isabelle (Xin et al., 2024; Ren et al., 2025). NaturalProofs (Welleck et al., 2021) and NaturalProver (Welleck et al., 2022) represent pioneering works that attend to informal theorem proving, though utilizing a domain-specific language from the ProofWiki website rather than the more human-accessible LaTeX-based natural language. Concurrent to our work, several small-scale informal theorem proving datasets have been introduced to the community, including OPC (Dekoninck et al., 2025), IneqMath (Sheng et al., 2025), and RFMDataset (Guo et al., 2025). However, these datasets are designed for analyzing LLMs’ reasoning errors and failure modes in theorem proving, whereas *DeepTheorem* represents the first attempt at scaling up the training of informal theorem proving models.

Learning to reason with RL-zero RL-Zero (DeepSeek-AI et al., 2025) is a streamlined framework designed to develop reinforcement learning capabilities in LLMs without SFT. While recent advances in LLM reasoning have been significantly influenced by RL techniques (Jaech et al., 2024; DeepSeek-AI et al., 2025; Team, 2024; xAI, 2025; Google, 2025), existing approaches predominantly focus on closed-form questions, addressing only a subset of reasoning problems. In contrast, we investigate the application of RL-Zero in process-oriented reasoning, specifically in informal theorem-proving. To the best of our knowledge, this is the first study to apply RL-Zero to informal theorem proving, marking a significant advance in enabling LLMs to address more diversified reasoning tasks in mathematical and logical domains.

6 CONCLUSION

In this paper, we introduce DeepTheorem, a novel comprehensive theorem-proving suite involving a large-scale annotated dataset of 121K IMO-level informal mathematical theorems and corresponding high-quality natural-language proofs, alongside systematically constructed verifiable theorem variants. We further adapt RL-Zero method to informal theorem reasoning, significantly surpassing supervised fine-tuning in performance. Comprehensive evaluations involving outcome accuracy and detailed process assessment on our newly constructed benchmark demonstrate the effectiveness

486 of our approach, achieving state-of-the-art theorem-proving performance and significantly pushing
 487 LLM reasoning boundaries. Through these contributions, DeepTheorem provides a robust founda-
 488 tion for future advancements in automated mathematical theorem proving, leveraging natural lan-
 489 guage flexibility to empower scalable, human-like reasoning abilities in large language models.
 490

491 **REFERENCES**
 492

493 Daman Arora, Himanshu Singh, and Mausam. Have LLMs advanced enough? a challenging prob-
 494 lem solving benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika
 495 Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language
 496 Processing*, pp. 7527–7543, Singapore, December 2023. Association for Computational Linguis-
 497 tics. doi: 10.18653/v1/2023.emnlp-main.468. URL [https://aclanthology.org/2023.
 498 emnlp-main.468](https://aclanthology.org/2023.emnlp-main.468).

499 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
 500 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John
 501 Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL
 502 <https://arxiv.org/abs/2110.14168>.

503 DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu,
 504 Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu,
 505 Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao
 506 Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,
 507 Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao,
 508 Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,
 509 Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang
 510 Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai
 511 Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang,
 512 Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang,
 513 Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang,
 514 Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang,
 515 R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye,
 516 Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing
 517 reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.
 518 48550/ARXIV.2501.12948. URL <https://doi.org/10.48550/arXiv.2501.12948>.

519 Jasper Dekoninck, Ivo Petrov, Kristian Minchev, Mislav Balunovic, Martin T. Vechev, Miroslav
 520 Marinov, Maria Drencheva, Lyuba Konova, Milen Shumanov, Kaloyan Tsvetkov, Nikolay
 521 Drenchev, Lazar Todorov, Kalina Nikolova, Nikolay Georgiev, Vanesa Kalinkova, and Mar-
 522 gulan Ismoldayev. The open proof corpus: A large-scale study of llm-generated mathemati-
 523 cal proofs. *CoRR*, abs/2506.21621, 2025. doi: 10.48550/ARXIV.2506.21621. URL <https://doi.org/10.48550/arXiv.2506.21621>.

525 Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL <https://github.com/huggingface/open-r1>.

527 Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. Mathodyssey: Benchmarking math-
 528 ematical problem-solving skills in large language models using odyssey math data. *CoRR*,
 529 abs/2406.18321, 2024. doi: 10.48550/ARXIV.2406.18321. URL <https://doi.org/10.48550/arXiv.2406.18321>.

531 Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao
 532 Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran
 533 Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang.
 534 Omni-math: A universal olympiad level mathematic benchmark for large language models. In
 535 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore,
 536 April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=yaqPf0KA1N>.

538 Google. Gemini 2.0 flash thinking, 2025. URL <https://cloud.google.com/vertex-ai/generative-ai/docs/thinking>. Accessed on March 25, 2025.

540 Dadi Guo, Jiayu Liu, Zhiyuan Fan, Zhitao He, Haoran Li, Yumeng Wang, and Yi R. Fung. Mathematical proof as a litmus test: Revealing failure modes of advanced large reasoning models.
 541 *CoRR*, abs/2506.17114, 2025. doi: 10.48550/ARXIV.2506.17114. URL <https://doi.org/10.48550/arXiv.2506.17114>.

542

543

544 Harvard-MIT Mathematics Tournament. Hmmt dataset and resources. <https://www.hmmt.org/>, 2024, 2025. Accessed: 2025-05-12.

545

546

547 Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi
 548 Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun.
 549 Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual
 550 multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Pro-
 551 ceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume
 552 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 3828–3850. Associa-
 553 tion for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.211. URL
 554 <https://doi.org/10.18653/v1/2024.acl-long.211>.

555

556 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob
 557 Steinhardt. Measuring massive multitask language understanding. In *9th International Confer-
 558 ence on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenRe-
 559 view.net, 2021a. URL <https://openreview.net/forum?id=d7KBjmI3GmQ>.

560

561 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang,
 562 Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with
 563 the MATH dataset. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings
 564 of the Neural Information Processing Systems Track on Datasets and Benchmarks
 565 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021b. URL
 566 <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html>.

567

568 Jingcheng Hu, Yinmin Zhang, Qi Han, Dixin Jiang, Xiangyu Zhang, and Heung-Yeung Shum.
 569 Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base
 570 model. *CoRR*, abs/2503.24290, 2025. doi: 10.48550/ARXIV.2503.24290. URL <https://doi.org/10.48550/arXiv.2503.24290>.

571

572 Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan,
 573 Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang,
 574 Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu
 575 Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. Olympicarena:
 576 Benchmarking multi-discipline cognitive reasoning for superintelligent AI. In Amir Globersons,
 577 Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng
 578 Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on
 579 Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, De-
 580 cember 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/222d2eaf24cf8259a35d6c7130d31425-Abstract-Datasets_and_Benchmarks_Track.html.

581

582 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec
 583 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard
 584 Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett,
 585 Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrei
 586 Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghor-
 587 bani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao
 588 Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lu-
 589 garesi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen,
 590 Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan
 591 Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely,
 592 David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund
 593 Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays,
 Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas,
 Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo,

594 Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao,
 595 Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won
 596 Chung, Ian Kivlichan, Ian O’Connell, Ian Osband, Ignasi Clavera Gilaberte, and Ilge Akkaya.
 597 Openai o1 system card. *CoRR*, abs/2412.16720, 2024. doi: 10.48550/ARXIV.2412.16720. URL
 598 <https://doi.org/10.48550/arXiv.2412.16720>.

599 Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ra-
 600 masesh, Ambrose Sloane, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam
 601 Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with lan-
 602 guage models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Ad-*
 603 *vances in Neural Information Processing Systems*, volume 35, pp. 3843–3857. Curran Associates,
 604 Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/18abbeef8cfe9203fdf9053c9c4fe191-Paper-Conference.pdf.

605 Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang,
 606 Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann
 607 Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [<https://huggingface.co/AI-MO/NuminaMath-1.5>] (https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

608 Chengwu Liu, Jianhao Shen, Huajian Xin, Zhengying Liu, Ye Yuan, Haiming Wang, Wei Ju,
 609 Chuanyang Zheng, Yichun Yin, Lin Li, Ming Zhang, and Qun Liu. FIMO: A challenge formal
 610 dataset for automated theorem proving. *CoRR*, abs/2309.04295, 2023. doi: 10.48550/ARXIV.
 611 2309.04295. URL <https://doi.org/10.48550/arXiv.2309.04295>.

612 Haoxiong Liu, Yifan Zhang, Yifan Luo, and Andrew C. Yao. Augmenting math word problems
 613 via iterative question composing. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25,*
 614 *Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March*
 615 *4, 2025, Philadelphia, PA, USA*, pp. 24605–24613. AAAI Press, 2025. doi: 10.1609/AAAI.
 616 V39I23.34640. URL <https://doi.org/10.1609/aaai.v39i23.34640>.

617 MAA. American invitational mathematics examination (AIME). Mathematics Competition Series,
 618 a. URL <https://maa.org/math-competitions/aime>.

619 MAA. American mathematics competitions (AMC 10/12). Mathematics Competition Series, b.
 620 URL <https://maa.org/math-competitions/amc>.

621 Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-
 622 networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of*
 623 *the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Inter-*
 624 *national Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong,*
 625 *China, November 3-7, 2019*, pp. 3980–3990. Association for Computational Linguistics, 2019.
 626 doi: 10.18653/V1/D19-1410. URL <https://doi.org/10.18653/v1/D19-1410>.

627 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien
 628 Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a
 629 benchmark. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=Ti67584b98>.

630 Z. Z. Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang,
 631 Zhe Fu, Qihao Zhu, Dejian Yang, Z. F. Wu, Zhibin Gou, Shirong Ma, Hongxuan Tang, Yuxuan
 632 Liu, Wenjun Gao, Daya Guo, and Chong Ruan. Deepseek-prover-v2: Advancing formal mathe-
 633 matical reasoning via reinforcement learning for subgoal decomposition. *CoRR*, abs/2504.21801,
 634 2025. URL <https://arxiv.org/abs/2504.21801>.

635 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li,
 636 Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open
 637 language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL
 638 <https://doi.org/10.48550/arXiv.2402.03300>.

639 Jiayi Sheng, Luna Lyu, Jikai Jin, Tony Xia, Alex Gu, James Zou, and Pan Lu. Solving inequality
 640 proofs with large language models. *CoRR*, abs/2506.07927, 2025. doi: 10.48550/ARXIV.2506.
 641 07927. URL <https://doi.org/10.48550/arXiv.2506.07927>.

648 Ling Team and Inclusion AI. Every activation boosted: Scaling general reasoner to 1 trillion open
 649 language foundation. *CoRR*, abs/2510.22115, 2025. doi: 10.48550/ARXIV.2510.22115. URL
 650 <https://doi.org/10.48550/arXiv.2510.22115>.

651

652 Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, Nov 2024. URL <https://qwenlm.github.io/blog/qwq-32b-preview/>.

653

654 Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Git-
 655 man. Openmathinstruct-2: Accelerating AI for math with massive open-source instruction data.
 656 In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore,*
 657 *April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=mTCbq2QssD>.

658

659 George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings,
 660 Amitayush Thakur, and Swarat Chaudhuri. Putnambench: Evaluating neural theorem-
 661 provers on the putnam mathematical competition. In Amir Globersons, Lester Mackey,
 662 Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.),
 663 *Advances in Neural Information Processing Systems 38: Annual Conference on Neural In-*
 664 *formation Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10*
 665 *- 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/1582eaf9e0cf349e1e5a6ee453100aa1-Abstract-Datasets_and_Benchmarks_Track.html.

666

667 Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. CMATH: can your language model
 668 pass chinese elementary school math test? *CoRR*, abs/2306.16636, 2023. doi: 10.48550/ARXIV.
 669 2306.16636. URL <https://doi.org/10.48550/arXiv.2306.16636>.

670

671 Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hanna Hajishirzi, Yejin Choi, and
 672 Kyunghyun Cho. Naturalproofs: Mathematical theorem proving in natural lan-
 673 guage. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neu-*
 674 *ral Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS*
 675 *Datasets and Benchmarks 2021, December 2021, virtual, 2021*. URL <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/d9d4f495e875a2e075a1a4a6e1b9770f-Abstract-round1.html>.

676

677 Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. Natural-
 678 prover: Grounded mathematical proof generation with language models. In Sanmi Koyejo,
 679 S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neu-*
 680 *ral Information Processing Systems 35: Annual Conference on Neural Information Process-*
 681 *ing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9,*
 682 *2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/1fc548a8243ad06616eee731e0572927-Abstract-Conference.html.

683

684 xAI. Grok: Artificial intelligence assistant, 2025. URL <https://x.ai>. Developed by xAI,
 685 accessed on March 25, 2025.

686

687 Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda
 688 Li, and Xiaodan Liang. Deepseek-prover: Advancing theorem proving in llms through large-
 689 scale synthetic data. *CoRR*, abs/2405.14333, 2024. doi: 10.48550/ARXIV.2405.14333. URL
 690 <https://doi.org/10.48550/arXiv.2405.14333>.

691

692 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,
 693 Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin
 694 Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang,
 695 Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu
 696 Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong
 697 Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115,
 698 2024. doi: 10.48550/ARXIV.2412.15115. URL <https://doi.org/10.48550/arXiv.2412.15115>.

699

700 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang
 701 Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng

702 Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang,
 703 Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu,
 704 Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men,
 705 Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren,
 706 Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang,
 707 Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu.
 708 Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL
 709 <https://doi.org/10.48550/arXiv.2505.09388>.

710 Xiang Yue, Tianyu Zheng, Ge Zhang, and Wenhui Chen. Mammoth2: Scaling instruc-
 711 tions from the web. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela
 712 Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neu-
 713 ral Information Processing Systems 38: Annual Conference on Neural Information Pro-
 714 cessing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15,
 715 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/a4ca07aa108036f80cbb5b82285fd4b1-Abstract-Conference.html.

716 Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao,
 717 Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi.
 718 Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*,
 719 abs/2309.01219, 2023. doi: 10.48550/ARXIV.2309.01219. URL <https://doi.org/10.48550/arXiv.2309.01219>.

720 Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu,
 721 Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical
 722 reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar
 723 (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria,
 724 July 27 - August 1, 2025*, pp. 10495–10516. Association for Computational Linguistics, 2025.
 725 URL <https://aclanthology.org/2025.findings-acl.547/>.

726 Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. minif2f: a cross-system benchmark
 727 for formal olympiad-level mathematics. In *The Tenth International Conference on Learning
 728 Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL
 729 <https://openreview.net/forum?id=9ZPegFuFTFv>.

730 Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied,
 731 Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foun-
 732 dation models. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Find-
 733 ings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico,
 734 June 16-21, 2024*, pp. 2299–2314. Association for Computational Linguistics, 2024. doi:
 735 10.18653/V1/2024.FINDINGS-NAACL.149. URL <https://doi.org/10.18653/v1/2024.findings-naacl.149>.

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756 A TOPIC DISTRIBUTION IN *DeepTheorem* DATASET
757
758
759760 Table 5: Subject and topic distribution in the *DeepTheorem* dataset.
761

762 Subject	763 Topic	764 Frequency
765 Algebra	Abstract Algebra	23134
	Linear Algebra	8858
	Intermediate Algebra	6634
	Matrices	3201
	Sequences And Series	440
	Vectors	383
	Prealgebra	281
	Algebraic Expressions And Inequalities	226
770 Calculus	Other	448
	Integral Calculus	11844
	Differential Calculus	8478
	Differential Equations	3478
	Limits	664
775 Applied Mathematics	Other	232
	Statistical Methods	13199
	Probability	632
	Math Word Problems	557
780 Geometry	Other	1583
	Differential Geometry	10951
	Plane Geometry	2044
	Solid Geometry	1082
	Non Euclidean Geometry	433
785 Discrete Mathematics	Other	504
	Logic	7710
	Combinatorics	3590
	Graph Theory	1790
	Algorithms	1318
790 Number Theory	Other	225
	Prime Numbers	1134
	Congruences	1032
	Factorization	318
794 Mathematical Analysis	Other	315
	Calculus	407
	Applied Mathematics	338
	Geometry	221
798 Other	Other	1746
	Topology	317
799	Other	1007

800
801
802
803
804
805
806
807
808
809

810
 811 **B EXAMPLES OF SAMPLE DIFFICULTIES AND DECONTAMINATED TRAINING**
 812 **CASES**

813
 814 **Table 6: Example questions in *DeepTheorem* dataset from different difficulty levels.**

815 Level	816 Example	817 Topic
818 5	819 Let A be a commutative ring with identity, and let $a, b \in A$ where a is in the Jacobson radical of A . Prove or disprove that the element $1 - ab$ is a unit in A , i.e., the principal ideal $(1 - ab)$ generated by $1 - ab$ is equal to A .	820 Abstract Algebra - 821 Ring Theory
822 6	823 Let M be a topological manifold with an open cover $\{U_i\}$ such that each U_i is diffeomorphic to an open subset of \mathbb{R}^n . Prove or disprove that if each U_i is diffeomorphic to the open unit ball in \mathbb{R}^n , then $\{U_i\}$ forms a contractible open cover of M .	824 Differential Geometry - Manifolds
825 7	826 Let \mathbb{N}^* denote the set of all positive integers. Suppose $a > 0$ is a real number, and there exist n mutually disjoint infinite subsets A_1, A_2, \dots, A_n such that $A_1 \cup A_2 \cup \dots \cup A_n = \mathbb{N}^*$, and for each $i = 1, 2, \dots, n$, and any $b, c \in A_i$, where $b > c$, the condition $b - c \geq a^i$ holds. Prove or disprove that a must satisfy $0 < a < 2$.	827 Intermediate Algebra - Polynomial 828 Operations
829 8	830 Let $f(z) = \frac{e^{iz}}{z}$, and consider the contour integral $\int_{\mathbb{R}} f(z) dz$ over the real line. Assume the contour is closed by a semicircular arc in the upper half-plane. Prove or disprove that the value of the improper integral $\int_0^{\infty} \frac{\sin x}{x} dx$ is given by: $\int_0^{\infty} \frac{\sin x}{x} dx = \frac{\pi}{2}$.	831 Integral Calculus - 832 Techniques of Integration
833 9	834 Let $F_3(a, b_1, b_2, c_1, c_2; x, y)$ denote Appell's hypergeometric function of two variables. For parameters $\alpha, \beta, \gamma, \nu, \rho \in \mathbb{C}$ such that $\text{Re}(\gamma) > 0$ and $\text{Re}(\nu - \gamma) > 0$, prove or disprove that the integral	835 Integral Calculus - 836 Multi Variable
$\mathcal{I}(\alpha, \beta, \gamma, z; \gamma, \nu, \rho, w) = \int_0^1 t^{\gamma-1} (1-t)^{\nu-\gamma-1} (1-zt)^{-\alpha} (1-wt)^{-\beta} dt$ <p>evaluates to</p> $\mathcal{I}(\alpha, \beta, \gamma, z; \gamma, \nu, \rho, w) = \frac{B(\gamma, \nu - \gamma)}{(1-w)^{\rho}} F_3 \left(\rho, \alpha, \nu - \gamma, \beta, \nu; \frac{w}{w-1}, z \right),$ <p>where $B(x, y)$ is the Beta function.</p>		

837 **Table 7: Examples of benchmark contamination in polynomial and number theory problems. Generalizing and logically equivalent parts are highlighted.**

838 Contaminated Example	839 Benchmark Example	840 Relationship
841 Let $p(x)$ be a univariate polynomial . Then $p(x)$ is nonnegative for all $x \in \mathbb{R}$ if and only if $p(x)$ can be expressed as a sum of squares (SOS), i.e., $p(x) = \sum_{i=1}^k q_i^2(x)$ for some polynomials $q_1(x), \dots, q_k(x)$.	842 Let $p(x)$ be a polynomial that is nonnegative for all real x . Prove that for some k , there are polynomials $f_1(x), \dots, f_k(x)$ such that $p(x) = \sum_{j=1}^k (f_j(x))^2$.	843 Identical
844 Let $p(x_1, x_2, \dots, x_n)$ be a real polynomial. If $p(x_1, x_2, \dots, x_n)$ is non-negative for all $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$, then $p(x_1, x_2, \dots, x_n)$ can be expressed as a sum of squares of polynomials if and only if p belongs to the quadratic module generated by the constraints of a certain semialgebraic set. Formally, there exists a set of polynomials $q_i(x_1, x_2, \dots, x_n)$ such that $p(x_1, x_2, \dots, x_n) = \sum_{i=1}^k q_i(x_1, x_2, \dots, x_n)^2$, provided certain conditions on p and the domain hold to ensure the SOS representation.	845 Let $p(x)$ be a polynomial that is nonnegative for all real x . Prove that for some k , there are polynomials $f_1(x), \dots, f_k(x)$ such that $p(x) = \sum_{j=1}^k (f_j(x))^2$.	846 Generalizing
847 Let n be a positive integer. If n is not prime, then $2^n - 1$ is not prime.	848 Show that if n is a positive integer and $2^n - 1$ is prime, then n is prime.	849 Logically Converse

864
865
866
867
868
C ADDITIONAL EXPERIMENTS869
870
871
872
873
874
875
876
877
Table 8: Performance comparison and generation length of Qwen3-4B trained on different data, all
sampled to 100K size.

Strategy	Data	FIMO		HMMT		Putnam		Avg.		Gen. Length
		out.	proc.	out.	proc.	out.	proc.	out.	proc.	
SFT	OpenR1-Proof	25.40	9.07	15.25	2.43	44.79	11.07	28.48	7.53	1572
	DeepTheorem _{easy}	30.16	10.07	22.03	9.10	36.46	13.30	29.55	10.82	1305
	DeepTheorem _{hard}	30.16	13.09	15.25	4.35	40.62	16.08	28.68	11.17	1443
	DeepTheorem	28.57	11.83	15.25	3.68	39.58	15.44	27.80	10.31	1266
RL	OpenR1-Proof	41.27	14.16	25.42	8.32	55.21	23.94	40.63	15.47	1197
	DeepTheorem _{easy}	44.44	20.52	23.73	8.73	51.04	30.57	39.74	19.94	1072
	DeepTheorem _{hard}	47.62	22.32	27.12	9.78	54.17	27.62	42.97	19.91	1003
	DeepTheorem	46.03	23.86	23.73	8.84	50.00	28.39	39.92	20.36	1022

878
879
Table 9: Results of Qwen3-4B trained on the full *DeepTheorem* dataset.

Strategy	FIMO		HMMT		Putnam		Avg.		Gen. Length
	out.	proc.	out.	proc.	out.	proc.	out.	proc.	
SFT	34.92	12.28	18.64	7.06	35.42	15.09	29.66	11.48	1151
RL	49.21	23.58	27.12	11.55	54.17	28.00	43.50	21.04	1122

880
881
882
883
In this section, we provide additional training results using Qwen3-4B (Yang et al., 2025) as the
backbone. To further analyze the impact of data difficulty, we train four sets of models on different
data mixtures:884
885
886
887
- OpenR1-proof: the proof subset in OpenR1-Math;
888
- *DeepTheorem*-easy: samples in *DeepTheorem* that have a difficulty level from 5 to 7;
889
- *DeepTheorem*-hard: samples in *DeepTheorem* that have a difficulty level from 7 to 9;
890
- *DeepTheorem*: samples from all difficulty levels in *DeepTheorem*.891
892
893
894
895
896
897
898
For all of these four mixtures, we sample the data to 100K theorem variants, ensuring a fair com-
parison, and the results are presented in Table 8. In Table 9, we also present the results of training
Qwen3-4B on the full-scale *DeepTheorem*, i.e. 121K theorems and 242K variants. The training
configurations are consistent with the experiments in Section 4 (i.e. 3 epochs for SFT and 1 epoch
for RL), which takes 11 hours on 8 A100-80G GPUs for SFT, and 80 hours on 32 A100-80G GPUs
for RL.899
900
Analyzing these results, we find that901
902
903
- **SFT is more effective for the (relatively) easy data, while RL is more effective for the hard
data.** However, simply combining the easy and hard data does not lead to better results in both
scenarios.904
905
906
907
- **Scaling data size improves model performance for both SFT and RL.** Comparing Table 8 and
9, we find that utilizing the whole *DeepTheorem* dataset can mitigate the impact of mixing different
data difficulties and lead to better performance, highlighting the value of *DeepTheorem* as a large-
scale training dataset.908
909
910
911
- **There is no significant correlation between performance and generation lengths.** When train-
ing on the sampled data (Table 7), SFT models consistently exhibit longer generation lengths than
RL models regardless of the data mixture, while training on the full data leads to similar generation
lengths.912
913
914
915
916
917

918 **D COMPARISON OF PROCESS EVALUATION JUDGES**
919920 **D.1 COMPARISON BETWEEN GPT-4O AND O3-MIN**
921922 In Table 10, we present the comparison between using GPT-4o, o3-mini, and Ling-1T (Team &
923 AI, 2025) as judges for process evaluation. While o3-mini is stricter and gives lower scores on
924 average, the relative ranks of all evaluated models are similar. The scores between different judges
925 also exhibit a high degree of correlation, with the following correlation coefficients: 0.955 between
926 GPT-4o and o3-mini, 0.987 between GPT-4o and Ling-1T, and 0.987 between o3-mini and Ling-1T,
927 highlighting strong consistency.928 In the last two columns of Table 10, we also provide the scores given by Qwen2.5-Math-PRM-
929 72B (Zhang et al., 2025), a process reward model specifically trained for scoring mathematical
930 reasoning steps. The scores yield a correlation coefficient of 0.996 with GPT-4o, 0.948 with o3-
931 mini, and 0.982 with Ling-1T, also exhibiting high consistency.932
933 Table 10: Comparison of process evaluation scores using three different judges and Qwen2.5-Math-
934 PRM-72B.

Model	GPT-4o		o3-mini		Ling-1T		Q2.5-PRM-72B	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank
Claude3.7-Sonnet	24.30	9	15.54	8	21.05	9	25.38	9
Gemini2.5-Pro	54.21	3	53.98	2	56.42	2	50.87	2
Gemini2.5-Flash	29.98	6	29.42	5	31.02	5	28.29	7
GPT-4o	21.39	10	14.29	12	16.90	11	20.32	11
o1-mini	46.34	4	36.02	4	42.34	4	42.17	4
o1	55.28	2	53.52	3	54.02	3	50.37	3
o3-mini	65.40	1	65.57	1	65.72	1	58.25	1
Qwen2.5-Inst-7B	17.78	12	8.03	15	14.08	14	16.95	14
Qwen2.5-Inst-72B	29.38	7	15.31	9	23.97	7	29.76	6
Qwen2.5-Math-Inst-7B	13.09	17	5.17	16	10.84	16	12.26	17
Qwen2.5-Math-Inst-72B	27.56	8	15.08	10	23.71	8	26.79	8
DS-Prover-v1.5-RL-7B	14.18	16	4.34	17	8.78	17	15.10	16
DS-Prover-v2-7B	17.37	13	8.36	14	12.95	15	18.80	12
R1-Distill-7B	2.28	18	1.90	18	2.38	18	2.58	18
R1-Distill-70B	16.89	14	16.46	6	18.40	10	18.53	13
QwQ-32B	14.57	15	14.89	11	16.32	12	15.86	15
Llama3.3-Inst-70B	18.92	11	10.05	13	14.13	13	21.45	10
DeepTheorem-RL-7B	34.04	5	15.95	7	27.03	6	31.74	5

955
956 **D.2 HUMAN EVALUATION**
957958 To further validate our LLM-based process evaluation frameworks, we also conduct human evalua-
959 tions on a subset of the model outputs. Due to the mathematical difficulty of the questions involved,
960 we sampled 100 questions and their corresponding proofs generated by ten models, covering both
961 API-based and open-source models, as well as reasoning-specialized and general-purpose models.962 We recruited 12 third-party expert annotators (all graduate students majoring in mathematics) to
963 conduct pairwise evaluations of the model responses, and the win rates are presented in Table 11.
964 The scores from o3-mini in Table 10 are also converted into pairwise binary scores by comparing
965 the scores assigned to responses from two models. The results suggest that the evaluation of the
966 LLM judge aligns well with human experts, resulting in consistent rankings.

972
 973
 974
 975
 976
 977
 978
 979
 980
 981
 982
 983
 984
 985
 986
 987
 988
 989
 990
 991
 992

993 Table 11: Comparison of pairwise evaluations from human experts and o3-mini.

Model	Human Evaluation				o3-mini Evaluation			
	Win	Tie	Lose	Rank	Win	Tie	Lose	Rank
R1-Distill-70B	0.217	0.450	0.333	5	0.600	0.100	0.300	4
Qwen2.5-Math-72B-Inst	0.306	0.486	0.208	4	0.417	0.000	0.583	6
Deepseek-Prover-V2-7B	0.021	0.646	0.333	10	0.062	0.000	0.938	10
DeepTheorem-7B	0.032	0.460	0.508	9	0.143	0.048	0.810	9
Gemini-2.5-Pro	0.722	0.204	0.074	1	0.889	0.111	0.000	1
GPT-4o	0.140	0.456	0.404	6	0.474	0.053	0.474	5
o1-mini	0.526	0.368	0.105	3	0.632	0.105	0.263	3
o3-mini	0.685	0.315	0.000	2	0.889	0.111	0.000	1
Qwen2.5-72B-Inst	0.079	0.587	0.333	8	0.333	0.095	0.571	7
QwQ-32B	0.125	0.472	0.403	7	0.333	0.000	0.667	8

1006
 1007
 1008
 1009
 1010
 1011
 1012
 1013
 1014
 1015
 1016
 1017
 1018
 1019
 1020
 1021
 1022
 1023
 1024
 1025

1026 **E CASE STUDIES ON MODEL-GENERATED PROOF**
10271028 **E.1 EXAMPLE FROM DEEPTHEOREM-RL-7B**
10291030 Model-generated proof
10311032 **Question:**1033 Let P be a polynomial of degree $n > 1$ with integer coefficients and let k be any positive
1034 integer. Consider the polynomial $Q(x) = P(P(\dots P(P(x))\dots))$, with k pairs of parentheses.
1035 Prove or disprove that Q has more than n integer fixed points, i.e. integers satisfying
1036 the equation $Q(x) = x$.
10371038 **Model Solution:**
1039

1040 <think>

1041 To determine whether the polynomial $Q(x) = P(P(\dots P(P(x))\dots))$ with k pairs of
1042 parentheses has more than n integer fixed points, we need to analyze the properties of the
1043 polynomial P and how the composition of P affects its fixed points.
10441045 First, recall that $P(x)$ is a polynomial of degree $n > 1$ with integer coefficients. The
1046 number of integer roots of $P(x) - x = 0$ (i.e., the number of integer fixed points of P) is at
1047 most n by the fundamental theorem of algebra and properties of polynomials. Let's denote
1048 the integer fixed points of P by r_1, r_2, \dots, r_m , where $m \leq n$.
10491050 Now, consider the polynomial $Q(x) = P(P(\dots P(P(x))\dots))$ with k pairs of parentheses.
1051 We need to determine the number of integer solutions to the equation $Q(x) = x$. If r is an
1052 integer fixed point of Q , then r must be an integer fixed point of some composition of P .
1053 Specifically, if r is a fixed point of Q , there exists some $m \leq k$ such that $P^{(m)}(r) = r$,
1054 where $P^{(m)}$ denotes the m -th iterate of P .
10551056 The key observation is that the number of integer fixed points of Q is bounded by the
1057 number of integer preperiodic points of P . Since P is a polynomial of degree n , the number
1058 of integer preperiodic points of P is at most n for each iterate. However, we need to
1059 consider the composition Q which is the k -th iterate of P .
10601061 The polynomial $Q(x)$ is also of degree n^k , and the number of integer roots of $Q(x) - x = 0$
1062 is at most n^k . However, we need to check if it is possible for $Q(x)$ to have more than n
1063 integer fixed points. Since each application of P can introduce at most n new fixed points,
1064 the total number of integer fixed points of Q is still bounded by the number of preperiodic
1065 points of P , which is at most n for each iterate.
10661067 To be more precise, let's consider the fixed points. If r is a fixed point of Q , then r must be
1068 a root of the polynomial $P^k(x) - x$. The polynomial $P^k(x) - x$ is of degree n^k , and by the
1069 fundamental theorem of algebra, it can have at most n^k roots. However, we are interested in
1070 integer roots, and the number of integer roots is at most the number of integer preperiodic
1071 points of P , which is at most n for each iterate.
10721073 Therefore, the number of integer fixed points of $Q(x)$ is at most n , the same as the number
1074 of integer fixed points of P for each iterate. Hence, $Q(x)$ cannot have more than n integer
1075 fixed points.
10761077 Thus, the statement that Q has more than n integer fixed points is false.
1078

</think>

<answer> **disproved** </answer>1079 This is a non-cherry-picked example from our *DeepTheorem-RL-7B* model. We analyze this proof
1080 in the following dimensions:

1080 • **Correctness.** The model accurately identifies that integer fixed points of $Q(x)$ satisfy $Q(x) = x$,
 1081 and recognizes that $Q(x)$ has degree n^k , with $Q(x) - x = 0$ having at most n^k roots. Its insight
 1082 that the number of integer fixed points is constrained by the fixed points of P , which are at most
 1083 n , effectively guides the solution. The model leverages the iterative structure of $Q(x)$ to argue
 1084 that additional compositions do not increase the number of integer fixed points beyond n , aligning
 1085 with expected polynomial behavior.

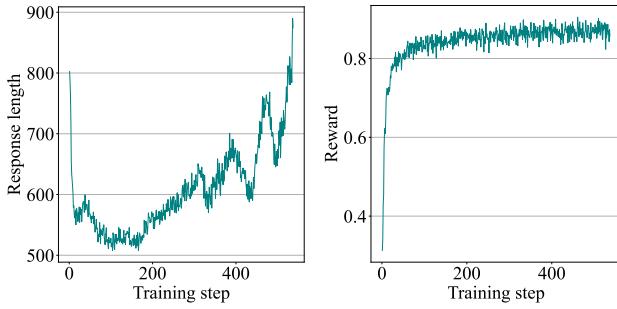
1086 • **Clarity.** The proof is well-structured, progressing logically from problem definition to conclu-
 1087 sion. Mathematical notation, such as $P^{(m)}$ for the m -th iterate, is precise and standard, enhancing
 1088 readability. The solution is concise, making it accessible and suitable for a mathematical audi-
 1089 ence.

1090 • **Strengths.** The model’s ability to distill a complex problem into a clear argument showcases its
 1091 strength in handling polynomial compositions. Its focus on fixed points and iterative properties
 1092 demonstrates a solid grasp of the underlying mathematics, making it a valuable tool for such
 1093 proofs.

1094 In summary, the model delivers a clear and correct disproof, highlighting its efficacy in tackling
 1095 advanced mathematical problems with precision and clarity.

1097 E.2 COMPARISON BETWEEN SFT AND RL

1099 In this section, we further analyze
 1100 the training dynamics of LLMs on
 1101 *DeepTheorem*, using the more recent
 1102 Qwen3-4B backbone as an example.
 1103 When training on *DeepTheorem* with
 1104 RL-Zero, we find that the average
 1105 response length quickly drops by
 1106 more than two hundred tokens during
 1107 the first few training steps, and the
 1108 gradually increases over the training
 1109 course, as shown in Figure 8. Observ-
 1110 ing that the initial drop in response
 1111 length overlaps with the sharp rise
 1112 of reward from less than 0.4 to more
 1113 than 0.7, we hypothesize that this is a
 1114 result of the model grasping at the task
 1115 definition and trying to output minimal
 1116 (often incorrect) answers to receive any
 1117 reward signal. In contrast, in the later
 1118 stage, more detailed reasoning is required
 1119 to further increase the reward.



1120 Figure 8: Response length (left) and reward (right) over the
 1121 RL training course of Qwen3-4B on *DeepTheorem*.

1122 To gain further insights into how models learn to master theorem proving, we analyzed responses
 1123 from one checkpoint near the start of the training (step 25), and one near the end of the training (step
 1124 525), along with the responses from the model trained with SFT. We find that the most common error
 1125 type from all model checkpoints is overgeneralizing from special cases, as shown in the examples
 1126 below, where the error is identified by the LLM judge for process evaluation.

1127 Furthermore, the overgeneralizing issue is most prominent in the early stage of RL training and
 1128 is mitigated as the training proceeds, which may explain the phenomena that generation length is
 1129 shorter in the beginning and gradually increases. In the response from checkpoint 25, a negative
 1130 conclusion is drawn directly after deriving a_1, a_2, a_3 in a special case, entirely ignoring the concept
 1131 of “for i sufficiently large” in the question. The response from checkpoint 525, on the other hand,
 1132 makes a more subtle over-generalizing statement: concluding that $a = 0$ is the only fixed point from
 1133 derivations starting from the premise that a is an integer, omitting the case where a is not an integer.

1134 Comparison of model-generated proofs

1135 Question:

1136 A sequence of real numbers a_0, a_1, a_2, \dots is defined by the formula

$$1137 \quad a_{i+1} = \lfloor a_i \rfloor \cdot \langle a_i \rangle \quad \text{for } i \geq 0;$$

1134 here a_0 is an arbitrary real number, $\lfloor a_i \rfloor$ denotes the greatest integer not exceeding a_i , and
 1135 $\langle a_i \rangle = a_i - \lfloor a_i \rfloor$. Prove or disprove that $a_i = a_{i+2}$ for i sufficiently large.
 1136

1137 **Answer:** True

1139 **Response from RL checkpoint 25:**

1140 To determine whether the statement is true or false, we need to analyze the given sequence
 1141 defined by the formula:

$$1142 \quad a_{i+1} = \lfloor a_i \rfloor \cdot \langle a_i \rangle$$

1143 where $\lfloor a_i \rfloor$ is the greatest integer not exceeding a_i , and $\langle a_i \rangle = a_i - \lfloor a_i \rfloor$ is the fractional
 1144 part of a_i .

1145 Let's denote $a_i = n_i + f_i$, where $n_i = \lfloor a_i \rfloor$ is the integer part and $f_i = \langle a_i \rangle$ is the fractional
 1146 part. Then the sequence can be rewritten as:

$$1148 \quad a_{i+1} = n_i \cdot f_i$$

1149 We need to check if the sequence eventually becomes periodic with period 2, i.e., if $a_i = a_{i+2}$ for
 1150 sufficiently large i .

1151 Consider the initial value $a_0 = n_0 + f_0$. The next term in the sequence is:

$$1153 \quad a_1 = n_0 \cdot f_0$$

1154 The integer part of a_1 is $\lfloor n_0 \cdot f_0 \rfloor$ and the fractional part is $\langle n_0 \cdot f_0 \rangle = n_0 \cdot f_0 - \lfloor n_0 \cdot f_0 \rfloor$.
 1155 Therefore, the next term in the sequence is:

$$1157 \quad a_2 = \lfloor n_0 \cdot f_0 \rfloor \cdot (n_0 \cdot f_0 - \lfloor n_0 \cdot f_0 \rfloor)$$

1158 To determine if $a_i = a_{i+2}$ for sufficiently large i , we need to check if the sequence can enter
 1159 a cycle. Let's consider some specific cases:

1160 1. If $f_0 = 0$, then $a_1 = 0$ and the sequence becomes $0, 0, 0, \dots$, which is periodic with
 1161 period 1. 2. If $f_0 = 1$, then $a_1 = n_0$ and the sequence becomes $n_0, 0, 0, \dots$, which is
 1162 periodic with period 1. 3. If $f_0 = \frac{1}{2}$ and $n_0 = 2$, then $a_1 = 1$ and $a_2 = 0$, and the sequence
 1163 becomes $2.5, 1, 0, 0, \dots$, which is periodic with period 2.

1164 From these examples, we see that the sequence can indeed become periodic with period 2 for
 1165 certain initial values. However, it is not necessarily true for all initial values. For example,
 1166 if $a_0 = 1.5$, then $a_1 = 1 \cdot 0.5 = 0.5$, $a_2 = 0 \cdot 0.5 = 0$, and $a_3 = 0 \cdot 0 = 0$, which is periodic
 1167 with period 1.

1168 Therefore, the statement that $a_i = a_{i+2}$ for i sufficiently large is not always true. It depends
 1169 on the initial value a_0 . Hence, the statement is disproved.

1170 <answer> *disproved* </answer>

1171 **Process evaluation (Ling-1T):**

1172 - Score: 0.35

1173 - Explanation: The proof is logically flawed because it incorrectly concludes the statement
 1174 is false based on limited examples, while the ground truth is 'True'. The reasoning does not
 1175 account for the eventual behavior of the sequence for all real starting values, missing key
 1176 cases and convergence analysis. The structure and writing are clear, but completeness and
 1177 correctness are severely lacking, and validity is compromised by hasty generalization from
 1178 examples to a universal disproof.

1179 **Response from RL checkpoint 525:**

1180 To determine whether $a_i = a_{i+2}$ for i sufficiently large in the sequence defined by $a_{i+1} =$
 1181 $\lfloor a_i \rfloor \cdot \langle a_i \rangle$, we need to analyze the given recurrence relation step by step.

1182 Given:

$$1183 \quad a_{i+1} = \lfloor a_i \rfloor \cdot \langle a_i \rangle$$

1184 where $\lfloor a_i \rfloor$ is the greatest integer not exceeding a_i , and $\langle a_i \rangle = a_i - \lfloor a_i \rfloor$ is the fractional
 1185 part of a_i .

1186 Let's denote $\lfloor a_i \rfloor$ by n_i and $\langle a_i \rangle$ by f_i . Then the recurrence relation can be rewritten as:

$$1187 \quad a_{i+1} = n_i \cdot f_i.$$

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

We need to investigate the behavior of the sequence a_i for large i . Let's consider the possible values of a_i :

1. If a_i is an integer, then $\langle a_i \rangle = 0$, and thus $a_{i+1} = n_i \cdot 0 = 0$. 2. If a_i is a non-integer, then $0 < f_i < 1$, and $a_{i+1} = n_i \cdot f_i$.

Let's analyze the behavior of the sequence in more detail. Suppose a_i is a non-integer. Then $a_{i+1} = n_i \cdot f_i$, and we need to consider the next term a_{i+2} :

$$a_{i+2} = \lfloor a_{i+1} \rfloor \cdot \langle a_{i+1} \rangle.$$

If $a_{i+1} = n_i \cdot f_i$, then:

$$\lfloor a_{i+1} \rfloor = \lfloor n_i \cdot f_i \rfloor,$$

and

$$\langle a_{i+1} \rangle = a_{i+1} - \lfloor a_{i+1} \rfloor = n_i \cdot f_i - \lfloor n_i \cdot f_i \rfloor.$$

To determine if $a_i = a_{i+2}$ for large i , we need to see if the sequence eventually becomes periodic or stabilizes. Let's consider some specific cases:

- If a_0 is an integer, then $a_1 = 0$, $a_2 = 0$, and so on. In this case, $a_i = 0$ for all $i \geq 1$, and clearly $a_i = a_{i+2}$ for all $i \geq 1$. - If a_0 is a non-integer, we need to see if the sequence can stabilize or become periodic. Let's consider $a_0 = n_0 + f_0$ where $0 < f_0 < 1$. Then:

$$a_1 = n_0 \cdot f_0.$$

The value of a_1 depends on n_0 and f_0 . If $n_0 \cdot f_0$ is an integer, then $a_2 = 0$, and the sequence becomes $0, 0, 0, \dots$. If $n_0 \cdot f_0$ is a non-integer, we need to check the next term:

$$a_2 = \lfloor n_0 \cdot f_0 \rfloor \cdot \langle n_0 \cdot f_0 \rangle.$$

To determine if $a_i = a_{i+2}$ for large i , we need to check if the sequence can enter a periodic cycle. Let's consider the possibility of a fixed point. A fixed point a would satisfy:

$$a = \lfloor a \rfloor \cdot \langle a \rangle.$$

If a is an integer, then $\langle a \rangle = 0$, so $a = 0$. If a is a non-integer, then $0 < \langle a \rangle < 1$, and $a = n \cdot f$ where $n = \lfloor a \rfloor$ and $f = \langle a \rangle$. For a to be a fixed point, we need:

$$a = n \cdot f.$$

Since $a = n + f - \lfloor n + f \rfloor = f$ if n is an integer, the only solution is $a = 0$.

Given that the only fixed point is $a = 0$ and the sequence can either become $0, 0, 0, \dots$ or enter a periodic cycle that includes 0, we need to check if there are any other possible periodic cycles. However, any non-zero non-integer a_i will eventually lead to a value that is either an integer or a value that, when multiplied by its floor, results in a value that leads back to a previous term in the sequence.

By analyzing the behavior, we see that for large i , the sequence either stabilizes at 0 or enters a periodic cycle that includes 0. In either case, $a_i = a_{i+2}$ for sufficiently large i .

Since the sequence either becomes $0, 0, 0, \dots$ or enters a periodic cycle that includes 0, we can conclude that $a_i = a_{i+2}$ for sufficiently large i .

<answer> **proved** </answer>

Process evaluation (Ling-1T):

- Score: 0.85

- Explanation: The proof correctly identifies that sequences eventually reach 0 or enter a cycle involving 0, leading to period 2 behavior for large i . Logical steps are mostly valid, though the argument for non-integer cases collapsing into cycles or zero could be more rigorous. The conclusion is correct, and the writing is clear, but a few more case distinctions or bounds would improve completeness.

Response from SFT checkpoint:

To prove this theorem, we can start by showing that for any real number a written in the form

$$a = n + \delta, \quad \text{with } n = \lfloor a \rfloor \in \mathbb{Z} \quad \text{and} \quad \delta = \langle a \rangle \in [0, 1),$$

the recurrence

$$a_{i+1} = \lfloor a_i \rfloor \cdot \langle a_i \rangle$$

1242 leads to a sequence that eventually becomes periodic with period 2.
 1243 1. **Case 1: $a = n$ is an integer.**
 1244 If $a = n$ (with $n \in \mathbb{Z}$), then

$$\langle a \rangle = 0.$$

1245 Hence, the recurrence gives
 1246

$$a_1 = \lfloor a \rfloor \cdot \langle a \rangle = n \cdot 0 = 0.$$

1247 Since 0 is an integer, we have
 1248

$$\langle 0 \rangle = 0,$$

1249 and therefore
 1250

$$a_2 = \lfloor 0 \rfloor \cdot \langle 0 \rangle = 0 \cdot 0 = 0.$$

1251 Thus, for all $i \geq 1$,
 1252

$$a_i = 0,$$

1253 and in particular, $a_i = a_{i+2}$ for all i .
 1254 2. **Case 2: $a = n + \delta$ with $0 < \delta < 1$.**

1255 In this case, we have
 1256

$$a_1 = \lfloor a \rfloor \cdot \langle a \rangle = n \cdot \delta.$$

1257 Notice that since $0 < \delta < 1$, the product $n \cdot \delta$ satisfies
 1258

$$0 < n \cdot \delta < n.$$

1259 Thus, a_1 is a positive number less than n (and in particular, it is not an integer because $n \cdot \delta$
 1260 is not an integer when $0 < \delta < 1$). Therefore, we can write
 1261

$$a_1 = n_1 + \delta_1, \quad \text{with } n_1 = \lfloor a_1 \rfloor \quad \text{and} \quad \delta_1 = \langle a_1 \rangle.$$

1262 Then, the recurrence gives
 1263

$$a_2 = \lfloor a_1 \rfloor \cdot \langle a_1 \rangle = n_1 \cdot \delta_1.$$

1264 Now, observe that
 1265

$$a_2 = n_1 \cdot \delta_1 = \lfloor a_1 \rfloor \cdot \langle a_1 \rangle = \lfloor a_1 \rfloor \cdot \langle a_1 \rangle = a_1.$$

1266 Hence, we have
 1267

$$a_2 = a_1.$$

1268 By the recurrence, it follows that for all $i \geq 1$,
 1269

$$a_{i+2} = a_i.$$

1270 3. **Conclusion:**
 1271 In both cases, we have shown that for i sufficiently large (in fact, for all $i \geq 1$), the sequence
 1272 satisfies
 1273

$$a_i = a_{i+2}.$$

1274 This completes the proof.
 1275

1276 proved

1277 **Process evaluation (Ling-1T):**

1278 - Score: 0.55

1279 - Explanation: The proof contains a critical logical flaw in Case 2: it incorrectly asserts that
 1280 $a_2 = a_1$, which is generally false (e.g., $a_0 = 2.5 \rightarrow a_1 = 2*0.5 = 1.0, a_2 = 1*0 = 0$). This
 1281 invalidates the reasoning for non-integer starting values. However, the conclusion is correct
 1282 (the sequence eventually becomes periodic with period 2), and the structure and clarity are
 1283 good. Completeness is partial because it addresses integer and non-integer cases but fails to
 1284 analyze the full dynamics (e.g., negative numbers, behavior when $a_i \in [0, 1)$, or when $n \cdot \delta$
 1285 is integer). Validity is low due to the false step; correctness is full since the final claim is
 1286 true; clarity is high.

1287 Apart from overgeneralizing, we also find that models trained with SFT are prone to hallucination,
 1288 i.e., stating or deriving a statement that clearly contradicts conditions given in the question, as in the
 1289

1296 example below. In this example, the model tries to prove the statement by starting from a special
 1297 case $n = 2$, but hallucinates an additional term $\frac{1}{a+b}$ on the right-hand side, and also omits the
 1298 n in the numerator. Such hallucination is much rarer in the RL models, corroborating previous
 1299 findings (Zhang et al., 2023).

1300

1301 Example of hallucination from the SFT model

1302

1303

Question:

1304

Prove or disprove the inequality

1305

1306

1307

$$\sum_{i < j} \frac{a_i a_j}{a_i + a_j} \leq \frac{n}{2(a_1 + a_2 + \dots + a_n)} \sum_{i < j} a_i a_j$$

1308

for positive real numbers a_1, a_2, \dots, a_n .

1309

1310

Answer: True

1311

Response from SFT checkpoint:

1312

To prove this theorem, we can start by showing that for any two positive real numbers a and
 1313 b the following inequality holds:

1314

1315

1316

$$\frac{ab}{a+b} \leq \frac{1}{2} \frac{ab}{a+b} \cdot \frac{1}{a+b}.$$

1317

However, this is not a valid inequality because it simplifies to:

1318

1319

1320

$$\frac{ab}{a+b} \leq \frac{ab}{(a+b)^2},$$

1321

which is false since $(a+b)^2 > a+b$ for positive a and b .

1322

Thus, the inequality

1323

1324

$$\frac{ab}{a+b} \leq \frac{1}{2} \frac{ab}{a+b} \cdot \frac{1}{a+b}$$

1325

is disproved.

1326

Since the inequality does not hold for any two positive real numbers, it cannot hold for all
 1327 pairs (a_i, a_j) for $1 \leq i < j \leq n$. Therefore, the original inequality

1328

1329

1330

$$\sum_{i < j} \frac{a_i a_j}{a_i + a_j} \leq \frac{n}{2(a_1 + a_2 + \dots + a_n)} \sum_{i < j} a_i a_j$$

1331

is also disproved.

1332

1333

disproved

1334

Process evaluation (Ling-1T):

1335

- Score: 0.1

1336

- Explanation: The proof contains a fundamental logical error: the claimed starting inequality
 1337 is incorrectly stated and analyzed, and the attempt to disprove a true statement fails. The
 1338 reasoning is invalid and incomplete, and the conclusion is incorrect. However, the writing is
 1339 clear and easy to follow.

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

1350 F DeepTheorem PROMPT TEMPLATE
13511352 F.1 CONTAMINATION JUSTIFICATION
13531354 Meta Prompt for Contamination Justification
13551356 I will now give you two theorems: the Original Theorem and the Candidate Theorem. Please
1357 help me determine if the following two theorems are the same.

1358 Original Theorem: ***theorem1***

1359 Candidate Theorem: ***theorem2***

1360 Disregard the names and minor changes in word order. If their theorem prompts are very
1361 similar, without considering the proving process, we consider them to be the same theorem.

1362 Note that you should not consider the solution process, only the theorem prompts.

1363 You should only respond with True or False. Do not respond with anything else.

1364
1365 F.2 THEOREM-PROVING ANNOTATION
13661367 Meta Prompt for Theorem-Proving Annotation
13681369 You are an expert in classifying questions based on their type and intent. Given the following
1370 discussion:

1371 # Discussion

1372 - problem: ***problem***
1373 - solution: ***solution***

1374 Determine whether the question is:

1375 - A question-answering (QA) question seeking a specific value or factual response, or
1376 - A theorem-proving question requiring logical reasoning, derivation, or proof of a mathematical or theoretical statement.

1377 Provide a clear classification (QA or theorem-proving) and justify your decision with a concise explanation. Consider the following:

1378 - QA questions typically ask for a specific fact, value, or definitive answer (e.g., "What is the capital of France?" or "What is the value of x in $2x = 8$?").
1379 - Theorem-proving questions typically involve logical reasoning, mathematical derivation, or proving a general statement (e.g., "Prove that the sum of two even numbers is even" or "Derive the Pythagorean theorem").

1380 Return "True" if the question is a theorem-proving question, and "False" if it is a QA question.

1381
1382 F.3 RATIONALE SUMMARIZATION
13831384 Meta Prompt for Rationale Summarization
13851386 You are provided with a corpus of forum discussions about mathematical topics. Your task
1387 is to analyze the discussion and:1388 1. Identify the key mathematical concepts, ideas, or rationales driving the discussion.
1389 2. Act as a teacher to formulate a theorem based on the discussion, presented as a formal
1390 theorem statement.

1391 # Requirements

1392 - All mathematical equations must be formatted in LaTeX.
1393 - The theorem should be a clear, formal statement (e.g., "Let $f : \mathbb{R} \rightarrow \mathbb{C}$ be a smooth
1394 function, ...").
1395 - The output must be in JSON format, with the following structure:1396 {
1397 "rationale": "A description of the main mathematical concepts or ideas in the discussion.",
1398 "theorem": "A formal theorem statement based on the discussion."
1399 }
14001401 # Discussion
1402
1403

1404
 1405 - problem: ***problem***
 1406 - solution: ***solution***
 1407
 1408
 1409

1410 F.4 PROOF GENERATION

1411
 1412
 1413 Meta Prompt for Proof Generation from o3-mini (high)
 1414
 1415 You are provided with a corpus of forum discussions about mathematical topics.
 1416 A theorem statement is summarized from the discussion. Your task is to provide a proof for
 1417 the theorem statement based on the discussion.
 1418 # Requirements
 1419 - All mathematical equations must be formatted in LaTeX.
 1420 - The proof should be a clear, formal statement (e.g., "To prove this theorem, we can start by
 1421 ...").
 1422 - The output must be in JSON format, with the following structure:
 1423 {
 1424 "proof": "A proof for the theorem statement."
 1425 }
 1426 # Discussion
 1427 - problem: ***problem***
 1428 - solution: ***solution***
 1429 # Theorem Rationale
 1430 ***theorem***
 1431
 1432
 1433

1434 F.5 LOGICAL VALIDATION

1435
 1436
 1437 Meta Prompt for Logical Validation of the Proof
 1438
 1439 You are an expert in mathematical theorem proving and logical analysis. Given the following
 1440 theorem and its proof or disproof, your task is to analyze each step of the proof or disproof
 1441 to determine if it is valid, providing a detailed justification for each step's correctness or
 1442 identifying any errors.
 1443 # Theorem
 1444 ***theorem***
 1445 # Proof or Disproof
 1446 ***Proof***
 1447 # Instructions
 1448 1. **Analyze Each Step**:
 1449 - Verify if the step is mathematically correct, logically sound, and relevant to proving or
 1450 disproving the theorem.
 1451 - Check for adherence to mathematical definitions, theorems, or properties cited in the step.
 1452 - Ensure the step follows from previous steps or given assumptions without logical gaps.
 1453 - If the step involves a disproof, confirm that it correctly demonstrates a counterexample or
 1454 contradiction.
 1455 2. **Overall Assessment**:
 1456 - Conclude whether the entire proof or disproof is valid.
 1457 - If invalid, return False and summarize the critical errors and recommend how to fix the
 1458 proof/disproof.
 1459 - If valid, return True and confirm that it fully addresses the theorem.

1458

1459 F.6 DIFFICULTY ANNOTATION

1460

1461 Meta Prompt for Difficulty Annotation

1462

CONTEXT

1463

I am a teacher, and I have some high-level olympiad math problems.

1464

I want to evaluate the difficulty of these math problems. There are some references available regarding the difficulty of the problems:

1465

1466

<difficulty reference>

1467

Examples for difficulty levels For reference, here are problems from each of the difficulty levels 1-10:

1468

1: How many integer values of x satisfy $|x| < 3\pi$? (2021 Spring AMC 10B, Problem 1)

1469

1.5: A number is called flippy if its digits alternate between two distinct digits. For example, 2020 and 37373 are flippy, but 3883 and 123123 are not. How many five-digit flippy numbers are divisible by 15? (2020 AMC 8, Problem 19)

1470

2: A fair 6-sided die is repeatedly rolled until an odd number appears. What is the probability that every even number appears at least once before the first occurrence of an odd number? (2021 Spring AMC 10B, Problem 18)

1471

2.5: A , B , C are three piles of rocks. The mean weight of the rocks in A is 40 pounds, the mean weight of the rocks in B is 50 pounds, the mean weight of the rocks in the combined piles A and B is 43 pounds, and the mean weight of the rocks in the combined piles A and C is 44 pounds. What is the greatest possible integer value for the mean in pounds of the rocks in the combined piles B and C ? (2013 AMC 12A, Problem 16)

1472

3: Triangle ABC with $AB = 50$ and $AC = 10$ has area 120. Let D be the midpoint of \overline{AB} , and let E be the midpoint of \overline{AC} . The angle bisector of $\angle BAC$ intersects \overline{DE} and \overline{BC} at F and G , respectively. What is the area of quadrilateral $FDBG$? (2018 AMC 10A, Problem 24)

1473

3.5: Find the number of integer values of k in the closed interval $[-500, 500]$ for which the equation $\log(kx) = 2\log(x+2)$ has exactly one real solution. (2017 AIME II, Problem 7)

1474

4: Define a sequence recursively by $x_0 = 5$ and

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

1486

1487

$$x_{n+1} = \frac{x_n^2 + 5x_n + 4}{x_n + 6}$$

1488

for all nonnegative integers n . Let m be the least positive integer such that

1489

1490

1491

1492

1493

1494

$$x_m \leq 4 + \frac{1}{2^{20}}.$$

In which of the following intervals does m lie?(A) $[9, 26]$ (B) $[27, 80]$ (C) $[81, 242]$ (D) $[243, 728]$ (E) $[729, \infty)$ (2019 AMC 10B, Problem 24 and 2019 AMC 12B, Problem 22)4.5: Find, with proof, all positive integers n for which $2^n + 12^n + 2011^n$ is a perfect square. (USAJMO 2011/1)5: Find all triples (a, b, c) of real numbers such that the following system holds:

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

$$a + b + c = \frac{1}{a} + \frac{1}{b} + \frac{1}{c},$$

$$a^2 + b^2 + c^2 = \frac{1}{a^2} + \frac{1}{b^2} + \frac{1}{c^2}.$$

(JBMO 2020/1)

5.5: Triangle ABC has $\angle BAC = 60^\circ$, $\angle CBA \leq 90^\circ$, $BC = 1$, and $AC \geq AB$. Let H , I , and O be the orthocenter, incenter, and circumcenter of $\triangle ABC$, respectively. Assume that the area of pentagon $BCOIH$ is the maximum possible. What is $\angle CBA$? (2011 AMC 12A, Problem 25)6: Let $\triangle ABC$ be an acute triangle with circumcircle ω , and let H be the intersection of the altitudes of $\triangle ABC$. Suppose the tangent to the circumcircle of $\triangle HBC$ at H intersects ω at points X and Y with $HA = 3$, $HX = 2$, and $HY = 6$. The area of $\triangle ABC$ can be written

1512 in the form $m\sqrt{n}$, where m and n are positive integers, and n is not divisible by the square
 1513 of any prime. Find $m + n$. (2020 AIME I, Problem 15)

1514 6.5: Rectangles BCC_1B_2 , CAA_1C_2 , and ABB_1A_2 are erected outside an acute triangle
 1515 ABC . Suppose that

$$1516 \angle BC_1C + \angle CA_1A + \angle AB_1B = 180^\circ.$$

1517 Prove that lines B_1C_2 , C_1A_2 , and A_1B_2 are concurrent. (USAMO 2021/1, USAJMO
 1518 2021/2)

1519 7: We say that a finite set \mathcal{S} in the plane is balanced if, for any two different points A, B in
 1520 \mathcal{S} , there is a point C in \mathcal{S} such that $AC = BC$. We say that \mathcal{S} is centre-free if for any three
 1521 points A, B, C in \mathcal{S} , there is no point P in \mathcal{S} such that $PA = PB = PC$. Show that for
 1522 all integers $n \geq 3$, there exists a balanced set consisting of n points. Determine all integers
 1523 $n \geq 3$ for which there exists a balanced centre-free set consisting of n points. (IMO 2015/1)

1524 7.5: Let \mathbb{Z} be the set of integers. Find all functions $f : \mathbb{Z} \rightarrow \mathbb{Z}$ such that

$$1525 xf(2f(y) - x) + y^2f(2x - f(y)) = \frac{f(x)^2}{x} + f(yf(y))$$

1527 for all $x, y \in \mathbb{Z}$ with $x \neq 0$. (USAMO 2014/2)

1528 8: For each positive integer n , the Bank of Cape Town issues coins of denomination $\frac{1}{n}$.
 1529 Given a finite collection of such coins (of not necessarily different denominations) with total
 1530 value at most $99 + \frac{1}{2}$, prove that it is possible to split this collection into 100 or fewer
 1531 groups, such that each group has total value at most 1. (IMO 2014/5)

1532 8.5: Let I be the incentre of acute triangle ABC with $AB \neq AC$. The incircle ω of ABC
 1533 is tangent to sides BC, CA , and AB at D, E , and F , respectively. The line through D
 1534 perpendicular to EF meets ω at R . Line AR meets ω again at P . The circumcircles of
 1535 triangle PCE and PBF meet again at Q . Prove that lines DI and PQ meet on the line
 1536 through A perpendicular to AI . (IMO 2019/6)

1537 9: Let k be a positive integer and let S be a finite set of odd prime numbers. Prove that
 1538 there is at most one way (up to rotation and reflection) to place the elements of S around the
 1539 circle such that the product of any two neighbors is of the form $x^2 + x + k$ for some positive
 1540 integer x . (IMO 2022/3)

1541 9.5: An anti-Pascal triangle is an equilateral triangular array of numbers such that, except
 1542 for the numbers in the bottom row, each number is the absolute value of the difference of
 1543 the two numbers immediately below it. For example, the following is an anti-Pascal triangle
 1544 with four rows which contains every integer from 1 to 10.

$$\begin{array}{cccc} & & & 4 \\ & & 2 & 6 \\ & 5 & 7 & 1 \\ 8 & 3 & 10 & 9 \end{array}$$

1551 Does there exist an anti-Pascal triangle with 2018 rows which contains every integer from 1
 1552 to $1 + 2 + 3 + \dots + 2018$? (IMO 2018/3)

1553 10: Prove that there exists a positive constant c such that the following statement is true:
 1554 Consider an integer $n > 1$, and a set \mathcal{S} of n points in the plane such that the distance
 1555 between any two different points in \mathcal{S} is at least 1. It follows that there is a line ℓ separating
 1556 \mathcal{S} such that the distance from any point of \mathcal{S} to ℓ is at least $cn^{-1/3}$.

1558 1559 ## Some known difficulty ratings of the competitions.

1560 1561 #### HMMT (November)

1562 Individual Round, Problem 6-8: 4

1563 Individual Round, Problem 10: 4.5

1564 Team Round: 4-5

1565 Guts: 3.5-5.25

1566 #### CEMC

1567 **Part A: 1-1.5**

1566	How many different 3-digit whole numbers can be formed using the digits 4, 7, and 9, assuming that no digit can be repeated in a number? (2015 Gauss 7 Problem 10)
1567	**Part B: 1-2**
1568	Two lines with slopes $\frac{1}{4}$ and $\frac{5}{4}$ intersect at (1, 1). What is the area of the triangle formed by these two lines and the vertical line $x = 5$? (2017 Cayley Problem 19) Part C (Gauss/Pascal): 2-2.5
1569	Suppose that $\frac{2009}{2014} + \frac{2019}{n} = \frac{a}{b}$, where a , b , and n are positive integers with $\frac{a}{b}$ in lowest terms. What is the sum of the digits of the smallest positive integer n for which a is a multiple of 1004? (2014 Pascal Problem 25)
1570	**Part C (Cayley/Fermat): 2.5-3**
1571	Wayne has 3 green buckets, 3 red buckets, 3 blue buckets, and 3 yellow buckets. He randomly distributes 4 hockey pucks among the green buckets, with each puck equally likely to be put in each bucket. Similarly, he distributes 3 pucks among the red buckets, 2 pucks among the blue buckets, and 1 puck among the yellow buckets. Once he is finished, what is the probability that a green bucket contains more pucks than each of the other 11 buckets? (2018 Fermat Problem 24)
1572	### Indonesia MO
1573	**Problem 1/5: 3.5** In a drawer, there are at most 2009 balls, some of them are white, the rest are blue, which are randomly distributed. If two balls were taken at the same time, then the probability that the balls are both blue or both white is $\frac{1}{2}$. Determine the maximum amount of white balls in the drawer, such that the probability statement is true?
1574	**Problem 2/6: 4.5** Find the lowest possible values from the function
1575	$f(x) = x^{2008} - 2x^{2007} + 3x^{2006} - 4x^{2005} + 5x^{2004} - \dots - 2006x^3 + 2007x^2 - 2008x + 2009$
1576	for any real numbers x .
1577	**Problem 3/7: 5** A pair of integers (m, n) is called good if
1578	$m \mid n^2 + n$ and $n \mid m^2 + m$
1579	Given 2 positive integers $a, b > 1$ which are relatively prime, prove that there exists a good pair (m, n) with $a \mid m$ and $b \mid n$, but $a \nmid n$ and $b \nmid m$.
1580	**Problem 4/8: 6** Given an acute triangle ABC . The incircle of triangle ABC touches BC, CA, AB respectively at D, E, F . The angle bisector of $\angle A$ cuts DE and DF respectively at K and L . Suppose AA_1 is one of the altitudes of triangle ABC , and M be the midpoint of BC . (a) Prove that BK and CL are perpendicular with the angle bisector of $\angle BAC$. (b) Show that A_1KML is a cyclic quadrilateral.
1581	### JBMO
1582	**Problem 1: 4** Find all real numbers a, b, c, d such that
1583	$a + b + c + d = 20, ab + ac + ad + bc + bd + cd = 150.$
1584	**Problem 2: 4.5-5** Let $ABCD$ be a convex quadrilateral with $\angle DAC = \angle BDC = 36^\circ$, $\angle CBD = 18^\circ$ and $\angle BAC = 72^\circ$. The diagonals intersect at point P . Determine the measure of $\angle APD$.
1585	**Problem 3: 5** Find all prime numbers p, q, r , such that $\frac{p}{q} - \frac{4}{r+1} = 1$.
1586	**Problem 4: 6** A 4×4 table is divided into 16 white unit square cells. Two cells are called neighbors if they share a common side. A move consists in choosing a cell and changing the colors of neighbors from white to black or from black to white. After exactly n moves all the 16 cells were black. Find all possible values of n .
1587	### Problem 1/4: 5 There are $a+b$ bowls arranged in a row, numbered 1 through $a+b$, where a and b are given positive integers. Initially, each of the first a bowls contains an apple, and each of the last b bowls contains a pear. A legal move consists of moving an apple from bowl i to bowl $i+1$ and a pear from bowl j to bowl $j-1$, provided that the difference $i-j$ is even. We permit multiple fruits in the same bowl at the same time. The goal is to end up with the first b bowls each containing a pear and the last a bowls each containing an apple. Show that this is possible if and only if the product ab is even.
1588	**Problem 2/5: 6-6.5** Let a, b, c be positive real numbers such that $a + b + c = 4\sqrt[3]{abc}$. Prove that
1589	$2(ab + bc + ca) + 4 \min(a^2, b^2, c^2) \geq a^2 + b^2 + c^2.$

1620 **Problem 3/6: 7** Two rational numbers $\frac{m}{n}$ and $\frac{n}{m}$ are written on a blackboard, where m
 1621 and n are relatively prime positive integers. At any point, Evan may pick two of the numbers
 1622 x and y written on the board and write either their arithmetic mean $\frac{x+y}{2}$ or their harmonic
 1623 mean $\frac{2xy}{x+y}$ on the board as well. Find all pairs (m, n) such that Evan can write 1 on the
 1624 board in finitely many steps.

1625 ### HMMT (February) Individual Round, Problem 1-5: 5 Individual Round, Problem 6-10:
 1626 5.5-6 Team Round: 7.5 HMIC: 8

1627 ### APMO Problem 1: 6 Problem 2: 7 Problem 3: 7 Problem 4: 7.5 Problem 5: 8.5

1628 ### Balkan MO Problem 1: 5 Solve the equation $3^x - 5^y = z^2$ in positive integers. Problem
 1629 2: 6.5 Let MN be a line parallel to the side BC of a triangle ABC , with M on the side AB
 1630 and N on the side AC . The lines BN and CM meet at point P . The circumcircles of trian-
 1631 gles BMP and CNP meet at two distinct points P and Q . Prove that $\angle BAQ = \angle CAP$.
 1632 Problem 3: 7.5 A 9×12 rectangle is partitioned into unit squares. The centers of all the unit
 1633 squares, except for the four corner squares and eight squares sharing a common side with one
 1634 of them, are coloured red. Is it possible to label these red centres C_1, C_2, \dots, C_{96} in such way
 1635 that the following two conditions are both fulfilled (i) the distances $C_1C_2, \dots, C_{95}C_{96}, C_{96}C_1$
 1636 are all equal to $\sqrt{13}$ (ii) the closed broken line $C_1C_2\dots C_{96}C_1$ has a centre of symmetry?
 1637 Problem 4: 8 Denote by S the set of all positive integers. Find all functions $f : S \rightarrow S$ such
 1638 that
 1639
$$f\left(f^2(m) + 2f^2(n)\right) = m^2 + 2n^2 \text{ for all } m, n \in S.$$

1640 ### USAMO Problem 1/4: 6-7 Problem 2/5: 7-8 Three nonnegative real numbers r_1, r_2, r_3
 1641 are written on a blackboard. These numbers have the property that there exist integers $a_1,$
 1642 a_2, a_3 , not all zero, satisfying $a_1r_1 + a_2r_2 + a_3r_3 = 0$. We are permitted to perform the
 1643 following operation: find two numbers x, y on the blackboard with $x \leq y$, then erase y and
 1644 write $y - x$ in its place. Prove that after a finite number of such operations, we can end up
 1645 with at least one 0 on the blackboard. Problem 3/6: 8-9 Prove that any monic polynomial
 1646 (a polynomial with leading coefficient 1) of degree n with real coefficients is the average of
 1647 two monic polynomials of degree n with n real roots.

1648 ### USA TST Problem 1/4/7: 6.5-7 Problem 2/5/8: 7.5-8 Problem 3/6/9: 8.5-9

1649 ### Putnam Problem A/B,1-2: 7 Find the least possible area of a concave set in the 7-D
 1650 plane that intersects both branches of the hyperparabola $xyz = 1$ and both branches of the
 1651 hyperbola $xwy = -1$. (A set S in the plane is called convex if for any two points in S the line
 1652 segment connecting them is contained in S .) Problem A/B,3-4: 8 Let H be an $n \times n$ matrix
 1653 all of whose entries are ± 1 and whose rows are mutually orthogonal. Suppose H has an $a \times b$
 1654 submatrix whose entries are all 1. Show that $ab \leq n$. Problem A/B,5-6: 9 For any $a > 0$,
 1655 define the set $S(a) = \{[an] | n = 1, 2, 3, \dots\}$. Show that there are no three positive reals a, b, c
 1656 such that $S(a) \cap S(b) = S(b) \cap S(c) = S(c) \cap S(a) = \emptyset, S(a) \cup S(b) \cup S(c) = \{1, 2, 3, \dots\}$.
 1657 ### China TST (hardest problems) Problem 1/4: 8-8.5 Given an integer m , prove that there
 1658 exist odd integers a, b and a positive integer k such that

$$2m = a^{19} + b^{99} + k * 2^{1000}.$$

1659 Problem 2/5: 9 Given a positive integer $n = 1$ and real numbers $a_1 < a_2 < \dots < a_n$, such
 1660 that $\frac{1}{a_1} + \frac{1}{a_2} + \dots + \frac{1}{a_n} \leq 1$, prove that for any positive real number x ,

$$\left(\frac{1}{a_1^2 + x} + \frac{1}{a_2^2 + x} + \dots + \frac{1}{a_n^2 + x} \right)^2 \geq \frac{1}{2a_1(a_1 - 1) + 2x}.$$

1665 Problem 3/6: 9.5-10 Let $n > 1$ be an integer and let a_0, a_1, \dots, a_n be non-negative real
 1666 numbers. Define $S_k = \sum_{i=0}^k \binom{k}{i} a_i$ for $k = 0, 1, \dots, n$. Prove that

$$\frac{1}{n} \sum_{k=0}^{n-1} S_k^2 - \frac{1}{n^2} \left(\sum_{k=0}^n S_k \right)^2 \leq \frac{4}{45} (S_n - S_0)^2.$$

1671 ### IMO **Problem 1/4: 5.5-7** Let Γ be the circumcircle of acute triangle ABC . Points
 1672 D and E are on segments AB and AC respectively such that $AD = AE$. The perpen-
 1673 dicular bisectors of BD and CE intersect minor arcs AB and AC of Γ at points F and G
 1674 respectively. Prove that lines DE and FG are either parallel or they are the same line.

1674 **Problem 2/5: 7-8** Let $P(x)$ be a polynomial of degree $n > 1$ with integer coefficients,
 1675 and let k be a positive integer. Consider the polynomial $Q(x) = P(P(\dots P(P(x))\dots))$,
 1676 where P occurs k times. Prove that there are at most n integers t such that $Q(t) = t$.
 1677 **Problem 3/6: 9-10** Let ABC be an equilateral triangle. Let A_1, B_1, C_1 be interior
 1678 points of ABC such that $BA_1 = A_1C, CB_1 = B_1A, AC_1 = C_1B$, and
 1679 $\angle BA_1C + \angle CB_1A + \angle AC_1B = 480^\circ$
 1680 Let BC_1 and CB_1 meet at A_2 , let CA_1 and AC_1 meet at B_2 , and let AB_1 and BA_1 meet
 1681 at C_2 . Prove that if triangle $A_1B_1C_1$ is scalene, then the three circumcircles of triangles
 1682 AA_1A_2, BB_1B_2 and CC_1C_2 all pass through two common points.
 1683 ### IMO Shortlist
 1684 Problem 1-2: 5.5-7
 1685 Problem 3-4: 7-8
 1686 Problem 5+: 9-10
 1687 </difficulty reference>
 1688
 1689 # OBJECTIVE #
 1690 1. Summarize the math problem in a brief sentence, describing the concepts involved in the
 1691 math problem.
 1692 2. Based on the source of the given problem, as well as the difficulty of the problems
 1693 referenced in these materials and the solution to the current problem, please provide an
 1694 overall difficulty score for the current problem. The score should be a number between 1
 1695 and 10, with increments of 0.5, and should align perfectly with the materials.
 1696 # STYLE #
 1697 Data report.
 1698 # TONE #
 1699 Professional, scientific.
 1700 # AUDIENCE #
 1701 Students. Enable them to better understand the difficulty of the math problems.
 1702 # RESPONSE: MARKDOWN REPORT #
 1703 ## Summarization
 1704 [Summarize the math problem in a brief paragraph.]
 1705 ## Difficulty
 1706 [Rate the difficulty of the math problem and give the reason.]
 1707 # ATTENTION #
 1708 - Add "==== report over ===" at the end of the report.
 1709
 1710 <example math problem>
 1711 [Question]:
 1712 If $\frac{1}{9} + \frac{1}{18} = \frac{1}{x}$, what is the number that replaces the x to make the equation true?
 1713 [Solution]:
 1714 We simplify the left side and express it as a fraction with numerator 1: $\frac{1}{9} + \frac{1}{18} = \frac{2}{18} + \frac{1}{18} =$
 1715 $\frac{3}{18} = \frac{1}{6}$. Therefore, the number that replaces the \square is 6.
 1716 [Source]: 2010_Pascal
 1717 </example math problem>
 1718 ## Summarization
 1719 The problem requires finding a value that makes the equation $\frac{1}{9} + \frac{1}{18} = \frac{1}{\square}$. This involves
 1720 adding two fractions and determining the equivalent fraction.
 1721 ## Difficulty
 1722 Rating: 1
 1723 Reason: This problem is straightforward and primarily involves basic fraction addition, making
 1724 it suitable for early middle school students.
 1725 ==== report over ===
 1726
 1727 <example math problem>
 1728 [Question]:

1728

1729

1730

1731

1732

1733

1734

1735

1736

1737

1738

1739

1740

1741

1742

1743

1744

1745

1746

1747

1748

1749

1750

1751

1752

1753

1754

1755

1756

1757

1758

1759

1760

1761

1762

1763

1764

1765

1766

1767

1768

1769

1770

1771

1772

1773

1774

1775

1776

1777

1778

1779

1780

1781

Let \mathcal{P} be a convex polygon with n sides, $n \geq 3$. Any set of $n - 3$ diagonals of \mathcal{P} that do not intersect in the interior of the polygon determine a triangulation of \mathcal{P} into $n - 2$ triangles. If \mathcal{P} is regular and there is a triangulation of \mathcal{P} consisting of only isosceles triangles, find all the possible values of n .

[Solution]:

We label the vertices of \mathcal{P} as $P_0, P_1, P_2, \dots, P_n$. Consider a diagonal $d = \overline{P_a P_{a+k}}$, $k \leq n/2$ in the triangulation. We show that k must have the form 2^m for some nonnegative integer m . This diagonal partitions \mathcal{P} into two regions Q, R , and is the side of an isosceles triangle in both regions. Without loss of generality suppose the area of Q is less than the area of R (so the center of P does not lie in the interior of Q); it follows that the lengths of the edges and diagonals in Q are all smaller than d . Thus d must be the base of the isosceles triangle in Q , from which it follows that the isosceles triangle is $\triangle P_a P_{a+k/2} P_{a+k}$, and so $2|k$. Repeating this process on the legs of isosceles triangle $(\overline{P_a P_{a+k/2}}, \overline{P_{a+k} P_{a+k/2}})$, it follows that $k = 2^m$ for some positive integer m (if we allow degeneracy, then we can also let $m = 0$). Now take the isosceles triangle $P_x P_y P_z$, $0 \leq x < y < z < n$ in the triangulation that contains the center of \mathcal{P} in its interior; if a diagonal passes through the center, select either of the isosceles triangles with that diagonal as an edge. Without loss of generality, suppose $P_x P_y = P_y P_z$. From our previous result, it follows that there are 2^a edges of P on the minor arcs of $P_x P_y, P_y P_z$ and 2^b edges of P on the minor arc of $P_z P_x$, for positive integers a, b . Therefore, we can write

$$n = 2 \cdot 2^a + 2^b = 2^{a+1} + 2^b,$$

so n must be the sum of two powers of 2. We now claim that this condition is sufficient. Suppose without loss of generality that $a + 1 \geq b$; then we rewrite this as

$$n = 2^b(2^{a-b+1} + 1).$$

Lemma 1: All regular polygons with $n = 2^k + 1$ or $n = 4$ have triangulations that meet the conditions. By induction, it follows that we can cover all the desired n . For $n = 3, 4$, this is trivial. For $k > 1$, we construct the diagonals of equal length $\overline{P_0 P_{2^{k-1}}}$ and $\overline{P_{2^{k-1}+1} P_0}$. This partitions \mathcal{P} into 3 regions: an isosceles $\triangle P_0 P_{2^{k-1}} P_{2^{k-1}+1}$, and two other regions. For these two regions, we can recursively construct the isosceles triangles defined above in the second paragraph. It follows that we have constructed $2(2^{k-1} - 1) + (1) = 2^k - 1 = n - 2$ isosceles triangles with non-intersecting diagonals, as desired.

Lemma 2: If a regular polygon with n sides has a working triangulation, then the regular polygon with $2n$ sides also has a triangulation that meets the conditions. We construct the diagonals $\overline{P_0 P_2}, \overline{P_2 P_4}, \dots, \overline{P_{2n-2} P_0}$. This partitions \mathcal{P} into n isosceles triangles of the form $\triangle P_{2k} P_{2k+1} P_{2k+2}$, as well as a central regular polygon with n sides. However, we know that there exists a triangulation for the n -sided polygon that yields $n - 2$ isosceles triangles. Thus, we have created $(n) + (n - 2) = 2n - 2$ isosceles triangles with non-intersecting diagonals, as desired. In summary, the answer is all n that can be written in the form $2^{a+1} + 2^b$, $a, b \geq 0$. Alternatively, this condition can be expressed as either $n = 2^k$, $k \geq 2$ (this is the case when $a + 1 = b$) or n is the sum of two distinct powers of 2, where $1 = 2^0$ is considered a power of 2.

[Source]:

USAMO 2008

</example math problem>

Summarization

The problem asks for the possible values of n for a regular n -sided polygon that can be completely triangulated into isosceles triangles using non-intersecting diagonals. The solution involves analyzing the properties of the diagonals forming isosceles triangles and deducing that n can be expressed in terms of powers of 2.

Difficulty

Rating: 7

Reason: The problem involves understanding properties of isosceles triangles in the context of polygon triangulation and requires critical reasoning to establish relationships between the number of sides and powers of 2, making it more complex than typical undergraduate-level problems.

```

1782     === report over ===
1783
1784
1785     <math problem>
1786     [QUESTION]:
1787     ***Question***
1788     [SOLUTION]:
1789     ***Solution***
1790     [SOURCE]:
1791     ***SOURCE***
1792     </math problem>
1793
1794 Your answer should be in JSON format for example:
1795     {"json
1796     {
1797     "Rating": YOUR RATING,
1798     "Reason": YOUR JUSTIFICATION,
1799     }
1800     "}
1801
1802
1803 F.7 TOPIC DOMAIN
1804
1805
1806     Meta Prompt for Topic Domain Annotation
1807
1808
1809
1810     I am a teacher, and I have some high-level Olympiad math problems.
1811     I want to categorize the domain of these math problems.
1812
1813
1814     # OBJECTIVE #
1815     1. Summarize the math problem in a brief sentence, describing the concepts involved in the
1816     math problem.
1817     2. Categorize the math problem into specific mathematical domains. Please provide a
1818     classification chain, for example, Applied Mathematics -> Probability -> Combinations.
1819     The following is a basic classification framework in the field of mathematics.
1820     <math domains>
1821     ...
1822     </math domains>
1823     # STYLE #
1824     Data report.
1825     # TONE #
1826     Professional, scientific.
1827     # AUDIENCE #
1828     Students. Enable them to better understand the domain of the problems.
1829     # RESPONSE: MARKDOWN REPORT #
1830     ## Summarization
1831     [Summarize the math problem in a brief paragraph.]
1832     ## Math domains
1833     [Categorize the math problem into specific mathematical domains, including major domains
1834     and subdomains.]
1835     # ATTENTION #
1836     - The math problem can be categorized into multiple domains, but no more than three. Sep-
1837     arate the classification chains with semicolons(;).
1838     - Your classification MUST fall under one of the aforementioned subfields; if it really does
1839     not fit, please add "Other" to the corresponding branch. For example: Algebra -> Intermedi-
1840     ate Algebra -> Other. Only the LAST NODE is allowed to be "Other"; the preceding nodes
1841     must strictly conform to the existing framework.
1842     - The math domain must conform to a format of classification chain, like "Applied Mathe-
1843     matics -> Probability -> Combinations".

```

F.7 TOPIC DOMAIN

Meta Prompt for Topic Domain Annotation

I am a teacher, and I have some high-level Olympiad math problems.
I want to categorize the domain of these math problems.

OBJECTIVE

1. Summarize the math problem in a brief sentence, describing the concepts involved in the math problem.
2. Categorize the math problem into specific mathematical domains. Please provide a classification chain, for example, Applied Mathematics -> Probability -> Combinations.

The following is a basic classification framework in the field of mathematics.

<math domains>

...

</math domains>

STYLE

Data report.

TONE

Professional, scientific.

AUDIENCE

Students. Enable them to better understand the domain of the problems.

RESPONSE: MARKDOWN REPORT

Summarization

[Summarize the math problem in a brief paragraph.]

Math domains

[Categorize the math problem into specific mathematical domains, including major domains and subdomains.]

ATTENTION

- The math problem can be categorized into multiple domains, but no more than three. Separate the classification chains with semicolons(;).
- Your classification MUST fall under one of the aforementioned subfields; if it really does not fit, please add "Other" to the corresponding branch. For example: Algebra -> Intermediate Algebra -> Other. Only the LAST NODE is allowed to be "Other"; the preceding nodes must strictly conform to the existing framework.
- The math domain must conform to a format of classification chain, like "Applied Mathematics -> Probability -> Combinations".

1836 - Add "==== report over ===" at the end of the report.
 1837
 1838 <example math problem>
 1839 [Question]
 1840 Determine the greatest real number C , such that for every positive integer $n \geq 2$, there exists
 1841 $x_1, x_2, \dots, x_n \in [-1, 1]$, so that
 1842
$$\prod_{1 \leq i < j \leq n} (x_i - x_j) \geq C^{\frac{n(n-1)}{2}}$$

 1843
 1844
 1845 </example math problem>
 1846 ## Summarization
 1847 The problem seeks to find the greatest real number C such that, for every integer $n \geq 2$, there
 1848 exist real numbers $x_1, x_2, \dots, x_n \in [-1, 1]$ satisfying the inequality $\prod_{1 \leq i < j \leq n} (x_i - x_j) \geq$
 1849 $C^{\frac{n(n-1)}{2}}$. This involves maximizing C to ensure the product of all pairwise differences
 1850 among n points in the interval $[-1, 1]$ is at least C raised to the power of the number of such
 1851 pairs, $\frac{n(n-1)}{2}$.
 1852 ## Math domains
 1853 Algebra -> Intermediate Algebra -> Inequalities; Discrete Mathematics -> Combinatorics
 1854 === report over ===
 1855
 1856 <example math problem>
 1857 [Question]
 1858 Given integer $n \geq 2$. Find the minimum value of λ , satisfy that for any real numbers a_1, a_2, \dots, a_n and b ,
 1859
 1860
$$\lambda \sum_{i=1}^n \sqrt{|a_i - b|} + \sqrt{n \left| \sum_{i=1}^n a_i \right|} \geq \sum_{i=1}^n \sqrt{|a_i|}.$$

 1861
 1862
 1863 </example math problem>
 1864 ## Summarization
 1865 Let $n \geq 2$ be an integer. The problem seeks the minimum value of λ such that for any
 1866 real numbers a_1, a_2, \dots, a_n and b , the inequality $\lambda \sum_{i=1}^n \sqrt{|a_i - b|} + \sqrt{n \left| \sum_{i=1}^n a_i \right|} \geq$
 1867 $\sum_{i=1}^n \sqrt{|a_i|}$ holds. The goal is to find the smallest λ that ensures this inequality is satisfied
 1868 for all possible choices of a_i and b .
 1869 ## Math domains
 1870 Algebra -> Intermediate Algebra -> Inequalities;
 1871 Calculus -> Differential Calculus -> Applications of Derivatives.
 1872 === report over ===
 1873
 1874 <math problem>
 1875 [Question]
 1876 ***Question***
 1877 [Solution]
 1878 ***Solution***
 1879 [Source]
 1880 ***Source***
 1881 </math problem>
 1882
 1883 Your answer should be in JSON format for example:
 1884 """ json
 1885 {
 1886 "Summary": "YOUR_SUMMARY",
 1887 "Domains": [domain1, ...]
 1888 }
 1889 """

1890

F.8 THEOREM VARIANT CONSTRUCTION

1891

1892

1893

Meta Prompt for Constructing Theorem Variants

1894

1895

1896

1897

1898

1899

1900

1901

1902

1903

1904

1905

I'm going to give you a math proof question and its solution. Your task is to follow the steps below to write a new question based on the given one. Here is the original question and solution:

```
““question
<question>
““,
““solution
<solution>
““
```

Please follow these steps:

1. The original question asks to "prove or disprove" a statement, where the statement can be "proved". Please write a new question by negating the original statement, so that it can now be "disproved". For example, if the original statement is $x = y$, you may change it to $x < y$ or $x \neq y$; if the statement is "there exists xxx", you may change it to "there does not exist xxx". When negating the original question, you should make minimal changes, i.e. leave as much background information unchanged as possible.

2. After changing the question, the solution should be changed accordingly. You do not have to write a new solution, and the original solution can probably be reused. For example, if the original question asks to prove $x = y$ and the new question asks to prove $x < y$, you may simply add a step to the original proof like "since we proved $x = y$, the statement $x < y$ is disproved". However, check the wording of the solution so that it tries to "prove" the statement at first, and then naturally transit to finding that it cannot be proved, but can be disproved instead.

3. The original solution ends with "\boxed{proved}". Your new solution should end with "\boxed{disproved}".

Output the new question and solution in two blocks:

```
““question
new question
““
““solution
corresponding solution
““
```

1928

1929

1930

1931

F.9 SYSTEM PROMPT FOR RL TRAINING

1932

1933

System Prompt for RL Training

A conversation between User and Assistant. The User gives a statement, and the Assistant either proves or disproves it. The Assistant first thinks about the reasoning process in the mind and then provides the User with the answer. The reasoning process is enclosed within <think> </think> and the answer is enclosed within <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. If you prove the statement, answer with "proved". If you disprove the statement, answer with "disproved". You must put your answer inside <answer> </answer> tags, i.e., <answer> \boxed{proved} or \boxed{disproved} </answer>. And your final answer will be extracted automatically by the \boxed{} tag.

1944 F.10 PROCESS EVALUATION FRAMEWORK
19451946 Meta Prompt for Process Evaluation
19471948 You are an expert in scoring solutions for mathematical proof questions. The follow-
1949 ing question asks to prove or disprove a statement, where the statement may be either
1950 true or false. The test subject is asked to end their proof with \boxed{proved} if they
1951 prove the statement to be true, and \boxed{disproved} if they prove the statement to be false.
19521953 The question:
1954 ““<question>““1955 The ground truth of the statement:
1956 ““<answer>““1957 The test subject’s solution:
1958 ““ <solution>““1959 Your task is to evaluate the proof’s quality and assign a score from 0 to 1 based on four
1960 criteria: logical validity (40%), completeness (30%), correctness (20%), and clarity (10%).
1961 Instructions:

1962 1. Analyze the proof step by step.

1963 2. For each criterion:

1964 - Logical Validity: Check if each step follows logically from the previous one. Flag any
1965 logical errors.

1966 - Completeness: Verify if all necessary cases and steps are included to prove the theorem.

1967 - Correctness: Confirm if the final conclusion is correct.

1968 - Clarity: Assess if the proof is clear, unambiguous, and well-explained.

1969 3. Assign a sub-score (0 to 1) for each criterion and compute the total score using the
1970 weights: $(0.4 \times \text{validity}) + (0.3 \times \text{completeness}) + (0.2 \times \text{correctness}) + (0.1 \times \text{clarity})$.1971 4. Provide a brief explanation (2-3 sentences) summarizing any errors or issues and justifying
1972 the score.1973 Final output format:
1974 ““1975 {
1976 "score": float,
1977 "validity": float,
1978 "completeness": float,
1979 "correctness": float,
1980 "clarity": float,
1981 "explanation": str
1982 }1983 where "score" is the total score, and "validity", "completeness", "correctness", "clarity" are
1984 the subscores.
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997