ByteDance Seed Prover Achieves Silver Medal Score at IMO 2025

The ByteDance Seed Team

July 23, 2025

Abstract

The ByteDance Seed team was officially invited to participate in the International Mathematical Olympiad (IMO) 2025. Our specialized model for formal mathematical reasoning, Seed Prover, after three days of attempts, successfully solved 4 out of the 6 problems and provided a partial proof for another. The IMO officially certified Seed Prover's score as 30 points, reaching the silver medal threshold. The individual scores for the six problems were 2/7/7/7/7/0. Following the competition, the team made additional attempts and successfully produced complete proofs for the first five problems.

Seed Prover is built upon the Lean proof assistant and is trained using multi-stage reinforcement learning. It proves mathematical problems through an interaction between natural language reasoning and formal code, ensuring that the proof process is 100% reliable. By leveraging long-chain-of-thought reasoning and expanding its computational power at test-time through a multi-agent setup, Seed Prover can engage in deep and broad thinking on a single mathematical problem for several days.

In addition to its strong performance at the IMO, Seed Prover has set new state-of-the-art results in formal mathematical reasoning across multiple benchmarks.

1 IMO 2025 Problem Solving Details

Seed Prover took three days to solve four of the IMO 2025 competition problems. Before the attempt, the problems were formalized by humans to ensure their accuracy. Ultimately, Seed Prover solved Problem 2 (Geometry) and Problem 3 (Number Theory) from the first day of the competition, and Problem 4 (Number Theory) and Problem 5 (Combinatorics/Algebra) from the second day. After the competition, Seed Prover continued its search and successfully proved Problem 1 (Combinatorics).

The specific details for each problem are as follows:

- **Problem 1 (Combinatorics):** During the competition, Seed Prover only found the correct construction. Post-competition, it completed a 4000-line proof after 4 days of search.
- **Problem 2 (Geometry):** The proof was automatically generated and verified by the Seed Geometry subsystem in 2 seconds.
- **Problem 3 (Number Theory):** The proof was completed by Seed Prover after 3 days of reasoning, resulting in 2000 lines of formal code.
- **Problem 4 (Number Theory):** This also took 3 days to prove, generating 4000 lines of formal code.
- **Problem 5 (Combinatorics/Algebra):** The proof was completed in 1 day and differs from the known public human solutions.

The complete formal proofs and their corresponding natural language translations for the problems mentioned above are available at the following links:

- **Problem 1 Proof:** `.../imo2025/p1.lean`
- **Problem 2 Proof:** `.../imo2025/p2_proof.pdf`
- **Problem 3 Proof:** `.../imo2025/p3.lean`

- **Problem 4 Proof:** `.../imo2025/p4.lean`
- **Problem 5 Proof:** `.../imo2025/p5.lean`

2 Seed Prover Test-Time Compute Scaling Settings

Seed Prover's success stems from its multi-stage reinforcement learning, which enhances its long-chain-of-thought reasoning capabilities for complex mathematical problems and enables diverse interactions with the Lean prover. Another critical factor is how Seed Prover scales its reasoning at test-time, for which the team designed a multi-level strategy.

2.1 Lightweight Setting

Previous research evaluated large language models equipped with long-chain-of-thought reasoning on automated theorem proving using the Pass@K metric. Our team discovered that by using feedback from the Lean compiler combined with the model's summarization capabilities to iteratively refine proofs, we could overcome the token budget limitations of a single inference pass and significantly enhance proving capabilities.

The lightweight setting for Seed Prover involves a maximum of 8 iterations per proof, evaluated under a Pass@8 standard, a process that can be completed within 1 hour. In this setting, Seed Prover can prove problems of the difficulty of IMO 2022 P2 (MOHS hardness = 15). Without the allowance for iterative refinement using Lean compiler information, solving IMO 2022 P2 would require a Pass@8192 standard.

2.2 Medium-weight Setting

Proving a high-difficulty competition problem can involve a lengthy and complex proof. The medium-weight setting for Seed Prover includes both an inner and an outer optimization loop. The outer loop optimizes the proof of the original problem, similar to the lightweight setting. The inner loop attempts to prove difficult lemmas that the outer loop failed to resolve, providing a sufficient token budget for each individual lemma.

Under this setting, Seed Prover can tackle more difficult problems, such as IMO 2003 P6 (MOHS hardness = 35) and IMO 2025 P5, with final proof code sometimes exceeding 1000 lines. The runtime for this setting ranges from several hours to several days.

2.3 Heavyweight Setting

While the medium-weight setting allows Seed Prover to think deeply about every detail of a proof, it does not broadly explore new conjectures surrounding a given problem. In the heavyweight reasoning setting, the system creates a conjecture pool containing the given problem and an empty lemma pool.

During the reasoning process, Seed Prover attempts to prove or disprove each problem in the conjecture pool. It utilizes the lemma pool and Lean compilation results to optimize its proofs. Once a conjecture is proven, it is moved to the lemma pool. If certain problems are difficult to prove, Seed Prover creates new conjectures that might help solve these challenges and adds them to the conjecture pool. After several days of thinking, the lemma pool may contain thousands of interesting mathematical facts. The system then evaluates the lemmas based on their proof difficulty and relevance, selecting several hundred of the most valuable ones to help complete the proof of the given problem. IMO 2025 P3 and IMO 2025 P4 were solved using this heavyweight setting.

3 Seed Prover Multi-Benchmark Results

To compare Seed Prover with previous state-of-the-art technologies, the team evaluated it on multiple formal reasoning datasets. Seed Prover successfully proved 79% of previously formalized IMO problems, solved the vast majority of problems in the MiniF2F benchmark, and also solved more than half of the historical problems from the North American university-level mathematics competition, the Putnam Competition. Seed Prover also demonstrated advantages on several other benchmarks, including CombiBench and MiniCTX-v2. This indicates that Seed Prover is not only proficient at high school competition problems but can also handle a wide range of mathematical concepts.

测试基准	Seed Prover	此前SOTA
IMO 2025	4/6（重量级，赛后5/6）	5/6（Gemini，自然语言）
MiniF2F-valid	100%（中量级，1题重量级）	90.6%（DeepSeek-Prover-V2）
MiniF2F-test	99.6%（中量级）	92.2%（Kimina-Prover）
PutnamBench	331（中量级）	64（Goedel-Prover-V2）
CombiBench	30%（中量级）	10%（Deepseek-Prover-V2）
MiniCTX-v2	81.8%（轻量级）	44.3%（O4-mini）

4 Summary and Outlook

We thank the IMO committee for inviting the ByteDance Seed team to participate, and for their help with grading and organizing the scoring. The final score for Seed Prover at IMO 2025 disclosed here was confirmed with the IMO committee prior to this announcement.

The performance of Seed Prover on math competition problems demonstrates the potential of combining large language models with formal verification. The reasoning-time scaling strategies adopted by the team significantly improved the system's performance through deep and broad thinking. However, competition mathematics does not represent all fields of mathematics, and the team's future goal is to explore a wider range of mathematical topics.