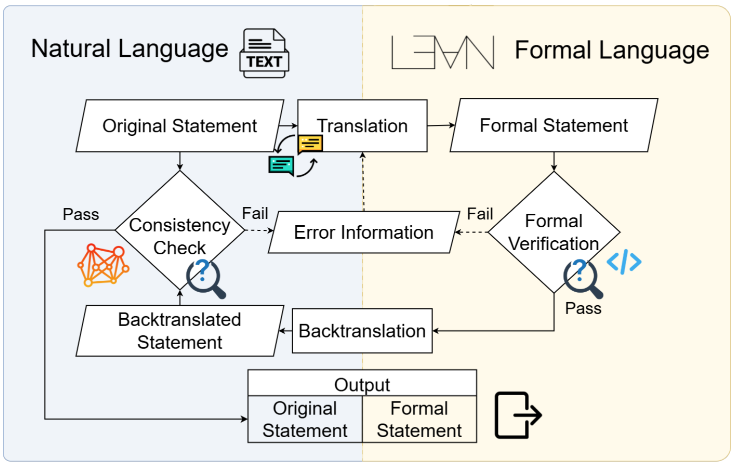
FMC: *F*ormalization of Natural Language *M*athematical *C*ompetition Problems

Jiaxuan Xie, Chengwu Liu, Ye Yuan, Siqi Li, Zhiping Xiao♤, Ming Zhang♤
Correspondence to: Zhiping Xiao<patxiao@uw.edu >, Ming Zhang<mzhang_cs@pku.edu.cn>,
First Author: Jiaxuan Xie<xiejiaxuan@stu.pku.edu.cn>



Our Main Pipeline

Formal Translation: Translate mathematical problems stated in natural language into formal expressions in Lean, using few-shot prompting initially and incorporating error feedback in the second round.

Formal Verification: Submit each translated formal theorem to Lean4 via DeepSeek-Prover's REPL interface to verify its syntax, and also use error feedback to refine prompts iteratively.

Back Translation: Back-translate formalized theorems into natural language to make semantic alignment checks more effective than directly comparing them with the original statements.

Consistency Check: Verify whether the backtranslated natural language theorem is mathematically consistent with the original theorem semantically.

Dataset Construction

Data Collection: collect **Olympiad level** math problems from IMOmath website, web crawling + OCR (Mathpix) + extracting statements

Data Preprocessing: excluding geometry problems + splitting problems with multiple goals

Data Construction: pass all data through our formalization pipeline

Table 1. The result of formalization.

Class	Number	Ratio
Total	4798	100%
Formal verification	4481	93.39%
Pass at one go	4287	89.35%
Pass with error feedback	194	4.04%
Consistency check	3922	81.74%
Pass at one go	3631	75.68%
Pass with error feedback	291	6.07%

Experiments

Autoformalization Capability of Different LLMs

DeepSeek-R1 at the forefront

Table 2. Model cross-validation results. The cell format is: *Formal verification pass rate/consistency check pass rate*.

Formalization model	Consistency check model		
	DeepSeek-R1	GPT-4o-mini	Claude 3.7 Sonnet
DeepSeek-R1	58% / 43%	58% / 31%	58% / 54%
GPT-4o-mini	34% / 10%	34% / 11%	34% / 22%
Claude 3.7 Sonnet	31% / 22%	31% / 14%	31% / 27%

Table 4. Evaluation matrix for consistency checks of different models. Experiments were based on formalization model Deepseek-R1.

Model name	Accuracy	Precision	Recall	F1 score
DeepSeek-R1	74.1%	69.8%	93.8%	80.0%
GPT-4o-mini	74.1%	77.4%	75.0%	76.2%
Claude-3-7	58.6%	57.4%	96.9%	72.1%

Testing as a Benchmark for Automated Theorem Provers

Table 8. Test results of different automated theorem provers. Each verification task was evaluated over 32 runs on 1,000 randomly sampled formal problems.

Dataset	Kimina-Prover	Goedel-Prover	DeepSeek-Prover-V1.5-RL
MiniF2F	63.1%	57.6%	50.0%
ProofNet	-	15.2%	16.0%
FormalMATH	16.5%	13.5%	10.2%
FMC	16.4%	15.7%	13.0%

Conclusions:

- **Autoformalization Pipeline:** A fully automated, training-free formalization framework enhanced with error feedback.
- **FMC Dataset:** A dataset of **3,922 Olympiad level** natural language problems aligned with 9,787 Lean statements.
- **Autoformalization Capbilities**: Evaluate formalization and reasoning capabilities of multiple LLMs.
- **SoTA Provers Evaluation**: Three automated theorem provers are tested on the FMC dataset.

Feel free to use our work!

Dataset Code Paper