# REAL-Prover: Retrieval Augmented Lean Prover for Mathematical Reasoning

**Ziju Shen**[1]* **Naohao Huang**[2]* **Fanyi Yang**[1]* **Yutong Wang**[3]* **Guoxiong Gao**[1]*

**Tianyi Xu**[1] **Jiedong Jiang**[1] **Wanyi He**[1] **Pu Yang**[1] **Mengzhou Sun**[3] **Haocheng Ju**[1]

**Peihao Wu**[4] **Bryan Dai**[4] **Bin Dong**[567]†

[1] Peking University [2] Renmin University of China [3] National University of Singapore [4] Ubiquant
[5] Beijing International Center for Mathematical Research and the New Cornerstone Science Laboratory, Peking University
[6] Center for Machine Learning Research, Peking University
[7] Center for Intelligent Computing, Great Bay Institute for Advanced Study, Great Bay University
dongbin@math.pku.edu.cn
Code: https://github.com/frenzymath/REAL-Prover

## Abstract

Nowadays, formal theorem provers have made monumental progress on high-school and competition level mathematics, but few of them generalize to more advanced mathematics. In this paper, we present REAL-Prover, a new open-source stepwise theorem-prover for Lean 4 to push this boundary. This prover, based on our fine-tuned large language model (REAL-Prover-v1) and integrated with a retrieval system (Leansearch-PS), notably boosts performance on solving college-level mathematics problems. To train REAL-Prover-v1, we developed HERALD-AF, a data extraction pipeline that converts natural language math problems into formal statements, and a new open-source Lean 4 interactive environment (Jixia-interactive) to facilitate synthesis data collection. In our experiments, our prover using only supervised fine-tune achieves competitive results with a 23.7% success rate (Pass@64) on the ProofNet dataset—comparable to state-of-the-art (SOTA) models. To further evaluate our approach, we introduce FATE-M, a new benchmark focused on algebraic problems, where our prover achieves a SOTA success rate of 56.7% (Pass@64).

## 1 Introduction

The formalization of mathematics has long been a great challenge at the intersection of programming and mathematical logic. Over the past decades, theorem provers have made steady progress. In particular, proof assistants have been used to formalize complex theorems such as the Four-Color Theorem [1] and the Feit-Thompson Odd-Order Theorem [2]. These achievements underscore the potential of formal methods to resolve research-level proofs.

Modern Interactive Theorem Provers (ITPs) like Lean [3] and Coq [4] have significantly improved usability and automation, lowering the barrier for adoption. Lean, for example, provides a user-friendly environment and a large community-driven library of mathematics (Mathlib) [5]. In early

*Equal contribution
†Corresponding author

Preprint. Under review.

2025, Lean's Mathlib had grown to encompass more than 210,000 formally proven theorems in various domains. Meanwhile, utilities like LeanSearch [6] also emerge to help users navigate easily in these numerous domains and facts, saving their efforts on formal proof drafting.

Recent advances in large language models (LLMs) have significantly advanced automated theorem proving, successfully generating formal proofs for high-school-level math competitions [7, 8, 9, 10, 11, 12, 13, 14, 15]. However, extending these approaches to undergraduate and graduate mathematics remains challenging due to data sparsity, domain-specific complexities, and the necessity of accurately recalling extensive lemma libraries, where even minor errors can derail purely neural proof generation.

To overcome these challenges, we propose a stepwise tactic synthesis approach in which each invocation of the language model is conditioned on the current proof state and enhanced by retrieving pertinent theorems from Mathlib. First, we introduce **LeanSearch-PS**, a semantic premise selection engine inspired by LeanSearch [6], designed to efficiently identify the $k$-most relevant theorems for any given proof state. Specifically, at each tactic step, we extract the current proof state from the Lean compiler and then query LeanSearch-PS to obtain candidate theorems. Second, we introduce REtrieval Augmented Lean Prover, named **REAL-Prover**, an integrated framework which combines the proof state information and these retrieved premises as inputs for the tactic-generation model. Specifically, it consists of an expert iteration framework and a translator pipeline. The former starts by imitating a small dataset derived from expert-guided proofs, then iteratively refines the prover through self-play interactions between the prover and compiler. The latter, **HERALD-AF**, is an extension of the HERALD translator model to be a whole translation pipeline [16], which generates and validates formal statements automatically from mathematical textbooks.

Furthermore, to support efficient interaction between prover and compiler in stepwise proving, we also developed **Jixia-interactive**, a Lean-based interactive environment. Specifically, the environment tool works within an all-Lean, CLI-driven scope with transparent environment and version management. It processes Lean tactic interactions internally, while collecting and presenting errors/warnings as the exploration or edition goes deeper. Compared to previous static-analysis tools like REPL, Jixia[17] and LeanDojo[18], Jixia-interactive is more usable, stable and efficient.

We finally propose a benchmark Formal Algebra Theorem Evaluation-Medium, named **FATE-M**, which is designed for formal reasoning in graduate-level abstract algebra. It enriches existing benchmarks emphasizing advanced algebraic structures (e.g., groups, rings, fields), making it an ideal testbed for exploring complex formal proofs. We evaluate our REAL-Prover on FATE-M, achieving a state-of-the-art accuracy of 56.7%, and further demonstrate its effectiveness on general benchmarks miniF2F [19] and ProofNet [20], where it achieves competitive results.

We summarize our contribution as follows.

- **Retrieval-Augmented Proof Search.** We propose REAL-Prover, a stepwise theorem proving system integrated with LeanSearch-PS, a semantic premise-selection engine. This approach significantly enhances prover effectiveness on graduate-level formal mathematics.
- **Automated Data Generation Pipeline.** We develop HERALD-AF, a novel pipeline for translating informal mathematical statements into formal Lean statements. This system facilitates large-scale auto-formalization, substantially enriching the available training dataset.
- **Lean Interactive Environment.** We introduce Jixia-interactive, an advanced interactive platform for Lean 4, designed to provide efficient, stable, and user-friendly interactions between the prover and compiler throughout both training and inference phases.
- **Specialized Benchmark for Abstract Algebra.** We present FATE-M (Formal Algebra Theorem Evaluation–Medium), a specialized benchmark comprising paired informal-formal problems in graduate-level abstract algebra. FATE-M complements existing benchmarks such as miniF2F [19] and ProofNet [20], establishing an effective framework for rigorous formal reasoning assessment.

## 2 Method

To construct a high-performance automated prover, we propose REtrieval-Augmented Lean Prover (REAL-Prover), a stepwise proof system for Lean 4. To supply REAL-Prover with a high-quality training corpus, we first introduce Hierarchy and Retrieval-based Translated Lean Dataset Auto-formalization (HERALD-AF) in Section 2.1, a pipeline that translates informal mathematical prob-

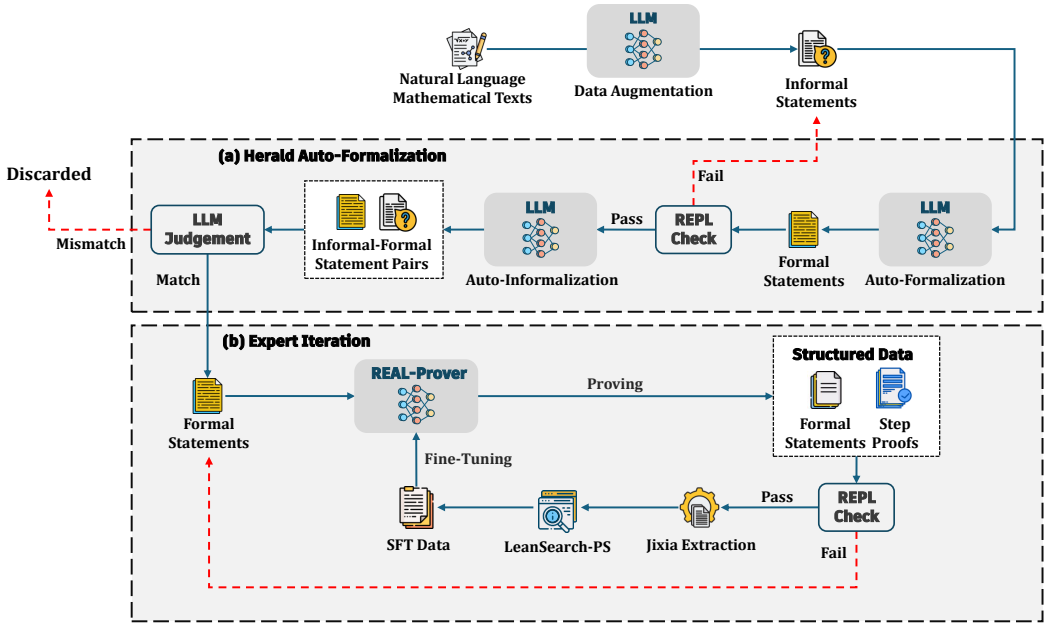



Figure 1: **Overview of the REAL-Prover model training pipeline.** The figure illustrates two components: (a) the HERALD-AF pipeline, which translates informal mathematical statements into formal ones; (b) the expert iteration pipeline, which iteratively refines the prover model.

lems into formal Lean 4 statements. HERALD-AF converts informal mathematical problems into formal Lean 4 statements, yielding a large corpus of candidate theorems. Since these statements still lack certified proofs, in Section 2.2, we build a complementary pipeline—again driven by REAL-Prover—that synthesizes and automatically verifies the required proofs, completing the training loop.

### 2.1 HERALD-AF

Our formal statement generation framework builds upon the Hierarchy and Retrieval-based Translated Lean Dataset (HERALD) [16], which establishes a protocol translating informal statements into formal statements. Figure 1(a) schematizes HERALD-AF pipeline.

Before applying HERALD-AF, we perform a preprocessing step to extract informal statements from a natural language mathematics corpus. This begins with the use of regular expressions to identify theorem blocks and exercises within the text. Subsequently, an LLM is employed to generate and refine the extracted informal statements.

Once the informal statements are obtained, we apply HERALD-AF to translate them into formal statements. This translation pipeline comprises three key stages: Auto-Formalization, Auto-Informalization, and LLM Judgement. In the Auto-Formalization stage, an open-source model Herald-translator [16] is used to generate multiple candidate formal statements from each informal input. During Auto-Informalization, these formal candidates are translated back into natural language using DeepSeek-V3 [21]. Finally, in the LLM Judgement stage, a general LLM evaluates whether each back-translated statement faithfully reflects the original input. Only those formal statements that pass this consistency check are retained.

### 2.2 REAL-Prover

From the HERALD-AF pipeline, we obtained a collection of formal statements without corresponding proofs. To generate complete proofs and achieve the second goal, we introduce the REAL-Prover system, illustrated in Figure 2. We developed a new Lean 4 interactive environment, Jixia-interactive, which receives Lean tactics and updates the proof state. For each formal statement, Jixia-interactive initializes the proof state. REAL-Prover constructs a search tree to explore possible proof paths. At

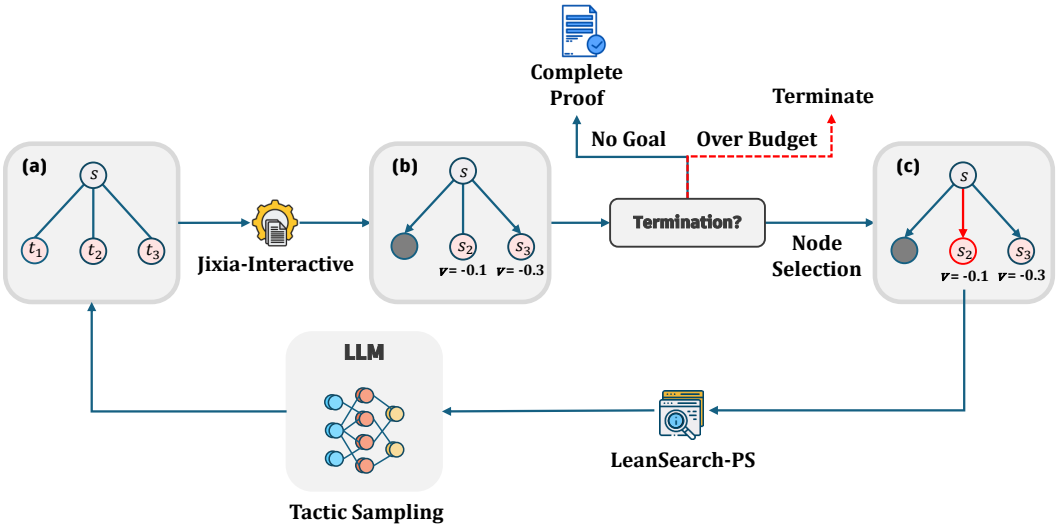



Figure 2: **Overview of REAL-Prover.** This figure illustrates three main processes: (a) to (b), the prover passes several tactics to the Lean 4 interactive environment (Jixia-interactive) and receives corresponding scored state nodes. (b) to (c), the prover selects a state node based on its score and checks whether the proof is complete. (c) to (a), the prover using LLM with retrieval information from LeanSearch-PS generate new tactics.

each node, a LLM generates multiple candidate tactics. Jixia-interactive then applies these tactics, discards invalid ones, and advances to new proof states. Similar to BFS-Prover [14], each node is scored using the scoring function: $\frac{\sum_{t=0}^{L-1} \log p(a_t|s_t)}{L^\alpha}$, where the $s_t$ represents the proof state and $a_t$ is the tactic applied at step t. $p(a_t|s_t)$ is the model's predicted probability of applying tactic $a_t$ at state $s_t$. $L$ is the total length of the path from the root to the current state $s_L$ and $\alpha$ is a hyper-parameter in $[0, 1]$. We adopt a best-first search strategy, selecting the node with the highest score to expand further by generating new tactics. This iterative process continues until either the proof is completed or a predefined search budget is exhausted. Once completed, the full proof can be reconstructed from the search tree. To enhance our prover's performace, we add retrieval information from Mathlib using LeanSearch premise selection (LeanSearch-PS), which extends the work of LeanSearch [6]. We will discuss the details of LeanSearch-PS.

### 2.2.1 LeanSearch-PS

In our work, premise selection addresses the challenge of identifying the most relevant theorems based on the current Lean proof state. To enhance this process, we developed LeanSearch-PS, a retrieval system built around a high-performance embedding model. This model projects both Lean proof states and Mathlib theorems into a shared semantic vector space. During proof generation, the current proof state is embedded, and a nearest-neighbor search is performed to retrieve the top-k most relevant theorems. These selected premises are then supplied as context to our language model, helping it produce more effective and accurate proof steps.

To train the embedding model, we construct a dataset of $(s, t_{\text{pos}})$ pairs. These pairs are extracted from Mathlib using the Jixia tool [17]. In this context, $s$ refers to a Lean proof state, and $t_{\text{pos}}$ refers to a theorem that was successfully applied at that step. The training process follows a two-stage approach, inspired by transfer learning techniques commonly used to improve text retrieval [22]:

- **Initial Training:** The model was initially trained on $(s, t_{\text{pos}})$ pairs using a contrastive loss function: $L(x_1, x_2) = yd(x_1, x_2)^2 + (1-y)\max(m - d(x_1, x_2), 0)^2$. where the $d(x_1, x_2)$ denotes a distance metric between embeddings, and $y = 1$ if $x_1$ and $x_2$ are similar, and $y = 0$ otherwise. The $m$ is hyperparameter that defines the minimum distance between dissimilar pairs. In our data, we treat state as $x_1$ and $t_{\text{pos}}$ as $x_2$. We adopt the in-batch negatives strategy, where other premises within the same training batch that are not associated with the current state are treated as negative examples for contrastive learning of premise relevance.

- **Hard Negative Enhanced Training:** To improve fine-grained discrimination, the initial model was used to mine hard negatives—challenging incorrect premises for each state. This process produces triplets of the form $(s, t_{\text{pos}}, t_{\text{hard\_neg}})$, where $t_{\text{hard\_neg}}$ refers to hard negative premise. We then train the model using a triplet loss function: $L(x, p, n) = \max(d(x, p) - d(x, n) + m, 0)$, where $x$ is state, $p$ is $t_{\text{pos}}$ and $n$ is $t_{\text{hard\_neg}}$. After this stage training, we obtain our final embedding model.

#### 2.2.2 Expert Iteration

Another key component of REAL-Prover is the prover model within the tactic generator. To train a strong prover model, we combine expert iteration to collect synthetic data, as shown in Figure 1(b). We begin by fine-tuning a prover model using open-source datasets. This initial model is then employed within REAL-Prover to solve formal statements produced by HERALD-AF, generating additional high-quality data. The newly collected data is incorporated into the supervised fine-tuning dataset to train an improved model. This refined model is then used to tackle challenging formal statements that the previous model could not solve. We update our model continuously in this iteration process.

## 3 Experiment Setup

### 3.1 Implementation Detail

**REAL-Prover Training** We perform supervised fine-tuning on the base models Qwen2.5-Math-7B [23] using a learning rate of $5 \times 10^{-5}$ with a cosine decay scheduler and a maximum context length of 8192 tokens. Our training data comprises the following sources, totaling 210,420 state-tactic pairs:

- Post-processed formal proofs collected via expert iteration. Each proof is parsed into state-tactic pairs using Jixia, and the corresponding prompts are enriched with natural language descriptions, the proof context, the current proof state, and retrieval information from LeanSearch-PS. The dataset comprises 15,071 algebra problems from undergraduate and graduate mathematics texts, and 19,633 NuminaMath problems [24], resulting in 25,818 and 30,589 state-tactic pairs, respectively, after post-processing.
- Post-processed, human-annotated datasets were constructed following the same procedure as expert-iteration proofs. Based on 196 annotated undergraduate-level algebra problems, the datasets contain 18,669 state-tactic pairs. Annotations were provided by undergraduate and graduate students majoring in mathematics, ensuring high quality.
- Formal proofs are extracted from Mathlib and parsed into state-tactic pairs using Jixia. Prompt inputs are further augmented with retrieval information from LeanSearch-PS, while natural language descriptions are omitted. The resulting dataset comprises 92,152 state-tactic pairs.
- The augmented Lean-Workbook dataset [25]. We enhance this dataset by adding retrieval information from LeanSearch-PS to each example, while keeping the original structure intact. The resulting dataset comprises 43,192 state-tactic pairs.

**LeanSearch-PS Training** We develop two dense retrievers from E5-mistral-7b-instruct [26] with Tevatron [27] training framework. For both initial training and hard negative enhanced training, we use LoRA to accelerate training, and set the learning rate as $2 \times 10^{-5}$, the batch size as 128, the number of epoch as 1, and other hyperparameters as their default values. Queries are informal statements with a maximum length of 128 and passages are formal theorems with a maximum length of 256. For the hard negative examples, we first embed all statements and theorems with the initial trained embedding model, and then randomly select one passage from the top 30 to top 100 most similar ones for each query as its hard negative premise.

**REAL-Prover Inference** In our experiments, considering the overall computational budget, we set the number of passes and samples per step to 64. We set our LLM's temperature to 1.5. Additionally, we set the hyperparameter $\alpha$ of the score function to 0.5, consistent with the configuration used in BFS-Prover [14]

### 3.2 Benchmark

As our prover highlights its capabilities on college-level mathematics, we created our own benchmark, FATE-M, which focuses specifically on algebraic formal statements to test model's performance on rich algebra structures. In addition, we also follow prior work [11, 28, 29, 30] to evaluate our prover's general capabilities using the ProofNet and MiniF2F benchmarks.

**ProofNet** ProofNet consists of 185 validation problems and 186 test problems, drawn from widely used undergraduate pure mathematics textbooks. The dataset spans a variety of topics, including real and complex analysis, linear algebra, abstract algebra, and topology.

**MiniF2F** MiniF2F includes 244 validation problems and 244 test problems, collected from a diverse set of mathematical sources such as the AIME, AMC, and IMO competitions. This benchmark focuses on Olympiad-level problems covering core areas of elementary mathematics, including algebra, number theory, and mathematical induction.

**FATE-M (Formal Algebra Theorem Evaluation-Medium)** FATE-M is a benchmark designed to evaluate theorem-proving capabilities in Lean4 for undergraduate-level abstract algebra. It consists of 141 problems formalized in Lean4, most accompanied by natural language comments in English or Chinese. Focusing on core topics in abstract algebra, FATE-M spans problems of simple to moderate difficulty, all of which rely on definitions and lemmas already formalized in Lean's Mathlib. To our knowledge, it is the first benchmark specifically targeting this domain, offering a specialized tool for assessing formal reasoning in algebra. Future expansions of the FATE series will include more advanced topics, such as commutative algebra, with increased difficulty.

The benchmark was created by extracting problems from 12 university-level abstract algebra textbooks (for a full list, see Appendix B), which were then formalized by students. Each formalized statement was rigorously verified by Mathlib contributors and PhD students in algebra to ensure it accurately represents the original mathematical meaning. Additionally, Lean-checked proofs are provided for all statements to guarantee their correctness.

We present two examples to illustrate the nature of the benchmark:

```
import Mathlib

example {G : Type*} [Group G] {A B : Subgroup G} :
    (∃ C : Subgroup G, C.carrier = (A.carrier ∪ B.carrier)) ↔ A ≤ B ∨ B ≤ A := by
```

This is a medium-level problem in the benchmark. This problem translates to the following natural language statement: Let $A$, $B$, and $C$ be subgroups of a group $G$. If $C = A \cup B$, then either $A \subseteq B$ or $B \subseteq A$.

```
import Mathlib

open Classical

/-Prove that a group of order $p^{2}, p$ a prime, has a normal subgroup of order $
    p$.-/
example {G : Type*} {p : ℕ} [Group G] [Fintype G] (pp : p.Prime) (ord :
    Fintype.card G = p ^ 2) : ∃ P : Subgroup G, (P.Normal) ∧ (Fintype.card P = p)
    :=
```

This problem represents one of the more challenging exercises in the benchmark. The natural language translation precisely matches the Lean code comment. An equivalent formulation would be: Let $p$ be a prime number. Then every group of order $p^2$ contains a normal subgroup of order $p$.

## 4 Quantitative Result

Our main results are on ProofNet and FATE-M, both of which consist of college-level mathematics problems that require the use of numerous theorems. To evaluate the generalization ability of our

Table 1: Comparison with state-of-the-art models (all 7B parameters) on ProofNet.

| Prove System | Sampling Budget | ProofNet Test |
|---|---|---|
| **Whole-proof systems** | | |
| Goedel Prover [30] | 32 | 15.6 % |
| DeepSeek-Prover-V1.5-SFT [11] | 128 | 15.9 % |
| DeepSeek-Prover-V1.5-RL [11] | 128 | 18.2 % |
| **Tree search systems** | | |
| DeepSeek-Prover-V1.5-RL + RMaxTS [13] | $1 \times 3200$ | 21.6 % |
| REAL-Prover (**ours**) | $64 \times 64$ | **23.7** % |

Table 2: Comparison with other system (all 7B parameters) on FATE-M.

| Prove System | Sampling Budget | FATE-M Test |
|---|---|---|
| **Whole-proof systems** | | |
| Goedel Prover [30] | 128 | 18.7 % |
| DeepSeek-Prover-V1.5-RL [11] | 128 | 31.2 % |
| **Tree search systems** | | |
| DeepSeek-Prover-V1.5-RL + RMaxTS [13] | $64 \times 64$ | 41.8 % |
| REAL-Prover (**ours**) | $64 \times 64$ | **56.7** % |

system, we also test it on MiniF2F, a benchmark in a different domain than ProofNet and FATE-M. The results are shown below.

### 4.1 ProofNet Result

In the ProofNet benchmark, we compare REAL-Prover with several leading provers, Geodel-Prover [30], DeepSeek-Prover-V1.5 [13]. For whole-proof system, the sampling budget in table means the number of generation. And for tree-search systems, the sampling budget is M×N, where M is the total number of passes and N is the number of every step generation. The experimental results demonstrate that REAL-Prover with retrieval archieved 23.7% success rate only using supervied fine-tune. Surpassing DeepSeek-Prover-V1.5-RL under a 3200-sampling budget and DeepSeek-Prover-V1.5-RL+RMaxTS under 1×3200-sampling budget, which are trained with reinforcement learning.

### 4.2 FATE-M Result

The performance on the FATE-M benchmark is presented in Table 2. Given the availability of open-source models and pipelines, we evaluate only Goedel-Prover and DeepSeek-Prover by using their official repositories. In this experiment, our prover achieves a state-of-the-art result with a Pass@64 of 56.7%. This demonstrates that our prover exhibits strong capability in solving college-level algebraic problems.

### 4.3 MiniF2F Result

The MiniF2F benchmark has been the focus of more extensive research, as summarized in Table 3. For tree search systems, serveral provers have K×W×N sampling budget, where K represents the number of passes, W the expansion width, and N the maximum number of proof state expansions per pass. As shown in the table, our model performs relatively poorly compared to state-of-the-art models. We consider two main factors that may contribute to this performance gap. First, MiniF2F primarily features high school-level olympiad problems, which often do not heavily rely on theorems from Mathlib. Consequently, the benefit of retrieval is diminished compared to its effectiveness on

Table 3: Comparison with other system on MiniF2F.

| Prove System | Sampling Budget | MiniF2F Test |
|---|---|---|
| **Whole-proof systems** | | |
| DeepSeek-Prover-V1.5-RL [11] | 102400 | 60.2% |
| Leanabell-Prover-GD-RL [31] | 128 | 61.1% |
| Goedel-Prover [30] | 25600 | 64.7% |
| STP [32] | 25600 | 67.6% |
| **Tree search systems** | | |
| LLMStep [33] | 1×32×100 | 27.9 % |
| GPT-f [7] | 64×8×512 | 36.6 % |
| Hypertree Proof Search [34] | 64 × 5000 | 41.0 % |
| DeepSeek-Prover-V1.5-RL + RMaxTS [13] | 32 × 6400 | 63.5 % |
| InternLM2.5-StepProver-BF [12] | 256 × 32 × 600 | 65.9 % |
| HunyuanProver v16 + BFS + DC [29] | 600 × 8 × 400 | 68.4 % |
| BFS-Prover [14] | 2048 × 2 × 600 | 70.8 % |
| REAL-Prover (**ours**) | 64 × 64 | 54.1 % |

Table 4: Ablation study on ProofNet Test and FATE-M Test. REAL-Prover-v1-NoRet w/o LeanSearch refers to use the REAL-Prover-v1-NoRet model without LeanSearch-PS. REAL-Prover-v1 w/ LeanSearch refers to use the REAL-Prover-v1 model with LeanSearch-PS enabled.

| Prove System | ProofNet Test | FATE-M Test |
|---|---|---|
| REAL-Prover-v1-NoRet w/o LeanSearch | 22.6% | 44.7% |
| REAL-Prover-v1 w/ LeanSearch | 23.7% | 56.7% |

college-level problems. Second, reinforcement learning plays a significant role in the training of current state-of-the-art provers. In contrast, our model is trained solely with supervised fine-tuning, making it challenging to compete without the additional boost provided by retrieval mechanisms.

### 4.4 Ablation Study

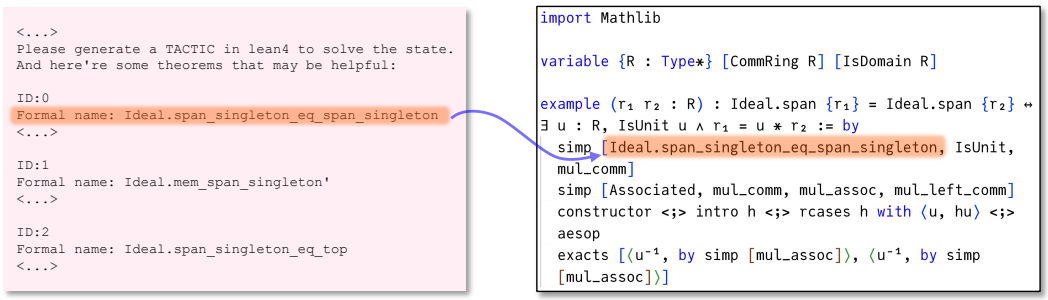
In this section, we will analysis the effective of LeanSearch-PS. To compare with REAL-Prover-v1, we trained a prover model, REAL-Prover-v1-NoRet, which used the same dataset but without retrieval information. And in REAL-Prover, we do not use LeanSearch-PS for REAL-Prover-v1-NoRet model. We evaluate both models on the ProofNet and FATE-M benchmarks with results presented in Table 4. The results show that REAL-Prover-v1 achieves a performance improvement over REAL-Prover-v1-NoRet on both benchmarks. This demonstrates that the retrieval system enhances the prover's performance on college-level mathematics problems.

We present a comparison between proofs with and without LeanSearch-PS. In Figure 3, the proof assisted by LeanSearch-PS successfully retrieves critical lemmas and completes verification, whereas the proof without LeanSearch-PS fails to verify. For problems that can be resolved both with LeanSearch-PS and without LeanSearch-PS, the one with LeanSearch-PS often leads to more accurate and human-readable proof, we provide 2 examples in Appendix E to illustrate this observation.

## 5 Related Work

Automated Theorem Proving in formal languages, particularly within the framework of Interactive Theorem Provers (ITPs) [3, 4], is a rapidly advancing research area in artificial intelligence. LLM-based provers can be broadly categorized by their proof generation strategies: whole-proof generation that aims to produce complete proof in a single pass, and proof-step generation that iteratively refines proofs through ITPs.

The whole-proof generation paradigm focuses on end-to-end synthesis of formal proof scripts from theorem statements. Early attempts like DeepSeek-Prover-V1 [35] demonstrated the feasibility of this approach in Lean 4. Subsequent improvements in DeepSeek-Prover-V1.5 [13] introduced hybrid techniques combining Reinforcement Learning from Proof Assistant Feedback (RLPAF) with Group Relative Policy Optimization [36], along with a 'truncate-and-resume' mechanism that allows error correction through partial stepwise verification. Goedel-Prover [30] addressed data scarcity through auto-formalization, first converting natural language mathematics into formal statements before generating proofs through iterative training. While these approaches benefit from large-scale supervised fine-tuning and reduced ITP interaction, they struggle with error accumulation in long logical chains and implicit state inference.




Prompt with Retrieved Theorems

Proof Using Retrieved Theorems

Figure 3: Proof using LeanSearch-PS and retrieval results. The left shows part of our input prompt; the right displays the proof generated by REAL-Prover.

This limitation has driven the development of interactive proof-step generation approaches that leverage ITPs feedback during the proving process. Systems like BFS-Prover [14] employ expert iteration with Best-First Search, using Lean compiler feedback to guide tactic generation. HunyuanProver [29] enhances search efficiency through specialized critic models that evaluate intermediate proof states. The LeanDojo framework introduces ReProver [18], which adopts a distinct retrieval-based approach using Dense Passage Retriever [37] to directly predict relevant premises. LeanAgent [38] extends this paradigm through lifelong learning, maintaining a dynamic knowledge base to progressively master theorem proving across multiple repositories. These interactive methods typically combine expert iteration, enabling real-time adaptation to proof state changes while managing computational complexity through guided tree search.

## 6 Conclusions

In this work, we propose REAL-Prover, a stepwise proof system designed to tackle challenging mathematical problems. To support the development of this system, we introduce the HERALD-AF pipeline, which translates informal mathematical statements into formal Lean 4 statements, to obtain a large-scale formal dataset. We also developed a new Lean 4 interactive environment, Jixia-Interactive, which facilitates both the training of our prover model, REAL-Prover-v1, and formal proof generation. To further enhance the prover's capabilities, we present LeanSearch-PS, a theorem retrieval system that boosts performance on the college-level mathematics problem benchmark. Additionally, to further evaluate the prover's effectiveness, we introduce the FATE-M benchmark, where REAL-Prover achieves a state-of-the-art success rate of 56.7%.

## Limitations

Despite its strong results, REAL-Prover still has two notable limitations. First, the current model is trained purely with supervised objectives. We refrained from adding reinforcement-learning (RL) stages because they demand substantially more compute and careful reward-design to avoid reward hacking. Given the evidence that RL can expose latent capabilities in large language models, incorporating RL objectives (e.g. GRPO[36]) is a promising avenue for unlocking further gains. Secondly, REAL-Prover emits only the final Lean tactics, whereas several leading provers first

generate a natural-language chain of thought (CoT) and then translate it into formal steps, enabling powerful test-time reasoning amplification. Introducing an informal-reasoning layer would let the model draft, filter, and organize candidate arguments before committing to formal tactics, potentially improving both premise selection and high-level proof planning. Addressing these two aspects—RL fine-tuning and CoT-based reasoning—forms the core of our future work.

## Acknowledgements

This work is supported in part by National Key R&D Program of China grant 2024YFA1014000, the New Cornerstone Investigator Program, and Ubiquant.

## References

[1] Georges Gonthier. The four colour theorem: Engineering of a formal proof. In Deepak Kapur, editor, *Computer Mathematics*, pages 333–333, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[2] Georges Gonthier, Andrea Asperti, Jeremy Avigad, Yves Bertot, Cyril Cohen, François Garillot, Stéphane Le Roux, Assia Mahboubi, Russell O'Connor, Sidi Ould Biha, Ioana Pasca, Laurence Rideau, Alexey Solovyev, Enrico Tassi, and Laurent Théry. A machine-checked proof of the odd order theorem. In Sandrine Blazy, Christine Paulin-Mohring, and David Pichardie, editors, *Interactive Theorem Proving*, pages 163–179, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[3] Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language. In André Platzer and Geoff Sutcliffe, editors, *Automated Deduction – CADE 28*, pages 625–635, Cham, 2021. Springer International Publishing.

[4] Yves Bertot and Pierre Castéran. *Interactive theorem proving and program development: Coq'Art: the calculus of inductive constructions*. Springer Science & Business Media, 2013.

[5] The mathlib Community. The lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, POPL '20. ACM, January 2020.

[6] Guoxiong Gao, Haocheng Ju, Jiedong Jiang, Zihan Qin, and Bin Dong. A semantic search engine for mathlib4. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8001–8013, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[7] Stanislas Polu and Ilya Sutskever. Generative language modeling for automated theorem proving. *arXiv preprint arXiv:2009.03393*, 2020.

[8] Albert Qiaochu Jiang, Sean Welleck, Jin Peng Zhou, Timothee Lacroix, Jiacheng Liu, Wenda Li, Mateja Jamnik, Guillaume Lample, and Yuhuai Wu. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. In *The Eleventh International Conference on Learning Representations*, 2023.

[9] Xueliang Zhao, Wenda Li, and Lingpeng Kong. Subgoal-based demonstration learning for formal theorem proving. In *Forty-first International Conference on Machine Learning*, 2024.

[10] Haiming Wang, Huajian Xin, Chuanyang Zheng, Zhengying Liu, Qingxing Cao, Yinya Huang, Jing Xiong, Han Shi, Enze Xie, Jian Yin, Zhenguo Li, and Xiaodan Liang. LEGO-prover: Neural theorem proving with growing libraries. In *The Twelfth International Conference on Learning Representations*, 2024.

[11] Huajian Xin, Daya Guo, Zhihong Shao, Z.Z. Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Advancing theorem proving in LLMs through large-scale synthetic data. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*, 2024.

[12] Zijian Wu, Suozhi Huang, Zhejian Zhou, Huaiyuan Ying, Jiayu Wang, Dahua Lin, and Kai Chen. Internlm2.5-stepprover: Advancing automated theorem proving via expert iteration on large-scale lean problems. *arXiv preprint arXiv:2410.15700*, 2024.

[13] Huajian Xin, Z.Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Haowei Zhang, Qihao Zhu, Dejian Yang, Zhibin Gou, Z.F. Wu, Fuli Luo, and Chong Ruan. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. In *The Thirteenth International Conference on Learning Representations*, 2025.

[14] Ran Xin, Chenguang Xi, Jie Yang, Feng Chen, Hang Wu, Xia Xiao, Yifan Sun, Shen Zheng, and Kai Shen. Bfs-prover: Scalable best-first tree search for llm-based automatic theorem proving. *arXiv preprint arXiv:2502.03438*, 2025.

[15] Z. Z. Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, Z. F. Wu, Zhibin Gou, Shirong Ma, Hongxuan Tang, Yuxuan Liu, Wenjun Gao, Daya Guo, and Chong Ruan. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801*, 2025.

[16] Guoxiong Gao, Yutong Wang, Jiedong Jiang, Qi Gao, Zihan Qin, Tianyi Xu, and Bin Dong. Herald: A natural language annotated lean 4 dataset. In *The Thirteenth International Conference on Learning Representations*, 2025.

[17] frenzymath. jixia: A static analysis tool for lean 4. `https://github.com/frenzymath/jixia`, 2024.

[18] Kaiyu Yang, Aidan M Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. Leandojo: Theorem proving with retrieval-augmented language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[19] Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *CoRR*, abs/2109.00110, 2021.

[20] Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*, 2023.

[21] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2025.

[22] Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. SFR-Embedding-Mistral: Enhance text retrieval with transfer learning. Salesforce AI Research Blog, 2024.

[23] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[24] Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. *Hugging Face repository*, 2024.

[25] Huaiyuan Ying, Zijian Wu, Yihan Geng, JIayu Wang, Dahua Lin, and Kai Chen. Lean workbook: A large-scale lean problem set formalized from natural language math problems. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[26] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[27] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Tevatron: An efficient and flexible toolkit for neural retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3120–3124, New York, NY, USA, 2023. Association for Computing Machinery.

[28] Ruida Wang, Jipeng Zhang, Yizhen Jia, Rui Pan, Shizhe Diao, Renjie Pi, and Tong Zhang. TheoremLlama: Transforming general-purpose LLMs into lean4 experts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11953–11974, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[29] Yang Li, Dong Du, Linfeng Song, Chen Li, Weikang Wang, Tao Yang, and Haitao Mi. Hunyuanprover: A scalable data synthesis framework and guided tree search for automated theorem proving. *arXiv preprint arXiv:2412.20735*, 2024.

[30] Yong Lin, Shange Tang, Bohan Lyu, Jiayun Wu, Hongzhou Lin, Kaiyu Yang, Jia Li, Mengzhou Xia, Danqi Chen, Sanjeev Arora, and Chi Jin. Goedel-prover: A frontier model for open-source automated theorem proving. *arXiv preprint arXiv:2502.07640*, 2025.

[31] Jingyuan Zhang, Qi Wang, Xingguang Ji, Yahui Liu, Yang Yue, Fuzheng Zhang, Di Zhang, Guorui Zhou, and Kun Gai. Leanabell-prover: Posttraining scaling in formal reasoning. *arXiv preprint arXiv:2504.06122*, 2025.

[32] Kefan Dong and Tengyu Ma. Stp: Self-play llm theorem provers with iterative conjecturing and proving. *arXiv preprint arXiv:2502.00212*, 2025.

[33] Sean Welleck and Rahul Saha. Llmstep: Llm proofstep suggestions in lean. *arXiv preprint arXiv:2310.18457*, 2023.

[34] Guillaume Lample, Timothee Lacroix, Marie anne Lachaux, Aurelien Rodriguez, Amaury Hayat, Thibaut Lavril, Gabriel Ebner, and Xavier Martinet. Hypertree proof search for neural theorem proving. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[35] Huajian Xin, Daya Guo, Zhihong Shao, Z.Z. Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Advancing theorem proving in LLMs through large-scale synthetic data. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*, 2024.

[36] Qihao Zhu Zhihong Shao, Peiyi Wang et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

[37] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020.

[38] Adarsh Kumarappan, Mo Tiwari, Peiyang Song, Robert Joseph George, Chaowei Xiao, and Anima Anandkumar. Leanagent: Lifelong learning for formal theorem proving. In *The Thirteenth International Conference on Learning Representations*, 2025.

# A Prompt for REAL-Prover

In Lean, a formal proof is a fully constructed proof term that is type-checked and verified by the kernel. It represents a complete and correct derivation of a proposition.
The state after tactics refers to the intermediate proof state during tactic-based proof construction. It includes the list of remaining goals and the local context at that point.

Relationship:

1. Tactics are procedural tools used to incrementally construct a formal proof.
2. Each tactic transforms the current proof state by solving or reducing goals.
3. The state after a tactic reflects the goals that still need to be proven after that tactic has been applied.
4. Once all goals are solved, Lean assembles the underlying proof terms generated by the tactics into a complete formal proof.
5. This final term is then type-checked by the kernel to ensure correctness.

In essence, the state after tactics shows where you are in the process of building a formal proof — it's a snapshot of what's left to do before the proof is complete.

Here is the FORMAL PROOF before the current state:

```
import Mathlib
open BigOperators
open Real Nat Topology
/-- Let \( A = (x_1, y_1) \), \( B = (x_2, y_2) \), and \( C = (x_3, y_3) \) be
    three vertices of a parallelogram. Prove that if \( D = (x_4, y_4) \) is
    the fourth vertex such that the midpoints of the diagonals coincide, then
    \( x_4 + x_1 + x_2 + x_3 = 2(x_{\text{mid}} + x_{\text{mid}}') \), where \(
    x_{\text{mid}} \) and \( x_{\text{mid}}' \) are the \( x \)-coordinates of
    the midpoints of any two pairs of opposite sides. -/
theorem parallelogram_midpoints_coincide_extracted {A B C D : ℂ} : (A + B + C +
    D) / 4 = (A + C) / 2 → (B + D) / 2 = (A + C) / 2 → B.re + D.re = 2 * ((A +
     C) / 2).re  := by
simp [add_comm, add_left_comm, add_assoc]
simp [div_eq_mul_inv, mul_add, mul_comm, mul_left_comm]
intro h₁ h₂
```

Here is the current STATE:

```
A : ℂ
B : ℂ
C : ℂ
D : ℂ
h₁ : (A + (B + (C + D))) * 4⁻¹ = (A + C) * 2⁻¹
h₂ : B + D = A + C
⊢ B.re + D.re = A.re + C.re
```

Please generate a tactic in lean4 to solve the state.
And here're some theorems that may be helpful:
ID:0
Formal name:`Complex.add_re`
Informal name: Real Part of Sum of Complex Numbers
Formal statement: `∀ (z w : ℂ), (z + w).re = z.re + w.re`
...
ID:5
Formal name: `add_add_add_comm`
Informal name: Commutativity and Associativity of Addition in Additive Commutative Semigroups
Formal statement: `∀ {G : Type u_3} [inst : AddCommSemigroup G] (a b c d : G), a + b + (c + d) = a + c + (b + d)`

Table 5: REAL-Prover training hyperparameters

| **Component** | **Setting** |
|---|---|
| Full Fine Tuning | Learning rate : $5 \times 10^{-5}$; Scheducler: Cosine Decay |
| Backbone | Qwen2.5-Math-7B |
| Length | Prompt Max Length: 8192 |
| Optimizations | `bf16, flash_attn` |
| Batch Sizes | `train_batch_size = 2` |
| Training Schedule | 3 Epochs |

Table 6: LeanSearch-PS training hyperparameters

| **Component** | **Setting** |
|---|---|
| LoRA Fine Tuning | Learning rate : $2 \times 10^{-5}$ |
| Backbone | E5-mistral-7b-instruct |
| Optimizations | `bf16, flash_attn` |
| Batch Sizes | `train_batch_size = 2` |
| Training Schedule | 1 Epoch |

# B Source Textbooks for FATE-M

The FATE-M benchmark draws problems from the following abstract algebra textbooks:

1. Pinter, C. C. *A Book of Abstract Algebra* (2nd ed., 2010)
2. Fraleigh, J. B. *A First Course in Abstract Algebra* (8th ed., 2021)
3. Anderson, M. & Feil, T. *A First Course in Abstract Algebra: Rings, Groups, and Fields* (3rd ed., 2015)
4. Rotman, J. J. *A First Course in Abstract Algebra with Applications* (3rd ed., 2005)
5. Hodge, J. K., Schlicker, S., & Sundstrom, T. *Abstract Algebra: An Inquiry-Based Approach* (2014)
6. Herstein, I. N. *Abstract Algebra* (3rd ed., 1996)
7. Grillet, P. A. *Abstract Algebra* (2nd ed., 2007)
8. Choudhary, P. *Abstract Algebra* (2008)
9. Saracino, D. *Abstract Algebra: A First Course* (2nd ed., 2008)
10. Hungerford, T. W. *Abstract Algebra: An Introduction* (3rd ed., 2014)
11. Dummit, D. S. & Foote, R. M. *Abstract Algebra* (3rd ed., 2004)
12. Feng, K. & Zhang, P. *300 Problems in Abstract Algebra* (2009)

# C Training Details

**REAL-Prover training** REAL-Prover training is performed on 8×A800 GPUs (80GB each) for approximately 18 hours, with hyperparameters detailed in Table 5.

**LeanSearch-PS training** LeanSearch-PS training is performed on 4×L40 GPUs (40GB each) for approximately 12 hours with hyperparameters detailed in Table 6.

# D Broader Impacts

This work explores the use of large language models (LLMs) for automated formal theorem proving. Potential positive impacts include:

```
example  {F : Type*} [Field F]
    [Fintype F] [IsAlgClosed F] :
    False := by
  set e : Fintype F :=
    Fintype.ofFinite F
  set L := (AlgebraicClosure F)
  have : Infinite L := by exact
    IsAlgClosed.instInfinite
  have := this
  have := Fintype.ofFinite F
  exact Fintype.false e
```

(a) Proof with LeanSearch-PS.

```
example  {F : Type*} [Field F] [Fintype
    F] [IsAlgClosed F] : False := by
  have : Fact (Nat.Prime 2) := ⟨by
    norm_num⟩
  exact not_finite F
```

(b) Proof without LeanSearch-PS.

Figure 4: Compare the proofs with and without LeanSearch-PS. The prover in (a) uses the existing instance 'IsAlgClosed.instInfinite' from Mathlib, resulting in a more readable proof.

```
example {G H : Type*} {p : ℕ} [Group
    G] [Group H] [Fact p.Prime] (gp
    : IsPGroup p G) (f : G →* H)
    (sf : Function.Surjective f) :
    IsPGroup p H := by
  exact IsPGroup.of_surjective gp f
    sf
```

(a) Proof with LeanSearch-PS.

```
example {G H : Type*} {p : ℕ} [Group G]
    [Group H] [Fact p.Prime] (gp :
    IsPGroup p G) (f : G →* H) (sf :
    Function.Surjective f) : IsPGroup p
    H := by
  rw [IsPGroup] at gp ⊢
  intro g
  obtain ⟨m, hm⟩ :=sf g
  rcases gp m with ⟨km, hkm⟩
  use km
  rw [← hm, ← map_pow, hkm, map_one f]
```

(b) Proof without LeanSearch-PS.

Figure 5: Compare the proofs with and without LeanSearch-PS. The prover in (a) uses the existing instance 'IsPGroup.of_surjective' from Mathlib, resulting in a more readable proof.

1. assisting mathematicians in constructing formal proofs more efficiently by automating tedious or routine steps;
2. facilitating the teaching of formal proof assistants such as Lean in undergraduate and graduate curricula;
3. potentially enabling LLMs to autonomously explore new conjectures or discover novel proofs, contributing to mathematical discovery.

However, several risks and limitations should be acknowledged. First, while LLMs may generate correct formal proofs, these proofs can often be opaque or unintuitive, reducing their explanatory value for human users. Second, excessive reliance on automated proof generation may hinder the development of deep mathematical understanding, particularly in educational settings. Third, the reinforcement learning and training process may be susceptible to reward hacking, especially given the complexity and non-determinism of the Lean environment—raising the risk of generating formally valid but semantically meaningless or misleading proofs that may be overlooked by human reviewers.

Careful evaluation, interpretability mechanisms, and human-in-the-loop verification are essential to ensure that such systems are used responsibly and contribute positively to both research and education.

## E Case study

We provide 2 cases in FATE-M to illustrate that even if both REAL-Prover-v1 with LeanSearch-PS and REAL-Prover-v1-NoRet without Leansearch-PS produce the valid results, the one with premise selection often produce faster and readable proofs, as illustrated in Figure 4 and Figure 5.