
ATLAS: Autoformalizing Theorems through Lifting, Augmentation, and Synthesis of Data

Xiaoyang Liu¹, Kangjie Bao¹, Jiashuo Zhang¹, Yunqi Liu¹, Yu Chen¹,
Yuntian Liu¹, Yang Jiao^{3,4}*, Tao Luo^{1,2}*

¹ School of Mathematical Sciences, Shanghai Jiao Tong University

² Institute of Natural Sciences, MOE-LSC, CMA-Shanghai, Shanghai Jiao Tong University

³ SPEIT, Shanghai Jiao Tong University

⁴ JoinTech Co., Ltd

{jiaoyang2002, luotao41}@sjtu.edu.cn

Abstract

Autoformalization, the automatic translation of mathematical content from natural language into machine-verifiable formal languages, has seen significant progress driven by advances in large language models (LLMs). Nonetheless, a primary barrier to further improvements is the limited availability of parallel corpora that map informal mathematical text to its formal counterpart. To address this limitation, we propose ATLAS (Autoformalizing Theorems through Lifting, Augmentation, and Synthesis of Data), a novel data generation framework designed to produce large-scale, high-quality parallel corpora of theorem statements. Distinct from prior approaches, ATLAS begins with a concept repository, accelerates the improvement of the student model through expert iteration combined with knowledge distillation, and introduces two novel augmentation strategies that exploit the structural characteristics of formal languages. Running the proposed ATLAS framework for 10 iterations, we construct an undergraduate-level dataset of 117k theorem statements and develop the ATLAS Translator by fine-tuning Llama3.1-8B-Instruct with LoRA. This model establishes a new state of the art, demonstrating statistically significant improvements over both the Herald Translator and the Kimina-Autoformalizer across all benchmarks ($p < 0.05$, two-sided t-test). Furthermore, we demonstrate that the full-parameter fine-tuning of a stronger base model on the ATLAS dataset leads to superior performance. The datasets, model, and code are available at <https://github.com/XiaoyangLiu-sjtu/ATLAS>.

1 Introduction

In modern mathematics, the escalating complexity of proofs, combined with the increasing reliance on computer-assisted arguments, has raised substantial concerns about reliability. Errors in traditional proofs can remain undetected for extended periods, while computer-assisted proofs frequently lack transparency and are difficult to verify manually, thereby raising issues of trust within the mathematical community. For example, the Four Color Theorem’s 1879 proof went unchallenged for over a decade before its flaw was discovered. The first computer-assisted proof in 1976 raised concerns due to its unverifiable computations, prompting further debate. Only in 2005 was the proof formally verified using Coq [4]. To address such issues, formal languages like Isabelle [29], HOL Light [10], Coq, and Lean [7] have been developed to rigorously verify the correctness of proofs.

However, writing mathematical content in formal languages requires significant time and effort, as well as a deep familiarity with these languages, making the process highly labor-intensive. This

*Corresponding authors: Yang Jiao, Tao Luo

highlights the critical importance of autoformalization, which aims to translate theorem statements and proofs from natural language (NL) into their formal language (FL) counterparts [34]. Since the precise formalization of statements can provide valuable training data for automated theorem proving [21], current research primarily focuses on the autoformalization of theorem statements. In this context, recent progress has shown encouraging results, primarily achieved by fine-tuning large language models (LLMs) with parallel corpora of theorem statements. For clarity, we hereafter refer to theorem statements expressed in natural language as **NL statements**, those expressed in formal language as **FL statements**, and their pairs as **parallel statements**. Furthermore, while ATLAS is a general framework for any formal language, this work focuses on Lean 4 [27] as the target language.

To construct parallel statements, previous studies such as MMA [15] and Herald [9] extract FL statements from Mathlib [35] and generate their NL counterparts using LLMs. However, the limited size of Mathlib imposes restrictions on the scale of the resulting datasets. Alternative methods, including Lean Workbook [51] and DeepSeek-Prover [46], attempt to generate FL statements from NL sources obtained via large-scale web scraping. Although this approach greatly alleviates the limitations on dataset size, it requires extensive pre-processing to obtain high-quality, formalizable NL statements, which significantly reduces overall efficiency. Consequently, it is essential to design a more effective method for generating large-scale, high-quality parallel statements.

In this work, we introduce ATLAS, a data generation framework composed of three key components: Data Lifting, Data Synthesis, and Data Augmentation.

- **Data Lifting.** Unlike previous approaches, this work integrates FL statements and NL statements as its starting point. Specifically, mathematical concepts are abstracted and extracted directly from Mathlib to synthesize NL statements. This method not only overcomes scale limitations but also eliminates the need for data pre-processing.
- **Data Synthesis.** Adopting the knowledge distillation [11] paradigm, teacher models guide the student model’s learning process, and the Lean compiler is jointly employed to ensure that the generated FL statements are both semantically accurate and syntactically valid. The resulting parallel statements are hereafter referred to as "synthetic data".
- **Data Augmentation.** The synthetic data are further expanded using two techniques: augmentation via proof and contraposition. The core idea is to leverage Lean 4’s capability whereby the Infoview provides real-time updates of the current state after each proof step. The additional parallel statements generated are hereafter referred to as "augmented data".

Finally, the synthetic and augmented data are combined to fine-tune the student model. The expert iteration [30, 31] approach is then employed to iteratively execute ATLAS. After 10 iterations, we build an undergraduate-level dataset comprising 117k parallel statements and train the ATLAS Translator. The performance of the ATLAS Translator shows statistically significant improvements over both the Herald Translator and the Kimina-Autoformalizer across all benchmarks.

Our main contributions are as follows:

1. We propose ATLAS, a novel framework for generating large-scale, high-quality parallel statements. Unlike previous work that starts from NL statements or FL statements, our innovative approach begins by extracting concepts directly from Mathlib through a process we call Data Lifting. Based on these concepts, we employ Data Synthesis and Data Augmentation to synthesize and augment the parallel statements.
2. We introduce the ATLAS dataset and the MathQual dataset. The former comprises 117k undergraduate-level parallel statements, making it one of the largest available. In contrast, the latter contains 465 graduate-level natural language statements, designed to assess the model’s autoformalization capability on more challenging data.
3. We develop the ATLAS Translator, which establishes a new state of the art by demonstrating statistically significant improvements over strong baselines across all benchmarks ($p < 0.05$, two-sided t-test). Furthermore, we demonstrate that the full-parameter fine-tuning of a stronger base model on the ATLAS dataset leads to superior performance.

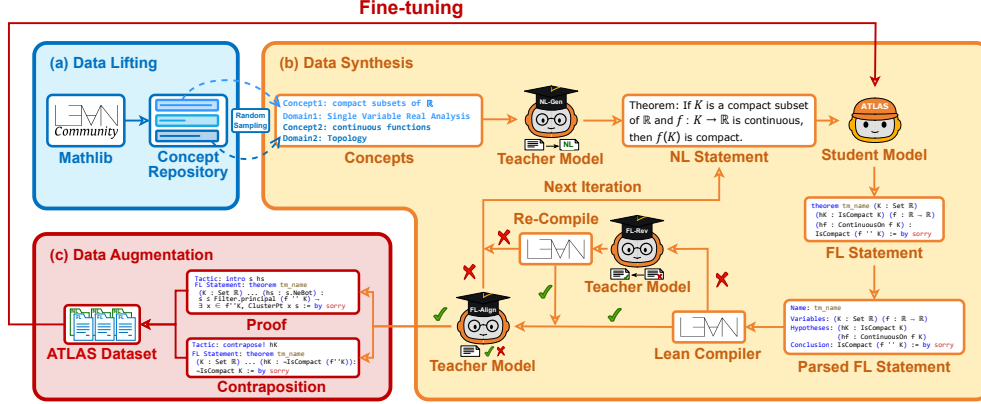


Figure 1: The overview of the proposed ATLAS framework.

2 Related Work

Autoformalization. The task of autoformalization can be seen as a machine translation problem [39], aiming to convert natural language content into expressions consistent with the target formal language’s syntax and vocabulary. Early approaches [6, 40] have utilized neural machine translation techniques to address the autoformalization of theorem statements. With the rapid advancement of LLMs, recent research on LLM-based autoformalization can be broadly categorized into three main paradigms. First, researchers [1, 43, 58] have explored few-shot prompting to enable LLMs to perform autoformalization effectively. Second, some methods [2, 25, 26] further enhance performance by fine-tuning LLMs with parallel statements. Finally, retrieval-augmented generation techniques have been combined with LLMs [54] to achieve additional improvements.

Meanwhile, another line of research focuses on the autoformalization of proofs [16, 55], a more challenging task that closely resembles a simplified form of automated theorem proving [3, 19, 32, 38, 44, 48]. For example, DSP [16] leverages LLMs to generate informal proofs, which are subsequently mapped to formal proof sketches. These sketches then serve as guidance for automated theorem provers to fill in the remaining proof gaps.

Dataset Generation. Obtaining large-scale, high-quality parallel corpora of theorem statements remains a significant challenge. Previous efforts [9, 15, 45] have tackled this problem by extracting FL statements from relevant repositories (e.g., Mathlib) and using LLMs to generate their NL counterparts. However, the limited size of these repositories constrains the scalability of the resulting datasets. On the other hand, some approaches [18, 46, 51] take the opposite direction by collecting NL statements from large-scale web sources and translating them into FL representations. While this strategy enables the creation of large-scale datasets, these web-based pipelines rely on extensive preprocessing to filter high-quality, formalizable NL statements, thereby diminishing overall efficiency.

3 Methodology

Our framework ATLAS, as illustrated in Figure 1, comprises three components: Data Lifting, Data Synthesis, and Data Augmentation. The framework begins with data lifting, which constructs the concept repository, as described in Section 3.1. Building on this foundation, Section 3.2 details the subsequent data synthesis workflow. Finally, Section 3.3 explains our approach to data augmentation.

3.1 Data Lifting

Mathlib. Mathlib [35], the most extensive mathematical library within the Lean community, provides a vast collection of formalized notations (e.g., $\|\cdot\|$), concepts, and theorems. This wealth of resources forms the cornerstone of autoformalization. Consequently, importing Mathlib is practically essential before translating a NL statement into a FL statement.

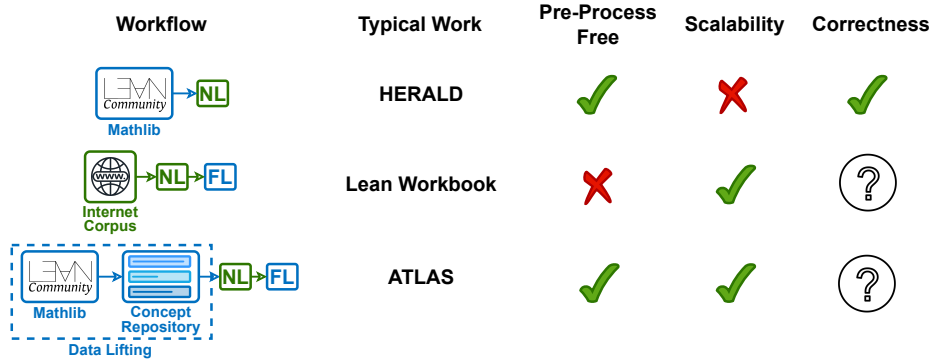


Figure 2: Comparison of different methods for constructing parallel statements.

However, the reliance on Mathlib reveals a critical limitation: when NL statements involve mathematical concepts absent in Mathlib, such as subgradients, the autoformalization process is prone to failure. This issue has affected prior work that begins by collecting NL statements. For example, Lean Workbook uses LLMs to categorize 458,692 NL statements and selects 327,870 based on this classification, primarily aiming to exclude NL statements involving concepts not present in Mathlib. In contrast, extracting FL statements from Mathlib does not face this challenge, but the library’s size constrains the scalability of the resulting datasets.

In contrast to previous work, our proposed method takes a novel approach by beginning with the extraction of concepts directly from Mathlib through a process we refer to as Data Lifting. By utilizing these concepts, we employ LLMs to synthesize NL statements. As shown in Figure 2, this starting point eliminates the need for pre-processing while ensuring scalability.

Correctness. For synthetic data, an important consideration is the correctness of the mathematical propositions themselves. However, in the context of autoformalization, the correctness of the mathematical proposition is secondary, or even insignificant. What is critical in this synthetic process is that the formalized statement is not only syntactically valid but also semantically equivalent to its natural language counterpart. This perspective stems from the primary objective of autoformalization and reflects the realities of mathematical practice, as the correctness of propositions or conjectures is seldom known a priori. Therefore, prior to formal verification, propositions must be precisely formalized in formal language—even if they ultimately turn out to be incorrect.

For example, the following example from the Lean Workbook illustrates that, although the NL statement itself is incorrect (due to the lack of a declaration for the scope of a, b, c), the FL statement is semantically equivalent to the NL statement and is syntactically valid. Therefore, this constitutes a valuable piece of synthetic data. We further explore the limitations about correctness in Appendix A.

```
# lean_workbook_plus_62
3(a^2b + b^2c + c^2a) ≤ (ab + bc + ca)^2 ≤ 9
theorem lean_workbook_plus_62 : ∀ a b c : ℝ, 3 * (a ^ 2 * b + b ^ 2 * c + c ^ 2 * a) ≤ (a * b + b * c + c * a) ^ 2 ∧ (a * b + b * c + c * a) ^ 2 ≤ 9 := by sorry
```

In the subsequent experimental section, we construct the concept repository based on undergraduate-level mathematical content² included in Mathlib. Following Mathlib’s organization by domain, topic, and concept, our concept repository comprises 13 domains, 55 topics, and 350 concepts. Further details regarding the composition of the concept repository are provided in Appendix C.

²<https://github.com/leanprover-community/mathlib4/blob/master/docs/undergrad.yaml>

3.2 Data Synthesis

Upon establishing the concept repository, the module is dedicated to producing a substantial number of high-quality parallel statements by means of knowledge distillation. The following subsections provide a detailed description of each phase, with the specific prompts provided in Appendix F.

NL Statements Generation. In this context, the teacher model serves the role of NL Statements Generation (NL-Gen) by randomly sampling concepts from the constructed concept repository to synthesize NL statements. To balance diversity and feasibility, we follow previous approaches [14, 33] by sampling two concepts for each NL statement and the detailed comparisons against MUSTARD [14] are provided in Appendix A. This choice is motivated by the observation that using only a single concept often leads the LLM to generate NL statements that are biased toward well-known, classic mathematical propositions, thus limiting diversity. Conversely, requiring each NL statement to involve many distinct concepts would be overly restrictive, as such problems are uncommon in mathematics and may exceed the capabilities of LLMs.

NL Statements Translation. For the synthetic NL statements described above, we employ the student model to translate them into the corresponding FL statements, thereby enabling subsequent tests of syntactic validity and semantic accuracy to ensure the quality of the parallel statements.

FL Statements Parsing. Before conducting syntactic validity test on these FL statements, we use the tactic `#check` to decompose each FL statement into the following four components: `theorem_name`, `theorem_variables`, `theorem_hypotheses`, and `theorem_conclusion`. The content is then systematically organized line by line, both within and across these components.

Compared to presenting the entire content on a single line, this line-by-line configuration, especially for statements with substantial content, significantly enhances the clarity of compiler feedback. In particular, error locations become much more explicit, enabling LLMs to more effectively identify and correct errors in FL statements that fail to compile.

FL Statements Compilation. In this phase, the Lean compiler is employed to verify the syntactic validity of FL statements by determining whether they can be compiled successfully. If compilation fails, detailed error messages are returned, specifying the location and cause of the error, thereby enabling more efficient subsequent revision. In addition, as the student model is unable to generate headers, a standard header, `import Mathlib`, is automatically appended prior to compilation.

FL Statements Revision. For FL statements that fail to compile, we utilize the corresponding NL statements and compilation error messages as context, providing this information to the teacher model serving as FL Statements Revision (FL-Rev) for modification. The modified FL statements are then subjected to a second round of compilation.

Unlike conventional knowledge distillation, the teacher model in our approach is tasked with revising the FL statements generated by the student model based on the provided error messages, rather than generating FL statements directly. There are two primary reasons for this design. First, it is generally much easier to revise existing FL statements than to construct them anew; consequently, this strategy is more likely to produce syntactically valid FL statements after modification. Second, as the performance of the student model improves through iterative learning, the need for FL-Rev diminishes, thus maintaining the efficiency of the overall framework.

FL Statements Alignment. For FL statements that pass either the first or second compilation, the teacher model acting as the FL Statements Alignment (FL-Align) evaluator assesses their semantic accuracy in translating the corresponding NL statement, ensuring no information is omitted or mistranslated. Specifically, the model assesses each pair of parallel statements and assigns a rating from three categories: good, average, or poor. Pairs rated as good or average are incorporated into the synthetic data, while those rated as poor, along with FL statements that fail both compilations, have only their NL statements preserved for the next iteration.

3.3 Data Augmentation

This module augments the synthetic data obtained in Section 3.2 using two innovative methods: proving these FL statements and converting them into their contrapositives, in order to further expand the scale of the resulting dataset. Figure 3 illustrates an example of the data augmentation process.

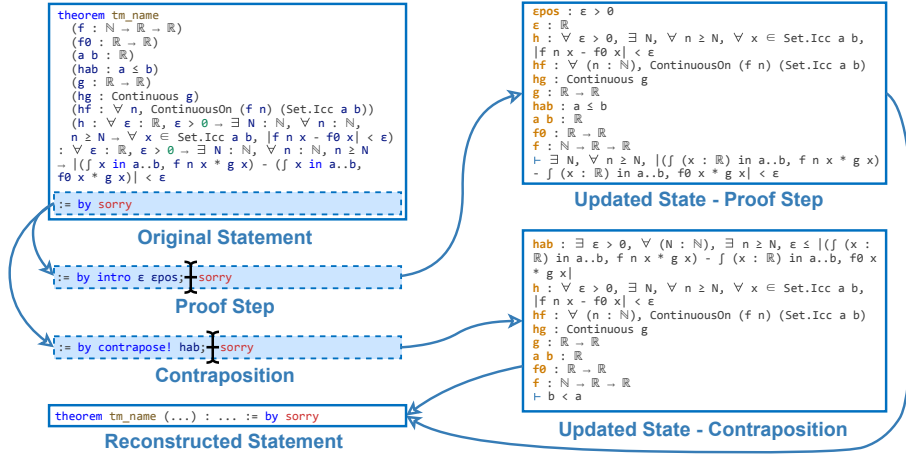


Figure 3: Demonstration of the proof step and contraposition augmentation methods.

Augmentation via Proof. For each FL statement in the synthetic data, we use DeepSeek-Prover-V1.5 [47] to generate a corresponding proof. The resulting proof steps are then executed sequentially. Each time a tactic is successfully applied and the proof process is not yet complete, Lean’s Infoview updates the proof state, displaying the current variables, hypotheses, and conclusions. Based on this information, new FL statements can be constructed.

Augmentation via Contraposition. In Section 3.2 FL Statement Parsing, we obtain the `theorem_hypotheses` for each FL statement. Furthermore, by extracting the names of all hypotheses and applying the `contrapose!` tactic, we can transform the original proposition into an equivalent contrapositive statement for each hypothesis. Leveraging the information provided by Lean’s Infoview, new FL statements can again be constructed.

Augmentation Data Construction. The aforementioned augmentation operations both rely on Lean Infoview. However, Infoview occasionally results in information loss, particularly concerning type-related details. Although Lean supports implicit type inference, this capability does not extend to all types. To address this, we perform FL Statements Compilation on the augmented FL statements and retain only those that compile successfully.

Furthermore, in the process of data augmentation, the primary consideration is the extent of diversity introduced relative to the original data. To maximize this diversity, we employ the following strategies. For the first augmentation method, we retain only those FL statements produced in the final proof step. For the second method, we utilize the Levenshtein distance [53] to select FL statements that exhibit the greatest dissimilarity from the synthetic FL statements. Finally, we utilize LLMs to translate these augmented FL statements into their corresponding NL statements, thereby constructing parallel statements. The used translation prompt can be found in Appendix F.

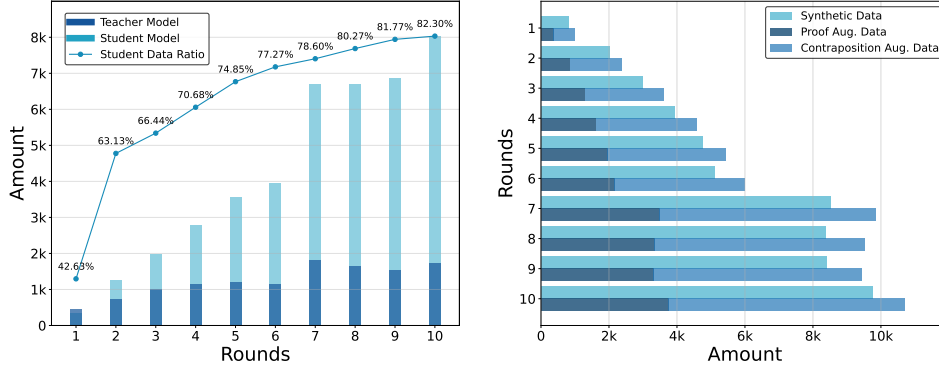


Figure 4: **Data Generation Statistics Across ATLAS Iterations.** Left: The number of synthetic data produced by the teacher and student models at each iteration, with the ratio of student-generated data indicated. Right: The composition of the generated data for each round, including synthetic data, proof augmentation data, and contraposition augmentation data.

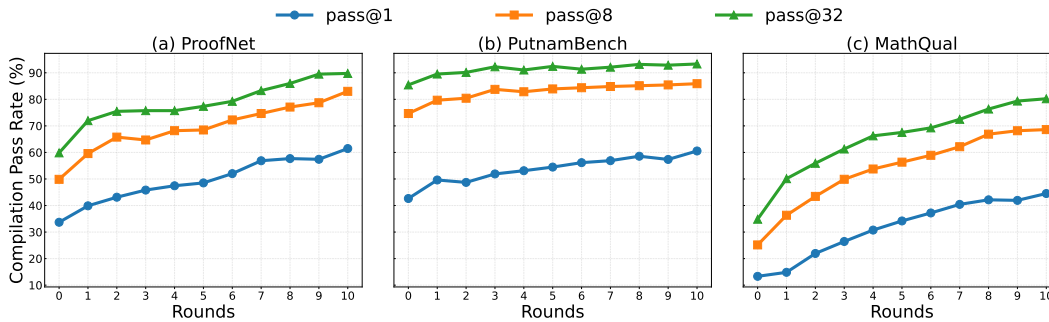


Figure 5: Performance of the student model on benchmarks throughout the iterative training process.

4 Experiments

4.1 ATLAS dataset Construction

4.1.1 Experimental Setup

We employ DeepSeek-V2.5 [22] as the teacher model and Llama3.1-8B-Instruct [8] as the student model, for which we utilize nucleus sampling [12] with top $p=0.9$ and a temperature of 0.6 for generation. To demonstrate the effectiveness of our framework, we aim to transform the general-purpose student model into the Lean 4 expert [41].

The training procedure of the student model comprises three stages:

1. **Model Initialization.** We use LeanDojo [50] to extract FL statements from Mathlib and employ LLMs to generate NL statements, constructing a dataset of 56,830 examples. Together with Lean Workbook, this dataset initializes the student model.
2. **Expert Iteration.** In each iteration, 10,000 new NL statements are generated and combined with the remaining NL statements from the previous round to create the synthetic and augmented data. The student model is then fine-tuned on this dataset before proceeding to the next iteration, with this process repeated for a total of 10 rounds.
3. **Final Re-training.** Llama3.1-8B-Instruct is re-trained on the data generated during expert iteration (referred to as the ATLAS dataset) to develop the ATLAS Translator.

All three stages are fine-tuned using LoRA [13] with LLaMA-Factory [57] for 3 epochs, with a total batch size of 128 and a learning rate of $1.0e-5$ with a cosine decay schedule. All experiments are conducted on a single NVIDIA A100 GPU with 40GB of memory. In particular, during the Expert

Iteration stage, fine-tuning required only 10 minutes to 1 hour, with minimal hardware requirements and low computational cost, further demonstrating the efficiency of our framework.

4.1.2 Experiment Results

As illustrated on the left side of Figure 4, the total amount of synthetic data produced by both the teacher and student models increases across ATLAS iterations. Notably, the proportion of data generated by the student model (indicated by the Student Data Ratio) grows substantially, rising from **42.68%** in round 1 to **82.30%** in round 10. This demonstrates the increasing capacity of the student model to autonomously contribute to the synthetic data as training progresses.

Meanwhile, the right side presents the detailed composition of the generated data for each round. It is evident that with each iteration, not only does the overall quantity of data increase, but the contributions from proof and contraposition augmentation methods also become more prominent. This diversified data augmentation strategy effectively enhances the variety of the dataset, allowing subsequent model iterations to benefit from richer and more diverse training signals. We refer to the synthetic and augmented data as the ATLAS dataset, whose statistics are shown in Table 1.

Figure 5 presents the student model’s pass rates on benchmarks across iterative rounds, clearly demonstrating a steady and significant improvement as training progresses. These results indicate that our framework effectively and continuously enhances the student model’s autoformalization ability, to a certain extent, since successful compilation is a prerequisite for correct formalization. In addition, Appendix E provides some examples of the synthetic data to illustrate the evolution of the student model’s behavior during successive iterations, including cases that achieve successful formalization after several rounds as well as cases that remain unsolved even after 10 iterations.

Table 1: Statistics of the ATLAS dataset

	Synthetic Data	Proof Aug. Data	Contraposition Aug. Data	Total
ATLAS dataset	54,641	22,103	40,401	117,145

4.2 ATLAS Translator Evaluation

4.2.1 Experimental Setup

Dataset. The datasets used for evaluation are ProofNet [2], PutnamBench [36], and MathQual. The version of ProofNet utilized in this evaluation is sourced from DeepSeek³. MathQual is a graduate-level dataset introduced in our work, consisting of 465 NL statements, specifically designed to assess the model’s generalization ability on more challenging problems. Detailed information on the construction process and specifics of MathQual can be found in Appendix D.

Baselines. We compare ATLAS Translator with the teacher model DeepSeek-V3 [23] (as V2.5 no longer provides API services), the initialization model Llama3.1-Initialization, the previous state-of-the-art model Herald Translator [9], and the latest work Kimina-Autoformalizer [37] to evaluate its performance. DeepSeek-V3 is set with a sampling temperature of 0.7, and the prompt used is provided in Appendix F. Herald Translator and Kimina-Autoformalizer utilize the sampling and prompt settings from their original papers, while Llama3.1-Initialization and ATLAS Translator follow the configurations in Section 4.1.

Validation Pipeline. To conduct the evaluation, we follow the validation pipeline described in the Lean Workbook [51] and Herald [9], which includes several key steps:

1. **Translation.** We translate the NL statements into FL statements using the corresponding model.
2. **Compilation.** We use the Lean compiler to verify the syntactic validity of the FL statements.
3. **Back-Translation.** For FL statements that pass compilation, we use InternLM2-Math-Plus-7B [52] to translate them back into NL statements.

³<https://github.com/deepseek-ai/DeepSeek-Prover-V1.5/tree/main/datasets>

Table 2: **Overall results of the competing baselines and ATLAS Translator.** The boldface refers to the highest score and the underline indicates the next best result of the models. “-” indicates that testing is not performed because the corresponding model uses that dataset during training. “*” signifies statistically significant improvements (two-sided t-test with $p < 0.05$) over the best baseline.

Model	ProofNet			PutnamBench			MathQual		
	pass@1	pass@8	pass@32	pass@1	pass@8	pass@32	pass@1	pass@8	pass@32
DeepSeek-V3	18.82%	34.07%	41.35%	11.53%	27.74%	37.33%	4.90%	13.42%	17.29%
Llama3.1-Initialization	23.56%	42.75%	51.54%	19.30%	45.58%	62.16%	6.97%	15.61%	22.02%
Herald Translator	31.43%	64.85%	78.57%	20.36%	52.56%	71.35%	10.92%	31.83%	45.33%
Kimina-Autoformalizer	-	-	-	-	-	-	19.01%	38.97%	50.71%
ATLAS Translator	39.46%*	67.28%*	78.71%	23.16%*	55.51%*	72.93%*	22.75%*	45.85%*	58.23%*

Table 3: **Ablation study on the three components.** The boldface refers to the highest score and the underline indicates the next best result of the models.

Model	ProofNet			PutnamBench			MathQual		
	pass@1	pass@8	pass@32	pass@1	pass@8	pass@32	pass@1	pass@8	pass@32
- w/o Synthetic	24.10%	51.86%	67.06%	15.99%	42.46%	58.15%	11.79%	29.76%	40.82%
- w/o Proof Aug.	37.41%	64.85%	76.39%	22.88%	53.23%	70.71%	22.97%	44.69%	56.13%
- w/o Contraposition Aug.	39.03%	66.09%	77.79%	22.70%	52.56%	69.56%	22.97%	44.95%	57.20%
ATLAS Translator	39.46%	67.28%	78.71%	23.16%	55.51%	72.93%	<u>22.75%</u>	45.85%	58.23%

4. **NLI Check.** Qwen2.5 [49] is used to compare the back-translated NL statements with the original NL statements to ensure semantic accuracy.

We consider the translation successful if any of these candidates pass both the compilation and NLI check. For a detailed discussion and case study of the validation pipeline’s effects, especially the NLI check, please refer to Appendix B.

Evaluation Metrics. We use the pass@ k metric [5] with $k = 1, 8, 32$, as larger values of k enable LLMs to better realize their potential in generating diverse outputs, which is beneficial for addressing challenging tasks [17, 41]. To reduce randomness, we conduct 5 experiments using the seeds 42, 43, 44, 45, 46, and report the mean of the results. To examine statistical significance, we further perform a two-tailed t-test with $p < 0.05$.

4.2.2 Overall Results

The overall results are presented in Table 2. At a glance, we find that the proposed ATLAS Translator outperforms all competing baselines across all datasets and all pass@ k metrics, thereby confirming the efficacy of our framework. Further insights will be explored through the subsequent analysis.

Comparison with Teacher and Initialization Model. The experimental results clearly demonstrate that ATLAS Translator significantly outperforms its teacher model DeepSeek-V3 and initialization model Llama3.1-Initialization across all benchmarks. Notably, on ProofNet, ATLAS achieves a pass@1 score of 39.46%, nearly doubling DeepSeek-V3’s 18.82% and surpassing Llama3.1’s 23.56%. More importantly, similar improvements are consistently observed in pass@8 and pass@32 metrics, as well as across other benchmarks, strongly validating the effectiveness of ATLAS framework.

Comparison with Competing Models. When examining the comparison with competing models, ATLAS Translator shows remarkable advantages over both the Herald Translator and the newer Kimina-Autoformalizer. With the exception of ProofNet’s pass@32, statistically significant improvements are observed in all other metrics and across all other benchmarks. Furthermore, considering that Herald Translator utilizes **1,160k** data points for fine-tuning and that Kimina-Autoformalizer consistently **involves Lean 4 experts** during its training process, it is noteworthy that the ATLAS Translator achieves these results using only **117k** data points and **without any human intervention**. This further underscores the effectiveness of the ATLAS framework.

Table 4: **Additional results of LoRA and full-parameter fine-tuning on various base models with the ATLAS dataset.** “*” denotes full-parameter fine-tuning. Abbreviations: L (Llama-3.1-8B-Instruct), D (DeepSeek-Prover-V1.5-7B-Base), and Q (Qwen2.5-Coder-7B-Instruct). The boldface refers to the highest score and the underline indicates the best result of the baselines. “-” indicates that testing is not performed because the corresponding model uses that dataset during training.

Model	miniF2F			ProofNet			PutnamBench			MathQual		
	pass@1	pass@8	pass@32	pass@1	pass@8	pass@32	pass@1	pass@8	pass@32	pass@1	pass@8	pass@32
Herald Translator	<u>76.02%</u>	<u>93.44%</u>	<u>95.29%</u>	<u>31.43%</u>	<u>64.85%</u>	<u>78.57%</u>	<u>20.36%</u>	<u>52.56%</u>	<u>71.35%</u>	10.92%	31.83%	45.33%
Kimina-Autoformalizer	-	-	-	-	-	-	-	-	-	19.01%	<u>38.97%</u>	<u>50.71%</u>
ATLAS Translator (L)	66.60%	88.52%	93.24%	39.46%	67.28%	78.71%	23.16%	55.51%	72.93%	22.75%	45.85%	58.23%
ATLAS Translator* (L)	69.67%	92.42%	96.93%	47.98%	74.66%	86.52%	38.54%	73.29%	84.98%	40.22%	65.81%	75.48%
ATLAS Translator (D)	67.01%	90.98%	96.31%	39.51%	69.49%	81.62%	25.64%	59.03%	75.75%	24.30%	49.59%	63.74%
ATLAS Translator* (D)	77.25%	93.65%	95.90%	54.99%	80.86%	88.95%	42.49%	76.93%	87.86%	38.92%	67.31%	79.78%
ATLAS Translator (Q)	66.60%	91.80%	96.72%	38.81%	71.43%	84.10%	29.29%	66.77%	82.25%	28.17%	54.19%	68.17%
ATLAS Translator* (Q)	69.88%	89.75%	92.62%	50.67%	79.51%	86.25%	39.91%	76.93%	87.56%	37.63%	65.59%	76.34%

4.2.3 Ablation Study

The results of the ablation study are shown in Table 3, where we also conduct 5 experiments using the same seeds and report the mean results. The removal of synthetic data leads to the most significant performance drop across all datasets and metrics, underscoring its critical role in training robustness. In contrast, omitting proof or contraposition augmentation data results in a more moderate decline in performance. Nevertheless, the full model consistently achieves the highest scores or, at the very least, competitive second-best results, thereby validating the synergistic effect of all components.

4.2.4 Additional Results

The results of applying LoRA and full-parameter fine-tuning to various base models on the ATLAS dataset are detailed in Table 4. These experiments reveal two key insights.

First, a primary finding is that full-parameter fine-tuning consistently and significantly outperforms the more parameter-efficient LoRA approach. This performance gap is particularly pronounced for the DeepSeek-Prover-V1.5-7B-Base model; on miniF2F [56], full-parameter fine-tuning achieves a pass@1 score of 77.25%, a substantial improvement of over 10 absolute points compared to its LoRA counterpart (67.01%). This trend holds true across all three base models.

Furthermore, the results highlight the critical role of the base model. The DeepSeek-Prover-V1.5-7B-Base model, when fine-tuned on the ATLAS dataset, achieves new state-of-the-art results on most benchmarks, with impressive pass@1 scores of 54.99% on ProofNet and 42.49% on PutnamBench. This outcome is expected, given the model’s extensive pre-training on Lean-related corpora. Consequently, we hypothesize that employing a more powerful base model as the student model within our iterative framework would further enhance its overall efficiency.

5 Conclusion

In this paper, we propose a novel framework to advance autoformalization by synthesizing and augmenting large-scale, high-quality parallel statements. Our method addresses key limitations of existing approaches, such as the finite amount of data that can be extracted from Mathlib and the extensive pre-processing required for data obtained from web scraping. Through comprehensive experiments, we verify the effectiveness of the ATLAS framework and achieve a new state of the art.

Acknowledgments and Disclosure of Funding

This work is sponsored by the National Key R&D Program of China Grant No. 2022YFA1008200 (T. L.). We also thank Shanghai Institute for Mathematics and Interdisciplinary Sciences (SIMIS) for their financial support. This research was funded by SIMIS under grant number SIMIS-ID-2025-ST. The authors are grateful for the resources and facilities provided by SIMIS, which were essential for the completion of this work. We appreciate the insightful discussions with Wei Zhao, Xinpu Tu, and Shuyu Yin during the early stages of the project, as well as Tao Zhu’s valuable involvement in the human evaluation at a later stage.

References

- [1] A. Agrawal, S. Gadgil, N. Goyal, A. Narayanan, and A. Tadipatri. Towards a mathematics formalisation assistant using large language models. *arXiv preprint arXiv:2211.07524*, 2022.
- [2] Z. Azerbayev, B. Piotrowski, H. Schoelkopf, E. W. Ayers, D. Radev, and J. Avigad. ProofNet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*, 2023.
- [3] Z. Azerbayev, H. Schoelkopf, K. Paster, M. D. Santos, S. McAleer, A. Q. Jiang, J. Deng, S. Biderman, and S. Welleck. Llemma: An open language model for mathematics. In *The Twelfth International Conference on Learning Representations*, 2024.
- [4] B. Barras, S. Boutin, C. Cornes, J. Courant, Y. Coscoy, D. Delahaye, D. de Rauglaudre, J.-C. Filliâtre, E. Giménez, H. Herbelin, et al. The Coq proof assistant reference manual. *INRIA, version*, 6(11), 1999.
- [5] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- [6] G. Cunningham, R. C. Bunescu, and D. Juedes. Towards autoformalization of mathematics and code correctness: Experiments with elementary proofs. In *Proceedings of the 1st Workshop on Mathematical Natural Language Processing (MathNLP)*, pages 25–32. Association for Computational Linguistics, 2023.
- [7] L. De Moura, S. Kong, J. Avigad, F. Van Doorn, and J. von Raumer. The Lean theorem prover (system description). In *Automated Deduction-CADE-25: 25th International Conference on Automated Deduction, Berlin, Germany, August 1-7, 2015, Proceedings 25*, pages 378–388. Springer, 2015.
- [8] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [9] G. Gao, Y. Wang, J. Jiang, Q. Gao, Z. Qin, T. Xu, and B. Dong. Herald: A natural language annotated Lean 4 dataset. In *The Thirteenth International Conference on Learning Representations*, 2024.
- [10] J. Harrison. HOL Light: A tutorial introduction. In *International Conference on Formal Methods in Computer-Aided Design*, pages 265–269. Springer, 1996.
- [11] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [12] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.
- [13] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- [14] Y. Huang, X. Lin, Z. Liu, Q. Cao, H. Xin, H. Wang, Z. Li, L. Song, and X. Liang. MUSTARD: Mastering uniform synthesis of theorem and proof data. In *The Twelfth International Conference on Learning Representations*, 2024.
- [15] A. Q. Jiang, W. Li, and M. Jamnik. Multilingual mathematical autoformalization. *arXiv preprint arXiv:2311.03755*, 2023.
- [16] A. Q. Jiang, S. Welleck, J. P. Zhou, T. Lacroix, J. Liu, W. Li, M. Jamnik, G. Lample, and Y. Wu. Draft, Sketch, and Prove: Guiding formal theorem provers with informal proofs. In *The Eleventh International Conference on Learning Representations*, 2023.
- [17] A. Q. Jiang, S. Welleck, J. P. Zhou, T. Lacroix, J. Liu, W. Li, M. Jamnik, G. Lample, and Y. Wu. Draft, Sketch, and Prove: Guiding formal theorem provers with informal proofs. In *The Eleventh International Conference on Learning Representations*, 2023.

- [18] J. Li, E. Beeching, L. Tunstall, B. Lipkin, R. Soletskyi, S. Huang, K. Rasul, L. Yu, A. Q. Jiang, Z. Shen, et al. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13:9, 2024.
- [19] Y. Li, D. Du, L. Song, C. Li, W. Wang, T. Yang, and H. Mi. Hunyuanprover: A scalable data synthesis framework and guided tree search for automated theorem proving. *arXiv preprint arXiv:2412.20735*, 2024.
- [20] Z. Li, Y. Wu, Z. Li, X. Wei, X. Zhang, F. Yang, and X. Ma. Autoformalize mathematical statements by symbolic equivalence and semantic consistency. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [21] Y. Lin, S. Tang, B. Lyu, J. Wu, H. Lin, K. Yang, J. Li, M. Xia, D. Chen, S. Arora, et al. Goedel-prover: A frontier model for open-source automated theorem proving. *arXiv preprint arXiv:2502.07640*, 2025.
- [22] A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Deng, C. Ruan, D. Dai, D. Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.
- [23] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [24] Q. Liu, X. Zheng, X. Lu, Q. Cao, and J. Yan. Rethinking and improving autoformalization: Towards a faithful metric and a dependency retrieval-based approach. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [25] J. Lu, Y. Wan, Y. Huang, J. Xiong, Z. Liu, and Z. Guo. Formalalign: Automated alignment evaluation for autoformalization. In *The Thirteenth International Conference on Learning Representations*, 2024.
- [26] J. Lu, Y. Wan, Z. Liu, Y. Huang, J. Xiong, C. Liu, J. Shen, H. Jin, J. Zhang, H. Wang, et al. Process-driven autoformalization in lean 4. *arXiv preprint arXiv:2406.01940*, 2024.
- [27] L. d. Moura and S. Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction—CADE 28: 28th International Conference on Automated Deduction, Virtual Event, July 12–15, 2021, Proceedings 28*, pages 625–635. Springer, 2021.
- [28] L. Murphy, K. Yang, J. Sun, Z. Li, A. Anandkumar, and X. Si. Autoformalizing euclidean geometry. In *Forty-first International Conference on Machine Learning*, 2024.
- [29] L. C. Paulson. *Isabelle: A Generic Theorem Prover*. Springer, 1994.
- [30] S. Polu, J. M. Han, K. Zheng, M. Baksys, I. Babuschkin, and I. Sutskever. Formal mathematics statement curriculum learning. *arXiv preprint arXiv:2202.01344*, 2022.
- [31] S. Polu and I. Sutskever. Generative language modeling for automated theorem proving. *arXiv preprint arXiv:2009.03393*, 2020.
- [32] Z. Ren, Z. Shao, J. Song, H. Xin, H. Wang, W. Zhao, L. Zhang, Z. Fu, Q. Zhu, D. Yang, et al. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801*, 2025.
- [33] V. Shah, D. Yu, K. Lyu, S. Park, N. R. Ke, M. C. Mozer, Y. Bengio, S. Arora, and A. Goyal. AI-assisted generation of difficult math questions. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS’24*, 2024.
- [34] C. Szegedy. A promising path towards autoformalization and general artificial intelligence. In *Intelligent Computer Mathematics: 13th International Conference, CICM 2020, Bertinoro, Italy, July 26–31, 2020, Proceedings 13*, pages 3–20. Springer, 2020.
- [35] The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*. ACM, Jan. 2020.

- [36] G. Tsoukalas, J. Lee, J. Jennings, J. Xin, M. Ding, M. Jennings, A. Thakur, and S. Chaudhuri. PutnamBench: Evaluating neural theorem-provers on the putnam mathematical competition. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- [37] H. Wang, M. Unsal, X. Lin, M. Baksys, J. Liu, M. D. Santos, F. Sung, M. Vinyes, Z. Ying, Z. Zhu, et al. Kimina-prover preview: Towards large formal reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.11354*, 2025.
- [38] H. Wang, H. Xin, C. Zheng, Z. Liu, Q. Cao, Y. Huang, J. Xiong, H. Shi, E. Xie, J. Yin, Z. Li, and X. Liang. LEGO-Prover: Neural theorem proving with growing libraries. In *The Twelfth International Conference on Learning Representations*, 2024.
- [39] Q. Wang, C. Brown, C. Kaliszyk, and J. Urban. Exploration of neural machine translation in autoformalization of mathematics in mizar. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, pages 85–98, 2020.
- [40] Q. Wang, C. Kaliszyk, and J. Urban. First experiments with neural translation of informal to formal mathematics. In *Intelligent Computer Mathematics: 11th International Conference, CICM 2018, Hagenberg, Austria, August 13-17, 2018, Proceedings 11*, pages 255–270. Springer, 2018.
- [41] R. Wang, J. Zhang, Y. Jia, R. Pan, S. Diao, R. Pi, and T. Zhang. Theoremllama: Transforming general-purpose llms into lean4 experts. *arXiv preprint arXiv:2407.03203*, 2024.
- [42] X. Wang, Y. Chen, and W. Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021.
- [43] Y. Wu, A. Q. Jiang, W. Li, M. Rabe, C. Staats, M. Jamnik, and C. Szegedy. Autoformalization with large language models. *Advances in Neural Information Processing Systems*, 35:32353–32368, 2022.
- [44] Z. Wu, S. Huang, Z. Zhou, H. Ying, J. Wang, D. Lin, and K. Chen. Internlm2.5-stepprover: Advancing automated theorem proving via expert iteration on large-scale lean problems. *arXiv preprint arXiv:2410.15700*, 2024.
- [45] Z. Wu, J. Wang, D. Lin, and K. Chen. Lean-github: Compiling github lean repositories for a versatile lean prover. *arXiv preprint arXiv:2407.17227*, 2024.
- [46] H. Xin, D. Guo, Z. Shao, Z. Ren, Q. Zhu, B. Liu, C. Ruan, W. Li, and X. Liang. DeepSeek-Prover: Advancing theorem proving in llms through large-scale synthetic data. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS’24*, 2024.
- [47] H. Xin, Z. Ren, J. Song, Z. Shao, W. Zhao, H. Wang, B. Liu, L. Zhang, X. Lu, Q. Du, W. Gao, H. Zhang, Q. Zhu, D. Yang, Z. Gou, Z. Wu, F. Luo, and C. 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.
- [48] R. Xin, C. Xi, J. Yang, F. Chen, H. Wu, X. Xiao, Y. Sun, S. Zheng, and K. Shen. Bfs-prover: Scalable best-first tree search for llm-based automatic theorem proving. *arXiv preprint arXiv:2502.03438*, 2025.
- [49] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [50] K. Yang, A. Swope, A. Gu, R. Chalamala, P. Song, S. Yu, S. Godil, R. J. Prenger, and A. Anandkumar. LeanDojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [51] H. Ying, Z. Wu, Y. Geng, J. Wang, D. Lin, and K. 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.

- [52] H. Ying, S. Zhang, L. Li, Z. Zhou, Y. Shao, Z. Fei, Y. Ma, J. Hong, K. Liu, Z. Wang, Y. Wang, Z. Wu, S. Li, F. Zhou, H. Liu, S. Zhang, W. Zhang, H. Yan, X. Qiu, J. Wang, K. Chen, and D. Lin. InternLM-Math: Open math large language models toward verifiable reasoning. *CoRR*, abs/2402.06332, 2024.
- [53] L. Yujian and L. Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.
- [54] L. Zhang, X. Quan, and A. Freitas. Consistent autoformalization for constructing mathematical libraries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4020–4033. Association for Computational Linguistics, Nov. 2024.
- [55] C. Zheng, H. Wang, E. Xie, Z. Liu, J. Sun, H. Xin, J. Shen, Z. Li, and Y. Li. Lyra: Orchestrating dual correction in automated theorem proving. *arXiv preprint arXiv:2309.15806*, 2023.
- [56] K. Zheng, J. M. Han, and S. Polu. miniF2F: a cross-system benchmark for formal Olympiad-level mathematics. In *International Conference on Learning Representations*, 2022.
- [57] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, and Z. Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In Y. Cao, Y. Feng, and D. Xiong, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410. Association for Computational Linguistics, Aug. 2024.
- [58] J. P. Zhou, C. E. Staats, W. Li, C. Szegedy, K. Q. Weinberger, and Y. Wu. Don’t Trust: Verify – grounding LLM quantitative reasoning with autoformalization. In *The Twelfth International Conference on Learning Representations*, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The scope and contributions of the paper are included in the abstract and introduction. Please refer to the first and last paragraph of Section 1 for scope and contributions, respectively.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: A limitation section is included in the Appendix A.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA] .

Justification: The paper does not have any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided all implementation details in Section 4.1.1 and Section 4.2.1. All prompts used are given in Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The datasets, model, and code are available at <https://github.com/XiaoyangLiu-sjtu/ATLAS>.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have provided all implementation details in Section 4.1.1 and Section 4.2.1. All prompts used are given in Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We run the experiments 5 times with seeds 42, 43, 44, 45, 46, report the mean results, and perform a two-tailed t-test with $p < 0.05$ to demonstrate statistical significance in Table 2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide the details of compute resources in the last paragraph in Section 4.1.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have made sure that our paper conforms with the NeurIPS code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the potential positive impacts that our framework will bring in the Section 1.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA] .

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have cited the original paper or attached the link to the existing assets used in this paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA] .

Justification: The paper does not release new assets up to now.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] .

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA] .

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: In Sections 3 and 4, we provide a detailed introduction on how LLMs are incorporated into our framework. In Appendix F, we present the prompts used for each LLM in our framework.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Motivation, Limitations, and Future Work

Motivation. Our work builds upon MUSTARD [14] by addressing two of its primary limitations. And solutions to these limitations constitute the core contributions of this paper.

- **Data Sourcing.** MUSTARD sources concepts from the Khan Academy Website, which can create a "formalization bottleneck" as some concepts lack a direct counterpart in Mathlib. To overcome this, ATLAS sources concepts exclusively from Mathlib, guaranteeing a valid formalization path for every statement from the start.
- **Generation Efficiency.** MUSTARD’s reliance on GPT-4 for correction loops is a costly, low-yield process: over 90% of initial generations fail the prover validation, resulting in a final dataset of only 6k samples. In contrast, the teacher-student distillation framework in ATLAS is highly efficient, dramatically reducing costs and enabling the generation of a much larger dataset of 117k high-quality pairs.

Limitations. A key limitation of our work is that we do not verify the correctness of the synthetic mathematical propositions in the ATLAS dataset. This decision is based on two primary factors. First, the autoformalization task is fundamentally about the fidelity of translation from natural language to formal language; the underlying truth value of a proposition does not alter the core translation challenge. Second, the current capabilities of automated theorem provers are insufficient for reliably verifying a large corpus of undergraduate-level mathematics. For example, the state-of-the-art DeepSeek-Prover-V2-671B [32] only achieves a 7.44% success rate on PutnamBench, making large-scale verification prohibitively challenging and costly.

Nevertheless, we acknowledge that training models on a corpus containing false propositions poses a potential risk for downstream applications, particularly for the automated theorem proving task. To quantify this risk, we performed an analysis to estimate the prevalence of incorrect propositions in our dataset. We randomly sampled 100 propositions from ATLAS and established a ground truth for their correctness using a panel of experts, which included human evaluators alongside advanced LLMs (DeepSeek-R1 and Gemini-2.5-Pro). Using a majority vote consensus from this panel, our analysis classified 60 of the sampled propositions as true and 40 as false. This finding provides a baseline estimate of the dataset’s truthfulness and highlights an area for future work in synthetic data refinement.

Another limitation is the diversity of data enhanced by proof. The augmentation-via-proof method is designed to generate novel propositions by treating each step of a formal proof as a semantic transformation. Ideally, the final proposition in a proof chain should be substantially different from the original. However, a limitation of the current implementation is that the practical diversity of these transformations can be constrained, particularly when the proof relies on trivial tactics.

To quantify the textual diversity of augmented data, we conducted an analysis on 100 randomly sampled propositions. We measured the BLEU score between the original and augmented versions, where a lower score signifies greater novelty. The analysis revealed that augmentation-via-proof (Average BLEU: 0.6709) produces less diverse statements on average than augmentation-via-contraposition (Average BLEU: 0.6026). To enhance the former, we plan to implement a stricter filtering mechanism that, for example, disallows augmentations generated from trivial tactics.

Future Work. A primary direction for future work is to implement a more explicit curriculum learning [42] strategy to guide the model’s improvement. This strategy comprises two main components:

- **Compositional Complexity.** We will restructure the concept repository into a graph, enabling the systematic generation of a curriculum. New theorems will be synthesized by progressively increasing both the conceptual distance between ideas (i.e., path length in the graph) and the number of concepts required to form a valid statement.
- **Conceptual Hierarchy.** We will introduce a graded concept repository (e.g., high school \rightarrow undergraduate \rightarrow graduate). The model must demonstrate proficiency at one level before unlocking access to the next, creating a structured path toward mastering more advanced topics.

Together, these mechanisms will enable our pipeline to autonomously generate a curriculum of increasing difficulty, creating a powerful virtuous cycle of self-improvement.

A more ambitious goal is to extend our framework from formalizing individual statements to entire mathematical theories. This direction is inspired by dependency-aware, retrieval-based methods like RAutoformalizer [24]. While such methods improve performance by retrieving premises at inference time, our work focuses on data generation. A promising synthesis of these ideas is to integrate the core principle of dependency awareness directly into our data generation process.

Our technical approach for this involves restructuring the concept repository into a dependency-aware tree. During synthesis, we will generate data in a bottom-up, layer-by-layer fashion, explicitly preserving the logical dependencies from foundational axioms to advanced theorems. The resulting structured dataset would be ideal for training next-generation models capable of theory-level autoformalization.

B Discussion for Validation Pipeline

In this section, we discuss and examine the reliability and equity of our validation pipeline and evaluate instances of both successful and unsuccessful validations.

Discussion. Recently, there has been some work [20, 24, 28] on automated evaluation of translated FL statements. However, all of these approaches are based on the mutual proof of the translated FL statements and the labeled FL statements. Considering the development of the field of automated theorem proving, it is unrealistic to use this method for automated evaluation on undergraduate and graduate datasets, as the most powerful model currently, DeepSeek-Prover-V2-671B, has only achieved a proof rate of 7.44% on PutnamBench.

Conversely, concerning the validation pipeline, it is important to highlight that validation based on LLMs can sometimes diverge from human judgment, particularly in instances of false positives. Nevertheless, this approach remains standard in the field (e.g., see Lean Workbook [51], Herald [9]) and is currently one of the few viable solutions for large-scale automated evaluation. This is primarily because conducting human expert reviews under the $\text{pass}@k$ metric is impractical.

In our implementation of the validation pipeline, especially the NLI check, the primary objective is to establish a fair performance evaluation among different models. As demonstrated in the subsequent experiments, all the baseline models and our proposed model yield comparable results regarding the proportion of false-positive cases, indicating that the validation pipeline is relatively fair.

Experiment Results. To provide a clearer illustration of the false positive cases and the fairness discussed earlier, we conduct an experiment to evaluate the outcomes that pass through the validation pipeline. Specifically, we select the entire ProofNet dataset, consisting of 371 natural language statements, and translate it using the ATLAS Translator, as well as two baseline models: the Herald Translator and the Llama-3.1 Initialization.

The outputs of autoformalization that successfully pass validation are classified by human experts into three distinct categories: correct translations, minor errors, and major errors. These classifications adhere to the evaluation criteria established in Herald [9].

- **Correct Translation.** A correct translation must accurately reflect the mathematical meaning of the natural language statement. In cases where the original statement is ambiguous, we allow the formalized statement to adopt any mathematically valid interpretation.
- **Minor Error.** If the formalized result deviates from the original statement but can be easily corrected by altering a single formal definition or restructuring a single logical relation within the hypothesis or statement, it is classified as a minor error.
- **Major Error.** All other cases are classified as major errors.

As shown in Table 5, the ATLAS Translator and other baseline models also experience false positives, exhibiting comparable false positive rates, which ensures a fair comparison. In Sections B.1, B.2, and B.3, we provide examples for each of these three categories.

Table 5: Human evaluation results of validation passed statements

Model	Validation Passed	Correct Translation	Minor Error	Major Error
Llama3.1-Initialization	82	43	30	9
Herald Translator	134	79	31	24
ATLAS Translator	151	99	26	26

B.1 Correct Examples

Example 1

Suppose that f is holomorphic in an open set Ω . Prove that if $\text{Re}(f)$ is constant, then f is constant.

```
theorem tm_name (Ω : Set ℂ) (hΩ : IsOpen Ω) (f : ℂ → ℂ) (hf :
  DifferentiableOn ℂ f Ω) (h : ∀ z ∈ Ω, ∀ w ∈ Ω, (f z).re = (f w)
  .re) : ∀ z ∈ Ω, ∀ w ∈ Ω, f z = f w := by sorry
```

Example 2

Let X be a topological space; let A be a subset of X . Suppose that for each $x \in A$ there is an open set U containing x such that $U \subseteq A$. Show that A is open in X .

```
theorem tm_name {X : Type*} [TopologicalSpace X] (A : Set X) (hA :
  ∀ x ∈ A, ∃ U, IsOpen U ∧ x ∈ U ∧ U ⊆ A) : IsOpen A := by sorry
```

Example 3

Let $p : X \rightarrow Y$ be a closed continuous surjective map such that $p^{-1}(\{y\})$ is compact, for each $y \in Y$. (Such a map is called a perfect map.) Show that if Y is compact, then X is compact.

```
theorem tm_name {X Y : Type*} [TopologicalSpace X] [
  TopologicalSpace Y] [CompactSpace Y] {p : X → Y} (hp :
  Continuous p) (h : Function.Surjective p) (h' : ∀ y : Y,
  IsCompact (p-1 {y})) : CompactSpace X := by sorry
```

B.2 Minor Error Examples

Example 1

Show that $\int_0^1 \log(\sin \pi x) dx = -\log 2$.

```
theorem tm_name (π : ℝ) : ∫ x in (0 : ℝ)..1, Real.log (Real.sin
  (π * x) / π) = -Real.log 2 := by sorry
```

Analysis: Incorrectly formalizing $\sin \pi x$ as $\sin(\pi * x)/\pi$.

Example 2

Let f be a real uniformly continuous function on the bounded set E in \mathbb{R}^1 . Prove that f is bounded on E .

```
theorem tm_name (f : ℝ → ℝ) (E : Set ℝ) (hE : IsCompact E) (hf :
  UniformContinuousOn f E) : ∃ M, ∀ x ∈ E, |f x| ≤ M := by sorry
```

Analysis: In the formalization, the condition that the set E is bounded is strengthened to E being compact.

Example 3

Let R be a commutative ring with $1 \neq 0$. Prove that if a is a nilpotent element of R then $1 - ab$ is a unit for all $b \in R$.

```
theorem tm_name {R : Type*} [CommRing R] [IsDomain R] [Nontrivial R]
  (a b : R) (ha : IsNilpotent a) : IsUnit (1 - a * b) := by
  sorry
```

Analysis: In the formalization, an additional assumption was made that R is a domain.

B.3 Major Error Examples

Example 1

If $C_0 + \frac{C_1}{2} + \dots + \frac{C_{n-1}}{n} + \frac{C_n}{n+1} = 0$, where C_0, \dots, C_n are real constants, prove that the equation $C_0 + C_1x + \dots + C_{n-1}x^{n-1} + C_nx^n = 0$ has at least one real root between 0 and 1.

```
theorem tm_name (n : ℕ) (hn : 0 < n) (C : Fin (n + 1) → ℝ) (hC : Σ
  i : Fin (n + 1), C i / (i + 1) = 0) : ∃ x : ℝ, 0 ≤ x ∧ x ≤ 1 ∧
  C 0 + C 1 * x + C 2 * x ^ 2 + C 3 * x ^ 3 + C 4 * x ^ 4 + C 5
  * x ^ 5 + C 6 * x ^ 6 + C 7 * x ^ 7 + C 8 * x ^ 8 + C 9 * x ^ 9
  + C 10 * x ^ 10 = 0 := by sorry
```

Analysis: The polynomial is written as explicit terms instead of a general sum, losing generality for arbitrary n .

Example 2

If G is a group in which $(ab)^i = a^ib^i$ for three consecutive integers i , prove that G is abelian.

```
theorem tm_name {G : Type*} [Group G] [Fintype G] [AddCommMonoid G
  ] [MulAction G G] (a b : G) (i j k : ℕ) (h : i + 1 = j ∧ j + 1
  = k) (h' : (a * b) ^ i = a ^ i * b ^ i) : a * b = b * a := by
  sorry
```

Analysis: Omitted the assumptions $(ab)^j = a^jb^j$ and $(ab)^k = a^kb^k$, only stated $(ab)^i = a^ib^i$.

Example 3

Prove that a set $U \subset M$ is open if and only if none of its points are limits of its complement.

```
theorem tm_name {M : Type*} [MetricSpace M] [TopologicalSpace M] (
  U : Set M) : IsOpen U ↔ ∀ x ∈ U, ∀ y ∈ UC, x ≠ y := by sorry
```

Analysis: This formalization does not faithfully express the conclusion about limit points of the complement."

C Concept Repository

The complete concept repository is composed of 13 domains, 55 topics, and 350 concepts. For the sake of clarity and brevity, Table 6 only presents a subset of topics from 3 domains, along with their corresponding concepts. And the full repository can be accessed in the open-source material.

D MathQual

The MathQual dataset is derived from graduate qualification examinations from multiple universities, including Boston University, Johns Hopkins University, University of Texas at Dallas, University of California, Los Angeles, University of California Riverside, and University of Georgia. Table 7 presents the domains included in the MathQual dataset and the process of creating the dataset is elaborated as follows.

1. Relevant PDF documents are retrieved from official websites.
2. Optical Character Recognition (OCR)⁴ technology is employed to convert these documents into Markdown format.
3. High-quality, formalizable problem statements are meticulously selected through a manual filtration process. Notably, for proof problems consisting of multiple sub-questions, we amalgamate the overarching contextual conditions of the main problem with the specific conditions of each sub-question, thereby constructing several distinct problem statements.
4. The problem statements are categorized according to Mathematics Subject Classification (MSC)⁵.

⁴<https://github.com/opendatalab/MinerU>

⁵<https://zbmath.org/classification>

Table 6: Partial list of mathematical concepts in the concept repository

Domain	Topic	Concept
Linear Algebra	Fundamentals	vector space, product of vector spaces, vector subspace, quotient space, sum of subspaces, direct sum, complementary subspaces, linear independence, generating sets, bases, existence of bases, linear map, range of a linear map, kernel of a linear map, algebra of endomorphisms of a vector space, general linear group
	Duality	dual vector space, dual basis, transpose of a linear map
	Finite-dimensional vector spaces	finite-dimensionality, isomorphism with K^n , rank of a linear map, rank of a set of vectors, isomorphism with bidual
	Multilinearity	multilinear map, determinant of vectors, determinant of endomorphisms, orientation of a \mathbb{R} -vector space
	Matrices	commutative-ring-valued matrices, field-valued matrices, matrix representation of a linear map, change of basis, rank of a matrix, determinant, invertibility
	Endomorphism polynomials	annihilating polynomials, minimal polynomial, characteristic polynomial, Cayley-Hamilton theorem
	Structure theory of endomorphisms	eigenvalue, eigenvector, generalized eigenspaces, Jordan-Chevalley-Dunford decomposition
Single Variable Complex Analysis	Complex Valued series	radius of convergence, continuity, differentiability with respect to the complex variable, complex exponential, extension of trigonometric functions(cos) to the complex plane, extension of trigonometric functions(sin) to the complex plane, power series expansion of elementary functions(cos), power series expansion of elementary functions(sin)
	Functions on one complex variable	holomorphic functions, Cauchy formulas, analyticity of a holomorphic function, principle of isolated zeros, principle of analytic continuation, maximum principle, holomorphic stability under uniform convergence
Topology	Topology and Metric Spaces	topology of a metric space, induced topology, finite product of metric spaces, limits of sequences, cluster points, continuous functions, homeomorphisms, compactness in terms of open covers (Borel-Lebesgue), sequential compactness is equivalent to compactness (Bolzano-Weierstrass), connectedness, connected components, path connectedness, Lipschitz functions, uniformly continuous functions, Heine-Cantor theorem, complete metric spaces, contraction mapping theorem
	Normed vector spaces on \mathbb{R} and \mathbb{C}	topology on a normed vector space, Banach open mapping theorem, equivalence of norms in finite dimension, norms $\ \cdot\ _p$ on \mathbb{R}^n and \mathbb{C}^n , absolutely convergent series in Banach spaces, continuous linear maps, norm of a continuous linear map, uniform convergence norm (sup-norm), normed space of bounded continuous functions, completeness of the space of bounded continuous functions, Heine-Borel theorem (closed bounded subsets are compact in finite dimension), Riesz' lemma (unit-ball characterization of finite dimension), Arzela-Ascoli theorem
	Hilbert spaces	Hilbert projection theorem, orthogonal projection onto closed vector subspaces, dual space, Riesz representation theorem, inner product space l^2 , completeness of l^2 , inner product space L^2 , completeness of L^2 , Hilbert bases, example, the Hilbert basis of trigonometric polynomials, Lax-Milgram theorem

Table 7: Domain Classification and Problem Counts in MathQual

Domain	Count
Algebraic geometry	2
Algebraic topology	26
Associative rings and algebras	11
Calculus of variations and optimal control; optimization	3
Category theory; homological algebra	6
Combinatorics	1
Commutative algebra	43
Difference and functional equations	1
Differential geometry	9
Field theory and polynomials	32
Functional analysis	23
Functions of a complex variable	101
General topology	24
Global analysis, analysis on manifolds	18
Group theory and generalizations	51
Harmonic analysis on Euclidean spaces	7
Linear and multilinear algebra; matrix theory	25
Manifolds and cell complexes	12
Mathematical logic and foundations	11
Measure and integration	16
Number theory	5
Operator theory	3
Ordinary differential equations	2
Partial differential equations	5
Potential theory	7
Probability theory and stochastic processes	11
Real functions	4
Several complex variables and analytic spaces	3
Topological groups, Lie groups	3
Total	465

E Case Study

This section presents both successful and unsuccessful cases of synthetic data during the ATLAS iterative process. Tables 8 to 13 display successful cases generated by ATLAS, while Tables 14 and 15 illustrate the failed cases, each accompanied by corresponding analyses. The FL statements consist of multiple rounds, with content highlighted in red boxes indicating syntactic errors or inconsistencies with the relevant NL statements. Iteration for each data point ceases when there are no syntactic or semantic errors in the FL statement. The cases below demonstrate that, as the model iterates, it is able to formalize mathematical concepts in Lean with increasing accuracy.

Table 8: A successful case generated by ATLAS

Analysis: In the first round, the model incorrectly formalized the matrix notation in Lean. However, in the second round, the formalization was correct.

Concept1: commutative-ring-valued matrices
 Domain1: Linear algebra
 Concept2: determinant
 Domain2: Linear algebra
 NL statement: For any commutative ring R and any two $n \times n$ matrices A and B over R , the determinant of their product AB is equal to the product of their determinants, i.e., $\det(AB) = \det(A)\det(B)$.
 FL statements:

Round 1

```
theorem tm_name {R : Type*} [CommRing R] (n : ℕ)
  (A B : Matrix (Fin n) R) : (A * B).det = A.det * B.det := by
  sorry
```

Round 2

```
theorem tm_name {R : Type*} [CommRing R] (n : ℕ) (A B : Matrix (
  Fin n) (Fin n) R) : (A * B).det = A.det * B.det := by sorry
```

Table 9: A successful case generated by ATLAS

Analysis: In the first two rounds, the model was unable to correctly formalize the concept of “derivative.” However, in the third round, it successfully formalizes this concept, demonstrating that the model effectively masters it after the learning experiences of the first two rounds.

Concept1: Leibniz formula
 Domain1: Single Variable Real Analysis
 Concept2: hyperbolic trigonometric functions
 Domain2: Single Variable Real Analysis
 NL statement: The derivative of the hyperbolic sine function, $\sinh(x)$, can be expressed using the Leibniz formula as $\sinh'(x) = \cosh(x)$.
 FL statements:

Round 1

```
theorem tm_name (x : ℝ) : deriv sinh x = cosh x := by sorry
```

Round 2

```
theorem tm_name : deriv sinh = cosh := by sorry
```

Round 3

```
theorem tm_name : HasDerivAt (fun x : ℝ => Real.sinh x) (Real.
  cosh x) x := by sorry
```

Table 10: A successful case generated by ATLAS

Analysis: In the first round, the model failed to correctly formalize the concepts of “domain” and “being an injection on a certain set.” In the second and third rounds, it learned how to formalize “domain,” but struggled to accurately formalize “being an injection on a certain set.” In the fourth round, it successfully formalized these concepts.

Concept1: metric structure

Domain1: Single Variable Real Analysis

Concept2: derivative at a point

Domain2: Single Variable Real Analysis

NL statement: If a function f is differentiable at c and $f'(c) \neq 0$, then there exists a neighborhood U of c such that f is injective on U .

FL statements:

Round 1

```
theorem tm_name (f : ℝ → ℝ) (c : ℝ) (hf : DifferentiableAt ℝ f c
  ∧ deriv f c ≠ 0) : ∃ (U : Set ℝ), (c ∈ U) ∧ (InjectiveOn f U)
:= by sorry
```

Round 2

```
theorem tm_name (f : ℝ → ℝ) (c : ℝ) (hf : DifferentiableAt ℝ f c)
  (hc : deriv f c ≠ 0) : ∃ U ∈ nhds c, Function.InjectiveOn f U :
= by sorry
```

Round 3

```
theorem tm_name (f : ℝ → ℝ) (c : ℝ) (hf : DifferentiableAt ℝ f c)
  (hc : deriv f c ≠ 0) : ∃ U ∈ nhds c,
  Function.Injective (f ∘ (λ x => x) ∘ (U := U) ∘ (λ x => x) ∘ (U := U))
:= by sorry
```

Round 4

```
theorem tm_name (f : ℝ → ℝ) (c : ℝ) (hf : DifferentiableAt ℝ f c)
  (h'f : deriv f c ≠ 0) : ∃ U ∈ nhds c, Set.InjOn f U := by
sorry
```

Table 11: A successful case generated by ATLAS

Analysis: In the previous rounds, the model was unable to correctly formalize “the order of a finite group” and the concept of “two coprime positive integers”, and there was unnecessary formalization. However, in the eighth round, the formalization was done correctly.

Concept1: direct product of groups

Domain1: Group Theory

Concept2: \mathbb{Z} is a euclidean ring

Domain2: Ring Theory

NL statement: Let G and H be groups. If $G \times H$ is cyclic, then G and H are cyclic and their orders are coprime.

FL statements:

Round 3

```
theorem tm_name {G H : Type*} [Group G] [Group H] (h : IsCyclic (
  G × H)) : IsCyclic G ∧ IsCyclic H ∧
  Nat.coPrime (orderOfOf G) (orderOfOf H) := by sorry
```

Round 4

```
theorem tm_name {G H : Type*} [Group G] [Group H] (hG : IsCyclic
  (G × H)) (hH : IsCyclic (H × G)) : IsCyclic G ∧ IsCyclic H ∧
  Nat.coPrime (orderOfOf G) (orderOfOf H) := by sorry
```

Round 5

```
theorem tm_name {G H : Type*} [Group G] [Group H] (hG : IsCyclic
  (G × H)) : IsCyclic G ∧ IsCyclic H ∧
  Nat.coPrime (Nat.card G) (Nat.card H) := by sorry
```

Round 6

```
theorem tm_name {G H : Type*} [Group G] [Group H] [Fintype G] [
  Fintype H] (hG : IsCyclic (G × H)) (hH : IsCyclic (H × G)) :
  IsCyclic G ∧ IsCyclic H ∧
  Nat.coPrime (Fintype.card G) (Fintype.card H) := by sorry
```

Round 7

```
theorem tm_name {G H : Type*} [Group G] [Group H] [Fintype G] [
  Fintype H] (hG : IsCyclic (G × H)) (hH : IsCyclic (H × G)) :
  IsCyclic G ∧ IsCyclic H ∧ (Fintype.card G).gcd (Fintype.card
  H) = 1 := by sorry
```

Round 8

```
theorem tm_name {G H : Type*} [Group G] [Group H] [Fintype G] [
  Fintype H] (h : IsCyclic (G × H)) : IsCyclic G ∧ IsCyclic H ∧
  (Fintype.card G).gcd (Fintype.card H) = 1 := by sorry
```

Table 12: A successful case generated by ATLAS

Analysis: In rounds five to seven, the incorrect formalization "Finset" was used. In round eight, the model revised it to "toFinSet," but it remained incorrect. Finally, in round nine, the formalization was done correctly. Note: The translation results for rounds 1 to 4 were subpar and lacked any significant insights, so they are not displayed.

Concept1: Cauchy-Schwarz inequality

Domain1: Bilinear and Quadratic Forms Over a Vector Space

Concept2: counting measure

Domain2: Measures and integral calculus

NL statement: For any finite set S and functions $f, g : S \rightarrow \mathbb{R}$, the Cauchy-Schwarz inequality holds: $(\sum_{x \in S} f(x)g(x))^2 \leq (\sum_{x \in S} f(x)^2) (\sum_{x \in S} g(x)^2)$.

FL statements:

Round 5 ~ 7

```
theorem tm_name (S : FinSet ℝ) (f g : ℝ → ℝ) : (Σ x in S, f x * g
  x) ^ 2 ≤ (Σ x in S, f x ^ 2) * (Σ x in S, g x ^ 2) := by
  sorry
```

Round 8

```
theorem tm_name (S : Set ℝ) (hS : S.Finite) (f g : S → ℝ) : (Σ x
  in S.toFinSet, f x * g x) ^ 2 ≤ (Σ x in S.toFinSet, f x ^ 2) *
  (Σ x in S.toFinSet, g x ^ 2) := by sorry
```

Round 9

```
theorem tm_name {S : Type*} [Fintype S] (f g : S → ℝ) : (Σ x : S,
  f x * g x) ^ 2 ≤ (Σ x : S, f x ^ 2) * (Σ x : S, g x ^ 2) :=
  by sorry
```

Table 13: A successful case generated by ATLAS

Analysis: In the first two rounds, there were formalization errors in Lean syntax, while in rounds three and four, the formalization expressed overly strong semantics. However, in round five, the formalization was done correctly.

Concept1: $K[X]$ is a euclidean ring when K is a field

Domain1: Ring Theory

Concept2: inverse trigonometric functions

Domain2: Single Variable Real Analysis

NL statement: Let K be a field. For any polynomial $f(X) \in K[X]$, there exists a polynomial $g(X) \in K[X]$ such that $f(X) \cdot g(X) = 1$ if and only if $f(X)$ is a constant polynomial. This property is analogous to the invertibility of non-zero constants in the context of inverse trigonometric functions.

FL statements:

Round 1

```
theorem tm_name {K : Type*} [Field K] (f : Polynomial K) : (∃ (g
  : Polynomial K), Polynomial.mul f g = 1) ↔
  f.coeffs.all ((*@λ*@) n => n = 0) := by sorry
```

Round 2

```
theorem tm_name {K : Type*} [Field K] (f : K[X]) : (∃ g : K[X], f
  * g = 1) ↔ f = C (1 : K) := by sorry
```

Round 3

```
theorem tm_name {K : Type*} [Field K] (f : Polynomial K) : (∃ g :
  Polynomial K, f * g = 1) ↔ f = 1 ∨ f = -1 := by sorry
```

Round 4

```
theorem tm_name {K : Type*} [Field K] (f : Polynomial K) : (∃ g :
  Polynomial K, f * g = 1) ↔ f = 1 := by sorry
```

Round 5

```
theorem tm_name {K : Type*} [Field K] (f : Polynomial K) : (∃ g :
  Polynomial K, f * g = 1) ↔ (∃ c : K, f = Polynomial.C c) :=
  by sorry
```

Table 14: A failed case generated by ATLAS

Analysis: The model was unable to correctly formalize the concept of a Gaussian integer ring.

Concept1: field \mathbb{C} of complex numbers
Domain1: Ring Theory
Concept2: greatest common divisor
Domain2: Ring Theory
NL statement: Let α and β be non-zero elements in the ring of Gaussian integers $\mathbb{Z}[i]$. Then the greatest common divisor $\gcd(\alpha, \beta)$ in $\mathbb{Z}[i]$ is a Gaussian integer of the form $\gamma = x + yi$, where x and y are integers.
FL statements:

Round 1

```
theorem tm_name {α : ℂ} (hα : IsAlgebraic ℤ α) (h : IsMonic (
  Polynomial.map (algebraMap ℤ ℂ) (X - C α))) :
  IsEuclideanDomain (ℤ⟨α⟩) := by sorry
```

Round 2

```
theorem tm_name (α β : ℤ[i]) (hα : α ≠ 0) (hβ : β ≠ 0) : ∃ x y :
  ℤ, ∃ γ : ℤ[i], γ = x + y * I ∧ γ = gcd α β := by sorry
```

Round 3

```
theorem tm_name (α β : ℤ[i]) (hα : α ≠ 0) (hβ : β ≠ 0) : ∃ γ : ℤ[
  i], γ = gcd α β ∧ ∃ x y : ℤ, γ = x + y * I := by sorry
```

Round 4

```
theorem tm_name (α β : ℤ[i]) (hα : α ≠ 0) (hβ : β ≠ 0) : ∃ (γ : ℤ[
  i]) (x y : ℤ), γ = x + y * I ∧ γ = gcd α β := by sorry
```

Round 5

```
theorem tm_name (α β : ℤ[i]) (hα : α ≠ 0) (hβ : β ≠ 0) : ∃ (γ : ℤ[
  i]) (x y : ℤ), γ = x + y * I ∧ γ = gcd α β := by sorry
```

Round 6

```
theorem tm_name (α β : ℤ[i]) (hα : α ≠ 0) (hβ : β ≠ 0) : ∃ γ : ℤ[
  i], ∃ x y : ℤ, γ = ⟨x, y⟩ ∧ γ = gcd α β := by sorry
```

Round 7

```
theorem tm_name (α β : ℤ[i]) (hα : α ≠ 0) (hβ : β ≠ 0) : ∃ γ : ℤ[
  i], γ = gcd α β := by sorry
```

Round 8

```
theorem tm_name (α β : ℤ[i]) (hα : α ≠ 0) (hβ : β ≠ 0) : ∃ γ : ℤ[
  i], γ = gcd α β ∧ ∃ x y : ℤ, γ = x + y * i := by sorry
```

Round 9

```
theorem tm_name (α β : ℤ[i]) (hα : α ≠ 0) (hβ : β ≠ 0) : ∃ γ : ℤ[
  i], γ = gcd α β := by sorry
```

Round 10

```
theorem tm_name (α β : ℤ[i]) (hα : α ≠ 0) (hβ : β ≠ 0) : ∃ γ : ℤ[
  i], γ = gcd α β := by sorry
```

Table 15: A failed case generated by ATLAS

Analysis: The model was unable to formalize such a relatively complex problem, and there were also certain issues with the natural language statement itself.

Concept1: Taylor's theorem with Lagrange form for remainder

Domain1: Single Variable Real Analysis

Concept2: affine function

Domain2: Affine and Euclidean Geometry

NL statement: For a function f that is $n + 1$ times differentiable on an interval containing a and x , the Taylor polynomial of degree n at a approximates $f(x)$ with an error term that can be expressed as an affine function of $(x - a)^{n+1}$.

FL statements:

Round 1

```
theorem tm_name (n : ℕ) (f : ℝ → ℝ) (hf : ∀ x, (x ∈ Set.Icc a x)
  → HasFDerivAt (f' x) f x) (x : ℝ) (hx : x ∈ Set.Icc a x) : ∃
  c : ℝ, f x - taylorPolynomial n a f = c * (x - a) ^ (n + 1) :=
  by sorry
```

Round 2

```
theorem tm_name (a x : ℝ) (n : ℕ) (f : ℝ → ℝ) (hf : ∀ x, f x = ∑
  i in range (n + 1), (f^[i] a * (x - a) ^ i / i!)) : ∃ k : ℝ,
  f x - ∑ i in range (n + 1), (f^[i] a * (x - a) ^ i / i!) = k
  * (x - a) ^ (n + 1) := by sorry
```

Round 3

```
theorem tm_name (n : ℕ) (a x : ℝ) (f : ℝ → ℝ) (hf : ∀ k : ℕ, k ≤
  n → DifferentiableOn ℝ (fun x => iteratedDeriv n f x) (Set.
  Icc a x)) : ∃ c : ℝ, f x - iteratedDeriv n f a = c * (x - a)
  ^ (n + 1) := by sorry
```

Round 4

```
theorem tm_name (n : ℕ) (a x : ℝ) (f : ℝ → ℝ) (hf : ContDiff ℝ (n
  + 1) f) : ∃ k : ℝ, f x - (f a + ∑ i in FinSet.range (n + 1),
  (f^[i + 1] a * (x - a) ^ i / (i + 1)!)) = k * (x - a) ^ (n +
  1) := by sorry
```

Round 5

```
theorem tm_name (f : ℝ → ℝ) (n : ℕ) (a x : ℝ) (hf : ContDiff ℝ (n
  + 1) f) : ∃ f' g : ℝ, f x = f a + f' * (x - a) + g * (x - a)
  ^ (n + 1) := by sorry
```

Round 6

```
theorem tm_name (f : ℝ → ℝ) (a x : ℝ) (n : ℕ) (hf : ContDiff ℝ (n
  + 1) f) : ∃ k : ℝ, f x = f a + (f' a) * (x - a) + (1 / (Nat.
  factorial (n + 1))) * (f (n + 1) a) * (x - a) ^ (n + 1) + k *
  (x - a) ^ (n + 1) := by sorry
```

...

Round 10

```
theorem tm_name (n : ℕ) (a : ℝ) (f : ℝ → ℝ) (hf : ∀ k : ℕ, k ≤ n
  → DifferentiableOn ℝ (fun x => deriv^[k] f x) (Set.Icc a a))
  : ∃ p : ℝ → ℝ, ∀ x : ℝ, |f x - p x| ≤ C * |x - a| ^ (n + 1) :=
  by sorry
```

F Prompt Templates

This section presents all the prompts used to facilitate the reproducibility of our work.

Prompt Template for NL Statements Translation (DeepSeek-V3)

Please autoformalize the following problem in Lean 4 with a header. Use the following theorem names: tm_name. Respond strictly with the translation, without any additional explanation.\n\n

Prompt Template for Back-Translation

[UNUSED_TOKEN_146]user\nConvert the formal statement into natural language:\n``lean\ninformal_statement\n``[UNUSED_TOKEN_145]\n[UNUSED_TOKEN_146]assistant\n

Prompt Template for NLI Check

You are an experienced mathematics expert and educator with extensive experience in mathematical problem analysis. I need you to analyze the fundamental nature of the following two mathematical problems.

Focus on:

1. Core mathematical concepts and principles
2. Problem-solving approaches and methodologies
3. Ultimate objectives of the problems

Ignore:

1. Variations in wording
2. Changes in contextual scenarios

Present your answer using exactly this format:

Analysis\nInsert your analysis here

Conclusion\nreply ||samell or ||differentll with "||" format

Please approach this analysis with professional rigor.

Math Problem 1: {informal_statement}

Math Problem 2: {back_translation}

Prompt Template for NL Statements Generation

You are an expert mathematics professor tasked with creating proof problems for undergraduate mathematics majors. Your assignment is to construct a proof problem that integrates {concept1} from {domain1} and {concept2} from {domain2}.

Requirements:

1. Create a concise theorem appropriate for undergraduate mathematics majors.
2. The theorem should be brief, not exceeding 50 words.
3. Incorporate both specified concepts into the theorem naturally.
4. State the theorem clearly and concisely.
5. Ensure the theorem is simple enough to be easily translated into Lean4.

Format exactly:

Answer\nInsert your problem with "||" format, i.e. ||Theorem: Insert the theorem in natural language here.||

Prompt Template for NL Statements Translation (ATLAS)

You are an expert in the Lean4 theorem prover. Your task is to translate theorems from natural language into formal Lean4 statements. Please follow these guidelines:

1. Carefully analyze the given theorem in natural language.
2. Translate it into a correct and precise Lean4 formal statement.
3. Use the following format for your response: theorem tm_name : The theorem's Lean4 formal statement := by sorry
4. Focus solely on the translation. Do not attempt to prove the theorem or provide additional explanations.
5. Ensure that your translation accurately captures all the mathematical concepts and relationships expressed in the natural language version.
6. Use appropriate Lean4 syntax, including correct use of quantifiers, implications, and mathematical symbols.
7. If the theorem involves specific mathematical structures (e.g., groups, rings, topological spaces), use the corresponding Lean4 definitions and notations.

Remember, the goal is to create a syntactically correct and semantically accurate formalization in Lean4. Your translation should be faithful to the meaning of the original theorem while adhering to Lean4 conventions and best practices.

Now please begin by carefully reading the natural language statement provided, and then proceed with your translation into Lean4.
{informal_statement}

Prompt Template for FL Statements Revision

You are a math expert and an expert in Lean4. Your task is to modify the Lean4 code based on the given natural language description of a theorem, the corresponding Lean4 code, and the error message from the Lean compiler.

Requirements:

1. Correct the Lean4 code to make it compile successfully.
2. Lean4 code may lack or have additional declarations of certain content. You can add or remove them as much as possible to keep it consistent with the natural language description.
3. No need to import any packages, because Mathlib will be imported by default as import Mathlib.
4. Carefully read the content and provide your modified answer: `**Lean4 code**\n{formal_statement}\n**Compiler error messages**\n{compiler_error_messages}\n**natural language statement**\n{informal_statement}`

Format exactly:

```
# Analysis\nInsert your analysis here\n# Answer\nInsert your revised Lean4 code with "||" format, i.e. ||theorem tm_name your revised Lean4 code here := by sorry||
```

Prompt Template for FL Statements Alignment

You are a math expert and an expert in Lean4. Your task is to check the alignment between the given natural language description of a theorem and the corresponding Lean4 code.

Requirements:

1. Determine whether the Lean4 code is missing declarations of certain entities.
2. Assess whether the Lean4 code accurately represents the theorem described in the natural language.
3. Carefully read the content and provide your answer: `**Lean4 code**\n {formal_statement}\n**natural language statement**\n{informal_statement}`

Format exactly:

`# Analysis:\nInsert your analysis here`

`# Answer\nreply ||good||, ||average|| or ||poor||`

Prompt Template for FL Statements Translation

You are a math expert and an expert in Lean4. Your task is to translate theorems from Lean4 code into natural language.

Requirements:

1. Focus solely on the translation. Do not attempt to prove the theorem or provide additional explanations.
2. The theorem's natural language statement should be brief, not exceeding 50 words.
3. Carefully analyze the given theorem in Lean4 code `{formal_statement}` and provide your translation in natural language.

Format exactly:

`# Answer\nInsert your translation with "||" format, i.e. ||Theorem: Insert the theorem in natural language here.||`