Learn from Failure: Fine-Tuning LLMs with Trial-and-Error Data for Intuitionistic Propositional Logic Proving

Chenyang An¹*, Zhibo Chen²*, Qihao Ye¹*, Emily First¹, Letian Peng¹
Jiayun Zhang¹, Zihan Wang¹†, Sorin Lerner¹†, Jingbo Shang¹†
University of California, San Diego¹ Carnegie Mellon University²
{c5an, q8ye, emfirst, lepeng, jiz069, ziw224, lerner, jshang}@ucsd.edu
zhiboc@andrew.cmu.edu

Abstract

Recent advances in Automated Theorem Proving have shown the effectiveness of leveraging a (large) language model that generates tactics (i.e. proof steps) to search through proof states. The current model, while trained solely on successful proof paths, faces a discrepancy at the inference stage, as it must sample and try various tactics at each proof state until finding success, unlike its training which does not incorporate learning from failed attempts. Intuitively, a tactic that leads to a failed search path would indicate that similar tactics should receive less attention during the following trials. In this paper, we demonstrate the benefit of training models that additionally learn from failed search paths. Facing the lack of such trial-and-error data in existing open-source theorem-proving datasets, we curate a dataset on intuitionistic propositional logic theorems and formalize it in Lean, such that we can reliably check the correctness of proofs. We compare our model trained on relatively short trial-and-error information (TRIALMASTER) with models trained only on the correct paths and discover that the former solves more unseen theorems with lower trial searches.

1 Introduction

Automated Theorem Proving is a challenging task that has recently gained popularity in the machine-learning community. Researchers build neural theorem provers to synthesize formal proofs of mathematical theorems (Yang et al., 2023; Welleck et al., 2021; Lample et al., 2022; Mikula et al., 2023; Wang et al., 2023; Bansal et al., 2019; Davies et al., 2021; Wu et al., 2021; Rabe et al., 2020; Kusumoto et al., 2018; Bansal et al., 2019; Irving et al., 2016). Typically, a neural theorem prover, given a partial proof and the current proof state,
gest the model to generate tactics similar to the failed ones sparingly. At the very least, the failure information should help the model easily avoid generating already failed tactics. See Figure 1.

In this paper, we wish to empirically verify this intuition. To conduct the experiment, we would compare the conventional model trained on correct proof paths, and TRIALMASTER, the model trained on the whole proof tree, containing both correct paths and incorrect paths. See Figure 2b. As such, TRIALMASTER can make predictions based on the failure information during inference time.

Since current open-source Automated Theorem Proving datasets do not contain complete proof trees, we create such a dataset, PropL, written in Lean. We focus on theorems of intuitionistic propositional logic. A propositional logic formula in intuitionistic propositional logic $A$ is true if and only if $\vdash A$ has a derivation according to the rules of the intuitionistic propositional logic. These rules are listed in Figure 6 in Appendix A with explanations. This is to be distinguished from the classical logic, where a theorem of classical propositional logic admits a proof if and only if under all truth value assignments to all propositions that appear in the theorem, the theorem statement is evaluated to be true.

We give two elementary examples of theorems of intuitionistic propositional logic with proofs. We then give an example of a theorem whose proof contains a backtracking step. The first theorem admits a proof, where as the second doesn’t.

```
theorem_1 thm : p1 → p1 ∨ p2 := by
  intro h1
  apply or.inr
  exact h1
```

The first theorem states that $p1$ implies $p1 \lor p2$, where $\lor$ stands for "or". To show this, we assume $p1$ and prove either $p1$ or $p2$. We prove $p1$ in this case. The last three lines are tactics (proof steps) in Lean used to prove this fact.

Our second example looks like the following.

```
theorem_2 thm : p1 → p1 ∧ p2 := by
```

The second theorem states that $p1$ implies $p1$ and $p2$. There is no proof to this theorem. When we assume $p1$, we cannot show both $p1$ and $p2$.

Proofs might include backtracking instructions. Here is another possible proof for theorem 1. After step 2, there is no possible proof step that can lead to final solution. Therefore we backtrack to step 1 and try a different proof step.

```
Note that "no solution, backtrack to step 1" is not a tactic in Lean. It simply tells the system to backtrack to step 2 and start reasoning from there.

Specifically, our PropL dataset is created through a two-stage process that first involves generating a comprehensive set of propositional logic theorems by uniformly sampling from all possible theorems, utilizing a bijection between natural numbers and propositions to ensure representativeness. Following theorem generation, proofs are constructed using a focusing method with polarization, incorporating a detailed trial-and-error search process that includes both successful and backtracked steps, thereby capturing the complexity and nuances of theorem proving in intuitionistic propositional logic. Thus, our dataset is complete, scalable, and representative. The proofs in our dataset are combined with trial-and-error information, which is generated by the Focused Proof Search (FPS) algorithm (McLaughlin and Pfenning, 2009; Liang and Miller, 2009; Pfenning, 2017).

We verify the effectiveness of incorporating the failed trials during training and inference by experiments on PropL, observing that TRIALMASTER achieves a higher proof search success rate and lower search cost over conventional model trained on correct proof paths. Our experiments further indicate that our model can perform backtracking without help from an external system.

Our main contributions are as follows:

• We establish, PropL, a complete, scalable, and representative benchmark for intuitionistic propositional logic theorems formalized in Lean. PropL includes proofs with trial-and-error information, generated by the FPS algorithm.

• We demonstrate that for intuitionistic propositional logic theorems proving, incorporating trial-and-error information into training and proving outperforms a conventional model that is trained on correct proofs only.

2 Related Work

Automated Theorem Proving. Automated Theorem Proving has evolved significantly since its inception, focusing mainly on developing com-
Figure 2: Method comparison. (a) A conventional system: The tactic generator (i.e., LLM) is fine-tuned on correct proof paths only. During inference, the trained tactic generator produces $N_{\text{sampled}}$ (e.g., 2 in the example) tactics at a time. If Lean decides that the current tactic is wrong, the system backtracks to the last valid state and tries other candidate tactics. (b) Our methodology: The tactic generator is fine-tuned on proofs with trial-and-error. During inference, we take the first tactic it generates and feed that into Lean for state checking at each step.

Propositional Logic Problem. Early implementations of ATP systems demonstrated the potential for computers to automate logical deductions, with notable examples including the Logic Theorist (Crevier, 1993; McCorduck, 2004; Russell and Norvig, 2010) and Gilmore’s program (Davis, 2001; Gilmore, 1960). These systems laid the groundwork for the resolution of propositional logic problems, showcasing the ability of automated systems to handle logical reasoning tasks. Recent advancements in Automated Theorem Proving have revisited propositional logic problems, integrating modern computational techniques. Sekiyama et al. (Sekiyama and Sue-naga, 2018) have employed Deep Neural Networks (DNNs) as a statistical approach to generate proofs for these theorems, while Kusumoto et al. (Kusumoto et al., 2018) have explored graph representations coupled with reinforcement learning to find proofs. Furthermore, sequence-to-sequence neural networks have been applied for deriving proof terms in intuitionistic propositional logic (Sekiyama et al., 2017). This area of research is particularly intriguing due to the simplicity and importance of propositional logic in mathematics, and there is a growing interest in evaluating the capability of LLMs in tackling this specific mathematical domain.

Trial-and-Error. The Chain-of-Thought (CoT) (Wei et al., 2022; Wang et al., 2022; Zhou et al., 2022; Fu et al., 2022; Chu et al., 2023; Yu et al., 2023) approach demonstrates that LLMs can be guided to perform step-by-step reasoning by incorporating intermediate reasoning steps in...
their prompts. This concept is expanded in later research, such as the Tree of Thoughts (ToT) (Yao et al., 2023), which organizes reasoning into a tree structure, and the Graph of Thoughts (GoT) (Besta et al., 2023), which adopts a graph format for thought structuring. Trial-and-error complements structured reasoning by allowing the model to empirically test hypotheses generated, thereby refining its reasoning process based on feedback from interactions or emulations. The Boosting of Thoughts (BoT) (Chen et al., 2024) prompting framework iteratively explores and evaluates multiple trees of thoughts to gain trial-and-error reasoning experiences, using error analysis from the LLMs to revise the prompt. Lyra introduces correction mechanisms to improve the performance of the tactic generator (Zheng et al., 2023).

3 PropL: A New Dataset for Intuitionistic Propositional Logic Theorems in Lean

Our aim is to experimentally validate that trial-and-error information can enhance models’ ability to do backtracking and tactic generation for theorem-proving tasks. Given that existing open-source theorem proving datasets lack information on trial-and-error processes, we have developed PropL, which is based on theorems of intuitionistic propositional logic. This dataset uniquely includes proofs that encapsulate the complete search process, incorporating the trial-and-error data generated by the FPS algorithm. Our dataset has two other benefits. It is formalized in Lean, so that the validity of the theorems and proofs are guaranteed. The tactics generated by the model trained on PropL can also be directly sent to Lean to be checked. PropL is also representative of all the intuitionistic propositional logic theorems, since by uniformly sampling integers, we can use a bijection between natural numbers and propositions to uniformly sample propositions. This bijection is explained in the Theorem Generation section.

3.1 Data Generation of PropL

PropL comprises theorems uniformly sampled from the entire set of propositional logic theorems. It includes various proof types for each theorem. We only report proof types that are used in this paper. For additional information about the dataset, please refer to the GitHub repository and Huggingface.

The construction of PropL involves two primary stages: the generation of propositional logic theorems and the generation of proofs for these theorems from an existing algorithm.

Theorem Generation. Consider the set of propositions $A$ with at most $p$ atomic propositions. $A$ can be inductively generated by the following grammar:

$$A, B ::= P_i \mid T \mid F \mid A \land B \mid A \lor B \mid A \to B,$$

where $P_i$ is the $i$-th atomic proposition with $1 \leq i \leq p$, $T$ stands for True, and $F$ stands for False. We note that an atomic proposition is just a statement like "Today is sunny". The concrete meaning of the statement is irrelevant with respect to the theorem proving task in this paper. A connective $\land, \lor, \to$ is called an internal node of the proposition. We assign the following lexicographic order to propositions:

1. Number of internal nodes (increasing order)
2. Number of internal nodes of the left child (decreasing order)
3. Top level connective, $T < F < P_1 < \cdots < P_p < \land < \lor < \to$
4. The (recursive order (2 - 5)) of the left child
5. The (recursive order (2 - 5)) of the right child

For a fixed upper bound $p$ for the number of atomic propositions, we establish a bijection between the natural numbers and the set of propositional logic formulas. The counting can be made efficient using the Catalan Numbers (Atkinson and Sack, 1992). Figure 3 gives an example of mapping between propositions and natural numbers. Details about the encoding and decoding algorithms are provided in Appendix B.

Proof Generation. Given a randomly sampled theorem, the proof of the theorem is constructed using the focusing method with polarization (McLaughlin and Pfenning, 2009; Liang and Miller, 2009; Pfenning, 2017). Proof search is divided into two stages: inversion and chaining. The inversion phase mechanically breaks down negative connectives (e.g. implications) in the goal and positive connectives (e.g. disjunctions) in the premises. After inversion, chaining will pick an implication in the premise or show one of the disjuncts in the conclusion, with backtracking. The proof search procedure terminates when the same atomic proposition appears in both the premise and the conclusion. An example of
proofs with trial-and-error information (backtracking) and with trial-and-error information removed are shown in Figure 4.

Once proofs are generated, we use them to fine-tune models and start the proof search on the test set.

The polarization of the connectives affects the behavior of the inversion and the search procedure. We choose to uniformly polarize conjunctions that occur negatively (e.g. on the right-hand side of an arrow) as negative and polarize conjunctions that occur positively (e.g. on the left-hand side of an arrow) as positive. Atomic propositions are assigned polarities based on the connective that they first appear under.

To improve the runtime of the search procedure, we make an additional assumption that once an implication is picked, the implication cannot be used to show its premise. In theory, this introduces incompleteness into the search procedure, but it only affects 1 theorem out of around 1000 provable theorems randomly sampled.

### 3.2 Construction of Training and Test Sets

In this section, we explain how we construct the datasets for training and evaluation. We want to avoid training and testing on similar data. In order to test the model performance on harder out-of-distribution (OOD) tasks, we need to ensure that the lengths of the proofs in the training data are shorter than the lengths of the proofs in the test data.

Given PropL, we fix the number of internal nodes in the theorem statement to be 16 (explained in the dataset generation section). We then uniformly randomly sample 200,000 theorems from PropL, which can be achieved by using the integer-theorem bijection as explained before. Our method ensures that the theorems we study in this paper are representative of the propositional logic theorems in general.

We first apply our deterministic algorithms to generate the proofs of the 200,000 theorems, and then remove the trial-and-error information of those proofs. We get a word-length distribution of those proofs without trial-and-error information. Next, to ensure the diversity of the trial-and-error information, we randomly select propositions to focus on during the chaining phase of the proof search, and then generate 10 different proofs with backtracking. By using the average length of the 10 different proofs, we have another word length distribution of proofs with trial-and-error information.

We then split the 200,000 theorems into training and test sets based on both of the word length distributions mentioned above. The word lengths of the proofs of the training data theorems fall within the lower 0.66 quantile of the two distributions of the word length of all the proofs of the 200,000 theorems (109,887 in total). The word lengths of the in-distribution test set also fall in that category (1000 in total). The word lengths of the proofs among the out-of-distribution test theorems are above 0.8 quantile (1000 total) of the two distributions of the word lengths of all the proofs of the 200,000 theorems. We note that the baseline model and TRIALMASTER are trained on proofs of the same set of theorems.
Table 1: The scale of training and test split in our PropL dataset.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Number of theorems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>109,887</td>
</tr>
<tr>
<td>In-dist. test set</td>
<td>1,000</td>
</tr>
<tr>
<td>Out-of-dist. test set</td>
<td>1,000</td>
</tr>
</tbody>
</table>

4 Methodology

4.1 LLM Fine-Tuning

We utilize the training set formed in PropL for training both TRIALMASTER and the tactic generator in the DFS system. The numbers of theorems used in the training and test datasets are presented in Table 1.

LLM fine-tuning with trial-and-error. In our approach, we randomly select two out of the shortest five among the ten proofs with trial-and-error information for each theorem in the training set and utilize them to train TRIALMASTER. Refer to Figure 2b for a description of this training process.

LLM fine-tuning in a DFS system. For the tactic generator of the DFS system, we employ the deterministic FPS algorithm to generate the proofs of the theorems in the training set. The trial-and-error information is removed from the proofs. The LLM is then fine-tuned on the proofs without trial-and-error information as the conventional methods do. Figure 2a illustrates the training process of the DFS system.

4.2 Inference

Inference method of model trained with trial-and-error. TRIALMASTER conducts inference on itself without any help from a backtracking system like DFS or BFS. It outputs two kinds of tactics: tactics in Lean and backtrack instructions. An example of a backtrack instruction would be like “no solution, return to state 2 [that leads to state 4]”, where state 4 is the current state. When TRIALMASTER is doing a proof search on the test set, it is prompted with all history paths, including previous tactics, states, the backtracking it made before, and the failed search path. It then outputs the entire proof path after. Nonetheless, we only utilize the first tactic in the output and employ Lean as a calculator to determine the next state, thereby ensuring the correctness of the state following the tactic. If the tactic outputted by TRIALMASTER is a backtrack instruction, it is then prompted with all the proof search history including the backtrack instruction and the state that the backtrack instruction says to return to. If that tactic is not a backtrack instruction, the tactic and the current state will be fed into Lean for producing the state after. TRIALMASTER is then prompted with the entire proof tree including the state that Lean calculated, and it should output a tactic again. This process is repeated until Lean identifies that the proof is complete or any Lean error occurs. We also note that TRIALMASTER only outputs one tactic at each state using greedy search.

Inference method of the DFS system. There are two hyperparameters in the DFS system: temperature $t$ and the number of sampled tactics $N_{\text{sampled}}$. The temperature $t$ decides the diversity of the model outputs. As $t$ increases, the outputs of the model become more varied. The second parameter determines how many tactics the tactic generator of the DFS system produces for a new proof state.

During inference, the LLM in the DFS system produces $N_{\text{sampled}}$ of candidate tactics at each new state. For each proof state, the DFS only makes one inference. If any two of the generated tactics for the same state are the same, we remove one of them to ensure efficiency. We also remove the tactic suggestions that fail to follow the grammar of Lean. The system follows the depth-first order to keep trying untried tactics. If the system exhausts all the tactics for a given state but has not found a valid one, the system returns to the parent state and then keeps trying untried tactics for the parent state. The overview is presented in the Figure 2a.

To fully exploit the ability of the DFS system, we varied the parameters of it, such as temperature and the number of sampled tactics. We count how many times Lean has been called to check tactics for both the DFS system and TRIALMASTER during the inference stage.

Why do we choose DFS over BFS? While the breadth-first-search (BFS) system is also popular for building neural provers in Automated Theorem Proving, we have opted for DFS as our baseline over BFS in the context of propositional logic theorem proving. This is due to the finite number (around 20) of tactics available at any step for the search process of intuitionistic propositional logic theorems, making DFS more efficient than BFS without compromising the success rate.
5 Evaluation

5.1 Experimental Setup

**Base LLM.** We used Llama-2-7b-hf (Touvron et al., 2023) as the backbone LLM for tactic generation. Models are trained on two A100 GPUs for a single epoch with batch size set to 4. Huggingface (Wolf et al., 2019) is used for fine-tuning the models, and VLLM (Kwon et al., 2023) is used for inference of the models for efficiency. The learning rate of all training processes is set to be $2 \times 10^{-5}$.

**Hyperparameters.** In the experiment, we vary temperature $t$ from $\{0.3, 0.7, 1.2, 1.5, 2.0\}$ and number of sampled tactics $N_{\text{sampled}}$ from $\{2, 5, 10, 15, 20\}$. We notice that in all experiments for temperature $t$ less than 1.5, there were only very few (less than 15) theorems that were still under search when the total steps for that search attempt reached 65. For $t$ higher than 1.5, $N_{\text{Lean}}$ starts to increase dramatically. At temperature 2 with $N_{\text{sampled}}=20$, $N_{\text{Lean}}$ climbs up to 32,171 when the number of total steps reaches 65. Therefore, in our experiment, we set 65 as the search step limit to control time complexity.

5.2 Evaluation Metrics

**Proof search success rate.** We use proof search success rate as our primary metric, which represents the ratio of successful searches by the model. A search attempt for a theorem is marked as successful if Lean outputs state “no goal, the proof is complete”, implying the theorem is effectively proved after applying a tactic produced by the model. For TRIALMASTER, a search attempt ends and fails immediately after one of the following conditions: 1) the word length of the proof tree with trial-and-error exceeds 1500 (for the sake of context length of the model), or 2) the tactic produced by the model at any step induces a Lean error. For the conventional DFS system, a search attempt fails when one of the following conditions happens: 1) the word length of the proof tree without trial-and-error exceeds 1500, or 2) all tactics generated for the initial states have been explored and failed, or 3) the total search steps exceed 65 (see Section 5.1 for the choice of this value). We note that the 1500-word limit is stricter for our method since it produces entire proof paths including trial-and-error and thus would be easier to hit the limit.

**Search cost.** We define a metric to assess the search cost—the total number of Lean calls for tactic checking during proof search for the entire test set, denoted as $N_{\text{Lean}}$. Given the same proof search success rate, a lower $N_{\text{Lean}}$ indicates a more efficient system for proof search. Note that backtracking instructions from our method do not require applying Lean to check the state and, consequently do not add search cost.

5.3 Results and Analysis

**TRIALMASTER outperforms conventional DFS system.** We begin by evaluating the methods of the in-distribution test set. Table 2 illustrates that both our method and the DFS system perform exceptionally well, achieving a success rate of nearly 100% in most configurations. This suggests that Llama-7b effectively masters in-distribution intuitionistic propositional logic theorems. Then, we compare the performance of the methods on the out-of-distribution task. The results are presented in Figure 5. Our method with trial-and-error significantly outperforms the DFS system across various hyperparameter configurations. Additionally, we observe that feeding more proofs without trial-and-error for LLM fine-tuning does not further improve the performance.

**Impact of hyperparameters in the DFS system.** As shown in Figure 5, on the OOD task, although the success rate of the DFS system gets higher when we increase the temperature $t$ or the number of sampled tactics $N_{\text{sampled}}$, the search cost (reflected by $N_{\text{Lean}}$) also goes up. Specifically, when we increase $N_{\text{sampled}}$, the DFS system explores a larger pool of candidate tactics during the search, leading to a higher number of Lean calls. In contrast, our method does a greedy search to generate only one tactic for each new state. Likewise, as $t$ increases, the tactic generator of the DFS system tends to produce more diverse tactics at each proof state, improving the system’s performance but leading to higher search costs.

---

**Table 2: Performance on in-distribution task.** Both methods perform well for propositional logic.

<table>
<thead>
<tr>
<th>Model</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRIALMASTER</td>
<td>100%</td>
</tr>
<tr>
<td>$t = 1.2$</td>
<td>$N_{\text{sampled}} = 2$ 99.5%</td>
</tr>
<tr>
<td>$N_{\text{sampled}} = 5$ 99.9%</td>
<td></td>
</tr>
<tr>
<td>$N_{\text{sampled}} = 10$ 99.6%</td>
<td></td>
</tr>
<tr>
<td>DFS</td>
<td></td>
</tr>
<tr>
<td>$t = 2.0$</td>
<td>$N_{\text{sampled}} = 2$ 75.9%</td>
</tr>
<tr>
<td>$N_{\text{sampled}} = 5$ 97.3%</td>
<td></td>
</tr>
<tr>
<td>$N_{\text{sampled}} = 10$ 99.0%</td>
<td></td>
</tr>
</tbody>
</table>
TriALMaster achieves high success rates at lower search cost. For a direct comparison on the OOD task. For both methods, we mark the search as failed if the tactic induces a Lean error, beside the marker. (c) Comparison of NLean among our method and top 3 DFS systems with the highest success rate. In summary, training with trial-and-error achieves a higher success rate with a relatively lower search cost compared to the DFS systems.

Table 3: Ablation study: Comparison of TrialMaster and model trained without trial-and-error information on OOD task

<table>
<thead>
<tr>
<th>Model</th>
<th>Success rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>TrialMaster</td>
<td>88.7%</td>
</tr>
<tr>
<td>Model - proof w/o t.a.e.</td>
<td>59.3%</td>
</tr>
</tbody>
</table>

**TriALMaster** achieves high success rates at lower search cost. For a direct comparison of search costs, we plot the $N_{\text{Lean}}$ values of our method alongside those of the top three DFS systems with the highest success rates among all the hyperparameters we experimented with, i.e., $t=1.5$, $N_{\text{sampled}}=20$ (87.2%), $t=1.5$, $N_{\text{sampled}}=10$ (86.1%), and $t=1.8$, $N_{\text{sampled}}=10$ (88.4%). This comparison is illustrated in Figure 5c. Notably, we observe that the DFS systems that closely approach our model’s performance exhibit significantly higher search costs. With $t=1.8$, $N_{\text{sampled}}=10$, $N_{\text{Lean}}$ of the DFS system, which has 0.3% lower success rate than our method, has reached 31,101, which is 72% higher than that of our method with trial-and-error. The high search cost makes the DFS system with high temperatures unfavorable. These results demonstrate that training with trial-and-error produces higher-quality tactics, achieving a higher success rate with relatively lower search cost.

Model learns backtracking capability from trial-and-error data. In the experiments, we find out that our TriALMaster successfully acquires the backtracking capability from proofs with trial-and-error information. This is evidenced by the fact that during TriALMaster’s proof search for theorems in the test set, all backtracking instructions produced by the LLM adhere to the correct format and point to existing state numbers.

**Including failed search paths helps TriALMaster to learn.** The following experiment shows that adding failed search paths to the training data for TriALMaster results in an overall gain. In this experiment, the model is only trained to learn the correct search paths and the backtracking instructions. The model is not trained to learn the failed search paths (we don’t compute the loss for the failed search paths during the training stage in this case). The proof success rate in this case is 75.6%, which is lower than TriALMaster’s proof success rate of 88.7%. The $N_{\text{Lean}}$ for the model is 13,600, which is lower than that of TriALMaster. This is expected since the model does not learn to predict failed search paths. Our explanation for why TriALMaster has a higher proof search success rate than the model trained in the previously mentioned experiment is that the failed search paths also contribute to improving the proof search success rate of the model. TriALMaster strategically tried some potentially failed search paths to gain a more comprehensive view of the problem, which then led to the correct search paths.

5.4 Ablation Study

To evaluate the effectiveness of training with trial-and-error, we craft an ablated version of our method where the LLM is finetuned with data of the correct path only and do inference in the same way as our method (i.e., producing one tactic at a time and applying Lean for state checking). We denote the ablated version as Model - proof w/o t.a.e.. For both methods, we mark the search attempt as failed if the tactic induces a Lean error,
or the search exceeds the 1500-word limit. The result is shown in the Table 3. The difference between the success rates of the two models is 29.4%, which is significant. This clearly shows that failed search states and trial-and-error information tremendously enhance the model’s capability to solve theorem-proving tasks.

5.5 Exploratory Study: Effect of Training Proof Length on Model Performance

Since the FPS algorithm of PropL dataset can generate multiple proofs with variable length, we conduct an exploratory study to assess the impact of proof length on model performance. We fine-tune two models using proofs with different lengths of trial-and-error information. For the first model, which is our TRIALMASTER, the training data is derived by randomly selecting two out of the shortest four proofs from the ten available proofs for each theorem in PropL. We denote it as Model - short proof w/ t.a.e. In contrast, the training data of the second model is formed by randomly selecting two proofs from the ten available for each theorem, irrespective of their lengths. We denote it as Model - long proof w/ t.a.e. For both models, we use greedy search to let them generate one tactic for each state. We evaluate the models on our 1000 OOD test set. The results are shown in the Table 4. A higher success rate is observed in the model trained with shorter proofs. This can be attributed to the fact that as the proof with trial-and-error information becomes longer, there is too much trial-and-error information that may detrimentally affect the model’s performance, as too many failed search paths may lower the quality of the training data.

6 Conclusion and Future Work

In this paper, we study Automated Theorem Proving in formalized environments. We create a complete, scalable, and representative dataset of intuitionistic propositional logic theorems in Lean. We demonstrate that leveraging information from failed search states and backtracking not only teaches models how to backtrack effectively, but also helps in developing better tactics than those generated by models trained without access to backtracking insights. We release our datasets on GitHub and Huggingface.1

A natural extension of our research involves investigating whether trial-and-error information is beneficial for more general mathematical theorem-proving settings. Exploring this avenue could provide valuable insights into the effectiveness of our approach across broader mathematical domains.

Limitations

One limitation of our study is that some proof attempts are forced to stop due to the prompt exceeding the context length of 1500 tokens. This constraint may potentially influence our results by truncating the available information during the proof search process.

Furthermore, our method was not evaluated on general mathematical theorems. This limitation arises from both the scarcity of proofs containing trial-and-error information in current math libraries and the intrinsic challenges associated with producing proofs, whether with or without backtracking, for general mathematical theorems in a formalized setting.

Automated theorem proving with LLMs is an emerging area in machine learning. There is still a lack of baselines on LLMs to compare with our method. We establish a fundamental baseline, but we still need accumulative work to provide methods for comparison.

Ethical Consideration

Our work learns large language models to automatically prove propositional logic theorems, which generally does not raise ethical concerns.

Acknowledgments

This work is supported by the National Science Foundation under grants CCF-1955457 and CCF-2220892. This work is also sponsored in part by NSF CAREER Award 2239440, NSF Proto-OKN Award 2333790, as well as generous gifts from Google, Adobe, and Teradata.

1PropL dataset is available at https://huggingface.co/datasets/KomeijiForce/PropL. Model weights are available at https://huggingface.co/KomeijiForce/l1lama-2-7b-propositional-logic-prover. Generation codes for theorems and proofs are available at https://github.com/ucsd-atp/PropL.
References


A Inference Rules for Intuitionistic Propositional Logic

Figure 6 shows the inference rules for intuitionistic propositional logic. Rules relate sequents of the form $\Gamma \vdash A$, where $\Gamma$ is an unordered list of assumptions. A derivation is constructed from the axiom (Assumption) using the derivation rules until $\Gamma$ is empty.

For example, the $(\rightarrow I)$ rule says that from a derivation of $\Gamma \vdash B$ under the assumption of $\Gamma, A$, we can get a derivation of $\Gamma \vdash A \rightarrow B$. The $(\rightarrow E)$ rule says that from a derivation of $\Gamma \vdash A \rightarrow B$ and a derivation of $\Gamma \vdash A$ we can derive $\Gamma \vdash B$.

Here are an example derivation.

\[
\begin{align*}
\Gamma, A & \vdash A \quad \text{(Assumption)} \quad \Gamma \vdash T \\
\Gamma \vdash F & \quad \Gamma \vdash A \quad \Gamma \vdash B \quad \Gamma \vdash A \wedge B \\
\Gamma \vdash A & \wedge B \quad \Gamma \vdash A \quad \Gamma \vdash B \quad \Gamma \vdash A \wedge B \\
\Gamma \vdash A \quad \Gamma \vdash B & \quad \Gamma \vdash A \wedge B \\
\Gamma \vdash A \wedge B & \quad \Gamma \vdash A \quad \Gamma \vdash B \\
\end{align*}
\]

B Uniformly Distributed Data Explanation

In this section, we discuss the uniform characteristics of our dataset, particularly emphasizing the one-to-one mapping between propositions and natural numbers. This bijection allows us to simply sample from the natural numbers to ensure the dataset exhibits uniformity.

Algorithm 1 Encoding

Require: A tree $T$ representing a proposition, with $n$ indicating the number of internal nodes and $p$ representing the number of atomic propositions

Ensure: Output a natural number

1: return $3^n(p+2)^{n+1} \times \text{SHAPE NUMBER}(T)$ + ASSIGNMENT NUMBER($T$)
2: function SHAPE NUMBER($T$)
3: if $T$ is a single node then return $0$
4: end if
5: $T_L, T_R \leftarrow$ left and right sub-trees of $T$
6: $n_l, n_r \leftarrow$ number of internal nodes in $T_L, T_R$ \hspace{1cm} \triangleright Total $n_r + n_l + 1$ internal nodes
7: return $\sum_{i=1}^{n_l} C_{i-1} C_{n_l+n_r+1-i} + C_{n_r} \times \text{SHAPE NUMBER}(T_L) + \text{SHAPE NUMBER}(T_R)$
8: end function
9: function ASSIGNMENT NUMBER($T$)
10: $N \leftarrow \text{NODE NUMBER}(\text{root node of } T)$
11: if $T$ is a single node then return $N$
12: end if
13: $T_L, T_R \leftarrow$ left and right sub-trees of $T$
14: $n_r \leftarrow$ number of internal nodes in $T_R$
15: $A_r \leftarrow 3^{n_r}(p+2)^{n_r+1}$ \hspace{1cm} \triangleright Compute all possible assignments in the right sub-tree
16: return $3A_r \times \text{ASSIGNMENT NUMBER}(T_L) + A_r \times N + \text{ASSIGNMENT NUMBER}(T_R)$
17: end function
18: function NODE NUMBER($N$)
19: switch $N$
20: case $\wedge$ or $\vee$: return $0$
21: case $\rightarrow$: return $1$
22: case $\rightarrow$: return $2$
23: case $P_i$: return $i + 1$ \hspace{1cm} \triangleright $1 \leq i \leq p$
24: end function

B.1 Catalan Number

The Catalan number $C_n$ is applicable for counting full binary trees that consist of $n+1$ leaf nodes, corresponding to exactly $n$ internal nodes (Stanley, 2015). Additionally, it can be calculated through recursion as shown:

$C_n = \sum_{i=1}^{n} C_{i-1} C_{n-i}.$
The first Catalan numbers for \( n = 0, 1, 2, 3, 4 \) are 1, 1, 2, 5, 14.
A concise interpretation of this recursion is as follows: it involves counting the number of internal nodes in the left sub-tree, amounting to \( i - 1 \), and then in the right sub-tree, amounting to \( n - i \), for each of the \( n \) scenarios in computing \( C_n \).

**Algorithm 2 Decoding**

**Require:** A natural number ID, with \( n \) indicating the number of internal nodes and \( p \) representing the number of atomic propositions

**Ensure:** Output a proposition tree

1: quotient, remainder ← DIVMOD(ID, \( 3^n(p + 2)^{n+1} \))
2: \( S \leftarrow \text{TreeShape}(\text{quotient}, n) \)
3: \( A \leftarrow \text{TreeAssignment}(S, \text{remainder}) \)
4: return a tree with shape \( S \) and assignment \( A \)
5: **function** TreeShape(\( N, n \))
6: if \( n \) is 0 then return a single node
7: end if
8: \( n_l \leftarrow \max(\{ k | \sum_{i=1}^{k} C_{i-1} C_{n-i} \leq N \} \cup \{0\}) \)
9: \( n_r \leftarrow n - n_l - 1 \)
10: remaining ← \( N - \sum_{i=1}^{n_l} C_{i-1} C_{n-i} \)
11: \( N_l, N_r \leftarrow \text{DIVMOD}(\text{remaining}, C_{n_r}) \)
12: return a tree with left and right sub-trees shaped by TreeShape(\( N_l, n_l \)) and TreeShape(\( N_r, n_r \)), respectively
13: **end function**
14: **function** TreeAssignment(\( S, N \))
15: Perform an inorder traversal of the tree with shape \( S \). For each leaf node, interpret it as a digit in base \( p + 2 \), and for each internal node, interpret it as a digit in base 3. Compute the assignment that gives \( N \) by utilizing the \text{NodeNumber} function introduced in Algorithm 1
16: **end function**

### B.2 Bijection between Propositions and Natural Numbers

As depicted in Figure 3, every proposition corresponds to a unique tree representation. Consequently, it only requires the identification of a bijection between full binary trees and natural numbers. For every full binary tree possessing \( n \) internal nodes and \( p \) atomic propositions, there exist \( 3^{\#\text{internal node}}(p + 2)^{\#\text{leaf node}} = 3^n(p + 2)^{n+1} \) distinct cases. Observe that the choices available for internal nodes include conjunction (\( \wedge \)), disjunction (\( \lor \)), and implication (\( \rightarrow \)); whereas for leaf nodes, the choices encompass true (\( T \)), false (\( F \)), and a set of propositions \( P_1, \ldots, P_p \).

This counting facilitates an efficient ranking of all full binary trees with \( n \) internal nodes. This ranking process inherently establishes a bijection with the set of natural numbers, allowing for a clear correspondence between each full binary tree and a unique natural number. Consequently, this sets the stage for a detailed examination of two critical processes: encoding (see Algorithm 1), which involves mapping a proposition tree to a natural number, and decoding (see Algorithm 2), which entails mapping a natural number back to a corresponding proposition tree.

Having established the bijection between full binary trees and natural numbers, it becomes apparent that uniformly sampling from the set of natural numbers will, in turn, result in a uniform sampling of full binary trees.

The inclusion of the parameter \( n \) in Algorithm 1 is not critical for the functionality of the algorithm as \( n \) can be counted from \( T \); however, it is retained to denote that the most basic encoding and decoding approaches are designed for a fixed number of internal nodes \( n \). For example, ID 0 denotes different propositions depending on the specified value of \( n \). As a result, the decoding algorithm, as outlined in Algorithm 2, is specifically intended for uniformly generating proposition trees with a fixed \( n \). To achieve a uniformly distributed proposition tree from a set of trees with various \( n \), one might simply merge the rankings of fully binary trees with different \( n \), a task which can be performed with relative ease. This approach of merging can be similarly applied to trees with varying \( p \) as well.

Given the uncertainty surrounding the proof lengths when generating propositions, our approach involves uniform sampling for proposition selection. This selection is later refined by excluding propositions according to their proof lengths as computed by the proof generation algorithm.

### C Examples

In Figure 7, we show an example Lean proof for a theorem in Figure 3. Lines preceded by ‘—’ are comments solely for explanatory purposes.
variable (p1 p2 p3 p4 p5 : Prop)
theorem : (((p1 ∨ p2) → False) → ((p1 → False) ∧ (p2 → False))) := by
  -- Implications on the right can always be decomposed.
  intro h1
  -- Now we want to show ((p1 → False) ∧ (p2 → False))
  -- Conjunctions on the right can always be decomposed.
  -- We then need to show (p1 → False) and (p2 → False) separately.
  apply And.intro
  -- We are showing (p1 → False).
  -- Implications on the right can always be decomposed.
  -- We introduce assumption h2 for p1. And we try to show False.
  intro h2
  -- We want to use the implication h1. So we show its premise.
  have h3 : (p1 ∨ p2) := by
  -- Show the left disjunct. (The right adjunct leads to an TAE)
  apply Or.inl
  -- One of the premise coincides with the conclusion.
  exact h2
  -- We have shown the premise of h1 (p1 ∨ p2),
  -- we can now drive its conclusion (False), denoted by h4.
  let h4 := h1 h3
  -- False on the left can always be used.
  apply False.elim h4
  -- We have shown (p1 → False) and now we show (p2 → False).
  -- Implications on the right can always be decomposed.
  -- We introduce assumption h2 for p2. And we try to show False.
  intro h5
  -- We want to use the implication h1. So we show its premise.
  have h6 : (p1 ∨ p2) := by
  -- Show the right disjunct. (The left adjunct leads to an TAE)
  apply Or.inr
  -- One of the premise coincides with the conclusion.
  exact h5
  -- We have shown the premise of h1 (p1 ∨ p2),
  -- we can now drive its conclusion (False), denoted by h7.
  let h7 := h1 h6
  -- False on the left can always be used.
  apply False.elim h7

Figure 7: An example of an intuitionistic propositional logic theorem with its proof in Lean