The Unreasonable Effectiveness of Nonoverlapping Failures in LLM Prover Ensembles

Nicholas Jiang[a,*], Joe Zhou[a] and Rishabh Sharma[a]

Abstract

Improving the reliability of mathematical reasoning in large language models (LLMs) is critical for applications in education, automated theorem proving, and formal verification. This paper investigates whether the functional diversity of prover models, specifically their non-overlapping failures, can be harnessed through ensembling to improve collective performance. We present a theoretical risk decomposition framework for an OR-aggregated ensemble of theorem provers, demonstrating that the ensemble's risk is equal to the average individual risk minus an 'ambiguity effect' that quantifies the diversity of the provers. Our analysis formalizes the intuition that diversity, defined as non-overlapping failures, is strictly beneficial in this context. We hypothesize that such ensembles may not only surpass the accuracy of any individual model but could also potentially generate proofs for statements previously unprovable by any single prover. Furthermore, we aim to investigate whether these techniques can be applied to current state-of-the-art models to push performance on more difficult, unsaturated benchmarks such as PutnamBench.

Keywords: Mathematical Reasoning, Large Language Models (LLMs), Automated Theorem Proving, Ensemble Methods, Model Diversity, Generalization, Formal Verification

1. Introduction

Improving mathematical reasoning in large language model (LLM) ensembles is essential due to its profound implications across education, automated theorem proving, and broader artificial intelligence (AI) reasoning tasks [1,2]. Reliable reasoning enables AI models to generate correct and verifiable solutions, significantly benefiting formal verification of critical software and scientific results, and strengthening AI's general reasoning capabilities [3].

*Corresponding author. E-mail address:

a: nicholas.jiang@uwaterloo.ca
b: yhzou@uwaterloo.ca
c: r342sharma@uwaterloo.ca

Source code and data available at: https://github.com/script-jpg/non-overlapping-failures

November 06, 2025

In the past few years, there has been remarkable progress on benchmarks such as the MiniF2F [4], with recent models achieving >95% accuracy on the test-set [5,6].

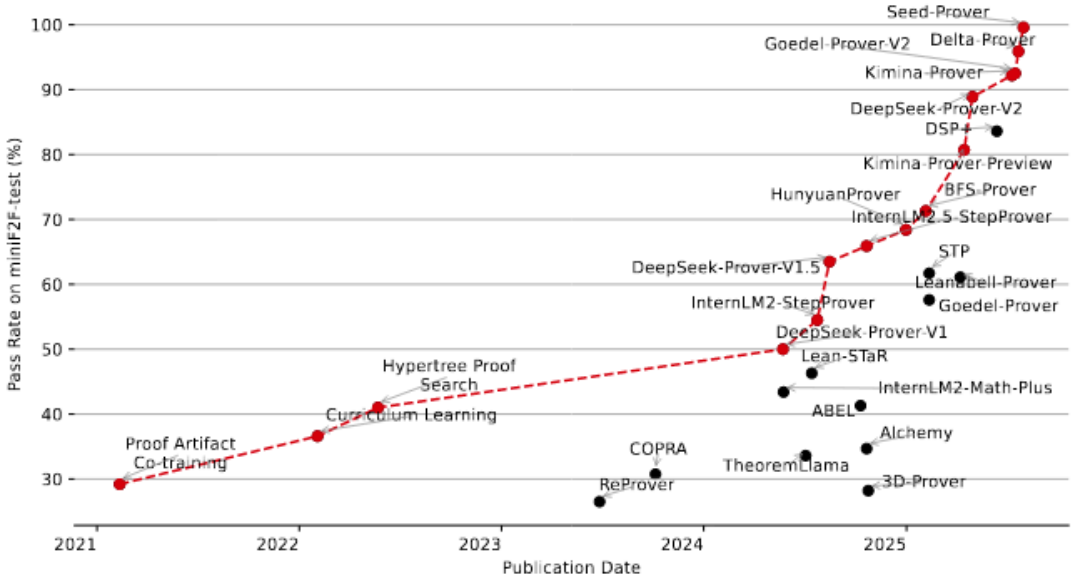

Figure 1: Growth in MiniF2F-Test performance over time from [6]

Here we see that as we improve on the benchmark, we leave many older models unused. As such, in this paper we ask the following questions:

1. Can we make use of non-overlapping failures (i.e. 'functional diversity') of prover models that do not saturate MiniF2F-Test to improve their collective accuracy?
2. Is it possible, via ensembling, to write validated proofs for propositions that no individual model can prove?
3. Can we apply these techniques to the current SOTA models to achieve stronger performance on as-yet unsaturated benchmarks such as PutnamBench? [7]

This paper answers question 1, sketches a theory for question 2, and defers an empirical study of question 3 to follow-on work.

2. Related Work

Previous ensemble strategies, such as majority-vote or confidence-based ensembles, typically measure diversity only at the output level (i.e. the final answers) [8]. There has also been some work to define and evaluate diversity measures on ensembles and determine their correlation to ensemble error [9], though findings have been mixed. Despite this, the crucial notion of "reasoning-path diversity" remains comparatively understudied.

A few recent works on LLMs touch on this idea. For example, Naik et. al [10] discuss reasoning-path diversity through the lense of prompting techniques that encourage diversity

of thought via different combinations of 'personas' and problem-solving approaches (e.g. the pair (Alan Turing, Action Rationale)) when attempting a problem.

We hypothesize that such diversity underpins robust generalization in structured tasks like mathematics.

3. Methodology

3.1. Ensemble Risk Decomposition

Assume $S = \{h_1, ..., h_m\}$ is a set of provers where

$$h_i(x) = \begin{cases} 1 & \text{if } h_i \text{ proves } x \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Define $h_S(x) = \bigvee_{i=1}^{m} h_i(x)$ be the (logical) OR-aggregator over the h_i

Using the 0/1 loss defined as $\ell(x, y) = 1\{x \neq y\}$,

Assume $(X, Y) \sim \mathcal{P}$ where X is a proposition and $Y \in \{0, 1\}$ is it's associated provability. The risk of a prover h is defined as:

$$\begin{aligned} \mathcal{R}(h) &= E_{XY}[\ell(h(X), Y)] && \text{(2a)} \\ &= P(h(X) \neq Y) && \text{(2b)} \\ &= P(h(X) = 1, Y = 0) + P(h(X) = 0, Y = 1) && \text{(2c)} \end{aligned}$$

Note that we can take $P(h(X) = 1, Y = 0) = 0$ because we assume the proof assistant (e.g. Lean) will not give "false positives" (it will not admit a proof when a statement is not provable). Giving:

$$\mathcal{R}(h) = P(h(X) = 0, Y = 1) \tag{3}$$

Now we introduce a simple identity inspired by Prop. 7 from [8]

$$\mathcal{R}(h_S) = \underbrace{\frac{1}{|S|}\sum \mathcal{R}(h_i)}_{\text{average risk}} - \underbrace{\frac{1}{|S|}\sum [\mathcal{R}(h_i) - \mathcal{R}(h_S)]}_{\text{ambiguity effect}} \tag{4}$$

For (4) it must be the case that $\mathcal{R}(h_i) \geq \mathcal{R}(h_S)$ since $h_i(x) \leq \bigvee_{i=1}^{m} h_i(x) = h_S(x)$ because the (logical) OR-aggregator is monotonic. If h_i fails to produce a proof then an OR aggregation of other provers in S can only improve the result (i.e. by successfully proving the statement). In other words, assuming $h_i \in S$:

$$\{(X, Y) \mid h_S(X) \neq Y\} \tag{5a}$$
$$\subseteq \{(X, Y) \mid h_i(X) \neq Y\}, \forall (X, Y) \sim \mathcal{P} \tag{5b}$$

meaning that $\mathcal{R}(h_i) \geq \mathcal{R}(h_S), \forall i \in \{1, \ldots, |S|\}$, so the ambiguity effect is non-negative:

$$\frac{1}{|S|} \sum (\mathcal{R}(h_i) - \mathcal{R}(h_S)) \geq 0 \tag{6}$$

Practically, this tells us that in order for $R(h_S)$ to decrease as new provers are added to S, either (*i*) the average individual risk must decrease, or (*ii*) the ambiguity effect must increase (or both).

In the proof assistant setting, diversity therefore corresponds to non-overlapping prover failures (or equivalently, the 'uniqueness' of prover successes). The OR aggregator is monotonic and ensures that any disagreement on the provability of a theorem in an ensemble is strictly beneficial. This formalizes the classical ensemble intuition that "two classifiers are diverse if they make different errors on new data points"[11] and frames it as strictly beneficial in the context of neural theorem provers.

4. Experiments

4.1. Experiment Design

First, we empirically verify that an OR-ensemble raises coverage under a fixed proof budget. Per problem we allow exactly 24 proof attempts. A single baseline model may consume all 24; the ensemble shares the same budget across three models:

- AI-MO/Kimina-Prover-Preview-Distill-7B
- Goedel-LM/Goedel-Prover-SFT, and
- deepseek-ai/DeepSeek-Prover-V2-7B

so each model contributes 8 attempts ($8 \times 3 = 24$). Those 8 attempts are drawn uniformly (without replacement) from the 24 independent attempts we previously generated for that model. Source code and data is available here.

4.2. Results

We repeat this sampling 10 000 times (Monte-Carlo) and report the resulting distribution of successes.

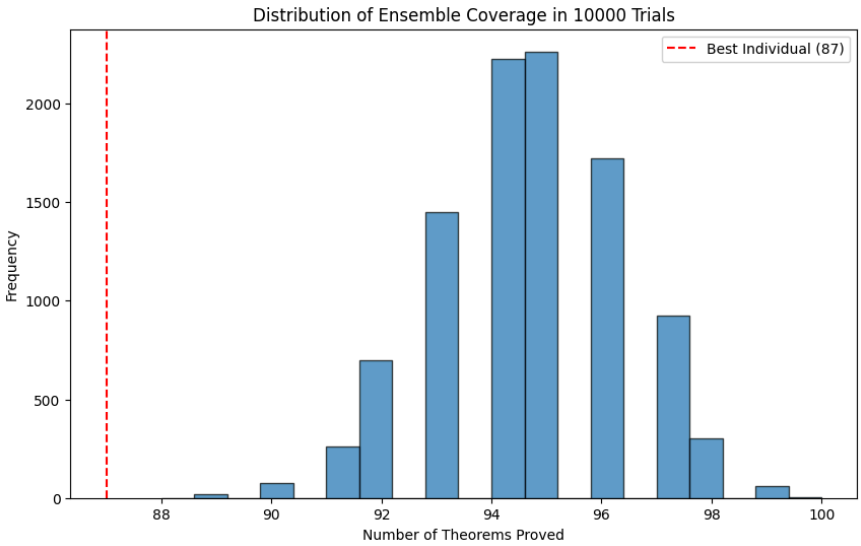

Figure 2: Distribution of number of theorems proved by ensemble and best individual model

We report the individual model performances below:

Model	**Accuracy**
AI-MO/Kimina-Prover-Preview-Distill-7B	72/244
Goedel-LM/Goedel-Prover-SFT	87/244
deepseek-ai/DeepSeek-Prover-V2-7B	83/244

On average, the ensemble proves 94.6 theorems from MiniF2F-test which is an absolute gain of 7.6 theorems over the best individual model (Goedel-LM/Goedel-Prover-SFT), or an improvement of 8.7%. In 10,000 sampling trials, the ensemble always proved strictly more theorems than the best individual model.

4.3. Interpretation

The ensemble theory we developed earlier works at an exact and distribution level. That is, it assumes that we can know for a prover model whether it can or cannot prove a theorem given unlimited attempts. Nonetheless, it's clear from the experiment that real implementations of prover models at the 7B size on a limited pass@24 budget possess sufficiently positive ambiguity effect that it becomes practically significant. This makes sense since different prover models often exhibit fundamental differences in their architecture, training data, and optimization methods We posit that these differences lead to differences in reasoning paths and consequently, non-overlapping failures.

5. Ensemble Lemma Proving

We now turn to the second question: *"Is it possible, via ensembling, to write validated proofs for propositions that no individual model can prove?"*. While the risk decomposition in (4) explains why an ensemble outperforms a single model on average, it does not directly address this more ambitious goal.

To see the limitations of our current framework, assume we have an ensemble of two provers $S = \{h_1, h_2\}$. Next, assume $\exists(x_1, 1) \in \mathcal{P} \mid h_1(x_1) = 1 \wedge h_2(x_1) = 0$ and $\exists(x_2, 1) \in \mathcal{P} \mid h_1(x_2) = 0 \wedge h_2(x_2) = 1$. That is, there exists a provable statement x_1 that only prove $h_1 \in S$ proves and another provable statement $x_2 \in S$ that only h_2 proves. Then there exists another statement that is simply the re-statement that both x_1 and x_2 hold. This statement is provable and the intermediate "lemmas" x_1 and x_2 are provable by an OR-ensemble h_S . However, we lack a meta-prover that can decompose problems into h_S -provable sub-problems. For this particular case, all we need is a

$$h_{\text{AND}}(x, S) = \begin{cases} 1 \text{ if } x = \bigwedge_{i=1}^{n} \varphi_i \text{ and } h_S(\varphi_i) = 1 \\ h_S(x) \text{ o.w.} \end{cases} \tag{7}$$

which will have strictly lower risk than h_S under the original distribution $\mathcal{P}$ when the ambiguity effect in the risk decomposition of h_S is non-zero.

So while an ensemble with a lower risk is more likely to successfully prove a greater number of constituent lemmas, we cannot construct a complete proof that is beyond the reach of any single prover until such a meta-prover is supplied.

Fortunately, a concrete instance of such a meta-prover already exists. In [12] the authors separate strategic reasoning from tactical reasoning: a high-level Planner (a general-purpose LLM) decomposes the theorem into a sequence of lemmas (subgoals) $\varphi_1, \ldots, \varphi_n$, while a step-level Prover model tries to close each lemma. The top-level statement is accepted only if every lemma is proved, i.e. you can conceptualize an h_{Planner} that combines the Planner and Prover as follows:

$$h_{\text{Planner}}(\varphi) = 1 \iff \bigwedge_{i=1}^{n} h_{\text{Prover}}(\varphi_i) = 1 \tag{8}$$

The implication is that assuming the existence of such a Planner, we can substitute h_S for h_{Prover} in (8) leading to

$$h_{\text{Planner}'}(\varphi) = 1 \iff \bigwedge_{i=1}^{n} h_S(\varphi_i) = 1 \tag{9}$$

which can fully take advantage of the non-overlapping failures benefit of OR-ensembles at the subgoal level.

6. Future Work

We return to the 3 questions we originally posed in the introduction:

1. Can we make use of non-overlapping failures (i.e. 'functional diversity') of prover models that do not saturate MiniF2F-Test to improve their collective accuracy?
2. Is it possible, via ensembling, to write validated proofs for propositions that no individual model can prove?
3. Can we apply these techniques to the current SOTA models to achieve stronger performance on as-yet unsaturated benchmarks such as PutnamBench? [7]

We have shown empirically that OR-aggregation under a fair budget for comparison improves collective accuracy. We have also sketched how using a planner can allow us to apply the non-overlapping failures of an existing ensemble at the subgoal level and how this may allow us to make more progress on as-yet unproved theorems from an ensemble. With our current work, we've answered Question 1 and we have argued theoretically that the answer to Question 2 can be 'yes' provided a suitable meta-planner. For a future version, we need to give empirical evidence that supports our ensemble lemma proving framework to answer the empirical aspect of Question 2 before moving on to Question 3.

In the next version, we plan to include empirical evidence that demonstrates that the non-overlapping failures exhibited for the MiniF2F-test can also be observed at the subgoal level for proof skeletons sketched by the planner. This will provide empirical evidence for Question 2.

We also plan to upgrade the 7B models we've used to newer/larger versions. The upgrades planned are as follows:

Model	**Upgraded Model**
AI-MO/Kimina-Prover-Preview-Distill-7B	AI-MO/Kimina-Prover-72B
Goedel-LM/Goedel-Prover-SFT	Goedel-LM/Goedel-Prover-V2-32B
deepseek-ai/DeepSeek-Prover-V2-7B	deepseek-ai/DeepSeek-Prover-V2-671B

We choose the set of upgraded models because even with greater parameter counts, the model architecture, optimization strategy, and training material will be similar. Doing this gives us confidence that our current results are more likely to hold when we scale up. That is, that these upgraded models will still exhibit non-overlapping failures albeit perhaps on different problems. This will allow us to better answer Question 3: whether we can improve

on SOTA using the techniques outlined in this paper. If diversity continues to deliver at scale, ensemble-plus-planner systems could become the cheapest upgrade path for strong neural theorem provers.

References

[1] A. Forootani, A Survey on Mathematical Reasoning and Optimization with Large Language Models, (2025). https://doi.org/10.48550/arXiv.2503.17726.

[2] P. Lu, L. Qiu, W. Yu, S. Welleck, K.-W. Chang, A Survey of Deep Learning for Mathematical Reasoning, (2023). https://doi.org/10.48550/arXiv.2212.10535.

[3] K. Yang, G. Poesia, J. He, W. Li, K. Lauter, S. Chaudhuri, D. Song, Formal Mathematical Reasoning: A New Frontier in AI, (2024). https://doi.org/10.48550/arXiv.2412.16075.

[4] K. Zheng, J.M. Han, S. Polu, MiniF2F: a cross-system benchmark for formal Olympiad-level mathematics, (2022). https://doi.org/10.48550/arXiv.2109.00110.

[5] Y. Zhou, J. Zhao, Y. Zhang, B. Wang, S. Wang, L. Chen, J. Wang, H. Chen, A. Jie, X. Zhang, H. Wang, L. Trung, R. Ye, P.N. Hoang, H. Zhang, P. Sun, H. Li, Solving Formal Math Problems by Decomposition and Iterative Reflection, (2025). https://doi.org/10.48550/arXiv.2507.15225.

[6] L. Chen, J. Gu, L. Huang, W. Huang, Z. Jiang, A. Jie, X. Jin, X. Jin, C. Li, K. Ma, C. Ren, J. Shen, W. Shi, T. Sun, H. Sun, J. Wang, S. Wang, Z. Wang, C. Wei, S. Wei, Y. Wu, Y. Wu, Y. Xia, H. Xin, F. Yang, H. Ying, H. Yuan, Z. Yuan, T. Zhan, C. Zhang, Y. Zhang, G. Zhang, T. Zhao, J. Zhao, Y. Zhou, T.H. Zhu, Seed-Prover: Deep and Broad Reasoning for Automated Theorem Proving, (2025). https://doi.org/10.48550/arXiv.2507.23726.

[7] G. Tsoukalas, J. Lee, J. Jennings, J. Xin, M. Ding, M. Jennings, A. Thakur, S. Chaudhuri, PutnamBench: Evaluating Neural Theorem-Provers on the Putnam Mathematical Competition, (2024). https://doi.org/10.48550/arXiv.2407.11214.

[8] D. Wood, T. Mu, A. Webb, H. Reeve, M. Luján, G. Brown, A Unified Theory of Diversity in Ensemble Learning, (2024). https://doi.org/10.48550/arXiv.2301.03962.

[9] L.I. Kuncheva, C.J. Whitaker, Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy, Machine Learning 51 (2003) 181–207. https://doi.org/10.1023/A:1022859003006.

[10] R. Naik, V. Chandrasekaran, M. Yuksekgonul, H. Palangi, B. Nushi, Diversity of Thought Improves Reasoning Abilities of LLMs, (2024). https://doi.org/10.48550/arXiv.2310.07088.

[11] T.G. Dietterich, Ensemble Methods in Machine Learning, in: Proceedings of the First International Workshop on Multiple Classifier Systems, Springer-Verlag, Berlin, Heidelberg, 2000: pp. 1–15.

[12] R. Xin, Z. Zheng, Y. Nie, K. Yuan, X. Xiao, Scaling up Multi-Turn Off-Policy RL and Multi-Agent Tree Search for LLM Step-Provers, (2025). https://arxiv.org/abs/2509.06493.