Diversity-Driven Generalization in Mathematical Reasoning Ensembles

Nicholas Jiang, Joe Zhou, Rishabh Sharma, Sarvesh Sivakumar

Summer 2025

1 Introduction

1.1 Motivation and Importance

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. Reliable reasoning enables AI models to generate correct and verifiable solutions, significantly benefiting education through scalable personalized tutoring, formal verification of critical software and scientific results, and strengthening AI's general reasoning capabilities. However, current models often suffer from brittle generalization, especially out-of-distribution (OOD), leading to correct outcomes derived through incorrect reasoning paths, thus limiting their practical reliability and adoption.

1.2 Problem Context and Prior Work

Recent advancements, particularly through stepwise or process supervision, have significantly improved in-distribution accuracy of math-solving LLMs, as demonstrated by process-supervised reward models (PRMs) significantly outperforming traditional outcome-supervised models. Nevertheless, these approaches remain costly in terms of annotation effort and still show limitations in handling distribution shifts. Prior ensemble approaches, like majority-vote or confidence-based ensembles, generally focus only on outcome-level diversity (final answers), neglecting the underlying reasoning path diversity crucial for robust generalization in structured tasks like mathematics. Ortega et al.'s foundational work on neural network ensembles established a relationship between ensemble accuracy, diversity, and error, but their approach lacks specificity in how to measure or induce meaningful diversity for symbolic-reasoning tasks. The Task2Vec framework, which uses Fisher-information geometry to embed tasks into a continuous space based on their gradient information, has successfully predicted transferability and model performance across different tasks. However, it has not yet been tailored for evaluating or inducing reasoning-level diversity within mathematical reasoning ensembles, highlighting a critical gap and opportunity.

1.3 Limitations of Existing Solutions

Current solutions fall short primarily due to:

- High cost and diminishing returns of detailed step-level supervision
- Insufficient reasoning-path diversity consideration in existing ensemble approaches
- Lack of measurable and inducible diversity metrics suited specifically to reasoning tasks

1.4 Specific Research Problem

The specific problem addressed by this research is the following:

> ”How can we automatically induce and measure reasoning-level diversity in math-solving LLM ensembles to achieve robust generalization, while minimizing additional annotation costs?”

2 Proposed Solution

Below is our detailed plan to *measure*, *induce*, and *exploit* reasoning-path diversity so that mathematical-reasoning LLM ensembles generalize robustly with *zero extra human annotation.*

2.1 Overview of the Workflow

Our approach unfolds in three tightly integrated phases that directly address the challenges identified in Section 1. In Phase 1, we construct diverse ensembles of LLM-based theorem provers that vary by seed and architecture. Then we introduce the *Ensemble Diversity Coefficient* (EDC), a quantitative metric based on Task2Vec embedding distances, to capture reasoning-path diversity. Phase 2 leverages that diversity through a self-supervised peer-learning protocol: whenever exactly one solver in an ensemble successfully proves a training problem, its verified proof is used to fine-tune the other members, converting diversity into a free, annotation-free learning signal. Finally, Phase 3 employs mixed-effects regression to test whether ensembles with higher EDC achieve greater out-of-distribution gains, providing statistical validation that reasoning-path diversity causally drives robust generalization. Together, these phases establish a measurable, low-cost pipeline from diversity quantification to practical performance improvements.

2.2 Phase 1: Ensemble Construction & Diversity Quantification

1. **Dataset & Splits.** Use the miniF2F benchmark with $K = 20$ independent 50/50 random splits (Train A, Test B).
2. **Solver Pool.** Assemble a diverse pool of LLM provers (varying seeds, architectures, and fine-tuning recipes).
3. **Sampling Ensembles.** Draw N equal-sized ensembles from the pool
4. **Diversity Metric (EDC).** For ensemble M, define the *Ensemble Diversity Coefficient*:

$$\mathrm{EDC}(M) = \frac{1}{\binom{|M|}{2}} \sum_{i<j} \mathrm{cmdiv}\big(m_i, m_j; D\big),$$

   where cmdiv is the mean Task2Vec embedding distance on sampled data batches.

2.3 Phase 2: Self-Supervised Peer Learning Protocol

For each ensemble i and split s:

1. **Baseline Measurement.** Record accuracy $S_{i,0,s}$ of each solver and the ensemble on Test B.
2. **Peer-Proof Fine-Tuning.** For every Train A problem that *exactly one* solver solves, fine-tune the other solvers on that verified proof.
3. **Post-Training Evaluation.** Re-evaluate on Test B to get $S_{i,1,s}$ and compute the gain $\Delta S_{i,s} = S_{i,1,s} - S_{i,0,s}$.

2.4 Phase 3: Evaluation & Statistical Validation

We fit the mixed-effects regression:

$$\Delta S_{i,s} = \beta_0 + \beta_1\,\mathrm{EDC}_i + \beta_2\,S_{i,0,s} + u_i + v_s + \varepsilon_{i,s},$$

where u_i and v_s are solver- and split-level random effects. A significant positive β_1 confirms that higher reasoning-path diversity correlates with better generalization.

2.5 Deliverables

- Open-source implementation of EDC (PyTorch/Task2Vec).
- Reproducible training scripts for peer-proof fine-tuning.
- Plots and analyses of diversity–performance curves over 20 splits.
- Ablation studies (e.g., seed vs. architecture diversity).

3 Evaluation and Implementation Plan

3.1 Evaluation Plan

We will assess success by measuring both quantitative performance improvements and statistical validation of our core hypothesis. First, we will compare out-of-distribution accuracy gains (ΔS) across ensembles with varying EDC, aiming for a significant positive correlation in our mixed-effects regression (i.e. a strictly positive β_1 at $p < 0.05$). Second, we will conduct ablations—disabling peer-proof updates or sampling low-diversity ensembles—to verify that any generalization gains vanish without the proposed diversity mechanisms. Finally, we'll produce precision-recall curves and diversity–performance scatter plots over all 20 splits to confirm consistent benefits and rule out split-specific artifacts.

3.2 Timeline

Our eight-week schedule is divided into four two-week sprints to ensure rapid iteration and clear milestones.
Weeks 1–2: Implement and validate the Ensemble Diversity Coefficient (EDC) computation, assemble the diverse solver pool, and write unit tests to guarantee reproducibility.
Weeks 3–4: Develop and integrate the self-supervised peer-proof fine-tuning protocol, build automated training pipelines, and run initial small-scale experiments to verify end-to-end functionality.
Weeks 5–6: Conduct full-scale experiments across all 20 data splits, collect accuracy and diversity metrics, and begin mixed-effects regression analysis to assess the impact of EDC.
Weeks 7–8: Perform ablation studies (e.g., removing peer-proof updates or low-diversity ensembles), refine statistical analyses, generate all plots and figures, and complete the draft manuscript alongside the open-source code release.

4 References

1. H. Lightman et al., "Let's Verify Step by Step," May 31, 2023, arXiv: arXiv:2305.20050. doi: 10.48550/arXiv.2305.20050.

2. A. Achille et al., "Task2Vec: Task Embedding for Meta-Learning," Feb. 10, 2019, arXiv: arXiv:1902.03545. doi: 10.48550/arXiv.1902.03545.

3. L. A. Ortega, R. Cabañas, and A. R. Masegosa, "Diversity and Generalization in Neural Network Ensembles," Feb. 16, 2022, arXiv: arXiv:2110.13786. doi: 10.48550/arXiv.2110.13786.