Abstract：

Evaluating the advanced mathematical reasoning of Large Language Models (LLMs) is a critical yet unsolved challenge, essential for progress towards Artificial General Intelligence. However, existing benchmarks are confined to simple, answer-based problems, while scalable evaluation of proof-based reasoning—the cornerstone of advanced mathematics—remains an open problem due to the high cost of human expertise and formal methods. To address this gap, we propose **Proof2Choice**, a novel, fully-automated, robust, and universally applicable methodology that uses a multi-LLM pipeline to convert complex proof-based statements into high-quality, easy-to-evaluate multiple-choice questions. Implementing this method on the algebraic geometry reference "The Stacks project," we create **AProofBench**, the first large-scale benchmark of its kind, featuring 456 challenging questions. The benchmark's rigor is underscored by our findings: even state-of-the-art models achieve a modest accuracy of only 20-60%, and human audits confirm that over 95% of its questions are of high quality, with mathematically flawed yet deceptive distractors. Our work thus provides two key contributions: a rigorously validated benchmark that uncovers fundamental deficits in current LLMs, and a scalable, domain-agnostic methodology poised to advance reasoning evaluation across numerous scientific fields. Code is available at this Repository: https: //github.com/xxx/xxx.