The pursuit of artificial general intelligence (AGI) hinges on developing systems capable of complex, abstract reasoning. Among all human intellectual endeavors, advanced mathematics stands as a paramount benchmark for this capability. While Large Language Models (LLMs) have achieved remarkable success on a wide array of tasks, their ability to comprehend, manipulate, and generate rigorous mathematical proofs—the very fabric of frontier scientific discovery—remains a largely uncharted and critical frontier. A fundamental obstacle impeding progress in this area is the lack of appropriate evaluation tools. Current mathematics benchmarks almost exclusively feature problems with definitive, easily verifiable answers, which are typically confined to elementary or competition-level mathematics. This leaves the core challenge of proof-based reasoning, which dominates advanced disciplines, largely unaddressed.

Evaluating mathematical proofs at scale presents a formidable challenge. The two prevailing methods are both critically flawed in terms of scalability and accessibility. Manual evaluation by human mathematics experts, while being the gold standard for quality, is prohibitively expensive, time-consuming, and inherently difficult to scale. On the other hand, leveraging formal verification systems like Lean or Isabelle requires a tremendous upfront investment of human effort to formalize both the mathematical problems and the generated proofs into a machine-readable language. This high barrier to entry has consequently limited the scope of formalization to relatively simple mathematical theorems. Thus, a scalable, automated, and accessible benchmark for evaluating LLMs on natural language mathematical proofs is urgently needed, yet conspicuously absent from the current research landscape.

To bridge this critical gap, we propose Proof2Choice, a novel and fully automated methodology for transforming hard-to-evaluate proof-based questions into easy-to-evaluate multiple-choice questions. The core mechanism of Proof2Choice is an orchestrated pipeline of powerful LLMs working in concert. This multi-LLM system intelligently generates a set of candidate distractors by strategically modifying key elements—such as keywords, conditions, and formulas—within the original mathematical definitions or propositions. Subsequently, the pipeline employs another set of LLM agents as judges to filter and refine these candidates, ensuring that the final distractors are not only mathematically incorrect but also deceptively plausible. This carefully engineered generation and verification process minimizes potential evaluation bias and guarantees the exceptional quality of the resulting questions.

Implementing Proof2Choice on the definitions and propositions of the open source Algebraic Geometry textbook and reference work “The Stacks project”, we produced AProofBench, the first large scale benchmark focusing on natural language math proof questions. It comprises [TODO] multiple-choice questions, each presenting six options: exactly two correct options and four carefully crafted distractors. Our evaluation of AProofBench on multiple open-source and closed-source LLMs reveals that even the best-performing model to date achieves only a moderate score of 60, with scores commonly as low as 20. This result vividly demonstrates the rigor and challenge of our benchmark, as well as its immense potential to advance LLM understanding of frontier mathematics.

To ensure the solidity of Proof2Choice, we commission mathematics experts to manually audit every question in AProofBench and the corresponding responses generated by leading LLMs, including o3 and Gemini-2.5-Pro. The result indicates that more than 98.75% of the LLM-generated distractors are both mathematically wrong and deceptively plausible—challenging enough that even state-of-the-art models still overlook a non-trivial fraction of them, while 95 % of the questions in AProof Bench meet the same standard, with every distractor satisfying the same rigorous criteria. On the other hand, the principal reason for the LLM’s erroneous response lies in either its failure to detect the flaw embedded within an incorrect choice or its hallucinatory conviction that an correct choice is mathematically inconsistent. Both reasons stem from the model’s intrinsic limitations rather than from any deficiency in the quality of the questions themselves.

To address the aforementioned gaps, the key contributions of our work are summarized belowrized as follows:

* **A Novel Methodology for Automated Proof-based Question Generation.** We introduce Proof2Choice, the first fully-automated, LLM-driven methodology that converts complex proof-based statements into high-quality multiple-choice questions. A principal advantage of this workflow is its universal applicability: it is domain-agnostic and can automatically generate challenging evaluation questions from seed examples of any difficulty. This drastically reduces the cost and human effort required for creating assessments, not only in mathematics but in any domain reliant on rigorous, proof-like reasoning such as law or the formal sciences.
* **The First Large-Scale Benchmark for Frontier Mathematical Reasoning.** We present **AProofBench**, the first large-scale benchmark featuring natural language proof questions from the advanced field of algebraic geometry. Validated by human experts, AProofBench serves as a rigorous and reliable tool for assessing, understanding, and ultimately advancing the mathematical reasoning capabilities of future language models.
* **An in-depth analysis of LLM failure modes in mathematical reasoning.** Beyond assessing performance, we provide the first systematic classification and quantitative analysis of how LLMs fail when handling advanced mathematical problems. By categorizing and counting a large corpus of errors, our work offers key insights into the deep reasoning deficits of current models, informing future efforts to build more robust and trustworthy AI.

Mathematical capability stands as one of the most pivotal yardsticks for Large Language Models (LLMs). Rooted in a discipline that epitomizes the distilled rigor of human intellect, mathematics provides a unique lens through which to assess a model's capacity for abstract reasoning, logical deduction, and structured thought—skills that are prerequisites for more general artificial intelligence. Consequently, a model's performance on advanced mathematical tasks serves as a crucial indicator of its true cognitive depth, distinguishing superficial pattern matching from genuine comprehension.

(中文: 数学能力是衡量大型语言模型（LLM）最关键的指标之一。)

Ensuring that LLMs develop this profound capability requires robust and targeted evaluation tools. However, the existing suite of mathematical benchmarks is systematically incapable of serving this purpose for most branches of modern mathematics. Contemporary mainstays, such as GSM8K and MATH, are confined to elementary and competition-level curricula, where model performance is already reaching saturation. They therefore provide a weak signal concerning a model's grasp of advanced concepts. At the other extreme, benchmarks like HLE-Math engage with frontier content but are designed as intractable stress tests rather than diagnostic tools. Their formidable difficulty makes them unsuitable for the fine-grained monitoring needed to guide an LLM's development. This leaves researchers with an untenable choice between elementary tasks and overwhelming challenges, highlighting a critical absence of domain-specific benchmarks pitched at an appropriate, moderate level of difficulty for frontier mathematics.

(中文: 确保LLM发展出这种深刻的能力，需要稳健且有针对性的评估工具。)

The bottleneck lies in the current methods of benchmark creation, which are all fundamentally limited in either scope or scalability. **Corpus-based harvesting** is effective only for elementary mathematics where "question-plus-answer" pairs are abundant; frontier mathematics, articulated through proofs and definitions, is intrinsically devoid of such easily extractable artifacts. **Expert authoring** remedies the content deficit but at a prohibitive cost, as each problem requires a labor-intensive process of conception, review, and validation that is inherently resistant to scaling. Lastly, **formal languages** like Lean promise machine-checkable rigor, but their ecosystems are still nascent, meaning any benchmark synthesis in this paradigm still hinges on extensive and costly human annotation. These limitations—an evaluative bottleneck on the consumer side and a creation bottleneck on the producer side—make it nearly impossible to generate the necessary benchmarks with existing tools.

(中文: 瓶颈在于当前的基准创建方法，这些方法在范围或可扩展性上都存在根本性的限制。)

Therefore, there is an urgent and unmet need for a scalable, cost-effective, and fully automated methodology capable of generating high-quality benchmarks across diverse branches of advanced mathematics. Such a method would need to overcome the expressive mismatch of natural corpora and sidestep the prohibitive costs of expert and formal intervention, providing a new path forward for assessing and improving the next generation of LLMs.

(中文: 因此，我们迫切需要一种可扩展、成本效益高且完全自动化的方法，能够在数学的各个分支中生成高质量的基准。)

**Our Solution and Contributions**

To address this critical gap, we propose **Proof2Choice**, the first scalable, fully automated, and domain-agnostic framework for synthesizing mathematical benchmarks from proof-based statements. The central idea of Proof2Choice is to transform hard-to-evaluate proofs into easily scorable multiple-choice questions. At its heart is a carefully orchestrated pipeline of powerful LLMs: a "generator" team crafts deceptive distractors by strategically altering keywords, conditions, or formulas in original mathematical statements, after which a "judge" team filters these candidates to ensure both mathematical invalidity and high plausibility.

(中文: 为了弥合这一关键差距，我们提出了Proof2Choice，这是首个可扩展、全自动、领域无关的，用于从基于证明的陈述中合成数学基准的框架。)

We instantiate Proof2Choice on the "The Stacks project," a comprehensive graduate-level textbook on algebraic geometry, to create **AProofBench**. This benchmark contains 456 challenging multiple-choice questions, each with two correct options and four carefully engineered distractors. AProofBench proves to be a rigorous diagnostic tool: our evaluations show that even state-of-the-art LLMs achieve a low accuracy of only 20-60%. Furthermore, a comprehensive audit by expert mathematicians confirms the benchmark's quality, with over 95% of its questions meeting a high standard of clarity and rigor. The audit also reveals that model failures stem primarily from an inability to verify foundational knowledge, not from flaws in the questions themselves.

Our work makes the following key contributions:

* We propose **Proof2Choice**, the first fully automated and domain-agnostic framework for creating proof-based evaluation benchmarks, breaking the scalability and cost bottlenecks of previous methods.
* We release **AProofBench**, the first large-scale, expert-validated benchmark for frontier mathematics that is specifically designed for diagnostic purposes.
* Through a detailed failure analysis on AProofBench, we provide critical insights into the fundamental deficits of current LLMs, demonstrating that a lack of robust knowledge verification is a primary obstacle to advanced mathematical reasoning.