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 strong 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 process of generation and verification minimizes evaluation bias and ensures the high quality of the resulting questions.

By implementing our Proof2Choice methodology, we introduce **AProofBench**, the first large-scale benchmark designed specifically for natural language proof comprehension in a frontier mathematical domain. We built AProofBench using content from "The Stacks project," a comprehensive and open-source reference work on algebraic geometry. The benchmark comprises **[在此处填写问题数量，例如：500]** challenging multiple-choice questions, where each question presents a proposition or definition and offers six options: two are correct restatements or direct consequences, while the other four are carefully crafted distractors. To validate the integrity of our benchmark, we commissioned mathematics experts to conduct a thorough audit. The results confirmed the exceptional quality of our method: over 98.75% of the LLM-generated distractors were deemed both mathematically flawed and highly plausible, with 95% of the questions in AProofBench meeting this rigorous standard.

Our experiments on AProofBench reveal the profound challenge it poses to even the most advanced LLMs. The current state-of-the-art model, such as Gemini 2.5 Pro achieves a modest accuracy of only ~60%, with other models scoring as low as 20%, indicating a significant gap in their deep reasoning capabilities. A detailed failure analysis, corroborated by our expert audit, shows that these errors stem from the models' intrinsic limitations—either failing to identify subtle flaws in incorrect choices or "hallucinating" inconsistencies in correct ones—rather than any ambiguity in the questions themselves. Our contributions are summarized as follows:

* **A Novel Methodology for Automated Proof-based Question Generation.** We introduce **Proof2Choice**, the first fully-automated, LLM-driven methodology that converts complex mathematical proofs into high-quality multiple-choice questions. This scalable workflow drastically reduces the cost and human effort required for evaluating advanced mathematical reasoning and is generalizable to other proof-centric domains.
* **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 Formal Reasoning.** We provide a comprehensive analysis of how and why state-of-the-art LLMs fail at complex, proof-based tasks. Our findings offer crucial insights into the current limitations of LLMs, providing a clear direction for future research aimed at developing more robust and trustworthy AI systems.

对通用人工智能（AGI）的追求，关键在于开发出能够进行复杂、抽象推理的系统。在所有人类智力活动中，高等数学是检验这种能力的最高基准。尽管大型语言模型（LLM）已在众多任务中取得了卓越的成功，但它们理解、运用和生成严谨数学证明的能力——这正是前沿科学发现的精髓所在——仍然是一个很大程度上未被探索的关键前沿。阻碍该领域取得进展的一个根本性障碍，是缺乏合适的评估工具。当前的数学基准几乎完全由那些答案确定、易于验证的问题构成，这些问题通常局限于初等数学或竞赛数学。这使得以证明为核心的推理挑战在很大程度上未被触及，而这正是高等学科的主导。

大规模地评估数学证明提出了一项艰巨的挑战。现有的两种主流方法在可扩展性和可及性方面都存在严重缺陷。由人类数学专家进行的人工评估，虽然是质量的黄金标准，但其成本高得令人望而却步，耗时且本质上难以规模化。另一方面，利用像 Lean 或 Isabelle 这样的形式化验证系统，则需要巨大的人力投入来将数学问题和生成的证明转化为机器可读的语言。这种高门槛因此将形式化的范围限制在了相对简单的数学定理上。因此，学界迫切需要一个可扩展、自动化且易于使用的，用于评估 LLM 在**自然语言**数学证明方面表现的基准，但这在当前的研究版图中却明显缺失。

为了弥合这一关键差距，我们提出了 **Proof2Choice**，一种新颖的、完全自动化的方法论，可将难以评估的证明题转化为易于评估的多项选择题。Proof2Choice 的核心机制是一个由多个强大的 LLM 协同工作的精心编排的流水线。这个多 LLM 系统通过策略性地修改原始数学定义或命题中的关键元素——例如关键词、条件和公式——来智能地生成一组候选干扰项。随后，该流水线利用另一组 LLM 代理作为评判者来筛选和优化这些候选选项，确保最终的干扰项不仅在数学上是错误的，而且具有以假乱真的迷惑性。这个精心设计的生成与验证流程最大限度地减少了评估偏差，并保证了最终问题的超高质量。

我们在开源的代数几何教科书与参考文献《The Stacks project》的定义和命题上实施了 **Proof2Choice**，从而生成了 **AProofBench**——这是首个专注于自然语言数学证明题的大规模基准。该基准包含了 **[TODO]** 道选择题，每道题提供六个选项：两个完全正确的选项和四个精心制作的干扰项。我们在多个开源及闭源的大语言模型上对 AProofBench 进行了评估，结果显示，即便是迄今为止数学性能测试最好的模型也仅获得了 60 分的中等分数，普遍低分在 20 分左右。这一结果鲜明地揭示了我们基准的严格性与挑战性，及其在推动 LLM 理解前沿数学方面所具备的巨大潜力。

为确保 Proof2Choice 的稳健性，我们委托了数学专家来人工审核 AProofBench 中的每一道题，以及包括 o3 和 Gemini-2.5-Pro 在内的顶尖 LLM 生成的相应回答。结果有力地证实了我们方法的有效性：超过 **98.75%** 由 LLM 生成的干扰项被认定为“在数学上错误，且具有高度迷惑性”，其难度足以让最先进的模型也频繁出错；同时，AProofBench 中 **95%** 的问题本身也达到了同样严格的标准。另一方面，我们发现 LLM 回答错误的主要原因并非源于题目质量的任何缺陷，而是根植于模型内在的推理局限：它们要么未能识别出错误选项中精心嵌入的逻辑瑕疵，要么“幻觉”出正确选项在数学上存在不一致。

为了弥补上面已知的差距，我们的贡献总结如下：

* **一种用于自动化生成证明题的新颖方法论。** 我们介绍了 **Proof2Choice**，这是首个完全自动化、由 LLM 驱动的方法论，可将复杂的数学证明转化为高质量的多项选择题。这个可扩展的工作流极大地降低了评估高等数学推理所需的成本和人力，并且可以推广到其他以证明为核心的领域。
* **首个面向前沿数学推理的大规模基准。** 我们推出了 **AProofBench**，这是第一个以高等代数几何领域的自然语言证明题为特色的大规模基准。经人类专家验证，AProofBench 为评估、理解并最终推动未来语言模型的数学推理能力提供了一个严谨而可靠的工具。
* **对 LLM 在数学推理中失败模式的深入分析。** 我们不仅评估了模型的最终得分，更通过对大量错误案例的归纳与统计，首次对大语言模型在处理高等数学问题时的失败模式进行了系统性分类与定量分析，为理解并修复当前模型的深层推理缺陷提供了关键洞见。

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.

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, our contributions are summarized as follows:

* **A Novel Methodology for Automated Proof-based Question Generation.** We introduce **Proof2Choice**, the first fully-automated, LLM-driven methodology that converts complex mathematical proofs into high-quality multiple-choice questions. This scalable workflow drastically reduces the cost and human effort required for evaluating advanced mathematical reasoning and is generalizable to other proof-centric domains.
* **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.

Abstract的原来的版本：

Current mathematics benchmarks largely overlook the critical area of proof-based problems, focusing instead on questions with definite answers, primarily because natural language proofs are difficult to evaluate automatically. To address this gap, we introduce Proof2Choice, an automatic, universally applicable, and robust methodology based on Large Language Models (LLMs) that transforms these hard-to-evaluate proof questions into an easy-to-assess multiple-choice format. Utilizing this approach, we developed AProofBench, the first large-scale benchmark dedicated to natural language math proofs, which is based on the open-source Algebraic Geometry textbook, “The Stacks project,” and consists of 456 six-choice questions. Evaluations on both open-source and closed-source LLMs show that even state-of-the-art models achieve only moderate scores of 20-60%, highlighting the benchmark's rigor and its capacity to advance LLM reasoning in cutting-edge mathematics. The outstanding quality of AProofBench is substantiated by human audits, which reveal that over 98.75% of its distractors are both mathematically incorrect and deceptive, and more than 95% of its questions are of high quality. This not only furnishes compelling evidence for the robustness of the Proof2Choice methodology but also demonstrates its unique capability to probe the knowledge embedded in LLMs, making it an essential tool for improving model performance on the most challenging reasoning tasks.

当前，多数数学基准测试侧重于具有唯一确定答案的计算题，而忽略了在数学研究中更为核心的证明题，其主要原因在于自然语言形式的证明难以进行自动化评估。为填补这一空白，我们提出了一种名为 Proof2Choice 的创新方法论。这是一种基于大语言模型（LLM）的自动化、通用且稳健的框架，能够将难以评估的数学证明题，高效地转换为易于评估的标准化多项选择题。利用 Proof2Choice 方法，我们基于开源代数几何教材“The Stacks project”创建了 AProofBench——这是首个专注于自然语言数学证明的大规模基准测试，共包含 456 道六选一的题目。在对多种前沿的开源及闭源 LLM 的评测中，模型的得分集中在 20% 至 60% 之间，这充分证明了该基准测试的挑战性及其推动前沿数学研究的潜力。更为重要的是，人工审核结果显示，AProofBench 中超过 98.75% 的干扰项在数学上是错误的同时具备高度迷惑性，且整体超过 95% 的题目达到了高质量标准，这为 Proof2Choice 方法的稳健性提供了强有力的证据。AProofBench 所展现出的独特能力也表明，Proof2Choice 框架非常适用于深度探测并揭示 LLM 在任何特定知识领域中的掌握程度，这对于提升模型在最具挑战性的推理任务上的表现至关重要。

Abstract后面我进行修改的版本：

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.

评估大型语言模型（LLM）的高级数学推理能力，是迈向通用人工智能（AGI）进程中一个关键但尚未解决的挑战。然而，现有的基准测试仅限于简单的、基于答案的问题，而对作为高级数学基石的、基于证明的推理能力进行可扩展的评估，由于高昂的人类专家成本和形式化方法的限制，至今仍是一个悬而未决的难题。为解决这一空白，我们提出了一种名为 Proof2Choice 的全新方法论。它是一种新颖的、全自动、稳健且普遍适用的方法，利用一个多LLM协作的流水线，将复杂的基于证明的命题转化为高质量且易于评估的多项选择题。

我们将此方法应用于代数几何参考文献《The Stacks project》，创建了同类中首个大规模基准测试 AProofBench，其中包含456个具有挑战性的问题。该基准测试的严谨性在我们的研究结果中得到了证实：即使是最先进的模型，其准确率也仅在20-60%之间；同时，人工审核确认超过95%的题目均为高质量，其干扰项在数学上存在谬误但又极具迷惑性。因此，我们的工作做出了两项关键贡献：一是一个经过严格验证的基准测试，它揭示了当前LLM存在的根本性缺陷；二是一种可扩展的、与领域无关的方法论，有望推动众多科学领域的推理评估工作向前发展。

7.21

我的写法：

**Evaluating the capabilities of Large Language Models (LLMs) on frontier scientific tasks, such as advanced mathematics, is a crucial step towards understanding their true reasoning abilities.** While LLMs have achieved remarkable success on a wide array of tasks, their ability to comprehend rigorous mathematical proofs—the very fabric of frontier scientific discovery—remains a critical and largely untested frontier. A fundamental obstacle 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.

**Using this methodology, we introduce AProofBench, the first large-scale benchmark for natural language proof comprehension, which reveals profound limitations in current state-of-the-art LLMs.** Created from the open-source textbook “The Stacks project,” it comprises [TODO] multiple-choice questions, each with two correct options and four carefully crafted distractors. Our evaluation on multiple open-source and closed-source LLMs shows that even the best-performing models achieve only a moderate accuracy of 20-60%, vividly demonstrating the benchmark's rigor and challenge.

**The quality of AProofBench and the robustness of our methodology are confirmed through a comprehensive manual audit by mathematics experts.** The results indicate that over 95% of the questions meet a high standard, with their distractors being both mathematically wrong and deceptively plausible. Furthermore, our analysis on the audit results reveals that LLM errors stem from intrinsic model limitations—either failing to detect flaws in incorrect choices or hallucinating inconsistencies in correct ones—rather than 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 mathematical proofs into high-quality multiple-choice questions. This scalable workflow drastically reduces the cost and human effort required for evaluating advanced mathematical reasoning and is generalizable to other proof-centric domains.
* **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.

评估大型语言模型（LLM）在前沿科学任务（如高等数学）上的能力，是理解其真实推理能力的关键一步。尽管LLM已在众多任务中取得了显著成功，但它们理解严谨数学证明的能力——这正是前沿科学发现的基石——仍然是一个至关重要且基本上未经检验的领域。一个根本性的障碍是缺乏合适的评估工具。当前的数学基准几乎完全采用那些具有确定性、易于验证答案的题目，而这些题目通常局限于初等数学或竞赛级数学。这使得在高等学科中占主导地位的、基于证明的核心推理挑战，在很大程度上未被触及。

大规模评估数学证明是一项艰巨的挑战。目前两种主流方法在可扩展性和可及性方面都存在严重缺陷。由人类数学专家进行手动评估，虽然是质量上的黄金标准，但成本过高、耗时巨大，且本质上难以规模化。另一方面，利用像Lean或Isabelle这样的形式化验证系统，则需要投入巨大的人力前期成本，将数学问题和生成的证明都形式化为机器可读的语言。这种高准入门槛因此将形式化验证的范围限制在了相对简单的数学定理上。因此，一个可扩展、自动化且易于使用的，用于评估LLM在自然语言数学证明方面能力的基准，是当前研究领域迫切需要却又明显缺失的。

为了弥合这一关键差距，我们提出了Proof2Choice，这是一种新颖且完全自动化的方法，能将难以评估的证明类问题转化为易于评估的选择题。Proof2Choice的核心机制是一个由多个强大的LLM协同工作的精心编排的流水线。这个多LLM系统通过策略性地修改原始数学定义或命题中的关键元素——如关键词、条件和公式——来智能地生成一组候选干扰项。随后，该流水线利用另一组LLM代理作为评判员，对这些候选选项进行筛选和提炼，以确保最终的干扰项不仅在数学上是错误的，而且还具有很高的迷惑性。这种精心设计的生成和验证过程最大限度地减少了潜在的评估偏差，并保证了最终生成问题的卓越质量。

利用这种方法，我们推出了AProofBench，这是首个针对自然语言证明理解的大规模基准，它揭示了当前最先进LLM的深层局限性。该基准源自开源教科书“The Stacks project”，包含了[TODO]道选择题，每道题有两个正确选项和四个精心设计的干扰项。我们对多个开源和闭源LLM的评估显示，即使是表现最好的模型也仅取得了20-60%的中等准确率，这生动地展示了该基准的严谨性与挑战性。

AProofBench的质量和我们方法的稳健性，通过数学专家进行的全面人工审计得到了证实。结果表明，超过95%的问题都达到了高标准，其干扰项既是数学上错误的，又具有迷惑性。此外，我们对审计结果的分析揭示，LLM的错误源于模型固有的局限性——要么是未能检测出错误选项中的缺陷，要么是在正确选项中幻觉出不一致性——而非问题本身质量的缺陷。

为解决上述差距，我们的工作主要贡献总结如下：

* **一种用于自动生成证明类问题的新方法。** 我们引入了Proof2Choice，这是首个完全自动化、由LLM驱动的方法，可将复杂的数学证明转化为高质量的选择题。这种可扩展的工作流极大地降低了评估高等数学推理所需的成本和人力投入，并可推广到其他以证明为核心的领域。
* **首个用于前沿数学推理的大规模基准。** 我们提出了AProofBench，这是首个以代数几何这一高等领域的自然语言证明题为特色的大规模基准。经人类专家验证，AProofBench可作为一个严谨可靠的工具，用于评估、理解并最终推动未来语言模型的数学推理能力。
* **对LLM在数学推理中失败模式的深入分析。** 除了评估性能，我们还首次对LLM在处理高等数学问题时的失败方式进行了系统性的分类和量化分析。通过对大量错误语料进行分类和统计，我们的工作为洞察当前模型深层推理能力的不足提供了关键见解，为未来构建更稳健、更可信的AI提供了方向。

Lzx的写法：

Background and Motivation

The ability to solve mathematical proof problems represents a critical frontier for large language models (LLMs). Unlike elementary mathematical tasks with definite answers, proof problems demand structured logical reasoning, multi-step argumentation, and the ability to connect abstract concepts — skills traditionally associated with advanced human mathematical ability. From that aspect, models excel at solving math proof questions are much more likely to possess human-like thinking skills beyond surface-level pattern recognition, a property that is attached great significance in developing and improving LLMs. Therefore, a scalable and high-quality benchmark specifically designed for evaluating the performance of current models on the basis of mathematical proofs is essential for systematically assessing their reasoning capabilities and driving progress towards more human-like, logically proficient language models.

Currently, there are two prevailing approaches to construct benchmarks for evaluation based on mathematical proofs. The first — and most direct — method relies on manual evaluation by expert mathematicians. Although this remains the gold standard for benchmark quality, it demands substantial human effort and is inherently difficult to scale. The second approach involves using formal proof languages, such as Lean, to encode both the mathematical problems and the proofs into a machine-verifiable language. However, this method faces similar limitations: each problem requires meticulously crafted formalization, making the process costly, labor-intensive, and impractical for building scalable benchmarks. Consequently, both methods face substantial barriers to produce large-scale, high-quality benchmarks.

To address the above challenges, we propose a clear design objective: to develop a high-quality benchmark that evaluates LLMs on advanced mathematical proof questions, with the aim of rigorously assessing their reasoning capabilities. Equally importantly, the benchmark must be fully scalable — its creation should require minimal human effort and incur low production costs, ensuring practical applicability for large-scale evaluation.

Our solution

To achieve our design goal, we propose Proof2Choice, the first fully automated, LLM-based framework adapting hard-to-evaluate math proof questions into easy-to-evaluate choice questions. The core mechanism of Proof2Choice is an orchestrated pipeline of powerful LLMs working in harmony, generating a set of candidate distractors by strategically modifying key elements—such as keywords, conditions, and formulae—within the original mathematical definitions or propositions. Subsequently, the pipeline employs another set of powerful LLMs as judges to filter and refine these candidates, ensuring that the final distractors are not only mathematically incorrect but also deceptively plausible. Finally, we implement Proof2Choice on the definitions and propositions of the open source Algebraic Geometry textbook and reference work “The Stacks project”, and thus produce \textbf{AProofBench}, \textbf{the first large scale benchmark focusing on math proof questions}. It comprises 456 multiple-choice questions, each presenting six options: exactly two correct options and four carefully crafted distractors.

Our framework and benchmark have the following key novelties:

1. We propose Proof2Choice, by far the first LLM-based, fully automated framework transforming math proof questions into multi-choice questions, which widens the traditional math-question-answering paradigm—extending from merely definite-answer problems to the richer, proof-centric area of frontier mathematics. In contrast to human-oriented workflow for creating benchmarks based on mathematical proofs, Proof2Choice generates choice questions automatically, thus drastically reducing the human effort involved in checking, filtering and formalizing math proof questions.
2. We highlight that Proof2Choice is a domain-agnostic framework, which innovatively transforms math proof questions into multi-choice questions. This workflow can be readily adapted to any domain where proof-like reasoning — such as logical deduction, formal argumentation, or structured problem solving — is more critical than surface-level answer generation.
3. We propose AProofBench, the first large scale benchmark focusing on math proof questions. Prior to our work, math proof benchmarks were manually made by expert mathematicians, and their size was greatly confined by the number of available high-quality, competion-level seed questions and the amount of available human resources. Our benchmark, on the contrary, is fully scalable thanks to its LLM-based, automatic generation nature and a much wider data source originating from open source, college-level reference work.
4. We highlight that AProofBench is a rigorous and challenging benchmark with great potential in boosting the mathematical reasoning ability of current models. Unlike many popular math benchmarks, for example, MATH and MMLU, where LLMs have reached performance saturation, 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 around 20. This result vividly demonstrates the rigorous and challenging nature of our benchmark, as well as its immense potential to advance LLM understanding of frontier mathematics.
5. We further highlight that AProofBench evaluates not only LLMs’ reasoning prowess but also their factual knowledge, uncovering the subtle, intentionally incorrect modifications hidden in the distractors demands extensive domain familiarity. When we are trying to push an LLM to excel on the hardest reasoning benchmarks—such as HLE-Math or FrontierMath — we first need assurance that the model has already mastered the requisite background knowledge during pretrain and SFT periods. Proof2Choice offers a direct solution: seed it with questions whose prerequisites mirror those of the target benchmarks — which are often straightforward to source — the resulting benchmark then becomes a precise probe for the model’s background competence, providing an unambiguous gauge of whether foundational knowledge gaps remain. Such characteristic is unseen in any of the mathematical benchmarks before, which almost solely emphasizes evaluating models’ reasoning ability.

解决数学证明问题的能力代表了大型语言模型（LLM）的一个关键前沿。与具有确定答案的初等数学任务不同，证明题要求结构化的逻辑推理、多步论证以及连接抽象概念的能力——这些技能传统上与人类高等数学能力相关。从这个角度来看，擅长解决数学证明问题的模型更有可能具备超越表层模式识别的、类人思维能力，这一特性在开发和改进LLM的过程中被高度重视。因此，一个专门为评估当前模型在数学证明方面表现而设计的、可扩展且高质量的基准，对于系统性地评估其推理能力、并推动语言模型向更具类人逻辑能力的方向发展至关重要。

目前，构建基于数学证明的评估基准主要有两种主流方法。第一种——也是最直接的——方法依赖于数学专家的手动评估。尽管这仍然是基准质量的黄金标准，但它需要大量的人力投入，并且本质上难以规模化。第二种方法涉及使用形式化证明语言（如Lean）将数学问题和证明编码为机器可验证的语言。然而，这种方法面临类似的局限性：每个问题都需要精心构建的形式化过程，这使得该过程成本高昂、劳动密集，且对于构建可扩展的基准来说不切实际。因此，这两种方法在生产大规模、高质量的基准方面都面临着巨大的障碍。

为了应对上述挑战，我们提出了一个明确的设计目标：开发一个高质量的基准，用于评估LLM在高等数学证明问题上的表现，旨在严格评估其推理能力。同样重要的是，该基准必须是完全可扩展的——其创建过程应仅需最少的人力投入和较低的生产成本，以确保其在大规模评估中的实际应用性。

**我们的解决方案**

为实现我们的设计目标，我们提出了Proof2Choice，这是首个完全自动化、基于LLM的框架，可将难以评估的数学证明题改编为易于评估的选择题。Proof2Choice的核心机制是一个由多个强大的LLM和谐协作的精心编排的流水线，通过策略性地修改原始数学定义或命题中的关键元素——如关键词、条件和公式——来生成一组候选干扰项。随后，该流水线利用另一组强大的LLM作为评判员，对这些候选选项进行筛选和提炼，以确保最终的干扰项不仅在数学上不正确，而且看似合理且具有迷惑性。最后，我们将Proof2Choice应用于开源的代数几何教科书和参考著作“The Stacks project”中的定义和命题，从而生成了**AProofBench**，**这是首个专注于数学证明题的大规模基准**。它包含456道选择题，每道题提供六个选项：两个完全正确的选项和四个精心设计的干扰项。

我们的框架和基准具有以下关键创新点：

1. 我们提出了Proof2Choice，这是迄今为止首个基于LLM的全自动化框架，可将数学证明题转化为多项选择题。这拓宽了传统的数学问答范式——从仅仅处理有确定答案的问题，扩展到前沿数学中内容更丰富、以证明为中心的领域。与基于人工流程创建数学证明基准相比，Proof2Choice能够自动生成选择题，从而极大地减少了在检查、筛选和形式化数学证明题方面的人力投入。
2. 我们强调，Proof2Choice是一个领域无关的框架，它创新地将数学证明题转化为多项选择题。在任何以类证明推理——如逻辑推导、形式论证或结构化问题解决——比表层答案生成更为关键的领域，此工作流都可以被轻松应用。
3. 我们提出了AProofBench，这是首个专注于数学证明题的大规模基准。在我们的工作之前，数学证明基准均由数学专家手动创建，其规模受到可用的高质量竞赛级种子问题的数量以及可用人力资源的限制。相反，我们的基准是完全可扩展的，这得益于其基于LLM的自动化生成特性，以及源自开源大学级别参考著作的更广泛数据源。
4. 我们强调，AProofBench是一个严谨且富有挑战性的基准，在提升当前模型的数学推理能力方面具有巨大潜力。与许多流行的数学基准（例如MATH和MMLU）中LLM已达到性能饱和的情况不同，我们对多个开源和闭源LLM在AProofBench上的评估显示，即便是迄今为止表现最好的模型也仅取得了60分的中等成绩，而普遍得分低至20分左右。这一结果生动地证明了我们基准的严谨性和挑战性，及其在推动LLM理解前沿数学方面的巨大潜力。
5. 我们进一步强调，AProofBench不仅评估LLM的推理能力，还评估其事实性知识；要发现隐藏在干扰项中那些微妙且被故意修改的错误，需要广泛的领域熟悉度。当我们试图推动一个LLM在最难的推理基准（如HLE-Math或FrontierMath）上表现出色时，我们首先需要确保该模型在预训练和监督微调（SFT）阶段已经掌握了必要的背景知识。Proof2Choice提供了一个直接的解决方案：用那些先决条件与目标基准相匹配的问题作为种子（这些问题通常很容易找到），由此产生的基准便成为探测模型背景能力的精确探针，为判断其是否存在基础知识差距提供了明确的衡量标准。这一特性是之前的任何数学基准中都未曾见过的，因为它们几乎完全侧重于评估模型的推理能力。

Pyb的写法：

\subsection{Background and Motivation}

Mathematical capability stands as one of the most pivotal yardsticks for large language models (LLMs). Rooted in an ancient and ever-refining discipline that epitomizes the distilled rigor of human intellect, mathematics has steered every major advance in human civilization. Consequently, virtually every leading LLM enterprise now allocates substantial resources to elevate their models' mathematical capability, thereby reaffirming the centrality of mathematics in contemporary LLMs.

（大意：数学是模型最重要的能力之一）

Modern mathematics extends across hundreds of distinct branches, and ensuring sufficiently solid training across the vast majority of them during pretrain and SFT periods is the decisive lever for boosting LLMs' overall mathematical performance. The most challenging mathematical benchmarks currently available, such as HLE-MATH [TODO: ref], deliberately interleave questions from various cutting-edge domains, such as algebraic geometry, higher algebra, and number theory. Only if the model has already been thoroughly grounded in these domains can one effectively amplify its latent mathematical capacity through processes such as RL.

（大意：数学有上百个分支，确保绝大多数分支在pretrain和sft阶段都得到了充分训练对提高模型数学性能是至关重要的）

Mathematical benchmarks are currently the only effective, reproducible, and widely accepted metric for determining whether a model has acquired domain-adequate mathematical competence, yet the existing suite is systematically incapable of serving this purpose for most branches of mathematics, especially frontier branches. Contemporary mainstays such as GSM8K, MATH, and AIME [TODO: refs], though frequently used, are confined to secondary-school curricula and introductory Olympiad material; their problems therefore provide virtually no signal concerning a model's grasp of modern mathematics. At the other extreme, benchmarks like FrontierMath and HLE-MATH [TODO: ref] do engage with cutting-edge content, but they were deliberately designed as stress tests rather than diagnostic tools: the tasks are so formidable that even specialist humans struggle, rendering the benchmarks unsuitable for the fine-grained, mid-training monitoring required during pretrain and SFT periods. Consequently, the current landscape leaves researchers with an untenable choice between elementary tasks and intractable grand challenges, highlighting a critical absence of domain-specific benchmarks pitched at an appropriate, moderate level of difficulty within frontier mathematics.

（大意：benchmark是最常用的衡量模型数学能力的方法，但是现有的数学benchmark在大部分数学领域不具备这个能力）

Current approaches to constructing mathematical benchmarks—namely, harvesting from natural corpora, commissioning expert-crafted items, or resorting to formal languages such as Lean—are all fundamentally limited in scope or scalability.

* Extracting problems from existing textual sources is effective only for comparatively elementary mathematics, such as introductory competition questions or standard undergraduate exercises. The scarcity of advanced problems in naturally-occurring corpora is attributable to two mutually reinforcing constraints. First, the evaluation protocols that are currently practical can only score tasks that terminate in a single, verifiable final answer; they are not yet able to appraise the logical coherence of entire proofs. Consequently, any problem that is automatically extracted must already possess this answer-key artefact, a condition that is overwhelmingly satisfied by elementary material but rarely by research-level material. Second, frontier mathematics is almost exclusively articulated through definitions and proofs whose value resides in logical structure rather than in any discrete numerical output. Since researchers do not, in the course of their work, generate the kind of “question-plus-definitive-answer” pairs that an extraction pipeline can harvest, the raw literature is intrinsically devoid of the very artefacts that the first constraint demands. These two limitations—an evaluative bottleneck on the consumer side and an expressive mismatch on the producer side—jointly confine corpus-based benchmarks to comparatively elementary domains.
* Resorting to domain experts to design items from scratch remedies the content deficit but at prohibitive cost: each problem must be individually conceived, reviewed, and validated, a process that is labor-intensive, economically inefficient, and inherently resistant to expansion.
* Formal languages promise machine-checkable rigour, yet Lean and its kin have existed for scarcely two decades, leaving their textual repositories minuscule; consequently, any benchmark in this paradigm still hinges on extensive human annotation, reproducing the same drawbacks of expense and limited scalability encountered in expert authoring.

（大意：现有的造数学benchmark的方法都在scope或者scalability上有不足）

Therefore, there is an urgent need for a scalable, cost-effective, and fully automated method capable of generating benchmarks across diverse branches of mathematics.

（大意：所以我们需要一套scalable，经济的，全自动的在数学 各个领域构造benchmark的方法）

\subsection{Our Solution}

To address this gap, we propse Proof2Choice, the first scalable, fully automated, domain-agnostic LLM-based framework for mathematical benchmark synthesis. The central idea of Proof2Choice is to transform hard-to-evaluate proof problems into easily scorable multiple-choice questions. At its heart is a carefully orchestrated pipeline of powerful LLMs. First, a generation team of models crafts distractors by strategically altering keywords, conditions, or formulas in the original statements. Then, a separate judging team filters and refines these candidates, discarding any that are obviously wrong while retaining those that are subtly and deceptively flawed. This carefully engineered generation and verification process minimizes potential evaluation bias and guarantees the exceptional quality of the resulting questions. A key strength of Proof2Choice is its domain-agnostic design: armed with nothing more than seed questions—which are relatively easy to obtain, regardless of domain or difficulty—it can automatically synthesize high-quality questions at just the right level to evaluate a model's performance in that domain.

To anchor Proof2Choice in a firm foundation, we instantiate it on the definitions and propositions of the open source Algebraic Geometry textbook and reference work “The Stacks project”. The result is AlgGeoTest, a multiple-choice benchmark designed to probe large language models' grasp of Algebraic Geometry—a frontier domain of modern mathematics that occupies a central position within the contemporary mathematical landscape. AlgGeoTest contains 456 items, each offering six options: exactly two correct answers and four carefully engineered distractors. Our evaluation of AlgGeoTest on multiple open-source and closed-source LLMs reveals that even the best-performing model to date achieves only a moderate score of around 60, with scores commonly lower than 20. This result underscore the rigorous and challenging nature of our benchmark.

To safeguard quality, we engaged expert mathematicians to audit every question of AlgGeoTest and the responses produced by leading models. The audit reveals that over 98.75$\%$ of model-generated distractors are mathematically incorrect yet deceptively plausible, while more than 95$\%$ of the benchmark questions meet the same standard, with every distractor satisfying the same stringent criteria.

Model failures on AlgGeoTest arise chiefly from two sources: an inability to detect subtle flaws in incorrect choices or a hallucinated belief that a valid choice is inconsistent. Both shortcomings trace to intrinsic limitations and gaps in background knowledge, not to any defect in AlgGeoTest itself. The benchmark's uncompromising quality thus offers compelling evidence for the robustness of Proof2Choice.

Our key novelties are summarized as follows:

* We propose the idea of adapting mathematical proofs into choice questions for mathematical benchmark synthesis.
* We propose the first scalable, fully automated, domain-agnostic framework for mathematical benchmark synthesis, named Proof2Choice. Proof2Choice is valuable for securing robust pretrain and SFT coverage across the majority of branches of mathematics, which is essential for elevating LLMs' overall mathematical competence.
* We instantiate Proof2Choice on the open source Algebraic Geometry textbook and reference work “The Stacks project”, yielding AlgGeoTest, a robust benchmark for gauging LLMs' comprehension of Algebraic Geometry—a frontier field that lies at the heart of modern mathematics.

数学能力是衡量大型语言模型（LLM）最关键的指标之一。数学植根于一门古老且不断精进的学科，它象征着人类智慧经过提炼后的严谨性，并引领了人类文明的每一次重大进步。因此，几乎所有领先的LLM企业现在都投入大量资源来提升其模型的数学能力，这再次印证了数学在当代LLM中的核心地位。

现代数学涵盖了数百个不同的分支，在预训练（pretrain）和监督微调（SFT）阶段确保模型在绝大多数分支上都得到足够扎实的训练，是提升LLM整体数学性能的决定性杠杆。目前最具挑战性的数学基准，如HLE-MATH [TODO: ref]，都有意地交织了来自各个前沿领域的问题，例如代数几何、高等代数和数论。只有当模型在这些领域已经有了扎实的基础，才能通过强化学习（RL）等过程有效放大其潜在的数学能力。

数学基准是目前唯一有效、可复现且被广泛接受的，用以判断模型是否已获得特定领域内足够数学能力的衡量标准。然而，现有的基准体系系统性地无法对大多数数学分支，特别是前沿分支，起到这一作用。尽管像GSM8K、MATH和AIME [TODO: refs]这样的当代主流基准被频繁使用，但它们仅限于中学课程和入门级奥林匹克竞赛内容；因此，它们的问题几乎无法提供任何关于模型对现代数学掌握程度的信号。在另一个极端，像FrontierMath和HLE-MATH [TODO: ref]这样的基准虽然确实涉及前沿内容，但它们被刻意设计为压力测试而非诊断工具：这些任务极其艰巨，即使是人类专家也感到吃力，这使得这些基准不适用于在预训练和监督微调期间所需的细粒度、训练过程中的监控。因此，当前的研究现状让研究人员不得不在初级任务和棘手的重大挑战之间做出一个难以为继的选择，这凸显了在前沿数学领域内，严重缺乏难度适中、针对特定领域的基准。

当前构建数学基准的方法——即从自然语料库中收集、委托专家创建题目，或求助于如Lean这样的形式化语言——都在范围或可扩展性上存在根本性的限制。

* 从现有文本源中提取问题仅对相对初等的数学有效，如入门级竞赛题或标准的本科生练习题。在自然语料库中，高级问题的稀缺性可归因于两个相互强化的制约因素。首先，目前实用的评估协议只能对那些有单一、可验证最终答案的任务进行评分；它们尚无法评估整个证明过程的逻辑连贯性。因此，任何自动提取的问题都必须已经具备这种“答案密钥”式的产物，这一条件在初等材料中绝大多数情况下都能满足，但在研究级别材料中却极为罕见。其次，前沿数学几乎完全是通过定义和证明来阐述的，其价值在于逻辑结构而非任何离散的数值输出。由于研究人员在工作中并不会生成那种可被提取流水线收集的“问题+确定答案”对，原始文献本质上就缺乏第一个制约因素所要求的那种产物。这两个限制——消费端（评估方法）的评估瓶颈和生产端（研究文献）的表达形式不匹配——共同将基于语料库的基准限制在了相对初等的领域。
* 求助于领域专家从头设计题目弥补了内容上的不足，但成本过高：每个问题都必须经过独立构思、审查和验证，这个过程劳动密集、经济效率低下，且本质上难以扩展。
* 形式化语言承诺了机器可检查的严谨性，但Lean及其同类语言存在至今不过二十年，其文本库微乎其微；因此，任何在此范式下的基准仍然依赖于大量的人工标注，重现了与专家编写题目时遇到的相同的高成本和有限可扩展性的缺点。

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基准。

**我们的解决方案**

为了弥合这一差距，我们提出了Proof2Choice，这是首个可扩展、全自动、领域无关的，基于LLM的数学基准合成框架。Proof2Choice的核心思想是将难以评估的证明题转化为易于评分的多项选择题。其核心是一个精心编排的、由强大的LLM组成的流水线。首先，一个生成团队通过策略性地修改原始陈述中的关键词、条件或公式来制作干扰项。然后，一个独立的评判团队对这些候选选项进行筛选和提炼，剔除那些明显错误的选项，同时保留那些存在微妙且具有迷惑性缺陷的选项。这种精心设计的生成和验证过程最大限度地减少了潜在的评估偏差，并保证了最终生成问题的卓越质量。Proof2Choice的一个关键优势是其领域无关的设计：仅需提供种子问题——无论领域或难度如何，这些问题都相对容易获取——它就能够自动合成出质量上乘、难度恰当的问题，以评估模型在该领域的表现。

为了将Proof2Choice建立在坚实的基础上，我们在开源的代数几何教科书和参考著作“The Stacks project”的定义和命题上对其进行了实例化。其成果是AlgGeoTest，一个旨在探究大型语言模型对代数几何掌握程度的多项选择基准——代数几何是现代数学的一个前沿领域，在当代数学格局中占据核心地位。AlgGeoTest包含456个项目，每个项目提供六个选项：两个完全正确的答案和四个精心设计的干扰项。我们对多个开源和闭源LLM在AlgGeoTest上的评估显示，即便是迄今为止表现最好的模型也仅取得了60分左右的中等成绩，而普遍得分低于20分。这一结果突显了我们基准的严谨性和挑战性。

为保证质量，我们聘请了数学专家来审计AlgGeoTest的每一个问题以及领先模型产生的回答。审计结果显示，超过98.75%由模型生成的干扰项在数学上是不正确的，但看似合理且具有迷惑性，同时超过95%的基准问题达到了相同标准，即每一个干扰项都满足同样严格的准则。

模型在AlgGeoTest上的失败主要源于两个方面：一是无法检测出错误选项中的微妙缺陷，二是幻觉出有效选项存在矛盾。这两个缺点都源于其固有的局限性和背景知识的欠缺，而非AlgGeoTest本身的任何缺陷。该基准不折不扣的质量因此为Proof2Choice的稳健性提供了有力的证据。

我们的关键创新点总结如下：

* 我们提出了将数学证明改编为选择题以用于数学基准合成的思想。
* 我们提出了首个可扩展、全自动、领域无关的数学基准合成框架，名为Proof2Choice。对于确保模型在绝大多数数学分支上获得稳健的预训练和监督微调覆盖而言，Proof2Choice极具价值，而这对于提升LLM的整体数学能力至关重要。
* 我们在开源的代数几何教科书和参考著作“The Stacks project”上实例化了Proof2Choice，从而产生了AlgGeoTest，这是一个用于衡量LLM对代数几何理解程度的稳健基准——代数几何是位于现代数学核心的前沿领域。

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.

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.

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.

Therefore, there is a 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.

**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.

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.

 **第一句：确立主题的重要性。** 开门见山，点明“数学能力”是衡量LLM的关键标尺。

 **第二句：引出对工具的需求。** 紧接着第一句，指出既然数学能力如此重要，那么我们就“需要”好的评估工具来衡量它。

 **第三句：点明当前工具的困境。** 顺着上一句的“需求”，直接指出“瓶颈”在于当前创造这些工具的方法存在根本性缺陷。

 **第四句：明确所需方案的特性。** 基于上一句的“瓶颈”，精准地定义出我们到底需要一种具备哪些特质（可扩展、自动化等）的新方法。

 **第五句：亮出本文的核心贡献。** 在清晰地铺垫好所有背景和需求之后，响亮地提出我们的 Proof2Choice，并强调它完美地满足了第四句中定义的所有要求。

数学能力是衡量大型语言模型（LLM）最关键的标尺之一。数学植根于一门浓缩了人类智识严谨性的学科，它提供了一个独特的视角，用以评估模型在抽象推理、逻辑推导和结构化思考方面的能力——这些都是实现更通用人工智能的先决条件。因此，一个模型在高等数学任务上的表现，成为其真实认知深度的关键指标，能够区分表面的模式匹配与真正的理解能力。

确保LLM发展出这种深层能力需要稳健且有针对性的评估工具。然而，现有的数学基准体系在现代数学的大多数分支上都系统性地无法实现此目的。当前的主流基准，如GSM8K和MATH，仅限于初等和竞赛水平的课程，模型在这些任务上的表现已接近饱和。因此，它们在反映模型对高级概念的掌握程度上提供的信号很弱。在另一个极端，像HLE-Math这样的基准虽然涉及前沿内容，但被设计为难以解决的压力测试，而非诊断工具。其极高的难度使其不适用于指导LLM发展所需的细粒度监控。这使得研究人员不得不在初级任务和艰巨挑战之间做出一个难以为继的选择，凸显了在前沿数学领域内，严重缺乏难度适中、针对特定领域的基准。

瓶颈在于当前的基准构建方法，这些方法在范围或可扩展性上都存在根本性的限制。基于语料库的收集方法仅对“问题+答案”对丰富的初等数学有效；而通过证明和定义来阐述的前沿数学，本质上就缺乏这类易于提取的产物。专家编写题目弥补了内容上的不足，但成本过高，因为每个问题都需要经历构思、审查和验证这一劳动密集型过程，这在本质上是难以规模化的。最后，像Lean这样的形式化语言承诺了机器可检查的严谨性，但它们的生态系统仍处于起步阶段，这意味着在此范式下合成任何基准仍然依赖于广泛且昂贵的人工标注。这些限制——消费端的评估瓶颈和生产端的创建瓶颈——使得用现有工具生成必要的基准几乎成为不可能。

因此，当前迫切且未能满足的需求是，需要一种可扩展、成本效益高且完全自动化的方法，能够在高等数学的各个分支中生成高质量的基准。这种方法需要克服自然语料库的表达不匹配问题，并规避专家和形式化介入所带来的高昂成本，为评估和改进下一代LLM提供一条新的前进道路。

**我们的解决方案与贡献**

为解决这一关键差距，我们提出了 **Proof2Choice**，这是首个用于从基于证明的陈述中合成数学基准的可扩展、全自动且领域无关的框架。Proof2Choice的核心思想是将难以评估的证明转化为易于评分的多项选择题。其核心是一个精心编排的、由强大的LLM组成的流水线：一个“生成者”团队通过策略性地修改原始数学陈述中的关键词、条件或公式来制作具有迷惑性的干扰项，之后一个“评判者”团队会筛选这些候选选项，以确保它们既在数学上无效又具有高度的合理性。

我们在《The Stacks project》——一本关于代数几何的研究生级别综合教科书——上实例化了Proof2Choice，创建了**AProofBench**。该基准包含456个具有挑战性的多项选择题，每道题有两个正确选项和四个精心设计的干扰项。AProofBench被证明是一个严谨的诊断工具：我们的评估显示，即便是最先进的LLM也仅取得了20-60%的低准确率。此外，由数学专家进行的全面审计证实了该基准的质量，超过95%的问题达到了清晰和严谨的高标准。审计还揭示，模型的失败主要源于其无法验证基础知识，而非问题本身的缺陷。

我们的工作做出了以下关键贡献：

* 我们提出了**Proof2Choice**，这是首个用于创建基于证明的评估基准的全自动且领域无关的框架，打破了以往方法在可扩展性和成本上的瓶颈。
* 我们发布了**AProofBench**，这是首个专为诊断目的而设计、经过专家验证的前沿数学大规模基准。
* 通过对AProofBench进行详细的失败分析，我们为当前LLM的根本性缺陷提供了关键洞见，证明了缺乏稳健的知识验证能力是实现高级数学推理的主要障碍。

Evaluating the mathematical reasoning of Large Language Models (LLMs) is a critical yet challenging frontier. Existing benchmarks fall short, particularly for proof-centric problems, as manual creation is unscalable and costly, leaving the true reasoning abilities of LLMs largely unassessed. To overcome these barriers, we propose **Proof2Hybrid**, the first fully automated and domain-agnostic framework that synthesizes high-quality, proof-centric benchmarks from natural language mathematical corpora. At the core of our framework is a novel **"m-out-of-n multiple judge"** question format, specifically designed to enable robust, automatic evaluation while being resilient to guessing and superficial pattern matching inherent in traditional formats. As a demonstration of our framework, we introduce **AlgGeoTest**, a rigorous benchmark for algebraic geometry comprising 456 challenging items. Our extensive evaluations on state-of-the-art LLMs using AlgGeoTest reveal profound deficits in their comprehension of formal mathematics, providing a more precise measure of their true capabilities. Our framework and benchmark pave the way for a new wave of scalable, in-depth research into the mathematical intelligence of AI systems.

评估大型语言模型（LLM）的数学推理能力是一项至关重要但又极具挑战性的前沿任务。现有基准，尤其是在“以证明为中心”的问题上，存在严重不足：人工创建的方法成本高昂且无法扩展，导致LLM的真实推理能力在很大程度上未经检验。为突破这些障碍，我们提出了 **Proof2Hybrid**，这是首个能够从自然语言数学语料库中全自动、领域无关地合成高质量、以证明为中心的基准的框架。我们框架的核心是一种新颖的“**n选m多重判断**”问题格式，该格式经过专门设计，旨在实现稳健的自动评估，同时能有效抵御传统题型中固有的猜测和浅层模式匹配问题。作为我们框架能力的展示，我们构建并发布了 **AlgGeoTest**，一个针对代数几何领域的、包含456个挑战性条目的严谨基准。我们利用AlgGeoTest对当前最先进的LLM进行的广泛评估，揭示了它们在形式化数学理解方面存在着深刻的缺陷，从而提供了对其真实能力更精确的衡量。我们的框架和基准为未来对人工智能系统进行可扩展的、深度的数学智能研究铺平了道路。