

从希尔伯特判定问题到推理智能体：定理自动证明的百年征程

高小山

中国科学院数学与系统科学研究院

2026 年 7 月 8 日

自希尔伯特 1928 年提出判定问题至今，定理自动证明发展已近百年，经历了理论奠基、符号推理、交互式证明、数学家机械化、大语言模型(LLM)推理几个主要阶段。如今，随着 LLM 的快速发展与智能体的使用，LLM 数学研究智能体已经成为数学家的得力助手。

本短文将简要回顾定理自动证明的主要成就，介绍我们团队开发的数学研究智能体 MechMath Agent Team，最后谈几点经验与感想。

一、定理自动证明百年征程

数学定理自动证明百年来历经四次范式迭代，每一次都重塑人类对“机器推理”或“AI 推理”的认知。

1. 理论奠基：希尔伯特判定问题

1900-1930 期间，希尔伯特逐步形成后世称之为希尔伯特纲领的宏大计划：将各个数学分支公理化，证明其相容性、完整性。特别是，发展统一的算法证明一个公理体系内所有的数学定理，即著名的判定问题 (Entscheidungsproblem)，本质上是追求定理的自动证明。

哥德尔、Herbrand、图灵等人对希尔伯特判定问题的突破性解答，搭建起 AI 推理的理论根基：

- 哥德尔 1931 年证明，包含皮亚诺算术的相容公理系统，必然存在“真但无法证明”的命题，击碎了“数学定理自动证明”的希望。
- 另一方面，Herbrand 也是在 1931 年给出特判定问题的正面结论。他给出了一阶逻辑半判定算法：只要命题证明存在，该算法即可自动生成。
- 1936 年，图灵为了精确描述“判定”与“计算”概念，引入了通用计算模型“图灵机”。

以上结果说明，希尔伯特判定问题在理论上完整解决了！对一阶逻辑公理系统，存在一台图灵机，可以自动生成其中所有证明！但现实远非这样简单：Herbrand 算法复杂度太高，即使在今天的计算机上也不能证明非平凡定理。

2. 热潮遇阻：符号 AI 的乐观、进展与困难

1956 年 Newell-Simon 发布“逻辑理论机器”，一次性自动证明了《数学原理》前 52 条定理中 38 条命题逻辑定理。该文是 AI 奠基工作之一，开启了符号主义人工智能。他们随后推出“通用问题求解器”，试图通过知识积累与自动

推理实现通用问题求解。Robinson 于 1965 年提出了一阶逻辑定理证明的“高效”方法-归结法。业界一度乐观认为通用人工智能近在眼前。

1971 年 Cook 证明：命题逻辑定理证明 (SAT) 是 NP 完全。这意味着除非 $P=NP$ (千禧年七大难题之一)，不存在完整且高效的定理证明算法，符号主义 AI 的理想受阻。

3. 砥砺前行：交互证明与数学机械化

定理自动证明的继续推进分化出两条新路线，简单讲：交互证明部分放弃了自动、数学机械化部分放弃了完整。

交互式定理证明 ITP (1968 AUTOMATH) 不再让机器自动生成证明，转而让计算机自动验证已有证明的逻辑严格性。由此诞生的 Isabelle、Rocq、Lean 等形式化验证工具，不仅被用于验证复杂证明的严格性，还被应用于芯片、操作系统、网络协议安全性验证。形式化数学是经过形式化验证的数学知识。有些出乎预料的是：已有数学证明的形式化也非常耗时：菲尔兹奖得主陶哲轩在 2024 年一个报告中讲到：中等难度证明的形式化耗时可能是人工证明的 10 倍。Szegedy 在 2020 年的悲观估计认为，当时只有 0.001% 数学知识被形式化。LLM 近期一系列重要进展，使得我们距离将现有数学知识的形式化的目标更近了一步。

数学机械化 (1978 吴方法) 是指针对有意义的数学分支或一类问题发展高效算法。例如，吴文俊提出的吴方法不仅“将几何定理证明从自动推理的一个不太成功的领域变为最成功的领域之一”，而且被广泛应用于 CAD、机器人视觉、密码分析、数控机床。沿着这条路线，出现了计算机代数、计算数论、计算代数几何、计算拓扑、计算群论、符号分析等新兴学科。

4. LLM 时代：AI 成为数学家强大助手

2022 年 OpenAI 推出 ChatGPT，标志人工智能迈入 LLM 时代。LLM 具备海量知识储备与强大泛化能力，结合智能体系统化调用 LLM 与其他工具，将 AI 推理发展推向全新阶段。依托 LLM 的数学智能体，如今已成为辅助数学家开展研究的重要工具。LLM 驱动的 AI 推理大致可以分为两个阶段：

第一个阶段是竞赛级别的定理证明，可以认为是其能力测试验证阶段。代表性成果包括：AlphaGeometry 与 AlphaProof 拿下 IM02024 年银牌。目前，LLM 智能体解决竞赛问题已经相对容易。

第二个阶段是研究级别的定理证明。一些代表性 LLM 智能体包括：AlphaEvolve (2025 年 6 月) 通过 LLM 驱动的进化式优化，刷新了拉姆齐数下界等若干数学计数问题；Aletheia (2026 年 2 月) 助力数学家完成 5 篇数学论文。

需要说明的是，尽管 LLM 数学推理相关的理论研究已取得显著进展，但现有理论仍难以刻画实际部署的 LLM 所具备的数学推理能力。

二、数学研究智能体：MechMath Agent Team

MechMath Agent Team (MMAT, <https://mechmath.github.io/>) 是中国科学院数学与系统科学研究院数学机械化实验室开发的基于 LLM 的数学研究智能体，其目标是在数学研究全流程上为数学家提供得力助手。

1. MMAT 简介

MMAT 主要由两部分构成：整体架构 (General Harness) 与三个智能体。

整体架构：可以形象地理解为 LLM 的操作系统，规划如何使用 LLM 及其他工具进行数学研究。该整体架构将数学研究的主要步骤或方法模板化为 30 多种功能不同的 SubAgent 与工具，并进行有效管理。一些典型 SubAgent 包括：

- **查询检查：**查询、检查论文；
- **证明生成：**Route 探索、Sketch 生成、Proof 生成、反例生成、结论验证、评估；
- **符号计算：**编程、调用符号计算工具
- **人机交互：**提供建议、明确难点、接受反馈
- **形式化验证：**Lean 编码生成、语义对齐检查

MMAT 智能体：在整体架构 (General Harness) 的基础上，通过使用不同角色的 SubAgent 与外部工具，形成三个智能体：

- **NL-Prover 自然语言证明器。**生成自然语言数学证明，遵循「生成-校验-修正」闭环：自主检索文献、拆解大命题为多层引理、并行探索多条证明路径、自动搜寻反例验证思路，失败路径自动归档沉淀经验等。
- **FL-Prover 形式化证明器。**自动生成或将自然语言证明转化为 Lean 4 编译器可核验的形式化代码，同时保证自然语言与形式化语言语义对齐。
- **KB-Manager 数学知识管理器。**构建面向研究课题的专属知识图谱，统一归档文献、定义、已证引理、形式化代码、未完成证明、失败思路障碍点，支持长期项目持续迭代，解决多轮研究知识碎片化问题。

2. 测试成果

MMAT 独立或与数学家交互解决了 11 个数学问题，包括代数计算理论、微分代数、数论领域 8 项长期公开问题，提交了 9 篇 arXiv 论文；其中 4 个问题由 MMAT 全自动独立解决，其余 7 项问题的核心关键引理由 MMAT 完成证明；MMAT 还给出了 2 篇论文的形式化验证。解决的问题包括：

- **Erdős - Rényi (1949) 稀疏多项式完整开方稀疏度猜想。**单变量情形由 Schinzel (1987)、Zannier (2008) 解决。我们解决了多变量情形。
- **Plaisted (1984) 等的稀疏多项式整除判定公开问题。**我们证明了有限域上稀疏多项式整除判定为 CoNP 难。
- **解决了 Katz (1987) 提出的广义 Airy 算子等价问题。**

- Davenport 等（2009）的输出敏感型 GCD 计算挑战问题。我们证明有限域上输出敏感型稀疏多项式 GCD 计算为 NP 难。
- Roche（2018）的稀疏多项式拟线性乘法公开问题。我们给出整系数稀疏多项式乘法的拟线性时间算法。
- 在 BSD 猜想相关的 CM 椭圆曲线 Shafarevich-Tate 群有限性问题得到阶段性结论。

实验验证表明 MMAT 确实可以用于数学研究全流程、且具有很强的解决数学问题的能力，可作为数学家的得力助手。

三、几点经验与感想

根据对 MMAT 的使用，我分享一下几点经验与感想。

1. 设计完善的 LLM 数学智能体有以下显著能力：首先，其数学知识覆盖面非常广、基本涵盖了现代数学前沿内容；其次，智能体表现出很强的数学问题求解规划、证明与使用工具能力；最后，与直接使用 LLM 不同，LLM 数学智能体通过引入专门的验证机制，能够更好的抑制幻觉发生；例如，MMAT 一旦生成证明，大多数情形下具有很强的逻辑严谨性。

2. 对于困难数学问题，LLM 智能体通常不能一次性给出完整证明。数学家在证明思路提示、证明正确性验证、理解 LLM 智能体的提示等方面，依旧起着关键作用；人机交互是 LLM 智能体的有效使用方式。

3. LLM Agent 生成自然语言证明的能力与它生成形式化证明的能力之间存在巨大鸿沟，目前 NL-Prover 的能力远远大于 FL-Prover 的能力。大多数情形下，NL-Prover 生成的证明仍然无法完全形式化，需要专家验证其正确性。

4. LLM 基座模型的能力与智能体 Harness 设计是 LLM 智能体能力的关键。另一方面，前期发展的逻辑求解器 (SAT、SMT)、数学机械化与符号计算方法，被 LLM 智能体用来增强 LLM 计算能力弱的短板；交互式形式化被用来验证证明的正确性。

5. LLM 数学智能体的发展，使得陶哲轩 2024 年提出的预言“AI Will Become Mathematicians’ Co-Pilot”已经变为现实。LLM 智能体已经成为数学研究的强大研究助手，大大提高了数学研究的效率。

6. 数学家仍将是数学研究的主体。在提炼核心研究问题、校验 AI 推导证明、搭建完整理论框架、创造全新数学概念、评判学术成果价值等关键环节，人类数学家具备不可替代的作用。