大型语言模型擅长什么样的数学?

原文作者：Timothy Gowers（剑桥大学，菲尔兹奖得主）

原文链接：https://gowers.wordpress.com/2026/08/12/what-sort-of-maths-are-llms-good-at/

为了便于任何可能在遥远的将来（比如说一个月后）读到这篇博客的人理解，先说明一下：我写下这篇文章的时间，是在 OpenAI 宣布其解决了数学与理论计算机科学中的十大难题之后的几天。其中包括首次构造出非 sofic 群，以及证明了多重色 Ramsey 数 $R(3,3,\ldots,3)$（其中有 k 个 3）关于 k 超指数增长。据我参加过的各种报告来看，前者是群论中最重要且尚未解决的问题之一；后者则是 Ramsey 理论中的一个重大开放问题——我原本并不指望在有生之年看到它被解决，当然，如今这样的预期都必须加以修正了。我想把时间点讲清楚，是因为我将以"LLM 的能力会继续快速变化"这一预期来讨论它们当前的状况。因此很可能过不了多久，如果我这篇东西里还有什么有趣之处，那将主要因为它记录下了 2026 年 8 月初时的情况。

这些结果，以及其他八个成果，都极其令人印象深刻；但即便如此，看起来也还不能说 LLM 在数学的方方面面都比所有人类更强。如果它们真的更强，那么相比于人类，它们巨大的速度优势就意味着应该涌现出多得多的成果洪流。所以，自然而然地会想弄清楚：LLM 擅长什么样的题目，又有哪里仍有提升空间。我并不假装对这一问题的答案有多好——一个好的答案，应是一种能很好地套用到当前这些例子上的清晰分类——但试着排除一些错误答案、并找出一些没有被证据明显否定的可能答案，这本身就是一件有趣的练习。

LLM 是否特别擅长找反例?

首先要说明的一点是：LLM 不只是擅长找反例，它们也能找到困难命题的证明。不过值得注意的是，它们解决的那些最著名的难题，几乎全部以"反例"（counterexample）的形式出现，而非证明。上面提到的两个问题如此，Jacobi 猜想和单位距离猜想也是如此。

如果有人想提出"LLM 特别擅长找反例"这一理论，那么要让这个理论更有说服力，有两件事值得去做。第一件听起来可能没什么问题：那就是确定"解决一个问题"在什么时候算是"找到了一个反例"。一旦这件事理清，第二件就是给出一个关于"LLM 为何特别适合解决这类问题"的潜在解释。

找到反例意味着什么?

我为什么说"找到反例意味着什么"并不完全显而易见？也许有人会说，这不就是说：一个命题形如"每一种某某类型的对象都具有某某性质"，然后你展示出一个该类型、却恰好不具有该性质的对象。然而，事情并不总是如此简单。

考虑一个著名结果——Vinogradov 定理，它断言：每一个足够大的正整数都能写成三个素数之和。这个命题的否定是（或者说等价于）：对每一个正整数 N，都存在一个整数 $n \geq N$，使得 n 不是三个素数之和。换句话说，它断言每一个正整数 N 都具有某种性质。从这个角度看，Vinogradov 找到了一个"不具有该性质的正整数 N"的例子。那我们是否要说 Vinogradov 找到了一个反例？显然不是——这个结果理所当然应被归类为一条定理，而不是反例。

因此，我们不能天真地说"LLM 特别擅长否定全称量化命题"：这里一定与全称量化的性质有关。就拿三素数例子来说，显然 Vinogradov 想的不是"我该怎么找到具有这种性质的 N？"，他想的更像是："我有一个很大的整数 n，我该怎么证明它能写成三个素数之和？"换句话说，他的全部注意力都集中在全称量化的 n 上，而存在量化的 N 只是在证明细节完成后的一种附属补充。

一般而言，许多有意思的结果，当被正式表述时，都以两个或三个（甚至更多）量词的交替开头。问题就变成：在某种意义下，哪个才是第一个"有意思的"被量化变量。

下面是另一个例子来阐明这一点，来自有限维赋范空间理论。对于好奇的读者我给出一些数学细节；如果不在意这些，可以跳过接下来三段，但依然能抓住我要说的要点。

设 X 和 Y 是两个 n-维赋范空间，设 T 是从 X 到 Y 的线性映射。我们说 T 是一个 C-同构，如果存在 $\lambda > 0$ 使得对每个 $x \in X$ 都有 $\lambda\|x\| \leq \|Tx\| \leq C\lambda\|x\|$。通过重新缩放，我们总可以取 λ 为 1，此时对每个 $x \in X$ 有 $\|x\| \leq \|Tx\| \leq C\|x\|$。若 $C = 1$，则这告诉我们 T 是一个等距（isometry）。一般地，X 与 Y 之间的 Banach–Mazur 距离 $d(X, Y)$ 定义为：存在一个从 X 到 Y 的 C-同构的最小 C。容易看出，Banach–Mazur 距离的对数，是 n-维赋范空间的等距类集合上的一个度量。另一个稍难、但也还不算

太难的事实是：由此得到的度量空间是紧的；事实上，它被称为 Banach–Mazur 紧空间（compactum）。

自然会问：Banach–Mazur 紧空间的直径是多少？这里开始变得有趣。Fritz John 的一个结果指出：每个 n-维空间 X 到 ℓ_2^n 的距离至多为 $\sqrt{n}$。（证明思路如下：在 X 的单位球内取出一个体积最大的 n-维椭球；它就是某个与 ℓ_2^n 等距的赋范空间 Y 的单位球；可以证明恒等映射是 X 与 Y 之间的一个 $\sqrt{n}$-同构。）由 Fritz John 定理与（乘性的）三角不等式，可推出对任意两个 n-维赋范空间都有 $d(X,Y) \leq n$。也就是说，Banach–Mazur 紧空间的直径至多为 n。但它是否会显著小于这个值呢?

一个表明答案并不显然的迹象，来自考察空间 ℓ_1^n 和 ℓ_∞^n。这两个空间之间的恒等映射是 n-同构，但我们可以做得好得多：不把标准基向量映到自身，而是映到单位立方体的顶点，且这些顶点尽量正交。特别地，若存在 $n \times n$ Hadamard 矩阵，则相应的线性映射是一个 $\sqrt{n}$-同构。可以进一步推进这一观察，推出：对任意 $p,q \in [1,\infty]$，ℓ_p^n 与 ℓ_q^n 之间的 Banach–Mazur 距离为 $O(\sqrt{n})$。同样容易证明 $d(\ell_1^n, \ell_2^n) = \sqrt{n}$，所以 ℓ_p-空间几乎无法改善那个容易得到的下界，而在存在 $n \times n$ Hadamard 矩阵的维数 n 上则完全无法改善。

1981 年，Gluskin 著名地解决了这一问题，确定了 Banach–Mazur 紧空间直径的正确渐近。通俗地说，他证明了这个直径与由 Fritz John 定理立即得出的上界相差一个常数。如果把量化明确写出来，我们得到的命题是

$$\exists c > 0 \ \forall n \ \exists X, Y \in K_n \ d(X,Y) \geq cn,$$

其中我把 K_n 记作所有 n-维赋范空间的集合。（如果你想争论说这并不是一个集合，那我再补充规定其底向量空间是 $\mathbb{R}^n$。）用语言说，存在一个正常数 c，使得对每个正整数 n，都存在 n-维赋范空间 X 和 Y，使得它们之间的 Banach–Mazur 距离至少为 cn。

我不能不简要描述一下 Gluskin 用来解决这个问题的那一漂亮而影响深远的思想。他取 X 和 Y 为这样的赋范空间：其单位球是如下定义的随机对称凸集——取标准基向量、以及少量其他随机单位向量，再加上所有这些向量的负向量，然后取它们的凸包。Gluskin 进而证明：若从这一分布中选取两个赋范空间，则以高概率它们的 Banach–Mazur 距离至少为 cn。

再回到主线：上述命题的逻辑形式，与 Vinogradov 定理的逻辑形式非常相似，后者是

$$\exists N \ \forall n \geq N \ \exists p_1, p_2, p_3 \in P \ \ p_1 + p_2 + p_3 = n$$

其中我把 P 记作素数集合。然而，Vinogradov 的结果毋庸置疑是一条定理，而 Gluskin 的结果毋庸置疑是一个反例，或至少是一个例子。这两个命题之间的重要差别是什么?

看来，在 Vinogradov 的三素数定理中，数 n 在"关于各被量化变量所要证明的那个陈述"里扮演着更本质的角色。在 Vinogradov 定理中，那个陈述是 $n = p_1 + p_2 + p_3$；而对于 Gluskin 的定理，要证明的陈述是

$\dim X = \dim Y = n$ 以及 $d(X, Y) \geq cn$，

我们可以等价地写为

$\dim X = \dim Y = n$ 以及 $d(X, Y) \geq c \dim X$。

在 Vinogradov 定理的情形中，全部挑战在于让那三个素数相加得到 n；而对 Gluskin，让 X 与 Y 的维数都等于 n 这一点毫不费力；挑战在于让 X 与 Y 相对于它们共同的维数彼此"相距甚远"。

这里还有一个需要牢记的进一步复杂之处：通过所谓的 Skolem 化（Skolemization）过程，一个全称量化命题形如 $\forall x \in X \; \exists y \in Y \;\; P(x, y)$ 可以被转化为一个存在量化命题 $\exists f : X \to Y \; \forall x \in X \;\; P(x, f(x))$。（要做到等价，需要选择公理，但它无疑是一个充分条件。）这并非只是逻辑上的小花招，它往往相当准确地反映了我们思考某些问题的方式。例如，把 Gluskin 的例子看作"给出一个针对任意维数 n 构造（或至少证明存在）一对合适赋范空间的配方"，换句话说，就是"构造一个从 $\mathbb{N}$ 到赋范空间对的合适函数，并给出在每个 n 处的取值"，比把它看作"每一个正整数 n 都具有某种复杂性质"的陈述要自然得多。

还有一个复杂之处：有些全称量化陈述会自然地由存在量化陈述推出，甚至可能与之等价。例如，"二维环面不同胚于二维球面"这条定理是一个全称量化陈述（从环面到球面的每一个映射都不是同胚），但证明它的自然方式，是证明一个存在性陈述——存在一个区分这两个空间的不变量。至于"全称陈述与存在陈述等价"的例子，考虑这样一类陈述：一个向量 $x \in \mathbb{R}^n$ 不属于某个紧集 A 的凸包。"A 中元素的任何凸组合都不等于 x"这个陈述，等价于"存在一个线性泛函 $\phi : \mathbb{R}^n \to \mathbb{R}$ 和一个 $\lambda \in \mathbb{R}$，使得 $\phi(x) > \lambda$ 且对每个 $a \in A$ 都有 $\phi(a) \leq \lambda$"。在这两种情形下，我们都很自然会把这个结果看作"通过一个存在性陈述来证明的定理"，也许是因为最终让我们感兴趣的正是那条定理。但用"什么让我们感兴趣"来判定什么是反例，似乎有点含糊；如果我们想令人信服地解释"为什么 AI 擅于找反例"，这是一个难用的判据。

一个反对"存在性陈述有某种特别适合 AI 之处"的更一般论点是：无论追求的头条结果是什么性质，建立存在性陈述的需要几乎贯穿所有数学研究。例如，若我想用归纳法证明一个命题，我很可能去寻找一个能更好地充当归纳假设的、被加强的命题。或者，若我想证明"每个具有性质 P 的 T 类型对象也都具有性质 Q"，那我很可能去找一个由 P 推出、并能用来证明 Q 的性质 R。这些更像是元数学的存在性问题，但其中的区别有时会模糊；更重要的是，当试图证明某个陈述 S 时，我们脑海中主要的问题常常不是"为什么 S 为真？"，而是"关于 S 的一个证明可能长什么样？"举个例子：我觉得自己相当理解为什么 Goldbach 猜想为真——一个极可信的素数概率模型能推出它，且与计算数据高度吻合——但若我真要严肃地尝试证明它，这种一个世纪以来许多数学家都有的理解，恐怕帮助有限。相反，我的主要任务是去找到足够强大的证明技巧，把这些启发式想法变成严谨的论证。

例子与反例的区别是什么？

逻辑上，每一个形如 $\exists x\ P(x)$ 的陈述，都是全称量化陈述 $\forall x\ \neg P(x)$ 的一个反例。但我们并不会把所有存在性陈述都描述为反例。例如，如果我说"ℓ_p-空间（其中 $1 \leq p < \infty$）都是可分的，c_0 也是可分的，但 ℓ_∞ 不可分"，我不会把后半个断言描述为"对'所有 Banach 空间都是可分的'这一断言的攻击的反例"，而会把它呈现为"一个不可分空间的最基本例子"。这里的关键点似乎是：本来没有任何特别理由认为所有 Banach 空间都可分，而举出一个不可分空间并不困难。

我认为第一个点更重要：当一个对象的存在推翻了一个我们本有相当充分理由相信的陈述时，我们更倾向于称它为反例。常常发生这样的情况：在反复尝试证明某个陈述未果之后，数学家开始觉得它没有特别理由为真，即便想给它找出反例似乎也很难。在这种情况下，如果最终找到了一个反例，它可能已经失去了一部分"反"的意味。我的印象是，非 sofic 群的构造就属于这一类。文献中已有好几个关于如何构造这样一个群的建议，而且我不认为有许多（甚至有任何？）专家强烈相信所有群都是 sofic。所以，说"OpenAI 提出了第一个非 sofic 群的例子"，比说"OpenAI 找到了 sofic 猜想的反例"要自然得多（尽管他们论文的那一节标题就是"sofic 猜想的一个反例"）。

类似地，在我看来，多重色 Ramsey 数的新下界，更像是一个例子，而非反例。我想相当多的人相信这个界应当是指数级的，所以对他们来说它是一个反例；但其他人（包括我自己）则对其持更中立的态度。事实上，我曾经（很久以前）在一个等价表述下研究过这个问题，那问的是：如果你想把这些图的并集做成完全图 K_n，那么在 n 个顶点上需要多少个三角形无关图。如果取二分图，那么容易看出需要 $\log_2 n$ 个；但如果你观察到"一个完全 5 分图可以写成两个三角形无关子图的并集"，那么这个界就可以改进，因此就有可能把完全图写成 $2\log_5 n$ 个三角形无关

图的并集。接下来自然会想做得更好，用密度较低、但以无界色数来弥补的三角形无关图——这是希望使用亚对数数量的图的一个必要条件，而这又等价于证明 $R(3,3,\ldots,3)$ 的超指数下界。所有这些都在说：当我研究这个问题时，我的努力都集中在了后来被证明是正确方向上，所以对我来说，OpenAI 找到的是一个我（弱弱地）预期的例子，而不是反例。

这把我们引向何方?

我想为下面几个事实的合取寻找一个连贯的解释。

- 由 LLM 证明的那些最引人注目的数学结果，往往倾向于被我们归类为例子或反例；其中，反例大略上就是"否定了我们预期为真之陈述的存在性陈述"。
- 许多陈述，当我们通常把它们想成全称陈述时，却可以被表述为存在陈述，反之亦然；因此我们认为什么是例子，既取决于一个陈述的数学语境，也取决于它的逻辑形式。
- LLM 在证明全称陈述方面也相当不错：只是它们所证明的那些、我们会当作"定理"的最强结论，大多没有达到我们会当作"反例"的最强结论的水平。

鉴于这些事实，看来 LLM 所擅长的其实是别的什么东西，而"擅长于此"的一个结果，恰好就是它们擅长于那种我们通常会归类为"要求找一个非平凡例子"的存在性问题。

让我们来考虑两件我们可以确信 LLM 擅长的事。其一是它们知道很多数学知识：如果一个难题能通过相当标准的论证来解决，那么 LLM 极有可能找到并运用那个论证。其二是 LLM 仅仅凭借其作为一台计算机就拥有的能力：它能以极快的速度工作（至少相比人类而言），因此它能够承担"在找到解法之前对一个问题做大量失败尝试"的成本。

即便完全不去看 LLM 实际解决了什么，人们也会猜测：这两个特征会导致它们拥有一种与人类数学家略有不同的风格。粗略地说，当证明寻找过程带有更多概率性成分时，LLM 会占上风：它们擅长那些"最佳方法就是尝试一大堆思路（未必特别新颖），直到某一刻走运"的题目。另一方面，人类（就目前而言）会更擅长找出那些更"出人意料"、更"概念性"的论证——那种需要不断深入挖掘一个问题、直到解法自行显现的方法。（很难确切说这指的是什么，但我希望任何读过此文的资深研究者都能明白我在说什么。）

这引出了两个问题：上述猜测是否与我们所观察到的现实相符；以及，是否有理由认为，我暂时称之为"LLM 做数学风格"的东西，会自然地导致 LLM 发现针对长期猜想的好几个反例（或只是例子），即便这远非它们能力所限。

对这两个问题，我并不假装有科学上的答案；但专家们对 ChatGPT 找到的几个非凡解的反应，确实在某种程度上支持"LLM 更像是一种'尝试大量东西直到走运'的工作方式"这一想法。人们常常似乎会这样反应："起初我很惊讶这个问题竟然被解决了，但仔细一看，我意识到这个方法其实并没有那么新颖，是那种只要给对人一个小小的提示、一位有足够水平的专家人类就很容易想出来的方法。"

对于第二个问题——LLM 的风格是否很适合找（反）例子——我认为情况不太明朗，因为在搜索反例时有多种方式，其中有些比我描述的那种风格更契合。这里给出几种一般方法。（我不声称这个清单是穷尽的。）

- **寻找现成的例子。** 这里，你有一个相当标准的例子库存，你只是把它们一个一个拿出来试，看看是否有一个不能满足给定陈述。例如，Ryan O'Donnell 在他那本关于布尔函数分析的精彩著作结尾给出一些建议，其中之一是："如果你对布尔函数有个猜想，就在独裁函数（dictator）、多数函数、奇偶校验、部族函数（tribes）上测试它（也许还要试试 3 的递归多数）。如果它对这些函数都成立，那它很可能成立。"
- **从基本例子和标准构造方法搭建一个例子。** 例如对一个代数问题，你可能从一些标准例子出发，然后取它们的积、商或极限。
- **大量使用元变量（metavariable）。**"元变量"这个词来自计算机科学，特别是自动定理证明；它对应的数学实践，就是写"其中 x 留待稍后选择"（此时 x 就是元变量）。在论文里，我们通常只在相当简单的情况下这么做，比如需要选择一个足够小、以便后续论证成立的数 $\epsilon > 0$。但当我们在搜索一个满足某性质 Q（这很可能是更简单性质 $Q_1, \ldots, Q_k$ 的合取）的对象 x 的例子时，先把 x 完全指定、然后再检查它是否满足 Q，往往并不是好策略。相反，更有成效的几乎恰恰相反：我们起初对 x 几乎什么也不说，直接开始证明它满足 Q。在此过程中，我们发现需要 x 满足一个性质 P_1。如果幸运，我们能很漂亮地描述一类非常一般的、满足 P_1 的对象 x。例如，我们也许能找到一个带参数类：我们确定某个函数 f，并证明对每一种特定类型的 y，$f(y)$ 都满足 P_1。于是问题就归结为：找一个 y，使得 $Q(x)$ 成立，而这是原问题的一个更具体版本。要最终找到例子，这个过程可能会有很多次迭代，或是这个过程与其他过程混用。
- **尝试证明相反的命题。** 如果想找满足 $Q(x)$ 的 x，那么先尝试证明 $\forall x\ \neg Q(x)$ 这个陈述，往往出人意料地有用。其理由在于：使用我们证明某

事的标准方法，我们最终可能找到起决定性作用的关键引理——即找到一个中间性质 R，它以一种非平凡方式推出 $\neg Q$，从而把问题 $\forall x\ \neg Q(x)$ 归结为 $\forall x\ R(x)$。再反过来看，找到 R 的反例很可能比找到 $\neg Q$ 的反例（也就是一个满足 Q 的例子）更容易。当然，并不保证 R 的反例会是 Q 的一个例子，但有时我们很走运，它恰好是。更常见的是，我们可以利用上一个方法的思路——注意，一个 Q 的例子至少应满足"不是 R 的例子"这一必要条件——从而尝试描述一类"不满足 R"的一般对象，以此简化问题。

- **逐次逼近。** 有时，当我们在寻找满足 $Q(x)$ 的 x 的例子时，我们会写下某个中等可信度的猜测 x_0，并非因为我们觉得它有机会成功（如果那样，我们就在用第一个策略了），而是因为我们希望：如果 x_0 不满足 Q，我们就能诊断出哪里出了错，并指定一个没有这个缺陷的新猜测 x_1。同样，这一策略可以被迭代，或与一个或多个其他策略结合。
- **"直接做"证明（just-do-it proofs）。** 有时我们需要 x 满足无穷多个性质 $Q_1, Q_2, \ldots$，而其中每一个单独看来都相当容易满足。在这种情况下，我们常常"一点点地"归纳地"建造" x，保证在这个过程的第 i 步，无论建造过程如何继续，x 都会满足 Q_i。
- **挑一个随机例子。** 常常很难给出一个满足 Q 的 x 的显式例子，但存在一个自然的概率分布，使得可以证明：若从这个分布中随机选取 x，则以高概率（或至少非零概率）它会满足 Q。
- **挑一个一般（generic）例子。** 在更无限的情境中，给出一个满足 Q 的 x 的显式例子同样可能很难，但我们可以证明：那些不满足 Q 的 x 构成的集合测度为零，或是个贫集（meagre set），或以其他某种方式很小。

没有特别理由认为 LLM 会对上述每种方法都同样擅长。所以我们观察到的可能并非"LLM 具有找例子的特殊能力"，而更像是"它们特别擅长以某种方式找例子（和证明）"。看看上面那些技巧，人们或许会想象：它们非常适合去核实现成例子、找"直接做"证明（因为这是一种相当标准、在其训练数据里有大量实例的方法）、使用概率方法（除非，像常常发生的那样，需要显著的新思想来证明那些概率能成立）、以及挑选一般的例子。上面描述的其他三种方法——使用元变量、尝试证明相反的命题、以及使用逐次逼近——则需要更多"判断自己所走的路线是否可能有效"的能力。在这里，人类至少有时似乎可能占优，只是命题需要满足的条件相当苛刻：需要一个例子，位于一棵极其巨大的搜索树的某个叶子处——大到无法用"中等数学能力＋暴力计算"来搜索——但它又能被一位嗅觉足够敏锐的数学家找到，使他能大幅地剪掉这棵搜索树。

为什么 LLM 就没有那种"嗅觉"呢？我并不排除"嗅觉"是 LLM 训练方式的一种涌现性质，也不排除一两年内它们会拥有与人类同等程度的嗅觉。但就目前而言，在我与 ChatGPT 的互动中，我有一个明确的印象：它们还没完全到达那一步。当我与 5.6 Pro 讨论一个开放问题时，它常会给出一些听起来很有希望、但在我仔细思考后就显得希望不大的方法。它们也常常会用这样的话来结束回答："我未能回答你提出的问题，但我已设法把它归结为下面这个更窄、更精确的问题"——这听起来非常有希望，直到这种情况发生了五次却没有任何明显进展。它们将如何在这方面变得更好，这一点并不完全明显，因为它们的训练数据里不会有满满当当的"在尝试解决问题时哪些方向有成效、哪些没有"的例子：它们通常只会看到被整理干净的证明，而这些证明隐藏了发现者的思维过程。当然，人类数学家也无法从别的数学家的经验中学到太多"如何做研究"，然而我们不知怎地还是把这门本事捡了起来。但对人类来说情况有一点不同：我们学到的很多，都是靠"做"而不是"模仿"。

"嗅觉"并非当 LLM 规模扩大时自然涌现的性质，这还有另一个原因：如果 LLM 大量依赖它们广博的知识、又能否担得起比人类多得多的暴力搜索，那么它们就会缺乏人类那种"无情地剪掉搜索树"的动机。可以想见的是，它们迄今的成功，是用那些对人类来说会被视为极端低效的方法取得的；但正因为它们拥有更快的速度和更广的知识，这些方法会带来的组合爆炸，其效应至今还未显现出来。

尝试在实验上检验这一点会很有意思，但也很困难，因为如果一个 LLM 提出了某种"看起来只可能来自对某个问题的'深度思考'"的想法，我们永远无法确定它真的进行了那种深度思考，还是它只是在文献中找到了一篇现成的典范论证——换句话说，是借用了某位人类数学家的深度思考。用那些不如最新模型强大、且在某种程度上被屏蔽了数学文献的模型来检验，可能会更容易些（但仍不容易）：可以给它们一套精心设计的问题，然后看看 LLM 能解决的那些问题是否具有某些共同特征。

这看起来可能像是我在拼命抓住"人类还能在相当一段时间内对数学发现作出有意义贡献"的希望不放；虽然我确实希望如此，但我并没有做任何形如"LLM 永远无法做 X"的断言。我认为它们很可能会做到，而且考虑到过去三年进展的步伐，这很可能很快就会发生。但我确实认为，LLM 可能需要跨过一道坎，而且至少有可能，这道坎不会像之前那些坎那么好跨。

在这个意义上，看看"不同的奖励结构是否会导致 LLM 能解决不同类型的题目"，也会很有意思。例如，如果在训练期间，一个 LLM（或其他某种机器学习系统）不仅在做出一道解时得到奖励，而且在探索了太多死胡同、或"作弊"（从文献里抄答案）时受到惩罚，那么也许它就会被激励着以更"像人"的方式去进行研究过程，从而在那些它尚未产生重大影响的问题类别上取得更好的结果。

如果这道坎被跨过——无论是靠纯粹扩大规模，还是靠某种更深思熟虑的方法——那么要准确知道它在何时被跨过，将相当困难，因为正如刚才提到的，一个似乎非常原创、非常出人意料的想法，可能只是潜伏在某个 LLM 的训练数据里。但我会有信心说它已被跨过的情况是：一个 LLM 提出了一条让我的惊讶程度不亚于 2016 年 cap-set 问题被解决时的证明——先前的已知最佳界被彻底超越，所用方法与我想过的任何尝试都截然不同，之后还掀起一阵热潮，人们竞相去理解这一美妙新技巧的威力。

结论

动笔之时，我并不十分确定自己会走向何处；而当我写到结尾时，我觉得自己的主要结论算不上特别新颖或出人意料，但我希望通往它们的路径多少有点趣味。我提出的主要论点如下。

- "找一个例子"实际上并不等于"证明一个以存在量词开头的陈述"。
- 如果说当前模型确实特别擅长找例子，那恐怕不是因为它们对存在性陈述有特别偏好，而更多的是因为：寻找某类例子时适用的那些"发现证明的方法"，恰好契合了 LLM 显而易见的长处——广博的知识，以及探索搜索树中那些"人类会判定为成功概率很低"的许多路径的能力。
- LLM 很可能将继续飞速进步。然而，如果与（至少是我的）预期相反，结果证明存在某一类余留问题（或其他数学活动），人类在其中继续保持一段时间的优势，那么那些问题很可能是"人类那种神秘的、剪掉证明发现搜索树的能力"特别占优的问题：也就是说，那些搜索树既深、分支又多的问题——若没有严格的剪枝，即便对计算机来说搜索也是不可行的。
- 一个表明 LLM 已在更广泛的问题类别上达到人类水平的良好信号是：它们开始用那些——像许多最优秀的人类数学一样——既新颖又出人意料、但在事后看来却显得美丽而自然的方法来证明定理。这些方法还应是"难以碰巧踩到"的。很难确切说怎样的证明才算数，但我想，我们会一眼认出的。

—— 全文完 ——

本文由 https://gowers.wordpress.com/2026/08/12/what-sort-of-maths-are-llms-good-at/ 的中文翻译整理而成。文中所有数学公式均按其原文 LaTeX 渲染。