大模型机理研究的方法论谱系

陆品燕
上海财经大学

被已有成果掩盖的基础问题

大模型已经取得了令人震撼的成功。从语言理解到多模态处理，从代码生成到数学推理，它在越来越多的任务上展现出超预期的能力。然而，越是成功，我们越容易忽视一个根本问题：**我们真的理解大模型吗**？

如果对当下的大模型研究方向做一个粗略归纳，笔者倾向于把它分成 4 个主要方向：**能力更强**、**应用更广**、**效率更高**、**基础更牢**。这 4 个方向都重要，但它们背后对“大模型是什么”的理解其实并不相同。

“能力更强”更多把大模型看成一个**系统工程**。研究者通过算力、数据、架构、训练策略和对齐方法的系统优化，持续推动模型能力增强。这一方向的核心目标很明确：让模型更强、更稳、更通用。

“应用更广”更多把大模型看成一个**工具**。当一个强大的通用模型已经存在，接下来的问题就变成：如何把它更好地接入真实任务、真实行业和真实流程？智能体（Agent）、工作流编排、行业垂直应用，本质上都属于这一视角。

“效率更高”则把大模型看成一种新的**计算负载**。在这里，研究者未必关心模型“在想什么”，而更关心“它如何更快地算”。训练和推理中的并行策略、内存优化、低精度计算、算子融合，都是这个方向的代表问题。

而“基础更牢”则不同，它把大模型看成一个需要被理解的**研究对象**。研究者追问的不是“它能不能更强”，而是“它为什么会强”“它到底学到了什么”“它的能力边界在哪里”“它的内部机制是否可解释”。

DOI：10.11991/cccf.202608004
通信作者：陆品燕，E-mail: lu.pinyan@mail.shufe.edu.cn

从短期看，能力、应用和效率更容易产生直接回报；但从长期看，真正决定这门学科是否成熟的，恰恰是第 4 个方向。因为如果一个系统越来越强，而人类始终停留在“会用但不懂”的状态，那么系统就更像是一套经验主义工程，而不是一门可积累、可验证、可传承的科学。

为什么“基础更牢”正在变得更加迫切

今天讨论大模型机理研究，并不是因为“理解”比“应用”更高贵，而是因为大模型的发展已经逼迫我们重新回答一些最基本的问题。

大模型仍然是高度黑箱化的系统　我们知道它会说话、会推理、会调用工具、会涌现出一些此前未被显式编程的能力，但对于这些能力究竟如何形成，我们并没有一个统一、令人信服的解释。哪怕是“为什么大模型可以生成如此流畅的自然语言”这样一个今天看似平常的问题，严格地说，我们也尚未真正回答清楚。

大模型正在进入高风险和高影响场景　当大模型被用于教育、医疗、金融、治理乃至科学研究时，仅仅依赖经验调参和外部评测是不够的。系统越重要，社会越要求它具有更强的可解释性、可预测性与可控制性。

机理研究是知识积累的前提　工程方法可以不断提高性能，但如果缺乏对原理的理解，我们就很难判断哪些经验是偶然有效，哪些规律是稳定可迁移的。没有机理，就难有边界；没有边界，就难有真正可靠的优化。

从这个意义上说，大模型机理研究首先是一种**探究真相导向的研究**。它的目标不是立刻带来某项可部

署的功能，而是探索“这件事情究竟是怎么回事”。至于这种理解是否有用，往往是它的副产品——但通常也是极其重要的副产品。因为更深的理解，最终总会反馈到更好的训练、更高的效率、更可靠的安全性和更稳健的应用设计之中。

大模型究竟是什么：一个被低估的理论问题

如果我们希望从根本上理解大模型，首先必须回答一个看似简单、实际上极难的问题：**大模型（在数学上）究竟是什么**？

目前不同研究社群对此给出的答案并不一致。做逼近理论的人，往往把它看成一个**函数逼近器**；做生成建模的人，更强调它是一个**概率采样机**；从表示学习和信息论的角度看，它又像是一个**信息压缩器**；而当我们讨论其推理、递归、工具使用和程序执行能力时，它又越来越像一个**通用图灵机**。

问题在于，这 3 种定义都没有错，但又都只抓住了大模型的一个侧面。如果只把大模型看成函数逼近器，就可能低估其在序列生成与交互过程中的动态性；如果只把它看成概率采样机，就可能忽视其内部结构化表示与算法性能；如果只强调信息压缩，就容易忽略训练动力学与任务执行机制；如果只把它看成通用图灵机，又可能遮蔽其统计学习本质。

因此，笔者更倾向于把大模型理解为一个“**四位一体**”**的对象**：它同时是函数逼近器、概率采样机、信息压缩器和通用图灵机。今天许多理论研究之所以彼此难以对话，很大程度上不是因为谁对谁错，而是因为它们分别站在这 4 个侧面中的某一个上讲话。未来大模型理论研究一个真正关键的任务，也许并不是继续在单一视角内打磨得更精细，而是尝试找到一种能够容纳这些不同侧面的统一数学表述。

从数学到心理学：理解大模型的四层方法论

如果说“大模型是什么”是研究对象的问题，那么“我们如何理解大模型”就是研究方法论的问题。对这一问题，我更愿意用一个连续谱系来描述：**大模型的数学、物理学、生物学和心理学**。这里提到的连续谱系主要是说这 4 个学科的研究方法论和这些学科的具体知识体系，特别是和物理学、生物学、心理学的具体研究内容没有关系。

大模型的数学：在理想化世界中刻画边界

所谓“大模型的数学”，不是指泛泛地使用数学工具，而是指在理想化假设下，给出严格定义、证明定理并刻画能力边界。

这一研究方向已经有一些重要成果。比如，神经网络的通用逼近定理告诉我们，它在表达上足以逼近任意函数。关于 Transformer 的理论研究则进一步讨论了大模型的计算能力：如果只看固定层数、一次前向输出的电路模型，其能力落在某些受限的并行复杂性 TC^0 类[1] 中；而如果允许思维链（chain of thought, CoT）递归展开，那么它可以到达图灵机的计算能力。

数学方法的优点是清晰、严格、可验证，缺点是它往往只能研究经过强烈简化后的对象。因此，数学给我们的通常不是完整现实，而是**边界、结构与原则**。

大模型的物理学：在受控实验中提炼规律

如果数学的代价是理想化，那么物理学的方法则试图在可控实验中获得更接近现实的规律。物理方法论得到的结论是“定律”而不是“定理”。

大模型领域已经出现了一些具有物理学意味的结果。例如缩放定律（scaling law）就是大模型最重要的定律，它通过大规模实验拟合模型性能与参数量、数据量、计算量之间的关系；双下降（double descent）和顿悟（grokking）则揭示了神经网络训练过程中一些反直觉的泛化现象。这些结果未必是严格证明的“定理”，但它们像自然科学中的经验定律一样，能够稳定地被观测、复现并指导实践。

这一方法论特别适合使用**形式语言**和**合成数据**。因为自然语言环境变量过多、噪声复杂、数据分布难以

[1] TC^0 类是计算复杂性与电路设计领域中的概念，是并行计算 / 电路复杂度中的经典复杂性类。

控制，很难进行真正严格的控制变量实验。相比之下，数学运算、形式文法、符号推理等任务虽然简单，却是非常理想的研究实验沙盒（playground）。它们让我们得以从头训练模型、精细控制数据分布、系统调节噪声水平，并由此观察学习、记忆和泛化之间的关系。

这些发现说明，受控实验并不是“玩具问题”，而是在为大模型建立一种实验科学。它们让我们有机会分离变量、复现实验、累积规律。尤其重要的是，那些传统机器学习并不太关心的形式化任务，在大模型时代反而可能变得重要，因为如果大模型真的是一种通用计算系统，它就不应只会统计模式匹配，也应能学习逻辑、递归和算法结构。

大模型的生物学：在真实模型中做“解剖学”

再进一步，我们会进入“大模型的生物学”视角。这里不再满足于小模型、合成数据和外部行为，而是试图像研究生物神经系统那样，直接“解剖”一个真实的大模型，分析其内部表征、功能模块与回路结构。

这个方法论当前最有代表性的发展方向是**机制可解释性**。这一思路本质上是一种逆向工程：模型已经训练完成，研究者再去分析其参数、激活和中间状态，理解某些语义概念、行为模式或推理步骤是如何在网络内部被编码和传递的。

围绕机制可解释性，叠加表征、稀疏自编码器等方法受到广泛关注。这些方法的一个核心设想是：模型内部许多概念并不是“一神经元一语义”式地存储，而是以叠加、分布式的方式存在；如果把激活映射到更高维且更稀疏的特征空间中，就有可能分离出更具解释性的“特征方向”。研究表明，通过剖析已经完成训练的大模型，确实可以发现相当明确的语义方向，甚至通过人工调节这些方向，可以在模型外部观察行为的系统变化。

这些研究非常重要，也让我们一定程度上打开了大模型的黑盒。但笔者相信这不能等同于机制可解释性的全部。机制可解释性研究至少面临 2 个值得认真追问的问题：第一，**线性表征是否足以表达模型内部真实的结构**？第二，**可解释性是否必然等同于稀疏性**？当前主流方法很重要，但不应成为唯一范式。

如果说大模型的数学提供了边界，大模型的物理学提供了规律，那么大模型的生物学则提供了**结构**。“结构”让我们第一次有机会把“黑箱”局部变成“半透明系统”。

大模型的心理学：从黑盒行为理解认知边界

还有一条路径常常被低估，那就是“大模型的心理学”。它的特点是**不看内部，只看输入输出**。这种方法看似“最不深入”，但实际上对大模型最早、最重要的一些发现，恰恰来自这种黑盒观察。

提示工程（prompt engineering, PE）、上下文学习（in-context learning, ICL）、CoT 等方法，最初都不是在理论推导中被预言出来的，甚至也不是研究人员训练出第一代大模型的时候事先设计的特征，而是在人们不断使用模型的过程中被“发现”的。换句话说，用户与模型交互的实践，本身就是一种认知实验。研究者通过提问方式、示例布局、推理诱导等外部操控，不断勾勒出模型的行为边界。

这类研究提醒我们：黑盒方法并不“浅”，它只是站在另一个层面理解系统。尤其当我们面对最前沿、最封闭、最难以被完全解剖的模型时，心理学方式的外部行为分析依然是不可替代的。

4 种方法论不是割裂的，而是一条连续谱系

大模型机理研究之所以经常遭到质疑，一个重要原因是：无论你使用哪一种方法论，都会被其他方法论的标准挑战。

做数学的人会被问：你的假设离真实模型这么远，证明出来有什么意义？做实验的人会被问：你的规律只在小模型和合成数据上总结，为什么可以泛化？做机制可解释性的人会被问：你只是用一个实验现象解释了另一个实验现象，凭什么算机理？做黑盒评测的人会被问：你连模型内部都没看，凭什么谈理解？

这些质疑之所以总是成立，不是因为这些研究没有价值，而是因为它们本身只覆盖了真实问题的一部分。真正合理的理解方式，不是让某一种方法单独承担全部任务，而是把它们放在一条连续谱系中来看。

数学与物理学之间，是“定理—定律”的互动。理

想化证明需要在更现实的实验环境中得到验证；反过来，实验中反复出现的规律，也会促使我们在适当假设下给出数学证明。

物理学与生物学之间，是“小模型规律—真实模型现象”的互动。受控实验中发现的规律，需要输送到真实大模型里去检验是否仍然存在；而在真实模型里观察到的复杂现象，也需要被抽取到更简单的系统中复现、归因和理解。

生物学与心理学之间，是“内部结构—外部行为”的互动。内部发现的特征或回路，只有在操控后真的引发行为变化时，才更有说服力；而外部行为模式的稳定差异，也常常能反过来引导我们定位其内部机制。

因此，这 4 种方法并不是 4 个彼此竞争的阵营，而是一套层层递进、相互校验的知识生产机制。它们共同构成了从最抽象的形式化证明，到最贴近现实的行为理解之间的一条完整通道。

未来的大模型基础研究路径规划

基于上述讨论，针对未来的大模型基础研究应当往哪走，笔者提出以下几点更明确的判断。

第一，大模型基础研究不应再被视为边缘工作，而应成为下一阶段人工智能发展的核心议题。 如果说过去几年是“把模型做出来”的阶段，那么未来几年更重要的问题将是“把模型解释清楚”。性能提升会继续发生，但真正决定人工智能是否走向成熟科学的，将是我们能否建立稳定、系统、可积累的理论与实验框架。

第二，大模型机理研究不能依赖单一方法。 只做数学，容易脱离现实；只做黑盒评测，容易流于现象；只做结构解剖，也可能陷入局部解释。未来更有价值的工作，应该是能够在不同层次之间来回穿梭：从现象中提出问题，在小模型中控制复现，在理论上抽象证明，再回到真实模型中验证边界。

第三，大模型基础研究既是好奇心驱动的，也是安全性驱动的。 我们研究机理，首先是因为这个系统足够神奇、足够陌生、足够值得理解；但与此同时，越强大的系统越需要被理解。好奇心与安全性治理需求在这里并不冲突，而是高度一致的。

大模型是被人类发明的，还是被人类发现的？

这是一个带有哲学意味，但笔者认为非常重要的问题。从起源上讲，大模型当然是人类发明的。它的架构、训练流程、硬件系统和应用生态，都是人为设计与工程堆叠的结果。自然界并不存在一个现成的“GPT”。

但从理解程度上讲，我们又不得不承认：人类对大模型的认识，更像是在**发现**它，而不是完全“掌控”它。它的很多关键能力并不是在设计之初就被清晰预见的，而是在训练之后、使用过程中逐渐显露出来的。我们像发现一种新自然现象那样，去观察它、命名它，复现实验、归纳规律、提出理论。

正因为如此，大模型研究今天面临的，不只是技术问题，更是方法论问题。我们也许已经“制造”出了这一对象，但距离真正“理解”它，还相当遥远。

结束语

大模型的发展正在推动人工智能从“做系统”走向“做科学”。在这一转变中，能力更强、应用更广、效率更高仍将是持续推进的方向，但“基础更牢”应当成为下一阶段更值得投入的研究方向。理解大模型，不能只依赖单一视角，而需要一条由数学、物理学、生物学和心理学共同组成的方法论谱系。只有当我们能够在理想化证明、受控实验、结构解剖和黑盒行为之间建立稳定联系时，大模型机理研究才可能真正走向深入，也才可能为可信人工智能奠定更坚实的基础。 ■

陆品燕
CCF 杰出会员、理论计算机科学专委会委员。上海财经大学“长江学者”特聘教授，计算机与人工智能学院创院院长。主要研究方向为理论计算机科学及其与其他学科的交叉、大模型算法和机理。
lu.pinyan@mail.shufe.edu.cn

Continuous Methodological Spectrum for Mechanistic Research on Large Models

Pinyan Lu
Shanghai University of Finance and Economics

Abstract: While the rapid advancement of Large Models has driven unprecedented successes in capabilities, efficiency, and real-world applications, a fundamental question remains: do we truly understand how they work? Establishing a solid theoretical foundation for these highly black-boxed systems is no longer just an academic curiosity, but a crucial prerequisite for developing safe, interpretable, and controllable AI. This article argues that mathematically, large models are multifaceted entities, acting simultaneously as function approximators, probability samplers, information compressors, and universal Turing machines. To fully capture this complexity, the study of AI mechanisms must transcend isolated viewpoints. The author proposes a continuous methodological spectrum that draws analogies from four fundamental scientific disciplines:

- Mathematics delineates the theoretical boundaries and absolute capacities of models under idealized assumptions through rigorous proofs.
- Physics extracts empirical laws (e.g., scaling laws) and statistical patterns via controlled experiments and synthetic data.
- Biology conducts mechanistic "anatomy" to decode internal representations, neural circuits, and functional modules within real-world models.
- Psychology explores cognitive boundaries and emergent behaviors through black-box testing, prompting, and input-output interactions.

These four paradigms are not mutually exclusive but deeply interconnected and mutually reinforcing. By bridging abstract theoretical proofs with empirical behavioral observations, this continuous spectrum provides a holistic framework for AI mechanism research. Ultimately, this paradigm shift will transition the field of large models from empirical engineering into a rigorous, cumulative scientific discipline.
Keywords: large models; mechanism study; methodological spectrum; mechanistic interpretability; science of AI

摘　要: 近年来，大模型研究迅速成为人工智能领域的中心议题。围绕这一技术体系，学术界与产业界大体沿着 4 个方向发力：能力更强、应用更广、效率更高、基础更牢。前 2 个方向直接推动模型能力进化和场景落地，第 3 个方向支撑大规模训练与推理的成本优化，而第 4 个方向——基础更牢——则关乎我们能否真正理解大模型、解释其行为、刻画其边界，并最终建立可信、可控、可持续发展的人工智能科学。本文认为，大模型机理研究不能被理解为单一方法的工作，而应被视为一条由"数学—物理学—生物学—心理学"构成的连续方法论谱系：数学方法论强调理想化假设下的定理证明与边界刻画；物理学方法论强调受控实验、统计规律与唯象定律；生物学方法论关注结构解剖、内部回路与功能定位；心理学方法论则通过黑盒测试与行为诱导理解模型的认知边界。4 种路径既相互区分，又彼此联通，共同构成理解大模型的完整框架。本文进一步提出，大模型在数学上不是一个单一对象，而是函数逼近器、概率采样机、信息压缩器与通用图灵机的"四位一体"。因此，未来的大模型基础研究应超越单一视角，建立跨方法、跨尺度、跨层次的研究范式。
关键词: 大模型；机理研究；方法论谱系；机制可解释性；人工智能科学

中图分类号: TP18

中文引用格式: 陆品燕. 大模型机理研究的方法论谱系 [J]. 计算, 2026, 2(8): 15−19.

英文引用格式: Pinyan Lu. Continuous Methodological Spectrum for Mechanistic Research on Large Models[J]. *Computing Magazine of the CCF*, 2026, 2(8): 15−19.