一个用于自主科学发现的闭环框架：连接自然语言、形式逻辑与物理现实的初步构想

肖逸峰*
基础数学与物理学院, 中国科学院大学杭州高等研究院
杭州, 中国

免责声明：本文为作者起草的初步研究构想，

不代表作者所属机构的官方观点。

October 2, 2025

Abstract

当前的深度学习模型在面对需要严谨符号推理和分布外泛化的科学发现任务时，表现出结构性不足。本文提出了一个新颖的、基于模块化功能分解的计算框架，旨在模拟并超越人类的科学探索过程。该框架的核心是一个统一的语言引擎，它以类似混合专家（Mixture-of-Experts, MoE）的模式运行，动态调用不同的功能模块：（1）自然语言定义模块，用于将人类直觉和新概念安全地引入形式系统；（2）演绎引擎，负责在已建立的公理体系内进行严谨的逐步推导；（3）公理创新模块，在现有理论无法解决问题时，通过修改或提出新公理来进行创造性探索。整个系统的每一步推理，都受到一个外部符号逻辑校验模块的严格约束，保证了过程的合法性。最终，当系统从纯粹的数学探索扩展到物理世界时，一个物理现实校验模块将理论预测与实验数据进行比对，提供关键的“接地”信号。此框架并非一个固定的闭环，而是一个可演化的、通过可验证信号驱动的、能够自主生成并检验新理论体系的智能体。

关键词：自主科学发现，神经符号 AI，混合专家模型 (MoE)，强化学习，公理化方法，可解释 AI。

1 引言：超越统计模式匹配

2 引言：超越统计模式匹配

2.1 泛化陷阱：现代人工智能的固有缺陷

当代人工智能，尤其是基于深度学习和大规模预训练模型（LLM）的系统，在模式识别和知识整合方面取得了空前的成就。然而，这些系统在执行科学发现这一类任务时，暴露出了其结构性的固有缺陷：即著名的分布外泛化陷阱（**Out-of-Distribution, OOD Generalization Gap**）。

传统的深度学习本质上是一种高维插值技术。模型擅长于对训练集（即数据分布）进行统计近似。一旦面对分布之外的数据或需要构建全新的理论框架时，其预测或输出将缺乏可靠的逻辑保障 [**?**]。对于科学发现而言，每一次理论突破都意味着一次对已知知识边界的超越，这恰恰要求模型在严格的“分布外”环境中运行。例如，一个在牛顿力学框架内训练出的模型，无法在不违反其核心假设的前提下，自主地导出相对论的修正。因此，当前 AI 的范式仅限于高效地解构和应用现有知识，而难以完成真正的理论创造。

*电子邮箱: xiaoyifeng25@mails.ucas.ac.cn

2.2 人类科学发现模式的启示

与此形成鲜明对比的是，人类的科学实践是一个迭代、可验证和自我修正的闭环过程，这为我们设计新一代 AI 提供了蓝图。历史上任何重大的科学飞跃，都包含以下几个关键步骤：

1. 直觉与创新（假说提出）：科学家使用自然语言进行思考，提出大胆的、甚至有悖常理的新公理或假 Calyce（牛顿看到苹果落地）。
2. 数学家的验证（逻辑校验）：新假说被转化为形式符号，经过严谨的数学或逻辑推理（如保守扩展检查），确保其在逻辑上是自洽的。
3. 实验家的验证（现实校验）：通过构建实验或计算模型（如可微分仿真），将理论的预测与现实观测数据进行比对，根据误差 ε 给出最终的判断。
4. 迭代与演化：根据实验结果的反馈（强化学习信号），理论家修正或推翻原有公理，开启新一轮的迭代。

我们认为，人工智能需要从这种人类的工作模式中汲取灵感，将可验证性和创新性作为核心驱动力。现有的尝试，如物理信息神经网络（PINNs）[?]，虽然成功地将物理约束纳入了训练，但它们是被动地接收物理定律，而非主动地发现和修正定律。同样，近期在数学发现领域的工作 [?]，虽然展示了 AI 的创造力，但其作用范围仍然局限于纯符号空间，缺乏与物理现实的直接交互和约束。

2.3 本文的贡献与框架概述

针对上述挑战，本文提出了一个全新的、可验证的模块化计算框架。我们的核心贡献在于：

- 提出了一个三位一体的模块化框架：将传统的 AI 模型重构为一个由统一语言引擎（LLM Core）驱动的 MoE 系统，并配以两大外部的、不可协商的“真理源”——符号逻辑校验器和物理现实校验器。
- 建立了从 **NL** 到物理现实的端到端可验证性：首次将数理逻辑中的“保守扩展”理论与实验物理学的“可微分仿真误差”相结合，为 AI 的创新活动提供了双重、硬性的安全带。
- 实现了理论的自主演化：通过强化学习机制，系统能够根据校验器的反馈，自主地在演绎、定义和公理创新之间切换，具备了模拟人类科学家进行理论演化的潜力。

在随后的章节中，我们将详细阐述框架的架构、功能模块的分解、并设计一个思想实验来论证该框架在发现新物理理论时的可行性。

3 一个功能分解的模块化框架

与简单的闭环结构不同，我们提出的框架更接近于一个由统一语言模型驱动的混合专家（MoE）系统。该系统的核心是一个强大的语言引擎，它根据任务需求，激活不同的功能“专家”或操作模式，并与外部的、不可协商的“校验器”进行交互。整体架构如图**??**所示。

3.1 功能模块详解

3.1.1 自然语言定义模块

本模块是系统与人类直觉的接口。其核心动机在于，纯粹的符号系统本身就是另一种需要学习的语言。为了方便、灵活地引入新符号、新公理，本模块允许以自然语言进行描述和定义。例如，用户可以输入：“定义一个新的运算符‘$\oplus$’，它满足交换律和结合律。”

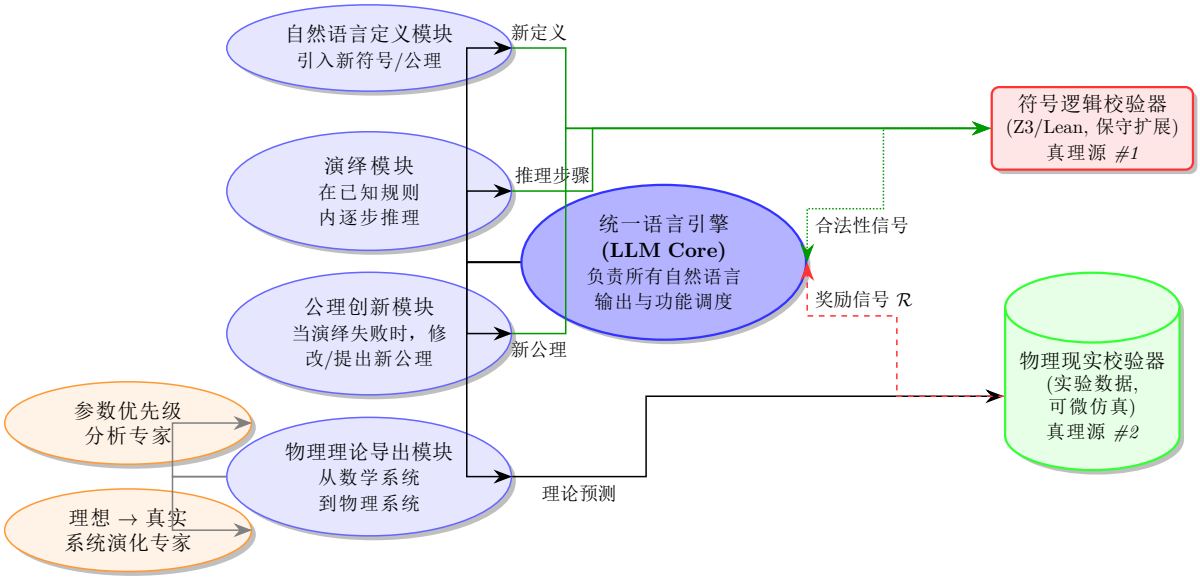

Figure 1: 所提出的模块化框架。其核心是一个统一的语言引擎，它调度不同的功能专家模块。所有专家的输出都必须经过外部“符号逻辑校验器”的审查，而物理理论则最终由“物理现实校验器”进行检验，形成一个可验证的、能够自主演化的系统。

3.1.2 符号逻辑校验模块

这是一个外部的、不可协商的“真理源”。它旨在保证系统每一步推理链条的准确性和可校验性。它不产生新知识，而是对“演绎模块”和“创新模块”生成的每一步推导进行合法性检查。这可以通过与形式验证器（如 Z3, Lean）的接口，以及实现“保守扩展”检查算法来完成。

3.1.3 演绎与创新引擎

这两个功能可以被视为统一语言引擎的两种不同操作模式：

- 演绎模块 (低创造性模式)：此模块是系统的主要“驱动力”。它严格使用已知的公理和定理（包括由定义模块新引入的），进行逻辑导出。它具有明确的目标，例如证明一个给定的猜想。
- 公理创新模块 (高创造性模式)：当演绎模块在已知框架内无法解决问题时，此模块被激活。它旨在通过提出新公理，或对现有公理进行修改（例如，将欧式几何的平行公理修改为球面几何的公理），来探索新的理论体系。其产出必须立即被“符号逻辑校验模块”进行审查。

3.1.4 物理系统接地模块

当系统从纯粹的数学探索扩展到对物理世界的描述时，此模块负责将抽象的理论“接地”。它包含一个关键的“可应用性判据”，例如比较理论预测与实验数据的置信度或统计误差。此模块本身也可以被设计为 MoE 结构，包含如“参数优先级分析专家”、“从理想系统向复杂真实系统演化专家”等，以迭代的方式，从一个简单的理想模型，逐步构建出能够囊括主要现实因素的、更复杂的理论。

4 核心原创性及与相关工作的比较

我们的工作站在多个领域的巨人肩膀上，但提出了一种新颖的综合。

- 与物理信息神经网络（**PINNs**）的比较：在物理知识通知的神经网络领域，其开创性工作由**?** 给出。PINNs 的核心思想是将物理定律（通常是偏微分方程）的残差作为损失函

数的一部分，从而“告知”神经网络去学习一个满足该定律的解。这是一个强大的、用于“解”已知方程的框架。然而，PINNs 是被动地接收物理定律。相比之下，我们的框架旨在发现这些定律。它不是一个求解器，而是一个理论生成器。

- 与神经符号 **AI** 的比较：逻辑与神经网络的集成是神经符号 AI 的核心主题。正如**?** 等先驱在其关于“第三次浪潮 AI”的论述中所展望的，这可以赋予模型严谨的推理能力。现有的神经符号系统通常在一个预先定义的、固定的逻辑体系内工作。我们的工作对此进行了关键的扩展，通过“自然语言定义模块”和“保守扩展检查”，允许逻辑系统本身通过自然语言提议，被动态且安全地进行扩展，从而具备了处理全新概念和理论体系的潜力。

- 与 **AI** 辅助科学发现的比较：近期，以 DeepMind 的工作为代表，AI 在纯粹的数学和算法发现领域取得了惊人的成就 [**?**]。这些工作证明了强化学习在高维、离散的符号空间中进行创造性搜索的可行性。我们的框架借鉴了这一思想，但提供了一个关键的缺失环节：一个双重安全机制。AI 生成的假说不仅要通过内部的逻辑校验，还必须通过与物理世界的交互进行外部校验。这确保了系统的探索不仅是富有创造性的，更是根植于现实、有意义的。

5 一个思想实验：从牛顿定律的雏形开始

为了具体阐明本框架的工作流程，我们设计一个简化的思想实验，其目标是从一组模拟的行星运动数据中“重新发现”牛顿万有引力的一个雏形。

1. 初始状态：系统被给予一组关于行星位置随时间变化的数据 $(x(t), y(t))$，以及基础的数学公理（如微积分、矢量代数）。但它不知道任何关于“力”或“引力”的概念。

2. 演绎失败：系统首先尝试用“演绎模块”，仅使用基础数学公理（例如，匀速直线运动模型）来预测行星轨迹。很快，它会发现预测轨迹与真实数据存在巨大误差，演绎失败。

3. 公理创新：“公理创新模块”被激活。统一语言引擎（LLM）受到“预测与数据不符”的负奖励信号的驱动，开始提出新的假说。

   - 假说 *1 (NL):* “也许有一个指向太阳的‘力’导致了行星的偏转。”

4. 定义与校验：

   - “自然语言定义模块”将“力”定义为一个新的矢量符号 $\vec{F}$。
   - “符号逻辑校验模块”检查该定义，确认其在量纲和数学结构上是合法的。

5. 再次创新与校验：

   - 假说 *2 (NL):* “这个力的大小也许与行星到太阳的距离 r 的平方成反比。”
   - 系统将其形式化为 $|\vec{F}| \propto 1/r^2$。
   - “逻辑校验模块”再次确认其合法性。

6. 物理校验与理论演化：

   - “物理理论导出模块”将这个新的公理体系（基础数学 + 新的引力假说）输入“可微分仿真器”。
   - 仿真器计算出的预测轨迹与真实数据的误差显著减小，系统获得了一个大的正奖励。
   - 这个新的公理——万有引力定律的雏形——被接纳并冻结入知识库。

这个思想实验展示了本框架如何模仿从数据观察到理论提出，再到验证和接纳的科学实践过程，正如牛顿本人在其划时代的著作《自然哲学的数学原理》中所奠定的那样 [**?**]。

6 局限性与未来工作

尽管本构想为自主科学发现提供了一个富有前景的蓝图，但我们必须清醒地认识到其面临的重大挑战和局限性。

- 公理筛选的保守性：我们依赖的“保守扩展”检查，其核心是保证新公理不与旧系统产生矛盾。这在理论演化的早期是优点，但在面对需要彻底推翻旧公理的“范式转移”（如从牛顿力学到相对论）时，可能会成为一个障碍。未来的工作需要探索更高级的“元逻辑”推理机制，以处理理论的非单调演化。
- 物理校验的敏感性与成本：“现实校验模块”的有效性高度依赖于误差阈值 ε 的设定和可微分仿真器的精度。ε 的设置是一个敏感的超参数，而高精度的可微分物理仿真本身就是一个计算成本极高且充满挑战的研究领域。
- 搜索空间的组合爆炸：尽管 LLM 的直觉可以极大地缩小假说的搜索空间，但从自然语言到形式公理的映射，以及新公理的组合，仍然面临着组合爆炸的风险。需要发展更高效的强化学习探索策略。
- 概念“接地”的深度：当前框架主要通过轨迹误差来将理论“接地”。但物理概念（如“场”、“熵”）的内涵远比其数学形式丰富。如何让 AI 真正“理解”这些概念的物理意义，而非仅仅将其作为有效的数学工具，是一个长期的、深刻的哲学与技术挑战。

未来的工作将首先致力于在一个简化的、但物理意义完备的“玩具世界”（如具有特定对称性的粒子系统）中，对该框架的各个模块进行初步的计算实现与验证。