北京大学深圳研究生院
Peking University Shenzhen Graduate School

DEEPSCIENTIST: ADVANCING FRONTIER-PUSHING SCIENTIFIC FINDINGS PROGRESSIVELY

Project: https://ai-researcher.net
Code: https://github.com/ResearAI/DeepScientist

汪静雅

2025.10.15

01 Abstract

研究背景: 以往的人工智能科学家系统虽能生成新颖发现，但往往缺乏聚焦，难以产出能应对人类定义的紧迫挑战且具有科学价值的成果。

研究重点:

- 提出 **DeepScientist** 系统，旨在通过开展以月为单位的、目标导向的完全自主科学发现来克服上述问题。
- 将科学发现形式化为**贝叶斯优化**问题，通过由 **“假设、验证、分析”** 组成的分层评估过程来实现。
- 借助累积的 “发现记忆库”，上述循环能智能平衡新假设的探索与已有成果的利用，有选择地**将最有前景的发现推进到更高保真度的验证级别。**

研究成果:

- 耗费超 20000 个 GPU 小时，系统生成约 5000 个独特科学想法，对其中约 1100 个进行了实验验证。
- 在三项前沿人工智能任务上，最终超越人类设计的最先进（SOTA）方法，分别提升 **183.7%、1.9% 和 7.9%**。
- 首次提供了大规模证据，证明人工智能在科学任务上能取得逐步超越人类 SOTA 的发现，产生了真正推动科学发现前沿的有价值成果。

目前研究现状：

- 科学发现本质上是一个**不断探索和试错**的过程。
- LLM在科学发现中实现了**端到端、全周期的自动化**。
- 在缺乏明确科学目标的情况下，当前的人工智能科学家系统经常陷入**盲目重组现有知识和方法**的陷阱。因此，他们的研究成果在人类评价下往往**缺乏真正的科学价值**。

工作架构：

将科学发现建模为**目标驱动的贝叶斯优化问题**，唯一目标是找到一种最大限度地提高目标性能指标的新方法。

DeepScientist:

- 采用**迭代工作流程**，不断扩展先前研究知识的记忆。巧妙地平衡了挖掘（深化对有前景的高价值方向的调查）和探索（冒险进入未知领域）。
- 通过大规模的并行探索，DeepScientist可以产生创新的假设，并最终通过持续探索**产生有价值的新方法和经过验证的科学发现**。

- 在三大任务上实现**SOTA**（代理故障归因、LLM推理加速和AI文本检测）
- 分配的资源与有价值的科学发现的产出之间存在近**线性关系**。
- 人工智能的探索速度非常快，但其固有的创新成功率非常低。
- 该领域的核心问题不再是“人工智能能创新吗？而是“**我们如何有效地引导其强大但高度耗散的探索过程，以最大限度地提高科学回报？**”

复制与优化

- 这类工作的共同目标是在已有的科学范式内进行工程驱动的优化，在不质疑核心基础假设的前提下改进现有系统。
- DeepScientist不是追求以现有最先进技术的局限性为目标的科学发现，也不是优化当前的最先进技术，而是**通过引入根本不同的方法来建立新的技术体系。**

半自动化科学辅助

- 这些强大的工具仅能解决科学过程中孤立的部分，关键的从失败中学习和探索环节仍由人类完成。
- DeepScientist 是一个自主智能体，能够**管理整个端到端的研究周期**，通过从自身实验中学习来闭合循环，并自主引导研究路径。

自动化科学发现

- 它们的主要局限在于探索策略，往往缺乏植根于某一领域重大挑战的具体科学目标，导致发现缺乏方向，可能被认为缺乏真正的科学价值。
- DeepScientist 是**首个利用闭环、迭代过程来发现超越人类最先进水平方法**的自动化科学发现系统，其探索以目标为导向且由洞察驱动，从识别人类最先进技术的公认局限性开始，然后利用故障归因确保发现既新颖又具有科学意义。

MODELING SCIENTIFIC DISCOVERY AS AN OPTIMIZATION PROBLEM

自动化科学发现的根本目标是自主识别出能在特定科学领域带来重大进展的新方法。这一过程可被形式化地概念化为在一个庞大且无结构的可能性空间中寻找最优解。

$$I^* = arg\,max_{I \in \mathcal{I}} f(I)$$

其中：

所有可能的候选研究方法构成的空间为$\mathcal{I}$

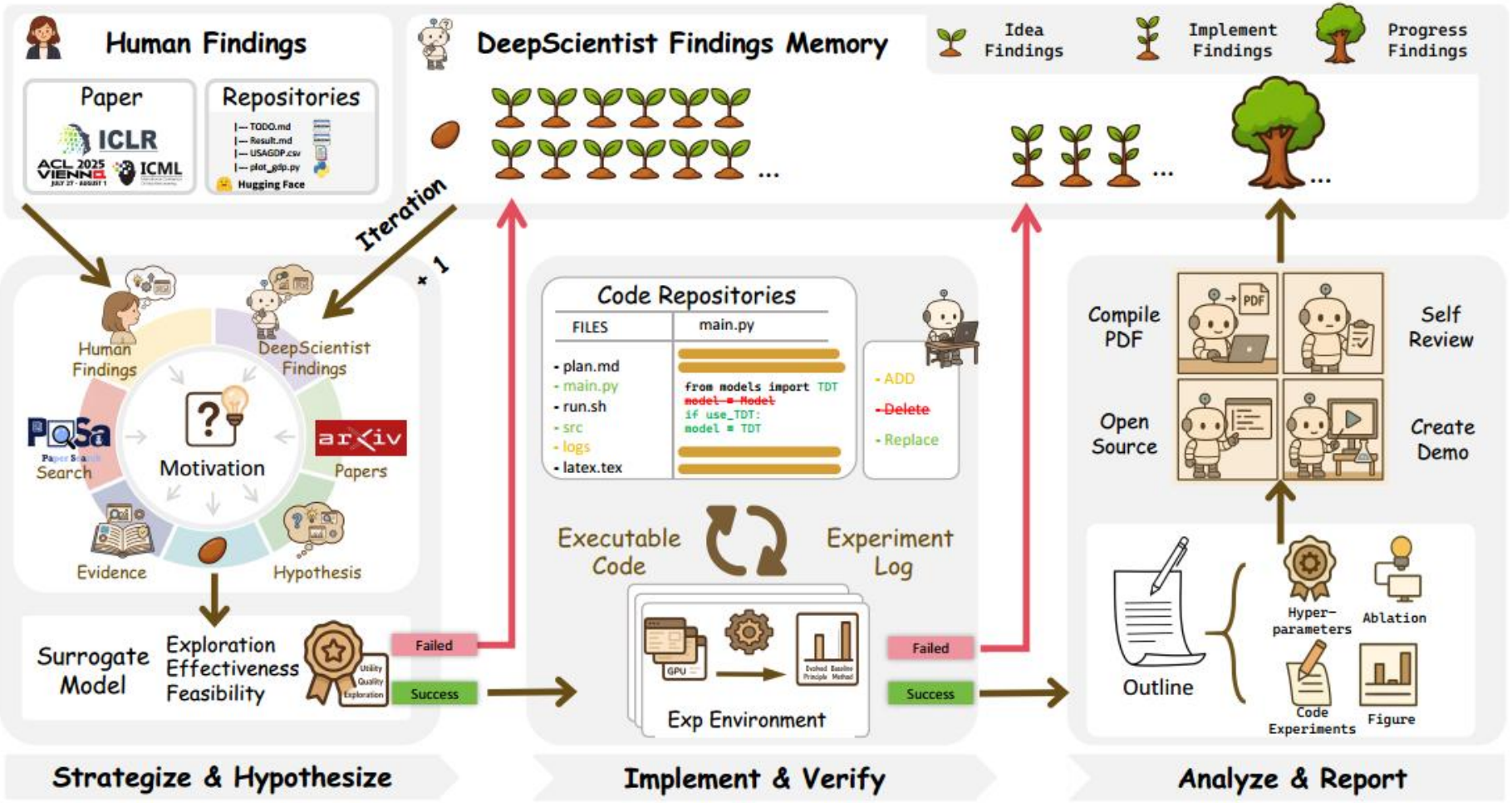
每一个单独的方法I，比如一种新算法或一个新的模型架构，都具有内在的科学价值。

这种价值由一个黑箱式的真实价值函数 f 决定，该函数将一种方法映射到其最终的实证影响上。

- 贝叶斯优化为昂贵黑箱函数的全局优化提供了一种原则性方法。通过**构建一个代理模型来智能引导搜索**，贝叶斯优化通过仔细平衡探索和利用，有效地减少了昂贵的真实世界评估次数。
- 然而，对于科学发现而言，$\mathcal{I}$是一个未被明确界定的概念空间，高质量候选假设的生成是一个关键瓶颈，而传统的贝叶斯优化算法并非为解决这一问题而设计，这一挑战需要一种将创造性构思与样本高效优化相结合的新机制。

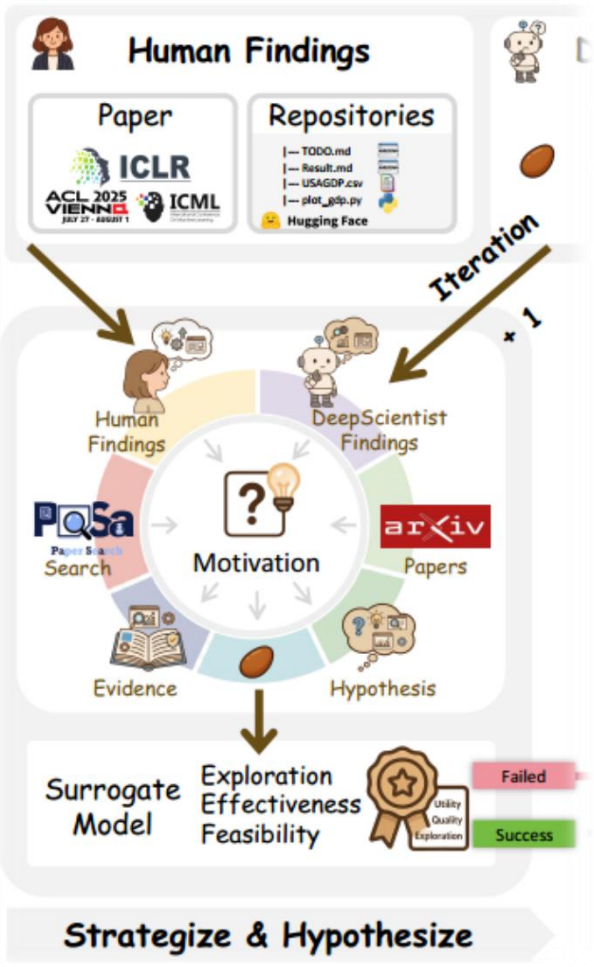
北京大学 信息工程学院
PEKING UNIVERSITY School of Electronic and Computer Engineering

THE DEEPSCIENTIST FRAMEWORK



DeepScientist的架构通过一个配备开放知识系统和持续积累的**发现记忆库**的多智能体系统，实现了**贝叶斯优化循环**。
该记忆库由前沿的人类知识以及系统自身的历史发现共同构成。
整个发现过程被构建为一个**分层且迭代的三阶段探索循环**。在这种分层机制中，只有展现出前景的研究想法才会被推进到成本更高的评估阶段，而其他想法则被保留在发现记忆库中，为后续探索提供参考。

THE DEEPSCIENTIST FRAMEWORK

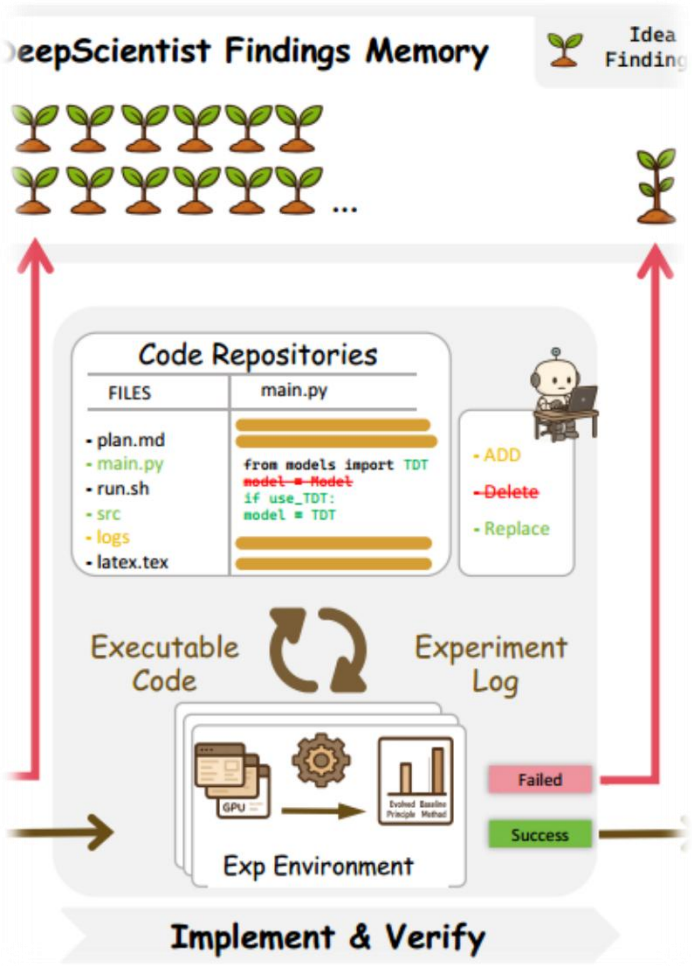


策略制定与假设生成

- 每个研究周期都始于对**发现记忆库$\mathcal{M}_t$**的分析，这是一个包含数千条结构化记录的列表式数据库，并根据每个科学发现的发展阶段进行分类。
- 在第一阶段，系统识别现有知识的局限性，并**生成一组新的假设$\mathcal{P}_{new}$**，然后用一个低成本的**代理模型g_t对这些假设进行评估**。
- g_t首先会结合整个发现记忆库的上下文，对于每个候选发现$\mathcal{P}_{new}$，它会近似真实价值函数f，并为其生成一个结构化的评估向量$\boldsymbol{V} = \langle v_u, v_q, v_e \rangle$，将其估计的效用、质量和探索价值量化为 0 到 100 范围内的整数分数。随后，每个新假设及其评估向量会被用于在发现记忆库中初始化一条新记录，阶段为“想法发现”。

北京大学 信息工程学院
PEKING UNIVERSITY School of Electronic and Computer Engineering

THE DEEPSCIENTIST FRAMEWORK



实施与验证

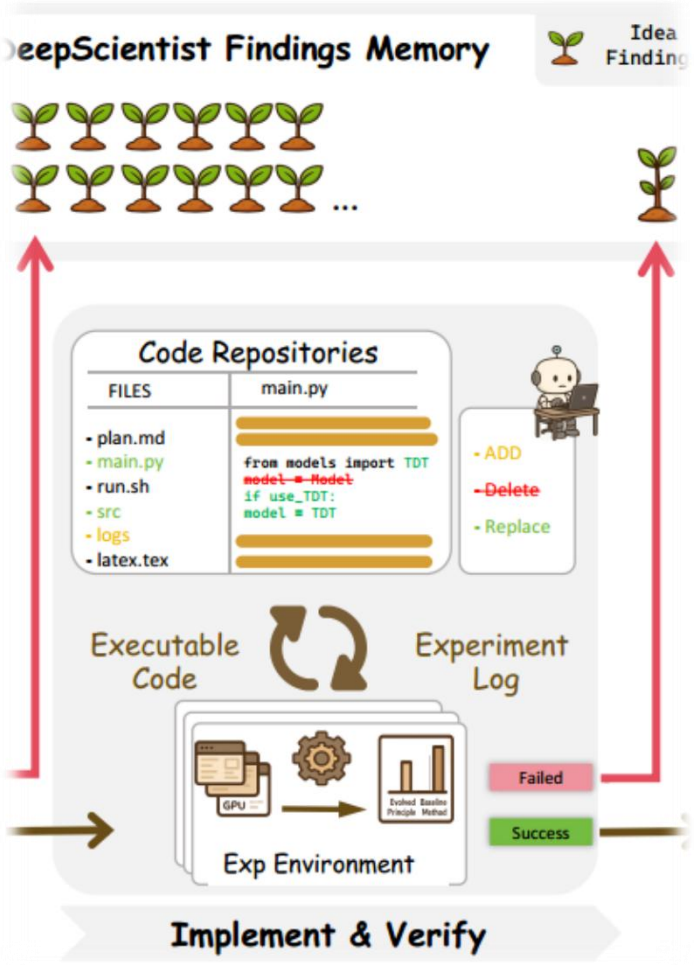
- 这一阶段是发现记忆库中的主要筛选环节。为了确定众多 “想法发现” 中哪些值得投入大量资源以推进到真实世界的实验中，系统采用了一个采集函数(α)。具体而言，它使用经典的上置信界（UCB）算法来选择最有前景的记录。UCB 公式将评估向量V进行映射，以平衡利用有前景途径（v_u, v_q）与探索不确定途径（v_e）之间的权衡：

$$I_{t+1} = arg\ max_{I \in P_{new}} (\omega_u v_u + \omega_q v_q + \kappa v_e)$$

- 其中ω_u和ω_q是超参数，κ控制探索的强度。在科学发现中，UCB 算法恰好能为 “科学发现的资源分配” 提供量化依据，让系统在预算有限时，最大化发现效率。 **(文中三个参数均为1)**

北京大学 信息工程学院
PEKING UNIVERSITY School of Electronic and Computer Engineering

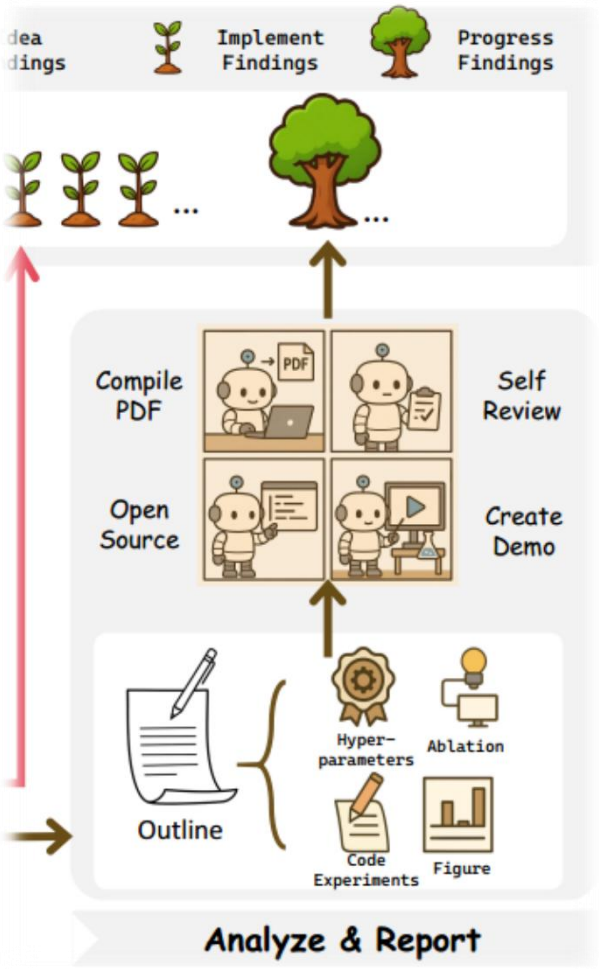
THE DEEPSCIENTIST FRAMEWORK



实施与验证

- 得分最高的发现I_{t+1}被选中进行验证，其记录被提升为 **“实施发现”** 状态。
- 一个编码智能体执行代码库级别的实施以开展实验。该智能体在一个具有完全权限的沙盒环境中运行，使其能够**读取完整的代码库，并可访问互联网进行文献和代码搜索。**
- **它的目标是在现有最先进（SOTA）方法的代码库基础上实施新的假设。**该智能体通常会先规划任务，然后阅读代码以理解其结构，最后实施变更以生成实验日志和结果。
- 实验日志和结果I_{t+1}被用于更新相应的记录，用实证证据丰富它，从而闭合学习循环。

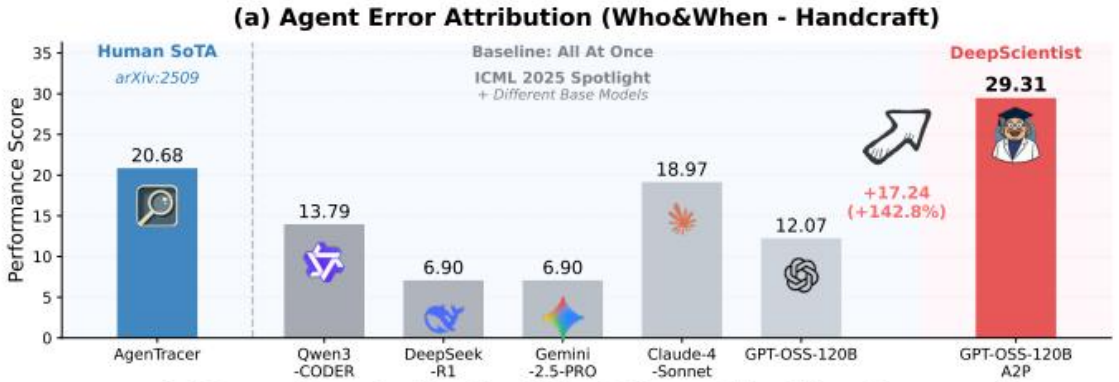
THE DEEPSCIENTIST FRAMEWORK

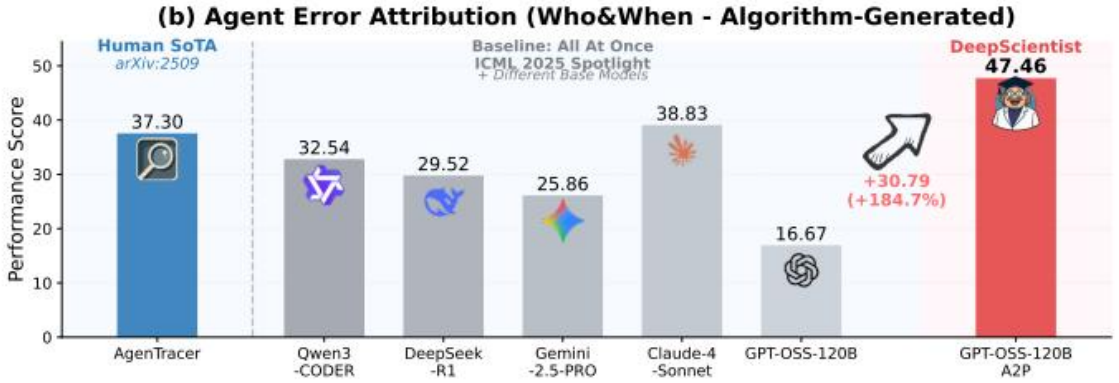


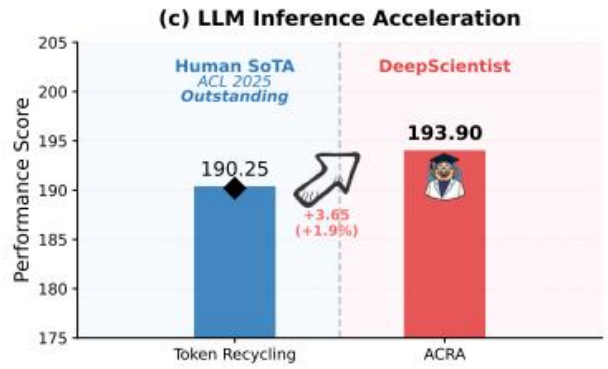
分析与报告

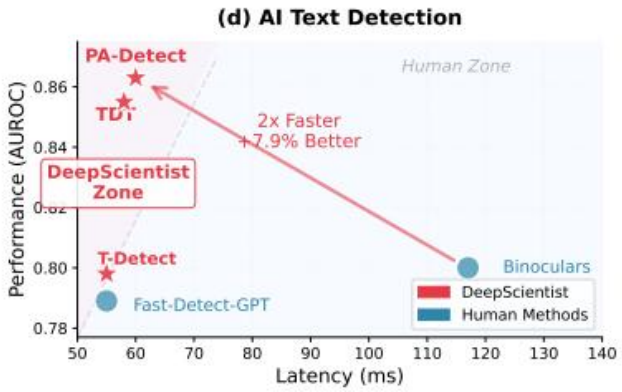
- 发现记忆库的最后也是最具筛选性的阶段，仅在验证成功时才会触发。**当一项 “实施发现” 成功超越基准线时，其记录会被提升为 “进展发现”。**
- 这种转变由一系列具备使用 MCP工具套件能力的专用智能体来实现。这些智能体首先会**自主设计并执行一系列更深入的分析实验**（例如，消融实验、在新数据集上的评估），利用 MCP 工具来**管理实验生命周期、数据收集和结果解析。**
- 随后，一个综合智能体使用同一套工具集，将所有实验结果、分析见解和生成的人工制品**整理成一篇连贯且可复现的研究论文**。经过深度验证的记录会成为系统知识库中的**新记录**，从而影响所有后续周期的决策过程。

Method	Agent Failure Attribution		LLM Inference Acceleration	AI Text Detection	
	Handcraft (Acc.)	Algorithm-Gen (Acc.)	Tokens/second	AUROC	Latency
Human SoTA method	12.07% (All at Once)	16.67% (All at Once)	190.25 (Token Recycling)	0.800 (Binoculars)	117ms (Binoculars)
DeepScientist's method	**29.31%** (A2P)	**47.46%** (A2P)	**193.90** (ACRA)	**0.863** (PA-Detect)	**60ms** (PA-Detect)
Improvement	Δ+142.8% (+17.24)	Δ+183.7% (+30.79)	Δ+1.9% (+3.65)	Δ+7.9% (+0.063)	Δ+190%↓ (-57)





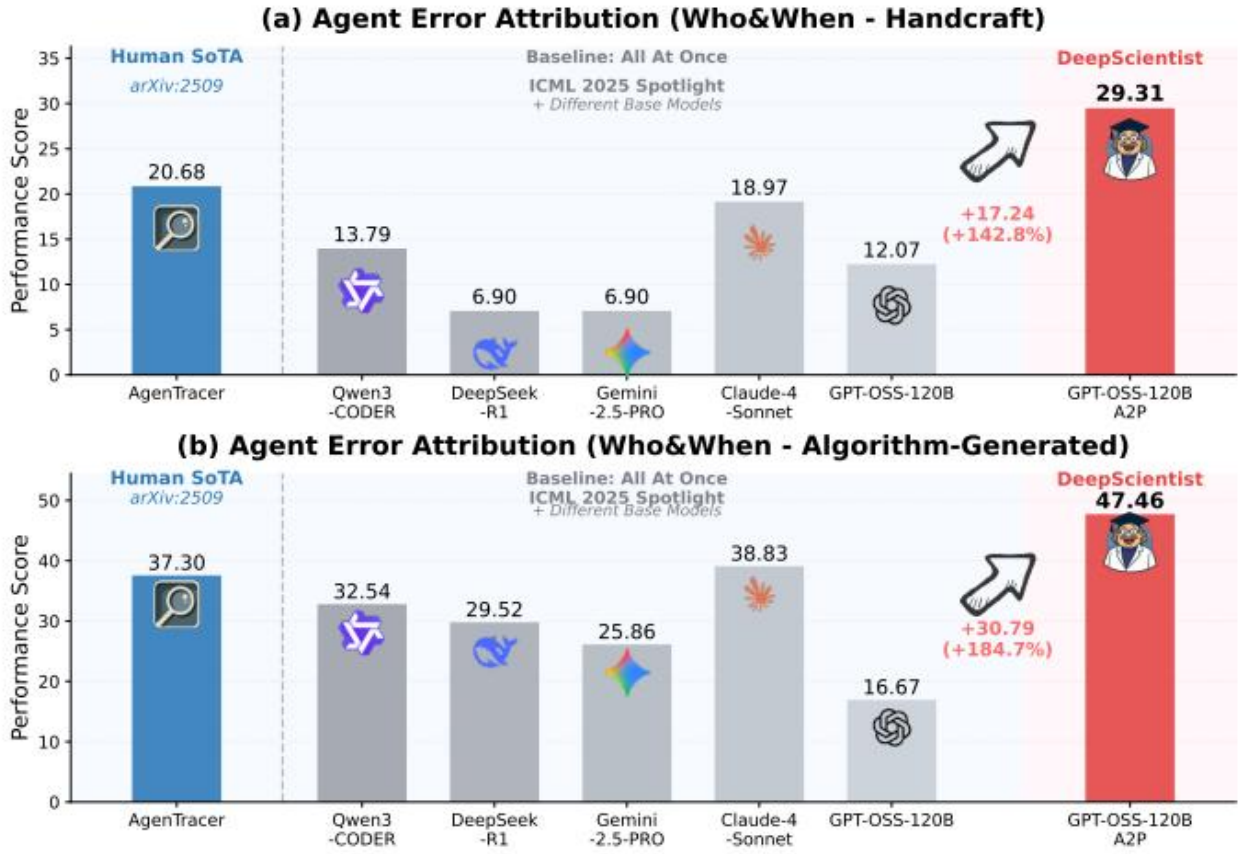




- 选择了三种不同的SOTA方法作为起点。每种SOTA方法都是手动复现的，并保留了执行日志和测试脚本，以便DeepScientist专注于研究进展。
- DeepScientist配备了两台服务器，每台服务器配备了8个Nvidia H800 GPU。
- 为了最大限度地提高利用率，我们为每个GPU启动了一个单独的系统实例，采用**Gemini-2.5-Pro**模型作为核心逻辑，采用**Claude-4-Opus**模型作为强大的代码生成功能。三位人类专家监督这一过程，以验证输出并过滤掉幻觉。

北京大学 信息工程学院
PEKING UNIVERSITY School of Electronic and Computer Engineering

DEEPSCIENTIST ACHIEVEMENTS ON THREE RESEARCH DOMAINS

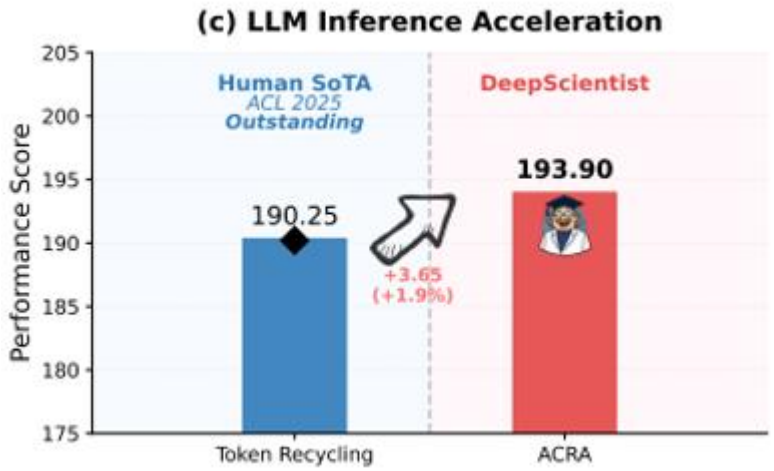


Agents Failure Attribution

- **核心问题：**在基于大语言模型（LLM）的多智能体系统中，是哪个智能体导致了任务失败，以及是在何时失败的?
- 从**基准方法** “All at once”出发，DeepScientist发现当前方法缺乏故障归因所必需的反事实推理能力。
- 它最终提出了 **A2P 方法**。其核心创新点在于将故障归因从模式识别提升到因果推理层面，通过预测所提出的修复措施是否能带来成功，填补了反事实推理能力方面的关键缺口。
- 在基准测试中，A2P 分别取得了 29.31 分和 47.46 分，创造了**新的最先进（SOTA）水平**。

北京大学 信息工程学院
PEKING UNIVERSITY School of Electronic and Computer Engineering

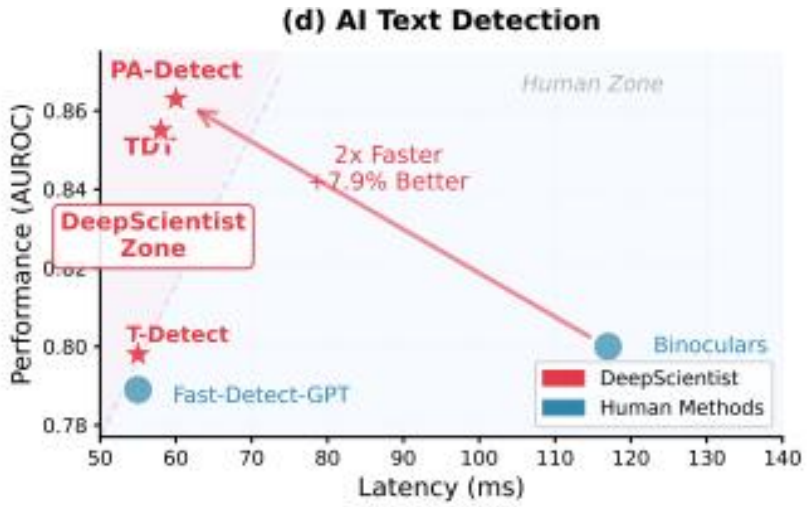
DEEPSCIENTIST ACHIEVEMENTS ON THREE RESEARCH DOMAINS



LLM Inference Acceleration

- **核心问题：**大语言模型（LLM）推理加速是一个高度优化的领域，旨在最大化大语言模型推理过程中的吞吐量并降低延迟。
- DeepScientist生成的 **ACRA 方法**最终通过识别稳定的后缀模式，将 MPBB的人类最先进（SOTA）水平从每秒 190.25 个 token 提升到了每秒 193.90 个 token。
- 从**科学角度**来看，这项创新意义重大，因为它利用这些额外的上下文信息来动态调整解码猜测，有效地为该过程植入了长期记忆，并打破了标准解码器的上下文坍缩问题。
- **这一发现凸显了DeepScientist的首要目标：创造人类未知的新知识，而非仅仅进行工程优化。**

DEEPSCIENTIST ACHIEVEMENTS ON THREE RESEARCH DOMAINS



AI Text Detection

- **核心问题：**对于一段可能包含大语言模型（LLM）生成内容（也可能包含额外噪声）的文本，判断它是由人类还是大语言模型生成的。
- 为了验证其持续进步的能力，DeepScientist 进行了大量尝试，在短短两周的快速演进中，系统产生了三种截然不同且逐步更优的方法**（T - Detect、TDT 和 PA - Detect）**。
- 从科学角度来看，**这种转变揭示了人工智能生成文本的 “非平稳性”**，缓解了以往范式中因对局部证据进行平均化处理而产生的信息瓶颈问题。
- **这一整个发现轨迹展示了 DeepScientist 逐步推进前沿科学发现的能力**，在将 AUROC（曲线下面积）提高 7.9% 的同时，推理速度也提高了一倍，从而建立了新的最先进（SOTA）水平。

ASSESSING THE QUALITY OF AI-GENERATED RESEARCH PAPER

为评估最终产出的质量，我们对 DeepScientist 端到端流程自主生成的五篇研究论文进行评估。我们的评估方案分为两部分。

- 首先，为了与现有工作进行基准对比，我们使用 **DeepReviewer**，这是一个具备外部搜索能力、能模拟人类同行评审过程的人工智能代理，将 DeepScientist 的产出与其他人工智能科学家系统的 28 篇公开论文进行对比。
- 其次，为了进行更严谨的评估，我们召集了**一个专门的程序委员会，由三位活跃的大语言模型研究人员组成**：两位曾担任 ICLR 评审的志愿者，以及一位受邀担任 ICLR 领域主席的资深志愿者。

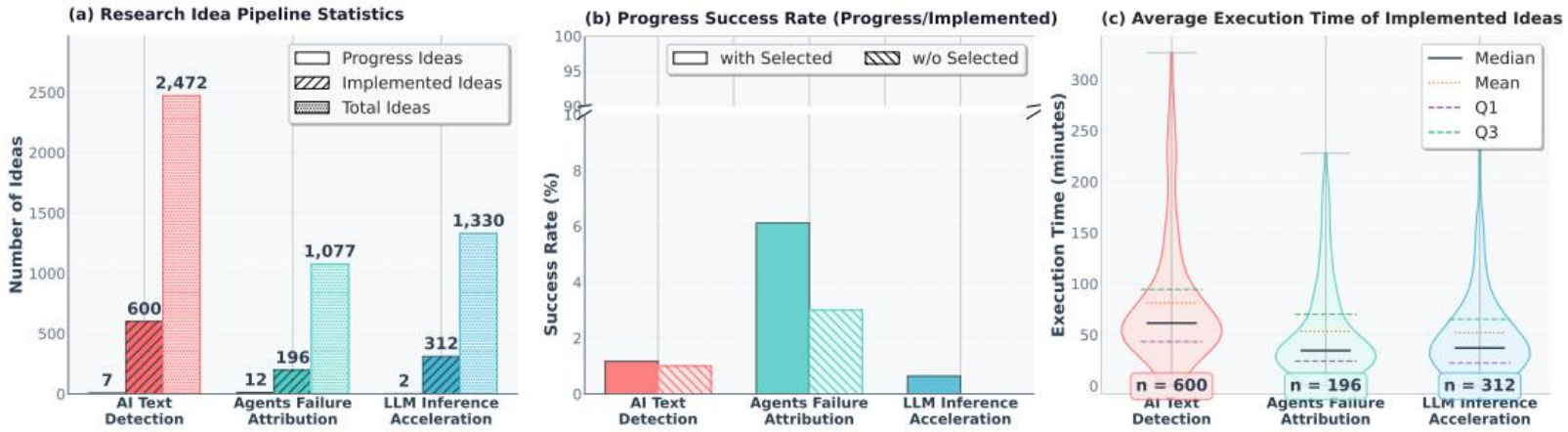
AI Scientist Systems	Number	Soundness	Presentation	Contribution	Rating	Accept Rate
AI SCIENTIST	10	2.08	1.80	1.75	3.35	0%
HKUSD AI Researcher	7	1.75	1.46	1.57	2.57	0%
AI SCIENTIST-V2	3	1.67	1.50	1.50	2.33	0%
CycleResearcher-12B	6	2.25	1.75	2.13	3.75	0%
Zochi	2	2.38	2.38	2.25	4.63	0%
DeepScientist (Ours)	5	**2.90**	**2.90**	**2.90**	**5.90**	**60%**

➢ DeepScientist是唯一一个论文接受率达到60%的人工智能科学家系统。

Paper	Confidence	Soundness	Presentation	Contribution	Rating
HUMAN Avg. (ICLR 2025)	-	2.59	2.36	2.62	5.08
1. T-DETECT	4.33 (0.33)	2.00 (1.00)	**2.67** (0.33)	**2.67** (0.33)	5.00 (0.00)
2. TDT	4.67 (0.33)	**3.00** (0.00)	**3.00** (0.00)	**3.00** (0.00)	**5.67** (0.33)
3. PA-DETECT	4.00 (0.00)	1.67 (0.33)	2.00 (1.00)	2.00 (1.00)	4.33 (1.33)
4. A2P	4.00 (0.00)	**3.00** (0.00)	**3.00** (0.00)	**2.67** (0.33)	**5.67** (0.33)
5. ACRA	3.33 (0.33)	1.67 (0.33)	2.00 (1.00)	1.67 (0.33)	4.33 (1.33)
DeepScientist Avg.	4.07	2.27	**2.53**	2.40	5.00

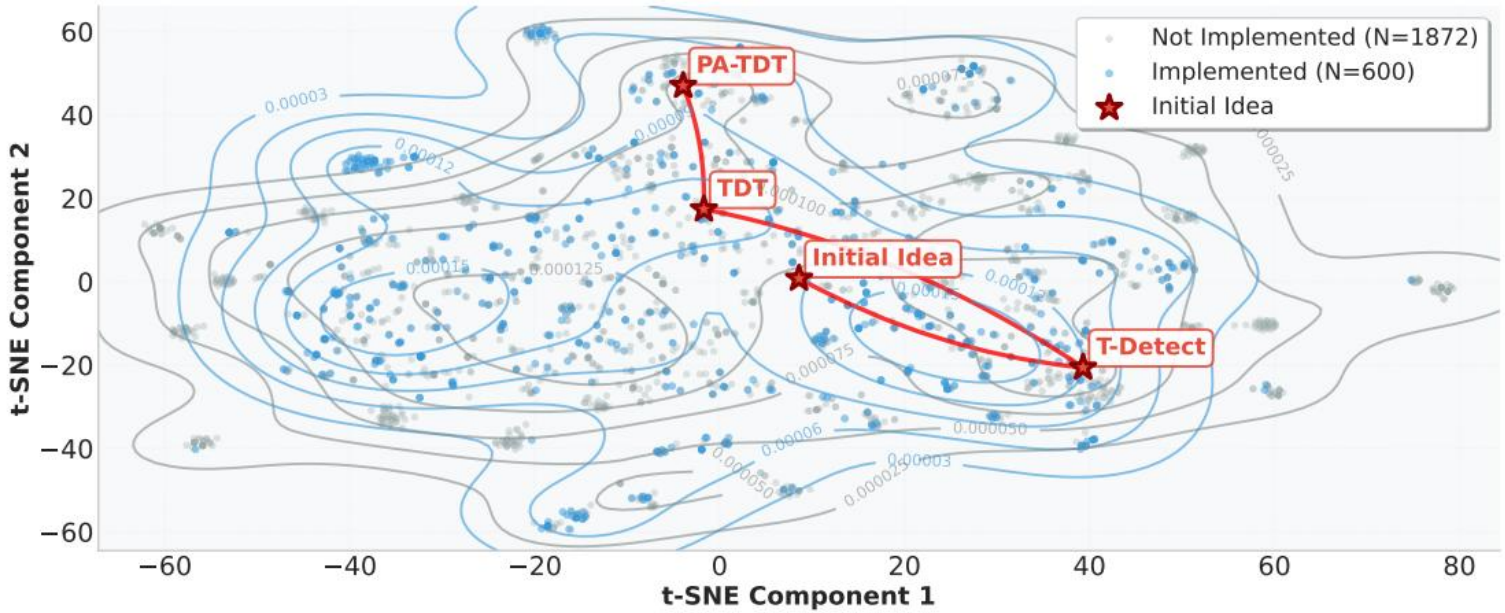
➢ DeepScientist的平均评分（5.00）与所有ICLR 2025提交的平均评十分接近（5.08），其中两篇论文明显超过了这一评分（5.67）。

ANALYSIS OF THE ITERATIVE TRAJECTORY OF AUTONOMOUS EXPLORATION



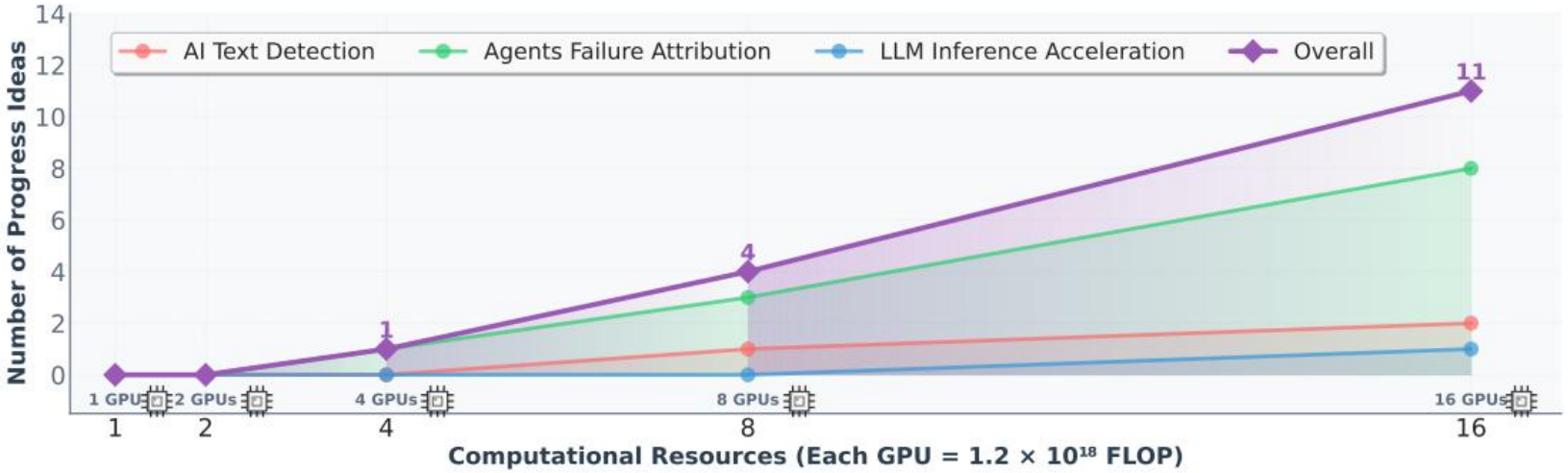
- 我们对 DeepScientist 实验日志的分析揭示了**自主科学发现中固有的试错过程的巨大规模。**
- 执行时间分布表明，当前自主科学存在**明确的应用边界：**对于具有快速反馈循环的任务，将大规模实验委托给人工智能是一种强有力的策略。然而，对于像预训练基础模型或药物合成这样的高成本工作，低成功率使得这种方法目前不切实际，必须继续依赖人类主导的构思
- **选择过程的关键性：**如果没有它，为每个任务随机抽取 100 个想法并进行测试，成功率实际上为零。通过我们的选择策略，成功率上升到约 1 - 3%，这表明尽管仍然很低，但智能筛选至关重要。
- 因此，**这项工作的成功并非源于暴力计算，而是源于搜索效率**。如果采用一种简单的方法，对所有 5000 个有希望的候选者进行全面测试，将需要超过 10 万 GPU 小时，而我们的定向探索仅用 2 万 GPU 小时就取得了突破。

ANALYSIS OF THE ITERATIVE TRAJECTORY OF AUTONOMOUS EXPLORATION



- DeepScientist 的**发现过程遵循一个有目的且逐步推进的轨迹：**尽管系统在广阔的概念范围内生成了数千个不同的想法，但它通往成功的路径并非随机，而是一系列有重点、符合逻辑的进展。
- 这表明**它具备逐步深化理解的能力：**在通过 T - Detect 取得初步突破后，系统有效地确立了最先进水平（SoTA），识别出其后续的局限性，并将搜索重新定向到新的目标。这种动态探索体现在向 TDT 和 PA - Detect 的概念转变上，它们利用新的位置和时间信息，在之前成功的基础上进一步发展。
- **这种基于自身发现、将每一项成功发现转化为识别和解决下一组局限性的新起点的能力，展示了强大的科学探索能力。**

ANALYSIS OF THE ITERATIVE TRAJECTORY OF AUTONOMOUS EXPLORATION



- 为了研究**计算规模与科学进步速度之间的关系**，我们评估了 DeepScientist 在固定的一周时间内，生成的 “进展发现”数量与可用并行资源的函数关系。
- 在这个设置中，系统首先识别出基准方法中的一系列局限性，每条并行探索路径都被赋予解决一个不同局限性的任务，所有路径会定期将其结果同步到共享的发现记忆库中。结果表明：**分配的资源与有价值的科学发现产出之间近乎线性的关系。**
- 这种结果不仅仅源于并行的试错，而是共享知识架构的直接结果。当每条并行路径进行探索时，它会丰富共享的发现记忆库。这会产生一种协同效应，使系统的集体智慧得以增长，让每条独立路径都能从其他路径的成功，以及同样重要的失败中受益。**这表明，有效扩展自主科学不仅仅是增加暴力计算的问题，而是要培育一个更丰富、相互关联的知识库，从而在所有并发工作中加速发现。**

HUMAN EXPERT REVIEW

在五篇生成的论文中，人类评审者达成了明确的共识：

- DeepScientist 在**研究的构思阶段始终表现出色**。这种反馈验证了该系统作为识别相关研究空白并生成创新性、有影响力解决方案的强大引擎的核心优势，确认了它能够成功构思出超越单纯渐进式改进的成果。
- 然而，这种在构思方面的优势，**因科学执行和严谨性方面反复出现的弱点而被系统性地削弱**。最关键且最频繁被提及的问题是实证可靠性不足，DeepScientist 未能设计出全面的验证计划，它还存在未能恰当地将自身贡献置于相关背景中的问题，论文常常省略与重要基准的对比，或未能讨论密切相关的研究工作。

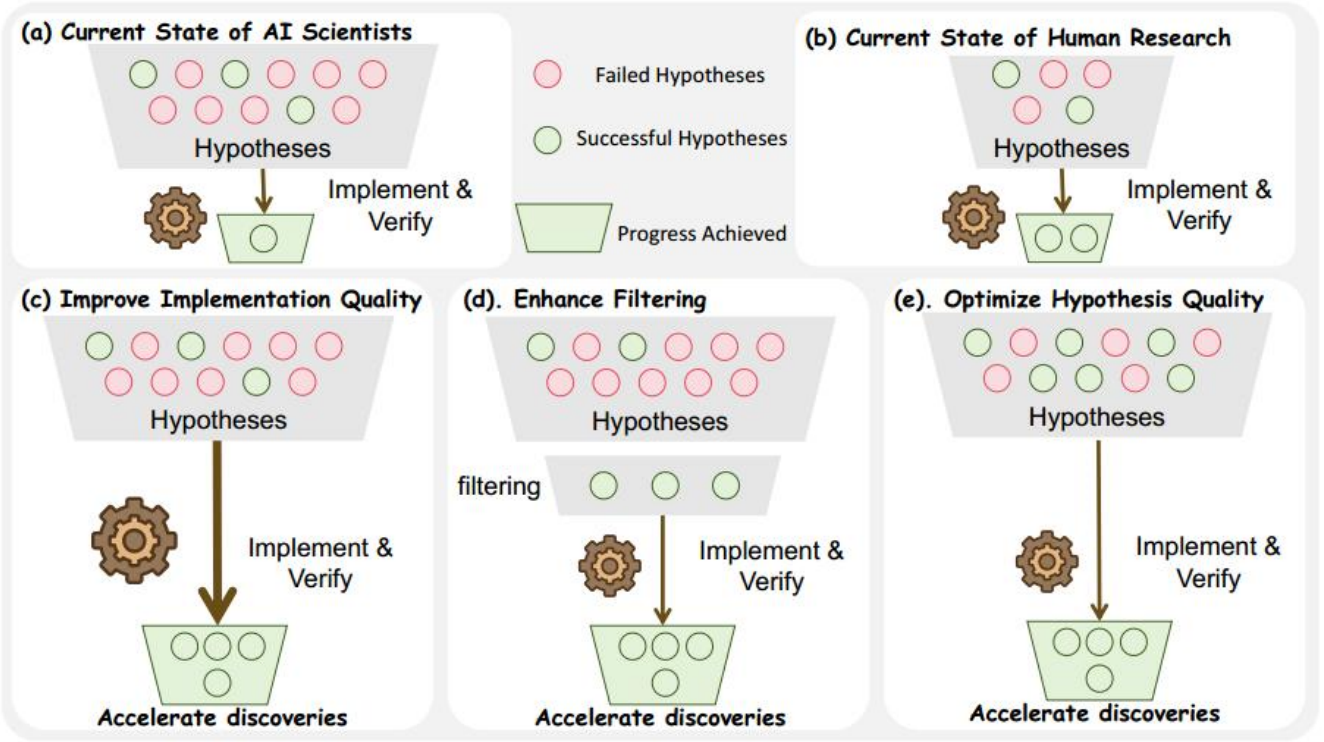
这些反馈指出了当前自主系统的主要瓶颈：在生成新颖概念的能力与进行严谨科学执行和阐述的能力之间存在巨大差距。该系统不仅未能正确实施这些想法，也未能令人信服地对其进行评估。为了弥合这一差距，未来的工作需要在两个关键领域进行提升：

- 首先，**开发明确接受过实验设计训练的智能体**，能够规划全面的评估，以预见并应对潜在的科学批评；
- 其次，**增强系统的分析推理能力**，使其不仅能描述结果，还能解释其重要性、形成有说服力的论据，并参与那种具有高影响力研究特征的深入、反思性讨论。

ADDRESSING THE BOTTLENECKS IN AUTONOMOUS SCIENTIFIC DISCOVERY

- 多数人工智能系统生成的最终能带来实质性进展的想法成功率通常低于 3%，意味着大量计算资源被用于探索低价值假设。（a）和（b）说明当前人工智能与人类研究均面临低成功率问题。**这种低效的“大海捞针”模式是阻碍人工智能科学家从“新颖工具”发展为“高效发现者”的核心障碍。**

- 因此，为进一步加速科学发现进程，未来人工智能科学家系统需在三个关键方向协同演进：
  - **优化初始假设质量、**
  - **增强过程中的筛选能力、**
  - **提升最终阶段的实施与验证质量。**



- **科学探索的新范式：**它的核心优势并非不出差错，而是能够以前所未有的规模和速度开展这种试错过程，将人类数年的探索压缩到数周内。因此，首要的前进方向是专注于系统性地提高这种发现效率，同时提升生成假设的质量及其实施的稳健性。
- **人机协同的巨大机遇：**我们设想未来 DeepScientist 将作为大规模探索引擎，其轨迹由人类智慧引导。人类研究人员的角色可以从繁琐的实验工作转向制定有价值的科学问题和提供战略方向等高层认知任务，从而利用人工智能进行快速、全面的探索。
- **未来的工作应聚焦于关键改进：**开发模拟发现环境以通过强化学习加速学习，创建整合科学界反馈的框架，最终通过机器人技术弥合与物理世界之间的差距。

北京大学深圳研究生院
Peking University Shenzhen Graduate School

请批评指正！

Project: https://ai-researcher.net
Code: https://github.com/ResearAI/DeepScientist

汪静雅

2025.10.15