揭秘科学计算新范式：什么是“可评分任务”与“实证软件”？

2025年10月13日　1 分钟阅读

#google　#可评分任务　#实证软件
#AI-assisted”，”科学探索

> *借助AI，现代计算科学正把”宏大科学问题”变成一场可打分的游戏*

TLDR

现代计算科学正把“宏大科学问题”变成一场可打分的游戏：先定义可量化指标（可评分任务），再让算法像学生刷题一样不断拟合观测数据（实证软件）。AI 借此 24×7 自动搜索最优程序，把探索周期从“年”压到“小时”，已在蛋白质结构预测、森林砍伐监测、疫情住院预测等领域大显身手，正引爆一场科研加速革命。本文将详细介绍这一范式的核心概念，以及它如何改变我们对科学探索的理解和实践。本文灵感来自Google论文【An AI system to help scientists write expert-level empirical software】

引言：当科学探索变成一场“可评分的游戏”

在当代科学研究的诸多领域中，为支持计算实验而手动创建和优化“经验性软件”已成为一个普遍存在的瓶颈。这一过程不仅耗时费力，而且往往依赖于研究者的直觉和经验，限制了探索解决方案的广度和深度。为了突破这一局限，Google系统性地阐述一个结合了大型语言模型（LLM）与**树搜索**（Tree Search）的创新型AI系统。该系统能够自动化地为**可量化评分的科学任务**创建达到甚至**超越专家水平的软件**，从而显著加速科学探索的进程。本报告旨在为跨学科的研究人员提供一个可复制、可扩展的**计算研究框架**，以应对各自领域中的复杂挑战。 从根本上说，该AI系统通过将软件创建这一开放性问题形式化为一个可在巨大方案空间内进行探索和优化的可

目录

文章信息

字数

阅读时间

发布时间

更新时间

标签

#google　#可评分任务　#实
#AI-assisted”，”科学探索

量化搜索问题，为自动化科学发现开辟了新的道路。 这种强大的新范式，其核心是两个紧密相连的概念：“可评分任务 (scorable task)”和 “实证软件 (empirical software)”。它们共同将宏大而复杂的科学探索，转变为一场规则明确、目标清晰、可计算、可优化的“游戏”。这种将科学转化为“可评分游戏”的范式已经变得如此强大和系统化，以至于我们正进入一个全新的时代：一个AI系统本身可以扮演专家科学家的角色，自动编写和优化赢得这些“游戏”所需的软件，从而加速科学发现。

那么，究竟什么是“可评分任务”和“实证软件”呢？让我们从它们的基本定义开始。

1. 核心概念解析：为科学问题设定“游戏规则”

1.1. 定义“实证软件”与“可评分任务”

要理解这个新范式，我们首先需要明确两个基本定义。

> 实证软件 (Empirical Software) 简单来说，这类软件的核心目标不是凭空创造，而是通过不断优化算法和参数，使其输出的结果尽可能地逼近和拟合现实世界中已经观察到的数据。它就像一个学生，通过不断做练习题（拟合观测数据）来提高自己的考试分数（质量分数）。简而言之：实证软件是学生，观测数据是教科书，而质量分数就是最终的考试成绩。
>
> 可评分任务 (Scorable Task) 我们可以为一项科学挑战设定一个明确的、可量化的评分标准。只要有了这个“分数”，我们就能判断哪一个软件解决方案做得更好，哪一个更差。

1.2. 核心洞察：从“宏大问题”到“可计算目标”

这种方法论的重大意义在于，它为我们提供了一座桥梁，将一个宏大、有时甚至有些模糊的科学问题，转化为一个具体、可执行、可优化的计算目标。

> Before (模糊的科学问题): “我如何分析全球的森林砍伐情况？”
>
> After (可评分的任务): “我如何编写一个软件，让它在识别卫星图像中的森林砍伐区域时，在 mIoU 指标上获得最高分？”

这种从问题到目标的转变至关重要，因为一旦一个任务变得“可评分”，它就为自动化和大规模计算优化打开了大门。这是将人类的科学直觉与机器不知疲倦的计算能力相结合的关键一步。

这种转变之所以如此强大，恰恰因为它为长期困扰传统科研软件开发的慢性瓶颈提供了直接的答案。

2. 为何需要这种方法？传统科研软件开发的挑战

2.1. 传统方法的瓶颈

在“可评分任务”范式出现之前，开发用于科学研究的软件通常面临以下几个核心痛点：

> 开发过程缓慢且繁重 传统科研软件的开发过程极其缓慢和繁重，通常需要耗费科学家数年时间投入到手动的软件创建工作中，这极大地拖慢了研究周期。
>
> 探索范围受限 由于开发耗时巨大，科学家们能够富有成效地探索的可能性受到了严重限制。他们没有足够的时间和精力去尝试所有可能的解决方案或算法组合。
>
> 依赖直觉而非穷举 在传统开发模式下，设计选择通常由直觉或权宜之计决定，而不是通过详尽的实验。这意味着最终的软件方案可能只是一个“足够好”的方案，而不一定是最优解。

2.2. “评分”带来的优势

与缓慢、依赖直觉的手动过程形成鲜明对比，“可评分任务”框架彻底改变了游戏规则。一旦定义了明确的“分数”，我们就可以释放计算系统的力量，来“不知疲倦地 (tirelessly)”和“详尽地 (exhaustively)”搜索巨大的解决方案空间。这种自动化的方法能够系统性地发现那些“大海捞针式”的高质量解决方案——而一个受限于时间、认知偏见或权宜之计的人类专家，可能永远也无法发现这些方案。

为了更具体地理解这一概念，让我们来看几个不同科学领域的真实案例。

3. “可评分任务”在真实科研领域的应用

“可评分任务”的框架并非纸上谈兵，它已经广泛应用于各个前沿科学领域，并取得了卓越的成果。下表展示了几个典型的应用案例：

科学领域	核心科学问题 (可评分任务)	软件的目标 (实证软件)	如何评分？(质量指标)
地理空间分析	准确识别卫星图像中的土地覆盖变化，特别是森林砍伐。	一个能自动分析全球范围、高分辨率卫星图像并为每个像素分配类别标签（如森林、建筑、水体）的程序。	通过将软件的预测标签与专家标注的"真实标签"进行比较来评分。一个常用的指标是"平均交并比 (mean Intersection over Union, mIoU)"，它衡量了预测区域与真实区域的重合度。
化学/生物学	根据蛋白质的氨基酸序列，准确预测其三维空间结构。	一个能根据一维的氨基酸序列信息，计算出其在物理上最稳定、最可能的三维折叠结构的程序。	分数通常用于衡量预测结构与通过实验（如X射线晶体学）验证的真实结构之间的吻合度。基于"实证软件"的蛋白质结构预测工作已获得了2024年诺贝尔化学奖。
公共卫生	提前数周准确预测与新冠病毒（COVID-19）相关的住院人数。	一个能分析历史数据（如感染率、住院人数、季节性趋势等），以预测未来几周住院人数变化趋势的程序。	使用"加权区间分数 (Weighted Interval Score, WIS)"进行评分。该指标不仅奖励预测的准确性，还奖励模型对其预测不确定性的良好校准，分数越低代表表现越好。

从这些例子可以看出，"可评分任务"的框架正广泛应用于各个前沿领域，并对科学探索的未来产生了深远影响。

4. 结论：加速科学发现的引擎

通过本文的解析，我们可以清晰地看到"可评分任务"和"实证软件"这对概念的核心思想：将复杂的科学研究问题，转化为一个寻找能获得最高分的软件程序的优化问题。

这种方法论并非个例，它在科学、应用数学和工程的"几乎每个子领域"都普遍存在。它的真正威力在于其加速效应。通过将一个想法的探索周期"从数周或数月，缩短到数小时或数天"，科学家们能够以前所未有的速度进行迭代、试错和创新。这种戏剧性的加速是由新型AI系统实现的，这些系统能够不知疲倦且详尽地搜索最优的软件解决方案，而这项任务过去一直受到人类时间和直觉的限制。

正如相关研究者所展望的那样，这一范式预示着一个激动人心的未来：

"我们相信，在那些解决方案可以被机器评分的科学领域，进步正处于一场革命性加速的前夕。"

参考

Google blog: Accelerating Scientific Discovery with AI-Powered Empirical Software

Google Paper: An AI system to help scientists write expert-level empirical software

Wikipedia: Search tree