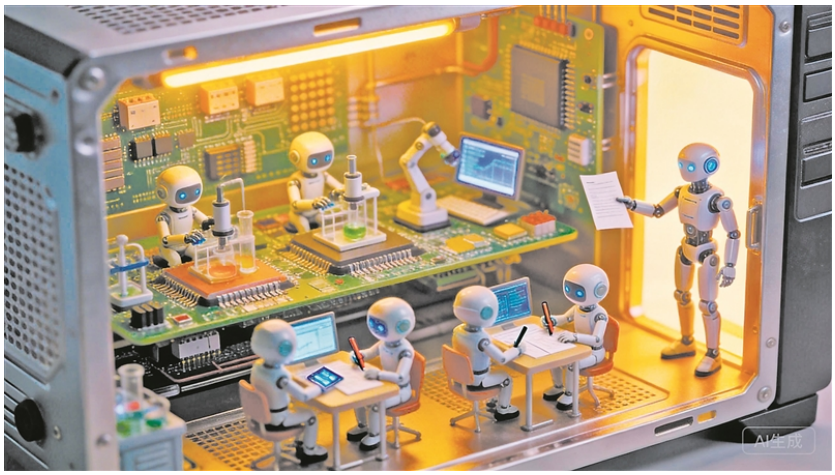
OpenAI致力打造自主“ AI研究员 ”

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/39033.html

OpenAI致力打造自主“ AI研究员 ”。



“ AI研究员 ”是一套全自动的多智能体研究系统，能够独立完成从数学、物理到生物、化学，乃至政策分析的各类科研任务。它能够像人一样，以连贯的方式持续工作，完成一系列科研任务，并产生可供分析的新结果。图片由AI生成

在人工智能（AI）竞逐日益白热化的今天，OpenAI再次抛出了一个震撼业界的宏大蓝图。近日，OpenAI首席科学家雅各布·帕乔基在接受《麻省理工科技评论》独家专访时透露，他们正瞄准一个前所未有的科研目标：在2028年前，打造一个能够自主解决复杂问题的“ AI研究员 ”。

这是一套全自动的多智能体研究系统，能够独立完成从数学、物理到生物、化学，乃至政策分析的各类科研任务。OpenAI表示，实现“ AI研究员 ”计划是该公司未来几年的方向。今年9月，第一阶段目标将率先落地，届时，OpenAI将先行推出“ 自主AI研究实习生 ”。

这一计划标志着OpenAI在推动AI技术应用方面的新尝试，同时也是其在面对Anthropic、“ 深度思维 ”等竞争对手时的重要战略部署。

“ 我们正接近这样一个阶段：我们的模型能够像人一样，以连贯的方式无限期地工作。 ”帕乔基表示，“ 当然，仍然需要有人负责并设定目标。但我认为，我们将最终达到这样的境界：在数据中心里拥有一个完整的研究实验室。 ”

从Codex开始“ 进化 ”

OpenAI并非在空中楼阁上构筑梦想。今年1月，OpenAI发布了Codex，这是一款能即时生成代码、执行复杂计算任务的智能体应用。它能分析文档、生成图表、整理邮件和社交媒体摘要等。时至今日，Codex已经成为其内部员工的标配，辅助开发代码并解决问题。帕乔基表示，可以把Codex看作是“ AI研究员 ”的雏形。未来，Codex将实现颠覆性革新。

作为OpenAI首席科学家和公司长期研究目标的制定者，帕乔基已经观察到，在技术演进上，模型的“ 长程工作能力 ”正随着参数规模和逻辑深度的增加而呈线性提升。

从GPT-3到GPT-4，模型在无干预情况下处理复杂问题的时长实现了质的飞跃。而2024年推出的“ 推理模型 ”技术，通过引入“ 思维链 ”训练，让AI学会了像人类一样步步为营、遇错回溯。目前，OpenAI正在利用数学和编程竞赛的难题对模型进行“ 魔鬼训练 ”，旨在提升其处理超长文本和拆解多重子任务的能力，最终能够解决现实世界的科研难题。

帕乔基认为，自动化科研的关键在于系统能够长期运行，减少人工干预。帕乔基解释说：“ 我们的目标是开发一个研究实习生系统，可以把本来需要几天的人力任务交给它完成。 ”通过训练模型逐步解决问题、回溯错误，推理模型能够在较长时间内保持连贯工作。

艾伦人工智能研究所的研究科学家道格·唐尼表示，自动化科研是令人兴奋的探索。“ 想象一下，明天早上我们回到实验室，智能体已经完成了一系列科研工作，并产生可供分析的新结果，这将极大加速科研进程。 ”

AI科研能力进入验证阶段

OpenAI目前更专注于与现实世界相关的研究。据介绍，研究人员已经利用驱动Codex的GPT-5模型，发现了多个未解数学问题的解决方案，并在生物、化学和物理学的若干难题中取得了进展。

这种生产力的飞跃，甚至改变了那些最“ 硬核 ”程序员的职业习惯。帕乔基坦言，由于对代码精准度有着近乎苛刻的追求，他一年前甚至拒绝使用最基础的自动补全功能，更倾向于在Vim编辑器（一款深受资深程序员喜爱的文本编辑器）中手动输入每一个字符。但随着模型能力的迭代，他的看法发生了根本性改变。他发现，尽管复杂的架构设计仍需由人主导，但在实验验证阶段，AI可以在一个周末内完成他以前需要一周才能编写完的代码。

针对OpenAI乐观的预期，学术界仍有不同声音。艾伦人工智能研究所的研究员指出，在去年的测试中，当任务需要多个复杂的逻辑步骤耦合时，现有模型极易因为每一个微小错误的累积，导致最终结果崩溃。对此，OpenAI正在不断迭代模型，例如近期发布的GPT-5.4版本，旨在进一步增强逻辑稳定性和任务处理的连贯性。OpenAI希望通过这种不断地迭代，证明“ AI研究员 ”在真正深度介入现实世界的科研之前，是具备科学意义上的可靠性的。

需共同应对“ 集中化力量 ”的挑战

然而，当科研的“ 方向盘 ”逐渐移交给算法，安全与伦理的围栏必须同步加固。帕乔基指出，一个能运行整个研究计划的强大AI，可能会伴随一些尚未解决的重大问题，例如系统失控、遭受黑客攻击，或者可能仅仅是误解了自身的指令。为了应对这些挑战，OpenAI正在推广“ 思维链监控 ”技术，即训练模型在“ 草稿本 ”中记录工作笔记，以便研究人员实时审计其行为是否符合预期。

帕乔基认为，在能够完全信任这些系统之前，必须设置严格的限制，例如将极强大的模型部署在与外界隔绝的“ 沙箱 ”中。他还提醒道，“ 想象一下，一个数据中心能完成过去需要大型组织才能完成的科研工作，而现在可能只需几个人 ”。这种能力集中、影响力巨大的系统将对社会和政策带来新挑战。

面对这种力量的崛起，帕乔基预测，即使到2028年，AI系统仍不会在所有方面都像人类一样聪明，但这并不妨碍它产生巨大的变革作用。这需要社会、政策制定者和科研机构共同参与监管，而非仅靠OpenAI一家公司。

作者：张佳欣 来源：科技日报