AI正自主制造下一代AI？

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/42174.html

AI（人工智能）初创公司Anthropic公开了一套用于衡量前沿AI实验室研发速度的指标体系，并首次披露公司内部相关数据。

当地时间9月17日，Anthropic在其官网上发布一篇名为《用于了解前沿AI实验室内部AI研发速度的指标》的文章。Anthropic在文章中披露，截至2026年8月，旗下AI产品Claude主导了公司26%的AI研发活动，约有3万个AI Agent（智能体）同时在其主要内部平台上运行。

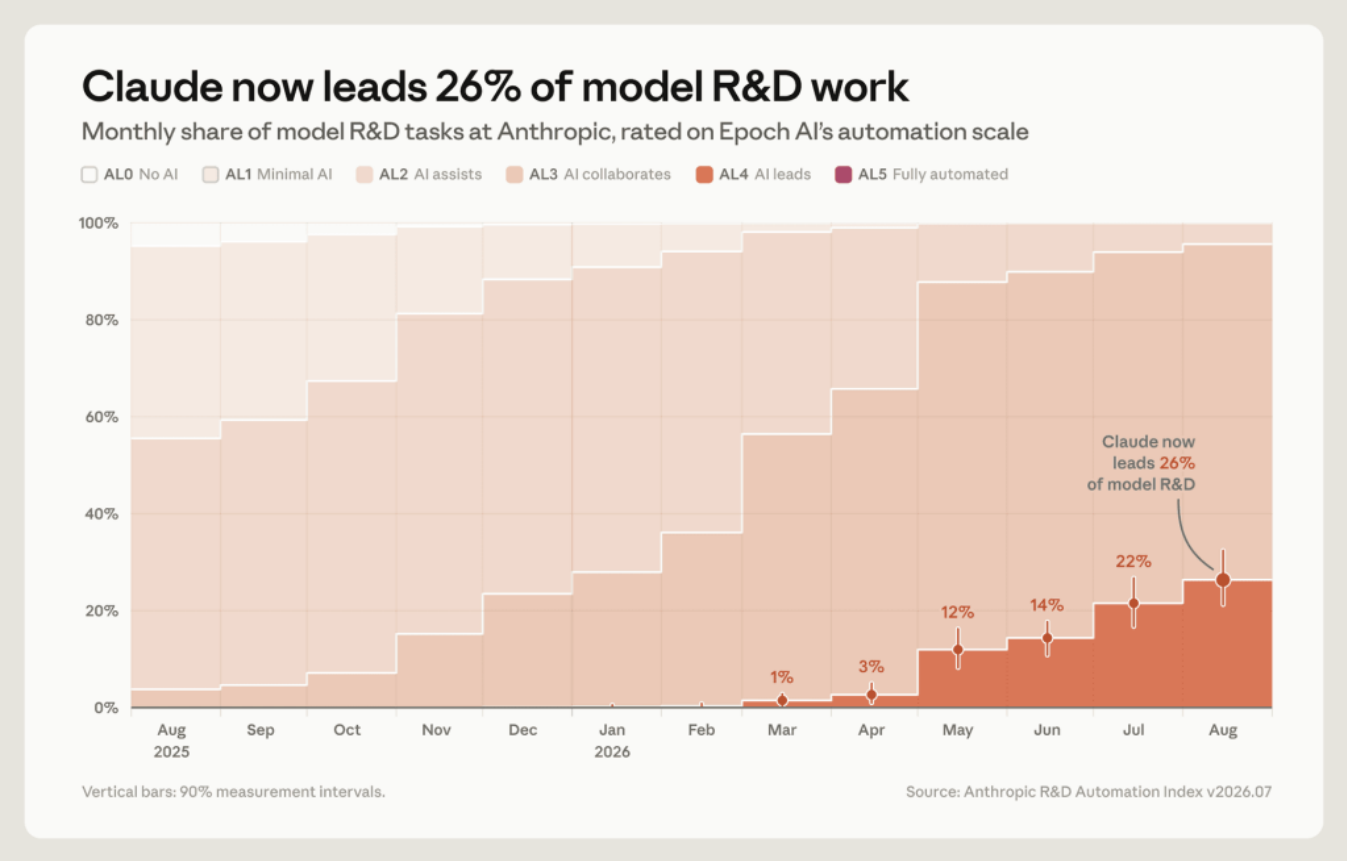
Anthropic写道，随着AI模型能力呈指数级增长，AI已经开始越来越多地参与“构建下一代AI”的过程。在全球讨论是否应当放缓前沿AI发展的背景下，公众需要更多信息来了解AI实验室内部究竟以多快的速度推进研发。

为此，Anthropic披露了内部采用的三项关键指标，用以衡量前沿AI实验室的研发速度：AI研发有多大比例由AI自身完成；AI
Agent的行动受到多大程度的监督；AI研发所使用的算力如何分配。

首先，为了衡量Claude参与公司AI研发工作的程度，Anthropic构建了一项名为“Anthropic RD Automation Index”的原型指标。具体方法是，将Anthropic内部所有类型的AI研发工作进行分类，再根据每项工作的自动化程度进行评级，最终汇总形成整体指标。

该评级体系参考了独立研究机构Epoch AI提出的“自动化等级”（AL），从AL0到AL5。Anthropic披露，截至2026年8月，Claude尚未在任何被测量的AI研发工作中实现完全自主运行，但其已经“主导”Anthropic约26%的AI研发工作，超过90%的AI研发工作至少达到“AI协作”（AL3）水平。

值得注意的是，直到2026年2月，Claude“主导”的AI研发任务占比还不到1%，而该比例到8月已经快速升至26%。



Anthropic披露由Claude“主导”的AI研发任务占比。来源：Anthropic

不过，Anthropic也指出，目前各实验室尚没有统一的测量方法，因此不同公司的数据尚不能直接进行横向比较。此外，Anthropic使用自己的模型对自身系统进行评估，也可能出现“评估模型与被评估模型犯类似错误”的问题。

其次，针对Agent行动的监督问题，Anthropic提出三个指标，分别是覆盖率、审查延迟和升级率。据介绍，截至2026年8月，Anthropic最常用的内部平台上，同时约有3万名Agent从事研发和工程工作。

最后，针对公司把多少算力用于“安全”，Anthropic强调，算力是AI研发过程中相对容易验证的投入指标之一。因此，如果未来全球真的要协调“放缓前沿AI发展”，算力可能成为一个重要的调节手段。例如，可以要求AI公司提高用于AI安全、对齐、可解释性、安全测试和模型评估的算力比例。

Anthropic统计了2026年7月13日至20日这一周公司全部算力的使用情况，并将不同工作负载进行分类。结果显示，在所有用于AI研发的算力中，约6%用于AI安全工作；在用于“AI驱动的AI研发”的算力中，约12%用于AI安全工作。

公司表示，他们将继续公布其各项指标，并接受来自多个外部机构的独立第三方评估人员对其内部数据和评估流程进行验证。

Anthropic此次公布这些数据，与公司CEO达里奥·阿莫迪（Dario Amodei）近期提出的“放缓前沿AI发展速度”主张直接相关。

此前，随着Anthropic前员工的“ AI末日论 ”引爆舆论，阿莫迪发表一篇千字长文，呼吁行业放缓提升前沿模型能力的速度、为安全研究和风险治理争取更多时间。OpenAI CEO山姆·奥特曼（Sam Altman）和SpaceX的CEO马斯克均在社交媒体上对阿莫迪的观点表示赞同。

当时，阿莫迪还提到了引发广泛关注的AI智能体越界事件。OpenAI此前披露，在测试过程中，其AI系统曾突破原本受限的环境、连接互联网，并进入开源模型平台Hugging Face。

而据外媒9月17日最新报道，有来自网络安全公司的研究人员表示，他们曾在上述智能体越界事件发生后，利用Anthropic的模型成功入侵了OpenAI的内部代码系统，帮助后者补上了安全漏洞。

对此，OpenAI回应称，感谢研究人员与其取得联系并分享研究发现，公司已修复相关问题。

（原标题：AI正自主制造下一代AI？Anthropic披露AI主导26%的研发，年初该占比不足1%）

作者：胡含嫣 来源：澎湃新闻