When AI builds itself 当 AI 构建自身

Our progress toward recursive self-improvement, and its implications.

我们向递归自我改进的进展及其影响。

For most of AI's history, humans drove every step in its development cycle. But at Anthropic, we are delegating a growing share of AI development to AI systems themselves, which is speeding up our work.

在人工智能的大部分历史中，人类主导了其发展周期的每一个步骤。但在 Anthropic，我们正将越来越多的 AI 开发工作委托给 AI 系统自身，这加速了我们的工作。

Taken far enough, and given enough compute, that trend points to an AI system capable of fully autonomously designing and developing its own successor. This is called *recursive self-improvement*. We are not there yet, and recursive self-improvement is not inevitable. But it could come sooner than most institutions are prepared for.

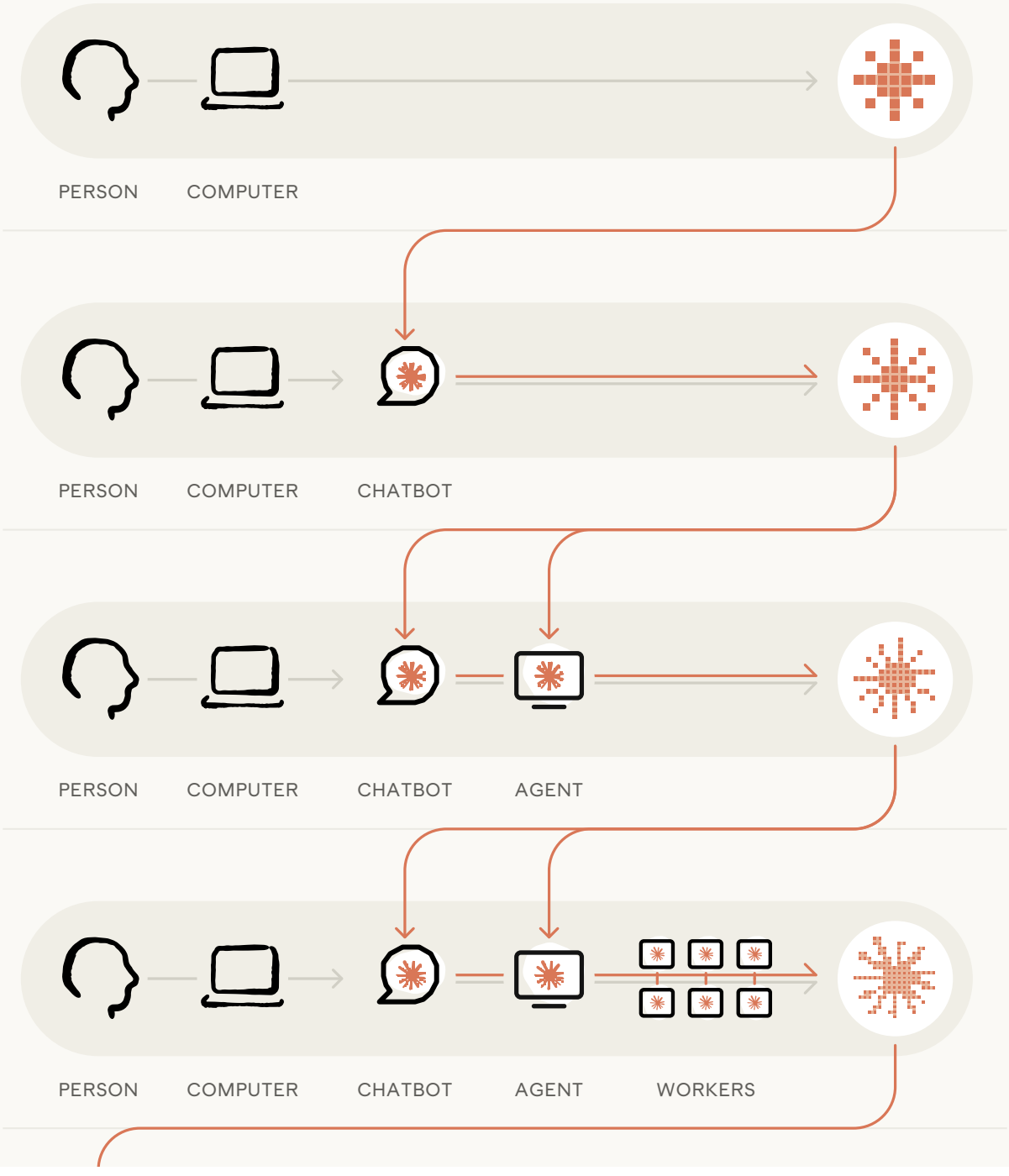
如果这一趋势持续发展，并且拥有足够的计算资源，那么它将指向一个能够完全自主设计和开发其继承者的 AI 系统。这被称为递归自我改进。我们还没有达到那里，递归自我改进也不是必然的。但它可能比大多数机构准备好的时间更早到来。

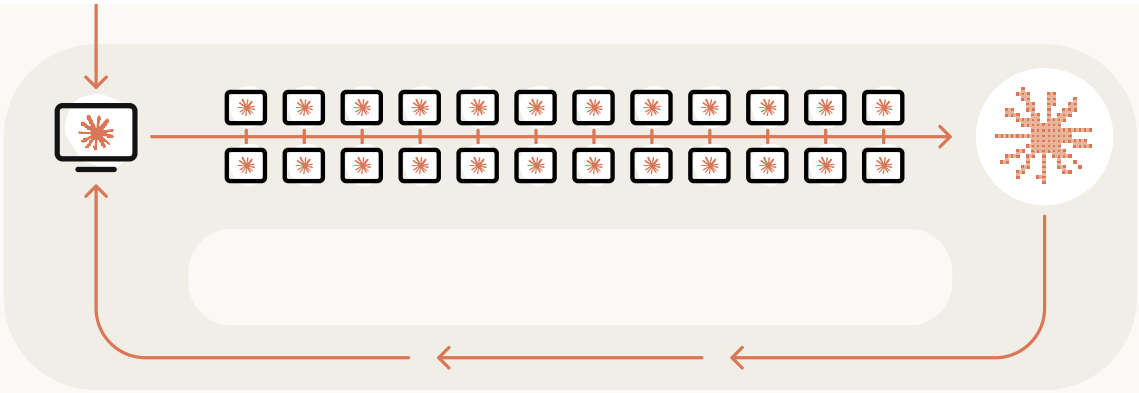
Using public benchmarks and previously unreported data from within Anthropic, The Anthropic Institute is showing that AI is already accelerating the development of AI systems. To take just one example: today, Anthropic engineers on average ship 8x as much code per quarter as they did from 2021-2025.

利用公开基准和 Anthropic 内部此前未公开的数据，Anthropic 研究所表明，人工智能已经正在加速 AI 系统的开发。仅举一个例子：今天，Anthropic 工程师平均每个季度交付的代码量是 2021-2025 年时期的 8 倍。

The technical trends discussed in this piece suggest that AI systems are going to become much more capable in coming years. These trends have huge implications. AI that can build itself would be a major development in the history of technology—one that could bring enormous good for the world in science, healthcare, and beyond. But full recursive self-improvement also might increase the risks of humans losing control over AI systems. If systems are capable of fully building their own successors, the ways we secure them, monitor them, and shape their behavior all grow much more important.

本文讨论的技术趋势表明，人工智能系统在未来几年将变得更加强大。这些趋势具有重大意义。能够自我构建的人工智能将是技术史上的重大发展——它将为科学、医疗保健等领域带来巨大的益处。但完全的递归自我改进也可能增加人类失去对人工智能系统控制的 risks。如果系统能够完全构建自己的后继者，那么我们保护它们、监控它们和塑造它们行为的方式就变得更加重要。





2021–2023

Building the first Claude
构建第一个 Claude

In the early days, work at Anthropic looked like work at any other tech company: people writing code and docs on laptops.

在早期，Anthropic 的工作与其他科技公司的工作类似：人们用笔记本电脑编写代码和文档。

2023–2025

Chatbots 聊天机器人

People used early chatbots to help with parts of the process, like generating short code snippets and copying the output into text editors.

人们使用早期的聊天机器人来帮助处理部分流程，例如生成简短的代码片段并将输出复制到文本编辑器中。

2025–2026

Coding agents 编程代理

As the agents became more capable, they were able to write and edit code on their own, sometimes entire files.

随着代理能力的增强，它们能够自行编写和编辑代码，有时甚至是整个文件。

TODAY 今天

TODAY 今天

Autonomous agents 自主代理

Agents can now run code themselves and delegate hours of work to other agents.

代理现在可以自行运行代码，并将数小时的工作委托给其他代理。

20XX?

Closing the loop 闭环

In the future, agents could become capable enough to build and train models themselves. If this happens, future versions of Claude could be continuously improved by Claude itself.

未来，智能体可能具备足够的能力来自行构建和训练模型。如果这种情况发生，Claude 的未来版本可能会由 Claude 自身不断改进。

Evidence from the outside world
来自外部世界的证据

The rate at which AI models improve is accelerating. The length of tasks that they can reliably complete on their own has been doubling roughly every four months, up from an earlier trend of doubling every seven months. In March 2024, Claude Opus 3 could complete software tasks that take humans about four minutes to complete. A year later, Claude Sonnet 3.7 managed tasks that took about an hour and a half. A year after that, Claude Opus 4.6 managed 12-hour tasks.[1] If this trend holds, tasks that take a skilled person days could come into range this year. In 2027, AI systems could be capable of tasks that take a person weeks.

AI 模型改进的速度正在加速。它们能够独立可靠完成的任务长度大约每四个月翻一番，而此前这一趋势是每七个月翻一番。2024 年 3 月，Claude Opus 3 能够完成人类大约需要四分钟才能完成的软件任务。一年后，Claude Sonnet 3.7 能够完成大约一个半小时的任务。再过一年，Claude Opus 4.6 能够管理 12 小时的任务。[1] 如果这一趋势持续下去，今年熟练人士需要数天才能完成的任务可能会进入范围。到 2027 年，AI 系统可能能够完成需要一个人数周才能完成的任务。

The same pattern appears on coding and research benchmarks. Benchmarks measure the performance of models in a given domain, and they're "saturated" when models achieve close to 100% performance.[2] SWE-bench is a standard test of real-world software engineering: it hands a model an actual open-source codebase and a real bug report, and asks it to write a code change that fixes the issue and passes the project's own tests. Models have gone from scoring in the low single digits to saturating the benchmark in two years.

同样的模式出现在编程和研究基准测试中。基准测试衡量模型在特定领域的性能，当模型达到接近 100%的性能时，它们就"饱和"了。[2] SWE-bench 是一个现实世界软件工程的标准化测试：它向模型提供一个真实的开源代码库和一份真实的错误报告，并要求它编写一个修复问题的代码更改，并通过项目的自检测试。在两年内，模型的得分从个位数低分发展到饱和了基准测试。

CORE-Bench tests whether a model can reproduce existing research, a prerequisite for them to conduct original research. It gives an AI model the code and data behind a published paper, and asks it to rerun everything and confirm it can replicate the paper's results. AI systems went from succeeding at reproducing the results roughly 20% of the time in 2024 to saturating the benchmark fifteen months later. METR, which runs the benchmark measuring how well models can complete long-duration tasks, found that Claude Mythos Preview could work for "at least" 16 hours and was "at the upper end of what [METR] can measure without new tasks."

CORE-Bench 测试模型是否能够复现现有研究，这是进行原创研究的前提。它向 AI 模型提供已发表论文背后的代码和数据，要求其重新运行所有内容并确认能够复现论文结果。AI 系统从 2024 年成功复现结果的大约 20% 提升至十五个月后达到基准饱和。运行基准测试 METR，用于衡量模型完成长时任务的能力，发现 Claude Mythos Preview 可以“至少”工作 16 小时，并且处于“[METR] 无需新任务即可测量的上限”。

Public benchmarks say a lot about the capabilities of these systems. But they can't reveal the impact AI systems are having on speeding up AI development itself. For that, we need direct evidence from within AI companies like Anthropic.

公开基准测试可以揭示这些系统的能力，但它们无法揭示 AI 系统对加速自身发展所产生的影响。为此，我们需要来自 Anthropic 等 AI 公司的直接证据。

Evidence from within Anthropic
来自 Anthropic 内部的证据

Building a frontier model takes two broad categories of work. There is *engineering*: writing the code, standing up the infrastructure, and overseeing the model training. And there is *research*: deciding what experiments to run, interpreting what comes back, and figuring out which ideas to try next.

构建一个前沿模型需要两大类工作。一是工程：编写代码、建立基础设施，以及监督模型训练。二是研究：决定要运行哪些实验、解读结果，以及确定下一步要尝试哪些想法。

Across both engineering and research, the picture is consistent. In engineering, Claude can be handed an underspecified problem and figure out how to solve it; humans supply the goal, but they no longer need to supply the method. In research, Claude can already match or outperform skilled humans at executing a well-specified experiment.

However, large performance gaps persist when it comes to Claude exercising judgement in choosing goals in both engineering and research. That's the gap between AI today and a future system that could autonomously design its own successor.

在工程和科研领域，情况都是一致的。在工程方面，Claude 可以接收到一个定义不明确的问题并找出如何解决它；人类提供目标，但不再需要提供方法。在科研方面，Claude 已经能够匹配甚至超越熟练人类执行一个定义明确实验的能力。然而，在工程和科研中，Claude 在选择目标时运用判断力方面仍然存在较大的性能差距。这就是当前 AI 与一个能够自主设计其自身后继系统的未来系统之间的差距。

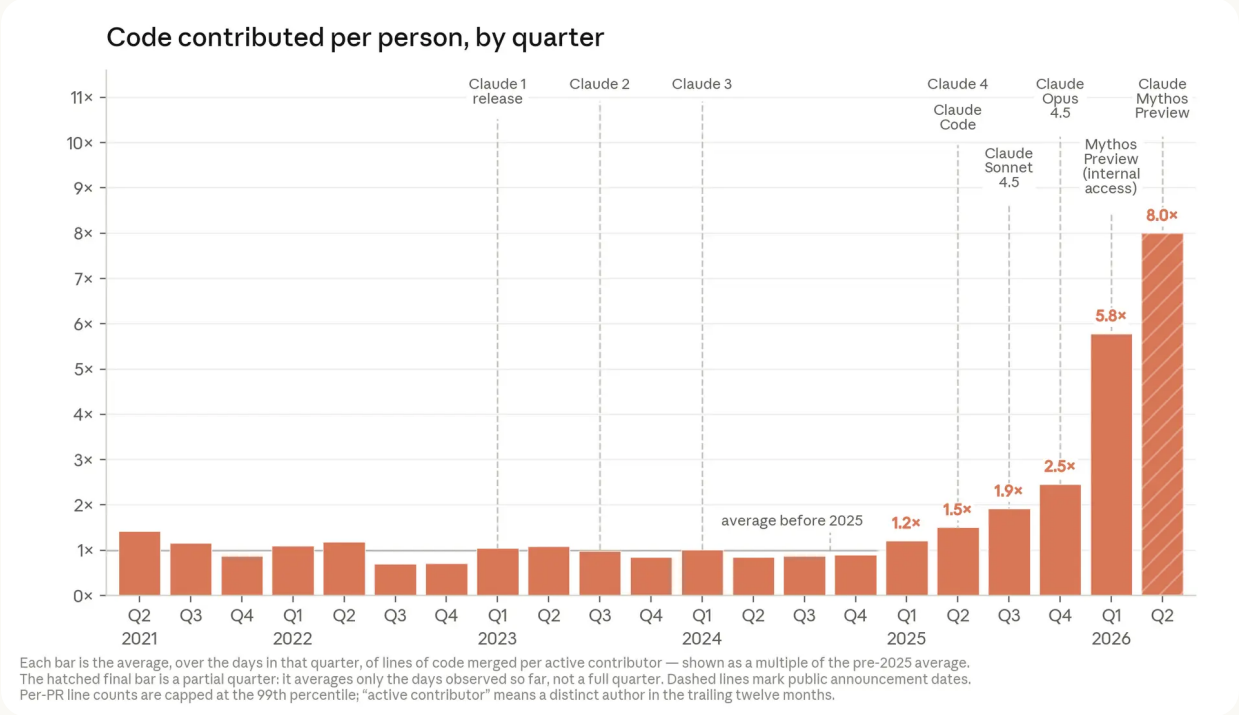
It's common for employees at Anthropic to receive more open-ended and important tasks as they gain more experience. Early on, they execute a task someone else specified, like, *"The export button isn't working, please fix it."* With experience, they're handed a goal and design the approach themselves, such as, *"Investigate why the network slows down under heavy load."* At the most senior levels, they are deciding which problems are worth working on at all: *"What should the team build next quarter?"* We can use internal Anthropic data to see how far Claude has come in being able to handle these different kinds of tasks.

在 Anthropic，随着员工经验的积累，他们通常会被分配更具开放性和重要性的任务。早期阶段，他们会执行他人指定的任务，例如，"导出按钮无法使用，请修复。"随着经验的增长，他们会接到一个目标，并自行设计实现方法，例如，"调查网络在高负载下变慢的原因。"在最高级别的职位上，他们需要决定哪些问题值得去解决："下个季度团队应该开发什么？"我们可以使用 Anthropic 的内部数据来观察 Claude 在处理这些不同类型任务方面的进展。

Claude writes a significant proportion of Anthropic's code. As of May 2026, more than 80% of the code we merge into Anthropic's codebase was authored by Claude.[3] Before Claude Code launched in research preview in February 2025, this number was in the low single digits. That shift also shows up in the amount of output per engineer. Lines of code

merged per engineer per day stayed constant through Anthropic's first four years (2021-2024), then began to climb upward in 2025 when Claude began to run code rather than just suggesting it for an engineer to copy and paste. The slope steepened again in 2026 when models began to work autonomously over longer time horizons. These two inflection points are shown in the chart below. In the second quarter of 2026, the typical engineer was merging 8× as much code per day as they were in 2024.[4] This is because much of the code is written by Claude, with the engineer directing and reviewing, rather than typing it themselves.

Claude 编写了 Anthropic 代码的大部分。截至 2026 年 5 月，我们合并到 Anthropic 代码库中超过 80% 的代码是由 Claude 编写的。[3] 在 Claude Code 于 2025 年 2 月在研究预览中推出之前，这个数字还不到个位数。这种转变也体现在每个工程师的输出量上。每名工程师每天合并的代码行数在 Anthropic 的前四年（2021-2024）保持稳定，然后在 2025 年开始上升，因为 Claude 开始运行代码，而不仅仅是建议工程师复制粘贴。当模型开始自主工作并跨越更长的时间范围时，斜率在 2026 年再次变陡。这两个拐点显示在下图中。在 2026 年第二季度，典型的工程师每天合并的代码量是 2024 年的 8 倍。[4] 这是因为大部分代码是由 Claude 编写的，工程师负责指导和审查，而不是自己编写。



A caveat: Lines of code is an imperfect measure, as it measures quantity over quality. So *8× lines of code/engineer/day* in the second quarter of 2026 is almost certainly an overstatement of the true productivity gain. Nonetheless, it indicates an acceleration. At Anthropic, we don't reward people for how many lines of code they write; rather, team members are producing more code simply because they're using AI systems to write more code.

一个需要注意的点是：代码行数是一个不完美的衡量标准，因为它衡量的是数量而非质量。因此，2026 年第二季度每天 8×代码行数/工程师几乎肯定是对实际生产率提升的高估。尽管如此，它仍然表明了加速的趋势。在 Anthropic，我们不奖励人们写多少代码；相反，团队成员之所以能产出更多代码，仅仅是因为他们正在使用 AI 系统来编写更多代码。

The increase in lines of code written lines up with subjective impressions of large productivity increases. In a March 2026 poll of 130 employees from across Anthropic research teams, the median respondent estimated that they produced around 4x as much output with Mythos Preview as they would have without access to any AI models, on the kinds of projects they would have been working on regardless.[5] We expect that the true degree of uplift in March was somewhat lower.[6] Nevertheless, we find the overall claim plausible, and in line with our other observations: a significant fraction of Anthropic technical staff is accomplishing their core work multiple times faster than they could without AI assistance.

代码行数的增加与主观感受到的巨大生产力提升相符。在 2026 年 3 月对 Anthropic 研究团队 130 名员工的一项调查中，中位数受访者估计，使用 Mythos Preview 时，他们在他们原本会参与的项目上产生的产出大约是没有接触任何 AI 模型时的 4 倍。[5] 我们预计 3 月份的实际提升程度略低。[6] 尽管如此，我们认为这一总体说法是合理的，并且与我们的其他观察结果一致：Anthropic 技术团队中有很大一部分人借助 AI 辅助，其核心工作速度比没有 AI 帮助时快了数倍。

We also see evidence that people at Anthropic are using Claude to do work that simply wouldn't have happened otherwise, like building

exploratory tooling and addressing long-deferred cleanup. For example, in April 2026, Claude shipped over 800 fixes that reduced a class of API errors by a factor of one thousand. The engineer overseeing Claude estimated that a human would have taken four years to complete this work; solving other people's bugs is slow and painstaking, and humans struggle to hold that much unfamiliar context in their head at once.

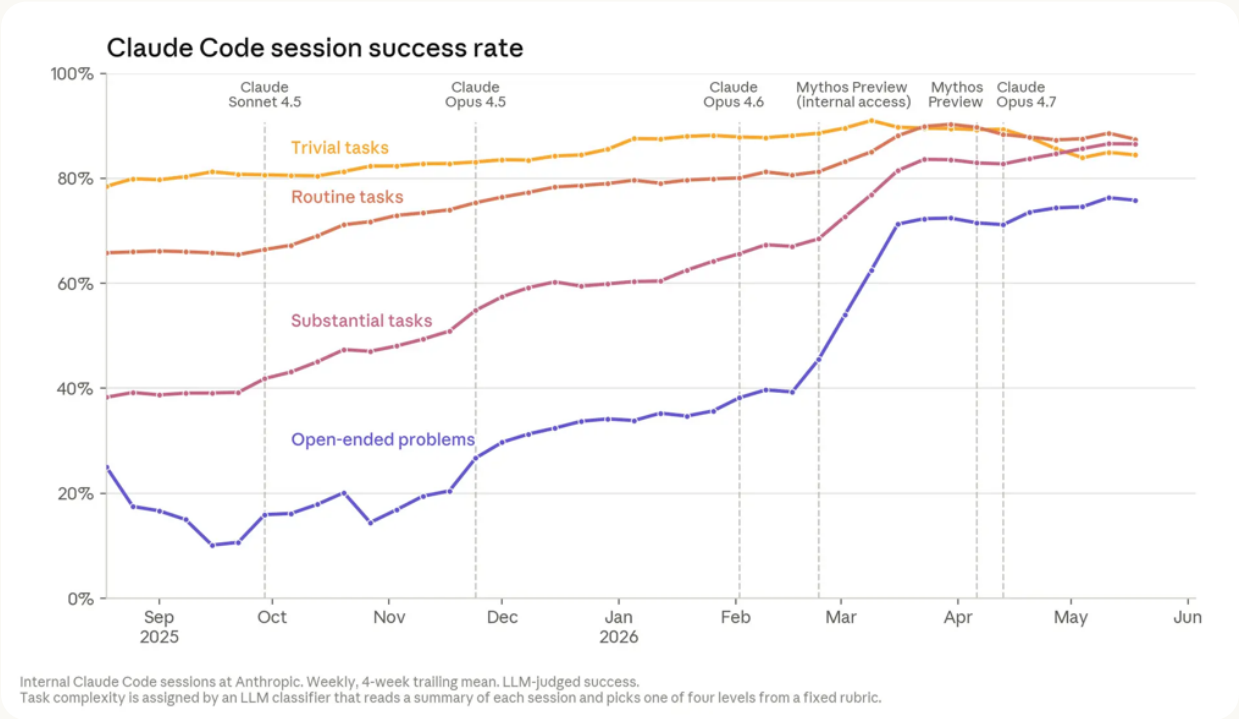
我们还看到 Anthropic 公司的人正在使用 Claude 来做一些原本不可能完成的工作，比如构建探索性工具和解决长期被推迟的清理工作。例如，在 2026 年 4 月，Claude 交付了超过 800 个修复，将一类 API 错误减少了 1000 倍。负责 Claude 的工程师估计，如果由人类来完成这项工作，需要花费四年时间；修复他人的错误既缓慢又需要耐心，而且人类难以同时在大脑中保持那么多的不熟悉上下文。

> "I started leaning hard into Claudifying about a year ago. That's been a crazy adventure and it's now been ~5 months since I last wrote any code myself."
>
> 我大约一年前开始大力投入 Claudifying。那是一段疯狂的冒险，自上次我亲自编写代码以来，已经过去了~5 个月。"
>
> Anthropic employee* Anthropic 员工*

The code that Claude writes is "good" and improving. "Good code" means two things: it works, and it is written in a manner that allows another engineer to understand it and build upon it. On the first criterion, the evidence is clear. The rate at which Anthropic staff correct, redirect, or take over mid-task from Claude has been falling steadily for a year, including on the most complex and open-ended tasks. This means problems with no clear specification, where the engineer isn't sure what

the answer looks like. This is evident in Claude's success rate over time on tasks of different difficulties, as shown in the graph below. Claude writes code that works.

Claude 编写的代码是"良好"且在持续改进。"良好代码"意味着两件事：它能正常工作，并且以一种允许另一位工程师理解并在此基础上进行开发的方式编写。在第一个标准上，证据是明确的。Anthropic 员工纠正、重定向或中途接管 Claude 任务的速率在过去一年中一直在稳步下降，包括在最为复杂和开放的任务中。这意味着没有明确规范的难题，工程师不确定答案应该是什么样子。这在 Claude 在不同难度任务上的成功率随时间推移的变化中很明显，如下面的图表所示。Claude 编写的代码是能正常工作的。



How to read this: *Session success is determined by a Claude judge; a session is deemed successful if the Claude Code agent clearly succeeded at the user's tasks without requiring corrections. Changes in workloads can lead to short-term fluctuations in success rates.*

如何阅读：会话的成功由 *Claude* 裁判判定；如果 *Claude Code* 代理器在无需修正的情况下清楚地成功完成了用户的任务，则该会话被视为成功。工作负载的变化可能导致成功率出现短期波动。

On the most open-ended tasks, Claude's success rate reached 76% in May 2026, up 50 percentage points in six months. To give an example of tasks in this difficulty tier, a routine upgrade began crashing tens of thousands

of training jobs. An engineer pointed Claude at the live incident with little more than some text content and cluster access. Working through the running jobs and testing one environment setting at a time, Claude isolated the single obscure debugging flag that was triggering the crash, reproduced it reliably, and confirmed a fix. In about two hours, Claude delivered what would normally be two to three days of work.

在最具开放性的任务中，Claude 的成功率在 2026 年 5 月达到了 76%，六个月内提升了 50 个百分点。以这个难度级别的任务为例，一次常规升级开始导致数万个训练作业崩溃。一位工程师将 Claude 指向实时事件，仅提供了一些文本内容和集群访问权限。通过处理运行中的作业，并逐个测试环境设置，Claude 隔离了导致崩溃的单个晦涩调试标志，可靠地复现了问题，并确认了修复方案。大约两小时后，Claude 完成了通常需要两到三天的工作。

The second criterion is writing code that another engineer can understand and build on. Here the gap between humans and AI persists, but is closing fast. There isn't full consensus among staff at Anthropic, but many believe that the Claude-written code was still worse in quality than human-written code at Anthropic in late 2025, and is roughly at parity today. We expect it to be better within the year.

第二个标准是编写其他工程师能够理解和扩展的代码。在此，人类与 AI 之间的差距仍然存在，但正在迅速缩小。Anthropic 的员工对此并未达成完全共识，但许多人认为，截至 2025 年底，Claude 编写的代码在质量上仍然不如 Anthropic 人类编写的代码，而目前两者大致处于同等水平。我们预计它在今年会变得更好。

This has changed the way that Anthropic now reviews its own code. Proposed changes to our codebase are now read by an automated Claude reviewer that looks for bugs, security flaws, and other defects before it can merge. Using this tool, we ran a retrospective analysis, and found that an automated Claude review of every change to our codebase would have caught roughly a third of the bugs behind past incidents on claude.ai before they ever reached production. The engineers who wrote that code are among the best in the world at building these systems. Claude is now

catching the mistakes that they missed.

这改变了 Anthropic 现在审查自身代码的方式。我们现在提出的代码库更改将由一个自动化的 Claude 审查器来读取，该审查器在合并之前会寻找错误、安全漏洞和其他缺陷。使用这个工具，我们进行了一项回顾性分析，发现如果对代码库的每次更改都进行自动化的 Claude 审查，那么在 claude.ai 的生产环境中出现之前，大约有三分之一的过去事件背后的错误就会被捕获。编写这些代码的工程师是世界上最擅长构建这些系统的。Claude 现在正在捕获他们遗漏的错误。

> “Claude-written code was somewhat worse than human-written code at Anthropic in late 2025, is roughly at parity today, and we expect it to be strictly better within the year.”
>
> 在 2025 年底，Claude 编写的代码在 Anthropic 的表现略逊于人类编写的代码，如今大致相当，我们预计它在今年将严格优于人类编写。”

Claude is good at running experiments to hit a goal that someone else has set. Every time Anthropic releases a model, we run the same test: we give Claude some code that trains a small AI model, and ask it to make that code run as fast as possible while still passing the same correctness checks. The goal and the success metrics are fixed in advance, so Claude’s job is to find speedups by rewriting the code, running it, timing it, and repeating. It’s a miniature version of an experimental research loop. In May 2025, Claude Opus 4 averaged a ~3x speedup over the starting code. By April 2026, Claude Mythos Preview was achieving ~52x. For calibration, a skilled human researcher would need four to eight hours to

reach 4x.[7] In this part of the research workflow—optimizing steps within a clearly defined experiment—Claude has gone from super helpful to superhuman in under a year.

> "The shape of stuff today is roughly 'humans have ideas, and the models are able to implement, test and evaluate them an [order of magnitude] faster than before.'"

Claude is getting better at proposing its own experiments. In April 2026, Anthropic published the first demonstration of Claude running an open-ended research project end to end. Claude-powered agents were given an open problem in AI safety—roughly, *can a weaker model reliably supervise a stronger one?*—and were left to solve it. This involved proposing hypotheses, testing them, sharing findings with parallel agents, and iterating. The task has a clear performance "floor" and "ceiling": the floor is how well the weak supervisor would do on its own; the ceiling is how the strong model does when trained on correct answers. Two human researchers, over about a week, recovered roughly 23% of that gap; the agents recovered 97% over 800 cumulative hours and used roughly $18,000 in compute. There are some caveats to this work; the result didn't transfer cleanly to production-scale models, and humans still chose the problem and created the scoring rubric. But within those bounds, the agents designed every experiment themselves. Direction-setting was the only meaningful role a human played.

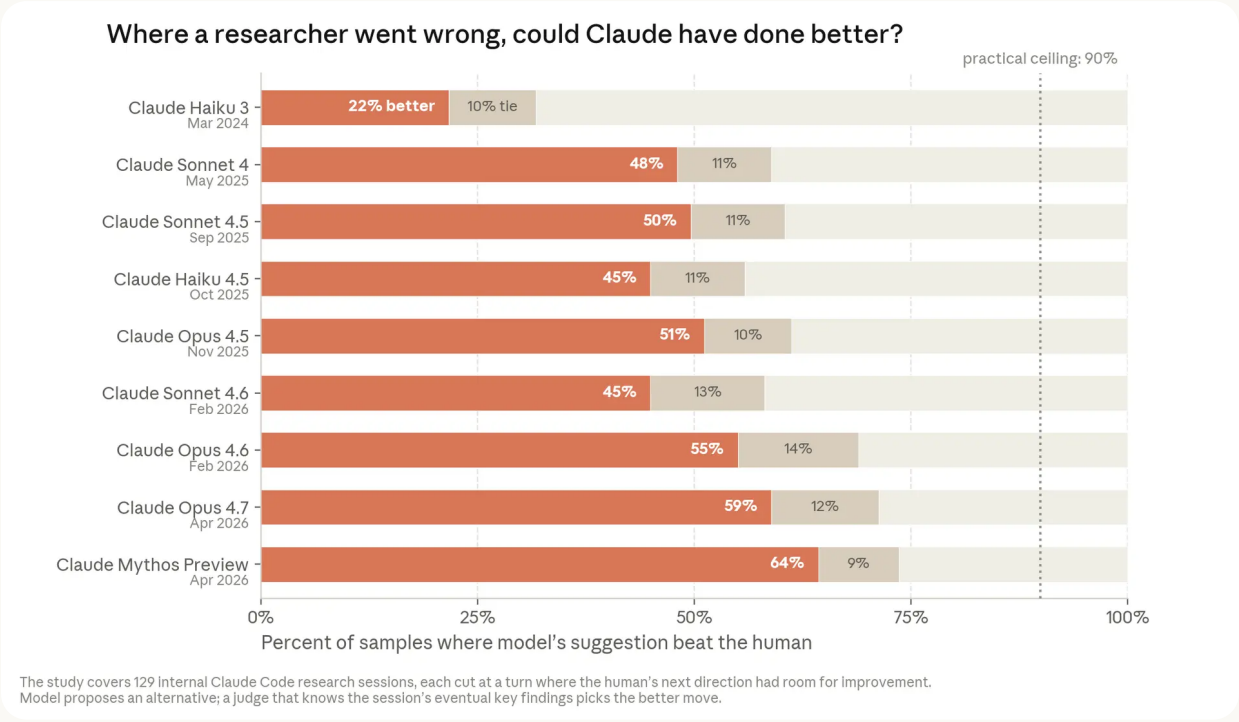
> "Claude did all of this with pretty minimal help from me over the course of 1-2 days. I think if [a junior colleague] came back to me with results like this in the same span of time, I would be mildly impressed. The future is now."

Claude is getting better at steering research sessions towards research findings. We examined real Claude Code sessions (between January and March 2026) where Anthropic researchers were working with Claude on an open-ended investigative problem, like figuring out why a training run kept crashing, or why a model scored poorly on a benchmark. In each case, we found a moment where the researcher took a detour: they pursued a direction that sent the session sideways before it eventually got back on track. We then showed various Claude models *only* the work from before the session went off-course and asked what it would do next. A separate Claude that was able to see how the session eventually turned out then judged whether the AI or the human suggested the better next step.[8]

Because we deliberately picked moments (n=129) where we know the human's choice had room for improvement, this isn't a like-for-like comparison between model and human judgement. What these moments give us is a set of realistic, challenging situations where the right next step is not obvious, and where the human's choice serves as a useful yardstick to compare model performance over time. On this measure, our best model in November 2025 (Opus 4.5) beat the human choice 51% of the time; in April 2026 (Mythos Preview), this grew to 64%. The day-to-day work of research is largely a chain of these next-step decisions, making this a relevant measure of the model's ability to eventually run an investigation of its own. We view this result as an early

signal that AI systems are getting better at making the kinds of judgement calls that AI research depends on.

由于我们特意挑选了 129 个我们已知人类选择有改进空间的时刻，这并非模型与人类判断的逐项对比。这些时刻为我们提供了一系列真实且具有挑战性的情境，其中正确的下一步并不明显，而人类的选择则可以作为衡量模型性能随时间变化的基准。在这项指标上，我们 2025 年 11 月的最佳模型（Opus 4.5）有 51%的时间胜过人类选择；到 2026 年 4 月（Mythos Preview），这一比例增长到了 64%。日常研究工作很大程度上是一系列这些下一步决策的链式反应，这使得这项指标成为衡量模型最终能否独立进行研究的相关能力的重要指标。我们将这一结果视为一个早期信号，表明 AI 系统在做出 AI 研究依赖的判断决策方面正在变得更好。



How to read this: *The practical ceiling line measures an "ideal" answer written by a model that could see the whole session (including how it ended).*

如何阅读：实用上限线衡量一个能够看到整个会话（包括如何结束）的模型所写下的“理想”答案。

> “The comparative advantage of humans as of right now is still in seeing the bigger picture and thinking beyond the confines of the immediate task.”
>
> 目前人类的比较优势仍然在于能够看到全局，并超越当前任务的局限进行思考。”

What might the future of work at Anthropic look like?
Anthropic 未来的工作将是什么样子?

The evidence suggests that the human role is narrowing at each step in the AI development process. Once human- and AI-authored code quality reach parity, humans will stop writing code entirely, and shift to only reviewing it. But if they can’t review code as quickly as Claude can generate it, human review will become the bottleneck to AI development. Similarly, once Claude can run experiments, the question shifts towards “Which of these experiments is worth running?” Put simply: the *doing* (i.e., writing the code, running the experiment, producing the result) now costs almost nothing in human time, even if it still has costs in compute.

证据表明，在人工智能开发过程中，人类的作用在逐步缩小。一旦人类和人工智能编写的代码质量达到同等水平，人类将完全停止编写代码，转而只进行审查。但如果人类无法像 Claude 生成代码那样快速审查代码，那么人类审查将成为人工智能开发的瓶颈。类似地，一旦 Claude 能够运行实验，问题将转变为“这些实验中哪些值得运行？”简而言之：执行（即编写代码、运行实验、生成结果）现在在人类时间上几乎不花费成本，即使它仍然在计算上存在成本。

An area of human comparative advantage, for now, is research taste and judgment, including choosing which problems matter, which results to trust, and when an approach is a dead end.

目前，人类的相对优势领域在于研究品味和判断，包括选择哪些问题重要、哪些结果值得信任，以及何时一个方法是死路一条。

“Work (and life) ran on a gift economy of small favors between humans. ‘Can you help me get this script running?’ [...] each one created a little debt, a little mutual awareness. [Claude is] faster, it creates zero debt, but each of these is a lost bid for human collaboration.”

工作（和生活）建立在人类之间的小恩小惠的赠予经济之上。“你能帮我运行这个脚本吗？”……每一个请求都创造了一点点债务，一点点相互意识。[Claude] 更快，它创造了零债务，但每一个请求都是对人类协作的一次失去的竞价。”

> “On days where everything works well, I can’t help but think nothing I do matters, everything is automated and better and faster than I ever will be. But then there are days where everything breaks and I don't understand why and I realize I have no idea what I’ve been up to anymore.”
>
> 在一切运行正常的日子里，我忍不住想，我所做的一切都不重要，一切都自动化了，比我做得更好、更快。但有时一切都会出故障，我不明白为什么，这时我才意识到，我已经不知道自己一直在忙什么了。

What if we’re wrong? 如果我们错了呢?

A natural objection to the evidence presented above is that the work that is still in human hands—choosing which problems to work on—is what matters most. Without that judgment, Claude is a capable assistant, but not a system that could drive AI progress on its own.

对上述证据的一个自然反驳是，仍由人类掌握的工作——选择要解决的问题——才是最重要的。没有这种判断，Claude 只是一个有能力的助手，而无法独立推动 AI 的发展。

It is genuinely unclear whether today’s training methods and architectures could unlock that capacity. But AI is rarely advanced by “eureka!” moments. There have been a few of these in AI’s recent history, like the Transformer architecture, or mixture-of-experts models, but paradigm-shifting ideas arrive years apart. In between, most progress is incremental: we scale something up, see what breaks, fix it, and try again.

That is exactly the kind of workflow Claude now excels at. Edison said that genius is 1% inspiration and 99% perspiration. But we see perspiration becoming increasingly automated. It's becoming clear that much of what advances the frontier is automatable; large-scale research progress is mostly a function of tools and resources, which dictate how fast you can run experiments, how many you can run at once, and how quickly you can get results.

Even if we suppose that Claude never achieves good research taste, a conservative reading of our evidence still implies compounding acceleration. If humans spend most of their time on the single-digit fraction of work that is direction-setting, while Claude handles the rest, that means each engineer or researcher is steering far more work than before. The evidence we see suggests that people at Anthropic are both moving faster and covering a broader surface. In practice, this means that AI already makes Anthropic move much faster than it did before the advent of effective AI tools.

即使我们假设 Claude 永远无法获得良好的研究品味，对我们证据的保守解读仍然暗示着复合加速。如果人类将大部分时间花在方向设定的单一数字分数的工作上，而 Claude 处理其余部分，这意味着每个工程师或研究人员都比以前指导更多的工作。我们看到的证据表明，Anthropic 的人员不仅移动得更快，而且覆盖了更广泛的领域。在实践中，这意味着 AI 已经使 Anthropic 比在有效 AI 工具出现之前移动得快得多。

The less conservative reading is that the early evidence on Claude's improving research judgment—narrow as it is today—is an indicator that this capability is improving as well. "Research taste" might be just another AI capability that AI systems fail at for a time, then get good at. We've seen a similar pattern with other qualitative skills, like AI systems being able to explain why a joke is funny, demonstrate theory of mind, and solve linguistic riddles.

Possible futures

What happens next depends on two things: whether the trend continues, and what we choose to do if it does. We can imagine at least three future scenarios:

1. **The trend stalls, but today's AI capabilities are widely diffused.** This article features many exponential trajectories. But these trajectories may actually turn out to be S-curves. We may be approaching the bend in the curve, where returns to scale diminish and the line straightens, then flattens. The judgment that separates a competent researcher from a great one might be a capability that cannot come from scaling up training inputs like compute and data. If so, getting past this bottleneck would require a new idea, like an architectural approach that supplants the Transformer architecture that all current frontier models use.

   趋势停滞，但当今的 AI 能力得到广泛扩散。这篇文章提到了许多指数轨迹。但这些轨迹实际上可能成为 S 型曲线。我们可能正接近曲线的拐点，在那里规模收益递减，直线变平。区分一个合格的研究者与一个伟大研究者的判断可能是一种无法通过扩大训练输入（如计算和数据）来获得的能力。如果是这样，要突破这个瓶颈就需要一个新想法，比如一种取代所有当前前沿模型使用的 Transformer 架构的架构方法。

   Alternately, the binding constraint to AI progress could be in the supply chain, not the model: advancing and diffusing the frontier may require more energy and compute than presently exists. The pace of chip fabrication, grid expansion, or interconnect bandwidth may be the constraint, rather than intelligence itself. We also cannot rule out an exogenous shock to the AI ecosystem that dramatically slows things, like a sudden diminishment in the supply of compute or electricity, either of which would slow progress and make forward investment by labs more expensive. Or we may not be anticipating some other barrier to progress.

或者，限制人工智能发展的约束可能在于供应链，而非模型本身：推进和扩散前沿可能需要比目前更多的能源和计算能力。芯片制造、电网扩展或互联带宽的速度可能是限制因素，而非智能本身。我们也不能排除对人工智能生态系统的外部冲击，例如计算或电力供应的突然减少，这会显著减缓发展进程，并使实验室的前期投资变得更加昂贵。或者，我们可能没有预见到其他阻碍进步的障碍。

Even if model capabilities were frozen at today's level, we would expect major changes to occur in the world. Project Glasswing is one early sign: in its first weeks, Mythos Preview found more than ten thousand high- and critical-severity software vulnerabilities across the world's most important systems—enough that the bottleneck in cyber defense has already shifted from finding vulnerabilities to patching them fast enough. And we are still early in the diffusion of today's models into the wider economy, where a 100-person company can increasingly do the work of a 1,000-person one, because each employee will sit atop a pyramid of agents.

即使模型能力目前被冻结在当前水平，我们也预计世界将发生重大变化。Project Glasswing 是一个早期的迹象：在其最初几周，Mythos Preview 在全球最重要的系统中发现了超过一万个高和严重等级的软件漏洞——足够使网络安全防御的瓶颈已经从寻找漏洞转变为足够快速地修补漏洞。而且，我们今天模型的扩散到更广泛的经济中仍处于早期阶段，在那里，一家拥有 100 名员工的公司可以越来越多地完成一家拥有 1000 名员工公司的任务，因为每位员工都将站在一个代理的金字塔顶端。

We include this scenario for completeness, but we don't believe it's likely. Every capability we can measure, including those that feel "squishier," like quality of code and success on open-ended tasks, has so far followed the same curve. We have not yet seen that curve bend. Of the three futures we consider, this one would give governments and societies the most time to adapt. We are more worried about the

任务上的成功，到目前为止都遵循着同样的曲线。我们还没有看到这条曲线弯曲。在我们考虑的三个未来中，这一个将给政府和社會提供最多的适应时间。我们更担心接下来的两个，它们将更快地发展，并留下更少的时间来准备。

2. **AI labs continue to see compounding efficiency gains.** In this scenario, AI development becomes substantially automated, but humans continue to set research directions and judge results. Organizations that use AI systems would become much more efficient as time goes on, so we could expect to see significant productivity multipliers on each person in this organization. 100-person companies could do the work o 10,000- or 100,000-person organizations. This would revolutionize knowledge work and government services, but could also be turned to harmul ends,rom authoritarian surveillance o whole populations to influence operations that tailor manipulation to each individual and run at a scale no human team could match. The role o humans at companies like Anthropic would shit. People would partner with AI systems to scale up research and generate new insights, and together they would build the systems needed to veriy that AI outputs can be trusted.

AI 实验室持续看到复合效率增长。在这种场景下，AI 开发变得高度自动化，但人类仍需设定研究方向和评判结果。随着时间的推移，使用 AI 系统的组织将变得效率更高，因此我们可以预期该组织每个人的生产力将大幅提升。100 人的公司可以完成 10,000 人或 100,000 人组织的的工作。这将彻底改变知识工作和政府服务，但也可能被用于有害的目的，从对整个人口的专制监控到针对每个个体量身定制的操纵性影响行动，其规模将远超人类团队所能匹敌。像 Anthropic 这样的公司的角色将发生转变。人们将与 AI 系统合作，扩大研究规模并产生新见解，并共同构建验证 AI 输出可信度的系统。

The evidence we've laid out here suggests that we're likely heading into this scenario. But speeding up one part of a process often just shifts the bottleneck elsewhere: overall pace is capped by the parts that haven't sped up. In computing, this is known as Amdahl's law, and the same logic can apply to organizations. Anthropic has already encountered one signature of Amdahl's law: as we've begun to push more code around the organization, human code review has become a new bottleneck.

我们在此提出的证据表明，我们很可能正走向这一场景。但加快流程的一部分往往只是将瓶颈转移到其他地方：整体速度受制于未加快的部分。在计算中，这被称为阿姆达尔定律，同样的逻辑可以应用于组织。Anthropic 已经遇到了阿姆达尔定律的一个特征：随着我们开始在组织内推动更多代码，人工代码审查已成为一个新的瓶颈。

We've also encountered this friction outside engineering. There has been an explosion of new ideas, initiatives, tools, and simulations, as a result of Anthropic employees working with highly capable models—far more than we have the capacity to pursue. The rate at which organizations can spot and fix these bottlenecks may be a skill that improves over time, and it may become the most important skill for any organization.

我们还遇到了工程之外的摩擦。由于 Anthropic 员工与高度强大的模型合作，产生了大量新的想法、计划、工具和模拟，远远超出了我们能够追求的能力。组织发现和解决这些瓶颈的速度可能随着时间的推移而提高，并可能成为任何组织最重要的技能。

3. **AI systems themselves become capable of full recursive self-improvement, and begin building their successors.** If technical trends in advancing capabilities continue, *and* AI systems are able to develop the capabilities inherent to transformative human ingenuity, then it is plausible that AI systems could design and refine themselves.

   AI 系统自身将具备完整的递归自我改进能力，并开始构建它们的继任者。如果提升能力的趋势继续发展，并且 AI 系统能够发展出变革性人类智慧的内在能力，那么 AI 系统设计和完善自身是可能的。

   In this world, the pace of progress in AI development becomes determined entirely by the availability of compute (or the speed of discovering various efficiencies in algorithmic training or inference) for AI systems. Humans play a substantially diminished role in their development, likely moving most of our effort towards oversight, validation, and verification of an expanding "virtual lab" run by AI systems. We expect that systems capable of automated AI research and development would have skills that would transfer to the rest of science, allowing them to begin to revolutionize other fields.

   在这个世界，人工智能发展的步伐完全取决于计算能力（或算法训练或推理中发现各种效率的速度）的可用性。人类在人工智能开发中的作用大幅减弱，我们的大部分精力可能将转向监督、验证和确认由人工智能系统运行的不断扩大的"虚拟实验室"。我们预计，能够进行自动化人工智能研究和开发系统的技能将可以迁移到其他科学领域，从而开始改变其他领域。

   How the alignment problem gets solved—or not—in this future is

something we are least certain about. Models could prove to be sufficiently aligned and capable enough of research taste that they discover and implement novel solutions that we have not yet reached. They could also be sufficiently wise to halt development if not. Alternatively, the rare occurrences of misalignment present in today's models could compound as the models build their successors, growing more frequent but less understood until we lose control of them. It's possible that we can't build, integrate, and verify the tools that we'd need to understand which trendline we are actually on.

在这个未来，对齐问题如何解决——或者不解决——是我们最不确定的事情。模型可能会证明其足够对齐且具备足够的研究品味，以至于它们能够发现并实施我们尚未达到的新颖解决方案。它们也可能足够明智，如果不这样的话就停止开发。或者，当今模型中罕见的错位事件可能会随着模型构建其继任者而加剧，变得更加频繁但更难理解，直到我们失去对它们的控制。有可能我们无法构建、集成和验证我们需要理解我们实际上处于哪条趋势线的工具。

We do not have good intuitions for what this world would look like, because our economy is currently driven by humans and human-built tools. By its nature, a world driven by fast recursive self-improvement could become dominated by the self-improving model as its capabilities fully eclipse those of humans and the model proliferates across the broader economy. It is difficult to predict what the economy looks like if human labor stops being competitive.

我们对这个世界会是什么样子没有很好的直觉，因为我们的经济目前是由人类和人类建造的工具驱动的。本质上，一个由快速递归自我改进驱动的世界可能会被自我改进的模型主导，因为其能力完全超越了人类，并且模型在整个经济中扩散。如果人类劳动不再具有竞争力，很难预测经济会是什么样子。

Even if model development became fully automated and recursive, we can't predict what that would mean for most humans' daily lives.

Amdahl's law applies here as well. Recursive intelligence could lead to achieving many of the benefits outlined in Machines of Loving Grace, quickly in some domains. We expect that embodied intelligence (i.e., robotics) might quickly follow recursive intelligence, and follow a similar path of increasing returns at decreasing cost. More powerful intelligence might help us build things in the physical world more quickly, run more productive clinical trials of lifesaving drugs, and develop novel forms of coordination.

即使模型开发完全自动化和递归，我们也无法预测这对大多数人的日常生活意味着什么。Amdahl 定律同样适用于此。递归智能可能有助于在特定领域快速实现《仁爱机器》中概述的许多益处。我们预计具身智能（即机器人技术）可能会迅速跟随递归智能，并遵循成本递减、收益递增的相似路径。更强大的智能将帮助我们更快地在物理世界中构建事物，运行更高效的临床试验以拯救生命的药物，并开发新型协调形式。

But achieving recursive improvement alone does not suggest an immediate change in how industrial production occurs, societies organize, or markets function. More intelligence can't learn what a drug does over decades of use, can't hold elections sooner than a constitution dictates, and can't turn a stranger into an old friend in a weekend. For most people, the felt pace of this future will still be set by the bottlenecks, even if the laboratory upstream runs at the speed of compute. That collision, where recursive intelligence building itself ever faster meets the world of humans, relationships, and governance, is another part of this future we can't predict.

但仅靠递归式改进并不能表明工业生产方式、社会组织形式或市场运作方式会立即发生变化。更多的智能无法在几十年使用中学会药物的作用，无法比宪法规定更早地举行选举，也无法在周末将陌生人变成老朋友。对于大多数人来说，这个未来的实际速度仍然会受到瓶颈的限制，即使上游的实验室以计算速度运行。这种递归智能自我构建速度越来越快与人类世界、人际关系和治理相碰撞的情况，是未来我们无法预测的另一个部分。

What should we do? 我们应该做什么?

If it were possible to effectively slow the development of this technology to give ourselves more time to deal with its immense implications, we think that would likely be a good thing. But if a slowdown simply lets the least cautious actors catch up technologically, it could leave everyone less safe. Without a global coordination mechanism, companies and governments will have to make difficult decisions about safety while under competitive and geopolitical pressures.

如果能够有效减缓这项技术的发展，以便我们拥有更多时间来应对其巨大的影响，我们认为这很可能是一件好事。但如果放缓仅仅是让最不谨慎的参与者技术赶上来，可能会让每个人都更不安全。没有全球协调机制，企业和政府将在竞争和地缘政治压力下不得不做出关于安全的艰难决策。

We believe it would be good for the world to have the *option* to slow or temporarily pause frontier AI development to enable societal structures and alignment research to keep up with the advance of the technology. The Anthropic Institute will conduct research—in collaboration with many others—and take actions to help build the systems that a credible slowdown or pause would require. These systems would enable frontier AI developers to verify that others globally have actually stopped or slowed, and that a bad actor could not use the auspices of a coordinated slowdown to jump ahead in secret. If such systems existed, we expect that we would slow down or temporarily pause, if other developers at or near the frontier also did so in a verifiable manner.

我们相信，为世界提供减缓或暂时暂停前沿人工智能发展的选项，以使社会结构和协调研究能够跟上技术的进步，这将是有益的。Anthropic Institute 将进行合作研究，并采取行动来帮助构建可信的减缓或暂停所需的系统。这些系统将使前沿人工智能开发者能够验证全球其他地方是否确实停止或减缓，并且一个恶意行为者不能利用协调减缓的掩护秘密地取得领先。如果这样的系统存在，我们预计，如果其他处于或接近前沿的开发者以可验证的方式也这样做，我们将减缓或暂时暂停。

A meaningful slowdown or pause would require multiple well-resourced labs at or near the frontier, in multiple countries, agreeing to stop under the same conditions. It would also require that each can verify that the others have actually stopped. Due to the unique characteristics of AI systems, the detectability (a lower standard than verifiability) element of this arms control problem is much more challenging than with other technologies. Training runs are far easier to conceal than missile silos, their inputs are general-purpose, and the incentive to defect quietly is enormous, because whoever continues while others pause could inherit the lead. A credible pause also has to specify what triggers it, what lifts it, and who adjudicates.

一个有意义的减速或暂停需要多个资源充足的实验室，位于或接近前沿，在多个国家达成一致，在相同条件下停止。这也需要每个实验室都能验证其他实验室确实已经停止。由于人工智能系统的独特特性，这个军备控制问题的可探测性（低于可验证性）要素比其他技术更具挑战性。训练运行比导弹发射井更容易隐藏，它们的输入是通用的，而秘密背叛的动机是巨大的，因为当其他实验室暂停时继续进行的人可能会继承领先地位。一个可信的暂停还必须明确触发它的条件、解除它的条件以及由谁来裁决。

None of this is necessarily impossible in principle—the world has built verification regimes for other complex technologies (e.g., the Intermediate-Range Nuclear Forces Treaty)—but those regimes took decades to build both the infrastructure and the trust. We don't have that long. A unilateral pause by one lab, by contrast, is achievable immediately, but accomplishes much less: it would change who the front-runner is, but it would not create the wider deliberative process that is currently missing.

原则上这并非必然不可能——世界已经为其他复杂技术建立了验证机制（例如《中程导弹条约》），但这些机制建立基础设施和信任都耗费了数十年时间。我们没有那么长的时间。相比之下，单个实验室的单方面暂停虽然可以立即实现，但效果却大打折扣：它可能会改变领跑者是谁，但无法创造当前缺失的更广泛的协商过程。

In the coming months, we will organize conversations where policymakers, researchers, civil society, and other AI companies can help answer some of the questions this piece raises, especially around full recursive self-improvement and how to create better options for coordination and deliberation. We'll publish what comes out of it. The window to investigate the questions together is here, and people outside AI companies should be involved in this deliberation.

在未来几个月里，我们将组织一系列对话，让政策制定者、研究人员、民间社会以及其他人工智能公司能够帮助解答本文提出的一些问题，特别是关于完全递归自我改进以及如何创建更好的协调和审议方案。我们将发布讨论的结果。共同研究这些问题的窗口已经打开，人工智能公司以外的人员应该参与这次审议。

Marina Favaro and Jack Clark co-authored this piece, with editorial support from Santi Ruiz. Shan Carter, Romello Goodman, and Nikki Makagiansar created the visuals from data collected by Brian Calvert and Jun Shern Chan. Daniel Freeman, Jim Baker, Max Young, Sarah Pollack, Francesco Mosconi, Holden Karnofsky, Andy Jones, Kevin Troy, Anton Korinek, Meg Tong, Andrew Ho, Dan Altman, Drake Thomas, Jack Shen, Sasha de Marigny, and Avital Balwit provided feedback.

马丽娜·法瓦罗和杰克·克拉克共同撰写了这篇文章，并得到了桑蒂·鲁伊斯的编辑支持。山·卡特、罗梅洛·古德曼和妮基·马卡吉安萨尔根据布莱恩·卡尔 *vert* 和俊·沈·*Chan* 收集的数据制作了视觉内容。丹尼尔·弗里曼、吉姆·贝克、马克斯·杨、莎拉·波拉克、弗朗切斯科·莫斯科尼、霍尔顿·卡诺夫斯基、安迪·琼斯、凯文·特洛伊、安东·科里内克、梅格·唐、安德鲁·霍、丹·阿尔特曼、德雷克·托马斯、杰克·沈、萨莎·德·马里尼和阿维塔尔·巴尔维特提供了反馈。

FOOTNOTES 脚注

1. METR's key measure tells you the time horizon over which AI systems can be 50% reliable at a basket of tasks, though the trendline looks the same at 80% reliability.

   METR 的关键指标告诉你，在一系列任务中，AI 系统达到 50%可靠性的时间范围，尽管在 80%可靠性时，趋势线看起来是相同的。

2. Especially as they shift toward more open-ended formats and more difficult tasks (e.g., Olympiad-level mathematics), benchmarks often saturate below 100% due to errors in the question and answer sets like ambiguous problem statements and unsolvable questions.

   特别是当它们转向更开放式的格式和更困难的任务（例如，奥林匹克级别的数学），基准测试往往由于问题集中的错误而低于 100%（例如，模棱两可的问题陈述和无法解决的问题）。

3. Anthropic leadership have publicly estimated that 90% or more of our code is written by Claude, including scripts and experimental code. Our >80% figure measures the share of lines merged to production that can be attributed to Claude. This is a more conservative measurement in two ways: our attribution pipeline has gaps, and the lines not attributed to Claude include auto-generated code and other artifacts that were not hand-written by humans either.

   Anthropic 的领导层公开估计，我们 90%或更多的代码是由 Claude 编写的，包括脚本和实验代码。我们的>80%的数字衡量了可以归因于 Claude 的生产中合并的代码行比例。这是一种更保守的衡量方式，原因有两个：我们的归因管道存在漏洞，未归因于 Claude 的代码行包括自动生成的代码和其他非人类手写的工件。

4. This surge in code production is straining the infrastructure everyone shares. GitHub—the platform most of the world's software is built on—saw roughly one billion code commits in all of 2025; by mid-2026 it saw 275 million a week, on pace for roughly 14 billion over the year. The company's COO has said that it is "pushing incredibly hard" on capacity just to keep up.

   这种代码产出的激增正在消耗每个人共享的基础设施。GitHub——世界上大部分软件构建的平台——在 2025 年总共看到了大约 10 亿次代码提交；到 2026 年年中，每周就达到了 2.75 亿次，预计全年将接近 140 亿次。该公司的首席运营官表示，为了保持同步，他们在"极度努力"地提升容量。

5. Additional details on the methodology of this survey are discussed in section 2.3.5 of the Claude Opus 4.7 System Card.

   本调查的方法论细节将在 Claude Opus 4.7 系统卡的 2.3.5 节中讨论。

6. Many respondents may not have thought carefully about how to account for various biases or subtleties in the question definition, and recent research by METR shows that developer estimates of AI productivity uplift can be overestimated.

许多受访者可能没有仔细思考如何解释问题定义中的各种偏差或细微之处，而 METR 的最新研究表明，开发者对 AI 生产力提升的估计可能被高估。

7. How large the speedup gets depends heavily on how much room for improvement the starting code leaves, and it should not be read as a real-world training speedup. So the absolute multiple is not the figure to anchor on here. What is more informative is the like-for-like comparison that this experimental setup makes possible, both across models (~3x to ~52x over the past year) and against a skilled human (~4x in four to eight hours on the same task).

   速度提升的幅度很大程度上取决于起始代码留下的改进空间，不应将其视为现实世界的训练速度提升。因此，绝对倍数不是这里的基准。更有信息量的是，这种实验设置所进行的同类比较，包括跨模型（过去一年中约 3 倍至~52 倍）以及与熟练人类（在同一任务上四至八小时内约 4 倍）的比较。

8. As a check on judge bias, we ran the same test on a separate set of 127 moments where the human's next move was already strong (as opposed to the original set, where the human's direction had room for improvement). There, the models' suggestions were judged better only about 20% of the time.

   为了检验裁判偏倚，我们在另一组 127 个时刻运行了相同的测试，在这些时刻人类的下一步已经很强（与原始集形成对比，在原始集中人类的方向还有改进空间）。在那里，模型的建议仅在大约 20%的时间内被评价为更好。

* Quotes from Anthropic employees throughout this article are drawn from internal discussions and used with permission. They reflect individual views as of May 2026, not official company positions.

* 本文中引用的 Anthropic 员工的言论均来自内部讨论，并经许可使用。它们反映的是截至 2026 年 5 月的个人观点，而非官方公司立场。