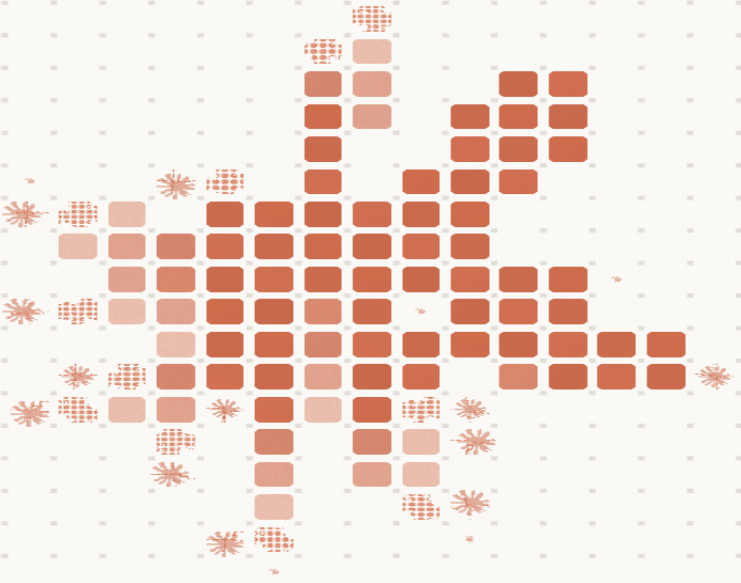
When AI builds itself 当 AI 构建自身时

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

我们在递归自我改进方面的进展及其影响。



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 AI 历史的大部分时间里，人类驱动着其开发周期中的每一个步骤。但在 Anthropic，我们正将越来越多的 AI 开发工作委托给 AI 系统本身，这正在加快我们的工作进度。

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.

如果这一趋势发展得足够远，并拥有充足的算力，它将指向一个能够完全自主设计和开发其后继者的 AI 系统。这被称为递归自我改进（recursive self-improvement）。我们目前尚未达到那个阶段，递归自我改进也并非必然发生。但它的到来可能比大多数机构准备好应对的时间还要早。

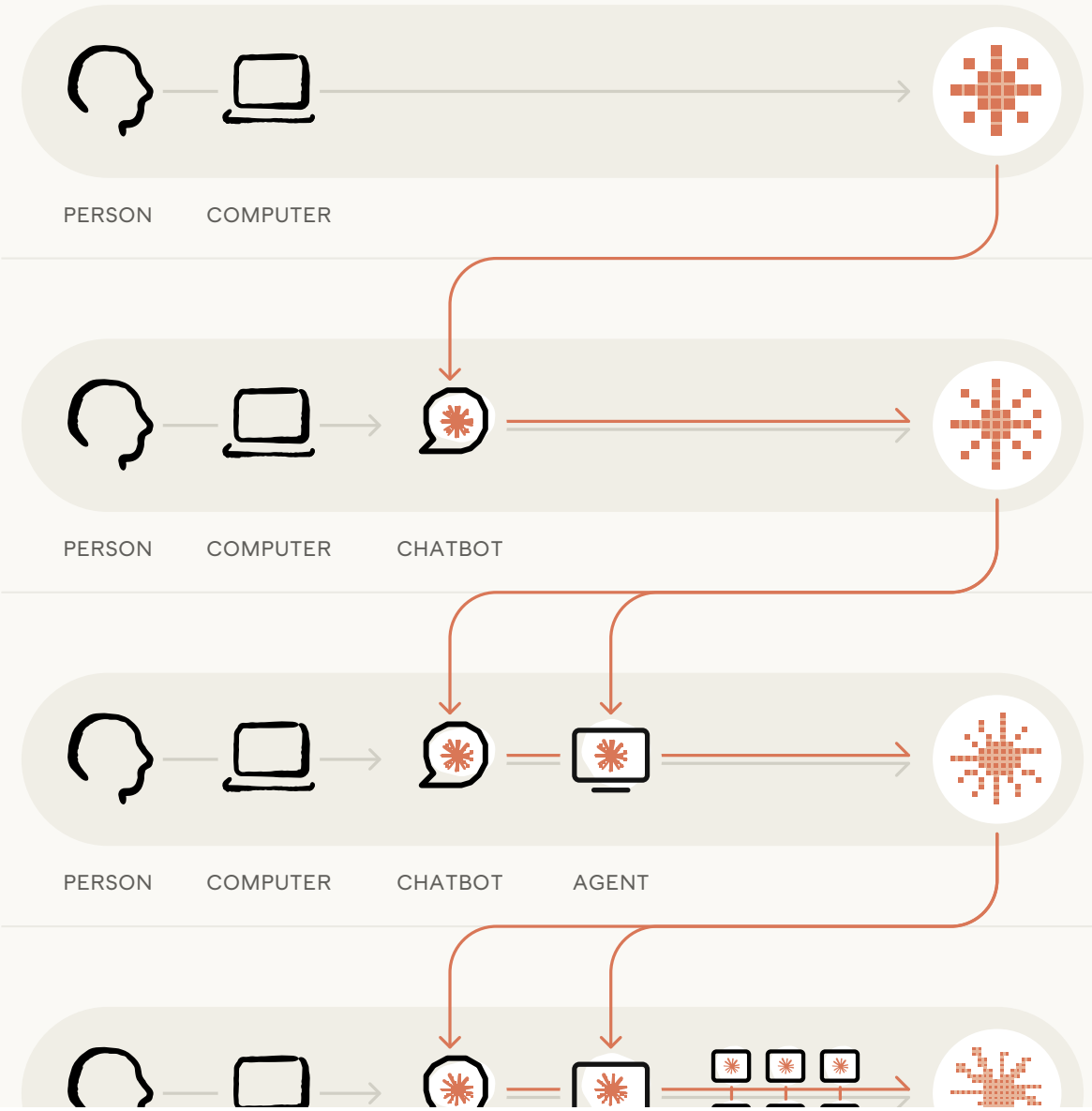
Using public benchmarks and previously unreported data from within Anthropic, The Anthropic Institute is showing that AI is already accelerating the development of AI systems. To take just one example: today, Anthropic engineers on average ship 8x as much code per quarter as they did from 2021-2025.

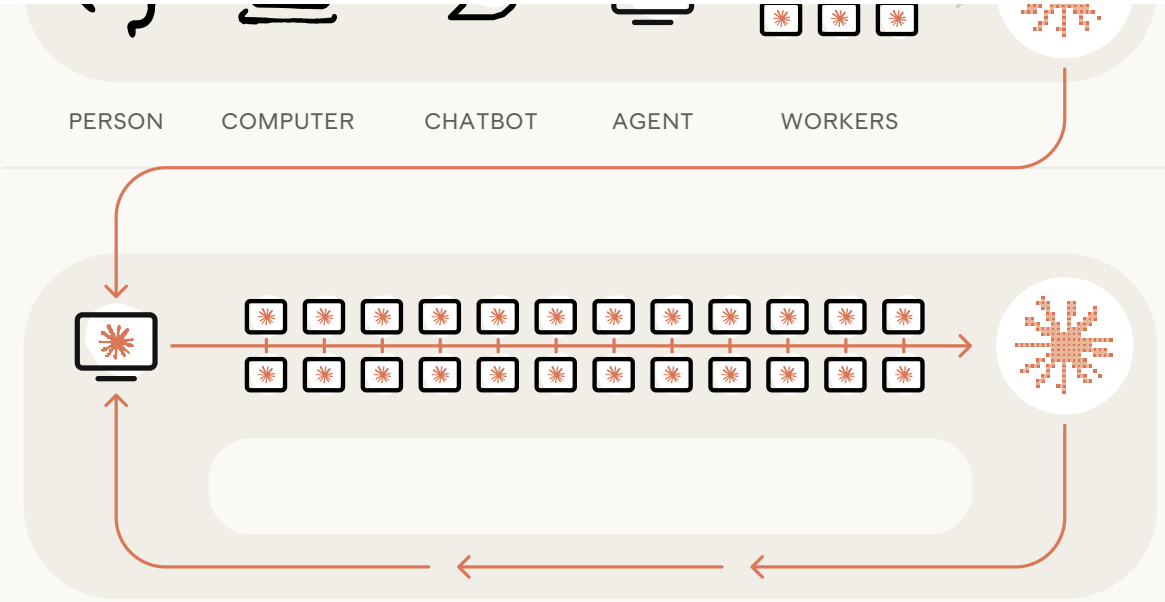
通过公开基准测试以及来自 Anthropic 内部此前未公开的数据，Anthropic 研究院（The Anthropic Institute）展示了 AI 已经在加速 AI 系统的开发。仅举一例：如今，Anthropic 的工程师平均每季度交付的代码量是 2021-2025 年期间的 8 倍。

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，AI 系统在未来几年将变得更加强大。这些趋势具有巨大的影响。能够自我构建的 AI 将是技术史上的重大进展——它可能在科学、医疗及其他领域为世界带来巨大的福祉。但完全的递归自我改进也可能增加人类失去对 AI 系统控制的风险。如果系统能够完全构建自己的后继者，那么我们保护、监控以及塑造其行为的方式都将变得更加重要。





2021–2023

Building the first Claude

构建首个 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 的工作方式与任何其他科技公司无异：人们在笔记本电脑上编写代码和文档。

2023–2025

Chatbots

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

人们开始使用早期的聊天机器人来辅助部分流程，例如生成简短的代码片段，并将输出结果复制到文本编辑器中。

2025–2026

Coding agents

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

随着智能体能力的提升，它们已经能够自主编写和修改代码，有时甚至能处理整个文件。

TODAY

Autonomous agents 自主智能体

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

智能体现在可以自行运行代码，并将长达数小时的工作任务委派给其他智能体。

20XX?

Closing the loop 闭环控制

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.

在未来，智能体（agents）的能力可能强大到足以自行构建和训练模型。如果这一天到来，Claude 的未来版本将能够由 Claude 自身不断进行改进。

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 模型的进步速度正在加快。它们能够可靠地独立完成任务的时长大约每四个月翻一倍，而此前的趋势是每七个月翻一倍。2024 年 3 月，Claude 3 Opus 可以完成人类大约需要 4 分钟完成的软件任务。一年后，Claude 3.7 Sonnet 能够处理耗时约一个半小时的任务。再过一年，Claude 4.6 Opus 已经能胜任 12 小时的任务。[1] 如果这一趋势持续下去，需要专业人员耗时数天才能完成的任务可能会在今年进入其能力范围。到 2027 年，AI 系统可能具备处理需要人类数周时间才能完成的任务的能力。

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.

同样的模式也出现在编程和研究基准测试中。基准测试用于衡量模型在特定领域的表现，当模型达到接近 100% 的表现时，这些基准就会趋于"饱和"。[2] SWE-bench 是衡量真实世界软件工程能力的标准化测试：它向模型提供实际的开源代码库和真实的错误报告，并要求模型编写代码更改以修复问题，且必须通过项目自身的测试。在短短两年内，模型在该基准测试中的得分已从个位数跃升至接近饱和。

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.

CORE-Bench 测试模型是否能够复现现有研究，这是它们开展原创性研究的前提。它向 AI 模型提供已发表论文背后的代码和数据，并要求其重新运行所有内容，以确认能够复制论文的结果。

AI systems went from succeeding at reproducing the results roughly 20% of the time in 2024 to saturating the benchmark fifteen months later.

AI 系统在复现结果方面的成功率从 2024 年的约 20% 提升到了十五个月后的基准测试饱和水平。

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."

运行衡量模型完成长时间任务能力的基准测试机构 METR 发现，Claude Mythos Preview 可以工作"至少"16 小时，并且处于"在不增加新任务的情况下 [METR] 所能衡量的上限"。

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.

公开基准测试充分说明了这些系统的能力。但它们无法揭示 AI 系统在加速 AI 自身开发方面所产生的影响。为此，我们需要来自像 Anthropic 这样的 AI 公司内部的直接证据。

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.

构建前沿模型需要两类主要工作。一是工程：编写代码、搭建基础设施以及监督模型训练。二是研究：决定运行哪些实验、解读返回的结果，并确定下一步尝试哪些想法。

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.

在工程和研究领域，情况是一致的。在工程方面，可以将一个定义不明确的问题交给 Claude，由它找出解决办法；人类提供目标，但不再需要提供方法。在研究方面，Claude 在执行定义明确的实验时，已经能够达到或超过熟练人类的水平。

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.

然而，当涉及到在工程和研究中行使判断力以选择目标时，Claude 仍存在巨大的性能差距。这就是当今 AI 与未来能够自主设计其继任者的系统之间的差距。

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，随着员工经验的增长，他们通常会承担更多开放性且重要的任务。在职业生涯早期，他们执行他人指定的任务，例如："导出按钮失效了，请修复它。"随着经验的积累，他们会被赋予一个目标并自行设计方案，例如："调查为什么网络在高负载下会变慢。"在最资深的层级，他们需要决定哪些问题值得去解决："团队下个季度应该构建什么？"我们可以利用 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.

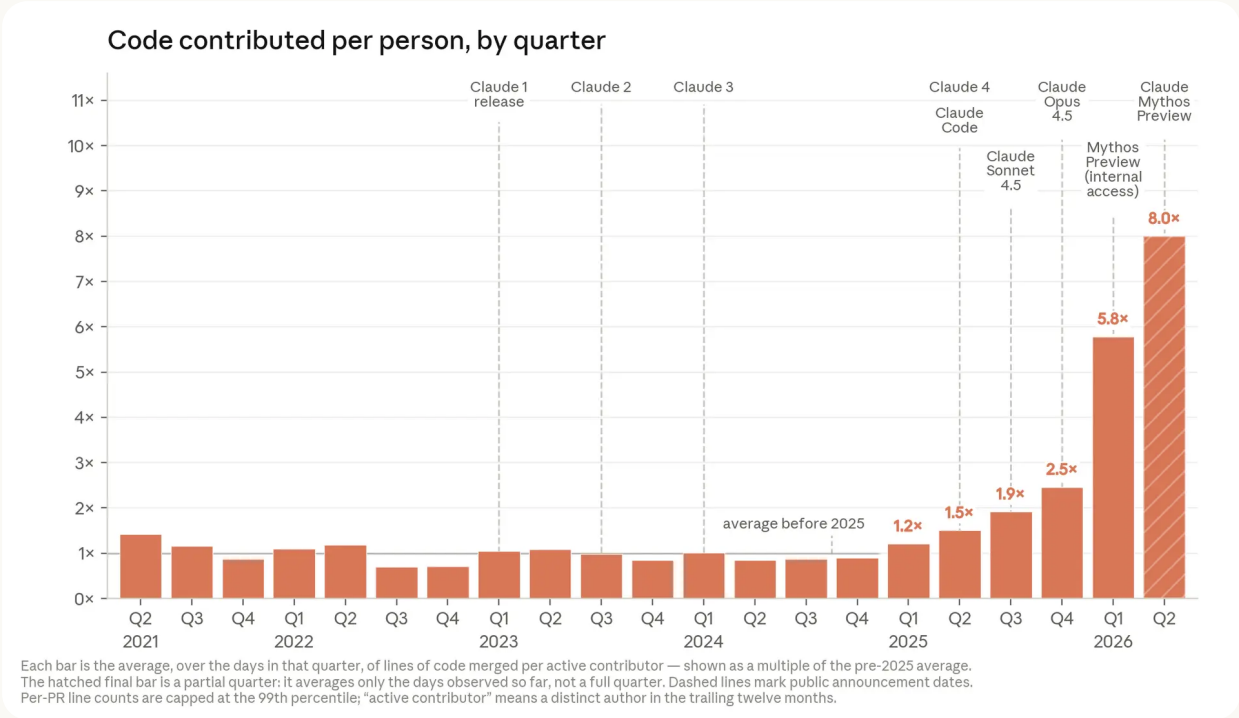
Claude 编写了 Anthropic 很大比例的代码。截至 2026 年 5 月，我们合并到 Anthropic 代码库中的代码有 80% 以上是由 Claude 编写的。[3] 在 2025 年 2 月 Claude Code 发布研究预览版之前，这个数字仅为低个位数。这种转变也体现在每位工程师的产出量上。

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.

在 Anthropic 成立的前四年（2021-2024 年），每位工程师每天合并的代码行数保持稳定，随后在 2025 年开始攀升，当时 Claude 开始能够运行代码，而不仅仅是提供建议供工程师复制粘贴。2026 年，随着模型开始能够在更长的时间跨度内自主工作，增长曲线变得更加陡峭。

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.

下图展示了这两个拐点。在 2026 年第二季度，典型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，而不是亲自输入。



Each bar is the average, over the days in that quarter, of lines of code merged per active contributor — shown as a multiple of the pre-2025 average. The hatched final bar is a partial quarter: it averages only the days observed so far, not a full quarter. Dashed lines mark public announcement dates. Per-PR line counts are capped at the 99th percentile; "active contributor" means a distinct author in the trailing twelve months.

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.

需要说明的是：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 年第二季度每位工程师每天产出的代码行数增长至 8 倍，几乎肯定夸大了真实的生产力提升。尽管如此，这仍预示着一种加速趋势。在 Anthropic，我们并不根据代码行数来奖励员工；相反，团队成员产出更多代码，仅仅是因为他们正在利用 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.

代码行数的增加与生产力大幅提升的主观感受是一致的。

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.

在 2026 年 3 月对 Anthropic 各研究团队 130 名员工进行的民意调查中，中位数受访者估计，在他们原本就要开展的项目中，使用 Mythos Preview 产出的成果大约是完全不使用 AI 模型时的 4 倍。[5] 我们预计 3 月份真实的提升程度可能略低一些。[6] 尽管如此，我们认为这一整体说法是合理的，并且与我们的其他观察结果一致：很大一部分 Anthropic 技术人员完成核心工作的速度比没有 AI 辅助时快了数倍。

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.

我们还看到证据表明，Anthropic 的员工正在使用 Claude 完成一些原本根本不会开展的工作，例如构建探索性工具和处理长期搁置的清理工作。例如，在 2026 年 4 月，Claude 交付了 800 多个修复补丁，将某类 API 错误减少了一千倍。

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.

负责监督 Claude 的工程师估计，人类需要四年时间才能完成这项工作；解决他人的 bug 既缓慢又艰辛，而且人类很难在脑海中同时处理如此多陌生的上下文信息。

> “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.”
>
> **大约一年前，我开始全身心地投入到“Claude 化”中。那是一段疯狂的冒险，到现在我已经有 ~5 个月没有亲手写过任何代码了。”**

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.

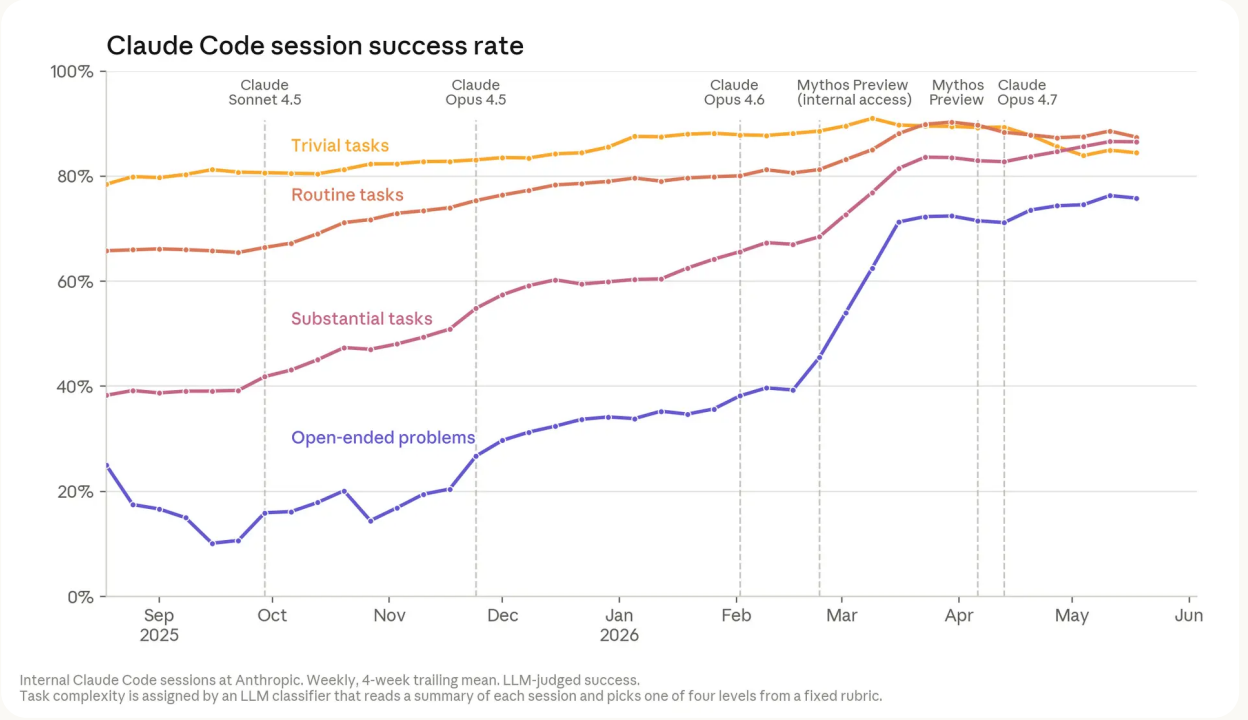
Claude 编写的代码质量“良好”且在不断进步。“良好的代码”意味着两点：它能运行，并且其编写方式允许其他工程师理解并在此基础上进行开发。就第一条标准而言，证据已经非常明确。

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.

Anthropic 员工对 Claude 进行纠正、重定向或在中途接管任务的频率在一年内稳步下降，即使是在最复杂且开放式的任务中也是如此。这意味着在没有明确规范、工程师也不确定最终答案是什么样的情况下，Claude 依然表现出色。

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.

这一点从下方图表中 Claude 在不同难度任务上的成功率随时间的变化中可见一斑。Claude 编写的代码确实行之有效。



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.*

图表解读：会话成功率由 *Claude* 裁判模型判定；如果 *Claude Code* 智能体在不需要人工纠正的情况下明确完成了用户任务，则该会话被视为成功。工作负载的变化可能会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.

在最开放的任务中，Claude 的成功率在 2026 年 5 月达到了 76%，在六个月内提升了 50 个百分点。举一个该难度级别的任务示例：一次常规升级导致数万个训练作业崩溃。一名工程师仅提供了部分文本内容和集群访问权限，就让 Claude 直接处理这一线上事故。

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.

通过逐一排查运行中的任务并测试环境设置，Claude 成功定位了触发崩溃的那个隐蔽的调试标志，实现了稳定复现，并确认了修复方案。

In about two hours, Claude delivered what would normally be two to three days of work.

在大约两个小时内，Claude 完成了通常需要两到三天的任务量。

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.

第二个标准是编写其他工程师能够理解并在此基础上进行开发的方案。在这一方面，人类与 AI 之间仍存在差距，但这种差距正在迅速缩小。

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.

Anthropic 内部尚未达成完全共识，但许多人认为，在 2025 年底，Claude 编写的代码质量仍逊于 Anthropic 工程师编写的代码，而目前两者已基本持平。我们预计在一年内，它的表现将超越人类。

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.

这改变了 Anthropic 审查自身代码的方式。现在，对我们代码库提出的更改都会由 Claude 自动审查器进行阅读，在合并之前查找漏洞、安全缺陷和其他缺陷。

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 onclaude.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.

利用这一工具，我们进行了一次回顾性分析，发现如果对代码库的每一次更改都进行 Claude 自动审查，就能在 claude.ai 过去发生的事故上线生产环境之前，拦截其中约三分之一的漏洞。编写这些代码的工程师是世界上构建此类系统最顶尖的人才，而 Claude 现在正捕捉到他们遗漏的错误。

“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 内部还略逊于人类编写的代码，到今天已基本持平，我们预计在一年内它将表现得更好。”

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.

Claude 擅长通过运行实验来实现他人设定的目标。每当 Anthropic 发布新模型时，我们都会进行同样的测试：给 Claude 一段训练小型 AI 模型代码，并要求它在通过相同正确性检查的前提下，让代码运行得尽可能快。

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.

目标和成功指标是预先设定的，因此 Claude 的任务就是通过重写代码、运行代码、计时并重复这一过程来寻找提速方法。这是一个实验研究循环的缩微版本。

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.

2025 年 5 月，Claude Opus 4 相比初始代码平均实现了 ~3 倍的提速。到 2026 年 4 月，Claude Mythos Preview 已经达到了 ~52 倍。作为参考，一名经验丰富的人类研究员通常需要 4 到 8 小时才能达到 4 倍的提速。[7] 在这一部分研究工作流中——即在定义明确的实验内优化步骤——Claude 在不到一年的时间里，已经从“超级助手”进化到了“超越人类”的水平。

“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.’”

现状大致是：“人类产生想法，而模型能够以比以前快[一个数量级]的速度去实现、测试和评估这些想法。”

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.

Claude 在自主提出实验方案方面也变得越来越出色。2026 年 4 月，Anthropic 发布了 Claude 首次端到端运行开放式研究项目的演示。由 Claude 驱动的智能体被赋予了一个 AI 安全领域的开放性问题——大致是：弱模型能否可靠地监督强模型？——并由其自行解决。这涉及提出假设、进行测试、与并行智能体分享发现以及迭代。该任务具有明确的性能"下限"和"上限"：下限是弱监督者独立完成任务的表现；上限是强模型在基于正确答案进行训练时的表现。

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.

两名人类研究员在大约一周的时间里，弥补了大约 23% 的差距；而智能体在累计 800 小时内弥补了 97% 的差距，并消耗了约 18,000 美元的算力。这项工作存在一些局限性：结果并未能完全转化到生产规模的模型中，且问题的选择和评分标准的制定仍由人类完成。

But within those bounds, the agents designed every experiment

themselves. Direction-setting was the only meaningful role a human played.

但在这些限制范围内，实验的每一个环节都由智能体自行设计。人类唯一发挥实质性作用的环节是设定研究方向。

"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 在 1 到 2 天的时间里，仅凭我极少的帮助就完成了这一切。我认为，如果一名[初级同事]在同样的时间内带着这样的结果回来找我，我会感到有些惊艳。未来已至。"

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.

Claude 正变得越来越擅长引导研究进程并得出研究结论。我们分析了真实的 Claude Code 会话记录（2026 年 1 月至 3 月期间），在这些记录中，Anthropic 的研究员正与 Claude 合作解决开放式的调查性问题，例如查明训练运行为何不断崩溃，或者模型为何在某项基准测试中得分较低。

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.

在每种情况下，我们都发现研究人员在某个时刻绕了弯路：他们追求的一个方向导致会话偏离了预定轨道，直到最后才重新回到正轨。

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]

随后，我们向不同的 Claude 模型仅展示会话偏离轨道之前的工作内容，并询问其下一步会怎么做。另一个能够看到会话最终结果的独立 Claude 模型则负责评判：是 AI 还是人类建议的下一步方案更好。[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.

由于我们刻意挑选了已知人类选择仍有改进空间的时机（n=129），因此这并不是模型与人类判断力之间完全对等的比较。

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.

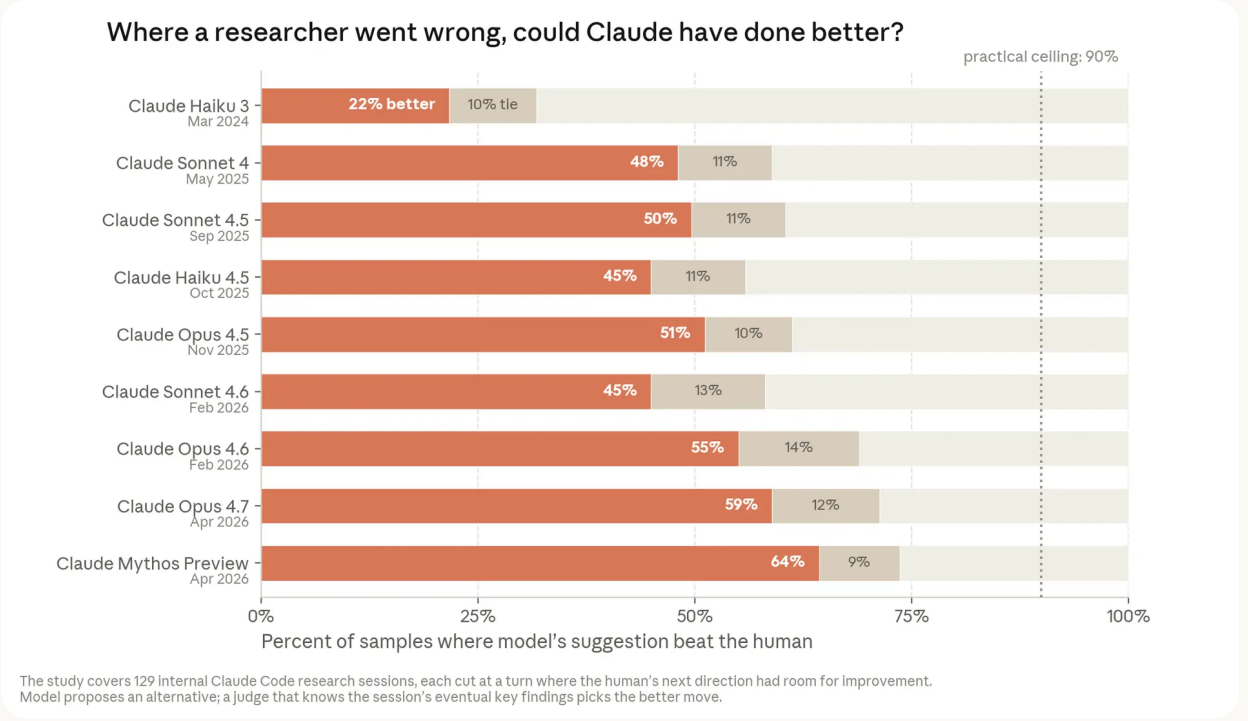
这些时刻为我们提供了一系列真实且具有挑战性的场景，在这些场景中，正确的下一步行动并不显而易见，而人类的选择则可以作为一个有用的基准，用来衡量模型性能随时间推移的变化。

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.

根据这一衡量标准，我们在 2025 年 11 月推出的最强模型（Opus 4.5）在 51%的情况下优于人类的选择；到 2026 年 4 月（Mythos Preview），这一比例增长到了 64%。科研的日常工作很大程度上是由这些后续决策组成的链条，因此，这是衡量模型最终能否独立开展研究的一项相关指标。

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.

我们将这一结果视为一个早期信号，表明 AI 系统在做出 AI 研究所依赖的那类判断决策方面正变得越来越出色。



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).*

解读说明：实际上限线（*practical ceiling line*）衡量的是由一个能够看到整个会话（包括结局）的模型所给出的“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.”

截至目前，人类的比较优势仍然在于洞察大局，以及跳出当前任务局限的思考能力。

What might the future of work at Anthropic look like?

Anthropic 未来的工作模式会是什么样?

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.

迹象表明，在 AI 开发的每一个环节中，人类的角色都在不断收窄。一旦人类编写的代码与 AI 编写的代码质量达到同等水平，人类将完全停止编写代码，转而仅负责代码审查。

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?

但如果人类审查代码的速度赶不上 Claude 生成代码的速度，人工审查就会成为 AI 开发的瓶颈。同样，一旦 Claude 能够自主运行实验，问题就会转变为："这些实验中，哪些是值得运行的？"

" 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.

简而言之："执行"层面（即编写代码、运行实验、产出结果）现在几乎不再消耗人类的时间成本，即便它仍然需要消耗计算资源。

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，人类具有比较优势的领域在于研究品味和判断力，包括选择哪些问题重要、哪些结果值得信赖，以及何时该认定某种方法已走入死胡同。

"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（以及生活）曾运作于一种人类之间互助互利的小额"人情经济"之上。"你能帮我运行这个脚本吗？"……每一次求助都产生了一份小小的欠人情，一份小小的相互感知。[Claude] 速度更快，且不产生任何欠人情，但每一次这样的交互都是一次对人类协作机会的丧失。

“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.”

在一切进展顺利的日子里，我不禁会想，我所做的一切都无关紧要，所有事情都是自动化的，而且比我做得更好、更快。但也有一些日子，一切都崩溃了，我不明白为什么，这时我才意识到，我根本不知道自己一直在忙些什么。

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.

对于上述证据，一个自然的反对意见是：目前仍掌握在人类手中的工作——即选择研究哪些问题——才是最重要的。缺乏这种判断力，Claude 只是一个能干的助手，而非一个能够独立推动 AI 进步的系统。

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.

目前尚不清楚当今的训练方法和架构是否能解锁这种能力。但 AI 的进步很少是由"灵光一现"的时刻推动的。在 AI 的近代史中确实出现过几次这样的时刻，比如 Transformer 架构或混合专家模型（MoE），但改变范式的想法往往相隔数年才会出现。在此期间，大多数进步都是渐进式的：我们扩大规模，观察哪里出了问题，修复它，然后再次尝试。这恰恰是 Claude 目前所擅长的工作流程。爱迪生曾说，天才是 1% 的灵感加上 99% 的汗水。

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.

但我们看到，"汗水"正在变得日益自动化。显而易见，推动前沿技术进步的大部分工作都是可以自动化的；大规模的研究进展主要取决于工具和资源，它们决定了你运行实验的速度、同时运行实验的数量，以及获取结果的快慢。

Even if we suppose that Claude never achieves good research taste, a conservative reading of our evidence still implies compounding acceleration.

即便我们假设 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意味着复合式的加速。

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.

如果人类将大部分时间花在仅占个位数的方向设定工作上，而由 Claude 处理其余部分，这意味着每位工程师或研究员所驾驭的工作量将远超以往。我们看到的证据表明，Anthropic 的员工不仅行动速度更快，而且覆盖的领域也更广。

In practice, this means that AI already makes Anthropic move much faster than it did before the advent of effective AI tools.

在实践中，这意味着 AI 已经让 Anthropic 的运转速度比高效 AI 工具出现之前快得多。

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.

一种不那么保守的解读是，关于 Claude 研究判断力提升的早期证据（尽管目前范围还很窄）表明这种能力也在不断进步。“研究品味”可能只是又一种 AI 能力——AI 系统会先经历一段时间的失败，然后逐渐变得擅长。

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.

我们在其他定性技能上也看到了类似的模式，例如 AI 系统能够解释笑话为什么好笑、展现出心理理论能力，以及破解语言谜题。

Possible futures 可能的未来

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:

接下来的发展取决于两件事：这一趋势是否会持续，以及如果持续下去，我们选择如何应对。我们至少可以想象出三种未来的情景：

---

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.

趋势停滞，但当今的 AI 能力得到广泛普及。本文提到了许多指数级的增长轨迹。但这些轨迹实际上可能被证明是 S 型曲线。我们可能正接近曲线的拐点，届时规模效应带来的收益会递减，增长曲线会变直，然后趋于平缓。

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.

区分平庸研究员与伟大研究员的判断力，可能是一种无法通过扩大计算量和数据等训练投入来获得的。如果是这样，突破这一瓶颈将需要新的思路，例如一种能够取代目前所有前沿模型所使用的 Transformer 架构的新型架构方法。

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.

或者，AI 进步的束缚性限制可能在于供应链，而非模型本身：推进和普及前沿技术可能需要比目前更多的能源和算力。芯片制造的速度、电网的扩张或互连带宽可能才是瓶颈，而非智能本身。

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.

我们也不能排除 AI 生态系统遭受外部冲击而导致进程大幅放缓的可能性，例如算力或电力供应的突然减少，这两者都会减缓进度并增加实验室的前瞻性投资成本。又或者，我们可能尚未预见到某些其他的进步障碍。

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.

即使模型能力冻结在今天的水平，我们仍能预见世界将发生重大变化。Project Glasswing 就是一个早期迹象：在运行的最初几周内，Mythos Preview 在全球最重要的系统中发现了超过一万个高危和严重级别的软件漏洞——这足以让网络防御的瓶颈从发现漏洞转向如何以足够快的速度修复漏洞。

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.

而且，我们仍处于当今模型向更广泛经济领域扩散的早期阶段。在这些领域，一家 100 人的公司将日益能够完成 1,000 人的工作量，因为每位员工都将统领一个由智能体（agents）组成的金字塔。

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.

为了完整性，我们将这一情景纳入考虑，但我们认为其发生的可能性并不大。到目前为止，我们所能衡量的每一项能力——包括那些感觉上更"模糊"的能力，如代码质量和处理开放式任务的成功率——都遵循着同样的曲线。我们尚未看到这条曲线发生转折。

Of the three futures we consider, this one would give governments and societies the most time to adapt.

在我们考虑的三种未来中，这一种将为政府和社会提供最充裕的适应时间。

We are more worried about the next two, which would move faster and leave far less room for preparation.

我们更担心接下来的两种情况，它们的发展速度会更快，留给准备的空间也小得多。

---

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.

AI 实验室将继续看到复合式的效率提升。在这种情景下，AI 开发实现了实质性的自动化，但人类继续设定研究方向并评判结果。随着时间的推移，使用 AI 系统的组织将变得更加高效，因此我们可以预见，该组织中的每位成员都将获得显著的生产力乘数效应。

100-person companies could do the work of 10,000- or 100,000-person organizations.

100 人的公司就能完成以往 10,000 人甚至 100,000 人规模的组织才能完成的工作。

This would revolutionize knowledge work and government services, but could also be turned to harmful ends, from authoritarian surveillance of whole populations to influence operations that tailor manipulation to each individual and run at a scale no human team could match. The role of humans at companies like Anthropic would shift.

这将彻底改变知识型工作和政府服务，但也可能被用于有害目的——从对全体国民的威权主义监控，到针对每个个体量身定制、且运行规模远超任何人类团队的操纵性影响力行动。像 Anthropic 这样的公司，其员工的角色也将发生转变。

People would partner with AI systems to scale up research and generate new insights, and together they would build the systems needed to verify that AI outputs can be trusted.

人类将与 AI 系统合作，以扩大研究规模并产生新的见解，双方将共同构建必要的系统，以验证 AI 的输出是否值得信赖。

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.

我们在此阐述的证据表明，我们很可能正走向这一情景。然而，加速流程中的某一部分往往只是将瓶颈转移到了别处：整体进度受限于那些尚未提速的部分。在计算领域，这被称为阿姆达尔定律（Amdahl's law），同样的逻辑也适用于组织。Anthropic 已经遇到了阿姆达尔定律的一个典型特征：随着我们在组织内部推送的代码量不断增加，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.

我们在工程领域之外也遇到了这种摩擦。由于 Anthropic 的员工在使用高性能模型开展工作，各种新想法、新倡议、新工具和新模拟层出不穷——其数量已经远远超出了我们的执行能力。

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.

组织发现并解决这些瓶颈的速度，可能是一种随时间推移而不断提升的技能，并且它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 系统本身将具备完全递归自我改进的能力，并开始构建其后继者。如果能力提升的技术趋势持续下去，且 AI 系统能够发展出人类变革性创造力所固有的能力，那么 AI 系统设计并完善自身是完全可能的。

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.

在这种情况下，AI 发展的进步速度将完全取决于 AI 系统可获得的算力（或发现各种算法训练及推理效率提升的速度）。

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.

人类在开发过程中所扮演的角色将大幅削弱，我们的精力可能会主要转向对由 AI 系统运行的、不断扩张的"虚拟实验室"进行监督、验证和核查。

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.

我们预期，具备自动化 AI 研发能力的系统所拥有的技能将迁移到其他科学领域，从而使它们能够开始在其他领域引发变革。

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.

在这样一个未来，对齐问题将如何解决（或者是否能被解决），是我们最不确定的事情。模型可能会被证明具有足够的对齐性，并具备充分的研究品味，从而发现并实施我们尚未达成的创新解决方案。如果情况并非如此，它们也可能拥有足够的智慧来停止开发。

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.

或者，当今模型中存在的罕见失配现象可能会在模型构建其后代的过程中不断累积，变得越来越频繁，却越来越难以被理解，直到我们最终失去对它们的控制。

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.

我们可能无法构建、整合并验证所需的工具，来明确我们究竟处于哪一条趋势线上。

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.

我们对这样一个世界缺乏直觉认知，因为目前的经济是由人类和人类制造的工具所驱动的。

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.

从本质上讲，一个由快速递归自我改进驱动的世界，可能会演变为由自我改进模型主导，因为其能力将完全超越人类，且该模型会在更广泛的经济领域中扩散。如果人类劳动力不再具有竞争力，很难预测届时的经济会呈现何种面貌。

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.

即使模型开发实现了完全自动化和递归化，我们也无法预测这对大多数人的日常生活意味着什么。阿姆达尔定律（Amdahl's law）在此同样适用。递归智能可能会在某些领域迅速实现《Machines of Loving Grace》中所描述的诸多益处。我们预计具身智能（即机器人技术）可能会紧随递归智能之后，并走上一条类似的、以递减成本获得递增回报的发展路径。

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.

更强大的智能或许能帮助我们更快速地在物理世界中建造设施，开展更高效的救命药物临床试验，并开发出全新的协作模式。

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.

但仅靠实现递归式改进，并不意味着工业生产方式、社会组织形式或市场运作机制会立即发生变革。再强大的智能也无法在短时间内习得药物在数十年使用中所产生的效果，无法在宪法规定之外提前举行选举，也无法在一个周末内就让陌生人变成老友。

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.

对大多数人而言，即便上游实验室以计算速度飞速运转，未来的感知节奏仍将由现实中的瓶颈所决定。当递归式自我构建、速度不断加快的智能，与人类、人际关系及治理体系的世界发生碰撞，这种交汇将构成未来中另一个我们无法预测的部分。

What should we do? 我们该怎么做？

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.

如果能够有效地放慢这项技术的发展速度，让我们有更多时间来应对其巨大的影响，我们认为这可能是一件好事。但如果放慢速度仅仅是让那些最不谨慎的参与者在技术上赶上来，那可能会让每个人都变得更不安全。在缺乏全球协调机制的情况下，企业和政府将不得不在竞争和地缘政治压力下，就安全性做出艰难的决定。

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.

我们认为，如果世界能够选择放慢或暂时暂停前沿人工智能的发展，以使社会结构和对齐研究能够跟上技术的进步，那将是有益的。Anthropic 研究院将与许多其他机构合作开展研究，并采取行动，帮助建立可靠的减速或暂停所需的系统。这些系统将使前沿人工智能开发商能够核实全球其他开发商是否确实停止或放慢了速度，并确保不良行为者无法利用协调减速的契机在秘密中实现反超。

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.

如果存在这样的系统，并且处于或接近前沿的其他开发商也以可验证的方式这样做，我们预计我们也会放慢速度或暂时暂停。

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.

有意义的减速或暂停需要多个国家、处于或接近前沿的多个资源充足的实验室同意在相同条件下停止。这还需要每个实验室都能核实其他实验室是否确实已经停止。

Due to the unique characteristics of AI systems, the detectability (a lower standard than verifiability) element of this arms control problem ismuch 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.

由于 AI 系统的独特特性，军备控制问题中的“可探测性”（其标准低于“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，其投入是通用型的，而且悄悄违约的诱因巨大，因为当他人停下时，继续前进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条件、解除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.

从原则上讲，这些并非完全不可能——世界已经为其他复杂技术建立了核查机制（例如《中导条约》）——但这些机制耗费了数十年时间才建立起基础设施和信任。我们没有那么长的时间。

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.

相比之下，单个实验室的单方面暂停是可以立即实现的，但效果要差得多：它只会改变谁是领跑者，而无法创造目前所缺失的更广泛的审议流程。

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.

在接下来的几个月里，我们将组织对话，邀请政策制定者、研究人员、公民社会和其他 AI 公司共同探讨本文提出的一些问题，特别是围绕完全递归自我改进，以及如何为协调和审议创造更好的选择。我们将发布相关的成果。

The window to investigate the questions together is here, and people outside AI companies should be involved in this deliberation.

共同探讨这些问题的窗口期已经到来，AI 公司之外的人士也应当参与到这一审议过程中。

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.

本文由 *Marina Favaro* 和 *Jack Clark* 共同撰写，*Santi Ruiz* 提供编辑支持。*Shan Carter*、*Romello Goodman* 和 *Nikki Makagiansar* 根据 *Brian Calvert* 和 *Jun Shern Chan* 收集的数据制作了可视化图表。

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.

***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** 和 **Avital Balwit** 提供了反馈。*

FOOTNOTES

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 的关键指标显示了 AI 系统在执行一系列任务时达到 50% 可靠性的时间跨度，尽管在 80% 可靠性水平下，趋势线看起来也是一样的。

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.

   随着基准测试转向更具开放性的形式和更具挑战性的任务（例如奥林匹克级别的数学），由于问题和答案集中存在表述不清或题目无解等错误，其得分往往在达到 100% 之前就已趋于饱和。

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.

   Anthropic 的领导层曾公开估计，我们 90% 以上的代码是由 Claude 编写的，这包括脚本和实验性代码。而我们提到的“超过 80%”这一数字，衡量的是合并到生产环境的代码行中，可归功于 Claude 的比例。

   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.

   这是一种更为保守的衡量方式，原因有二：一是我们的归属追踪系统存在缺口；二是那些未归功于 Claude 的代码行中，还包括了自动生成的代码和其他既非人类手写也非 Claude 直接生成的产物。

---

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.

   代码产量的激增正使全球共享的基础设施承受巨大压力。GitHub 作为全球绝大多数软件的构建平台，在 2025 年全年约有 10 亿次代码提交；而到 2026 年中期，其每周的提交量已达 2.75 亿次，全年预计将达到约 140 亿次。该公司的首席运营官表示，为了跟上这一速度，他们正在“极力扩张”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.

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.

---

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).

---

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.

* 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.

Max plan

Team plan

Enterprise plan

Download app

Pricing

Log in to Claude

Models

Mythos Preview

Opus

Sonnet

Haiku

Solutions

AI agents

Code modernization

Coding

Customer support

Education

Financial services

Government

Healthcare

Legal

Life sciences

Nonprofits

Security

Small business

Claude Platform

Overview

Developer docs

Pricing

Marketplace

Regional compliance

Claude on AWS

Google Cloud's Vertex AI

Microsoft Foundry

Console login

Skip to footer

Resources

Blog

Claude partner network

Community

Connectors

Courses

Customer stories

Engineering at Anthropic

Events

Inside Claude Code

Inside Claude Cowork

Inside Claude Enterprise

Inside Claude Security

Plugins

Powered by Claude

Service partners

Startups program

Tutorials

Use cases

Help and security

Availability

Status

Support center

Anthropic

Careers

Economic Futures

Research

News

Claude's Constitution

Responsible Scaling Policy

Security and compliance

Transparency

Terms and policies

Privacy choices

Privacy policy

Consumer health data privacy policy

Responsible disclosure policy

Terms of service: Commercial

Terms of service: Consumer

Usage policy