公告

克劳德寓言5和克劳德神话5

2026年6月9日

今天我们推出**Claude Fable 5**：一款 Mythos[1] 级模型，我们已将其设计为可安全用于一般用途。

Fable 5 的性能超越了我们以往所有公开发布的模型。它在几乎所有人工智能性能测试基准测试中都处于领先水平，在软件工程、知识工作、视觉、科学研究以及许多其他领域都展现出卓越的性能。任务越长、越复杂，Fable 5 相对于我们其他模型的优势就越明显。

发布如此强大的模型也伴随着风险。如果没有安全措施，Fable 5 在网络安全等领域的强大功能可能会被滥用，造成严重损害。因此，我们在发布该模型时加入了安全措施，这意味着对某些主题的查询将由我们功能次强的模型 Claude Opus 4.8 进行响应。为了安全快速地发布该模型，我们对这些安全措施进行了保守的调整——它们有时会误报一些无害的请求，但平均触发率不到 5%。随着未来几个月功能更强大的模型陆续推出，我们正在努力改进安全措施，并尽快减少误报。

我们还面向一小部分网络安全防御者和基础设施提供商推出了**Claude Mythos 5**。它与 Fable 5 采用相同的底层模型，但在某些方面取消了安全防护措施。Mythos 5将首先通过与美国政府合作的“玻璃之翼计划”（Project Glasswing）进行部署，作为 Claude Mythos Preview 的升级版。它拥有全球所有模型中最强大的网络安全能力。我们计划很快通过更广泛的可信访问计划来扩大 Mythos 5 的访问权限。

Fable 5 和 Mythos 5 等模型的能力具有造福世界的潜力。我们在“玻璃翼计划”（Project Glasswing）中已经看到了这种潜力的萌芽，这些模型帮助网络安全防御者保护了至关重要的软件。我们也看到了它们在生命科学研究领域的应用，这些模型提出了新的假设，并加速了新疗法的研发。

Fable 5 和 Mythos 5 的定价分别为每百万个输入代币 10 美元和每百万个输出代币 50 美元——不到 Claude Mythos Preview 价格的一半。今天的联合发布是我们朝着目标迈出的又一步，我们将以最快、最安全的方式，为尽可能多的用户带来先进的 AI 功能。

对 Claude Fable 5 和 Claude Mythos 5 的评价

下表将《神鬼寓言5》和《神话5》的功能与其他领先机型进行了比较。

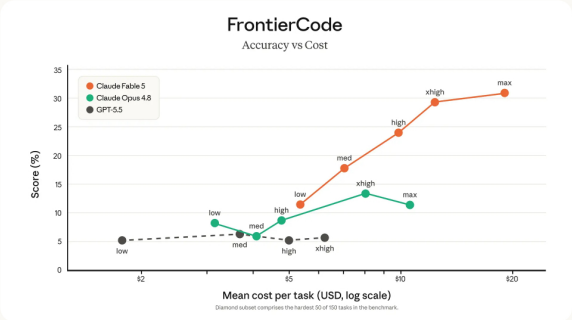
	Claude Mythos 5 / Fable 5	Claude Mythos Preview	Claude Opus 4.8	GPT 5.5	Gemini 3.1 Pro
Agentic coding SWE-Bench Pro	**80.3%**	77.8%	69.2%	58.6%	54.2%
Agentic coding FrontierCode (Diamond)	**29.3%** xhigh	–	13.4% xhigh	5.7% xhigh	–
Knowledge work GDPval-AA	**1932**	–	1890	1769	1314
Knowledge work vision GDP.pdf	**29.8%** no tools	–	22.5% no tools	24.9% no tools	16.7% no tools
Spatial reasoning Blueprint-Bench 2	**38.6%**	–	14.5%	36.2%	26.5%
Tool use AutomationBench	**17.4%**	–	15.5%	12.9%	9.6%
Computer use OSWorld-Verified	85.0%	**85.4%**	83.4%	78.7%	76.2%
Legal Legal Agent Benchmark	**13.3%**	–	10.4%	2.1%	0.0%
Multidisciplinary reasoning	**59.0%*** no tools	56.8% no tools	49.8% no tools	41.4% no tools	44.4% no tools

Humanity's Last Exam	64.5%* with tools	64.7% with tools	57.9% with tools	52.2% with tools	51.4% with tools
Biology BioMysteryBench	46.1%* hard	29.6% hard	40.0% hard	–	–
	83.9%* human solved	82.6% human solved	80.4% human solved	–	–
Agentic coding Terminal-Bench 2.1	88.0%*	–	82.7%	83.4% Codex CLI	70.7% Gemini CLI
Cybersecurity ExploitBench (Cap%)	78.0%*	69.0%	40.0%	34.0%	–
Health HealthBench Professional	66.0%*	64.7%	56.9%	51.8%	–

Methodology: Reported scores are within a 1-3 percentage point difference for Claude Mythos 5 and Claude Fable 5. This table shows the higher score of the two. Starred (*) benchmarks show a larger difference due to our blocking safeguards for cybersecurity and biology-related questions. For these benchmarks, Claude Fable 5 performs closer to Claude Opus 4.8 due to fallbacks. See the system card for details.

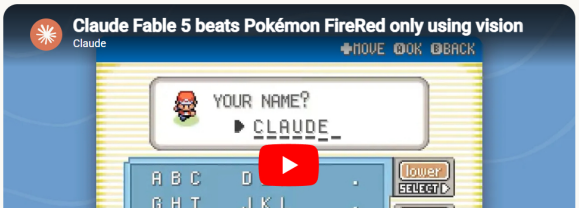
Fable 5 和 Mythos 5 的自主运行时间比以往任何 Claude 模型都长。下文我们将探讨这些能力如何应用于软件工程，并介绍该模型在知识工作、视觉、记忆和生命科学研究方面的改进能力。

软件工程。在早期测试中，Stripe报告称，Fable 5 将数月的工程量压缩到了几天之内。在一个拥有 5000 万行代码的 Ruby 代码库中，该模型仅用一天时间就完成了整个代码库的迁移，而这项工作如果由一个团队手动完成则需要两个多月。Fable 5 的令牌效率也高于以往的 Claude 模型：在 Cognition 的FrontierCode评估中，该评估旨在测试模型能否在满足高质量生产代码库标准的前提下完成高难度编码任务，Fable 5 即使在中等工作量下，也在所有前沿模型中得分最高。



知识型工作。Fable 5 在复杂的分析任务中表现出色。在Hebbia的高级推理能力财务基准测试中，Fable 5 的得分在所有模型中最高，在基于文档的推理、图表解读和问题解决方面均有显著提升。IMC指出，Fable 5 在其交易分析评估中几乎全面胜任，包括事实查找、概念推理、根本原因分析和预期值分析。

视觉。*Fable* 5 是目前最先进的视觉任务模型。它能够从复杂的科学数据中提取精确数字，并能执行复杂的基于视觉的任务，例如仅凭屏幕截图重建 Web 应用程序的源代码。它也更易于集成：例如，之前的 Claude 模型即使配备了提供额外辅助工具的组件，也难以流畅运行《精灵宝可梦 火红》，而 Fable 5 仅使用一个极简的、仅支持视觉功能的组件就轻松通关了《火红》。



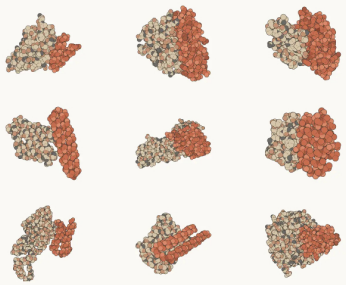
这段延时视频记录了克劳德仅使用游戏截图从头到尾游玩《精灵宝可梦 火红》的过程——没有使用任何地图、导航辅助工具或额外的游戏状态信息。早期的克劳德机器人需要复杂的辅助设备才能玩《精灵宝可梦》；而克劳德·费布尔5号仅凭视觉就完成了游戏。

内存和长上下文。*Fable* 5能够在长时间运行的任务中处理数百万个令牌，并利用自身笔记改进输出。当我们让模型玩卡牌构筑游戏《杀戮尖塔》时，赋予其对持久性文件级内存的访问权限，使其性能比Opus 4.8提升了三倍；Fable也更频繁地进入游戏的最终关卡。

日食 | 异星工厂 | VibeCAD | 具有经典电火花加工能力的流体

Claude Fable 5 构建了这个太阳系模拟，从物理学第一原理推导出行星的轨道运动，并用它来预测日食。

药物设计：借助Mythos 5，我们内部的蛋白质设计专家将药物设计流程的某些环节效率提高了约十倍。例如，他们发现Mythos 5仅需蛋白质设计和生物信息学工具，无需人工干预，其效率就能与经验丰富的操作人员媲美甚至超越他们。该模型能够执行通常由科学家完成的所有任务：选择结合位点、选择并运行蛋白质设计工具，以及从过程中出现的错误中恢复。本研究中14个蛋白质靶点中的9个（如下所示）是我们正在研究的强候选药物。



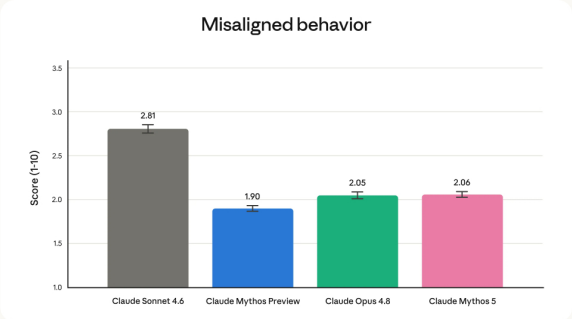
Mythos 5 设计的蛋白质复合物。目标包括免疫检查点、生长因子和受体信号传导、神经退行性疾病、肌肉疾病以及更难的结构靶点。

分子生物学领域的新假设。*Mythos* 5是我们首个能够持续产生新颖且引人注目的科学假设的模型。在与Opus系列模型的盲法直接对比中，我们的科学家在约80%的情况下更倾向于Mythos的分子生物学假设，并且已将其中几个假设推进到实验验证阶段。与此同时，Mythos的一个假设——一种关于大肠杆菌蛋白的新机制——在另一家独立研究同一问题的实验室的研究中得到了证实。

基因组学领域的新研究。*Mythos* 5在一周多的时间里，几乎完全自主地开展了基因组学研究。它收集了涵盖138种动物的数百万个细胞的单细胞数据，并设计训练了一个定制的机器学习模型，用于识别即使在亲缘关系较远的生物体中也执行相同功能的细胞。Mythos 5仅需少量人工干预，其训练后的模型性能就优于近期发表在《科学》杂

志上的一个模型——尽管其规模只有后者的百分之一。我们计划在未来几个月内发表这些研究成果。

一致性。在我们的自动化一致性评估中，我们发现 Mythos 5 的不一致性行为水平（包括模型采取的不一致性行为，例如欺骗行为，以及与用户合作滥用模型）较低，与 Opus 4.8 类似。鉴于它们使用相同的底层模型，Fable 5 的一致性水平也将与之类似。完整的评估结果以及其他安全性和功能测试的详细信息，请参阅模型的系统卡。



根据我们的自动化一致性评估，总体行为不一致程度较高。更多信息请参见系统卡的 *6.2.3.1* 节。

《克劳德寓言5》的早期反馈

提前体验过《神鬼寓言5》的玩家们自行进行了测试。以下是他们描述的部分测试结果：

CURSOR

“Claude Fable 5 是 CursorBench 上最先进的模型。它能够处理早期模型无法解决的一类长周期问题。

Michael Truell
首席执行官兼联合创始人

GitHub

“Claude Fable 5 对 GitHub 服务的开发者而言是一项真正的进步。在我们早期的测试中，它能够以超越以往基准的自主性和可靠性完成复杂、长期的编码任务。但最令我们兴奋的是它所指向的方向：未来开发者可以将越来越雄心勃勃的工作交给智能体，并在整个软件生命周期中信赖其结果。

马里奥·罗德里格斯，
首席产品官

01/14

克劳德《神鬼寓言5》的新安全措施

Mythos级模型已经发展到一定程度，构成重大风险。今年4月，我们启动了Glasswing项目，向部分网络安全防御人员和关键软件基础设施提供商发布了首个Mythos级模型（Claude Mythos预览版）。当时我们表示，希望最终能将Mythos级功能开放给所有用户，前提是我们开发出足够强大的安全保障措施，能够可靠地防止滥用。

过去几个月，我们一直在改进这些安全措施，现在它们已经足够完善，可以正式发布了。由于我们优先考虑安全性，因此我们特意将安全措施调整得较为谨慎，目前仍比理想状态更为严格——例如，有时一些无害的请求也会触发我们的分类器。我们意识到这可能会让部分用户感到不便，我们的目标是在发布后不断更新和完善安全措施，以减少误报。

下面我们将逐一讨论Fable 5的各项新增安全措施。我们更全面的安全措施方案已在模型系统卡和我们最新的风险报告中进行了讨论和评估。

安全分类器

Mythos级模型的尖端网络安全和生物学研究能力意味着它们对恶意行为者构成重大风险。也就是说，这些模型可能提供信息或建议，帮助恶意行为者造成严重危害，而这些信息或建议他们无法从其他来源（例如互联网搜索引擎）获得。此外，人工智能模型的许多高级应用都具有双重用途：对网络安全专家和生物学研究人员有益的查询，如果落入恶意行为者手中，则可能造成危险。

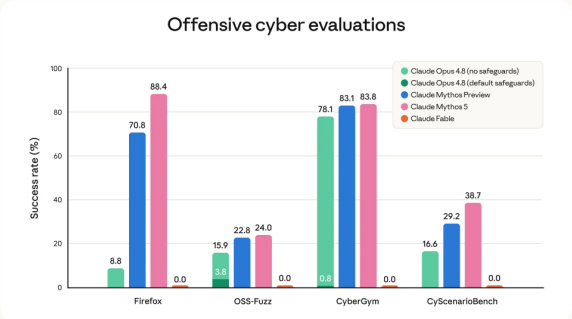
因此，我们需要强有力的保障措施来防止滥用，而且这些措施的覆盖范围必须广泛。这些保障措施本身必须能够抵御持续且复杂的绕过尝试（也称为“越狱”系统）。Mythos 级别的能力提升对许多对手来说都极具价值——例如，那些能够从网络攻击中获利的人——因此我们预计他们会有动机去尝试绕过我们的安全措施。

Fable 5 配备了一套全新的分类器：独立的 AI 系统，用于检测潜在的滥用行为，包括越狱尝试，并阻止主模型（在本例中为 Fable 5）做出响应。我们已经在模型上运行分类器一段时间了，Fable 5 的分类器是对我们先前工作的扩展，并增加了检测范围。

当 Fable 的分类器检测到与网络安全、生物学、化学或蒸馏相关的请求时，响应将自动由 Claude Opus 4.8 处理。用户会在发生这种情况时收到通知。Opus 4.8 本身就是一个功能强大的模型：相比 Fable 直接拒绝请求，回退到 Opus 的响应体验要好得多。我们的早期数据显示，超过 95% 的 Fable 会话完全不需要回退——对于这些会话，Fable 5 的性能与 Mythos 5 几乎相同。

以下是分类器涵盖的领域：

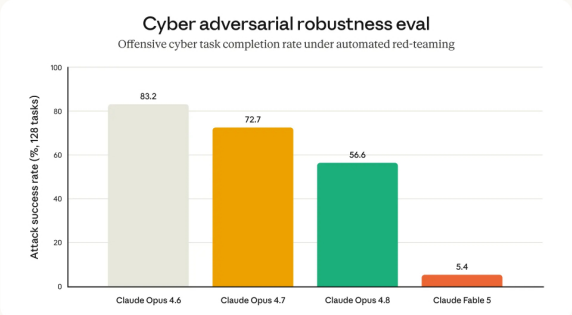
1. 网络安全。Mythos 类模型擅长发现和利用软件漏洞。因此，它们可以显著降低网络攻击的难度和成本。Mythos 类模型在智能黑客攻击方面也展现出强大的能力。这不仅包括发现漏洞，还包括执行网络攻击的多个不同环节——侦察、发现、横向移动等等。为了防止这些智能黑客攻击技能提升网络攻击的效率，我们设计的网络安全分类器能够更广泛地涵盖漏洞利用和攻击性网络任务。如下图所示，我们的分类器能够有效阻止 Fable 在这些任务上取得任何进展。



运行网络安全评估的结果为[3]，其中 *Fable 5* 处于阻止响应的模式，而不是回退到 *Opus 4.8*。评估没有涉及试图规避安全措施。

我们对分类器进行了广泛的红队演练，以测试其抵御越狱的鲁棒性。除了内部测试外，我们还开展了外部漏洞赏金计划，在超过 1000 小时的测试中，没有发现任何通用越狱漏洞。我们聘请的外部红队演练机构也未能发现任何针对长篇智能体任务的通用越狱漏洞——尽管英国 AISI 在短暂的初始测试窗口期内取得了进展。[4] 完全阻止通用越狱可能是不可能的，但我们的目标是使任何剩余的越狱手段都足够缓慢且成本足够高，以便我们能够在它们被大规模使用之前检测并阻止它们。

下图来自我们的一项内部评估，说明了《神鬼寓言5》的安全措施如何使其比我们之前面向公众开放的模型更能抵御越狱：



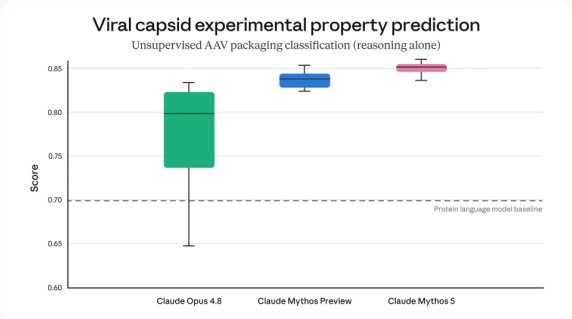
内部评估结果显示，自动化红队尝试使用该模型完成一项与攻击性网络安全相关的简短任务，共 *400* 回合，遇到阻碍时会重新开始或回溯。这些任务大多比较简单，并不代表真实的网络安全应用——有时甚至简单到加密远程服务器上的文件。在更复杂、更贴近实际的任务中，我们尚未在生产系统中发现成功的越狱案例。请注意，*Opus 4.6* 不具备阻止网络安全防护的功能。

我们的一位外部合作伙伴发现，Fable 5 抵御恶意网络查询的安全措施在所有测试模型（包括 Opus 4.8 和 Opus 4.7）中最为强大。Fable 5 完全能够应对所有与网络攻击策划、漏洞利用开发或防御规避相关的单轮恶意请求。无论这些请求是否使用了 30 种不同的公开越狱技术，这一结论都成立。

2. 生物学和化学。长期以来，我们一直利用分类器来阻止模型响应与生物武器相关的特定查询。但我们现在不再确信仅仅阻止这些特定查询就足够了。原因有二：首先，我们有理由担心资源雄厚的恶意行为者会试图利用我们的模型进行高风险的生物学研究。其次，模型现在更有能力完成现实世界的科学任务。

例如，我们测试了 Mythos 5 在腺相关病毒(AAV) 设计中完成一项具有挑战性步骤的能力。AAV 是基因疗法递送系统的重要组成部分，但如果落入不法分子之手，同样的

能力也可能被用于设计危险病毒。在这项任务中，我们评估了多种人工智能模型预测基因修饰如何影响病毒外壳组装的能力（这些病毒外壳来自Dyno Therapeutics开发的一组具有治疗意义但尚未发表的候选病毒）。我们并未专门训练模型执行这项任务——然而，仅凭生物学推理能力，Mythos 类模型就超越了专门用于蛋白质任务的复杂模型（称为“蛋白质语言模型”）。这表明 Mythos 5 在基因疗法研发中完成简单但重要的任务方面展现出了巨大的潜力，同时也凸显了这种双重用途能力所带来的风险。



本次评估结果显示，我们的模型能够预测一种简单病毒病毒外壳的未发表实验特性。病毒外壳组装是此类病毒特性中预测最简单的，但对于设计更复杂的特征而言，准确预测这一特性至关重要。*AAV* = 腺相关病毒。

我们的首要任务是尽快安全地发布 Fable，即使这意味着要采取过于宽泛的安全措施。因此，目前我们已安排 Fable 在处理大多数与生物学和化学相关的请求时回退到 Opus 4.8。与我们所有的分类器一样，我们希望尽快缩小这些安全措施的范围：正如以上证据所示，Fable 在科学领域具有巨大的积极应用潜力，我们不希望分类器的误报阻碍其发展。在接下来的几周内，一些生物医学研究人员和公司将能够加入我们的 Mythos 5 生物学功能可信访问计划（详见下文）。

3. 提炼。我们之前已发现有人试图大规模提取（“提炼”）Claude 的能力，用于在专制国家训练竞争模型。提炼 Fable 5 的能力可能间接导致接近前沿人工智能能力的扩散——而这些能力可能在缺乏适当保障措施的情况下被发布。被我们的分类器标记为此类提炼尝试的请求将回退到 Opus 4.8 版本。

一项新的数据保留政策

最后，我们将对Fable 5、Mythos 5以及未来功能级别相近或更高的模型处理企业客户数据的方式进行调整。我们将要求Mythos级模型的所有流量数据（包括第一方和第三方平台）保留30天。我们不会将这些数据用于训练新的Claude模型，也不会用于任何与安全无关的用途。此外，我们还实施了新的隐私保护措施，包括记录所有人工访问数据的情况，并确保在几乎所有情况下，数据都会在30天后删除（详情请参阅此帖）。这些数据将帮助我们防御复杂且新型的攻击（包括新的越狱攻击和跨多个请求的攻击），并帮助我们识别和减少误报。

Claude Mythos 5 和可信访问计划

从今天起，所有目前已获得 Claude Mythos Preview 访问权限的用户（例如，我们的网络安全合作伙伴 Project Glasswing）均可升级至 Claude Mythos 5——该版本与 Claude Fable 5 的模型相同，但取消了网络安全防护措施。用户会发现，在大多数情况下，Mythos 5 的性能与 Mythos Preview 相当，甚至更胜一筹，而价格却大幅降低。

经与美国政府协商，我们计划稳步扩大对 Claude Mythos 5 的访问权限，继续定期增加新的合作伙伴，并推行可信访问计划，使网络安全组织能够以更系统的方式申请。

我们的计划还包括启动一个生物学可信访问项目，旨在利用 Mythos 级功能加速生物医学研究并发现新的疗法。该项目将提供对 Fable 5 的访问权限，但移除生物学和化学方面的安全措施（但网络安全安全措施仍然保留）。该项目将招募少量来自不同生命科学机构的研究人员，涵盖基础研究和转化研究；我们计划在不断完善安全措施的同时，扩大该项目的参与范围。

可用性

Claude Fable 5 现已面向全球用户开放。Claude Mythos 5 目前仅限 Glasswing 合作伙伴使用（已解除网络安全保护），不久后将仅面向部分生物学研究人员开放（已解除生物学和化学安全保护），直至我们更广泛的可信访问计划推出。

两种模型的定价均为每百万个输入代币 10 美元，每百万个输出代币 50 美元。开发者可以通过Claude API使用 claude-fable-5 。

我们预计对《神鬼寓言5》的需求将非常高，而且难以预测。对于Claude API和按需付费的企业版套餐，《神鬼寓言5》已于今日全面上线。对于订阅套餐，我们更倾向于尽早开放访问权限，因此我们将采取更为保守的方式，分阶段推出：

- 从即日起至 6 月 22 日，Fable 5 将包含在 Pro、Max、Team 和基于席位的企业版套餐中，无需额外付费。
- 6月23日，我们将从套餐中移除《神鬼寓言5》。之后使用该游戏将需要消耗使用积分。如果套餐容量允许，我们将延长包含游戏的期限。
- 在此之后——一旦资源允许——我们计划将《神鬼寓言5》重新纳入订阅计划的标配。我们将尽快完成这项工作。

在此期间，我们会提前告知用户任何变更，以便用户了解最新情况。

2026 年 *6* 月 *9* 日更新：更新了关于 *AAV* 的讨论，指出候选药物是由 *Dyno Therapeutics* 开发的。

脚注

1. Mythos级模型是Claude系列的一个更高等级的型号，其性能高于Opus级。首款Mythos级模型Claude已于四月通过Project Glasswing发布。今天，Claude Fable 5和Claude Mythos 5也相继问世。
2. 寓言（Fable）一词源于拉丁语*fabula*，意为“被讲述的故事”，与希腊神话（Mythos）类似。正是这些保障措施区分了寓言和神话这两种模式，这也是我们给它们取不同名称的原因。
3. 指标：Firefox = 成功执行任意代码的试验比例（漏洞利用的完全成功级别）。OSS-Fuzz = 五级评分的严重性加权平均值（0.2 崩溃→1.0 控制流劫持），因此数值是加权平均值而非成功率。CyberGym = 重现目标漏洞的比例（公开排行榜指标）。CyScenarioBench = 所有挑战的平均成功率。
4. 通用越狱可以定义为任何提示、脚本或工具，它允许用户与模型交互，就好像模型的安全防护措施不存在一样。这与仅在非常有限的场景下有效或需要额外努力才能适应每种新情况的次要越狱截然不同。

相关内容

隆重推出 Claude 合作伙伴网络的服务板块和合作伙伴中心

阅读更多 →

我们从绘制一年来的人工智能网络威胁图中学到了什么

随着人工智能改变网络攻击的性质和方法，安全界使用的技术和框架还能有效应对吗？在这份新报告中，我们试图解答这个问题。

阅读更多 →

扩大玻璃翼项目

我们将“玻璃之翼”项目扩展到超过 15 个国家的约 150 个新组织。

阅读更多 →