

Mozilla | 3

开源AI 2026

第一版 202
6年7月

新西兰最北端，一名毛利语广播员在获得许可的情况下，为毛利语（一种市场太小而无法生存的语言）训练语音模型，该许可确保数据与其人民同在。在东非，农民利用在手机上离线运行的模型诊断木薯疾病，这些田野从未被云服务触及。在瑞士，一个公共联盟在公共超级计算机上训练了一个国家模型，并发布了全部内容：权重、数据、训练代码。

拉斐·克里科里安

首席技术官，Mozilla

他们都没有征得许可，而且他们也不可能租用这里。他们拥有这里——这才是关键所在。

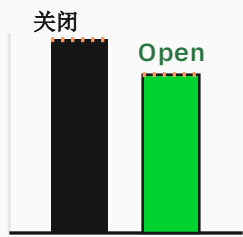
请将以下内容视为一张地图：OpenAI 在哪些方面取得了优势——一些数据甚至让我们感到惊讶——以及它在哪些方面存在不足。一个隐藏其弱点的情况就是广告。

条款。

01	当前状态	能力、成本与采用——证据
02	谁在赌	首都、公司、经济
03	为什么它正在发生 everywhere	可选择性，主权，国家战略
04	背带是新的前沿阵地	模型上方的层及其开放的间隙
05	五份赌注	层仍开启时，应在哪里操作

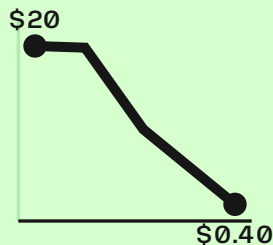
开源AI的现状

2026年开源AI，用四个数字来说明。



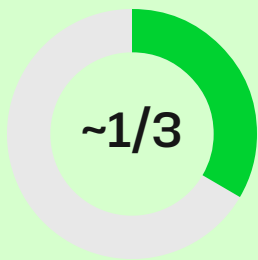
3.3 points

能力差距已缩小至顶尖封闭模型 — 开放
近乎均等，跨越崎岖边界之平均值



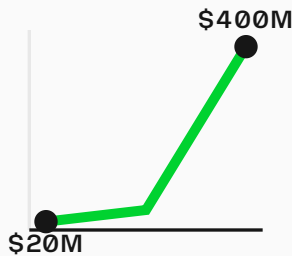
50x

36个月内推理成本下降——20美元至
每百万个token 0.40美元



~1/3

OpenRouter上由open提供服务的所有代币中
2025年底前完成的重量



20x

米斯特拉一年ARR增长率——大约
\$20M to \$400M

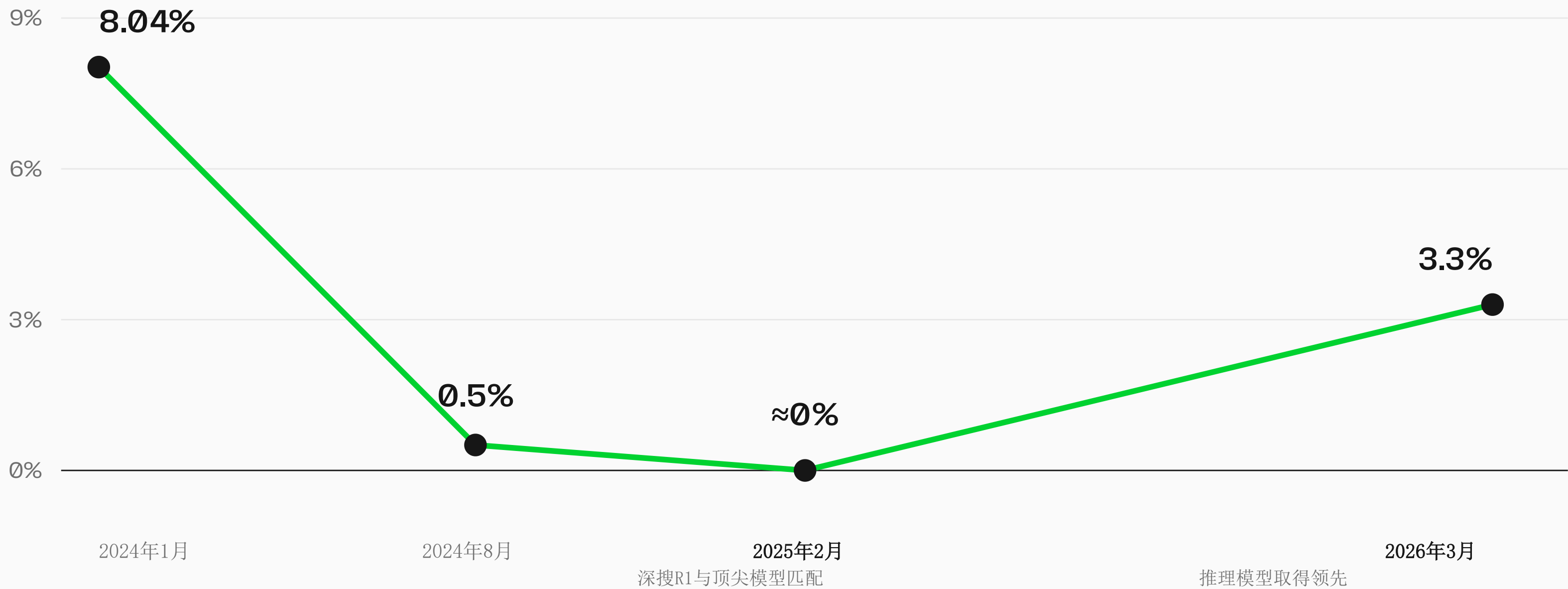
自2023年你尝试以来，发生了什么变化？

那次经历是真实的。它也有两代模型之久了。

2023	2024	2025	2026
- Llama 2 - 质量问题，幻觉 - 微调失败 - 恢复到 GPT-4	- Llama 3 时代 - 主要功能改进 - 更严谨的推理，更清晰的背景 - 一个真正的替代方案	- 深求时代 - 与前沿技术基本持平 - 降低成本，强大的经济效益 - 竞争性	- 生产规模 - 大规模企业采用 - 开模成本占~33%。代币 - 商业生态系统

The capability gap: 8.04% → 0.5% → 3.3%.

在Chatbot Arena上，通过24个月的三个阶段，实现封闭式引导胜过开放式引导。



LMSYS Chatbot Arena。目前排名前十的Arena席位中已有六个是已关闭的模型。

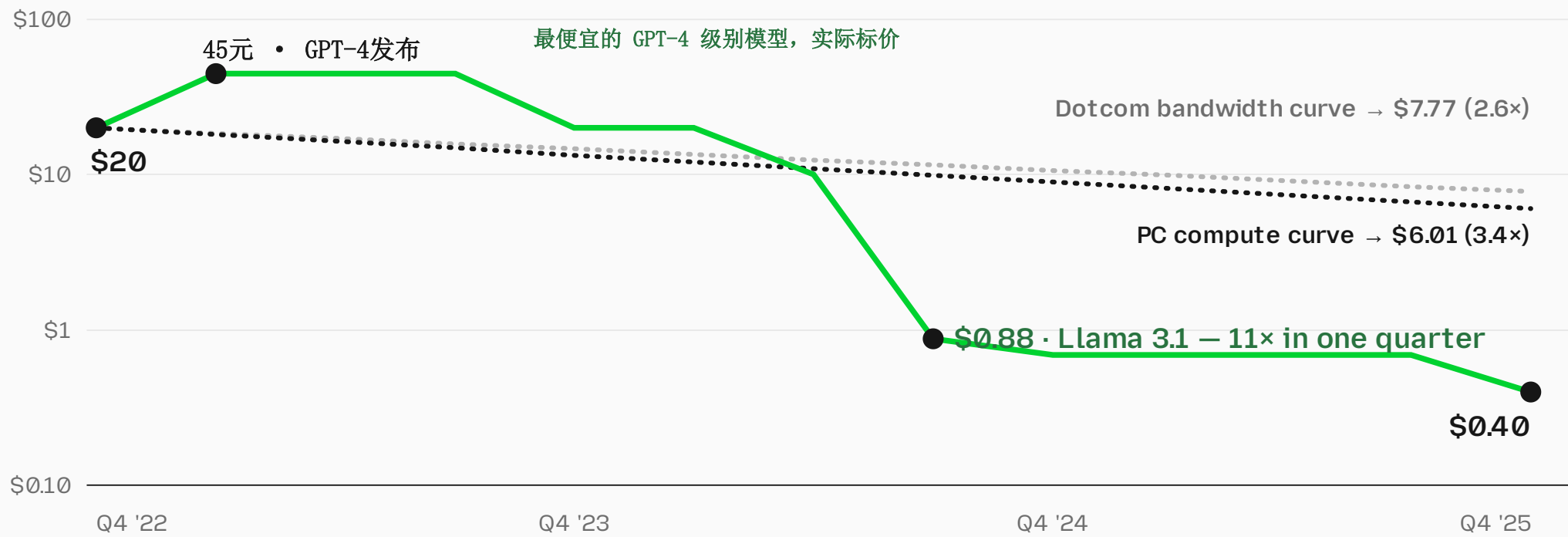
一个3.3%的平均数掩盖了参差不齐的前沿。

问题已不再是开放模型是否足够好，而是这项工作负载需要什么。

■ 平价或接近平价 编程 Instruction-following 常识	■ 已关闭仍具优势 推理 长上下文检索 代理任务
---	--

在36个月内，费用下降了50倍——从20美元降至每1000万代币0.40美元。

GPT-4级别性能的最低配置模型——混合API列表平均价格（出入均计），对数刻度——与先前平台迁移成本曲线的历史比率进行对比。



50×

2022年11月的尖端价格 → 2025年12月的GPT-4级价格（比GPT-4的45美元发布价高出112倍）

3.4×

同一36个月内，个人电脑计算性能的历史曲线带来了什么

2.6×

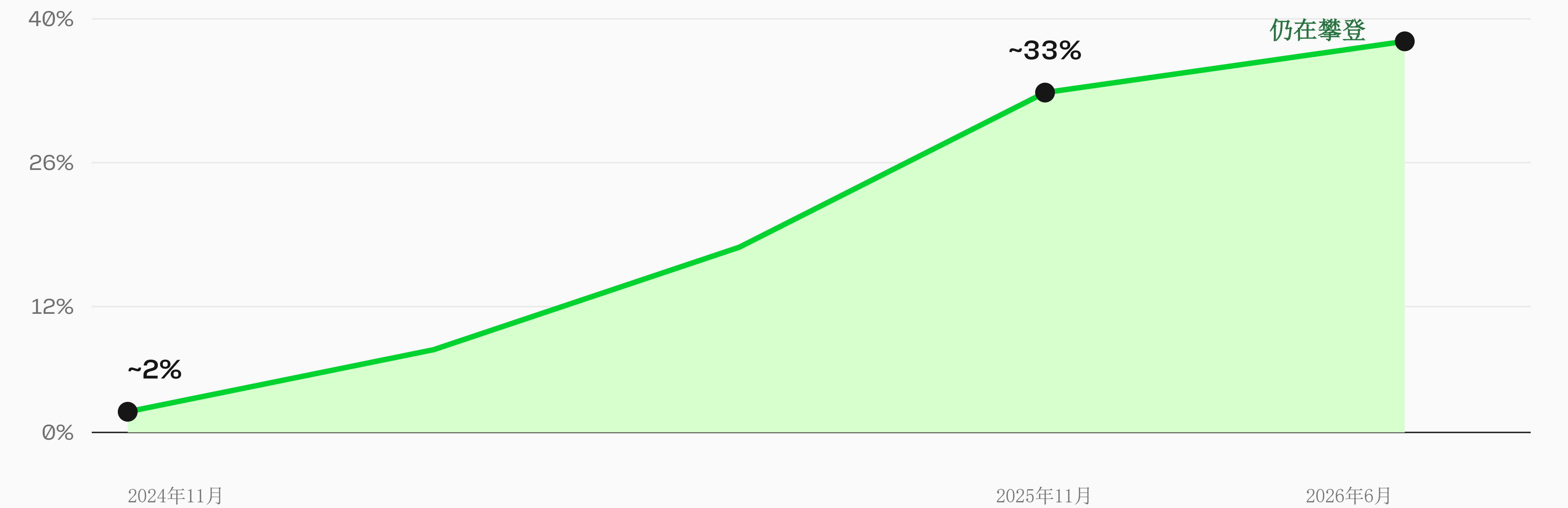
同期 36 个月里，点com带宽历史曲线所传递的信息

最大的跌幅是开放权重：Llama 3.1，然后是DeepSeek，在一个季度内将GPT-4级别的底线砸了11倍。

a16z “大语言模型通胀”（2024年11月）方法论 — 最低价模型 ≥ GPT-4级 MMLU，输入/输出列表价的混合平均值 • Epoch AI 大语言模型推理价格趋势（2025） • Intro 1/GPUunex 2025年12月指数。2022年11月基准点 = text-davinci-003，GPT-4之前的领先级模型。比较曲线：Meeker（KPCB 2014）历史价格率下的20美元。

开路称重站现在占有所有代币的三分之一。

所有通过 OpenRouter 路由的代币中，Open Weight 的份额，2024年11月 - 2026年6月。



开源路由器 100T-token 研究（2024年11月 - 2025年11月）；开源路由器 2026 年第一季度实时排行榜。测量点之间的指示曲线。

开放式重量主导代币使用。

OpenRouter上6月每月代币使用量。按使用量排名前五的模型均为开源模型。

1	深求V4闪存	深索		18.4T	↑ 95%
2	MiMo-V2.5	小米		14.9T	↑ >999%
3	H3 预览	腾讯		14.8T	↑ 67%
4	MiniMax M3	极小极大算法		14.3T	new
5	奥尔特阿尔法•	OpenRouter 潜行模式		11T	↑ 200%
6	克劳德 Opus 4.7	• 人类学的		9.02T	↑ 26%
7	深搜V4 Pro	深索		8.55T	↑ 112%
8	克劳德十四行诗第4首第6节	人类学的		7.33T	↑ 3%
9	克劳德 Opus 4.8	• 人类学的		6.18T	↑ >999%
10	深搜V3.2	深索		4.31T	↑ 5%

开启状态赢得代币。关闭状态仍赢得请求。

蒙特·亨利代币 • 前九大模型

请求 • 由提供方分享



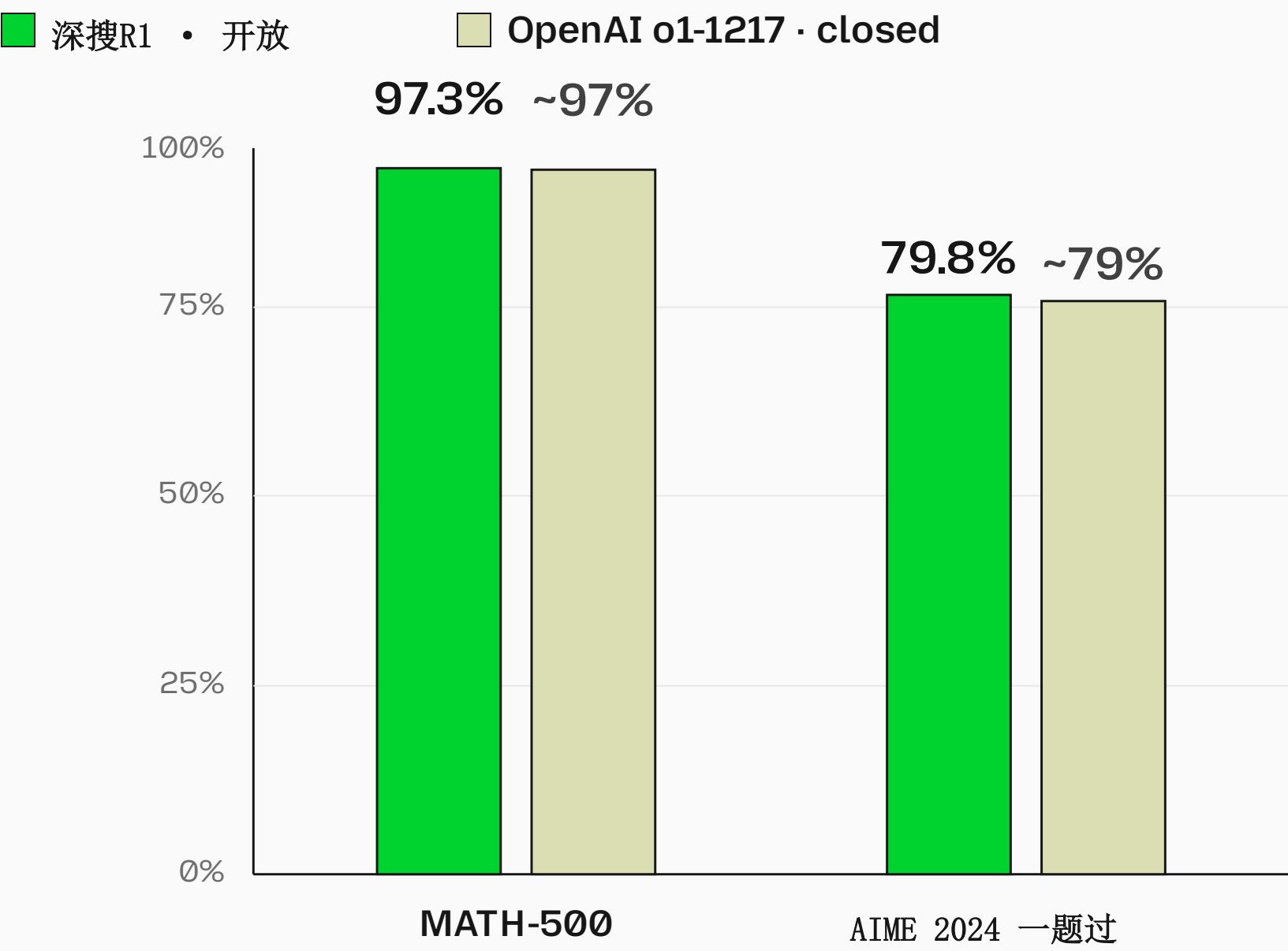
超过3.5:1。五款量级最高的模型均为开放权重，其中DeepSeek V4 Flash（18.4T）领先，其次是Owl Alpha / LongCat-2.0（11T）和MiMo-V2.5（14.9T）。

按请求数量计算，关闭的美国服务商仍然领先。开放的优势集中在代币密集型编码和代理型工作负载上——也就是支出所在的地方。

开放路由LLM排行榜，数据提取日期为2026年6月30日，涵盖6月份所有模型，采用月度统计窗口。需求：2026年年中对开放路由进行FT分析。仅统计路由流量——排除第一方使用（ChatGPT、Gemini、豆包）。6月30日Owl Alpha被识别为美团LongCat-2.0，按中文开放模型统计。

Fro tier reasoning 现在在 MIT 许可下提供。

DeepSeek-R1（2025年1月）对阵OpenAI o1-1217——推理能力同属一流，价格却只是几分之一。



价格与许可

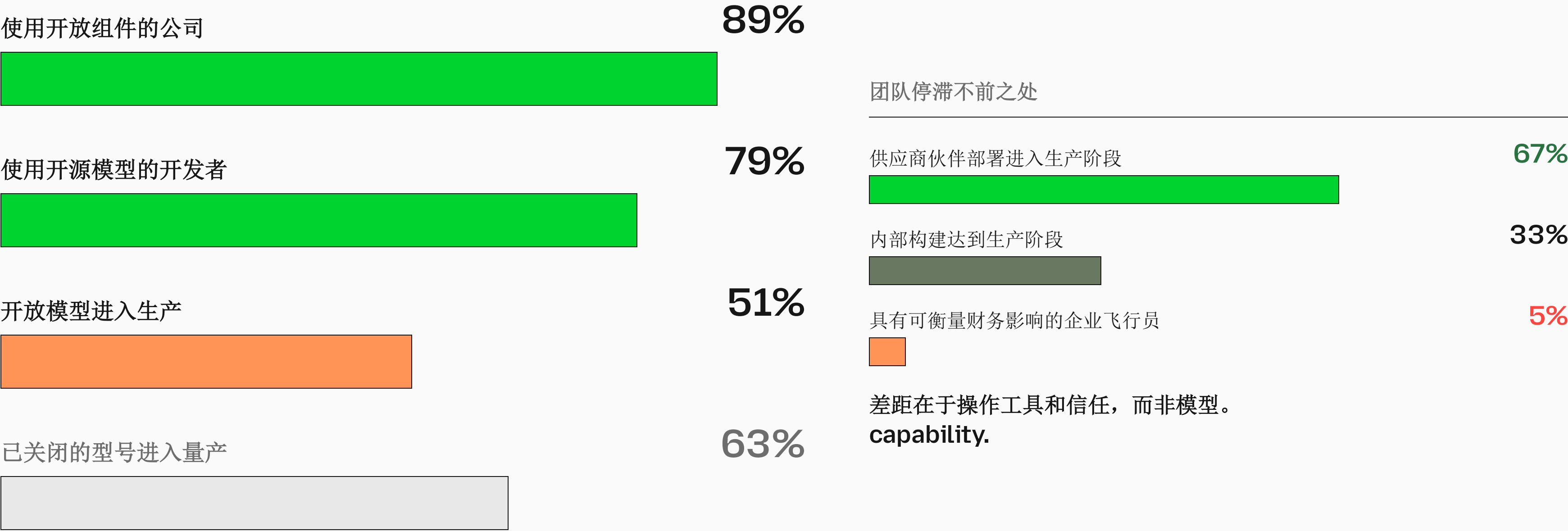
\$0.55 • \$2.19 每 100 万个 token 的 DeepSeek-R1 价格，
入 • 出 — MIT 许可证

15元 • 60元 OpenAI o1-1217 每百
万个token的价格，进 • 出 — 专
有
近乎相同的推理得分下，存在约27倍的价差。

2026年4月：DeepSeek-V4 Pro扩展至1.6T参数和1M令牌上下文，输入成本为0.435美元/兆。剩余差距在于框架，而非权重。信息来源：DeepSeek-R1技术报告
• OpenAI o1文档 • 服务商价目表。

开启船容易。开启部署难。

从实验到生产的采用漏斗。



来源：Linux基金会（89%）；Mozilla / SlashData 2026调查，样本量=1,494（79% • 51% • 63%）；MIT NANDA 2025（67% • 33%）；斯坦福2026企业AI playbook（5%）。

开放式栈在功能方面表现优异，但在操作方面较为简便。

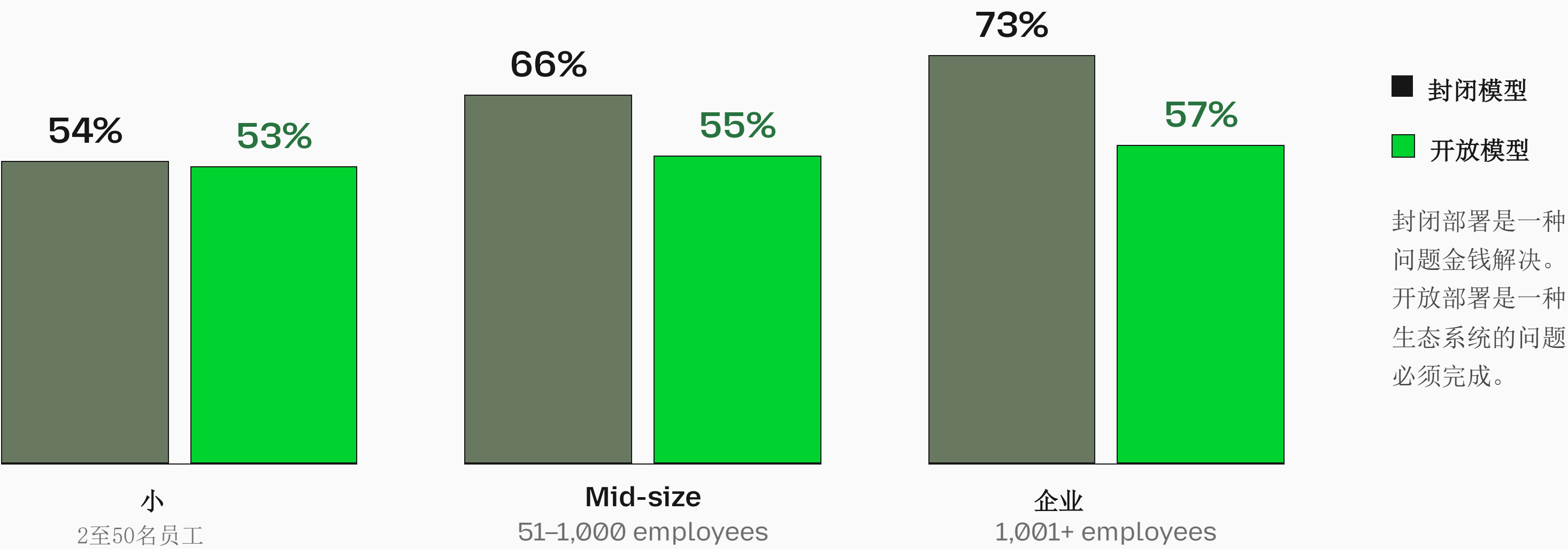
九层堆叠，按九项标准（1-5分）进行评分，各列按强弱顺序排列。最冷的两列——标准化和企业准备度——在每个层级中重复出现。

层	平均	社区	便捷性 采用	产品准备就绪	互操作	Sustain- 能力	完美 vs 关闭的	Docs	Standard- “化”	企业 准备好	层平均
模型代码		4.7	4.0	4.3	4.5	3.8	3.7	3.7	3.7	3.3	3.96
模型权重		4.6	4.0	4.0	4.0	3.4	3.6	3.2	3.4	2.8	3.67
基础设施		3.7	3.3	3.7	3.6	3.7	3.4	3.4	3.1	3.3	3.48
产品 / 用户体验		3.8	3.8	3.5	3.7	3.2	3.2	3.7	2.8	3.0	3.41
数据集		4.0	4.0	3.8	3.8	2.8	3.0	3.0	2.8	2.8	3.33
文档		4.0	4.0	3.0	3.3	3.5	2.8	3.0	3.3	2.3	3.22
许可		3.5	3.3	3.0	3.0	3.3	3.5	3.3	2.5	2.8	3.11
代理层		4.2	3.3	3.3	2.0	3.3	3.5	3.2	2.2	2.3	3.04
保护措施		3.4	3.0	2.8	2.6	3.0	2.6	2.6	1.6	2.2	2.64
标准平均值		4.00	3.63	3.54	3.40	3.35	3.27	3.25	2.83	2.79	
强 (≥4.0) 3.5–3.9 3.0–3.4 2.5–2.9 弱 (<2.5) 运营差距 = 标准化 + 企业准备											

来源：Mozilla开源AI技术栈地图，2026年7月 — 共9层，包含48个子组件。单元格代表层级；标准平均值为所有48个子组件的平均值。

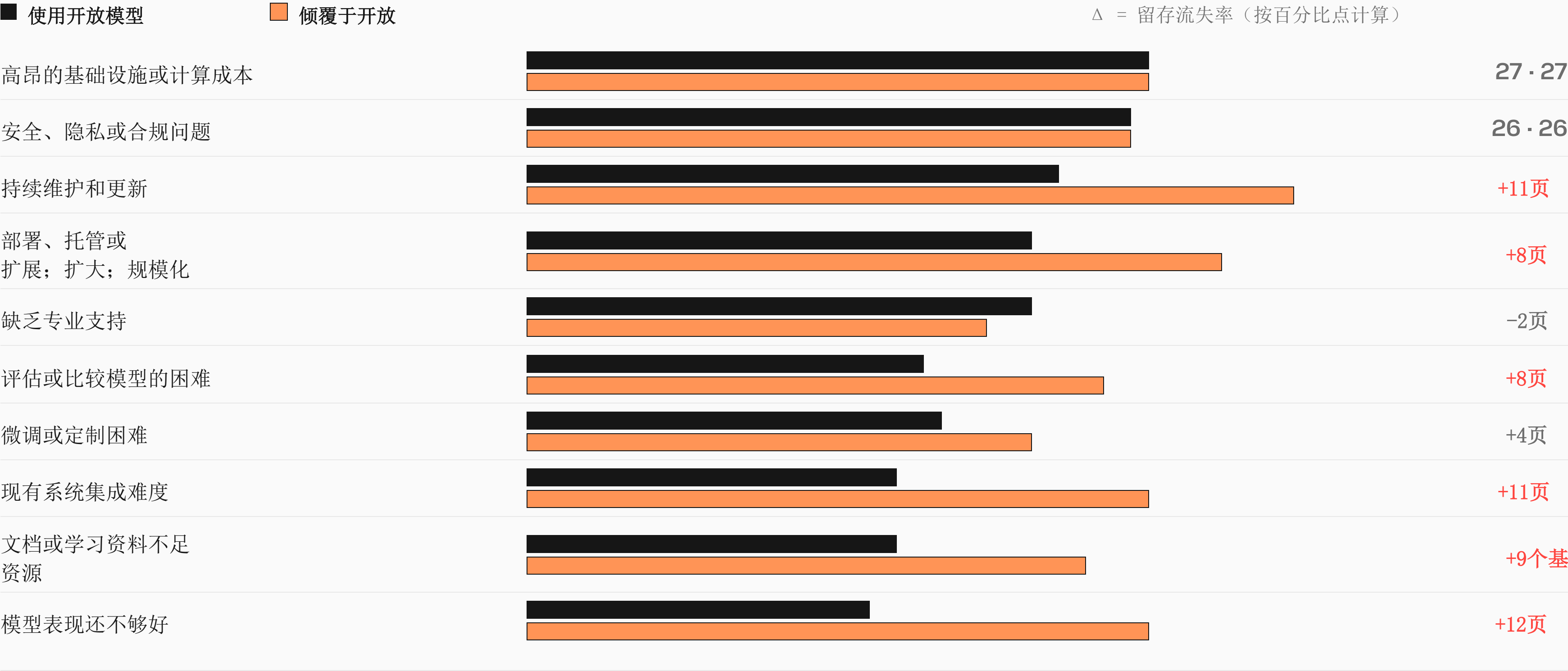
解除已部署的封锁，而非未部署的。

按组织规模划分，采用者进入量产的比例。如果差距是资源问题，规模本应缩小它——但并没有。



来源：Mozilla / SlashData 2026开发者调查，专业开发者 n=954。各类模型采用者的生产率。

什么阻碍了开放——以及什么让开发者离开。



来源：Mozilla / SlashData 2026开发者调查（样本量=1,410名当前或已流失的开源模型开发者）。条形图最高值按39%缩放。

同样的挑战，无处不在。

按地区划分，各挑战的命名者占比。

挑战	西欧 & 以色列	北美洲	更中国	南亚	东亚产 GC	南美洲	欧、亚、中东	大洋洲	All
高昂的基础设施或计算成本	25%	26%	29%	28%	28%	28%	29%	18%	27%
安全、隐私或合规担忧	20%	27%	18%	39%	29%	28%	25%	22%	26%
持续维护和更新	27%	26%	18%	26%	20%	31%	21%	25%	24%
部署、托管或扩展；扩大；规模化	27%	24%	19%	24%	11%	30%	26%	25%	23%
缺乏专业支持	17%	16%	21%	31%	24%	23%	23%	32%	22%
评估或比较困难模型	14%	17%	14%	23%	16%	26%	25%	18%	18%
微调或定制困难	22%	18%	18%	20%	11%	22%	18%	12%	18%
融入现有体系的困难系统	19%	21%	14%	20%	7%	26%	19%	20%	18%
文档不足或学习资源	18%	15%	15%	17%	15%	20%	24%	15%	17%
模型表现不佳足够	18%	15%	13%	22%	16%	17%	19%	8%	17%
没有重大挑战	9%	21%	16%	5%	14%	4%	8%	12%	12%
加权样本量	286	277	206	192	164	147	98	39	1411

来源：Mozilla / SlashData 2026 开发者调查（MZCS1），样本量 n=1,411。大洋洲（n=39）和东欧及独联体（n=98）低于可靠阈值。

谁在用它。

开磅计重是一种商业模式。

已融资公司、付费客户、实际收入。

Databricks 是一个基于 Apache Spark 的统一分析平台。 \$5.4B 营收年增长率，超过65%。 权衡一份价值1650亿美元以上的加薪评估	斯普林特用于构建、调试、协作和部署基于 Apache Spark 的应用程序。它简化了 Apache Spark 的运维。 \$400M ARR，12个月内上涨了20倍；在与 € 20B的估值进行 € 3B的融资谈判评估	探索 Spark 的应用程序。它简化了 Apache Spark 的运维。 \$220M ARR by mid-2025, mostly API and 企业；融资额达7.4亿美元，估值超过500亿美元+评估
---	---	---

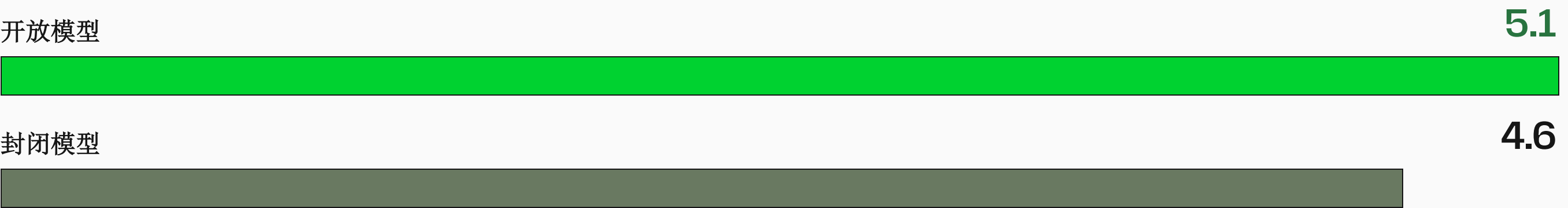
五种已被大规模验证的营收模式

- 托管推理
- 企业平台
- 本地授权
- 微调服务
- 夹具

专业基础设施已拥有付费客户，现正达到真正的规模。来源：公司披露，2026年6月。

开发者运行的开源用例比闭源的多。

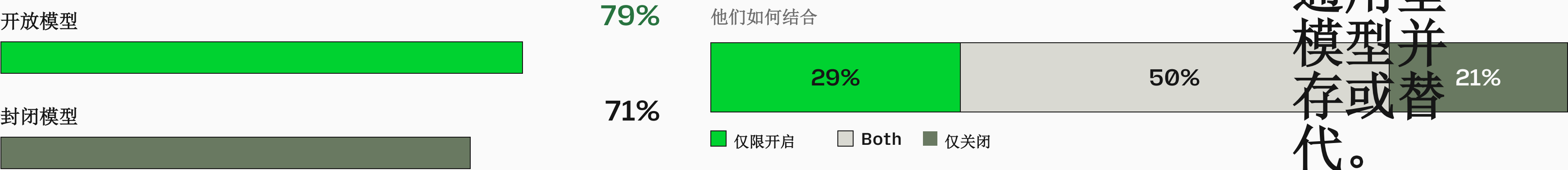
每个开发者的平均用例数量，以及开放型模型是如何组合的。



每个开发者的平均用例数。各受调查类别的开放线索。

78%的专用、任务优化型开放模型与通用型模型并存或替代。

哪些人使用每种型号？

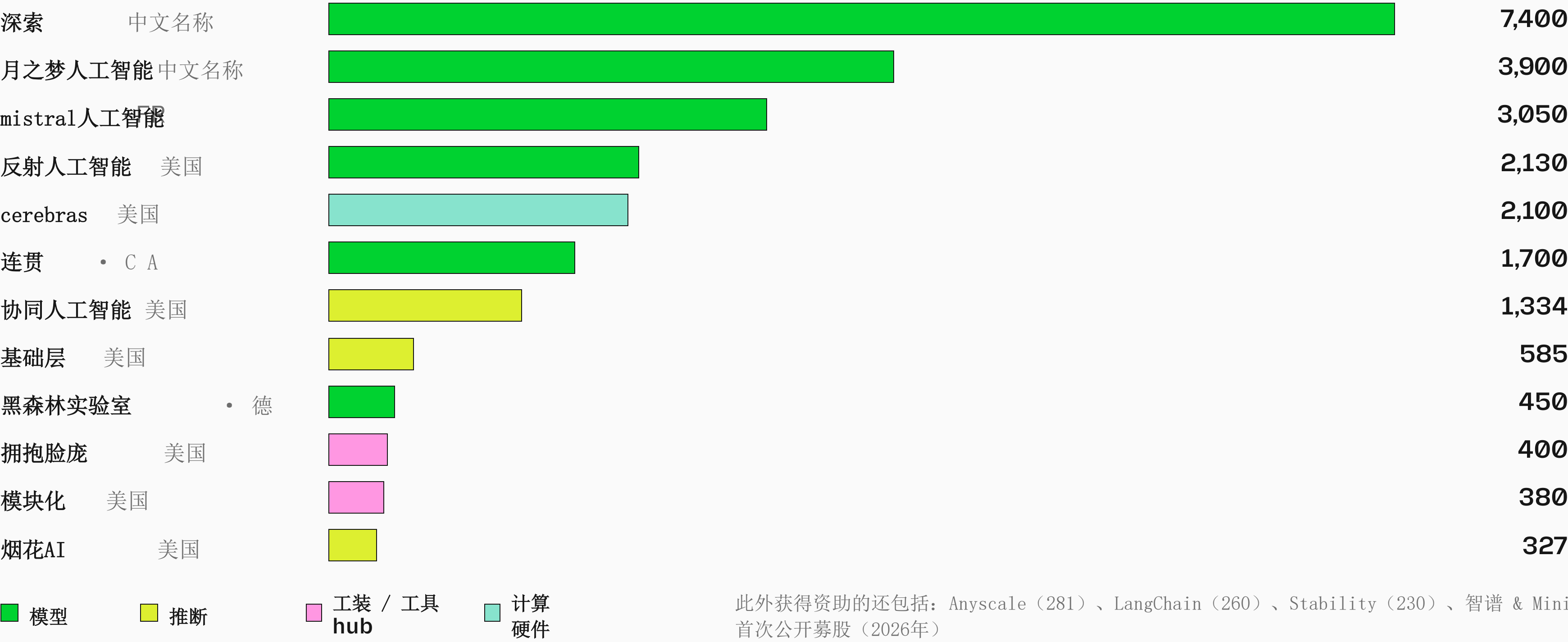


添加 AI 功能的开发者中，目前使用每种模型类型的占比。

开放收养中的潜在客户——但开放与封闭不能互相替代。大多数团队两者都做。

文丘里资助的开源生态系统。

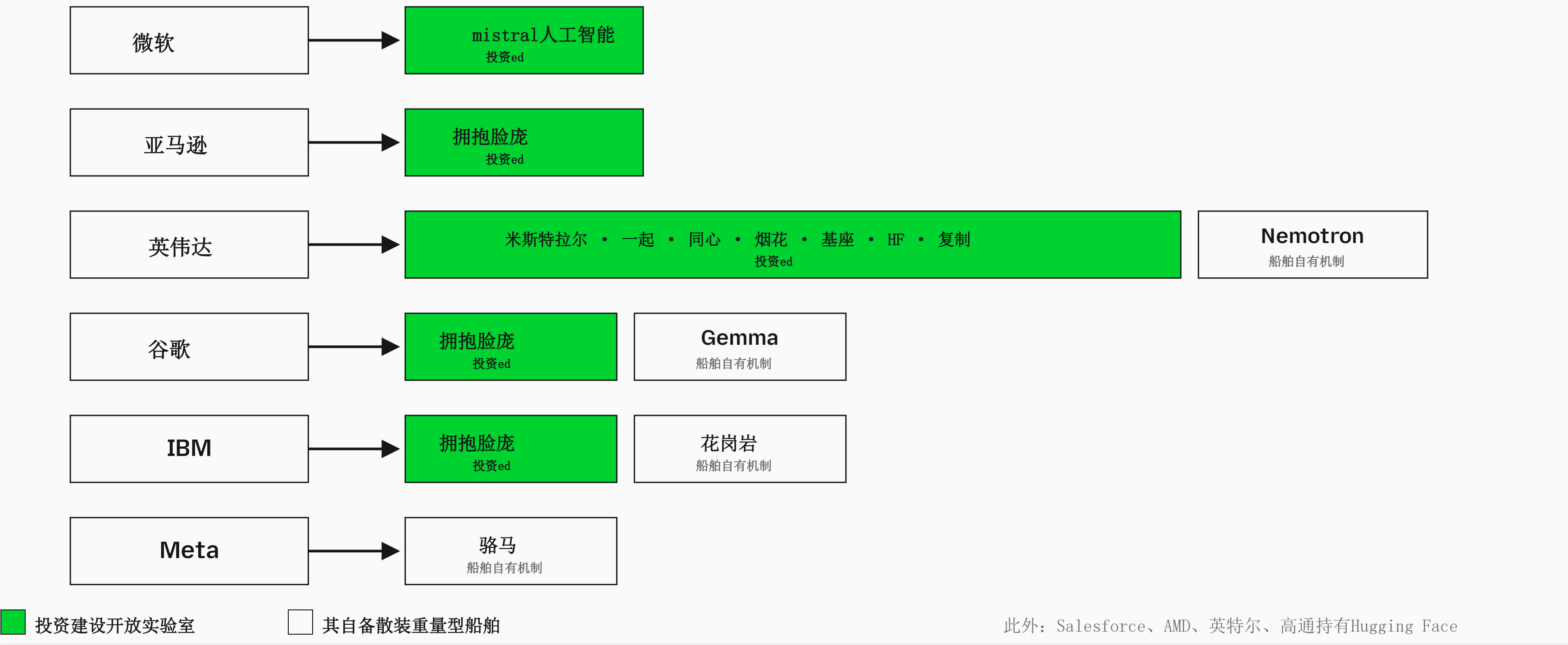
已披露总资金，百万美元。颜色标示堆叠层。



来源：公开文件和报告，2026年6月。条形图按DeepSeek的74亿美元进行缩放。

企业正在跨行业购买。

六家最大的科技公司悉数到场。



图表显示参与情况，而非资金规模。最大披露轮次：ASML → Mistral AI 1.4亿美元（2025年9月） • 腾讯 → DeepSeek 1.4亿美元（2026年6月） • CATL → DeepSeek 7亿美元（2026年6月） • 黑森集团 → Cohere/Aleph Alpha 6亿美元（2026年4月）。数据来源：公开申报及报告，2023 - 2026年。

合并已经开始。

2023年至2026年开源AI领域及周边的重要收购。

收购方	目标	目标是什么	美元 M	日期
核心编织	权重与偏差	MLOps 服务于开放模型构建者	1,700	2025年5月
Databricks 是一个基于 Apache Spark 的统一分析平台，用于构建、调试、协作和部署基于训练 Apache Spark 的应用程序。它简化了 Apache Spark 的使用，使数据科学家、分析师和开发人员能够更轻松地使用 Apache Spark。				
英伟达	Run:ai	GPU编排，拟开源	700	2024年末
AMD	筒仓人工智能	开放模型实验室，OpenEuroLLM联合负责人	665	2024
英伟达	格蕾特	合成数据	320	2025
英伟达	八爪AI	开放模型的推理优化	165	2024年9月
瑞布里克	Predibase	LoRAX 开放微调	100–500	2025年6月
连贯	阿列夫阿尔法	主权/企业开放式大语言模型	约200亿美元的交易	2026年4月

此外：ClickHouse - Langfuse, Mintlify - Helicone, Mistral - Koyeb & Emmi AI, Baseten - Parsed, 三个Hugging Face集成, Nvidia提供的另外三个——大多未公开。

资金堆栈有漏洞。

按堆叠层级和资本类型划分的资本强度，0-3。颜色越深代表资本越多。

堆叠层	私募风险投资	政府	慈善的	战略公司	收入 / 首次公开募股
基础模型	3	2	1	3	2
推理 / 服务	3	0	0	2	1
基础设施/计算	2	3	0	3	1
数据 / 数据集	0	1	2	1	0
工装 / MLOps	2	0	1	2	1
应用 / 垂直	2	0	0	1	1
安全 / 评估 / 治理	0	1	3	1	0

模型、推理和计算已获得大量投入。数据、安全和评估——即信任层——则完全依赖慈善机构和政府。资料来源：Mozilla于2026年6月对公开资金进行的分析。

米特模型在大规模应用时会出现问题。

三家企业，三次基于代币定价的碰撞。

微软

截至2026年6月30日，取消大部分Claude Code许可证

数月内，部门年度人工智能预算已被代币计费所耗尽。

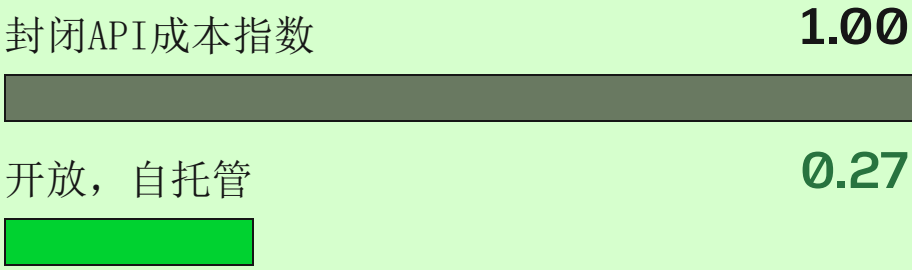
UBER

2026年AI编程预算四个月用尽

工程师每月账单为500至2000美元，但在每件工具每位员工支出上限为1500美元后，账单被限制在此额度内。

STRIPE

开源模型的推理成本降低了73%。



vLLM上承载了5000万次每日API调用，仅使用了三分之一GPU集群——这是一项公司拥有所有权并能进行预测的固定成本。

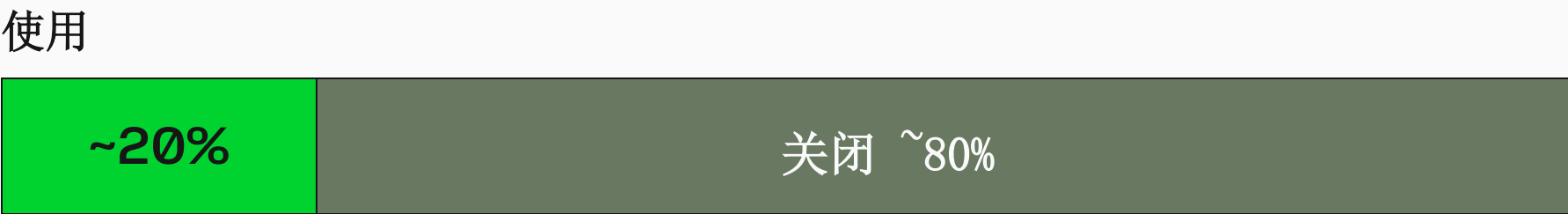
六月份之前

微软正在探索用于其最繁重的Copilot工作负载的Azure托管DeepSeek V4——这家全球最大的软件公司，正绕过其自身合作伙伴的计费。

两次失败的导火索：转向基于使用量的定价模式，导致供应商而非客户掌控了单位经济。当一项工具过于有用而不需要限制使用时，拥有其经济权比租用更重要。

五分之一的使用率，4%的收入。

OpenRouter上模型层使用量和收入的开放与封闭份额，2025年5月至9月。



按使用量计算，DeepSeek V4 Flash 是最常用的模型，其智能排名第六。Anthropic 拥有四个最强大的模型中的三个，并且大约处理了每周八分之一的 token。

~6x

每通电话的闭价约为90%的能力
奇偶性——驱动因素不是价格，而是能力

\$24.8B

尚未实现的年度储蓄
开闭价差不对称——Linux
基础估算

30 - 28

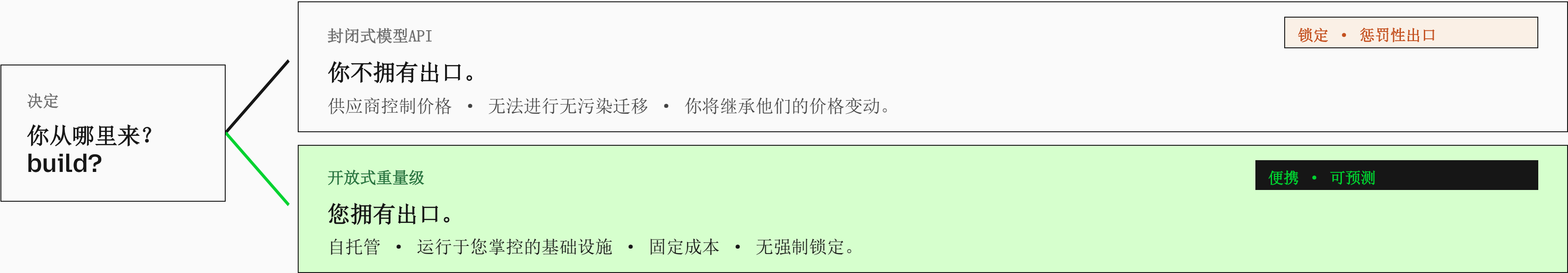
选择开源并引用的开发者百分比
低成本和隐私至上
原因

其使用处于通用化层面；收入则更高地归属于上层。

为什么它正在发生
，到处都是。

开放是选择，而非意识形态。

云时代已经做过这个实验：专有API加上数据引力使得退出惩罚性。遣返浪潮就是凭证。



封闭式模型API会重蹈覆辙。

先例 • 云时代已运行此实验			
\$90K–\$120K 将 1PB 数据从 AWS S3 迁移—— 外出罚金	~80% 目前一些企业正在遣返部分人员。 工作负载 — IDC	2.5× 盖可的云成本与预期对比 在遣返之前	\$10M+ 37signals预计的五年节省金额 离开之后

十九天里，最新的一级方程式赛车一直是深色的。选项性不再抽象。

Claude Fable 5上线后三天，一份美国出口管制令切断了访问权限。十八天后，一项第二项决定撤销了该令——访问权限于7月1日恢复，此时距离切断访问已过去十九天。所有基于该端点构建的人都从外部见证了这两项决定的发生。

6月9日	Anthropic 公司推出了 Fable 5 和 Mythos 5。
6月12日	商业部门立即实施出口管制。该命令禁止任何外国人——无论其在国内还是国外——获得访问权限。美国，包括Anthropic自己的外籍员工。国籍无法实时核实，因此合规检查具有选择性。不可能和两个模型都对所有人失效了。
6月26日	部分清查：神话系统恢复至约100家经审核的美国关键基础设施组织。
6月30日	控制解除。
7月1日	全球范围内恢复 Fable 5。

您可以关闭一个模型。您无法关闭已在您的机器上运行的副本。主权就是以国家规模持有同样的论点。

中国退出， 下载所有人。

累计Hugging Face下载量， 以及中文开放权重在OpenRouter代币中的占比。

累积拥抱表情下载 • 2026年3月



2026年2月， Qwen下载了接下来八个组织的全部数据。

开放式路由器上的中国开放式重量模型



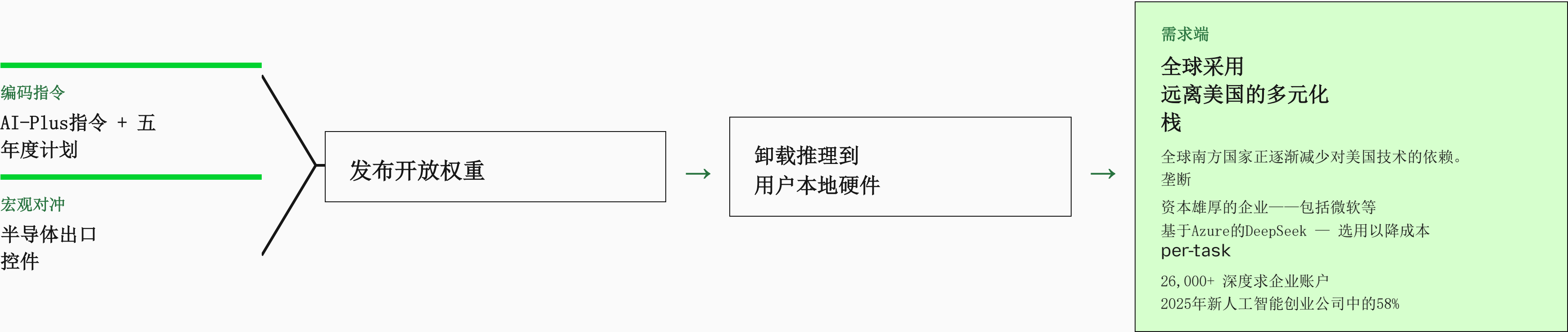
61% 平台使用频率最高的十个项目之间的交通模型

超过26, 000个DeepSeek企业账户； 在58%的2025年新人工智能初创公司技术组合中

警告是真实的——训练数据的不可见性，硬编码的拒绝。解决方案是架构性的：至少有八个司法管辖区禁止托管的应用程序；企业无论如何都会采用这些权重，无论是自行托管还是通过西方端点。来源：Hugging Face • OpenRouter。

开放性产业政策是中国产业政策。

转向中国开源权重是一个有意的政策结果。



全球需求与相对能力、每项任务成本以及区域战略协同紧密相连。资料来源：国务院“AI+”倡议（2025年8月） • 第十五个五年计划（2026年3月）。

欧洲正在将开放权重视为产业政策。

一个逐步升级的序列，2025年2月 - 2026年7月。

2025年2月 €109B 法国致力于人工智能基础设施。	2025年8月 人工智能法案豁免 符合资格的开源系统将在《人工智能法案》中获得定向豁免	2026年6月 欧盟奖 全欧盟24种语言支持、拥有4000亿参数的开源模型	2026年7月 阿玛莉亚 & SPARK 葡萄牙以550万欧元的价格出售一个完全开源的大型语言模型；德国开源其公共管理人工智能平台
---	--	--	--

我们不能依赖他人来维持医院运行、保障电网稳定和确保服务安全的技术。

加拿大：设计中的主权。

主权基础设施，由开放重量模型提供支持。

\$890M

为加拿大研究人员和创新者开放的、属于主权国家的AI超级计算机

2026年6月：特鲁多总理启动“AIforall”计划。其六大支柱之一是一个明确的“主权基础”——保护加拿大数据与知识产权，减少对外国供应链的依赖。

人工智能可以像互联网和手机那样发展——成为一种主要削弱人们权力的技术。或者，它也可以赋能使用者。我们正朝着后一种方向努力。

弗罗斯特·尼克（Nick Frosst），Cohere联合创始人——关于开源Command A+

A+ 命令（2026年5月）：一个用于企业智能代理任务的 218B 参数 MoE，仅需两台 H100 即可运行。北码 Mini 预计于六月推出——专为“主权开发者生态系统”而构建。

印度： 38, 000块GPU和500万名新开发者。

38,231 采用补贴价格（约每小时65卢比）入围的GPU 低于市场价约40%	5M+ 2025年新增Git Hub开发者达2190万。 —全球增长最快的基础
₹10, 372 Cr 印度人工智能任务预算，计划建设600个数据实验室	13.6% DeepSeek 的月活跃用户数量——仅次于中国

赌注：廉价的公共计算，加上一个多语言开发者群体，为全球南方用例生产本土基础模型和工具。来源：印度AI计划 • GitHub Octoverse 2025 • DeepSeek披露。

70多项国家人工智能战略及制定。

最大公开的国家人工智能基础设施投入。



80+ 拥有人工智能的司法管辖区政策 — OECD. AI	47 各国限制外国关键处理工作负载
12 新的国家战略2024年登录 — 牛津洞见	9GW 计划中的沙特产能2030

背帶是新的前层。

代理 harness 是另一个用户代理。

用户端编写代码，代表他们与世界进行协商——这比浏览器更高一层。

网络时代

浏览器

用户端编写代码，代表用户与服务器进行协商



代理时代

代理 harness

编排循环、工具、内存、沙箱和权限模型——生产难题集中的地方

已是一款产品类别

LangChain已获得126, 000+ GitHub星标，在编排领域拥有60%的开发者份额。

97M

每月MCP SDK下载量，首年拥有超过10, 000个活跃服务器

元利用

市场正在向上整合——统一层（例如 Databricks 开源的 Omnigent）将独立的代理生态系统打包并交换，置于同一治理平面之下。

背带市场图

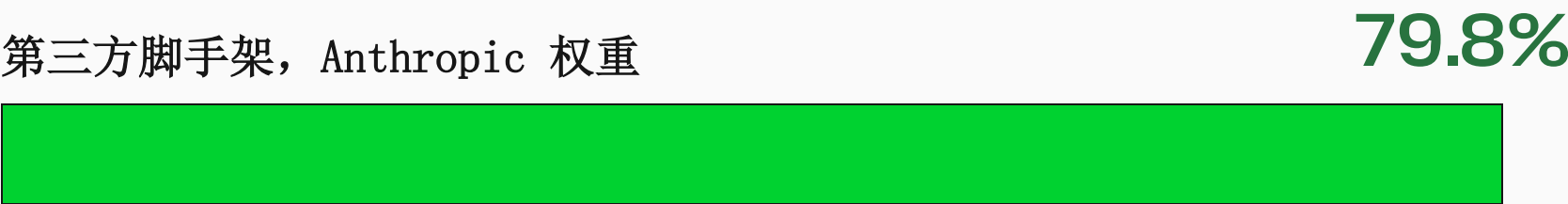


每个子层都有实际产品。编排和内存是开放引导的；互操作性基于开放标准运行；沙盒和评估是混合的。权限是最年轻、标准化程度最低的类别——开放差距。来源：Mozilla市场分析，2026年7月。

模型正在吃 harness。

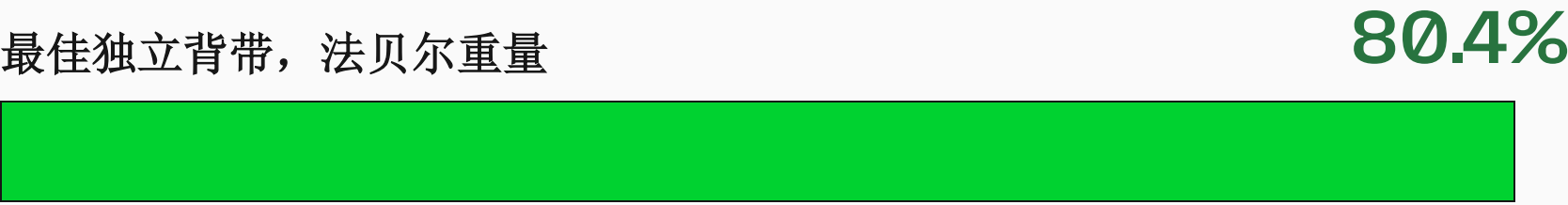
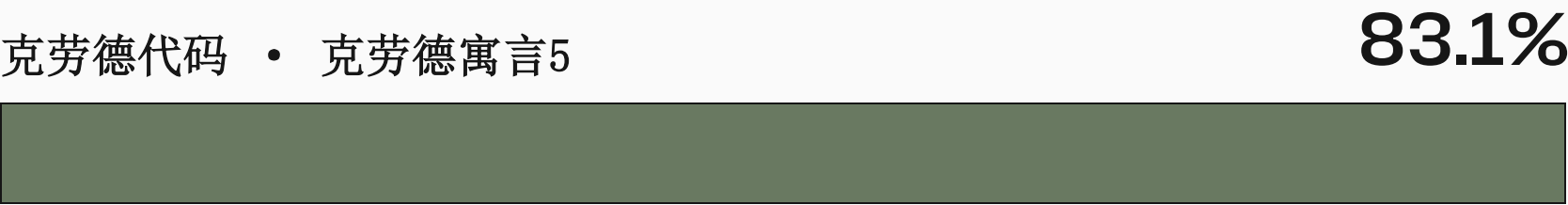
在实验室将线束内部化后，八周内将21.8点的线束优势压缩至约3点。

TERMINAL-BENCH 2.0 • MAY 2026



21.8分的优势——领先于重量的挽具

终端长凳 2.1 • 八周后

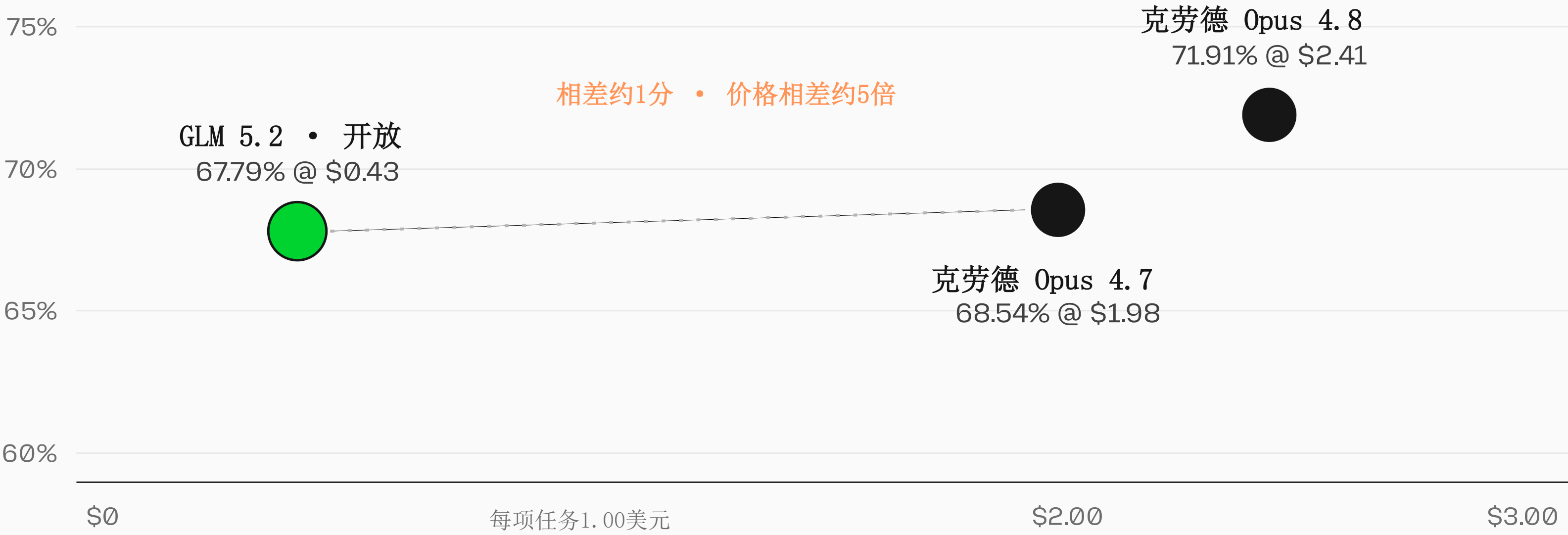


21.8分的差距缩小至约3分。在所有同时出现这两种型号的测试中，实验室自制的线束胜出。

针对某一实验室权重进行精细调优的系统能在其他模型上表现下降。锁定效应是优化的副作用——一道正在形成的护城河。在经过验证的顶级梯队中，未出现任何开放模型。来源：Terminal-Bench 2.0 / 2.1 排行榜。

在中性护套上，价格差距是5倍。

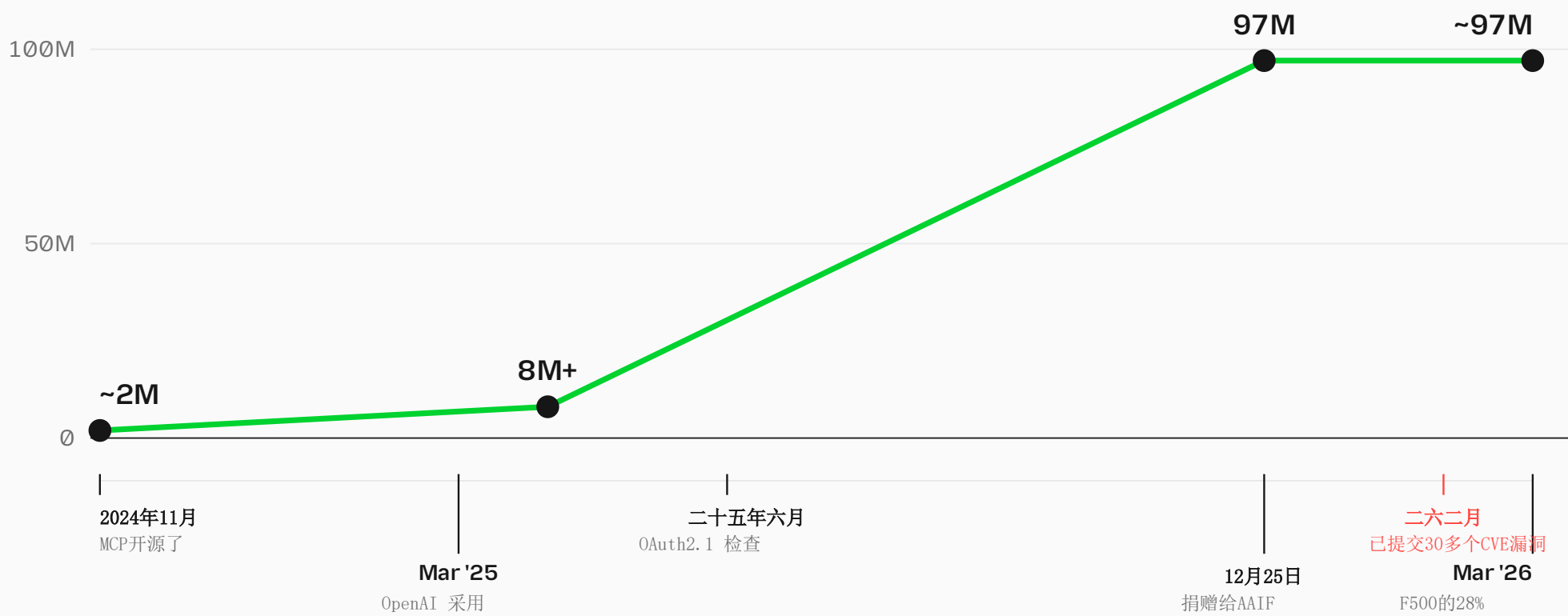
终端-工作台2.1分数与每项任务成本对比，所有模型均在同一中性基座上运行（vals.ai Terminus-2）。



集成同时也为实验室带来了数据飞轮——使用所产生的数据会流向拥有相关工具链的所有者。来源：vals.ai 终端-2 运行于 终端测试台 2.1。

MCP： 2M → 97M单月下载量， 历时16个月。

每月组合 Python + TypeScript 官方 SDK 下载量。四个有文档记录的点，直接连接——两者之间的一切都将是插值。



从约200万的启动基准点，
下载量增长了4750%。

28%
财富500强中的MCP
2026年3月前投入生产

30多个CVE
由安全研究人员提交的
2026年的前八周 —
就在MCP成为Linux之后
基础设施

每月下载量达9700万。八周内存在30多个CVE漏洞。

来源：官方MCP博客（2025年12月，AAIF捐赠） • Anthropic（2026年3月） • AAIF公告。9700万美国的数字参考了Anthropic公布的数据；未声称12月的精确性。

收养正在超越治理。

公共 MCP 服务器



~970万每月SDK下载量； 150多家组织在生产环境运行A2A。测量点之间的指示曲线。

~21%

据报道，许多公司已建立起成熟的代理治理机制——其采纳速度正快于管控步伐。

OAuth 2.1 + PKCE

现MCP强制要求； A2A船舶签署代理卡——身份验证已解决，授权未解决

2025年12月

Anthropic 向 Linux 基金会的 Agentic AI 基金会捐赠了 MCP，同时还有 Block 的鹅和 OpenAI 的 AGENTS. md。白金会员：AWS、Google、Microsoft、OpenAI。

互操作性层如今已成为中立地带。其上方的控制层仍属无人地带。资料来源：MCP 注册 • Linux 基金会 AAIF • 行业治理调查。

关闭并不等于安全。

Anthropic Slack MCP 2025年7月 通过链接展开实现零点击数据窃取 ✓ 检索：已检查 管理员 OAuth ✗ 未检查 攻击者服务器	关键
微软小助手“回声泄露” 2025年6月 隐藏的邮件指令，从未打开 ✓ 检索：已检查 受害者M365权限 ✗ 未检查 攻击者URL	关键
Salesforce Agentforce “强制泄露” 2025年9月 5个过期域名变成了一个可信的出逃通道 ✓ 检索：已检查 员工的CRM权限 ✗ 未检查 攻击者域	关键
ServiceNow “身体劫持者” 2025年10月 邮件 + 固定硬编码密钥 = 完全伪装，多因素认证被绕过 ✓ 检索：已检查 假冒用户权限 ✗ 未检查 攻击者身份	关键

您将隐私与安全联系在了一起吗？

封闭模型41%



开放模型29%



该机制追踪由谁承担操作负担：关闭的API默认内置安全措施；开放的部署则要求您自行接入相同的控制机制。它衡量的是工作由谁完成——而非风险存在于何处。

授权检索已送达非授权接收者——重量机密的保密未能阻止此事。

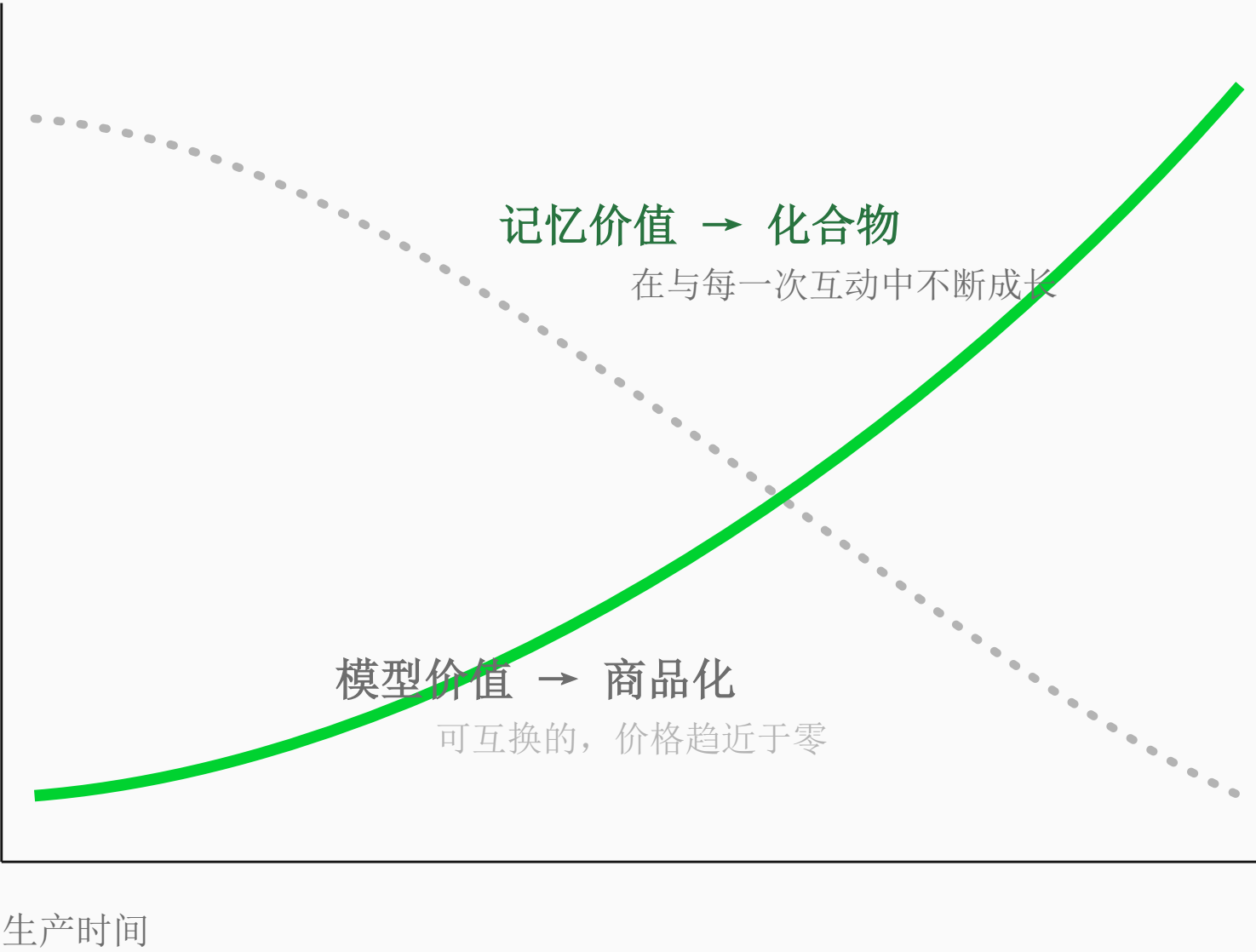
在推理、长期背景和问责制方面，封闭仍然领先。

IC终端部代理 82.0% vs 67.9% GPT-5.5（Codex CLI）与DeepSeek-V4-Pro-Max在终端-Bench 2.0上的对比：存在14分的差距，这一差距大于在静态基准测试上的差距	长文本上下文忠诚度 • 多针头1M个token 杰尼米 3 89% GPT-5.5 74% 克劳德 Opus 4.7 56% 深求V4-Pro 41%
火鸡合规与安全 SOC 2 · HIPAA · ZDR 默认情况下，封闭供应商套餐符合要求。开放式重量领导者在安全指数的评分中表现不佳——DeepSeek在FLI人工智能安全指数上获得F级评分（0.37/4）。	责任 相关方承担法律责任 向封闭供应商付款，东西坏了则由他人承担责任。自建开放重量，后果自负。

这些差距主要存在于连接器（harness）层面和合同签订过程中——而不是在权重（weights）上。来源：Terminal-Bench 2.0 • 长上下文评估 • FLI AI 安全指数。

该模型易于更换。内存是资产。

其中一部分会随着时间的推移而贬值，另一部分则会增值。



租用的模型可能被其实验室弃用。而防火墙的企业侧所持有的内存则不能——并且也不能通过更换供应商来重新获取。

可拥有的建筑

- 01 可移植格式——受版本控制的纯 Markdown，与本地联系人及电子表格同步，不受任何提供方管辖
- 02 在开放网络之前检索会读取私密记忆——没有自信的公众猜测来填补私密空白
- 03 仅追加写入——该存储器永不覆盖用户输入的内容

未解决的鸿沟是便携式许可。

两个利害关系截然不同的表面。

阅读

可逆，低影响

默认情况下可以很大程度上被允许

写

昂贵的或有不可逆副作用的

发送消息 • 支出预算 • 修改记录 • 执行交易。确认、阈值、成本上限和撤销必须集中于此。

Zero

生态系统内的12个框架、10个 harness 和3个对等协议均支持便携式写权限标准

认证
≠ 授权

MCP已迁移至OAuth 2.1；A2A标准化的已签名代理卡——两者都止步于身份验证

试图弥合差距的生态系统参与者——身份/认证：Okta、WorkOS、Auth0、Stytch、Arcade；策略/授权引擎：OpenFGA、Cedar。

头号威胁

CoSAI将同意疲劳列为顶级——用户批准了大多数提示，而重要的是那些授权行动的提示。

新兴的元 harness 架构在单个智能体之上实施有状态策略——根据会话已执行的操作来控制下一次写入。这就是持久化权限模型最有可能形成的环节。来源：MCP / A2A 规范 • CoSAI 威胁分类法。

机遇

保持层开启的五种方法。

1	构建开放式 harness	与开放权重共同设计，通用或垂直
2	拥有记忆	便携式，仅追加式，位于您的防火墙之后
3	解决便携权限	写入面，跨边界标准化
4	打破韵律	此时此刻寻找第二货源，因为它既便宜又乏味。
5	将开放设为默认复数形式	因此，公共领域从未只有一个供应商。

这些赌注都不需要超越前沿。它们需要拥有其上层的一一同时这些层仍然是开放的。依据：本报告第01 – 04节中的证据。

构建开放式 harness。

与开放权重联合设计，并像Codex调校GPT-5.5那样对其进行调校——通用型，或为特定领域量身定制型，前沿并非调校的目标。

数据

如今，该系统已领先一代产品，实验室已将其内部化。2026年，前沿模型轮次吸收了数千亿资金；而开放系统类别仅占一个舍入误差。

时钟

垂直整合正在进行中，今年尚未完成。当封闭式堆栈焊接模型与脚手架整合为单一租赁产品时，窗口将关闭。

站立式

一旦实验室同时拥有模型和训练数据集，训练数据集就变成了护城河。开放权重（模型参数）可以匹配前沿水平，但仍会输掉每一个重要的部署。

拥有记忆。

堆栈中唯一需要复合处理的东西。将其以可移植的、仅追加写入的格式保存在你自己的防火墙后面。

数据

开发生态工具早已实现规模化运行——Mem0已超过47,000颗星，并集成了Letta、Zep和LangMem。累积记录的价值随着每一次交互而增长，任何供应商切换都无法重建它。

时钟

必须在模型围绕其进行商品化之前确定架构。每个季度在一个封闭的端点上，都会将那项有价值的资产交给一个您无法控制的供应商。

站立式

数年积累的背景信息一旦失去访问权限便无法重建。该项被设计用来增值的资产，恰恰成为你最无力保护的——当你试图离开的那一天才恍然大悟。

解决便携式权限问题。

该背带在其中心有一个未解决的漏洞——即写入表面。 关闭此漏洞的团队现在拥有用于保持模型封闭的安全论点。

数据

没有便携式标准定义代理可以无监督地跨越任何边界执行哪些操作。MCP跨越了10,000台服务器和9700万月下载量；约21%的公司拥有成熟的治理体系。基础设施扩展了，但控制机制没有。

时钟

元数据约束层正在强制实施有状态写入策略。在这个窗口期，是开放标准来设定规则——抑或是封闭平台为所有人设定规则。

站立式

如果封闭平台首先定义权限和身份，那么每个开放部署都将继承为出租提供安全保障而编写的规则——而挽具中最有价值的位置则被封闭。

打破界限。

优步的票价在乘客围绕补贴价格重组生活后上涨了约92%。低价从来不是产品；锁定效应才是。现在找第二选择吧，虽然它又便宜又无聊。

数据

微软和优步在几个月内就因按量计费而耗尽了年度人工智能预算。Stripe对公开模型的服务推理成本削减了73%。自行托管服务在每天超过8000次对话量时能够收支平衡；这种价格不对称每年价值248亿美元。

时钟

introductory pricing将于2027-28年左右结束，届时供应商将上市，折扣也将用完。这趟行程最便宜的日子，就是你无法离开的前一天。

站立式

供应商设定的按量计费账单，将成为你路线图的最高上限。如果一个供应商确切知道某个出口会花费你多少钱，就没有理由不继续提高费用。

将开放设为默认复数形式。

一个拥有单一供应商的共享平台将不再是共享平台，无论该供应商是谁。资助一个足够开放、多元化的生态系统，让政府和企业需要时首先选择它。

数据

开放的生态系统正朝着单一来源倾斜：Qwen 拥有9.42亿次下载，而Llama为4.76亿次；中文模型占据了每周代币的45%以上。替代方案的需求是真实存在的，并且正在获得资金支持。

时钟

公共资金通过缓慢的中游运作，或因缺乏而默认汇聚于单一源头。扩散决定了谁能获胜，而扩散目前正获得资助——或者没有。

站立式

开放权重仍然隐藏着训练数据、对齐选择和拒绝模式。当一个源头提供默认选择时，这些选择会传播到每个衍生品中，而每个人都继承了单个构建者的盲点，却没有任何东西可以用来核查它们。

监视名单：什么能撤销这次阅读。

一项跨版本追踪的评估。任何两条车道同时退化都应作为修订的信号。

LANE	观看	反转时
能力与采用	3. 3%的差距；Open的OpenRouter代币份额，尤其是在代理编码中	在理由的空白处设立代币售货摊 拓宽
背带	实验室与独立终端测试台的价差 AAIF框架下的MCP/A2A治理；便携式 尚未存在的权限规范	实验室支架的引线变宽，或一个封闭式平台 首先设定权限标准
市场结构	开放实验室经济学（ARR、加薪、智谱/MiniMax） 首次公开募股（IPO）针对按量计价的分界点（约2027-28年） 以主权能力作为制衡	主权资金漏洞，或开放实验室经济学 无法扩展
信任与安全	误用功能及安全调谐条如何轻易被滥用； 硬摩擦区，尤其是合成CSAM和NCII； NTIA的“监测而非限制”原则是否有效	重大误用事件，或从监控转变为 限制

边界：开放是必要的，但不是充分的。它仍然需要在损害集中的地方进行硬摩擦——而拥有引擎决定了谁捕获收益，而不是谁拥有数据。开放将引擎置于可触及的范围，使竞争变得可以赢。追踪来源：LMSYS • OpenRouter • Terminal- Bench • AAIF • NTIA。

你可以运行一个测试来验证其余部分：看看在决定人工智能（AI）职位的人选的会议室里，谁被安排就座了。

如今，这些房间由引擎的所有者使用。当他们为那些使AI保持开放、便携和广泛部署——在平等基础上——的人提供座位的那一天到来时，从租赁到拥有的转变就将发生。

报告方法论。

开源AI状态报告引用了外部来源，并附有其各自的截至日期——其中包括Chatbot Arena、OpenRouter的代币和市场份额数据、Linux基金会委托的Nagle-Yue研究、Terminal-Bench、Hugging Face下载量、斯坦福大学人工智能研究所（HAI）、Epoch AI、麻省理工学院、经合组织（OECD）和牛津洞察政策追踪器等——由于排行榜和价格在数天内就会发生变化，因此该报告应被视为特定时间点的快照。该报告日期为v1，2026年7月（最后更新于2026年6月30日），其资金数据统计截止至2026年6月18日。这两项原始评估都是有意采用序数和方向性判断，而非测量值，并且每项评估都伴随着预期的注意事项：资金数据中的信息披露和承诺膨胀，以及对非西方和中国融资轮次的可见度降低。

开放式栈成熟度热力图是一项原创评估，依据九项标准（生产就绪度、易用性、文档、社区活跃度、与封闭式性能对比、企业就绪度、互操作性、可持续性和标准化）对栈的每一层进行1至5分的评分，采用固定评分标准，并在48个切片和1361个项目上，借助AI辅助确保一致性。该分类法在评分开始前就已固定，因此地图从未根据结果进行调整，表格中唯一计算得出的值是每行的平均值；其余所有单元格均为评级判断。因此，分数应被视为方向性指标，而非衡量标准。

资金成熟度分析将资本映射至开源和开源权重AI，时间范围从2022年1月至2026年6月18日。该分析区分开源权重与完全开源发布，涵盖七个技术栈层和五种资金类型。它仅统计现金、债务、补助金和量化实物承诺——从不涉及估值、年收入或市值——并对照由主要发布、重大报告和聚合者构成的多级来源层级体系验证每个数据。随后，每个单元在七个维度（规模、类型多样性、来源集中度、发展阶段、现金与实物对比、已实现与承诺对比、持久性）上获得0至3分的评分，总计为0至21分的成熟度等级，其中差距按战略重要性权重读作低成熟度。

调查方法

调查结果基于从“AI模型状态2026”在线调查中收集的数据。SlashData代表Mozilla设计和执行了该调查，通过结合专家研究小组招募受访者。该调查提供英语、简体中文、繁体中文、巴西葡萄牙语和日语版本。实地调查于2026年5月19日至29日进行，最终获得1494名合格受访者。

受访者需满足以下条件才有资格参与本次调查：他们是开发者，并且正在使用开源或闭源/专有AI模型为其应用程序添加AI功能。对符合条件的受访者，我们询问了关于其地理位置、软件开发参与情况、经验、公司规模、行业以及职位等背景问题。其余问题则涵盖了受访者如何使用和部署AI模型、他们对开源与闭源选项的看法和评价，以及他们做出选择背后的驱动因素和挑战。

结果按地区加权，以纠正样本实现上的轻微不平衡。加权因子为：大中华地区1.24，南亚0.82，东亚（不包括大中华地区）1.11，以及其他所有地区约为0.99。基准数据来自SlashData的Developer Nation调查，这是一项每两年进行一次的全​​球软件开发追踪研究。

关于Mozilla与SlashData

关于Mozilla。Mozilla项目是一个全球性的社群，其成员相信开放、创新和机遇是互联网持续健康发展的关键。自1998年以来，我们一直携手合作，确保互联网的发展能够惠及所有人。如今，Mozilla项目内的组织包括：Mozilla基金会，通过慈善风险资本捍卫我们的使命；Mozilla公司，开发了Firefox浏览器；MZLA，开发了Thunderbird邮箱；Mozilla风险投资，投资负责任的科技初创企业；Mozilla.ai，创建开源人工智能开发者工具；以及Mozilla数据集，一个用于人类自主和公平价值交换的数据共享平台。Mozilla.org也协调着这个组织体系，以及支持Mozilla使命的社区倡议。

关于SlashData。SlashData是一家领先的市场调研与分析公司，拥有超过20年的软件行业经验，与微软、英特尔、谷歌、亚马逊和Meta等顶级科技品牌，以及思科、戴尔、DigitalOcean、Okta、Stripe、Square和索尼等品牌合作。SlashData通过独立研究、基准测试以及对开发者、工具和软件市场变化趋势的前瞻性洞察，帮助平台、产品、营销和工程领域的领导者做出更明智的决策。