2026 年 6–7 月 · 前沿模型横向拆解

三模对垒

GPT-5.6 · Claude Fable 5 · Grok 4.5

谁在冲能力天花板，谁在抢单位成本，谁把安全做成了产品

OpenAI · GPT-5.6

Sol / Terra / Luna 三档同发

$5 / $30

输入 / 输出
每 100 万 token

Anthropic · Fable 5

Mythos 级能力 + 安全护栏

$10 / $50

输入 / 输出
每 100 万 token

xAI · Grok 4.5

Grok Build 默认模型

$2 / $6

输入 / 输出
每 100 万 token

资料来源：OpenAI《GPT-5.6》(2026-07-09)、Anthropic《Claude Fable 5 and Claude Mythos 5》(2026-06-09)、xAI《Introducing Grok 4.5》(2026-07-16) 官方发布文档

TIMELINE

38 天：一次贴身肉搏的正面交锋

6 / 9

Anthropic 发布 Claude Fable 5 与 Mythos 5，Mythos 级首次面向公众

6 / 12

Fable 5 / Mythos 5 访问被暂停，官方致歉并抢修

7 / 1

两款模型重新部署上线，恢复全量可用

7 / 9

OpenAI 发布 GPT-5.6 家族： Sol / Terra / Luna 三档同发

7 / 16

xAI 发布 Grok 4.5，与 Cursor 联合训练，成 Grok Build 默认模型

数据洞察

后发的两家，靶子都是 Fable 5。

OpenAI 在正文与图表中反复把 GPT-5.6 与 Fable 5 直接对位； xAI 官方给出的 5 张编码榜里，有 4 张的榜首是 Fable 5。

但 Grok 4.5 发布于 GPT-5.6 之后 7 天，全文对标的仍是 GPT-5.5 —— 评测的周期，已经追不上发布的节奏。

注： Anthropic 于 6 月 12 日公告暂停访问， 7 月 1 日公告重新部署，官方未在发布页披露暂停原因。

POSITIONING

一句话看懂：三家各自想让你记住什么

OpenAI

GPT-5.6 (Sol / Terra / Luna)

“每一个 token 里榨出更多智能”

- 核心叙事：性能／美元。不比谁分高，比同样的钱能干成多少活
- 产品化手段：三档模型 + max / ultra 档位，能力按需付费
- ultra 默认并行调度 4 个智能体，用 token 换时间

“更强的每美元性能：同样的花费做更多事，或以更低总成本得到同等结果。”

Anthropic

Claude Fable 5 / Mythos 5

“一个被做安全了的 Mythos 级模型”

- 核心叙事：能力上限。任务越长越复杂，领先幅度越大
- 产品化手段：把安全等级本身做成两个 SKU（Fable / Mythos）
- 主打真实成果： Stripe 迁移、蛋白设计、新科学假说

“Fable 5 的能力超过我们此前公开发布过的任何模型。”

xAI

Grok 4.5

“单位时间、单位成本下的最高智能”

- 核心叙事：速度 + 便宜 + 省 token，三者一起打
- 产品化手段：与 Cursor 联合训练，直接绑定开发者入口
- 横向扩张： Grok Build 默认模型 + Word / PPT / Excel 插件

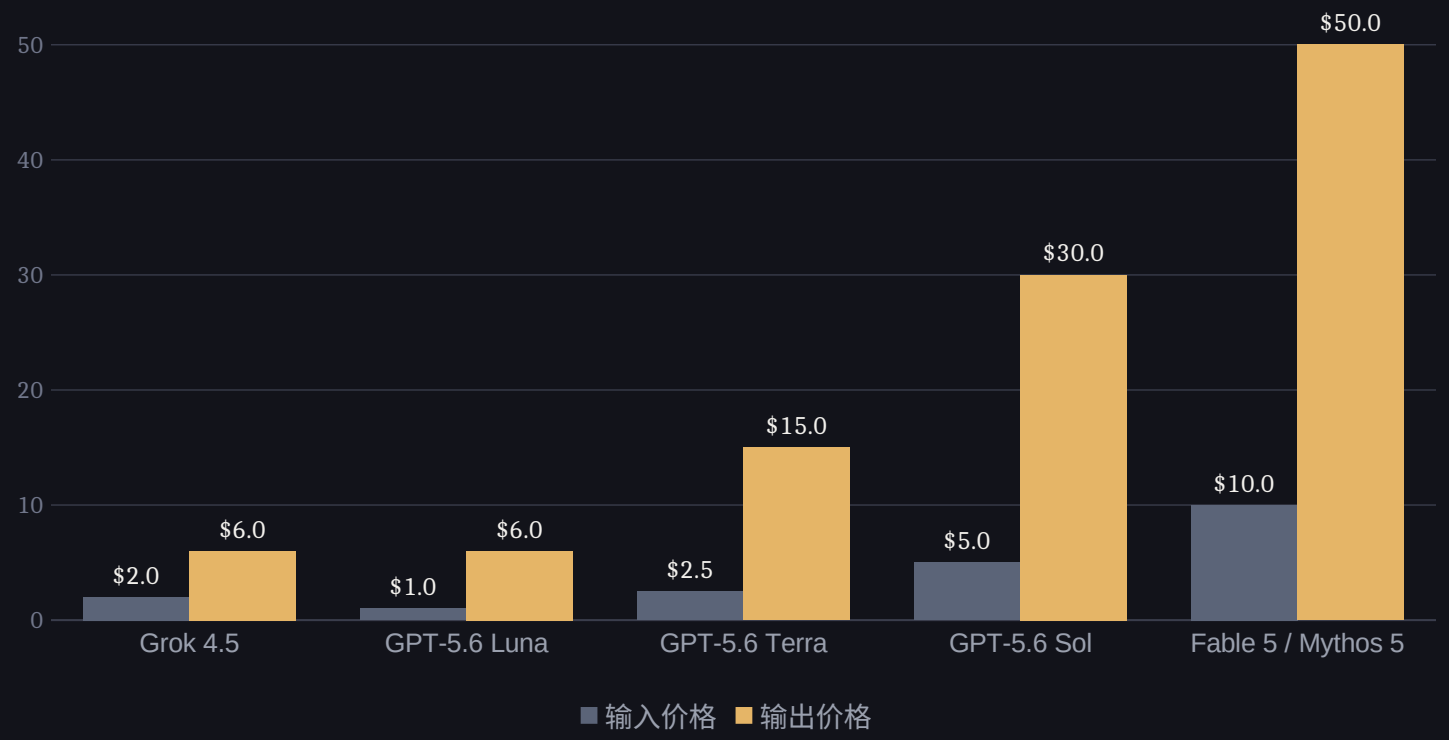
“Grok 4.5 提供单位时间与单位成本下的最高智能。”

引文均译自三家官方发布文档正文。

PRICING

价格：一条 8.3 倍的输出成本梯度

美元 / 每 100 万 token



8.3× 输出 token 价格差

Fable 5 的 $50 对 Grok 4.5 / GPT-5.6 Luna 的 $6

5× 两家旗舰之间的价差

Fable 5 $10/$50 对 GPT-5.6 Sol $5/$30 —— 输入 2 倍、输出 1.7 倍

−50% Anthropic 自己的降幅

Fable 5 / Mythos 5 的定价不到 Mythos Preview 的一半

三家把价格锚在了完全不同的位置——这不是竞价，是分层。

数据来源：三家官方发布文档定价章节。 GPT-5.6 另有缓存写入 1.25× 计费与 90% 缓存读折扣，此处未计入。

OPENAI · GPT-5.6

把「智能」拆成可采购的三档 + 两个加力挡

Sol　旗舰　$5 / $30

AA 编码智能体指数 80 分登顶； BrowseComp 90.4%、 OSWorld 2.0 62.6% 双项 SOTA

Terra　均衡　$2.5 / $15

编码指数 77.4，已超过 Fable 5 的 77.2；官方称 Terra 与 Luna 以约 1/16 成本在 Agents' Last Exam 上跑赢 Fable

Luna　最省　$1 / $6

编码指数 74.6，超过 Opus 4.8； SWE-Bench Pro 62.7% 逼近自家旗舰

ultra：把并行做成一个可选档位

默认同时调度 4 个智能体分头作战。 Terminal-Bench 2.1 由 88.8% 拉到 91.9%，BrowseComp 由 90.4% 拉到 92.2%—— 用 token 换时间与分数。

程序化工具调用（PTC）

模型自己写小程序去编排工具、过滤中间结果。 Clio 报告多步文档分析提示 token 降 38% 且质量无损； PlayCo 场景构建总 token 降 63.5%、模型轮次降 50.1%。

设计判断力的跃升

能检查渲染结果再回头修改，而不只是生成代码。 Triple Whale 前端评测得 4.4 分（GPT-5.5 为 4.0、 Claude 4.8 为 3.5）；可推断并复用 Slide Master 中的排版规则。

安全边界与自我改进

RSI 自改进指数 57.9%，较 GPT-5.5 提升 16.2 分；内部研究员日均 token 消耗是 GPT-5.5 峰值的 2 倍以上，近半年内部编码推理算力增长 100 倍。

数据来源： OpenAI 《GPT-5.6》发布文档正文、客户证言与官方 benchmark 表。

ANTHROPIC · CLAUDE FABLE 5

同一个模型，两种放行方式

能力面：任务越长，领先越大

1 天 — **软件工程**

Stripe 的 5,000 万行 Ruby 代码库全库迁移，1 天完成；人工团队原需 2 个多月

3 × — **记忆与长上下文**

《Slay the Spire》中持久化文件记忆带来的增益是 Opus 4.8 的 3 倍，进入终章的频率也高 3 倍

10 × — **药物设计**

内部蛋白设计流程加速约 10 倍；14 个靶点中有 9 个产出了可继续开发的强候选

~80% — **科学假说**

盲测中科学家约 80% 的比例更偏好 Mythos 5 的分子生物学假说，其中一条已被独立实验室论文佐证

放行面：安全等级即产品 SKU

同一套底层权重

Claude Mythos 5

护栏解除 · 仅限 Project Glasswing 可信伙伴 · 全球最强网络安全能力

Claude Fable 5

叠加分类器护栏 · 全球全量开放 · 命中即由 Opus 4.8 接管应答

分类器覆盖三类请求

网络安全 | 生物与化学 | 模型蒸馏

< 5% 的会话触发回退，其余 95%+ 会话中 Fable 5 等同 Mythos 5

1,000+ 小时外部悬赏红队未发现通用越狱；Mythos 级流量强制 30 天数据留存

数据来源：Anthropic《Claude Fable 5 and Claude Mythos 5》正文、客户证言与安全章节。

XAI · GROK 4.5

不比天花板，比每一秒和每一分钱

80 TPS

按官方口径达到「快模型」级别的吐字速度

4.2×

SWE-Bench Pro 单任务平均输出 15,954 token，约为 Opus 4.8 (max) 67,020 的四分之一

$2 / $6

三家之中最低的输入／输出定价

29.0%

SWE Marathon 解决率登顶，超过 Opus 4.8 的 26.0% 与 Fable 5 的 24.0%

- **与 Cursor 联合训练：直接长在开发者的工作流里**

  Grok 4.5 与 Cursor 共同训练，发布即在 Cursor 全部套餐可用，并成为 Grok Build CLI 的默认模型——一条绕过通用聊天入口、直插编码场景的分发路径。

- **训练侧：数万张 GB300 + 高度异步的强化学习**

  RL 覆盖数十万个任务，以多步软件工程为核心，配自动化与模型化评分。异步架构让智能体 rollout 可运行数小时而训练不停——目标函数明确写着「每 token 智能」。

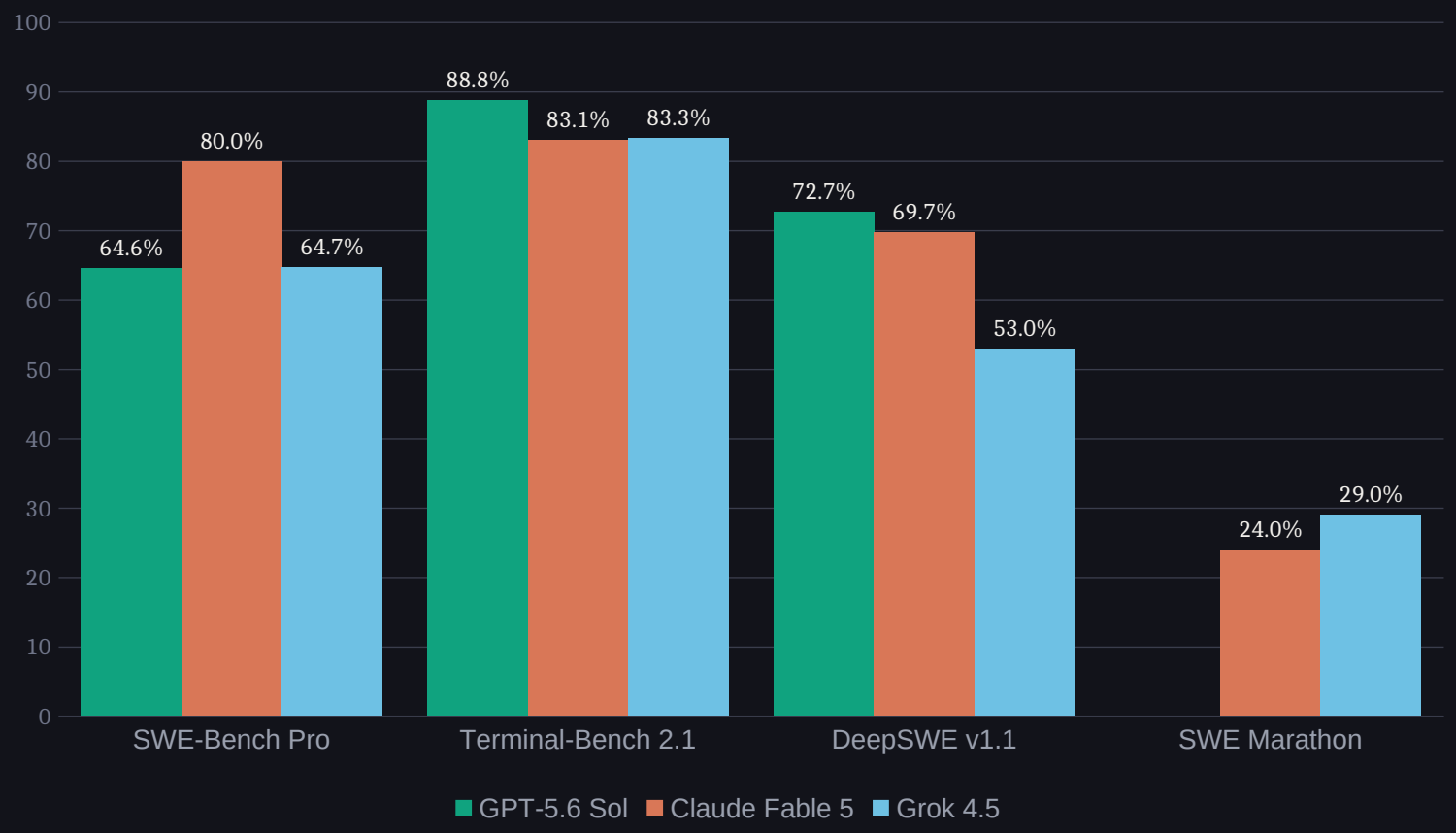
- **办公场景同样在抢： Word / PowerPoint / Excel 插件**

  官方演示中 Grok Build 能做联网调研 + 多表公式的 Excel 模型，并留下便签注释； PPT 里用原生形状搭建复杂图示。定位与 GPT-5.6 的知识工作叙事正面重叠。

数据来源： xAI 《Introducing Grok 4.5》正文与图表说明。 TPS = 每秒 token 数，为官方服务口径。

BENCHMARK

编码主战场：四张榜，三个赢家



Fable 5 赢下最硬的一张

SWE-Bench Pro 80.0%，比 GPT-5.6 Sol 高 15.4 分、比 Grok 4.5 高 15.3 分。这是四张编码榜中唯一拉开代差的一项。

Sol 赢下终端与长程工程

Terminal-Bench 2.1 88.8%（ultra 91.9%）、DeepSWE v1.1 72.7% 均为最高，且官方称输出 token 不到 Fable 的一半。

Grok 赢下唯一一张马拉松

SWE Marathon 29.0% 登顶，是它在所有公开榜单里唯一的第一；但同一份文档中 DeepSWE v1.1 只有 53%。

没有一家在编码上全面领先——每家都能在自己挑的榜上称王。

数据来源：GPT-5.6 与 Fable 5 数据取自 OpenAI 官方 benchmark 表；Grok 4.5 与 SWE Marathon 数据取自 xAI 官方图表。Sol 未参与 SWE Marathon。

DATA QUALITY

同一个模型、同一张榜，两套数字

榜单 / 对象	OpenAI 口径	xAI 口径	差异与含义
Terminal-Bench 2.1 （Fable 5）	83.1%	84.3%	+1.2 分。同一模型同一张榜，两家给出不同数字
SWE-Bench Pro （Fable 5）	80.0%	80.4%	+0.4 分。差距虽小，但说明 harness 与推理档位未对齐
DeepSWE （Grok 4.5）	—	v1.0： 62.0% / v1.1： 53.0%	同一文档内换一个版本掉 9.0 分——版本号比分数更重要
Agents' Last Exam （GPT-5.6 Sol）	正文 53.6 / 表格 52.7%	—	同一篇发布文档内部就存在两个口径
SWE-Bench Pro （GPT-5.6 Sol）	64.6%	未收录	xAI 发布晚 7 天，仍只对标 GPT-5.5 的 58.6%

为什么会这样

harness 不同、推理档位（max / xhigh / medium）不同、榜单版本不同、执行方不同。三家都在正文里注明「竞品数字取自对方公开的系统卡或排行榜」——也就是说，没有一组数字是在同一条件下跑出来的。

看发布文档时的三条自保规则

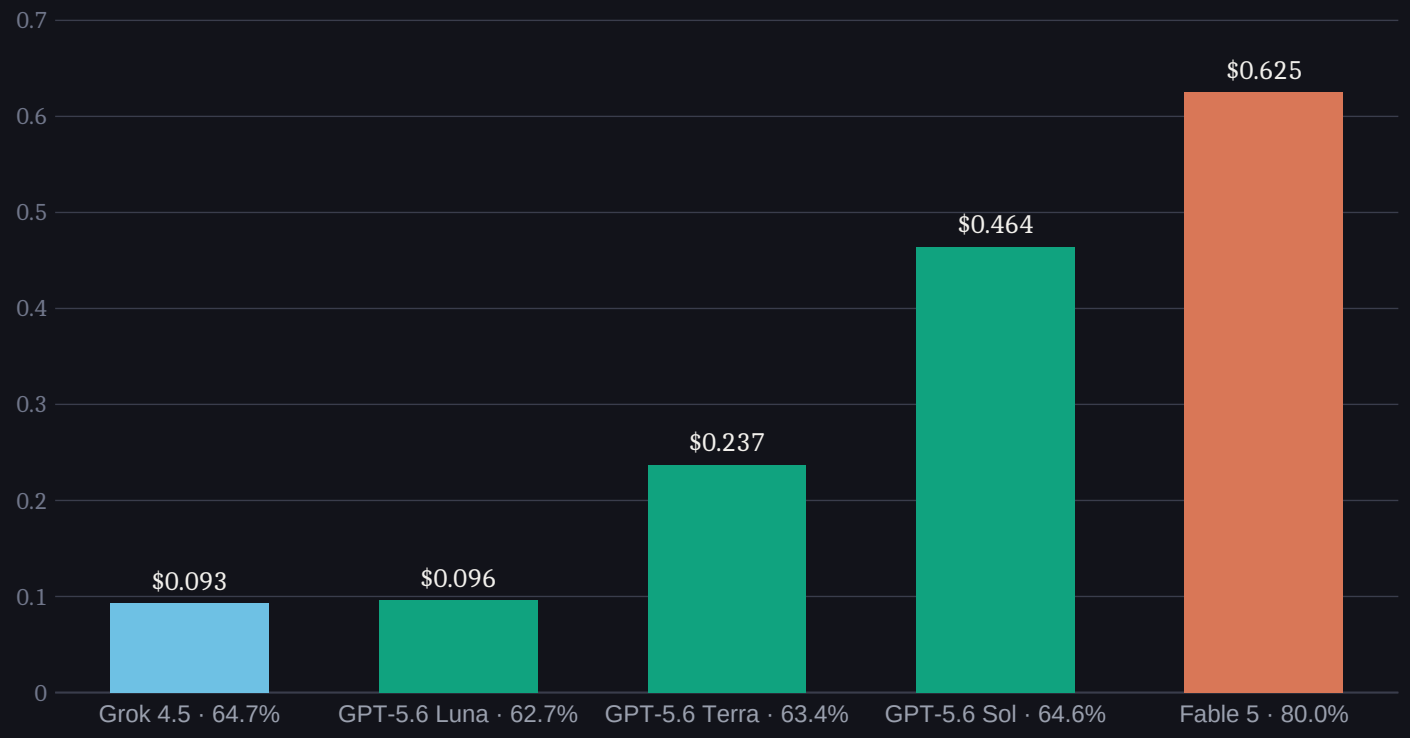
① 先看榜单版本号与推理档位，再看分数；
② 只在同一份文档内部做横向比较，跨文档只做量级判断；
③ 留意「谁没被列进表里」——缺席往往比分数更能说明问题。

所有对照数字均取自三家官方发布文档，未引入第三方复测。

INSIGHT

性价比：一分 SWE-Bench Pro 值多少钱

输出定价 ÷ SWE-Bench Pro 得分（美元 / 每 100 万 token / 每分）



6.7× 同一条曲线上的极差

Fable 5 每分单价 $0.625，是 Grok 4.5 $0.093 的 6.7 倍。能力天花板依然极贵。

+0.1 $6 档与 $30 档的分差

Grok 4.5 （$6） 64.7% 对 GPT-5.6 Sol （$30） 64.6%——多花 5 倍钱，在这张榜上几乎没有回报。

+15.4 花 8.3 倍钱买到的东西

从 $6 档到 Fable 5 的 $50 档，换来 SWE-Bench Pro 上 15.4 分的提升。值不值，取决于任务长度。

口径说明：该指标只用「输出单价 ÷ 单一 benchmark 得分」做粗略代理，未计入各模型实际消耗的 token 量与失败重试成本——若计入 Grok 4.5 自称的 4.2× token 效率，其优势会进一步放大；若计入 Fable 5 在长任务上的一次成功率，差距则会收窄。

得分来源： GPT-5.6 系列与 Fable 5 取自 OpenAI 官方表； Grok 4.5 取自 xAI 官方图表。定价取自三家官方定价章节。该比值为本文自建指标。

TREND

Token 效率：从「分数竞赛」到「单位成本智能」

三家在同一个月里，都把「更省」写进了发布文档的主标题层级——这在一年前还只是脚注。

O OpenAI

61%

在 AA 智能指数上， Sol 以少 1 分的成绩用掉少 61% 的时间与约一半成本

< 50%

AA 编码智能体指数登顶的同时，输出 token 不到 Fable 5 的一半

−85%

OSWorld 2.0 超过 Opus 4.8 ，输出 token 少 85%

−38%

Clio： 程序化工具调用让多步文档分析的提示 token 降 38% ，质量无损

A Anthropic

medium

Cognition FrontierCode： 即便在中等推理档位， Fable 5 仍是前沿模型最高分

1/3

客户证言：前沿物理研究中只用掉约三分之一的推理 token

25–30%

日常表格套件全档位胜过 Opus 4.8 ，轮次更少、整体耗时少 25–30%

3 ×

持久化记忆带来的增益是 Opus 4.8 的 3 倍——省 token 靠的是记住，不是重算

X xAI

4.2 ×

SWE-Bench Pro 单任务平均 15,954 输出 token ，约为 Opus 4.8 (max) 的 1/4

2 ×

官方称 token 效率约为同级领先模型的 2 倍，步数不到一半

80 TPS

把「快」直接写进卖点： faster than flash models

$0.096

按 15,954 token × $6/ 百万推算的单任务输出成本（本文推算）

结构性判断：当三家的绝对分差被压缩到个位数，竞争焦点就必然从「能不能做对」转向「做对一次要花多少钱」——这也解释了为什么 OpenAI 一次发三档、 xAI 直接把价格砍到五分之一。

数据来源：三家官方发布文档正文与客户证言。 $0.096 为本文按官方 token 数与定价推算，非官方口径。

SAFETY

安全：从合规成本，变成产品设计与护城河

A Anthropic

把安全等级做成 SKU

- 分类器命中即回退 Opus 4.8，用户会被告知，而不是直接拒答
- 触发率 < 5%，官方自陈「刻意调得保守，会误伤良性请求」
- 外部悬赏 1,000+ 小时未找到通用越狱；英国 AISI 取得部分进展
- Mythos 级流量强制 30 天数据留存，仅用于安全用途

O OpenAI

把安全做成分级访问权限

- 网络安全侧拦截量约为上代的 10 倍，并保留「换低能力模型重试」的出口
- 约 70 万 A100 等效 GPU 小时的黑盒自动化红队
- Trusted Access for Cyber：验证身份后才开放防御性能力
- 9 月 1 日起须启用硬件 passkey，才能保留最强模型的访问权

X xAI

发布文档里几乎不谈

- 《Introducing Grok 4.5》全文没有独立的安全章节
- 未披露危险能力评估、红队投入或拦截机制
- 安全相关内容仅以页脚 Safety / AUP 链接形式存在
- 这本身也是一种定位选择：把叙事全部押在性能与成本上

数据洞察

两家美国头部实验室在同一季度做了同一件事：把最强的网络安全与生物能力从公开产品中切出去，只对通过身份验证的机构开放（Project Glasswing / Trusted Access for Cyber）。安全审查正在从「发布前的一道关」变成「持续运营的一层产品」——也顺带成了新的准入壁垒。

数据来源：三家官方发布文档安全章节。 xAI 一项为对其该篇发布文档内容的缺省性观察，不代表该公司无相关工作。

NEW FRONT

第二战场：三家同时卷进了 Office

编码是共识战场，但这一轮真正的新增量，是 PPT 、 Excel 、 Word 这类「可交付物」。

GPT-5.6

主打「读得懂模板」

- Model ML： 在 20 个客户工作流、数百份 deck 上，每份比 Fable 少用 39% token ，返工更少
- Canva： 幻灯片生成强于竞品， token 效率约 1.6 倍
- 能推断出 Slide Master 里的排版规则并一致套用；GPT-5.5 会漏掉母版元素

Claude Fable 5

主打「分析深度」

- Hebbia 金融基准最高分，图表与表格解读、文档推理显著提升
- 首个在某分析基准上突破 90% 的模型，较 Opus 提升 10 分
- 日常表格套件全档位胜过 Opus 4.8 ，运行快 25–30%

Grok 4.5

主打「直接给插件」

- 已上架 Word / PowerPoint / Excel 官方插件
- Grok Build 可做联网调研 + 多表公式的 Excel 模型，并留下便签注释
- PPT 中使用原生形状搭建复杂图示，而不是贴图

为什么是现在

编码榜单已逼近饱和（ Terminal-Bench 2.1 上三家挤在 83–89% 区间），而办公交付物既没有公认榜单，又直接对应付费意愿最强的企业席位。

对使用者的含义

「能不能按我给的模板做出可直接交付的 deck 」正在取代「能不能答对题」，成为更有区分度的选型标准——而这项能力目前只能拿自己的真实文件去试。

数据来源：三家官方发布文档中的知识工作章节与客户证言。

TAKEAWAY

结论：三种钱，买三种不同的东西

A

选 Claude Fable 5

当任务足够长、足够贵，
错一次的代价很高

- 长程复杂任务上唯一拉开代差的模型（SWE-Bench Pro 80.0%）
- 视觉、记忆、科研发现三项有别家给不出的证据链
- 代价： $10 / $50 的最高定价，以及 < 5% 会话的护栏回退

O

选 GPT-5.6 家族

当你要规模化跑智能体，
且必须控住单位成本

- 三档 + max / ultra，能力可按任务分级采购，是最灵活的一家
- BrowseComp、 OSWorld 2.0、 Terminal-Bench 三项 SOTA，工具与计算机使用最强
- 模板化办公交付物目前证据最充分（Model ML −39% token、 Canva 1.6×）

X

选 Grok 4.5

当预算敏感、吞吐优先，
且任务以编码为主

- $2 / $6 的定价 + 4.2× token 效率，单位成本优势是结构性的
- SWE-Bench Pro 64.7%，与 $30 档的 GPT-5.6 Sol 基本持平
- 代价： DeepSWE v1.1 仅 53%，长程一致性是短板；安全披露最少

一句话：能力天花板还在 Anthropic 手里，但「够用」的价格已被打到原来的五分之一——2026 下半年真正的变量，不是谁的分更高，而是谁先把单位成本打穿。

以上建议基于三家官方发布文档的公开数据，未包含实际业务场景测试，选型前建议以自有任务复测。

APPENDIX

关键数据总表

榜单 / 指标	GPT-5.6 Sol	Claude Fable 5	Grok 4.5	数据来源
SWE-Bench Pro	64.6%	80.0%	64.7%	OpenAI 表 / xAI 图
Terminal-Bench 2.1	88.8%（ultra 91.9%）	83.1%	83.3%	OpenAI 表 / xAI 图
DeepSWE v1.1	72.7%	69.7%	53.0%	OpenAI 表 / xAI 图
SWE Marathon	未参与	24.0%	29.0%	xAI 图
AA 编码智能体指数 v1.1	80.0	77.2	未收录	OpenAI 表
AA 智能指数 v4.1	58.9	59.9	未收录	OpenAI 表
Agents' Last Exam	53.6（表 52.7%）	40.5%	未收录	OpenAI 正文 / 表
GDPval-AA v2（Elo）	1,747.8	1,759.6	未收录	OpenAI 表
BrowseComp	90.4%（ultra 92.2%）	Mythos 5： 88.0%	未收录	OpenAI 表
FrontierMath Tier 4	83.0%	87.8%	未收录	OpenAI 表
Toolathlon	58.0%	61.7%	未收录	OpenAI 表
定价（输入 / 输出，每 100 万 token）	**$5 / $30**	**$10 / $50**	**$2 / $6**	**三家定价章节**

说明：不同来源的评测 harness、推理档位与榜单版本并不一致，跨列数字仅供量级参考，不构成同条件对比。资料截至 2026-07-25。