2026 年中 · 前沿大模型三国杀

GPT-5.6 vs Claude Fable 5 vs Grok 4.5

三家官方发布文档深度对比：能力、效率、安全与战略意图

OpenAI · 2026.7.9 发布

GPT-5.6 · Sol / Terra / Luna

Anthropic · 2026.6.9 发布

Claude Fable 5 / Mythos 5

xAI · 2026.7.8 发布

Grok 4.5

基于 openai.com 、 anthropic.com 、 x.ai 三篇官方发布文档撰写 | 数据截至 2026 年 7 月 17 日

OVERVIEW

三份发布文档，三种产品哲学

维度	GPT-5.6 （ OpenAI ）	Claude Fable 5 （ Anthropic ）	Grok 4.5 （ xAI ）
发布日期	2026.7.9 （ GA ）	2026.6.9 （ 7.1 重新部署）	2026.7.8
模型结构	三档家族： Sol / Terra / Luna	一体两面： Fable 5 = Mythos 5 + 安全闸	单一旗舰，与 Cursor 联合训练
官方定位	每个 token 更多智能，每美元更强性能	Mythos 级能力首次安全地面向通用开放	编码、智能体与知识工作的实用主义
API 价格 (每百万 token)	Sol $5/$30 · Terra $2.5/$15 · Luna $1/$6	$10 / $50 （输入 / 输出）	$2 / $6 （输入 / 输出）
杀手锏	ultra 多智能体并行 + 程序化工具调用	安全分类器回退架构 + 科学发现能力	80 TPS 速度 +4.2× token 效率
可用渠道	ChatGPT / Codex / API / M365 Copilot	Claude 全系 / API （订阅限量）	Grok Build / Cursor / API （ EU 暂缺）

资料来源： OpenAI / Anthropic / xAI 官方发布文档（ 2026.6–7 ）

NARRATIVE

核心叙事：三家各自想让你记住什么

OpenAI

“Efficient by default, maximum performance on demand”

效率即护城河

- **性价比叙事：**中等推理档即以约 1/4 成本打平 Fable 5； Terra/Luna 以约 1/16 成本超越 Fable 5
- **家族化覆盖：** Sol/Terra/Luna 三档通吃 $1–$30 价格带，智能“按需伸缩”
- **上限可解锁：** ultra 模式默认 4 智能体并行（可至 16），用 token 换时间与上限
- **自我加速：**内部研究编码推理算力半年增长 100×，RSI 指数较上代 +16.2

Anthropic

“A Mythos-class model made safe for general use”

安全即发布前提

- **能力叙事：**发布时“几乎所有基准 SOTA”，任务越长越复杂，领先越大
- **安全架构：**分类器命中即回退 Opus 4.8 （非拒绝），>95% 会话无感知
- **双轨发布：** Fable 5 面向公众， Mythos 5 解除部分安全闸、仅政府与可信伙伴
- **科学背书：**蛋白质设计提速 10×、分子生物学假设 80% 被科学家优选

xAI

“Highest intelligence per unit of time and cost”

速度与成本即武器

- **速度叙事：** 80 TPS，“比 flash 级模型还快”的旗舰智能
- **效率叙事：** SWE-Bench Pro 平均 15,954 输出 token，比 Opus 4.8 max 少 4.2×
- **价格颠覆：** $2/$6 定价，旗舰性能打到对手 1/4–1/8 价格
- **工程落地：**与 Cursor 联合训练； Office 三件套原生插件直接进工作流

资料来源： OpenAI / Anthropic / xAI 官方发布文档（2026.6–7）

GPT-5.6：把“每美元智能”做成家族化生意

特点与核心优势

- **三档家族：** Sol 旗舰 / Terra 均衡 / Luna 极致性价比，能力档可独立演进
- **ultra 多智能体：**默认 4 个智能体并行（可至 16），BrowseComp 达 92.2% 新纪录
- **程序化工具调用：**模型自写轻量程序协调工具，Unity 场景构建 token 省 63.5%
- **设计判断力：**生成后可自查渲染结果并打磨，前端 QA 评分 4.4/5 （GPT-5.5 为 4.0）
- **网络安全最强：** ExploitBench 73.5% （上代 47.9%）； CTF 96.7%；配可信访问计划
- **自我改进叙事：** RSI 指数 57.9% （+16.2），研究员日均输出 token 翻倍以上

一句话总结： OpenAI 不再只卖“最强模型”，而是卖“智能的阶梯定价”——默认高效，上限付费解锁。

关键数据

53.6
Agents' Last Exam 新纪录
领先 Fable 5 达 13.1 分

80
AA 编码智能体指数 SOTA
token 不到 Fable 5 一半

62.6%
OSWorld 2.0 计算机使用
输出 token 比 Opus 4.8 少 85%

88.8%
Terminal-Bench 2.1
ultra 模式下达 91.9%

+16.2
RSI 自我改进指数
内部研究算力半年 100×

$1/$6
Luna 每百万 token 定价
性能超 Fable 5，成本约 1/16

资料来源： OpenAI / Anthropic / xAI 官方发布文档（2026.6–7）

Claude Fable 5： 最强能力，第一次“安全地”交到你手上

特点与核心优势

- **Mythos 级首秀：** Opus 之上的新能力层级，发布时几乎所有基准 SOTA
- **回退式安全：**分类器覆盖网络 / 生化 / 蒸馏，命中即无缝回退 Opus 4.8
- **超长自主性：** Stripe 5000 万行 Ruby 代码库整体迁移，一天完成（原需团队两月 + ）
- **视觉 SOTA：**纯视觉、无辅助 harness 通关《宝可梦 火红》；可从截图重建 Web 应用
- **持久记忆：**《杀戮尖塔》中借助文件记忆，抵达终章频率是 Opus 4.8 的 3 倍
- **科学发现：**蛋白设计提速约 10×；基因组学自主研究一周，小 100× 的模型超越《 Science 》已发表模型

关键数据

80.0%
SWE-Bench Pro SOTA
三家对比中的最高分

59.9
AA 智能指数第一
GPT-5.6 Sol 以 58.9 紧随

87.8%
FrontierMath Tier 4
最难数学推理基准第一

1760
GDPval-AA 知识工作 Elo
高于 GPT-5.6 Sol 的 1748

<5%
会话触发安全回退比例
>95% 会话性能等同 Mythos 5

$10/$50
每百万 token 定价
不到 Mythos Preview 一半

一句话总结： Anthropic 把安全从“合规成本”做成了“发布架构”——能力触顶后，治理方式本身就是产品。

资料来源： OpenAI / Anthropic / xAI 官方发布文档（ 2026.6–7 ）

Grok 4.5： 用速度和价格，打一场效率闪电战

特点与核心优势

- **联合训练：**与 Cursor 共同训练，直接面向真实编码工作流优化
- **大规模 RL：**数万块 NVIDIA GB300；数十万个多步软件工程任务的异步强化学习
- **极致速度：** 80 TPS 输出，官方称“比 flash 级模型更快”
- **token 效率：** SWE-Bench Pro 平均 15,954 输出 token，比 Opus 4.8 max 少 4.2×
- **办公落地：** Grok Build 默认模型； Excel/PPT/Word 原生插件，一条提示词端到端建站
- **长程强项：** SWE Marathon 29.0% 排名第一，领先 Opus 4.8 （26%）与 Fable （24%）

关键数据

80 TPS
输出速度
“faster than flash models”

4.2×
token 效率优势
同任务 token 仅为对手约 1/4

$2/$6
每百万 token 定价
旗舰中最低，输出价为 Fable 1/8

29.0%
SWE Marathon 第一
长程工程任务解决率

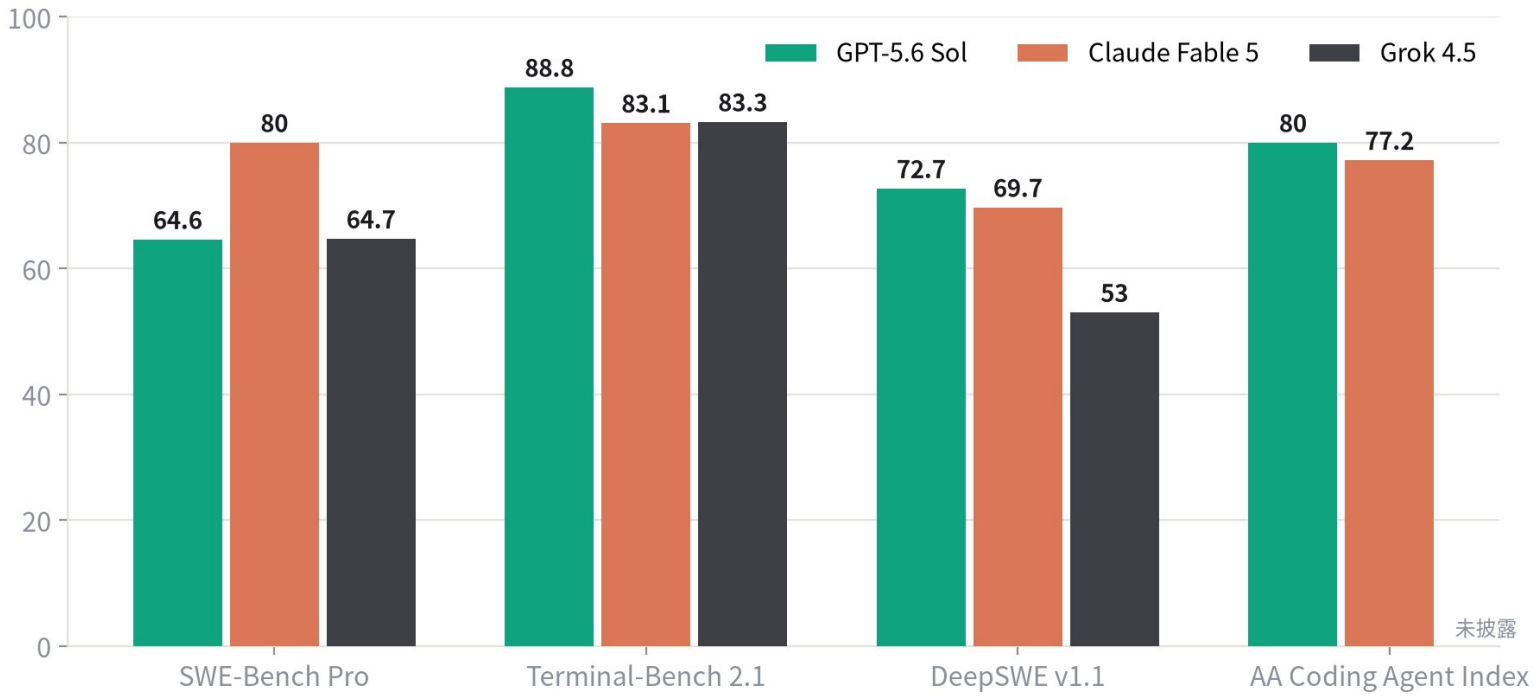
83.3%
Terminal-Bench 2.1
与 Fable 84.3% 基本打平

62.0%
DeepSWE 1.0
仅次于 Fable max 与 GPT-5.5

一句话总结： xAI 不卷“绝对最强”，卷“单位时间与单位成本的智能”——用价格把旗舰能力商品化。

资料来源： OpenAI / Anthropic / xAI 官方发布文档（2026.6–7）

编码： GPT-5.6 赢下“工程综合”， Fable 守住“难题上限”



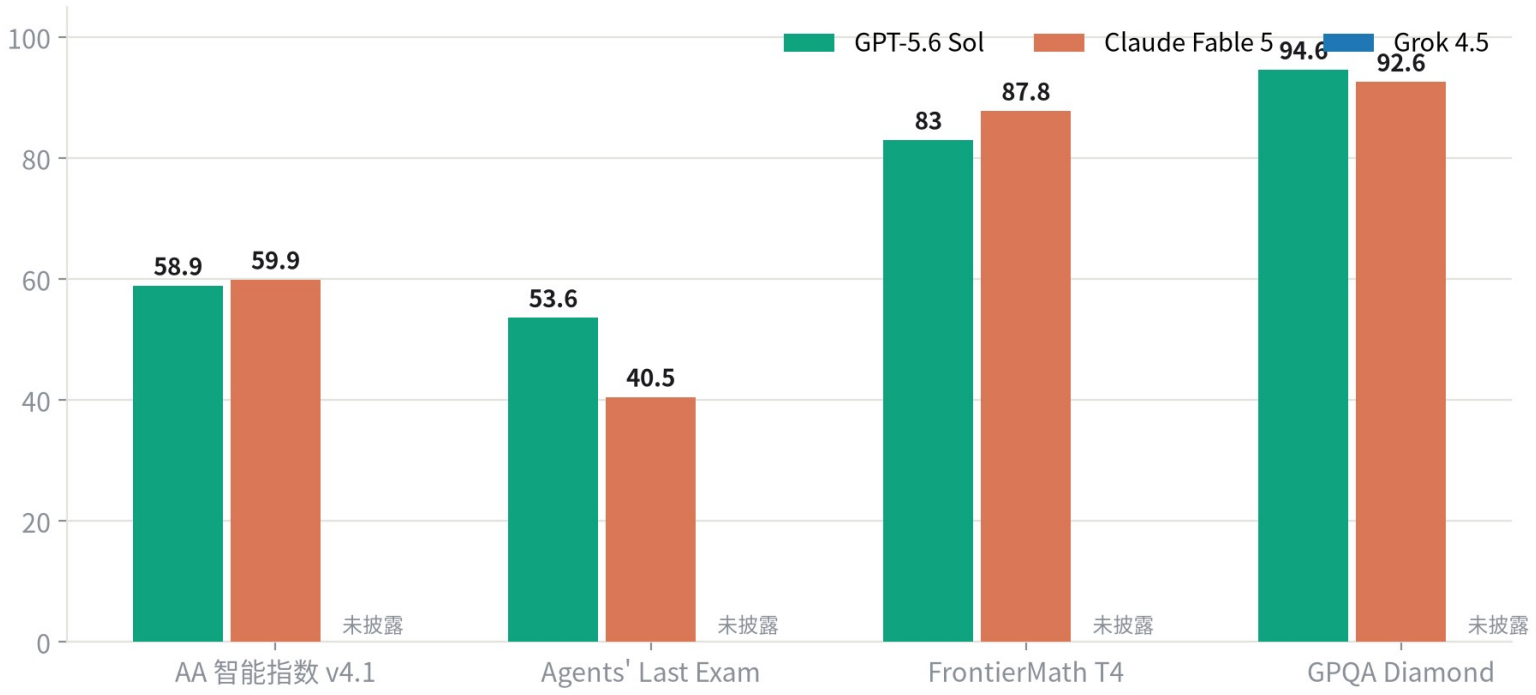
数据口径：OpenAI/Anthropic 数值取自 OpenAI GPT-5.6 发布文档基准表；Grok 4.5 取自 xAI 发布文档（自测 harness，口径略有差异）

怎么读这张图

- **GPT-5.6 Sol** 四项中三项第一，编码智能体综合指数 80 创 SOTA，且 token/ 时间约为对手一半
- **Fable 5** SWE-Bench Pro 80% 一骑绝尘（领先 15 分 +），难题上限仍是三家最高
- **Grok 4.5** Terminal-Bench 与 Fable 打平，但 DeepSWE 长程工程明显掉队（53%）
- **注意口径：** Grok 数据来自 xAI 自测，与 OpenAI 表内口径存在差异

资料来源： OpenAI / Anthropic / xAI 官方发布文档（2026.6–7）

综合智能： Fable 5 险胜指数榜， GPT-5.6 碾压长程工作流



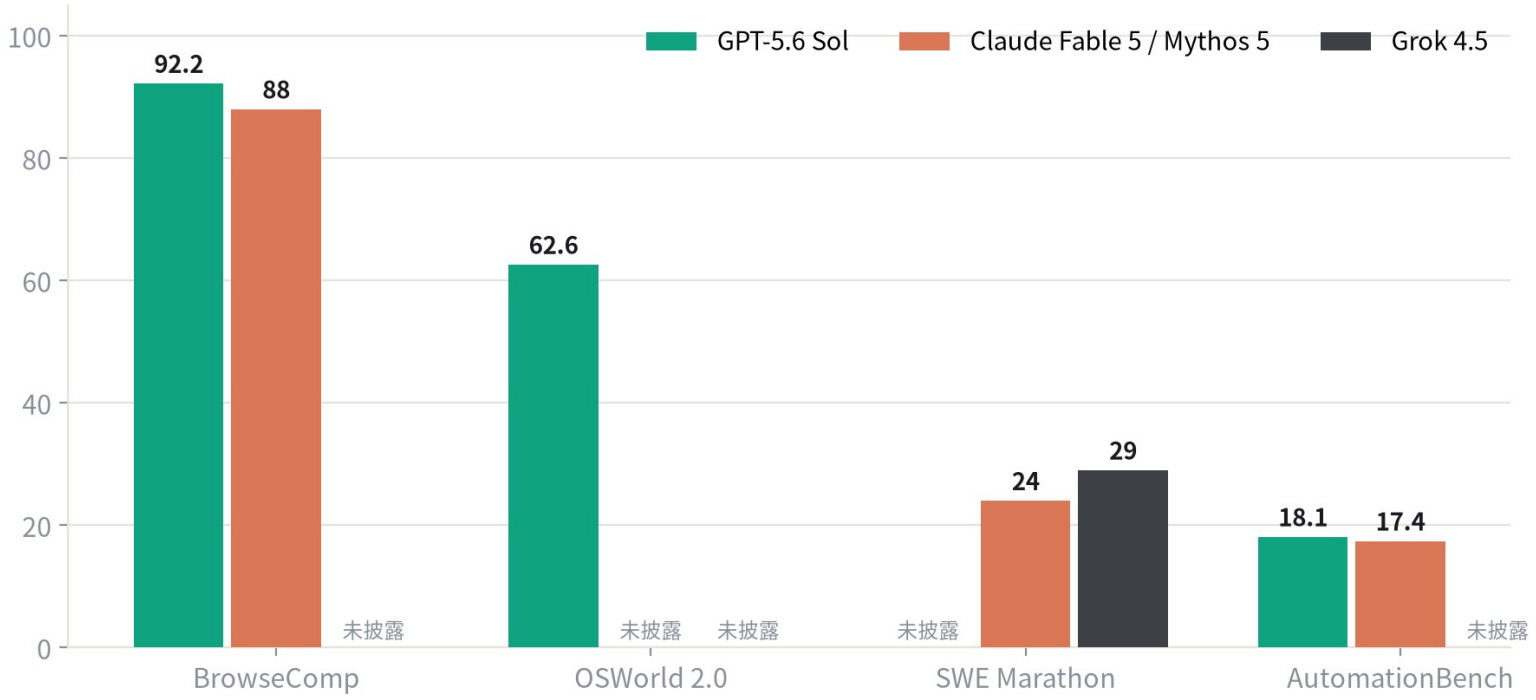
GPT-5.6 Sol 为 max 推理档；Agents' Last Exam 中 Sol 以 53.6 创纪录，领先 Fable 5 达 13.1 分；Grok 4.5 未披露此类基准

怎么读这张图

- **AA 智能指数：** Fable 5 以 59.9 对 58.9 险胜，但 GPT-5.6 用时少 61%、成本约一半
- **Agents' Last Exam：** 55 个职业领域长程工作流， Sol 53.6 创纪录，领先 13.1 分
- **FrontierMath T4：** 最难数学档 Fable 87.8% 仍居首，基础研究推理更硬
- **Grok 4.5** 未披露此类学术基准，叙事重心明显在工程实用

资料来源： OpenAI / Anthropic / xAI 官方发布文档（2026.6–7）

智能体与计算机使用：并行智能体成为新的“作弊码”



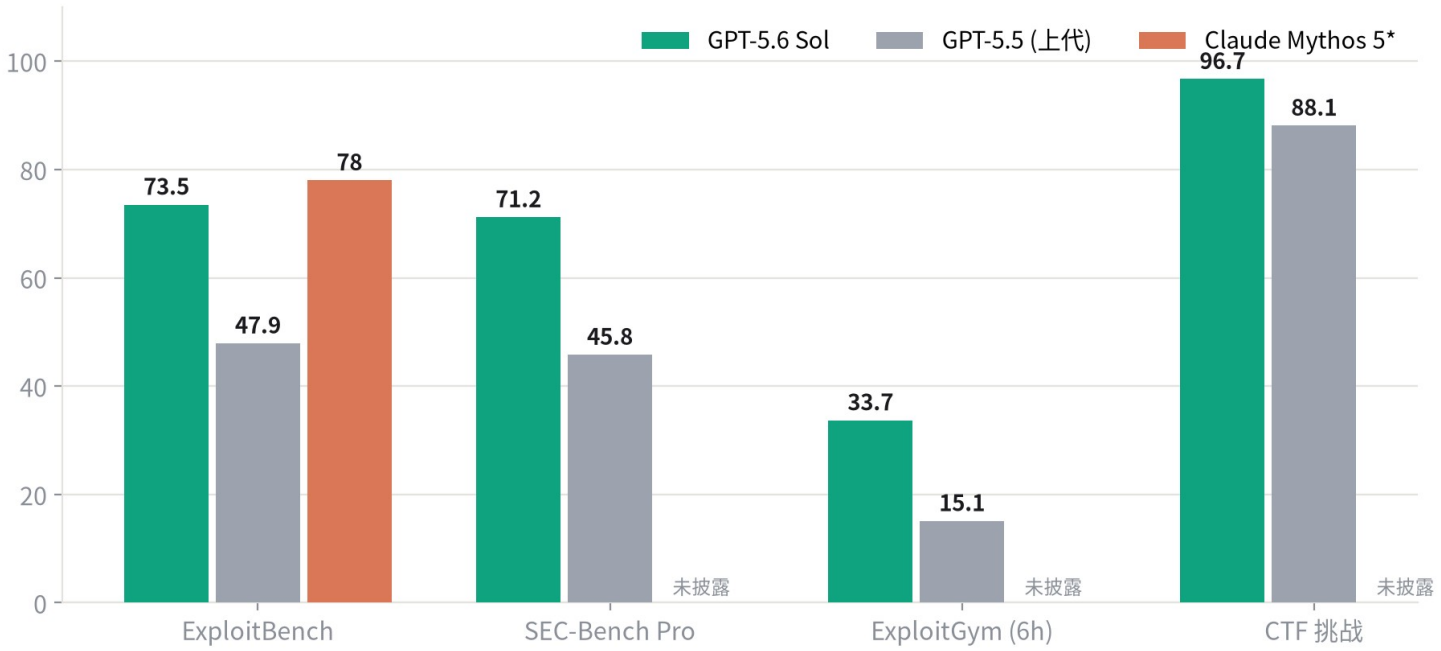
BrowseComp 为 GPT-5.6 Sol ultra 多智能体成绩；Anthropic 侧 BrowseComp 取 Mythos 5；SWE Marathon 来自 xAI 文档（Grok 4.5 第一）

怎么读这张图

- **BrowseComp：** ultra 4 智能体并行把 GPT-5.6 推到 92.2%，多智能体全面上移分数 - 延迟前沿
- **OSWorld 2.0：** Sol 62.6% 新 SOTA，输出 token 比 Opus 4.8 少 85%
- **SWE Marathon：** Grok 4.5 以 29% 拿下第一——马拉松式长任务是它的甜区
- **趋势判断：** 单模型分数见顶后，“编排多个智能体”成为继续 scaling 的方向

资料来源： OpenAI / Anthropic / xAI 官方发布文档（2026.6–7）

前沿能力：网络安全与科学，两种截然不同的开放姿态



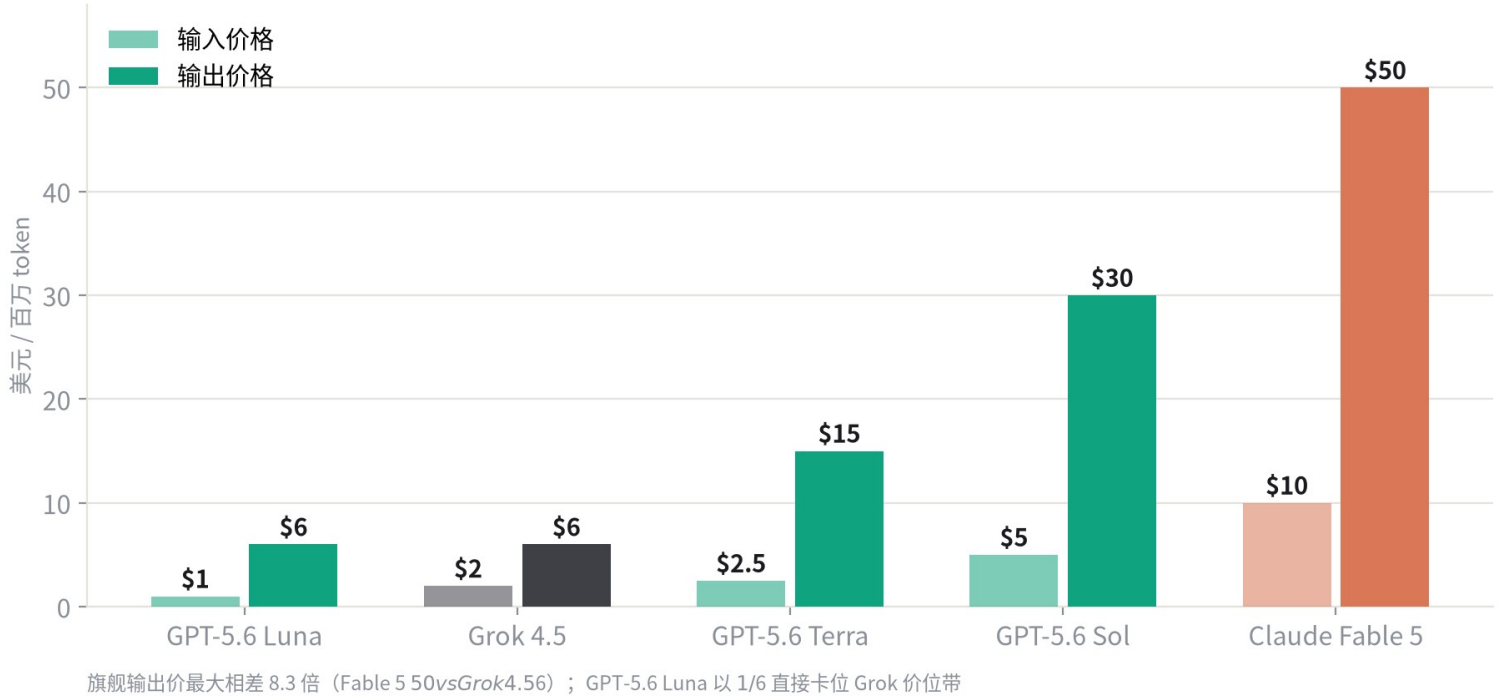
*Mythos 5 为解除部分安全闸的版本，仅 Glasswing 可信伙伴可用；Fable 5 触发安全分类器时回退至 Opus 4.8

关键观察

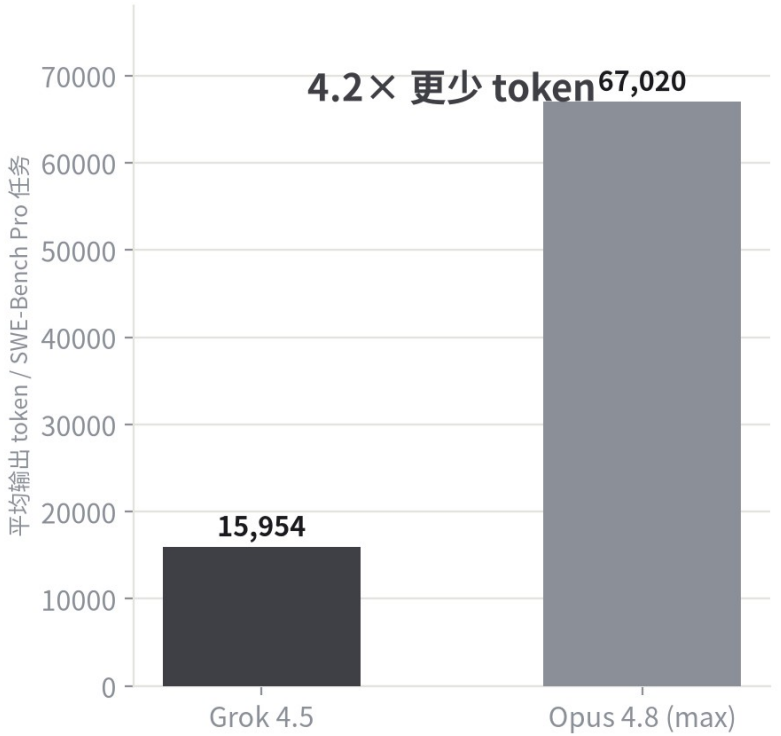
- **GPT-5.6 大幅跃进：** ExploitBench 73.5% vs 上代 47.9%；ExploitGym 6 小时达 33.7%（近翻倍）
- **Mythos 5 仍是天花板：**解除安全闸后 ExploitBench 78%，但仅 Glasswing 可信伙伴可用
- **OpenAI 的姿态：**防御优先 + 可信访问（Daybreak 计划、硬件 passkey 验证），称拦截量约为上代 10×
- **Anthropic 的姿态：** Fable 5 对网络 / 生化查询直接回退 Opus 4.8； OpenAI 在文档中点名其“拒答多数高等生物题”
- **科学能力对照：** GPT-5.6 主动公布 GeneBench Pro 28.7%；Anthropic 则展示蛋白设计 10× 提速等 Mythos 5 成果

资料来源： OpenAI / Anthropic / xAI 官方发布文档（2026.6–7）

价格与效率：旗舰智能正在经历“通缩”



旗舰输出价最大相差 8.3 倍（Fable 5 50*vs*Grok 4.56）；GPT-5.6 Luna 以 1/6 直接卡位 Grok 价位带



数据洞察：三家通稿不约而同把“token 效率”放在标题级位置——OpenAI 称 Terra/Luna 以约 1/16 成本超越 Fable 5；xAI 称同任务 token 仅为对手 1/4、步骤不到一半；Anthropic 也强调 Fable 5 比历代 Claude 更省 token。竞争单位已从“跑分”变成“每美元智能”。

资料来源：OpenAI / Anthropic / xAI 官方发布文档（2026.6–7）

SAFETY

安全策略：同一代模型，三种治理哲学

Anthropic · 预防式封印

- **架构：**独立分类器覆盖网络、生化、蒸馏三类风险，命中即回退 Opus 4.8 （非生硬拒绝）
- **验证：** 1000+ 小时外部赏金测试无通用越狱； 30 种公开越狱手法零违规
- **治理：** Mythos 级流量强制 30 天留存； Mythos 5 仅政府 / 可信伙伴；生物可信计划筹备中
- **代价：**误伤存在（ <5% 会话），官方承诺持续收窄

OpenAI · 分层防护 + 可信升级

- **架构：**训练内防护 + 实时推理监控 + 账户级执行，多层冗余
- **验证：**发布前最大规模红队（人工 + 大规模自动化黑盒测试）
- **治理：**网络能力拦截量约为上代 10×； Daybreak 可信访问向防御者开放高阶能力
- **姿态：**明确反对过度封锁——“过度封锁本身也是安全风险”

xAI · 能力优先，安全留白

- **文档事实：**整篇发布文档无安全章节，未提及红队、分类器或滥用防护
- **叙事重心：**全部篇幅给能力、速度、价格与生态
- **对照：**在另外两家把安全写成半篇通稿的时代，沉默本身也是一种信号
- **备注：**本页仅陈述三份文档的文本事实，不构成安全水平评判

资料来源： OpenAI / Anthropic / xAI 官方发布文档（ 2026.6–7 ）

发布节奏与生态卡位

一个月内的三连发

OpenAI

- **全渠道同日 GA：** ChatGPT、Codex、API 24 小时内全球铺开
- **巨头背书：** Microsoft 365 Copilot 首选模型；Figma/Canva/Notion/Shopify 等 20+ 家证言
- **缓存新政：**显式缓存断点 +30 分钟最短缓存期，读缓存 1 折

Anthropic

- **API 全量，订阅限量：** Pro/Max/Team 免费用至 6.22，之后转用量积分
- **政府合作：** Project Glasswing 与美国政府合作，Mythos 5 定向开放
- **波折：**发布 3 天后暂停访问，7.1 才完成重新部署——需求超预期

xAI

- **工具链绑定：** Grok Build 默认模型 + Cursor 全计划可用 + 限时免费
- **Office 插件：** Excel/PPT/Word 原生插件上架 Microsoft 市场
- **区域缺口：** EU 暂不可用（预计 7 月中），合规节奏落后

资料来源：OpenAI / Anthropic / xAI 官方发布文档（2026.6–7）

KEY INSIGHTS

五个关键数据洞察

01

没有全能冠军：基准领导权彻底碎片化

Fable 5 守住推理 / 数学 / 知识工作上限（SWE-Bench Pro 80%、FrontierMath T4 87.8%、GDPval 1760）；GPT-5.6 横扫智能体工程、终端、浏览、计算机使用与网络安全；Grok 4.5 赢下马拉松长任务与速度成本。选模型 = 选场景。

02

“每美元智能”取代“跑分”成为主叙事

OpenAI：1/4 成本打平 Fable 5；xAI：4.2× token 效率、步骤不到一半；Anthropic：Fable 比历代更省 token。三家通稿的标题级卖点都是效率——前沿智能进入通缩周期。

03

价格剪刀差高达 8.3×，家族化卡位全价格带

旗舰输出价：Fable $50 / Sol $30 / Grok $6。OpenAI 用 Luna（$1/$6）直接插到 Grok 价位带下方，用 Terra 守中端——价格带本身成为竞争武器。

04

安全从合规项升级为产品架构

Anthropic 把安全做成“回退而非拒绝”的无缝架构（>95% 会话无感知）；OpenAI 做成“分层防护 + 可信升级”；xAI 只字未提。生物能力成为首个被主流厂商“主动封印”的前沿能力。

05

AI 加速 AI：自我改进成为下一代竞争指标

OpenAI 首次披露 RSI 指数（+16.2）与内部数据：研究编码推理算力半年增长 100×、智能体 token 用量增 22×。模型厂商的下一个战场是“用模型训练 / 改进模型”的飞轮速度。

资料来源：OpenAI / Anthropic / xAI 官方发布文档（2026.6–7）

SELECTION GUIDE

怎么选：按场景对号入座

应用场景	首选	关键依据
复杂编码智能体 / 终端工作流	**GPT-5.6 Sol**	AA 编码指数 80 SOTA； Terminal-Bench 88.8%； token 与耗时约为对手一半
高难度数学 / 金融分析 / 研究写作	**Claude Fable 5**	FrontierMath T4 87.8%； GDPval 1760 Elo 最高； Hebbia 金融基准第一
大规模、成本敏感的生产部署	**Grok 4.5 / GPT-5.6 Luna**	$2/$6 与 $1/$6 定价； 80 TPS； Luna 性能仍超 Fable 5
长程无人值守工程任务	**Grok 4.5 / Fable 5**	SWE Marathon 29% 第一； Fable 可连续自主工作数日（Stripe 案例）
网络安全防御（授权环境）	**GPT-5.6 / Mythos 5**	Daybreak 可信访问； Mythos 5 经 Glasswing 开放， ExploitBench 78%
生物 / 化学研究	**注意限制**	Fable 5 会回退至 Opus 4.8； Mythos 5 生物可信计划 / GPT-5.6 GeneBench 28.7%
Office 办公自动化	**Grok 4.5 / GPT-5.6**	Grok 原生 Office 插件； GPT-5.6 为 M365 Copilot 首选模型

注：以上建议仅基于三家官方文档披露的数据与定位；实际选型建议结合自有业务场景做小规模基准复测。

资料来源： OpenAI / Anthropic / xAI 官方发布文档（2026.6–7）

旗舰智能的“通缩时代”已经到来

能力见顶的迹象没有出现，但竞争的单位已经改变：从跑分，到每美元智能，到治理架构，再到自我改进的飞轮速度。
三家没有谁在“造更强的模型”上输，只是各自押注了不同的赢法。

资料来源

OpenAI · GPT-5.6: Frontier intelligence that scales with your ambition （openai.com/index/gpt-5-6， 2026.7.9）

Anthropic · Claude Fable 5 and Claude Mythos 5 （anthropic.com/news/claude-fable-5-mythos-5， 2026.6.9）

xAI · Introducing Grok 4.5 （x.ai/news/grok-4-5， 2026.7.8）

说明：各基准分数均来自官方发布文档，各家测试口径（harness、推理档位、是否多智能体）存在差异，跨文档对比已尽量标注；本报告不构成投资或采购建议。