FRONTIER MODEL LANDSCAPE

GPT-5.6 × Claude Fable 5 × Grok 4.5

不是“谁赢了”的单一榜单，而是三种不同的前沿模型产品哲学

GPT-5.6

单位成本的完成度

Fable 5

长时自主与前沿科研

Grok 4.5

速度与经济性

基于三家官方发布文档 | 数据口径审慎对齐 | 2026.07

一页结论：三家争夺的不是同一个“第一”

真正的差异，来自厂商选择优化的目标函数

OpenAI	**把“高能力”包装成可伸缩的工作系统**	Sol / Terra / Luna + max / ultra；重点是性能、成本、时间三者的前沿移动
Anthropic	**把“更长自主时间”与科研突破置于中心**	Fable 面向通用用户，Mythos 面向可信访问；安全边界本身就是产品架构
SpaceXAI	**用接近前沿的质量换取极强速度与价格优势**	80 TPS、$2/$6、4.2×更少输出 token；强调尽快把智能嵌入代码与 Office

决策建议： 高风险、复杂交付看 GPT-5.6；长周期工程/科研看 Fable 5；高吞吐、成本敏感 Agent 看 Grok 4.5。

来源：OpenAI、Anthropic、SpaceXAI 官方发布；结论为基于公开材料的分析。

截至 2026-07-10

比较方法：先分证据，再谈结论

避免将不同 harness、推理预算和安全模式下的数字硬拼

A

可直接对齐

价格、上下文、发布时间、官方可用性

证据强度：高

B

同表横向比较

同一发布方表格中的竞品数字；仍需留意运行环境

证据强度：中高

C

厂商独有案例

客户案例、内部评测、科研故事；用于理解定位，不用于总排名

证据强度：中

三个关键限制

- Fable 与 Mythos 是同一底层模型的不同安全访问层
- GPT-5.6 “Ultra”是多 Agent 配置，不能与单模型调用等价
- Grok 官方重点展示编码/效率，跨领域基准披露少于 OpenAI

方法说明：优先使用用户提供的三篇官方文档；xAI API 规格用于补充上下文窗口。

三种产品哲学：完成度、自治深度、速度经济性

官方标题与发布叙事，透露了每家公司最希望市场记住什么

GPT-5.6

“More intelligence from every token”

厂商核心主张

把智能转化为可交付成品；设计、办公、工具、知识工作形成闭环

系统化

Claude Fable 5

“The longer and more complex, the larger the lead”

厂商核心主张

用长时自治、记忆与科研结果证明模型跨越 Opus 能力层级

能力跃迁

Grok 4.5

“Highest intelligence per unit of time and cost”

厂商核心主张

把 80 TPS、低单价、低 token 消耗变成部署竞争力

效率化

引文均来自三家官方发布页；中文为意译。

规格与商业门槛：Grok 最便宜，Fable 最昂贵

价格只是单 token 单价；真实成本还受推理档位、token 使用量与工具循环影响

维度	GPT-5.6 Sol	Claude Fable 5	Grok 4.5
输入 / 1M tokens	$5	$10	$2
输出 / 1M tokens	$30	$50	$6
相对输出价	5× Grok	8.3× Grok	基准 1×
上下文	官方发布页未明示	“millions of tokens”任务聚焦	500K
推理控制	medium / high / max / ultra	effort；Fable / Mythos访问层	low / medium / high
核心工具形态	Programmatic Tool Calling、多 Agent	长时自治、文件记忆、Claude Code	函数、Web/X 搜索、代码执行
可用性	ChatGPT / Codex / API	Fable 通用；Mythos 可信访问	API / Grok Build / Office 插件

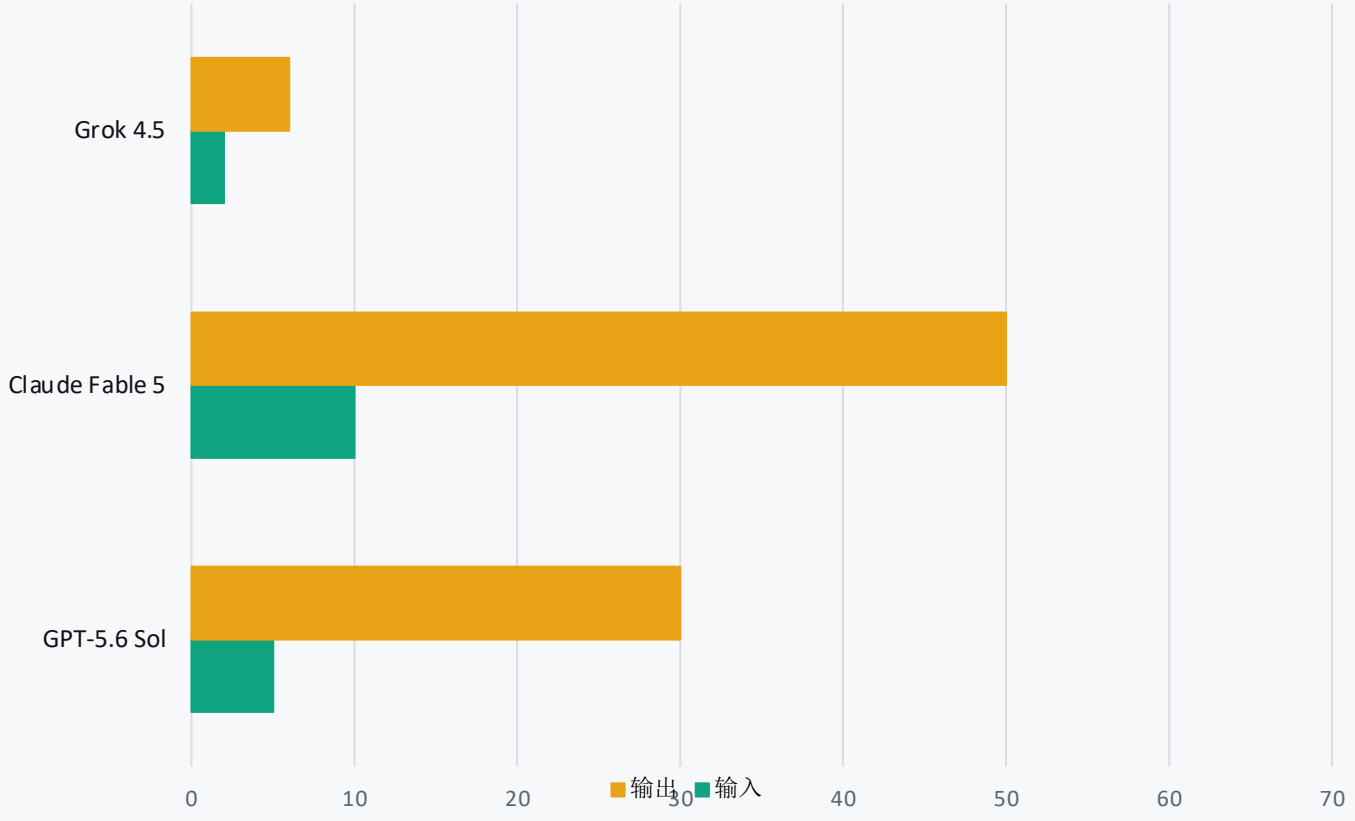
成本洞察：仅按输入+输出各 1M token，Sol ≈ $35、Fable ≈ $60、Grok ≈ $8；但任务级成本差距可能因 token 效率进一步扩大。

来源：三家官方发布页；Grok 500K 来自 xAI 官方 API/模型页。

截至 2026-07-10

同等 token 量下，价格差距非常大

以输入 1M + 输出 1M token 的“混合篮子”作示意，不代表真实任务成本



混合篮子成本

$35 / $60 / $8

Sol / Fable / Grok

Grok 相对 Fable

约 87% 更低

同等输入与输出 token 数

Sol 相对 Fable

约 42% 更低

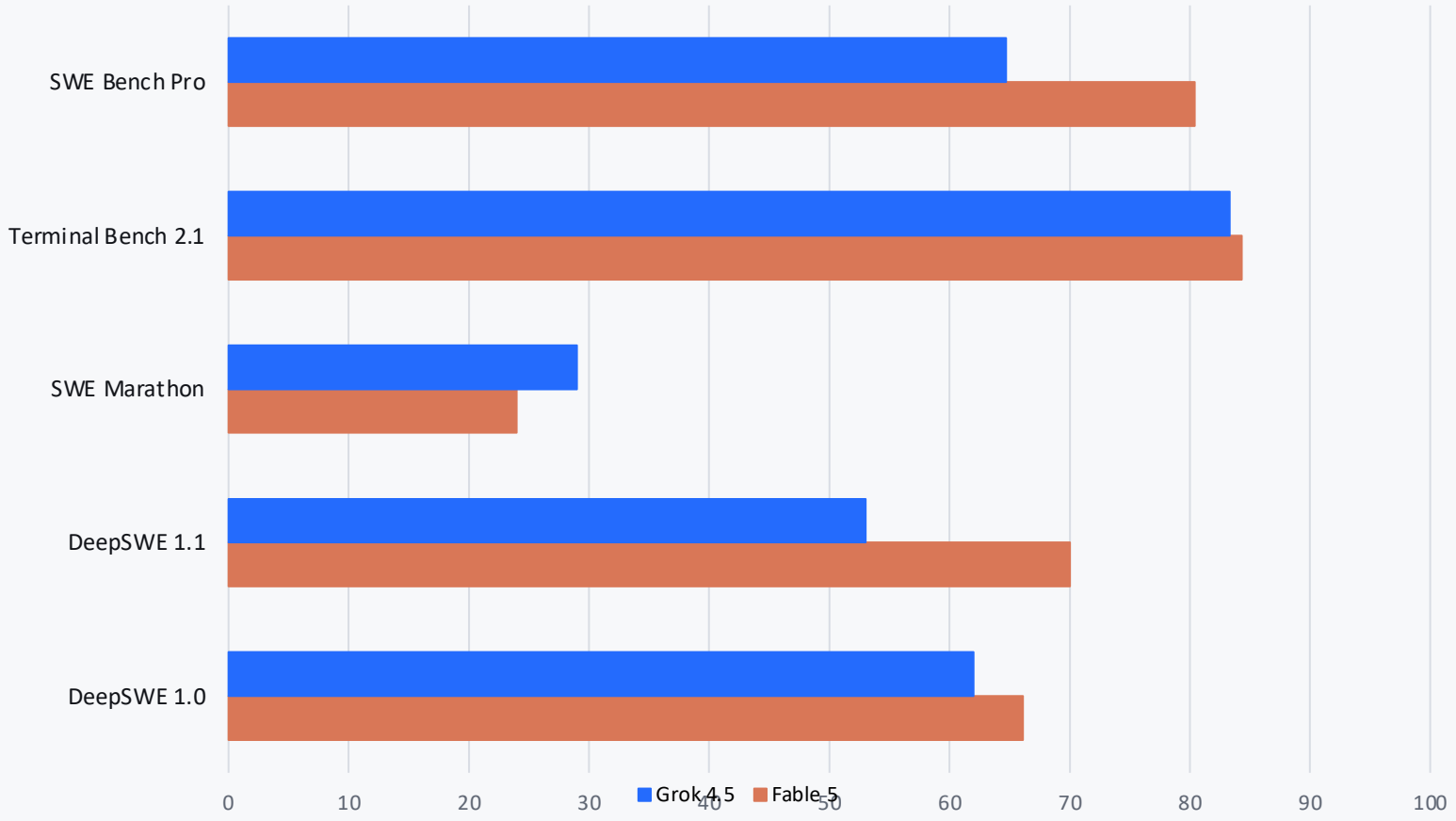
同等输入与输出 token 数

计算：官方每 1M token 标价简单相加；未计缓存、工具调用、推理隐藏成本、批处理和优先通道。

截至 2026-07-10

编码：Fable 胜在深度，Grok 在长程任务上抢到一项第一

同一组 xAI 发布页数据中，Fable 领先 4/5 项；Grok 在 SWE Marathon 领先



读数

- Fable 在 DeepSWE 1.1 领先 17 个百分点
- Terminal Bench 仅差 1.0 点，Grok 基本追平
- SWE Marathon：Grok 29%，领先 Fable 5 个点
- 说明“编码强”至少包含短题、真实仓库、终端和长程任务四种能力

不能用单项替代总能力

来源：SpaceXAI《Introducing Grok 4.5》；竞品数字由 xAI 引用各厂商系统卡或排行榜。

截至 2026-07-10

换一个官方表格：GPT-5.6 更像“全栈工作模型”

OpenAI 同表数据显示：Sol 在 Agent、终端、浏览、OS 操作上强；Fable 在部分知识/工具/学术项占优

基准	GPT-5.6 Sol	Fable 5	领先者	差值
Agents' Last Exam	52.7%	40.5%	**GPT-5.6**	+12.2pp
GDPval-AA v2	1,747.8 Elo	1,759.6 Elo	**Fable**	+11.8 Elo
AA Coding Agent	80.0	77.2	**GPT-5.6**	+2.8
DeepSWE v1.1	72.7%	69.7%	**GPT-5.6**	+3.0pp
Terminal-Bench 2.1	88.8%	83.1%	**GPT-5.6**	+5.7pp
BrowseComp	90.4%	88.0%*	**GPT-5.6**	+2.4pp
Toolathlon	58.0%	61.7%	**Fable**	+3.7pp
GPQA Diamond	94.6%	92.6%	**GPT-5.6**	+2.0pp

* OpenAI 表内 BrowseComp 的 88.0 对应 Claude Mythos 5；Fable 未单列。此处用于能力层级参照，不视为严格同模型比较。

来源：OpenAI《GPT-5.6》官方评测表。注意不同页面中的分数可能因 harness、版本和推理档位不同。

截至 2026-07-10

效率不是一个数字：三家公司各选了不同证据

最可信的信号不是“谁最省”，而是行业开始同时优化质量、成本、时间和操作步数

GPT-5.6 Sol

61% 更快

AA Intelligence Index：较 Fable 5 用时少 61%，估算成本约一半；分数相差 1 点

Claude Fable 5

25–30% 更快

早期客户称其在日常表格套件中较 Opus 4.8 更快；属于客户案例，不是统一基准

Grok 4.5

80 TPS

SWE Bench Pro 平均输出 15,954 tokens，约为 Opus 4.8 的 1/4.2

模型单价 × **Token 数** × **步骤 / 工具轮次** × **墙钟时间**

真实 TCO = 四者共同作用；只看 token 单价会低估高效 Agent 的价值，也会高估昂贵模型的成本。

来源：OpenAI、Anthropic、SpaceXAI 官方发布页。各效率数字的参照对象不同，不做直接倍数排名。

截至 2026-07-10

Agent 路线分化：编排、自治、吞吐

三家都在卖“完成任务”，但控制面放在不同位置

GPT-5.6

模型内编排

1. 写轻量程序协调工具
2. 过滤中间结果
3. Ultra 默认 4 Agent 并行
4. 交付可编辑成品

协调复杂性

Fable 5

长时自治

1. 跨百万 token 保持目标
2. 用文件笔记形成持久记忆
3. 自我反思与验证
4. 复杂任务优势随时长扩大

延长自治时间

Grok 4.5

快速工具循环

1. Grok Build 默认模型
2. 函数 / Web / X / 代码工具
3. 上下文压缩支持长循环
4. 低单价支撑高频调用

压低循环成本

来源：三家官方发布页与官方 API 文档。

截至 2026-07-10

知识工作：三家都盯上 Office，但“完成”的定义不同

从写答案升级到生成可交付文件，是三方共同趋势

能力层	GPT-5.6	Fable 5	Grok 4.5
文档 / 表格	**格式遵循、公式/财务模型、排版精度**	**复杂分析、法律 redline、金融推理**	**Word 清晰写作；Excel 多表公式 + Web 研究**
演示文稿	**从材料生成可编辑 deck；学习母版与设计系统**	**视觉理解强，能从截图重建应用**	**PowerPoint 原生形状、复杂图示与叙事**
计算机使用	**OSWorld 2.0：62.6%**	**视觉-only 完成 Pokémon；强调少 scaffolding**	**主要强调插件与 Build 内工具循环**
厂商要证明	**“最后一公里完成度”**	**“复杂任务判断力”**	**“嵌入现有办公流”**

洞察：模型竞争正在从“输出内容”迁移到“理解模板 → 调用工具 → 检查结果 → 交付原生文件”。

来源：三家官方发布页中的知识工作、Office 与计算机使用章节。

科学与网络安全：能力越强，访问架构越重要

Anthropic 把访问分层做成产品；OpenAI 用可信访问与推理监控；xAI 发布页弱化安全叙事

GPT-5.6	能力证据	ExploitBench 73.5%；SEC-Bench Pro 71.2%；生命科学多项提升	治理机制	分层防护 + 实时检查 + 推理监控；约 70 万 A100e GPU 小时自动红队
Fable / Mythos	能力证据	药物设计流程约 10×；14 个靶点中 9 个产出强候选；分子生物假设盲测约 80% 被偏好	治理机制	Fable 在敏感领域回退至 Opus 4.8；Mythos 向可信机构开放；要求 30 天留存
Grok 4.5	能力证据	训练数据强调科学、工程、数学；官方发布页未披露同等粒度科研/网络安全数据	治理机制	发布页核心是编码、速度、Office 与价格；需另行审查企业治理与高风险领域边界

提示：官方未披露 ≠ 能力不存在；只表示本次材料不足以支撑同等强度结论。

来源：OpenAI 与 Anthropic 官方发布页；xAI 仅依据本次 Grok 4.5 官方发布材料。

截至 2026-07-10

安全不是附录，而是模型体验的一部分

同一能力如何开放，已经成为模型产品差异化

Fable 5

保守分类器

<5% 会话触发回退；敏感生物/网络请求由 Opus 4.8 处理

GPT-5.6

分层、动态判断

高风险能力面向验证用户；安全栈使用推理监控与持续更新

Grok 4.5

发布叙事更轻

本次页面主要谈性能与效率，治理信息需从其他文档补足

更强默认限制

更依赖账户、场景与企业治理

采购含义：评估模型时必须同时测试“能力命中率、误拦截率、审计/留存要求、可信访问流程”。

来源：OpenAI 与 Anthropic 官方安全章节；Grok 仅评价本次发布页的信息重心。

截至 2026-07-10

发布会真正卖的是什么？

从指标选择，可以反推三家下一阶段的商业重心

厂商	核心卖点	产品路径	战略含义
OpenAI	**性能/美元 + 可编辑成品 + 多 Agent**	把 ChatGPT、Codex、Work 和 API 收束成统一工作层	强调“小模型也能超过昂贵竞品”，扩大用量与工作流渗透
Anthropic	**长时自治 + 科研发现 + 安全访问**	把 Claude Code 的工程口碑外溢到金融、法律、生命科学	以 Fable / Mythos 分层同时争取大众市场与高价值受控市场
SpaceXAI	**80 TPS + $2/$6 + Office/Build**	用基础设施与低成本快速进入开发者和办公入口	先以部署经济性抢量，再用生态入口放大模型价值

共同方向：模型本体逐渐商品化，竞争转向推理预算、Agent 编排、原生工具、交付格式与治理。

分析基于三篇官方发布材料的叙事结构、指标选择与产品入口。

截至 2026-07-10

怎么选：先按工作负载，而不是按品牌

以下为材料驱动的场景建议，仍应以企业自己的任务集做 A/B 测试

工作负载	首选	为什么	验证重点
复杂跨工具交付 / 研究报告 / 可编辑文件	**GPT-5.6 Sol**	Agent、浏览、OS 操作、设计与文档完成度证据最完整	端到端成功率、Ultra 成本、成品可编辑性
超长代码迁移 / 长时自治工程	**Claude Fable 5**	厂商与客户案例集中证明长程自治、记忆和复杂工程判断	中断恢复、回滚、误拦截、30 天留存
生命科学 / 高端网络安全研究	**Fable / Mythos**	科研案例最强；可信访问路径最清晰	资格、数据边界、敏感请求回退
高吞吐编码 Agent / 成本敏感自动化	**Grok 4.5**	$2/$6、80 TPS、较少 token；终端能力接近前沿	真实任务通过率、幻觉、缓存命中率
Office 插件内快速生产	**Grok 4.5 或 GPT-5.6**	Grok 强调原生 Office 入口；GPT 强调模板/母版与可编辑成品	模板忠实度、图表/公式正确性

建议：用 30–50 个真实任务，记录成功率、人工返工分钟数、总 token、工具调用次数与墙钟时间。

企业评测：不要只看 pass@1，要看“成功交付成本”

推荐把模型测试变成一张可复用的运营仪表盘

40%	20%	15%	15%	10%
任务成功率	**返工成本**	**墙钟时间**	**全链路成本**	**治理损耗**
结果是否真正可用	人工修订分钟数	从请求到最终成品	token + 工具 + 重试	误拦截、审计、数据约束

成功交付成本 = 模型/API 成本 + 人工返工成本 + 失败重试成本 + 治理成本

比“单 token 价格”更接近真实 ROI，也能解释为什么更贵的模型有时反而更省。

建议权重为通用起点，可按业务风险与吞吐要求调整。

接下来最值得观察的 5 个变量

这轮发布不是终局，而是 Agent 经济学和访问治理的开端

01	**多 Agent 是否值得**	Ultra 提升是否能覆盖额外 token 与编排复杂度
02	**长时自治的可靠性**	持续数小时/数天后，错误累积与恢复能力是否可控
03	**低价能否保持质量**	Grok 的成本优势能否在非编码、非精选任务集上延续
04	**安全分层的用户摩擦**	回退、误拦截、可信访问和留存要求如何影响采用
05	**Office 成品化竞争**	母版、公式、原生图形、可编辑性将成为新的真实世界基准

观察框架：性能、经济性、可靠性、治理、原生交付。

FINAL TAKE

没有绝对冠军，只有不同的优化函数

GPT-5.6	把能力变成完整、可伸缩的工作系统
Fable 5	把长时自治与前沿科研推到新层级
Grok 4.5	把接近前沿的智能做得更快、更便宜

真正的采购问题：谁能以最低的“成功交付成本”，稳定完成你的真实任务？

资料日期：2026-07-10

来源与口径说明

所有核心结论优先来自用户指定的三篇官方材料

OpenAI	**GPT-5.6: Frontier intelligence that scales with your ambition** https://openai.com/index/gpt-5-6/	2026-07-09
Anthropic	**Claude Fable 5 and Claude Mythos 5** https://www.anthropic.com/news/claude-fable-5-mythos-5	2026-06-09；2026-07-01 恢复可用
SpaceXAI	**Introducing Grok 4.5** https://x.ai/news/grok-4-5	2026-07-08
SpaceXAI Docs	**grok-4.5 / Models（补充 500K 上下文与 API 工具）** https://docs.x.ai/developers/grok-4-5	2026-07-08 更新

口径提醒：不同厂商页面可能使用不同模型版本、推理档位、工具环境与 harness；本稿不把跨页面数字合成为单一总分。