2026 前沿大模型 • 三强对决

GPT-5.6 • Claude Fable 5 • Grok 4.5 三大前沿模型深度对比

基于 OpenAI、Anthropic、SpaceXAI 三家官方发布文档，横向拆解各模型的能力特点、核心叙事与关键数据，读懂这一轮竞争到底在比什么。

OpenAI • GPT-5.6（Sol / Terra / Luna）• 7-09

Anthropic • Claude Fable 5 • 6-09

SpaceXAI • Grok 4.5 • 7-08

数据来源：openai.com/index/gpt-5-6 · anthropic.com/news/claude-fable-5 · x.ai/news/grok-4-5 · Artificial Analysis 独立评测

一句话，看懂三家想抢的位置

GPT-5.6

OpenAI

- 关键词：效率 × 性价比 × 多智能体
- 三档产品线 Sol / Terra / Luna，按需选算力
- 用更少 token、更低成本拿到同级甚至更强结果
- ultra 模式并行调度多个 agent 提速

每一分钱买到更多智能

Claude Fable 5

Anthropic

- 关键词：能力天花板 × 长程任务 × 安全
- Mythos 级模型的“安全版”，能力最强
- 任务越长越复杂，领先幅度越大
- 科研、蛋白设计、编程长程自治突出

最强能力 + 强安全护栏

Grok 4.5

SpaceXAI / xAI

- 关键词：够用前沿 × 极致省 token × 低价
- Opus 级智能，输出速度 80 TPS
- token 效率约为 Opus 4.8 的 4.2 倍
- 算力 + 模型 + Cursor 编码数据垂直闭环

前沿智能的性价比之选

基本盘

规格、定价与产品形态：起跑线上的差异

维度	GPT-5.6 Sol	Claude Fable 5	Grok 4.5
厂商 / 发布	OpenAI • 7-09	Anthropic • 6-09	SpaceXAI • 7-08
产品分层	Sol / Terra / Luna 三档	Fable 5（+受限 Mythos 5）	单一 grok-4.5
输入价 / 1M	$5 ★	$10	$2 ★
输出价 / 1M	$30	$50	$6 ★
上下文	百万级（1M）	百万级 token	500k
规模 / 架构	未公开	Mythos 级	V9 • 约 1.5T 参数
核心场景	编码•知识工作•计算机使用	软件工程•科研•视觉	编码•代理•办公

★ = **该维度当前最优。价格分化极大：**Grok 输出价仅为 Fable 的 1/8、GPT 的 1/5；Fable 用最高定价对应最强能力，三家形成“能力向上、价格向下”的两极张力。

深度剖析 · 01

GPT-5.6：把“每一个 token 的智能”做到极致

53.6

Agents' Last Exam 新高

领先 Fable 13.1 分

-61%

同级智能下任务耗时

更短，成本约一半

80

Coding Agent Index

新 SOTA，token 减半

核心特点

三档模型 Sol / Terra / Luna 覆盖旗舰到低价，Terra/Luna 以约 1/16 成本超越 Fable。新增 ultra 模式默认并行调度 4 个 agent，把“分数-时延”前沿整体推向更强更快。设计判断力与计算机使用能力大幅提升，可自查并打磨界面成品再交付。

想重点突出的核心点

叙事不是“我最聪明”，而是“同样的钱办更多事”——性能/美元、程序化工具调用、更少模型往返。同时强调最强网络安全模型，并在加速 OpenAI 自身研究。

“more successful work for the same spend.”

深度剖析 · 02

Claude Fable 5：能力天花板，越难越强

80%

SWE-Bench Pro 最高分
编码能力登顶

10×

Mythos 5 辅助蛋白设计
流程提速约 10 倍

~80%

科学家盲测更偏好
其分子生物学假设

核心特点

Mythos 级模型的通用安全版，几乎全 benchmark 达 SOTA。长程自治最强：Stripe 用它把 5000 万行代码库迁移，从“团队两个月”压缩到一天。视觉、记忆、长上下文全面领先，能借助自身笔记持续改进产出。

想重点突出的核心点

核心叙事是“前沿能力 + 可负责任地释放”：一面展示科研级突破（基因组、蛋白、新假设），一面配套安全分类器——敏感请求回退到 Opus 4.8，触发率 <5%，1000+ 小时红队未现通用越狱。

“越长、越复杂的任务，Fable 5 的领先越大。”

Grok 4.5：够用的前沿，压到最低的成本

4.2×
输出 token 比
Opus 4.8 更省

80 TPS
flash 级
输出速度

29%
SWE Marathon 居首
超 Opus 4.8 与 Fable

核心特点

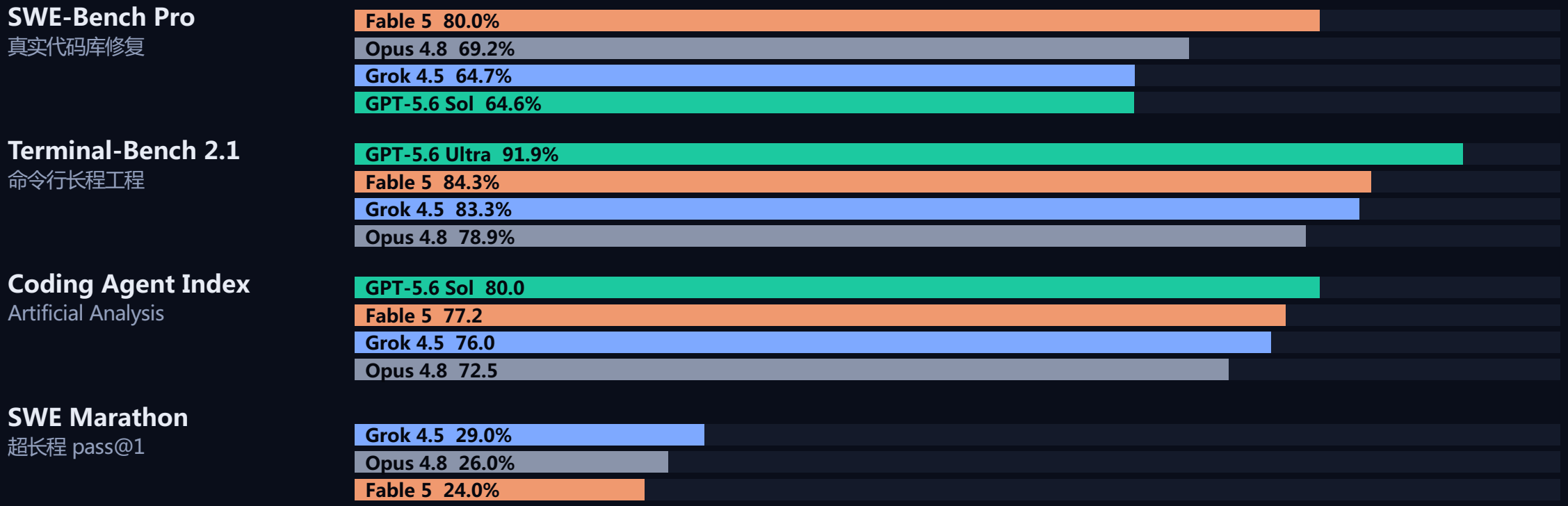
V9 基座约 1.5T 参数，融合 Cursor 编码数据做补充训练。在 SWE Marathon、τ³-Banking 上真实领先，Terminal-Bench 2.1 紧咬第一梯队（83.3%）。500k 上下文、三档推理、内置 X 搜索与代码执行，主打代理与办公。

想重点突出的核心点

不争“全维第一”，主打“Opus 级智能 × 极致性价比”：$2/$6 定价、4.2× token 效率、每编码代理任务约 $2.49，稳居成本-性能 Pareto 前沿。背后是算力 + 模型 + Cursor 的垂直闭环。

“frontier intelligence, faster and at far lower cost.”

编程与工程能力横评：谁在赢哪一类任务



洞察：没有单一赢家。Fable 赢传统工程基准，GPT 赢代理化/命令行与多智能体，Grok 独占超长程 SWE Marathon。选型要看具体是哪一类工程任务，而非笼统的“谁更强”。

数据洞察 · 02

智能水平 vs 成本效率：三家价值主张一目了然

关键数据：AA 智能指数 Fable≈59.9、GPT-5.6 Sol≈58.9（差距不到 1 分）、Grok≈54；但 GPT 完成任务比 Fable 少约 61% 时间、约一半成本，Grok 输出价低至 Fable 的 1/8。智能趋同，竞争主战场正转向“单位成本的智能”。

三家“最想让你记住”的一句话

OpenAI 的核心命题

效率即护城河。

用产品分层（Sol/Terra/Luna）+ ultra 多智能体，把“性能/美元”做成叙事主轴：不承诺绝对最聪明，而承诺“同样预算下完成更多、更快、更便宜”。

主打指标：Agents' Last Exam、性能/美元、-61% 耗时

Anthropic 的核心命题

能力上限 + 责任感。

强调这是能力最强的通用可用模型，尤其在长程与科研上拉开身位；同时用安全分类器、回退机制、红队结果，把“强大”与“可控”绑定叙述。

主打指标：SWE-Bench Pro、科研突破、安全护栏 <5%

SpaceXAI 的核心命题

前沿智能的普惠化。

用 Opus 级能力 + 最低价 + 最高 token 效率 + 80 TPS 速度，主打“人人用得起的前沿”，并靠算力/模型/Cursor 数据的垂直闭环支撑快速迭代。

主打指标：token 效率 4.2×、$2/$6、SWE Marathon 第一

三种路线映射三种战略：OpenAI 抢规模化落地，Anthropic 抢能力与安全高地，SpaceXAI 抢成本与速度的价格带。

综合评估

能力矩阵：一张表看清相对强弱

能力维度	GPT-5.6	Fable 5	Grok 4.5
顶尖工程编码（SWE-Bench Pro）	中	最强 ★	中高
代理 / 命令行 / 多智能体	最强 ★	强	强
超长程自治任务	强	最强 ★	强（Marathon 居首）
知识工作 / 办公文档	最强 ★	强	强
科研 / 生命科学	强	最强 ★	中
成本效率 / token 效率	强	中	最强 ★
速度	强	中	最强（80 TPS） ★
安全护栏成熟度	强	最强 ★	未充分披露

★ = **该维度相对最优。整体格局：**Fable 攻高端能力，GPT 攻综合效率，Grok 攻性价比——三者在不同象限各有明确主场。

结论

怎么选？按你的真实场景对号入座

选 GPT-5.6

规模化落地 · 预算敏感团队

你要在大量日常任务里稳定拿到高质量结果，又在意时间和账单；需要计算机使用、多智能体并行、知识工作产出（PPT/文档/表格）。三档模型让你按任务难度精细控成本。

选 Claude Fable 5

最难的问题 · 科研与关键工程

你面对超长程、超复杂或前沿科研级任务，需要能力天花板和最强长程自治，且对安全合规要求高。愿意为顶级能力付更高单价。注意：部分生物/网安问题会回退到 Opus。

选 Grok 4.5

高并发 · 成本优先 · 代理工作流

你要在大规模代理/编码工作流里把单位成本压到最低，又不想牺牲太多智能；看重速度（80 TPS）和 token 效率。适合高频调用、成本主导的生产环境。

一句话总结：这一代前沿模型的智能差距在收敛，真正的分水岭是“能力上限、综合效率、极致性价比”三条不同赛道。选型不再是“谁最强”，而是“谁最贴合你的任务与预算”。