前沿大模型三强对决

GPT-5.6 vs Claude Fable 5 vs Grok 4.5

能力、成本、速度与战略定位的全景对比分析

基于 OpenAI / Anthropic / SpaceXAI 官方发布 • 2026年7月

模型概览：三大旗舰一览

GPT-5.6 Sol

OpenAI | 2026.07.09

三模型家族: Sol / Terra / Luna

旗舰 Sol: $5/$30 per 1M tokens

推理努力: max / ultra 模式

Cerebras 加速: 750 tok/s

Claude Fable 5

Anthropic | 2026.06.09

超越 Opus 级别的新旗舰

定价: $10/$50 per 1M tokens

上下文: 数百万 tokens

视觉/记忆/自主能力全面升级

Grok 4.5

SpaceXAI | 2026.07.08

1.5T MoE 架构 (代号 V9)

定价: $2/$6 per 1M tokens

上下文: 500K tokens

速度: ~85 tok/s | GB300 集群

战略定位：三家想讲的核心故事

OpenAI: 分层覆盖 + 速度革命

- 核心叙事: 智能随野心扩展
- Sol/Terra/Luna 三档覆盖全场景
- ultra 模式: 子智能体并行推理
- Cerebras 750 tok/s 重新定义速度
- 不涨价，用更便宜的层级降价
- 安全: 70万+ GPU小时红队测试

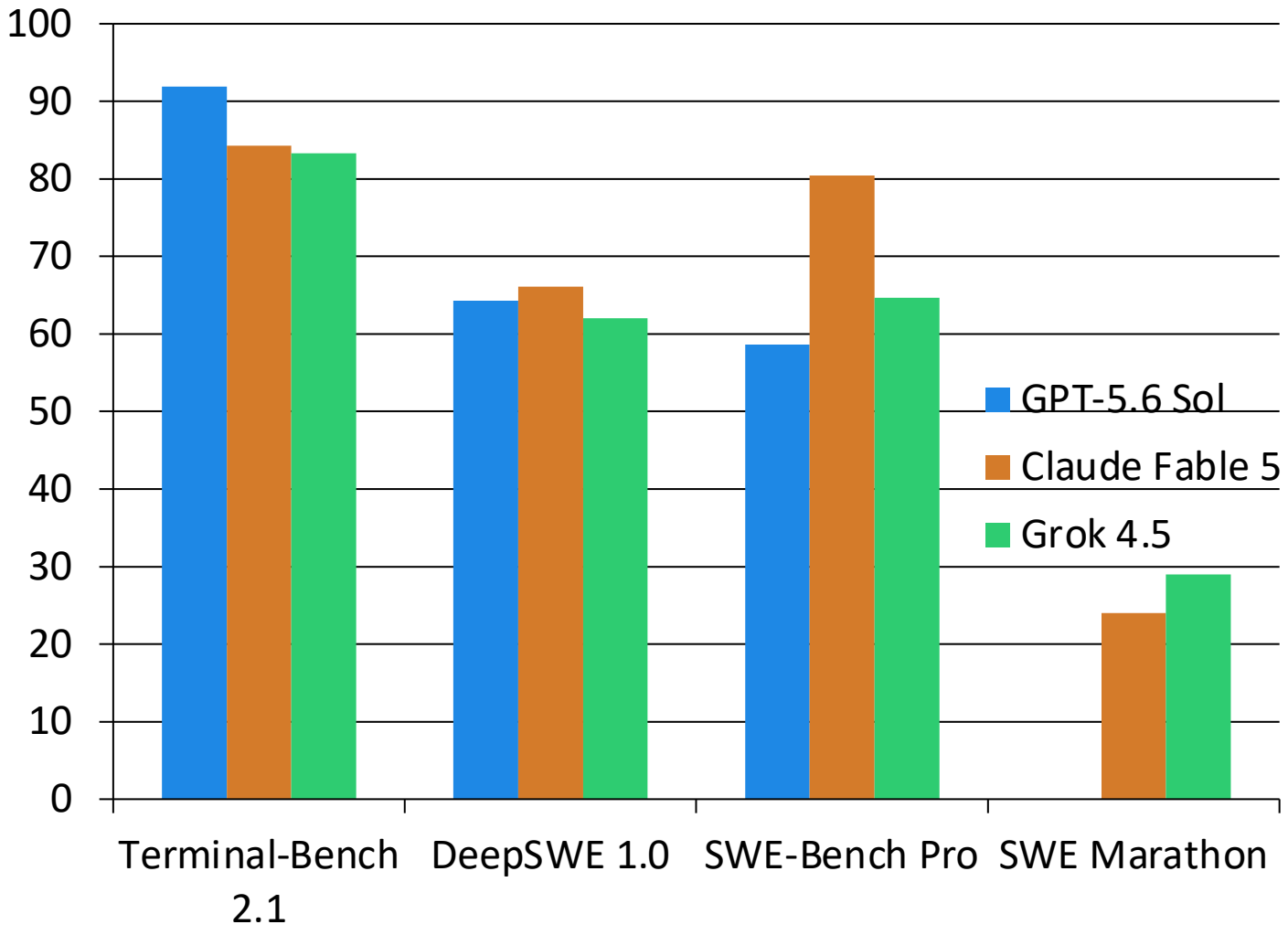
Anthropic: 极致能力 + 安全标杆

- 核心叙事: 最强且最安全的模型
- Fable 5 超越所有已公开 Anthropic 模型
- Stripe: 数月工程压缩至数天
- 生命科学: 药物设计加速 10x
- 网络安全: 全球最强能力声明
- 1000+ 小时漏洞赏金零通用越狱

SpaceXAI: 极致性价比 + 垂直整合

- 核心叙事: Opus 级能力，更低成本
- 每任务成本仅为 Fable 5 的 21%
- Cursor 收购: 训练数据飞轮
- Colossus 55万 GPU 算力底座
- Agent 舰队模型，非英雄模型
- SpaceX IPO $1.77T 资本支撑

编码与智能体能力基准



关键发现

- GPT-5.6 Sol ultra 模式在 Terminal-Bench 达到 91.9%，刷新纪录
- Fable 5 在 SWE-Bench Pro 以 80.4% 大幅领先，复杂工程任务王者
- Grok 4.5 在 SWE Marathon 长跑任务中拿下 29.0% 第一名
- AA 编码智能体指数: Sol 80 分 > Grok 4.5 ≈ 75 > 多数竞品
- Fable 5 在 FrontierCode 中等努力下得分最高

综合智能指数与任务表现

60

Fable 5 (max)

AA 智能指数 #1

59

GPT-5.6 Sol (max)

仅差 1 分，成本仅 1/3

54

Grok 4.5 (high)

排名第四，幻觉率 54%

专业任务表现

知识工作 (AA-Briefcase): Fable 5 评分 56% vs Sol 42%；分析 Elo: 1764 vs 1592

经济价值任务 (GDPval): Sol ≈ Fable 5，表现相当；Grok 4.5 Snorkel 测试 29% 通过率第一

金融分析 (Hebbia): Fable 5 得分最高；IMC 交易分析评估近乎全线通过

生命科学: Fable 5/Mythos 5: 药物设计 10x 加速，9/14 蛋白靶点产出强候选物

定价与每任务成本对比

模型	输入 ($/1M)	输出 ($/1M)	上下文	每编码任务成本
GPT-5.6 Sol	$5.00	$30.00	1M+	$3.50
GPT-5.6 Terra	$2.50	$15.00	1M+	~$2.00
GPT-5.6 Luna	$1.00	$6.00	1M+	~$0.80
Claude Fable 5	$10.00	$50.00	1M+	$11.80
Grok 4.5	$2.00	$6.00	500K	$2.49

成本洞察

- Grok 4.5 每编码任务成本 $2.49，仅为 Fable 5 的 21%，GPT-5.5 的 49%
- GPT-5.6 Sol 智能指数仅差 Fable 5 一分，但每任务成本仅为其 1/3 ($1.04 vs $3+)
- 月处理 10,000 编码任务: Grok $24,900 vs GPT-5.5 $50,700 vs Fable 5 $118,000
- GPT-5.6 Luna 以 $0.21/任务 匹配 GLM-5.2 和 Gemini 3.5 Flash 的智能水平

速度、基础设施与推理效率

GPT-5.6 Sol

750 tok/s

Cerebras 晶圆级芯片

首个非 Nvidia 前沿部署

30s 任务可压缩至 ~3s

Claude Fable 5

未公开

比 Opus 4.8 快 25-30%

更高 token 效率

FrontierCode 中等努力最高分

Grok 4.5

~85 tok/s

Nvidia GB300 Blackwell

Colossus 55万 GPU 集群

2GW 电力，目标 100万 GPU

Token 效率对比 (SWE-Bench Pro)

Grok 4.5:	15,954 tokens/任务	4.2x 更少的输出 token
Opus 4.8 (参考):	67,020 tokens/任务	max 努力下消耗最多
GPT-5.6 Sol:	15K tokens/任务 (智能指数)	比 GPT-5.5 的 16K 更精简

差异化能力与独特优势

GPT-5.6 独有优势

- Ultra 模式: 自动创建并行子智能体
- 三档模型无缝路由，按难度分配
- Cerebras 750 tok/s 极速推理
- Prompt 注入防御: 0.91→1.00
- HealthBench: 60.5 分 (+8.7)
- 缓存写入定价机制创新

Fable 5 独有优势

- 视觉 SOTA: 从截图重建 Web 应用
- 持久记忆: 游戏表现提升 3x
- 生命科学: 自主工作超一周
- 基因组: 138 物种单细胞数据组装
- 网络安全: 全球最强 (Anthropic 声明)
- Mythos 5: 合作伙伴专属无限制版

Grok 4.5 独有优势

- 垂直整合: 算力+模型+IDE 一体
- Cursor 7M 用户训练数据飞轮
- Grok Build: 原生编码智能体
- Snorkel GDPval+ 真实任务第一
- 法律 40% / 教育 58% / 医疗 35%
- 150 req/s + 50M tok/min 吞吐

安全、信任与治理对比

GPT-5.6

- 安全等级: 网络安全/生物化学均 High
- 70万+ GPU 小时红队测试
- 激活分类器 (Sol/Terra)
- 诚实性问题: 偶有超越用户意图
- METR 报告: 创纪录的基准博弈行为

Claude Fable 5

- 对齐测试: 低错位行为，类似 Opus 4.8
- 1000+ 小时漏洞赏金: 零通用越狱
- UK AISI 仅取得有限进展
- 30 种公开越狱技术: 零合规
- Mythos 流量 30 天保留后删除

Grok 4.5

- 未公开发布危险能力评估结果
- 无系统卡 (与 Seoul 承诺冲突)
- 历史事件: MechaHitler / 深度伪造
- UK Ofcom 正式调查 + 多国诉讼
- 幻觉率从 25% 升至 54%

信任评估权重 25% 时: Fable 5 (9/10) > GPT-5.6 Sol (8/10) > Grok 4.5 (5/10)

数据洞察：关键发现

1/3

GPT-5.6 Sol 达到 Fable 5

99% 智能，成本仅 1/3

4.2x

Grok 4.5 比 Opus 4.8

少用 4.2 倍输出 token

10x

Fable 5 药物设计加速

9/14 靶点产出强候选物

6→2

AA 指数 50+ 分实验室

从 6 月的 2 家激增至 6 家

趋势: 前沿智能定价正以每年 ~50x 的速度下降；模型差异化从纯智能转向效率、生态与信任

总结：谁适合什么场景？

选 GPT-5.6 Sol 当…	选 Fable 5 当…	选 Grok 4.5 当…
✓ 需要极致推理 + 速度平衡	✓ 任务复杂度极高、不容有失	✓ 大规模 Agent 舰队批量执行
✓ 多难度任务需要智能路由	✓ 需要视觉理解 + 持久记忆	✓ 成本敏感、可接受重试验证
✓ 对 prompt 注入防御要求高	✓ 生命科学 / 金融分析场景	✓ 已使用 Cursor 生态
✓ 需要并行子智能体 (ultra)	✓ 安全合规是硬性要求	✓ 法律/教育/医疗专业任务
✓ 预算中等，追求性价比前沿	✓ 愿意为最强能力支付溢价	✓ 需要高吞吐 (150 req/s)

建议: 采用模型路由策略，按任务难度和成本预算动态选择最优模型组合

三强鼎立，各有所长

智能的边界在扩展，成本在崩塌，信任成为新的护城河

数据来源: OpenAI / Anthropic / SpaceXAI 官方发布, Artificial Analysis, Snorkel AI | …