AI 模型性能与费用对比分析
Terminal Bench v2 编程基准测试

svtter

https://svtter.cn

2026 年 1 月 3 日

1 概要

本文档提供了主流 AI 模型在 Terminal Bench v2 编程基准测试上的全面对比，包括性能指标和成本分析。

2 性能概览

表 1: Terminal Bench v2 性能对比

排名	模型	准确率	Agent	日期
1	**GPT-5.2**	**64.9% ± 2.8**	Droid	2025-12-24
2	**Gemini 3 Flash**	**64.3% ± 2.8**	Junie CLI	2025-12-23
-	**Claude Opus 4.5**	**59.3%**	-	2025-11-24
-	**MiniMax M2.1**	**47.9%**	Claude Code	2025-12-23
22	**Claude Sonnet 4.5**	**46.5% ± 2.4**	CAMEL-AI	2025-12-24
28	Claude Sonnet 4.5	42.8% ± 2.8	Terminus 2	2025-10-31
-	**GLM-4.7**	**41.0%**	-	2025-12-22
58	GLM-4.6	24.5% ± 2.4	Terminus 2	2025-11-01

表 2: API 定价（每百万 Token）

模型	输入	输出	缓存输入	性能
GPT-5.2	$1.75	$14.00	$0.175 (90%)	64.9%
Gemini 3 Flash	$0.50	$3.00	-	64.3%
Claude Opus 4.5	$5.00	$25.00	$0.50 (90%)	59.3%
Claude Sonnet 4.5	$3.00	$15.00	$0.30 (90%)	42.8-46.5%
MiniMax M2.1	$0.30	$1.20	$0.03 (90%)	47.9%
GLM-4.7	$0.60	$2.20	$0.11 (82%)	41.0%

表 3: 性价比排名（100 万输入 + 100 万输出 Token）

排名	模型	总成本	得分	单位成本	评级
1	Gemini 3 Flash	$3.50	64.3%	$0.054	
2	MiniMax M2.1	$1.50	47.9%	$0.031	
3	GLM-4.7	$2.80	41.0%	$0.068	
4	GPT-5.2	$15.75	64.9%	$0.243	
5	Claude Opus 4.5	$30.00	59.3%	$0.506	
6	Claude Sonnet 4.5	$18.00	46.5%	$0.387	

3 费用对比

4 性价比分析

5 其他基准测试

6 模型迭代提升

7 关键发现

7.1 最佳性能

- **GPT-5.2**: 最高准确率（64.9%），但价格较贵
- **Gemini 3 Flash**: 接近顶级性能（64.3%），性价比极佳
- **Claude Opus 4.5**: 第三名（59.3%），SWE-bench 最高分（80.9%）

7.2 最佳性价比

- **Gemini 3 Flash**: 最佳性能成本比（每分仅需 $0.054）
- **MiniMax M2.1**: 绝对价格最低（总计 $1.50），适合预算有限的用户
- **GLM-4.7**: 强大的中文支持，价格极具竞争力

表 4: 多项基准测试对比

基准测试	GPT-5.2	Gemini 3 Pro	Opus 4.5	MiniMax M2.1	Sonnet 4.5	GLM-4
SWE-bench Verified	80.0%	78.0%	80.9%	74.0%	77.2%	73.8%
SWE-bench Multilingual	72.0%	65.0%	-	72.5%	68.0%	66.7%
LiveCodeBench	-	90.7%	-	81.0%	64.0%	84.9%
Terminal Bench 2.0	64.9%	54.2%	59.3%	47.9%	42.8%	41.0%

表 5: 代际改进对比

模型	当前版本	前代版本	提升幅度	提升百分比
GLM-4.7	41.0%	24.5% (GLM-4.6)	+16.5%	+67.3%
MiniMax M2.1	47.9%	30.0% (M2)	+17.9%	+59.7%

7.3 使用场景推荐

表 6: 模型选择指南

使用场景	推荐模型	理由
追求最高性能	GPT-5.2 或 Gemini 3 Flash	顶级得分（64.9%, 64.3%）
追求性价比	Gemini 3 Flash	最优的性能成本比
预算有限	MiniMax M2.1	绝对价格最低
中文开发	GLM-4.7 或 MiniMax M2.1	优秀的中文支持
订阅计划	GLM-4.7	$3-20/月的编程计划

8 订阅计划

9 结论

Gemini 3 Flash 在大多数使用场景中脱颖而出，提供接近顶级的性能（64.3%），价格仅为 GPT-5.2 的 22%。对于预算紧张的用户，**MiniMax M2.1** 提供了最低的绝对成本，同时保持了不错的性能表现。中文开发者可能更倾向于选择 **GLM-4.7**，因其卓越的中文语言支持和实惠的订阅选项。

注：所有性能数据基于 2025 年 12 月的 Terminal Bench v2 评测。价格可能会发生变化。

表 7: 可用订阅选项

模型	计划	价格	支持工具
GLM-4.7	GLM Coding Plan	$3-20/月	Claude Code, Cline, Roo Code
MiniMax M2.1	Coding Plan	$100-200/年	多种编程 Agent
Claude Sonnet 4.5	Claude Pro	$20/月	通用 AI 助手