

2025 主要大语言模型（GPT-5.1、Gemini 3.0、Grok 4.1 和 Claude 4.5）能力与价格研究

概述

随着 2025 年末新一代大语言模型（LLM）推出，开发者和产品团队可以根据任务复杂度、预算及安全需求选择不同模型。本文结合 Clarifai 2025 年 11 月的比较文章和其它公开资料，汇总了几款领先模型的能力、上下文窗口、价格及适用场景。使用此表可以帮助系统按任务动态选取模型以降低成本并提升效果。

GPT-5.1（OpenAI）

OpenAI 的 GPT-5.1 是 2025 年的旗舰模型，综合能力强，且价格相对低廉，适合通用写作与自动化工作流。

- **特性与优势：**
- **工具调用与代码执行：**支持 `apply_patch` 和 `shell` 等安全工具，允许模型生成补丁文件、运行代码或搭建 DevOps 流水线 ¹。
- **长上下文与缓存：**16 K-196 K 的上下文窗口能处理较长对话；提示缓存可以在重复查询时减少令牌消耗 ²。
- **推理与输出控制：**新增 `reasoning_effort` 参数可指定快速或深度推理；`verbosity` 调节输出冗长程度 ³。
- **自定义工具和安全限制：**支持声明自定义工具并限制允许调用的工具 ⁴。
- **价格：**OpenAI 官方定价为输入约 **1.25 美元/百万 token**、输出 **10 美元/百万 token** ²。
- **上下文窗口：**Clarifai 文章指出 GPT-5.1 的实际上下文范围在 **16-196 K** ²。相比其他模型，上下文较小，但对大多数对话和文档任务足够。
- **推荐任务：**成本敏感的研究与知识查询、快速迭代的代码生成与补丁打包、一般内容创作或与向量搜索结合的检索 ⁵。需要深推理时可调整推理参数。

Gemini 3.0 Pro（Google）

Google 于 2025 年发布的 Gemini 3.0 Pro 注重推理和多模态理解，是复杂项目的首选。

- **特性与优势：**
- **多模态与深层推理：**Clarifai 指出 Gemini 3.0 在推理和多模态理解方面领先，适合跨文本、代码和其他媒体的任务 ⁶。
- **Antigravity 平台与 Deep Think：**该平台允许模型直接访问开发环境——打开文件、搜索网页、运行命令并测试代码；Deep Think 切换让模型分配更多算力来处理复杂任务 ¹。
- **大量上下文窗口：**上下文长达 **1-2 M tokens**，适用于长篇研究、复杂规划与多步骤推理 ²。
- **强大的数学与算法能力：**在 LiveCodeBench/SWE-Bench 测试中取得 2439 Elo 和 76.2 % 准确率 ⁷。教育与辅导任务中，Gemini 3 无需工具即可在数学问题上表现优异 ⁵。

- **价格**：Clarifai 报告成本约为 **2 美元/百万输入 tokens，12 美元/百万输出 tokens** ²。Google 提供免费层带有检索功能。
- **推荐任务**：深度研究与知识工作、跨模态分析、算法设计及商业策略规划等高复杂度任务；适用于长时间仿真或多步骤 workflow ⁵。因其与 Google 生态深度集成，应考虑供应商锁定及数据治理风险 ²。

Grok 4.1 (xAI)

由 Elon Musk 旗下的 xAI 推出的 Grok 4.1 将情感理解与实时检索相结合，专注对话和创意写作。

- **特性与优势**：
- **大容量上下文与实时检索**：Grok 4.1 拥有大约 **2 M token** 的上下文窗口，并结合实时浏览功能，可在对话中融入最新信息 ⁶。
- **情绪智力与创意写作**：Yahoo 新闻指出新模型的创意写作能力显著提升，在 Creative Writing v3 测试中击败 GPT-5.1；训练加入情绪智力数据，使其在 EQ Bench 上表现较好 ⁸。
- **错误率与幻觉率降低**：与 Grok 4 Fast 相比，Grok 4.1 的误答率下降约 7 个百分点，幻觉率从 **12.09 %** 降至 **4.22 %** ⁹。
- **Agentic 功能**：虽然不提供补丁或内存机制，但结合实时浏览可监控数日社交媒体趋势并生成摘要 ¹。
- **价格**：Clarifai 指出 Grok 4.1 的成本约 **3 美元/百万输入 tokens，15 美元/百万输出 tokens**，与 Claude 4.5 相当 ²。
- **推荐任务**：情感支持、创意写作、社交媒体趋势监测和对话类应用 ⁵。需要事实准确时应搭配推理型模型，因为 Grok 偶尔会在逻辑问题上出错 ²。

Claude 4.5 (Anthropic)

Anthropic 的 Claude 4.5 专长于代码和文本的长时处理，强调安全性与稳定性。

- **特性与优势**：
- **上下文编辑与记忆工具**：模型可自动删除无关数据以保持 token 限制内，配备外部 **memory 工具** 持久存储信息，且支持 **checkpoints** 回滚到早期状态；这些能力让模型可以自运行数小时 ¹。
- **长上下文窗口**：基本上下文达 **200 K tokens**，结合内存工具可扩展到更长 ²。
- **安全性与对齐**：Clarifai 文章提到 Claude 4.5 引入 ASL-3 安全过滤器，专注财务、法律和医疗领域的安全与合规 ²。
- **价格**：与 Grok 4.1 类似，每百万输入 token 约 **3 美元**，输出 **15 美元** ²。
- **推荐任务**：长时间编码和大型重构、科研报告撰写、复杂项目的安全运行及政策/法规敏感领域的文档整理 ⁵。不过过度使用可能导致表面化或含瑕疵的代码输出 ²。

上下文窗口与价格对比

模型	约可用上下文窗口	价格（每百万 token）	关键特性简述
GPT-5.1	约 16-196 K tokens ²	输入 \$1.25 ；输出 \$10 ²	通用写作、代码执行、长对话缓存、成本最低
Gemini 3.0 Pro	1-2 M tokens ²	输入 \$2 ；输出 \$12 ²	深度推理、多模态理解、Antigravity 平台与 Deep Think ¹

模型	约可用上下文窗口	价格（每百万 token）	关键特性简述
Grok 4.1	~2 M tokens ⁶	输入 \$3；输出 \$15 ²	实时检索、情绪智力、创意写作，幻觉率降至 4.22 % ⁹
Claude 4.5	200 K tokens+内存扩展 ²	输入 \$3；输出 \$15 ²	上下文编辑、持久记忆工具、checkpoint 回滚、适合安全敏感任务 ¹