厂商	官方平台 / 文档入口	推荐通用模型（截至 2025-11）	调用方式（核心信息）	典型价格*（USD / 1M tokens）
MiniMax	开发者平台：`https://platform.minimax.io`（platform.minimax.io）文档：`/docs`	`MiniMax-M2`（通用+代码+agent，M2-Stable 为商用稳态版）（Minimax Ai）	**OpenAI 兼容**：`base_url="https://api.minimax.io/v1"`，`model="MiniMax-M2"` 或 `"MiniMax-M2-Stable"`（Minimax Ai）	官方文本接口：`MiniMax-M2` 输入约 **$0.30/M**，输出 **$1.20/M** tokens（不含缓存折扣）（platform.minimax.io）
Kimi / Moonshot	开放平台：`https://platform.moonshot.ai`（platform.moonshot.ai）文档：`/docs` → Chat API / Pricing	Kimi K2 系列：`kimi-k2` / `kimi-k2-0711-preview` / `kimi-k2-thinking`（长期推的是 K2 + K2 Thinking）（moonshotai.github.io）	**OpenAI 兼容**：`base_url="https://api.moonshot.ai/v1"`，模型可用 `kimi-latest` / `kimi-k2-*` 系列，接口路径 `/chat/completions`（Kimi Ai）	Kimi K2 / K2 Thinking：输入命中缓存最低约 **$0.15/M**，cache miss 约 **$0.60/M**，输出约 **$2.50/M**；普通 K2 API 文档也给出约 $0.15/Min$，**2.5/M out** 的量级（Kimi Ai）
智谱 GLM	中国站：`https://bigmodel.cn`（open.bigmodel.cn）国际：`https://z.ai` 文档：`https://docs.z.ai`	`GLM-4.6`（最新旗舰，200K context，推 agent & code）（open.bigmodel.cn）	官方 HTTP：`POST https://api.z.ai/api/paas/v4/chat/completions`，`Authorization: Bearer <API_KEY>`，`model: "glm-4.6"`（docs.z.ai）	Z.ai 官方价：`GLM-4.6` 输入约 **$0.60/M**，缓存输入 **$0.11/M**，输出 **$2.20/M** tokens（docs.z.ai）
DeepSeek	平台：`https://platform.deepseek.com`（api-docs.deepseek.com）文档：`https://api-docs.deepseek.com`	通用：`deepseek-chat`；推理：`deepseek-reasoner`（统一 Chat Completion 接口）（api-docs.deepseek.com）	**OpenAI 兼容**：`base_url="https://api.deepseek.com"`，模型名如 `"deepseek-chat"` / `"deepseek-reasoner"`，路径 `/chat/completions`（api-docs.deepseek.com）	官方表：`deepseek-chat` 输入 cache miss **$0.27/M**，cache hit **$0.07/M**，输出 **$1.10/M**；`deepseek-reasoner` 输入 **$0.55/M**，cache hit **$0.14/M**，输出 **$2.19/M**（api-docs.deepseek.com）；新 V3.x 系列有更便宜档位，整体仍明显低于 GPT-5 价位（Venturebeat）
Qwen（阿里 / 通义千问）	Qwen 官网：`https://qwen.ai`（Qwen）模型 & 价格：阿里云 Model Studio 文档 `Model Studio → Models & pricing`（阿里云）	国际推荐：`qwen3-max`（旗舰推理）+ `qwen-plus`（性价比通用）+ `qwen-flash`（极低价高速）+ 代码向 `qwen3-coder-plus` / `qwen3-coder-flash`（阿里云）	**OpenAI 兼容**：`base_url="https://dashscope-intl.aliyuncs.com/compatible-mode/v1"`（国际）或 `https://dashscope.aliyuncs.com/compatible-mode/v1`（北京）；模型如 `"qwen3-max"` / `"qwen-plus"` / `"qwen-flash"`（阿里云）	代表档位（国际，新 Qwen3 系列）：`qwen3-max` 输入约 **$0.86–1.2/M**，输出 **$3.4–6/M**（按 context 区间分档）；`qwen-plus` 低档输入 **$0.40/M**、输出 **$1.20/M**（non-thinking）；`qwen-flash` 低档输入 **$0.05/M**、输出 **$0.40/M**，极便宜（阿里云）