

# 模型网关 · 测试说明与配额调度指南

感谢各位开发者参与现阶段技术测试并提供使用反馈。本文档说明测试期间的客户端控制台功能、统一美金额度计费规则（官方 API 原价 × 通道倍率，缓存读取 0.1x）、每日 00:00 自动重置与额度用满后的处理以及多轮会话上下文管理建议。

## 【置顶入口】Web 客户端控制台（申领密钥 / 查询用量 / 邀请扩容）

- 客户端控制台网址：<https://aghub.lol>
- 管理员激活与技术支持：WeChat: [Huaodnn](#) | Telegram: [+60 1125261626](#) | X: [@ha\\_hua64139](#)
- 主要功能说明：
  - 自助申领测试密钥（激活后起算周期）：在 `/client` 切换到「申请新密钥」，选择 微信（WX:）、Telegram（TG:）或 X（X:）并填写联系账号（邀请码可填受邀码 `INV-XXXX` 或留空），生成专属 `sk-...` 密钥；复制网页显示的 `AG-...` 账户编号发送给管理员（微信 [Huaodnn](#)、TG [+60 1125261626](#) 或 X [@ha\\_hua64139](#)），后台激活后正式起算 9 天两阶段测试周期（前 48 小时 \$5.00 USD/天，后 7 天 \$2.00 USD/天，全程不限请求次数与旗舰模型调用次数）。
  - 实时用量面板：查看今日美金消耗进度（每日 00:00 自动重置）、充值余额、旗舰模型今日调用次数（仅统计，不设上限）、分模型用量与最近请求明细；额度用满时页面顶部会直接提示原因。
  - 测试邀请配额扩展：网页提供专属邀请码（`INV-XXXX`）与邀请链接。每有 1 位受邀开发者完成激活，系统自动为双方各追加 +3 天有效期 与 +\$1.00 USD / 天 美金额度。
  - 按需扩容与更高并发：如需增加每日美金额度、充值余额或申请多并发 / 企业专属通道，请直接联系管理员办理。

## 一、计费规则：统一按美金额度扣费，没有层级

所有密钥只看美金额度，不再区分旗舰层级、也不限制旗舰调用次数：你选择什么模型，网关就调用什么模型，并按该模型的价格扣减美金额度；额度够用时不会因为用量多而被自动换成其他模型。

| 项目     | 规则                                                                                                                                                                  |
|--------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 扣费方式   | 每次请求按所选模型的官方 API 原价 × 通道倍率折算美金，从当日额度（或充值余额）中扣减；缓存命中的输入 Token 按 0.1x 计                                                                                               |
| 通道倍率   | Claude 全系：主力编码通道 0.9x，纯净零注入通道（ <code>sk-pure-...</code> 或在 <code>/client</code> 切换）1.6x；Gemini 4 Argon 与 Gemini 常规系列 0.9x；GPT / Codex 全系 0.2x；Gemini 3.8 Flash 0.9x |
| 每日美金额度 | 测试卡前 48 小时 \$5.00 / 天，后 7 天 \$2.00 / 天；每邀请 1 位好友激活 +\$1.00 / 天；每日 <code>00:00 (UTC+8)</code> 自动重置                                                                   |
| 充值余额   | 充值卡按余额扣减（1 RMB = \$1.00 USD，全模型共享），余额用完后续费即可恢复                                                                                                                      |

| 项目       | 规则                                                                                                  |
|----------|-----------------------------------------------------------------------------------------------------|
| 请求次数     | 不限                                                                                                  |
| 旗舰模型调用次数 | 不限（ <code>claude-opus-*</code> 、 <code>gemini-4-argon-*</code> 、 <code>gpt-6-astra</code> 同样只按美金扣费） |
| 额度用满后    | 返回 HTTP 429 并注明原因：今日美金额度已用完（次日 00:00 自动重置）或余额已用完（续费后恢复）；不会自动降级或替换模型                                 |
| 并发       | 测试卡默认 1（深度测试资格为 2），可联系管理员申请更高并发                                                                     |

用量核算细则

1. Prompt Cache 缓存命中按 0.1x（1折）计费：

- 网关内置会话粘滞路由（Sticky Session Cache Routing），自动按会话指纹将同一多轮任务固定在同一上游节点（多轮 Agent 会话缓存命中率通常为 90% ~ 96%）。
- 命中上下文缓存的输入 Token（`cache_read`）对齐官方计费标准，按输入单价的 10%（**0.1x**）计入美金消耗。

2. 官方 API 原价（输入 / 输出 / 缓存读取，每百万 Token）：

- Claude Opus 5.5（`claude-opus-5-5*`）：\$4.00 / \$20.00 / \$0.20
- Claude Opus 5 / 4.x（`claude-opus-5` / `claude-opus-4-*`）：\$5.00 / \$25.00 / \$0.50
- Claude Sonnet 5.5（`claude-sonnet-5-5*`）：\$2.00 / \$10.00 / \$0.10
- Claude Sonnet 4.6（`claude-sonnet-4-6*`）：\$3.00 / \$15.00 / \$0.30
- Claude Haiku 5.5（`claude-haiku-5-5`）：\$0.10 / \$0.50 / \$0.01（超 100K 上下文：\$0.50 / \$2.50）
- Claude Haiku 4.5（`claude-haiku-4-5-20251001`）：\$1.00 / \$5.00 / \$0.10
- GPT-6 Astra（`gpt-6-astra`）：\$10.00 / \$50.00 / \$1.00
- OpenAI / Codex 主力（`gpt-6.1-sol` / `gpt-6-sol` / `codex-auto-review`）：\$2.00 / \$10.00 / \$0.10 ~ \$0.20
- GPT-5.x（`gpt-5.6-*` / `gpt-5.5`）：\$2.50 / \$15.00 / \$0.25
- GPT-6 Luna（`gpt-6-luna`）：\$0.10 / \$0.50 / \$0.01
- Gemini 4 Argon（`gemini-4-argon*`）：\$3.50 / \$17.50 / \$0.35
- Gemini 3.8 Flash（`gemini-3.8-flash*`）：\$0.75 / \$3.75 / \$0.075
- Gemini 常规系列（`gemini-3-flash` / `gemini-2.5-flash` 等）：\$0.15 ~ \$0.50 / \$0.60 ~ \$3.00

3. 计算示例：在主力编码通道调用 `claude-opus-5-5-high`，输入 100K Token（其中 90K 命中缓存）、输出 5K Token：官方原价 = 10K × \$4.00/M + 90K × \$0.20/M + 5K × \$20.00/M = \$0.158，乘以通道倍率 0.9x 后扣减 \$0.142。

4. 异常响应与预检请求 0 计入：

- 任何异常响应（HTTP >= 400）以及 Claude Code 发送的 `/v1/messages/count_tokens` 预检请求，均不计入美金消耗。

## 二、测试模型总表与选型建议

所有模型均支持在统一接口（`https://api.aghub.lol`）下直接调用（完整模型列表可在 `/client` 页面点击复制）：

| 模型系列              | 可用模型 ID                                                                                                                                              | Cursor 专用别名                                                         | 官方原价（输入 / 输出 每百万 Token）             | 通道倍率                                        | 适用开发场景                                            |
|-------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------|-------------------------------------|---------------------------------------------|---------------------------------------------------|
| Claude 5.5 旗舰     | <code>claude-opus-5-5-high</code><br><code>claude-opus-5-5-medium</code><br><code>claude-opus-5-5-low</code><br><code>claude-opus-5-5</code>         | <code>ag-opus</code><br><code>gw-claude-opus-5-5</code>             | <b>\$4.00 / \$20.00</b> （缓存 \$0.20） | 编码 <code>0.9x</code> · 纯净 <code>1.6x</code> | 旗舰深度推理：复杂系统架构设计、底层疑难排障、跨多文件重构                     |
| Gemini 4 Argon 旗舰 | <code>gemini-4-argon-high</code><br><code>gemini-4-argon-medium</code><br><code>gemini-4-argon-low</code><br><code>gemini-4-argon</code>             | <code>ag-argon</code><br><code>gw-gemini-4-argon</code>             | <b>\$3.50 / \$17.50</b> （缓存 \$0.35） | <code>0.9x</code>                           | Google 4.0 旗舰推理：1M 上下文窗口 + 64K 输出，适合跨库代码理解与深度逻辑推理 |
| Claude 4.x 旗舰     | <code>claude-opus-4-8 / claude-opus-4-7</code><br><code>claude-opus-4-6 / claude-opus-4-6-thinking</code>                                            | <code>gw-claude-opus</code><br><code>gw-claude-opus-thinking</code> | <b>\$5.00 / \$25.00</b> （缓存 \$0.50） | 编码 <code>0.9x</code> · 纯净 <code>1.6x</code> | 专项深度推理：兼容指定 4.x Opus 版本的评测与复杂推理任务                 |
| Claude 5.5 主力     | <code>claude-sonnet-5-5-high</code><br><code>claude-sonnet-5-5-medium</code><br><code>claude-sonnet-5-5-low</code><br><code>claude-sonnet-5-5</code> | <code>ag-sonnet</code><br><code>gw-claude-sonnet-5-5</code>         | <b>\$2.00 / \$10.00</b> （缓存 \$0.10） | 编码 <code>0.9x</code> · 纯净 <code>1.6x</code> | Claude 5.5 日常主力：兼顾代码质量与响应速度，适合日常业务开发与前后端联调        |

| 模型系列                | 可用模型 ID                                                                                       | Cursor 专用别名                                   | 官方原价（输入 / 输出 每百万 Token）                                    | 通道倍率               | 适用开发场景                                        |
|---------------------|-----------------------------------------------------------------------------------------------|-----------------------------------------------|------------------------------------------------------------|--------------------|-----------------------------------------------|
| Claude 4.6 主力       | claude-sonnet-4-6<br>claude-sonnet-4-6-thinking                                               | gw-claude-sonnet<br>gw-claude-sonnet-thinking | \$3.00 / \$15.00（缓存 \$0.30）                                | 编码 0.9x<br>纯净 1.6x | Claude 4.6 专项兼容                               |
| Claude Haiku 轻量     | claude-haiku-5-5<br>claude-haiku-4-5-20251001                                                 | 直接填原名                                         | \$0.10 / \$0.50（缓存 \$0.01）<br>4.5: \$1.00 / \$5.00         | 编码 0.9x<br>纯净 1.6x | 高频日常编码：1M 上下文窗口 + 极低单价，适合大量简单改动与快速问答          |
| OpenAI 6.x 编码与推理    | gpt-6.1-sol<br>gpt-6-sol<br>gpt-6-astra                                                       | ag-gpt6<br>(或直接填原名)                           | \$2.00 / \$10.00<br>gpt-6-astra : \$10.00 / \$50.00        | 0.2x               | OpenAI 编码与算法主力：倍率最低、最省额度，适合高频编码与 Agent 任务     |
| OpenAI 轻量与专项        | gpt-6-luna<br>gpt-5.6-sol /<br>gpt-5.6-luna /<br>gpt-5.6-terra<br>gpt-5.5 / codex-auto-review | 直接填原名                                         | gpt-6-luna : \$0.10 / \$0.50<br>gpt-5.x : \$2.50 / \$15.00 | 0.2x               | 代码审查与常规重构：适合批量代码检视（codex-auto-review）与中大型模块迭代 |
| Gemini 3.8 Flash 轻量 | gemini-3.8-flash-high<br>gemini-3.8-flash-medium<br>gemini-3.8-flash-low                      | ag-flash<br>(注：gemini-2.5-pro 自动映射至此)         | \$0.75 / \$3.75（缓存 \$0.075）                                | 0.9x               | 轻量高速编码与问答：适合日常问答、文档总结与单测生成                    |
| Gemini 常规系列         | gemini-3-flash /<br>gemini-2.5-flash<br>gemini-3.5-flash-low /<br>gemini-3.5-flash-lite       | 直接填原名                                         | \$0.15~\$0.50 /<br>\$0.60~\$3.00                           | 0.9x               | 轻量多模态与快速检索：适合长文本摘要与快速检索                       |

关于「思考强度（Thinking / Reasoning Effort）」的说明

1. 支持四档思考强度调节（-low / -medium / -high / -xhigh）：
- -low：低思考 Token 开销，优先响应速度；

- `-medium`：速度与推理深度均衡；
  - `-high`：高深度推理预算（旗舰模型默认档位）；
  - `-xhigh`（或 `-max` / `-ultrathink`）：深度推理档位，适合复杂算法与架构分析。
2. 客户端参数优先级：若在客户端（如 Codex、Claude Code、Cherry Studio）中显式设置了思考档位（`reasoning_effort`、`thinking.budget_tokens`、`output_config.effort`），网关优先采用客户端参数；也可直接在模型名末尾添加 `-low` / `-medium` / `-high` / `-xhigh` 后缀（如 `gpt-6.1-sol-xhigh`），由网关自动注入对应思考参数。
- 

### 三、多轮会话上下文管理建议

在 API 调用模式下，客户端（如 Claude Code、Cursor、Cherry Studio）每次发起请求都会携带当前会话的完整历史上下文：– 第 1 条消息（含少量代码文件）：约 `10K Token` – 同一窗口连续对话至第 10 轮：累计历史上下文可达 `150K Token` – 此后在该窗口继续提问，客户端每次均会重发前序历史（其中命中缓存的历史部分按 `0.1x` 计入，未命中部分按原价计入）。

#### 降低无效上下文消耗的建议

1. 完成独立子任务后及时新建或压缩会话：同一任务内连续追问可利用 `0.1x` 缓存权重；当一个独立问题解决或切换至无关模块时，建议新建会话（Claude Code 中可执行 `/clear` 清空上下文，或执行 `/compact` 压缩历史记录）。
  2. 按任务复杂度选择模型：`claude-opus-5-5-high` 用于核心架构与复杂排障；日常代码检索、常规函数编写与测试生成可使用倍率最低的 `gpt-6.1-sol`（`0.2x`），或 `claude-sonnet-5-5-high`、`claude-haiku-5-5`、`gemini-3.8-flash-high`。
  3. 精简长日志输入：粘贴报错信息时建议截取核心 Traceback 段落，避免单次注入数百 KB 的原始未过滤日志。
- 

### 四、上下文无损转发、协议自愈与并发规则

1. 180k 原样转发与超限自动裁剪：
  - 常规会话无损转发：单次请求在 `180k Tokens`（`1.6MB`）以内时，网关原样转发全部消息，保持缓存前缀完整。
  - 超限窗口适配：仅当单次请求超过 `180k Tokens` 或上游返回 `400` 上下文超限错误时，网关自动保留首轮设定与最近对话进行滑动窗口裁剪。
  - 跨模型协议自愈：兼容 `claude-cli 2.1.280+` 对话中途注入的 `system / developer / tool` 角色块与 `tool_reference`（`ToolSearch`），并在请求中断时自动清理冲突思考签名、配平悬空的 `tool_use` / `tool_result`。
2. 单密钥并发控制：
  - 标准测试密钥默认并发上限为 `1`（深度测试资格为 `2`），并发请求按到达顺序排队执行。
3. 测试反馈与配额扩展：

- 测试期间如遇额度用满、需要增加每日美金额度或反馈调用问题，可联系管理员（WeChat: [Huaodnn](#) | Telegram: [+60 1125261626](#) | X: [@ha\\_hua64139](#)）并提供 [AG-...](#) 账户编号办理。