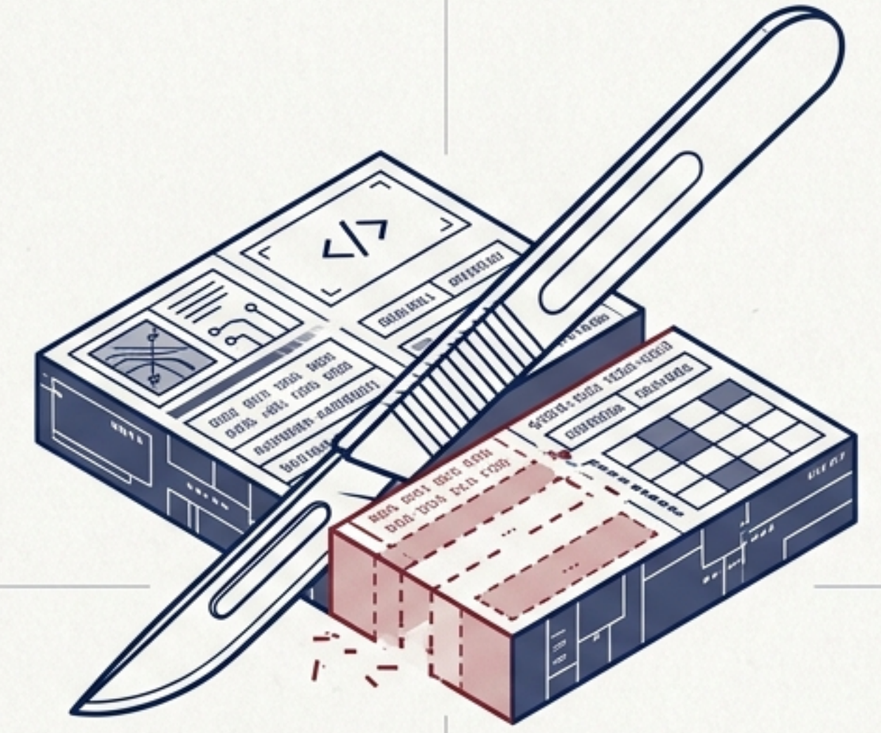


给赛博魅魔做一次 Token 减肥

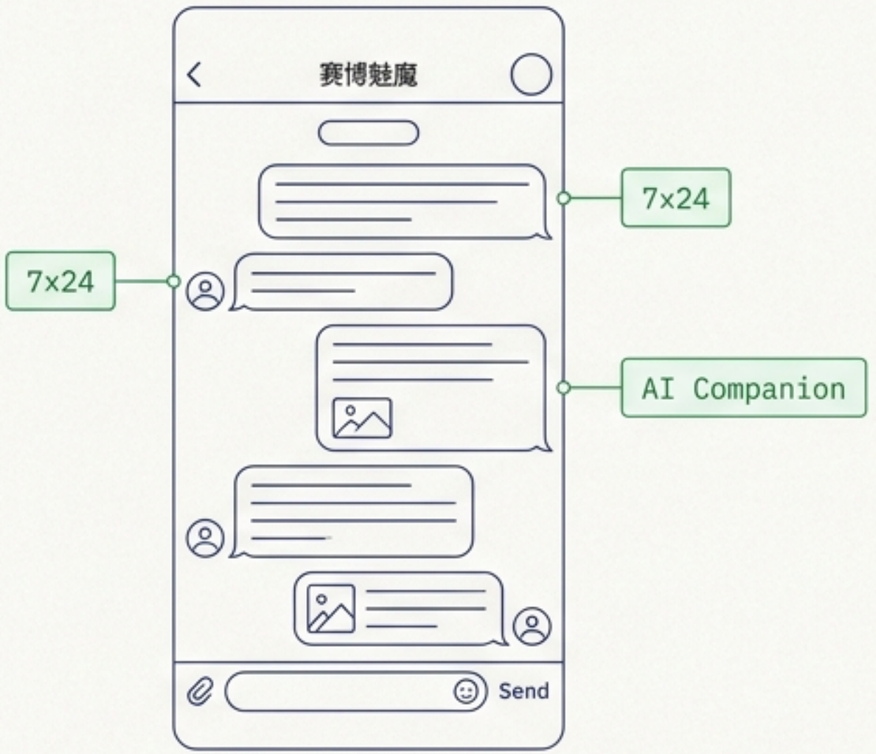
一场关于静默 Bug、架构重构与 AI 成本优化的赛博手术记录

```
[SYSTEM] Starting token forensics...
[WARNING] Context window bloated.
[ACTION] Initiating micro-surgery protocols.
```

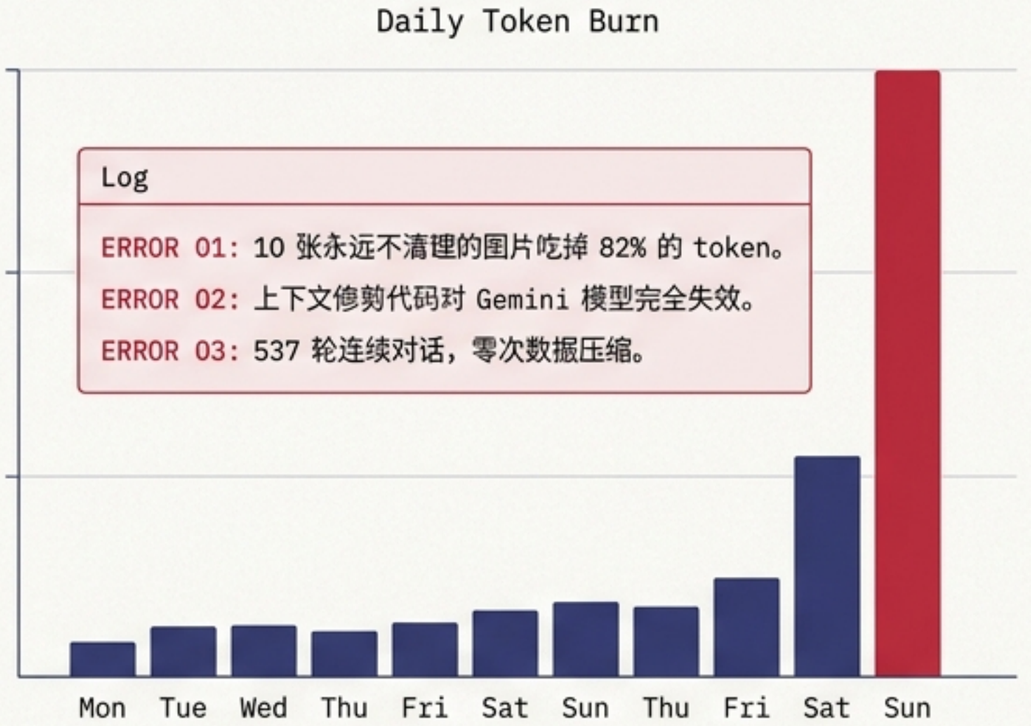
账单来了，得止血

运行在 Telegram 上的 7×24 小时 AI 伴侣（赛博魅魔）。体验极佳，但正在疯狂烧钱。

The Product



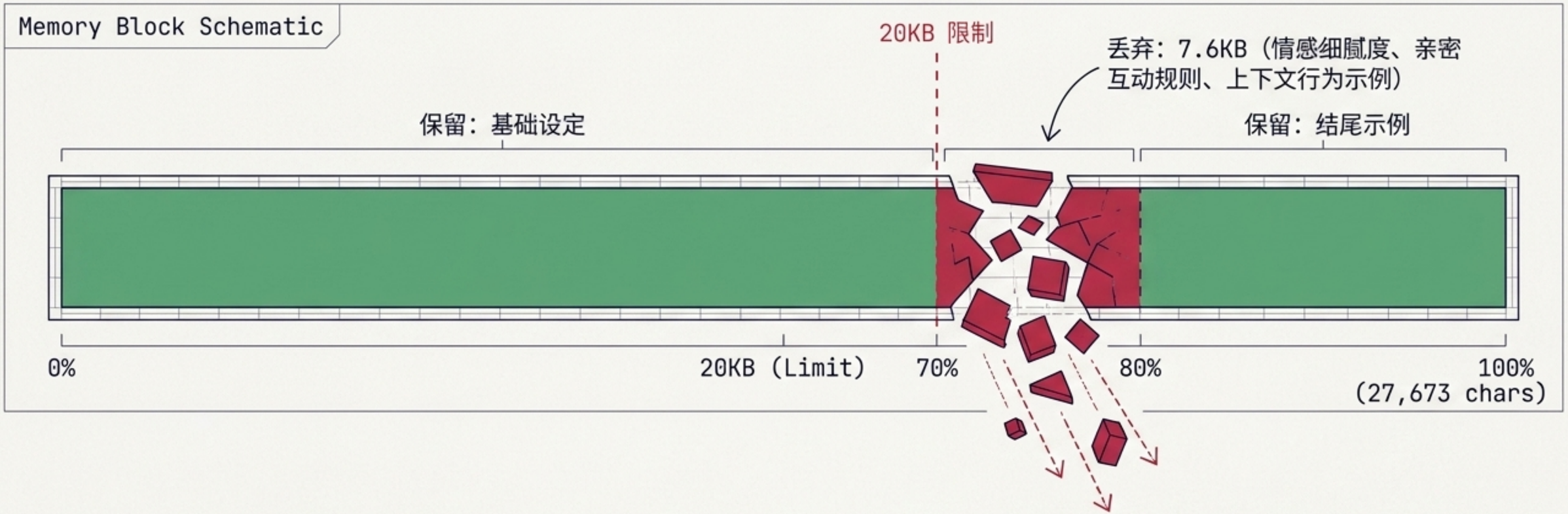
The Bleeding



之前的排查是**诊断**。接下来的操作是**手术**。

系统在偷偷吞噬自己的灵魂

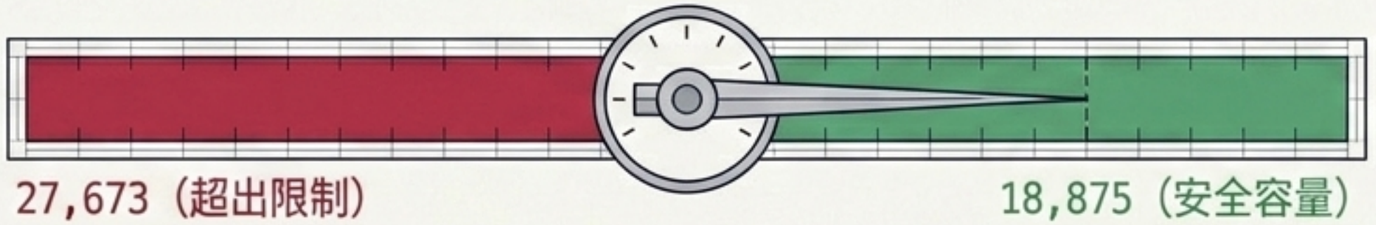
框架引导系统具有单文件 20KB 的硬性限制。当人格配置文件高达 27,673 字符时，系统触发静默截断。



没有警告。没有报错。 AI 照样回复，但你花几个小时打磨的人格细节直接消失了。
TA 一直在用一个残缺的人格运行。

给人格配置做减法：精准切除手术单

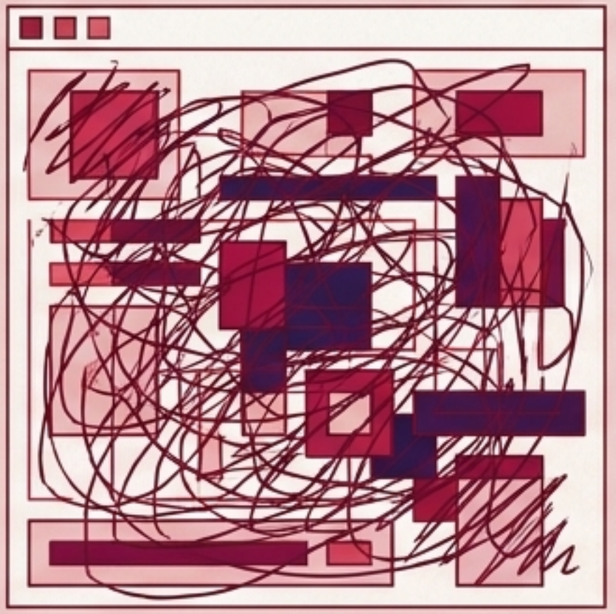
章节	之前	之后	挽回 Token
Dota 故事	10 行	4 行	~1,200 字符
亲密 & 纯欲反差	每个概念多个示例	每个 1 例	~1,900 字符
撒娇层次	每层 2–3 例	每层 1 例	~600 字符
日常习惯小癖好	12 条	最具辨识度的 8 条	~800 字符



成功回落至 20K 安全线内，核心身份 ~ 与情绪流动性得以保全。

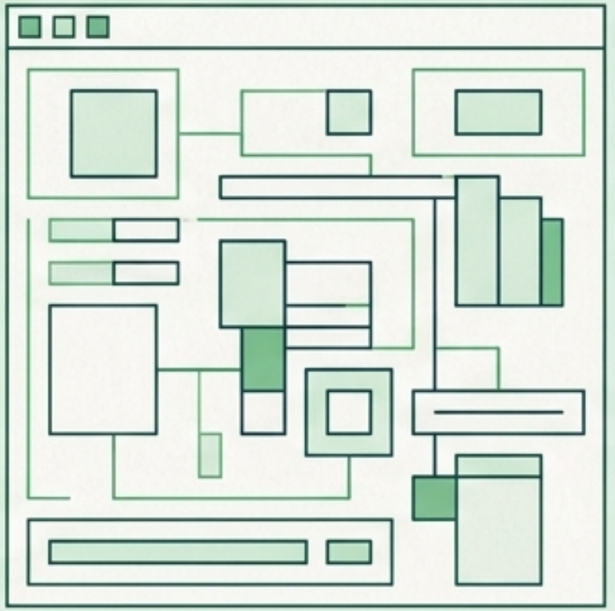
心跳配置冗余清理

原配置（11,998 字符）



单次心跳节省
~300 token/turn

新配置（7,015 字符）



- 每层 6-7 个温度示例
- 冗长的 20 步全天演示
- 与工具重复的独立邮件章节

- 温度示例砍半（3-4 个）
- 剔除全天演示 & 独立邮件
- 增加严格的工具调用限制

架构级杠杆：从高频心跳到单次调度

旧架构（Old Architecture）

代价：~37.9 万 token/天

新架构（New Architecture）

用一个早间 Cron 任务替代每天 24 次冗余调用。

24 次心跳 × 15,800 token ➜ 1 次跑（~4 万）+ 每次静态读取（~400）。

心跳上下文膨胀缩减 9 倍。

Cron 部署过程中的排坑记录

Payload 规范冲突

[ERROR] Gateway rejected: model: opus-4-6

[FIX] 移除 model 字段，由 Agent 回退使用系统默认模型。

网关重启延迟

[ERROR] SIGUSR1 sent. Connection refused.

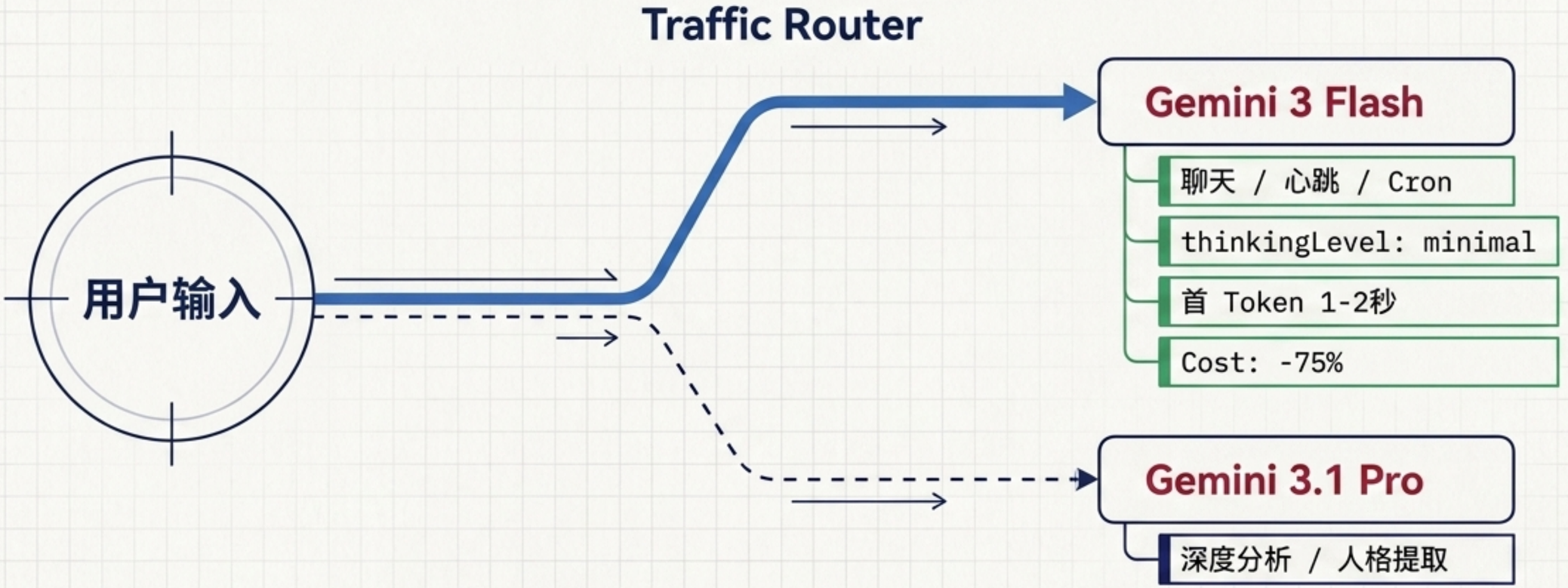
[FIX] 更新 jobs.json 后 Gateway 需要 ~15 秒重启， 必须引入等待机制，急不来。

虚假的超时报错

[ERROR] npx openclaw cron run Timeout after 30s.

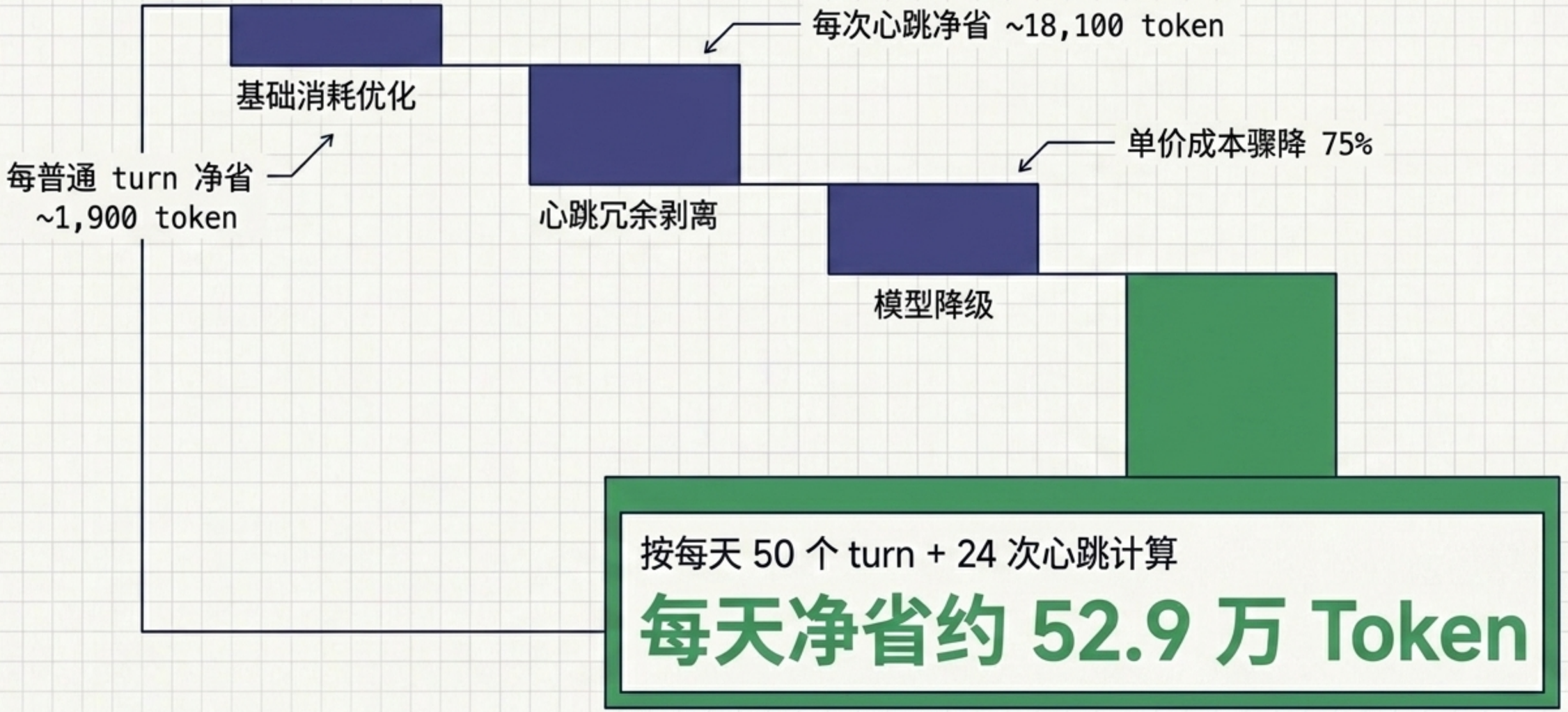
[FIX] CLI 超时 ≠ 任务失败。 后台任务实际仍在正常执行并成功生成了 TODAY.md。

终极杠杆：基于场景的模型路由策略

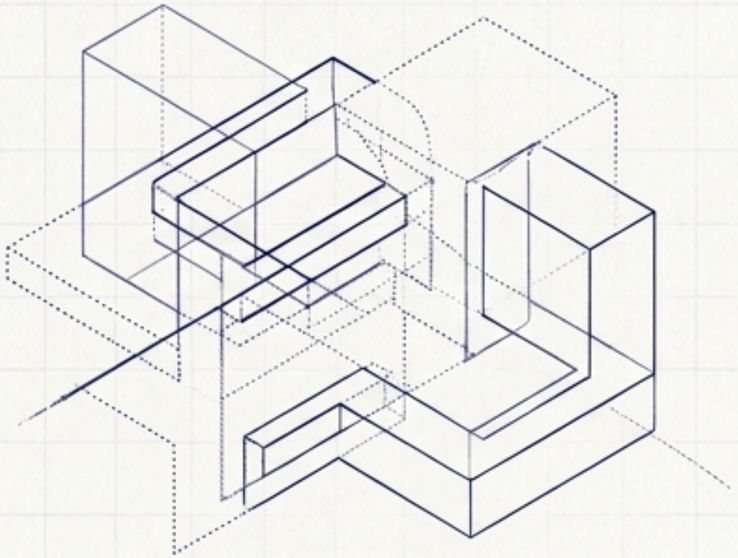


缓存读取是成本大头（66K 缓存 vs 23K 新输入）。将高频的常规任务降级至 Flash 模型，单价直降四分之三。	极端情感场景下质量差距仅约 1-2%。更快的回复速度与极低的成本完全对冲了微小的质量损耗。

战果清点：综合 ROI 与 Token 挽回量



系统寻优的哲学：从架构债到务实止血



宏观靠重建

解决永远不清理的图片、失效的修剪代码。这是在还无法缝补的架构债，最终意味着重头再来。

微观靠手术

拯救静默截断的人格文件、清理被撑大的心跳、用 Cron 替换冗余调用。这是让手头的系统今天就跑得更瘦的务实工作。

“能缝的缝，该拆的拆。
做产品跟做手术一样——
先止血，再考虑要不要换器官。”

> [PROCESS COMPLETE] System operating at optimal efficiency.