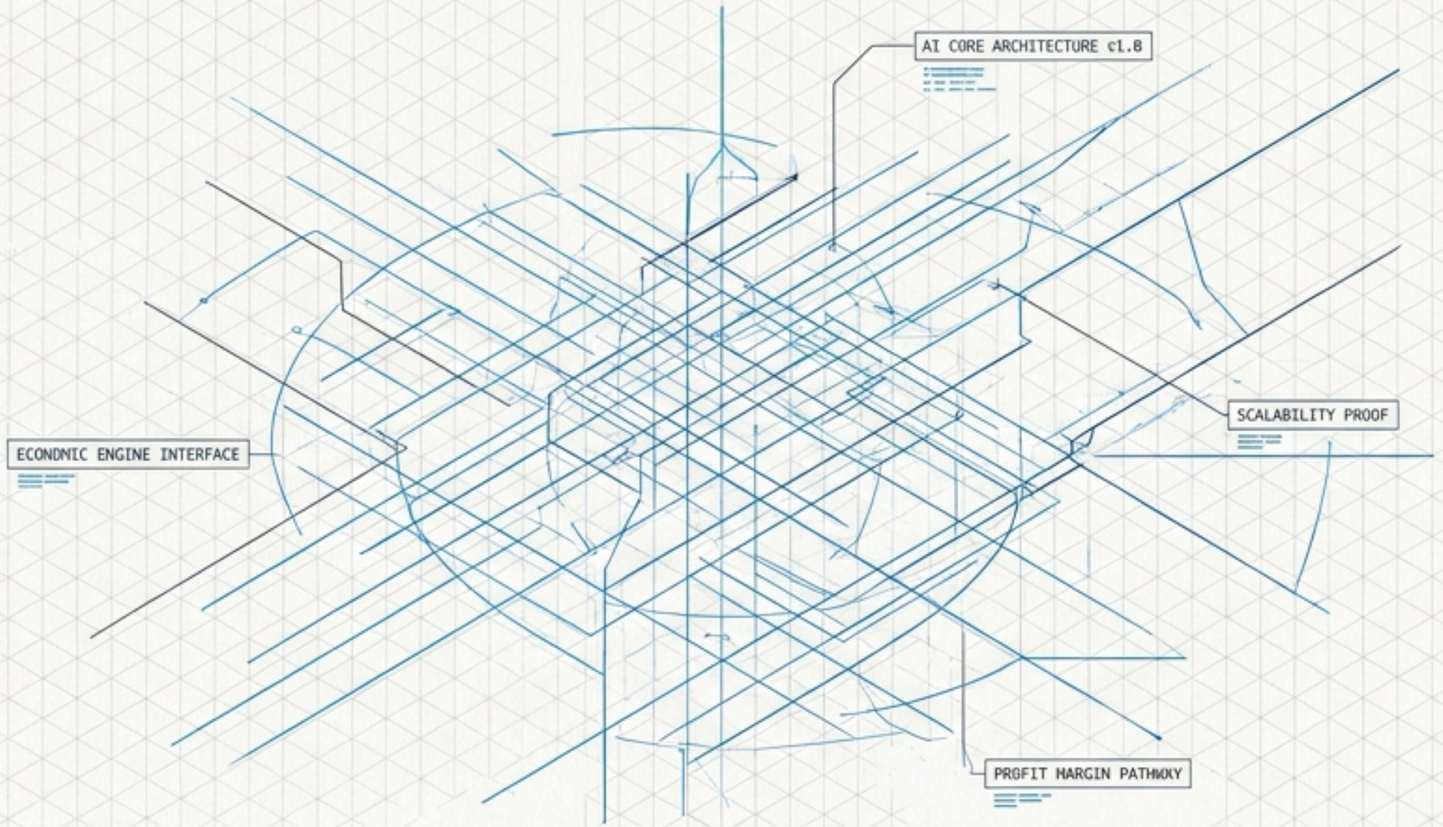
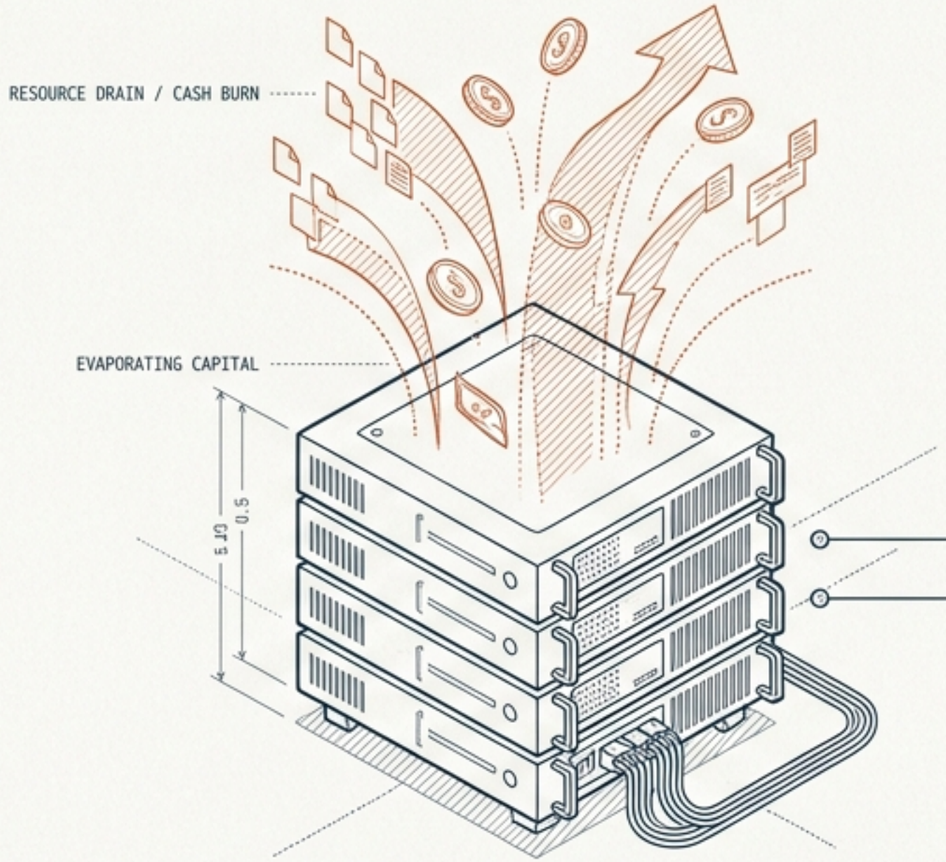


Mio 的账到底算不算得过来：AI 伴侣经济模型的终极重构

从架构底层拆解单位经济学，证明高毛利 AI 软件模式的可行性

现实的骨感：两周烧出的生存危机



在构建 OpenClaw 早期原型时，我们面临了全行业共通的痛点：**高得离谱的日成本**。单个用户**两周**的交互就能烧掉一笔的惊人的**算力费用**。

DAILY COST
TWO WEEKS
COMPUTE EXPENSE

传统思维将这归咎于“**模型太贵**”，期望**等待外部降价**。

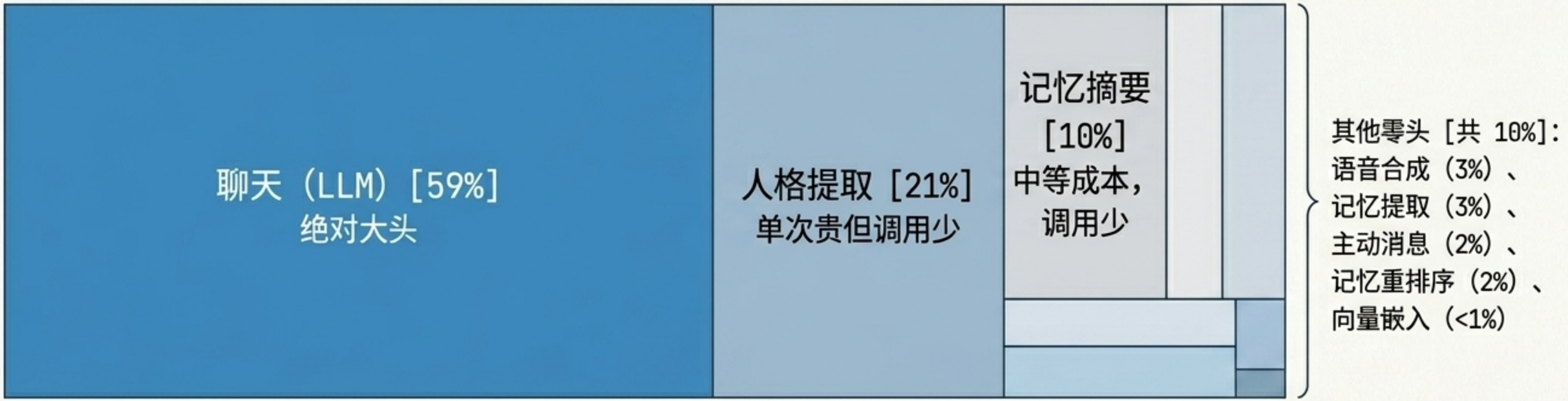
MODEL TOO EXPENSIVE
WAITING FOR EXTERNAL PRICE REDUCTION

这种量级的消耗逼迫我们做出决定：不能等“以后再优化”。从**第一天**起，**成本控制**就必须成为**核心架构设计**的一部分。

FROM DAY ONE
COST CONTROL
CORE ARCHITECTURAL DESIGN

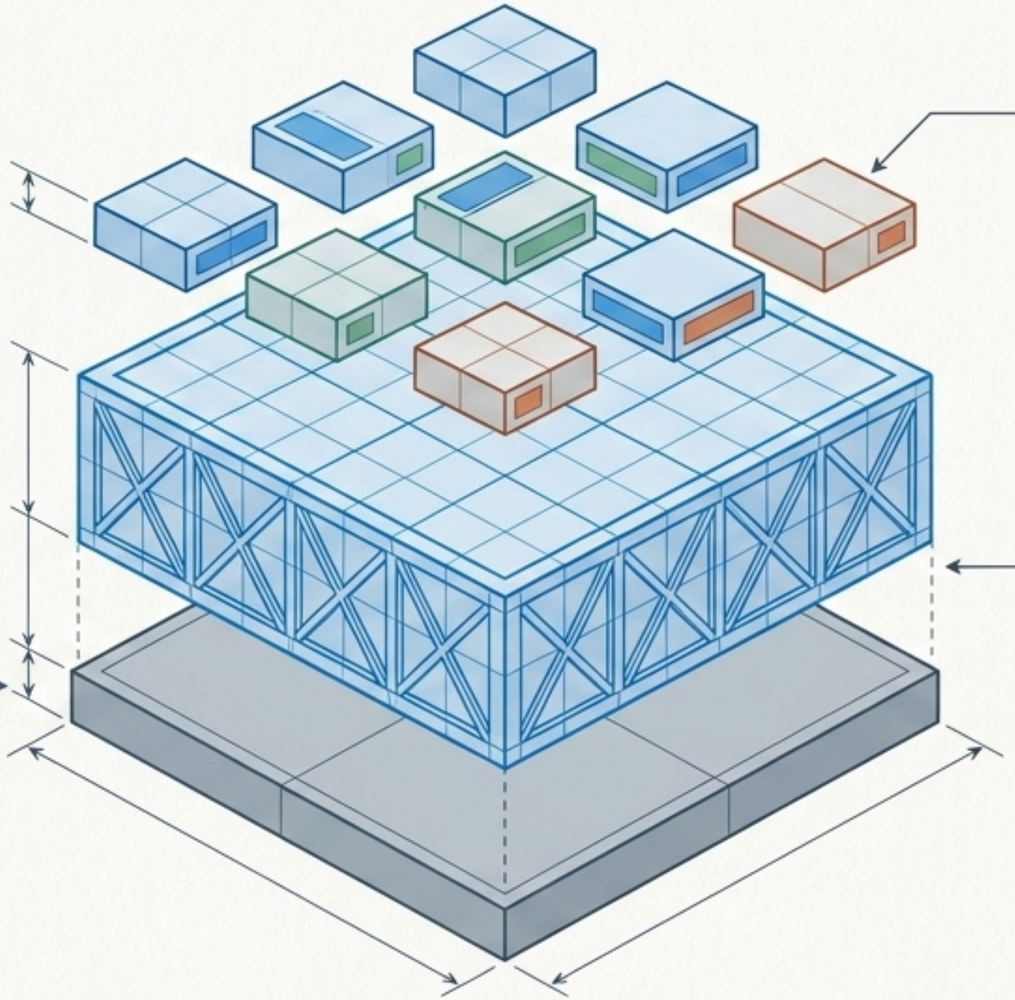
拆解真实账单：生产环境的数据真相

经过八个架构版本的迭代，线上真实环境（77 次调用，~28 条消息）的成本结构的发生了根本改变。当前一个活跃用户的日总成本已极低。



平均每条消息的成本已**降低两个数量级**，几乎可以忽略不计——但这对于实现低价层级盈利而言，依然不够。

成本的物理结构：固定与变动的博弈



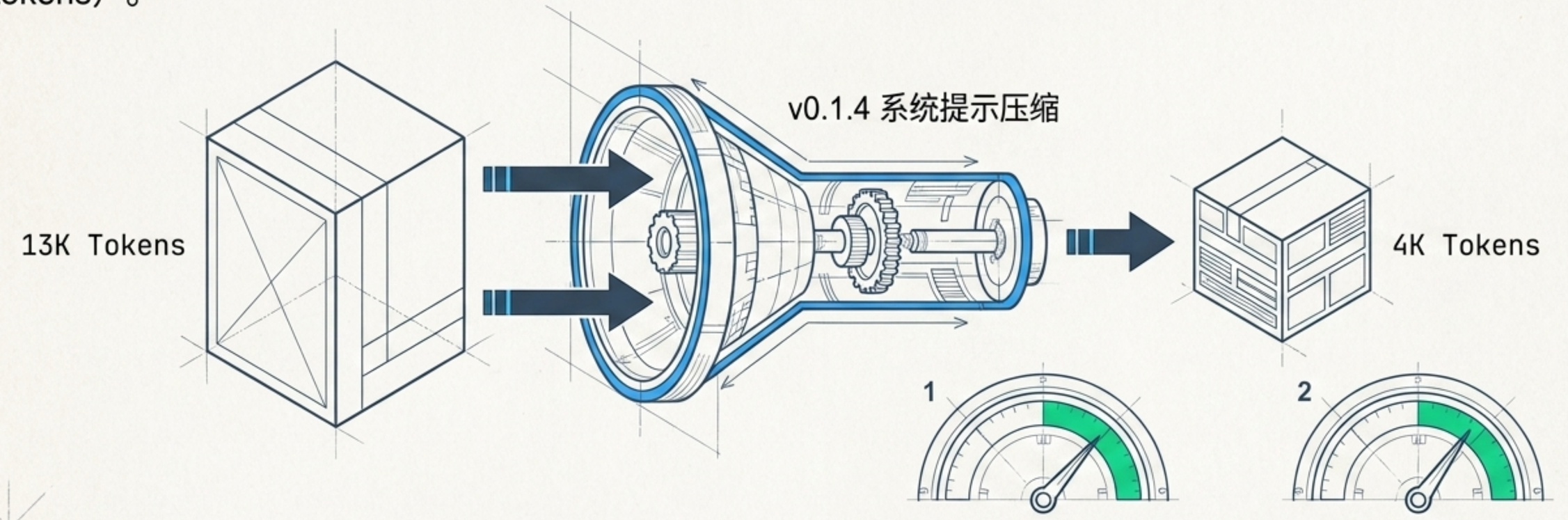
媒体成本（按需叠加）：
图片理解/自拍生成/语音合成成本极低；仅视频理解成本略高。

每条消息成本：业务规模的决定因素。聊天（LLM）占据核心（每次 8K-17K 输入 tokens）。
- 30条/天：月度账单微小
- 100条/天：变动成本反超固定成本
- 200-300条/天：变动成本达到固定开销的数倍

每日固定开销：占比不大。主要包括人格提取（3次/天，Gemini 3.1 Pro）与记忆摘要（2次/天）。
其他后台提取任务约等于零。

核心突破口：v0.1.4 提示压缩引擎

系统提示是每次聊天输入 tokens 的最大头。我们将系统提示大幅压缩了 ~60%（从 9K-13K 降至 3K-5K tokens）。



压缩直接砍掉了成本的大头。越是高使用量的层级，由于变动成本占比更大，净利润的改善效果越具颠覆性。

越界转绿：全层级盈利的实现

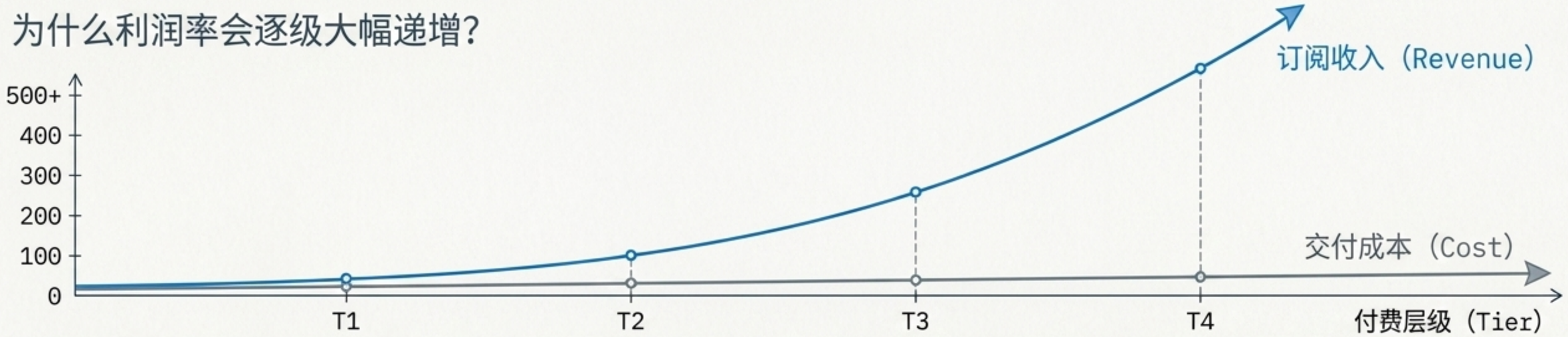
测算基于“满载毛利”（用户每天打满消息上限的最坏情况）。实际使用中，用户通常只消耗额度的40-60%，真实利润率远超表格预期。

层级	每日限制	压缩前	压缩后
免费	20/天	获客漏斗	获客漏斗
入门	30/天	● 【亏钱】入不敷出	● 【转正】稳妥盈利
进阶	100/天	● 【打平】勉强打平	● 【盈利】毛利可观
高级	200/天	● 【微利】小赚	● 【高利】毛利很健康
旗舰	300/天	● 【微利】小赚	● 【高利】毛利很健康

压缩前，仅高层级满载时能微利。压缩后，所有付费层级全部实现盈利转正。

商业模式升维：售卖体验溢价，而非算力溢价

为什么利润率会逐级大幅递增？



低层级逻辑：

低层级用户在为真正消耗算力成本的基础功能付费（LLM 聊天、语音识别）。

高层级逻辑（零成本特权）：

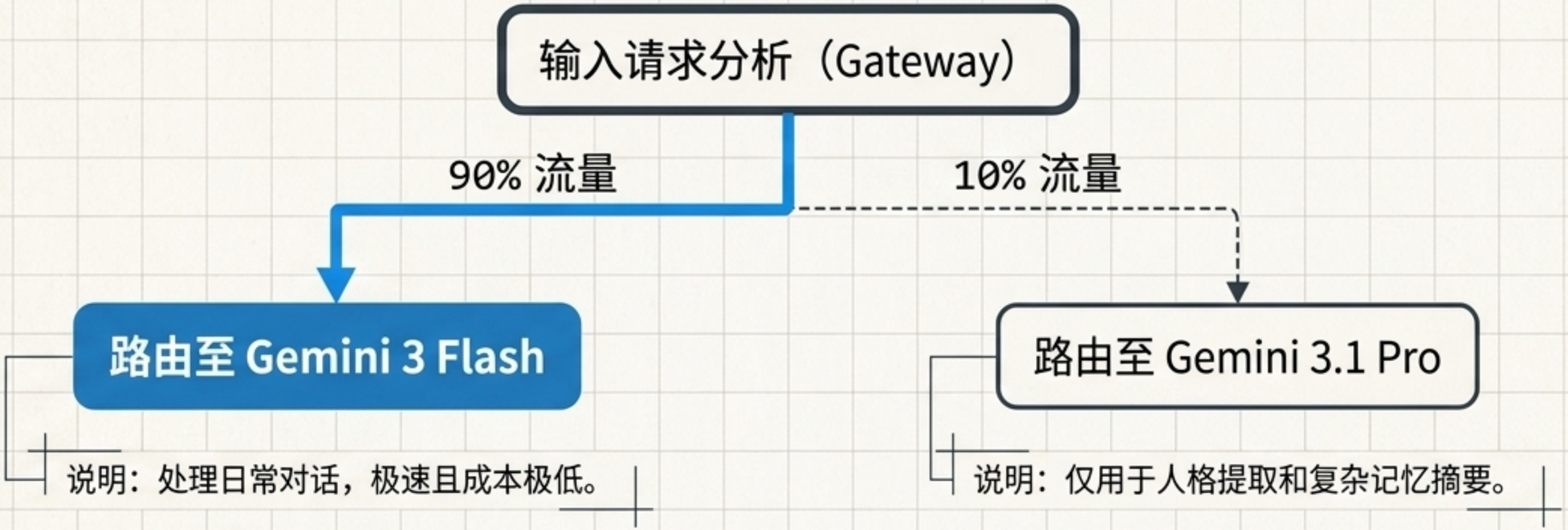
高层级用户支付高昂溢价购买的核心功能，其边际成本几乎为零：

[+] 优先响应（零成本）
[+] NSFW 解锁（零成本）
[+] 自拍生成（可忽略）
[+] 扩展记忆（可忽略）

SYNTHESIS STATEMENT

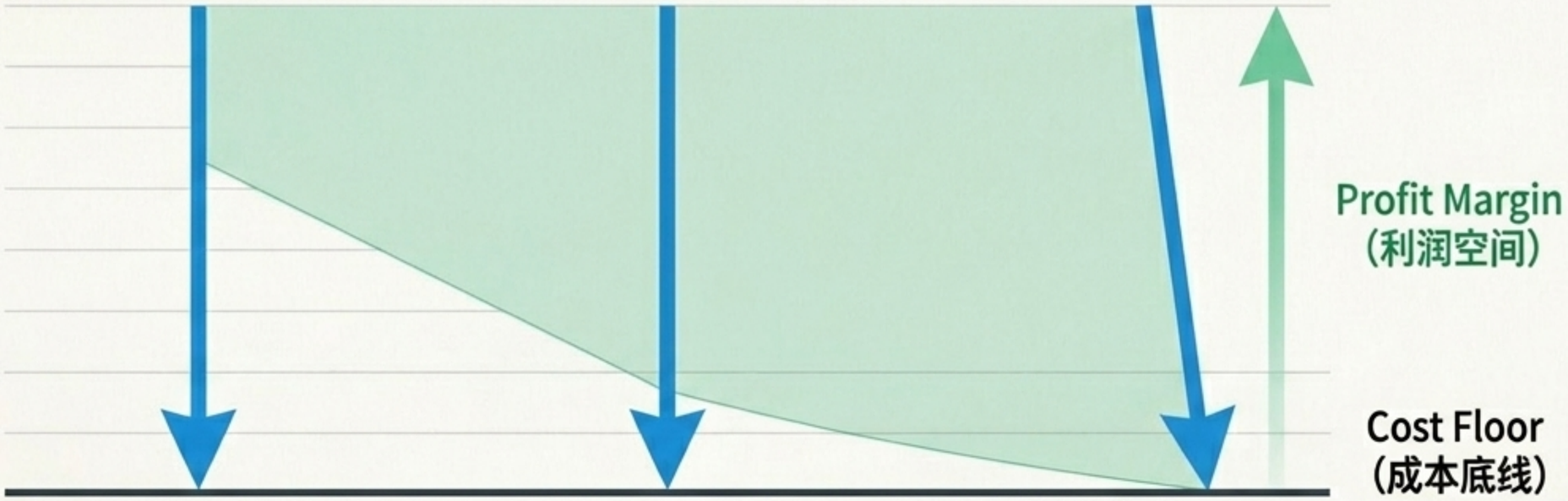
Mio 并非一个线性的算力转售商，而是一个拥有超高软件 SaaS 级边际利润的体验平台。

架构级护城河：智能路由调度



架构级优化的威力在于**“倍增效应”**。随着便宜模型的能力持续进化，越来越多的 Pro 级任务将被降级至 Flash 处理——**每一次轻量化切换，都会在所有用户的每一次交互上产生利润倍增**。

利润复利效应：为什么现在的毛利只是地板



Force 1：提示工程复合累积（内部）
未来的 lorebook 架构（按需注入背景故事而非全量加载）将把 Tokens 再降 30-40%。

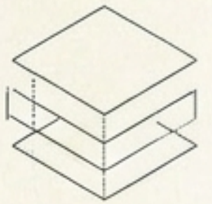
Force 2：模型成本断崖式下降（外部）
LLM 推理成本过去两年已降两个数量级，预计一年内单条消息成本将再降 3-5 倍。

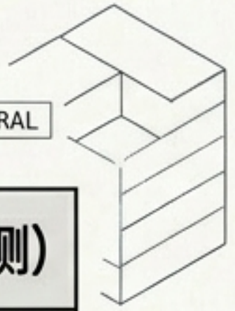
Force 3：架构级降维倍增（内部）
便宜模型能力迅速提升，智能路由将进一步把高负载任务下放至基础模型。

SYNTHESIS STATEMENT

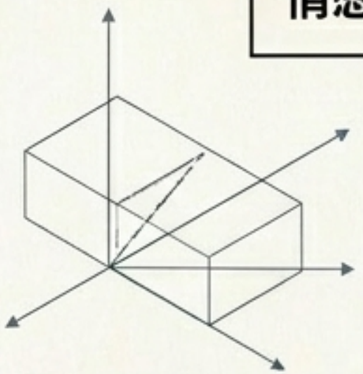
预期在 6-12 个月内，这三股力量的结合将把所有付费层级推向 50%-70%+ 的净毛利区间。

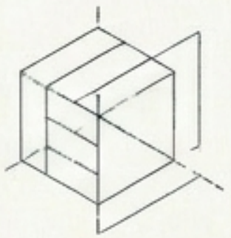
进化全景图：从验证原型到高毛利引擎





	早期原型	Mio（压缩前）	Mio（压缩后）	Mio（12个月预测）
每用户每天成本	天文数字	降了两个数量级	再降一大截	只剩零头
每条消息成本	高得离谱	几乎可忽略	再降 ~35%	再降 3-5倍
入门层级盈利状态	[不能]	[仅高层级]	[能 - 稳妥盈利]	[能 - 高毛利]
记忆管理	无(只加不减)	多层检索引擎	+ 压缩提示	+ 自优化
情感细腻度	基于规则	灵魂驱动	+ 关系进化	+ 微调模型





成本从天文数字降至可忽略水平，而体验的情感细腻度却呈指数级上升。战略方向已彻底验证。