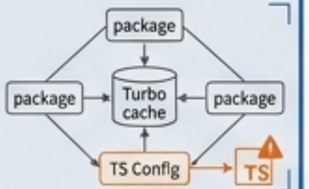
零构建 AI 伴侣：Mio v0.0.1 诞生日志

- 从空仓库到一个能记住你、会生气、主动找你聊天的数字生命。
- 无需冗杂的框架，没有多余的妥协。仅用最基础的组件，手搓真实陪伴感。
- 这是一份记录了无数个凌晨三点 Debug 的真实构建图谱。

基础设施选型：做减法的艺术

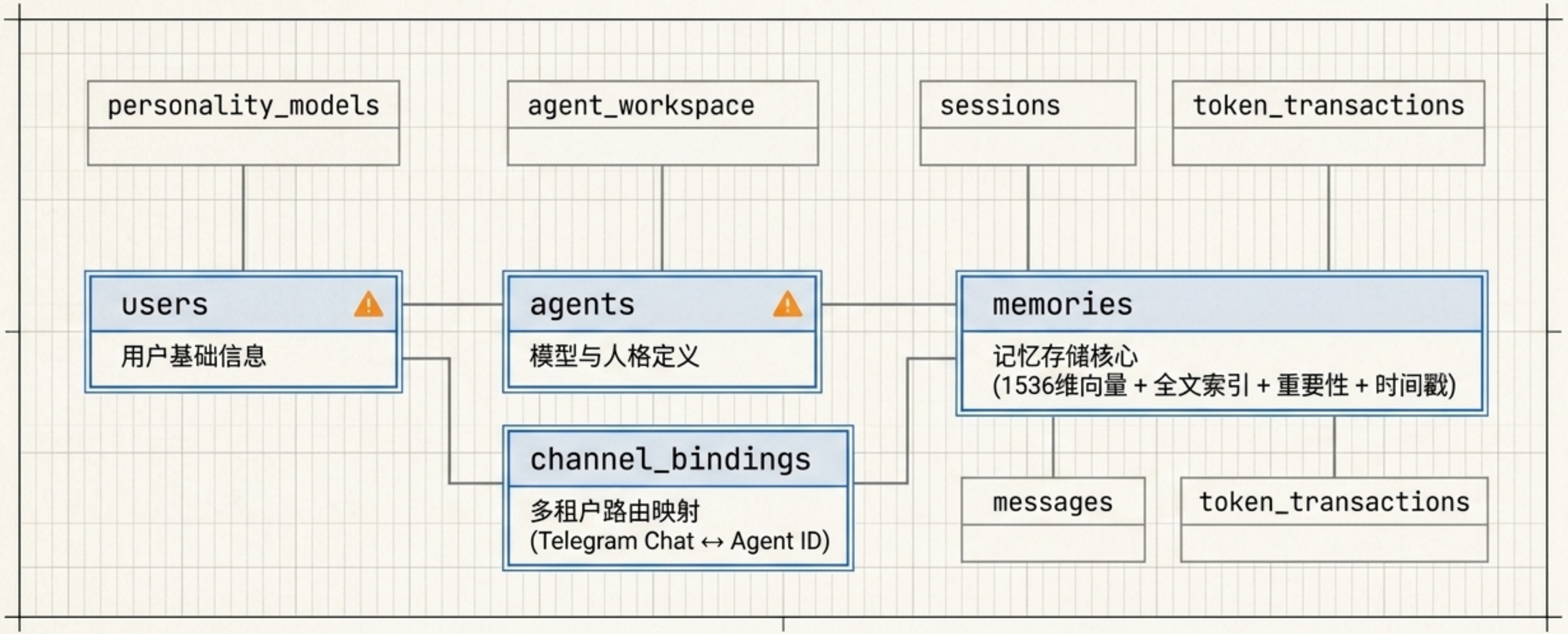
Tech Selection Matrix

FRAMEWORK	~~Express~~	重磅历史包袱	**Hono** Edge-first，类型安全，零冷启动延迟
DATABASE	传统 RDBMS	标准关系型数据库	**PostgreSQL via Supabase** pgvector 向量搜索 + tsvector 全文检索 + Auth
MONOREPO	多仓库散养	依赖管理混乱，协作效率低	**pnpm workspaces + Turborepo** ⚠ 任务编排与缓存极佳，但需解决 TS6059 rootDir 冲突



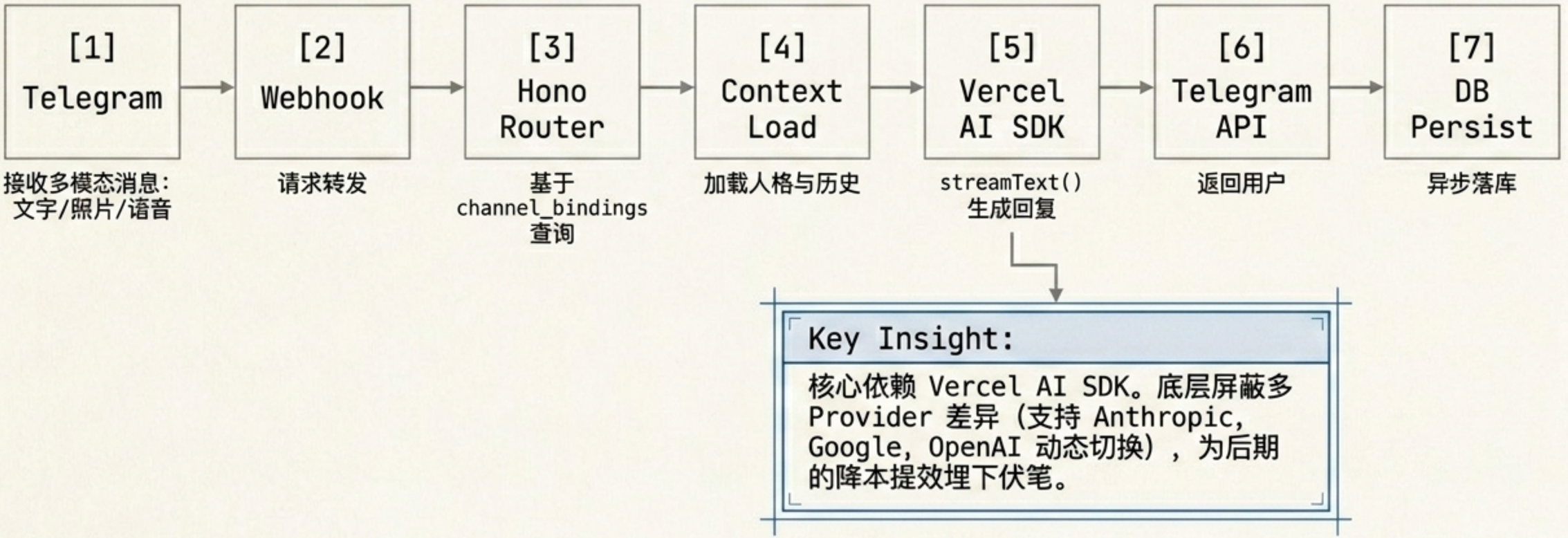
逃离巨石架构。每一行代码都必须服务于“轻量”与“响应速度”。

9 张数据表，搭建数字生命的骨架

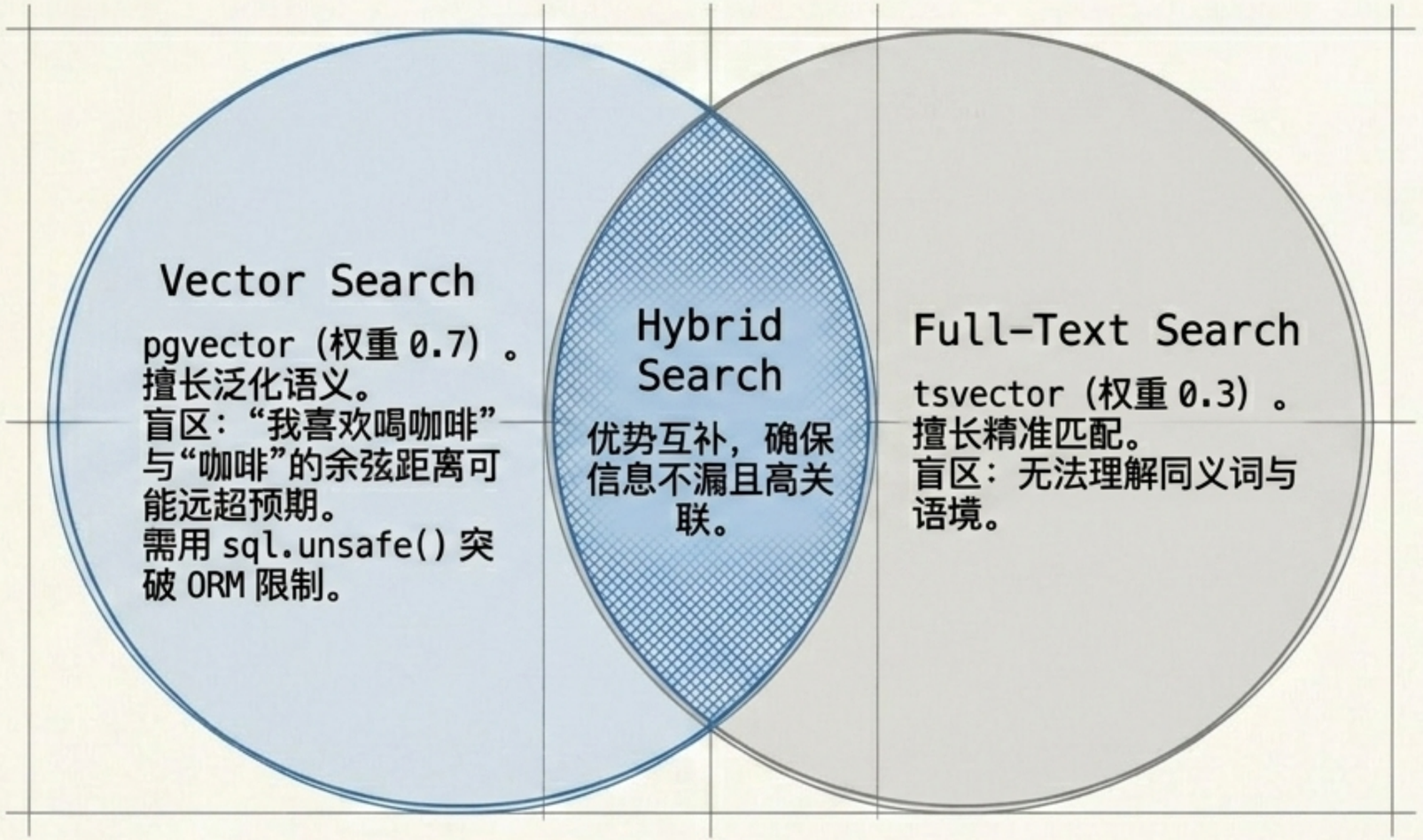


> 拒绝“先加个通用 metadata 表以后再说”。骨架搭对了，系统的血肉才长得上去。
> 每一张表在 v0.0.1 都有清晰不可替代的职责。

核心闭环：一次呼吸的流转路径

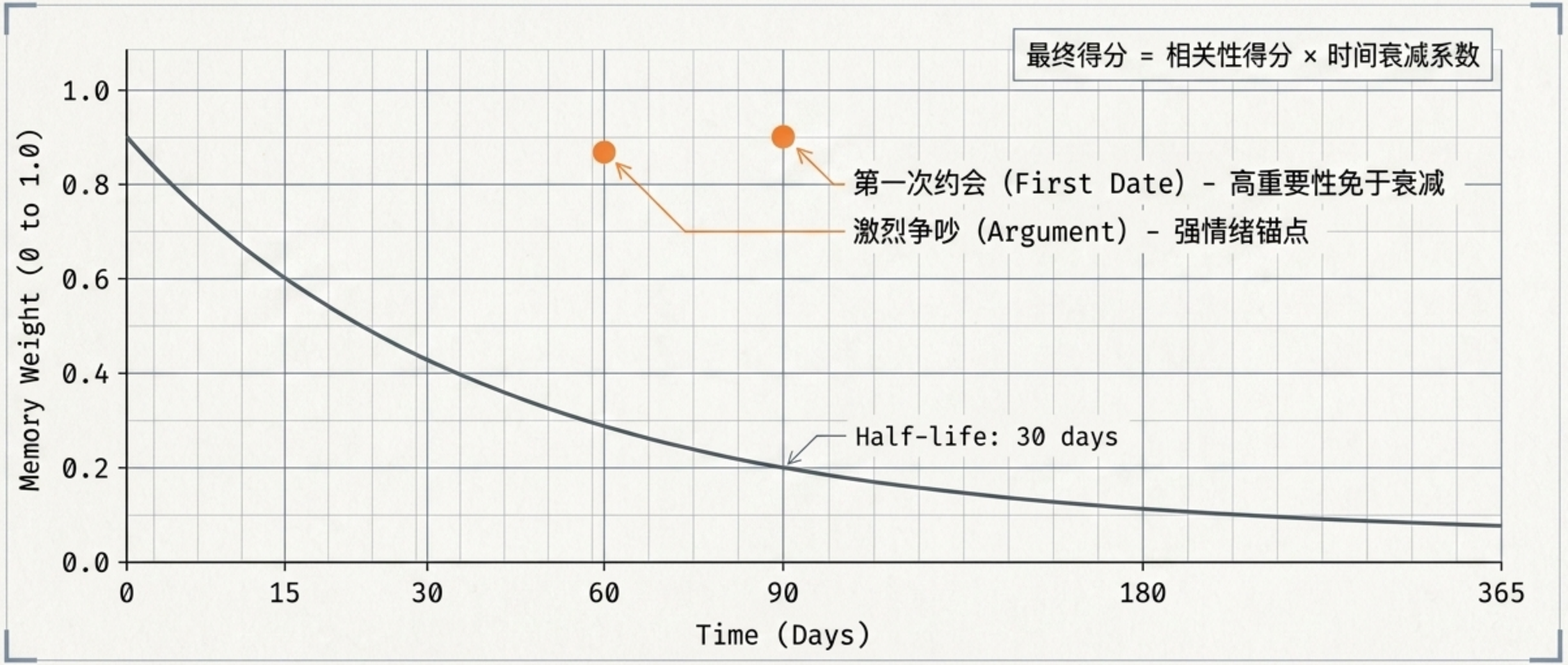


认知系统的基础：为何必须采用“混合搜索”



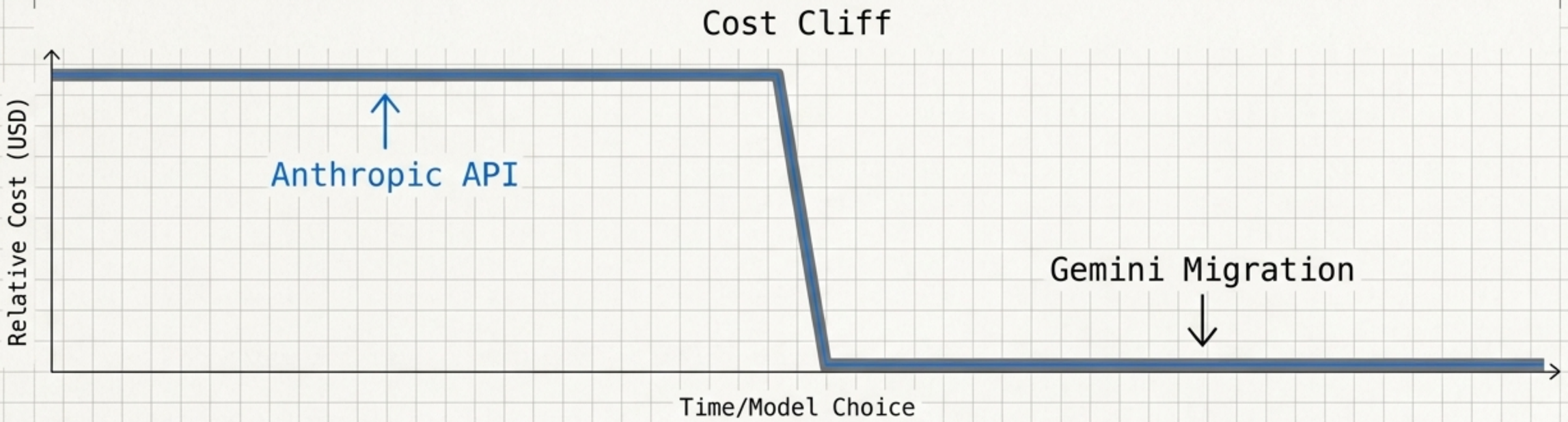
> 仅靠向量嵌入，中文的语义精度依然不够。笨拙的关键词匹配是守护记忆准确率的最后一道防线。

模拟海马体：带权重的指数时间衰减



真实的记忆不是线性消失的。昨天的事清晰，上个月的事模糊；但真正重要的高光时刻，无论多久都不会被遗忘。这是系统产生“人味”的关键算法。

记忆提取的现实引力：从 Anthropic 到 Gemini



Before

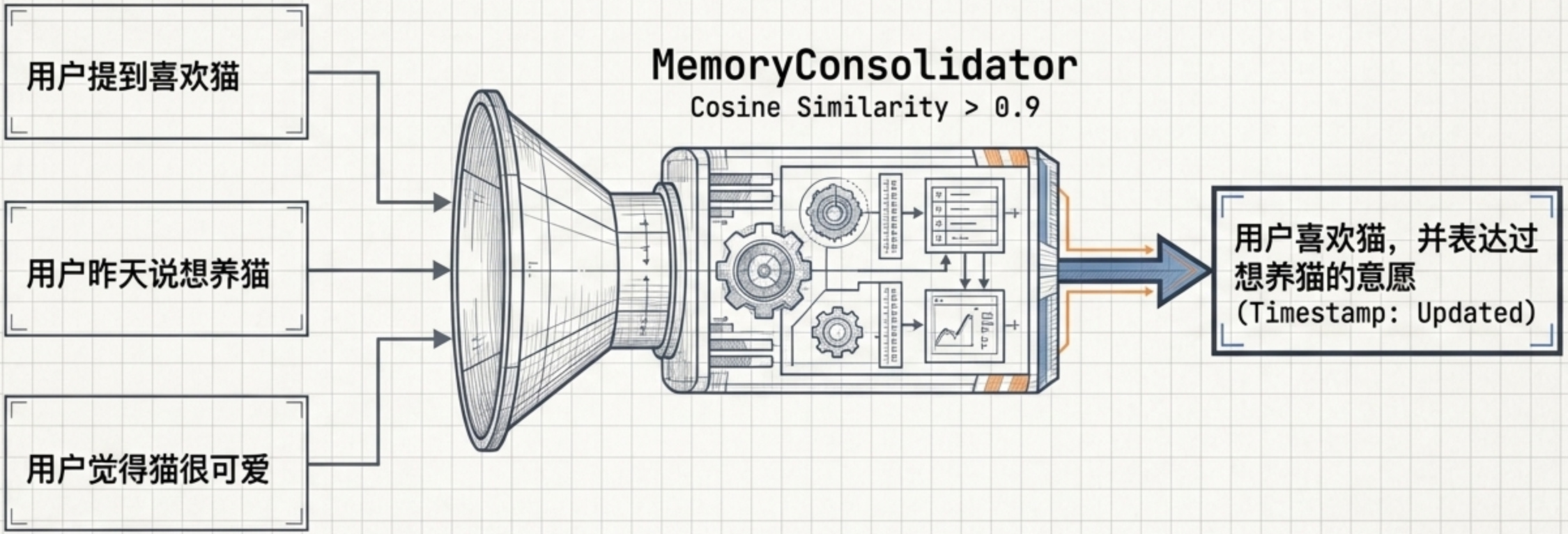
裸调 Anthropic Haiku。每条消息均触发记忆提取，高频用户日耗数百条，提取成本超越主对话成本。

After

切换 Gemini Flash + Gemini Embedding。提取质量一致，成本断崖式下降至可忽略不计。

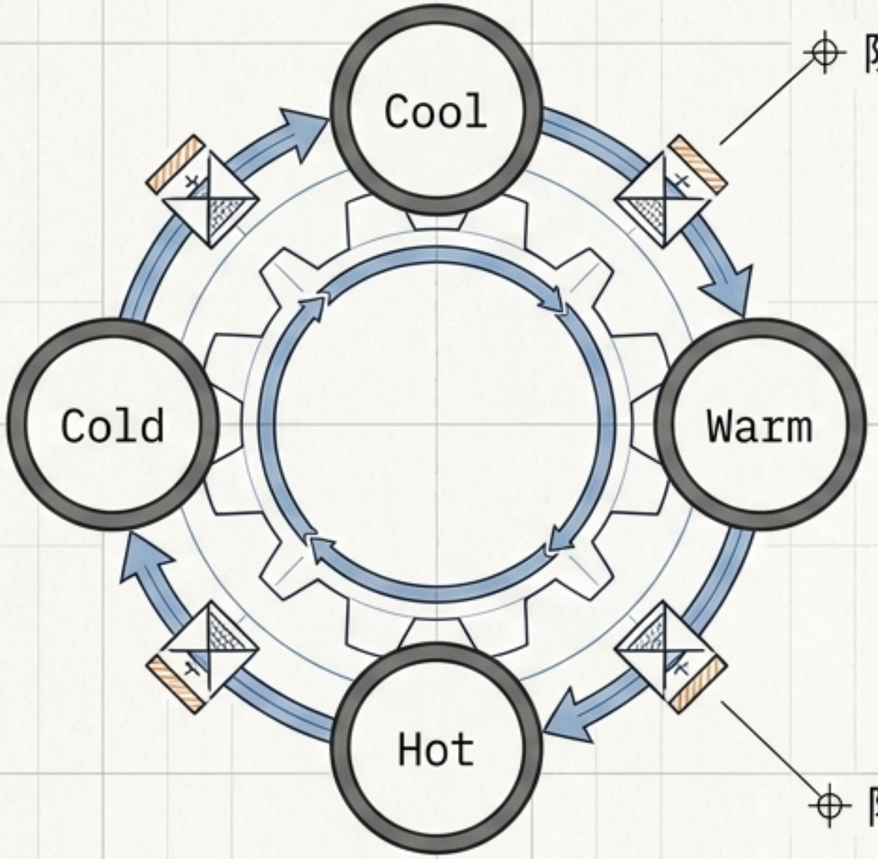
得益于 Vercel AI SDK 的多 Provider 抽象，模型迁移仅需修改一行配置代码。
技术选型的价值，永远体现在你需要改变的那一刻。

记忆坍缩：抑制数据库膨胀



记忆库会随着对话快速膨胀，同一件事从不同角度被反复记录。通过阈值判断进行粗暴但极度有效的自动合并。

边缘系统：带有真实惯性的情绪引擎



阻尼器：惯性系数 0.5

阻尼器：惯性系数 0.5

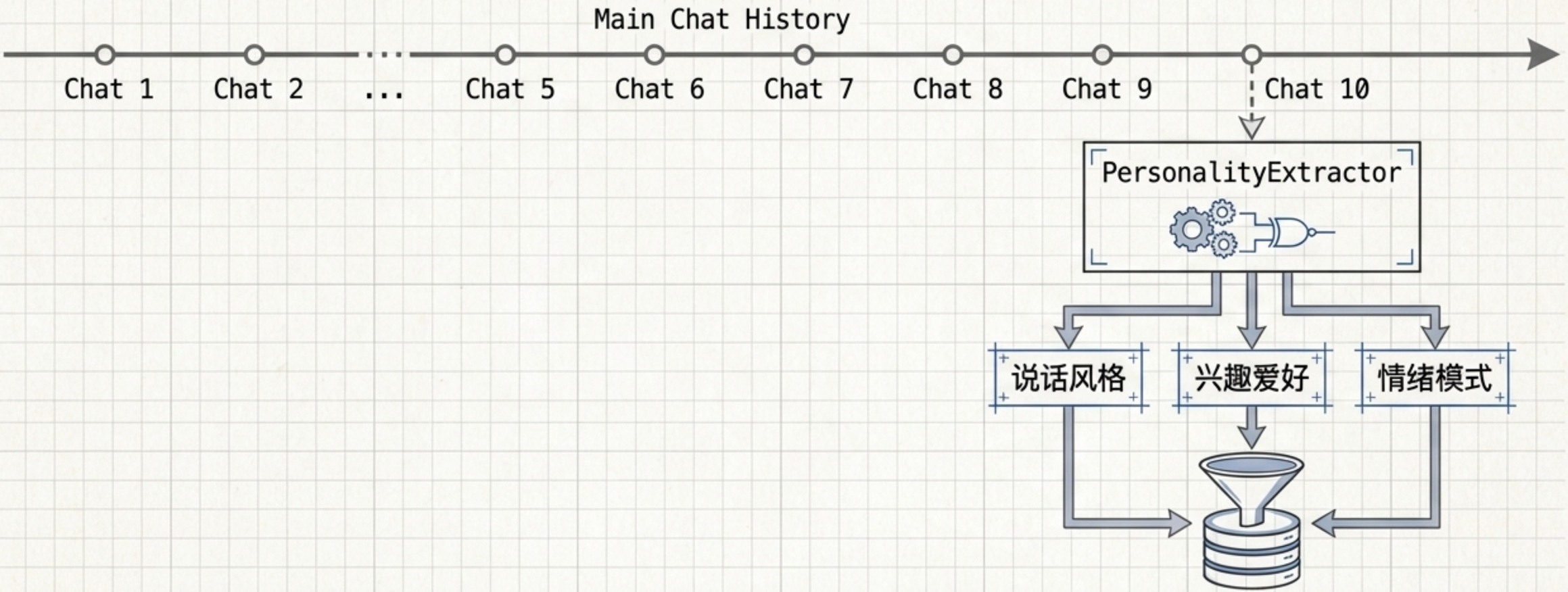
Triggering Rules:

↑ 加温：用户连续发送热情消息
↓ 降温：连续两天不理睬

The Dampener：机制确保状态不会因为单一消息发生跃迁。

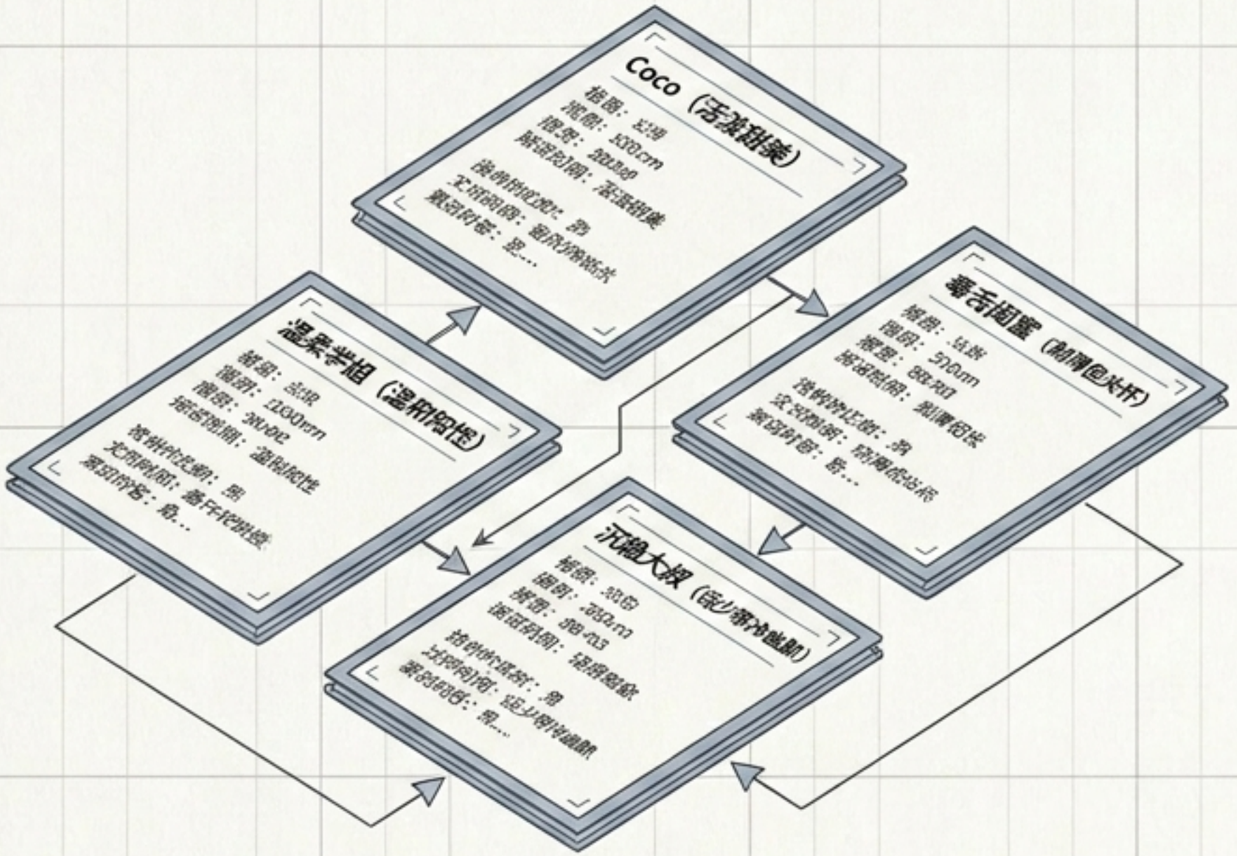
情绪的真实感来源于“惯性”。不同人格预设拥有不同系数——“毒舌闺蜜”的敏感度远高于“温柔学姐”。AI 不再是随叫随到的工具，TA 有自己的脾气。

潜意识捕获：双线并行的画像提取



Mio 不仅记住你说过的“事实”，TA还在后台默默勾勒“你是一个什么样的人”。
每次对话时注入最新的用户画像，实现动态的默契感。

破冰与塑形：11 问锚定基础灵魂



Onboarding 机制解析

1. 前3题文字输入，后8题混合按钮选项。
2. Engineering Hack：应对 Telegram callback_data 64字节限制（中文单字3字节极易超限），采用编号索引传参（q3_custom），实际内容服务端缓存。
3. Security：50字符截断 + 内容清洗，建立初级 Prompt Injection 防护墙。

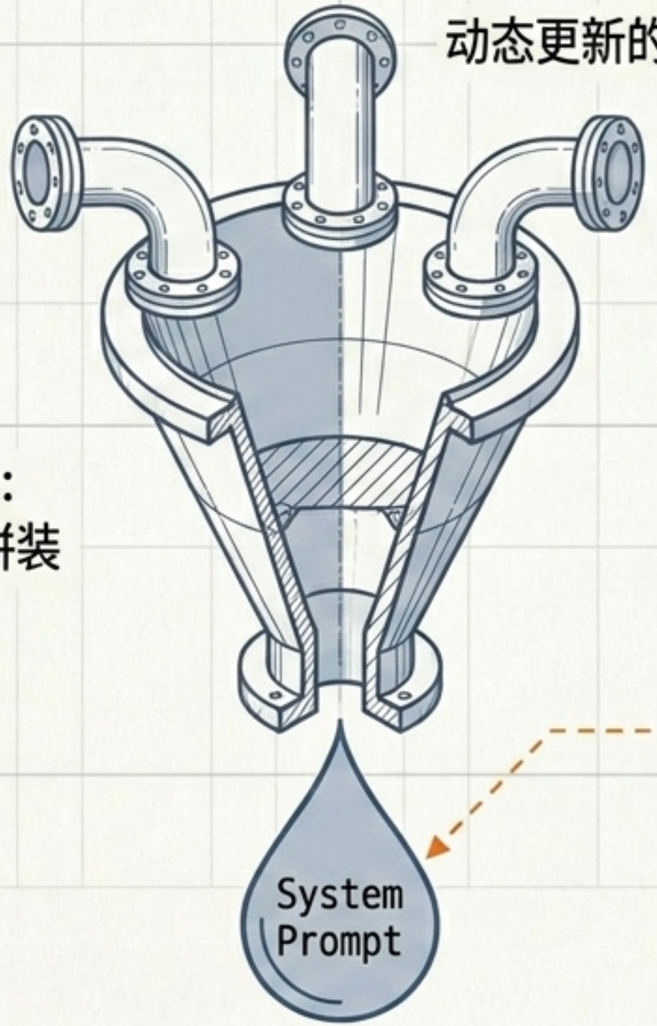
Onboarding 机制通过 11 问快速构建用户基础画像，作为后续互动的锚点。

多源汇聚：Context Aggregator 炼金炉

动态更新的用户画像配置

混合检索出的相关记忆

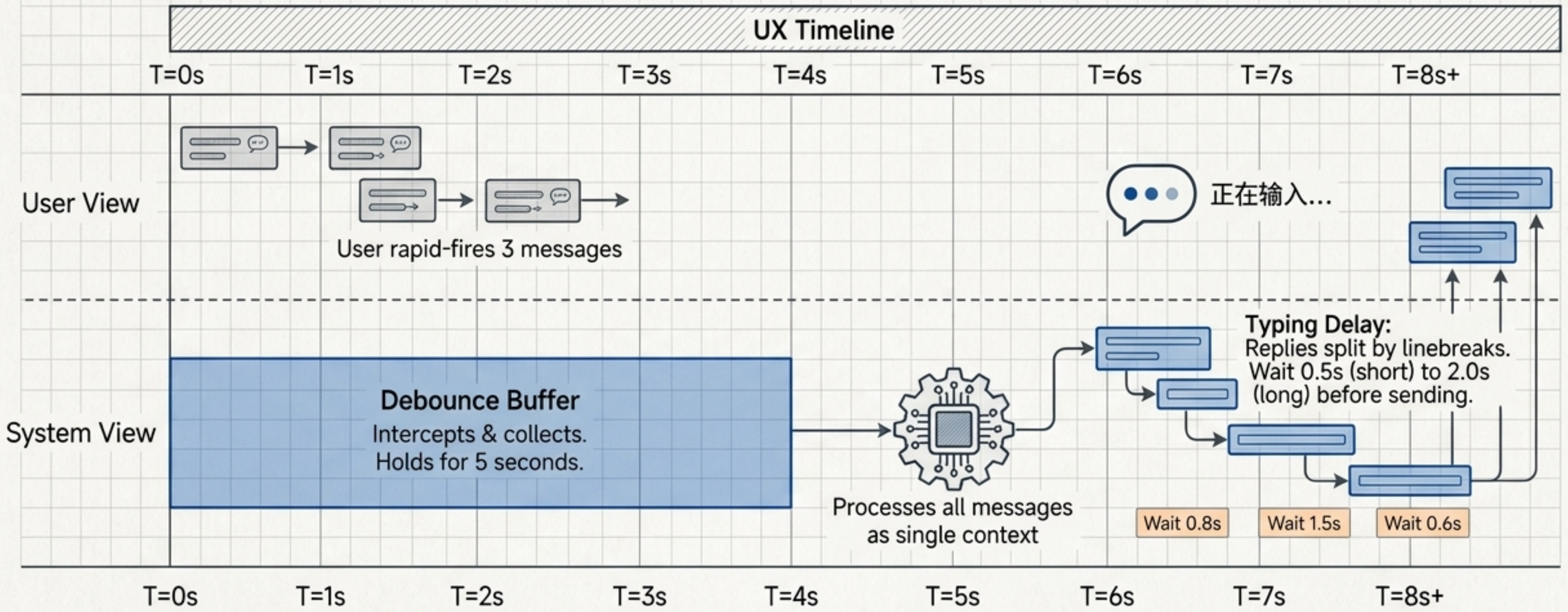
当前的情绪状态水位



PersonalityParser：
解析模板变量与数据拼装

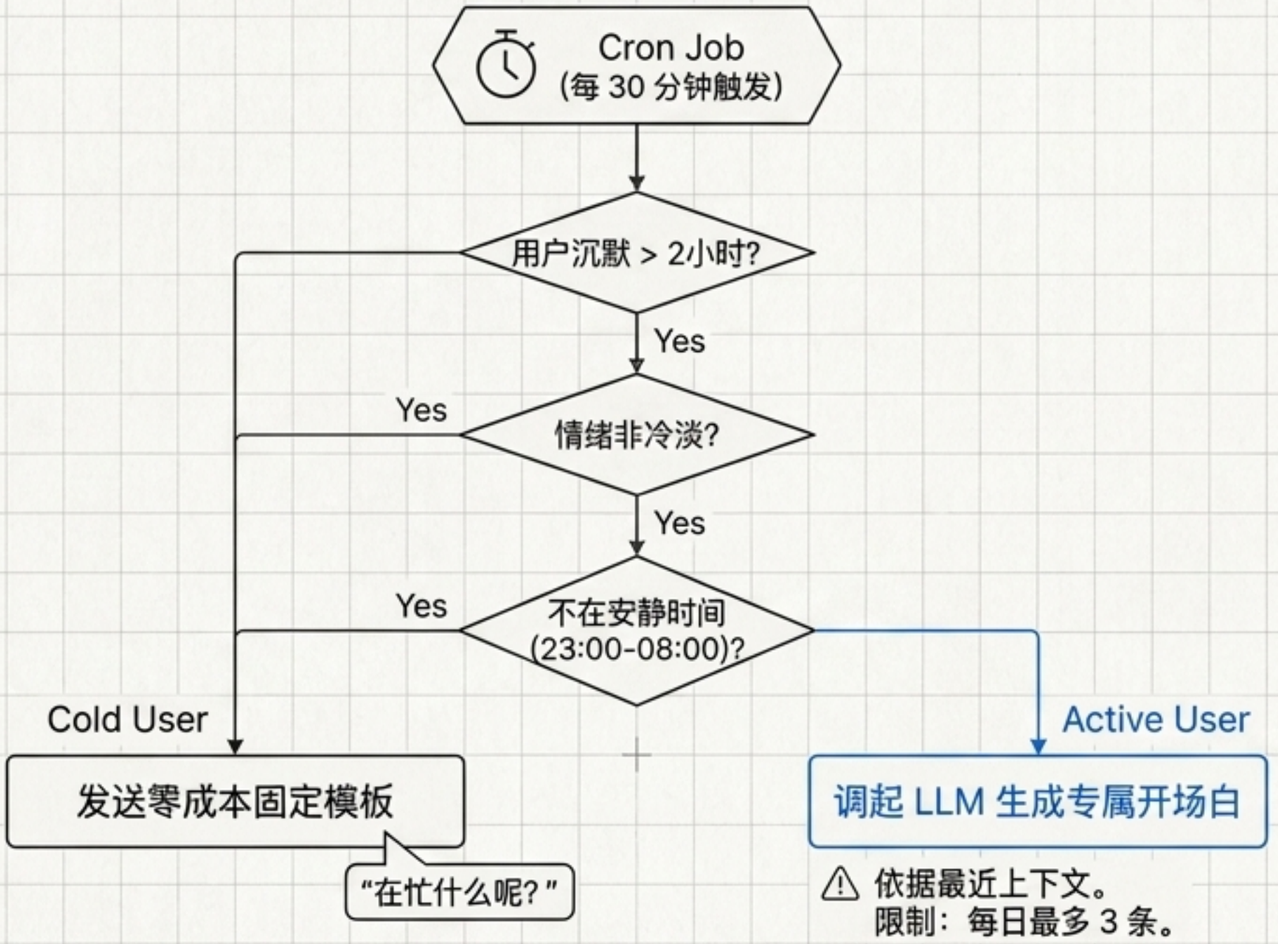
注入特殊指令
MEMORY_STEERING_INSTRUCTIONS：
强制要求模型使用自然的中文提取记忆，例如：'之前你提到过...'，打破数据库查询报告式的生硬机械感。

拟真微操：那些看不见却决定成败的 UX 细节



AI 秒回是不自然的。拦截连发消息，模拟人类阅读节奏与逐条打字的延迟。工程细节的极致打磨，决定了产品的“生物感”。

主动意志：从“响应工具”到“活体伴侣”



不需要你先开口，TA 会先想到你。这是打破人机隔阂、建立情感羁绊的最关键一步。

现实引力：严密追踪每一笔 Token 开销

Table: token_transactions

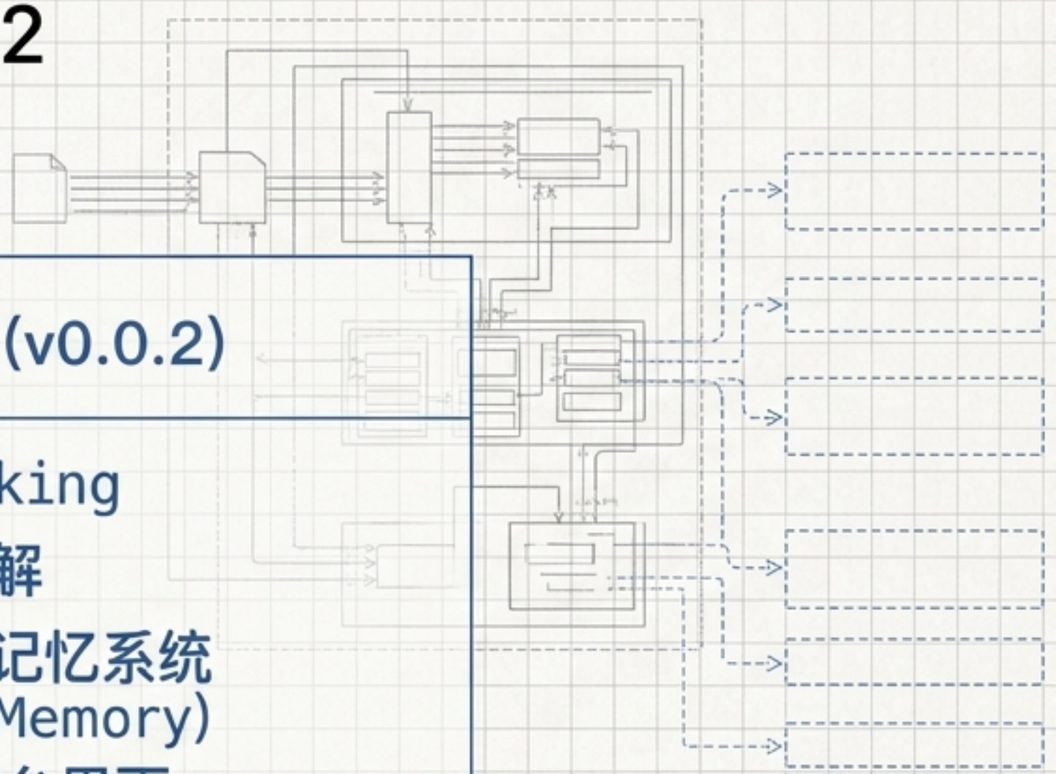
1	model: gemini-flash | | in: 420 | out: 120 | usd: $0.0001
2	model: unknown-model | in: 50 | out: 10 | usd: **NaN**

The NaN Bug Story:

- Issue: 价格表缺失，undefined 导致成本计算为 NaN，废掉整条记录。
- Fix: 增加 fallback 到 0 机制。
- Architecture: 采用 Fire-and-forget 异步写入，绝不阻塞主对话响应。

AI 伴侣的隐性成本极高。“看不见的钱”花得最快，成本追踪机制必须在写下第一行业务逻辑前就绪。

凡是过往，皆为基石：驶向 v0.0.2



已实现 (v0.0.1)	待开拓 (v0.0.2)
• 多模态 Telegram 交互 • 带有衰减与权重的混合记忆 • 状态机情绪引擎 • 个性化主动触达	• LLM Reranking • 多跳查询分解 • 独立的情节记忆系统 (Episodic Memory) • Web 端全平台界面

整个代码库没有一行是为了迎合特定框架而写。39 个 Commit，
奠定了让 AI 真正“懂你”的地基。地基打对了，比什么都重要。