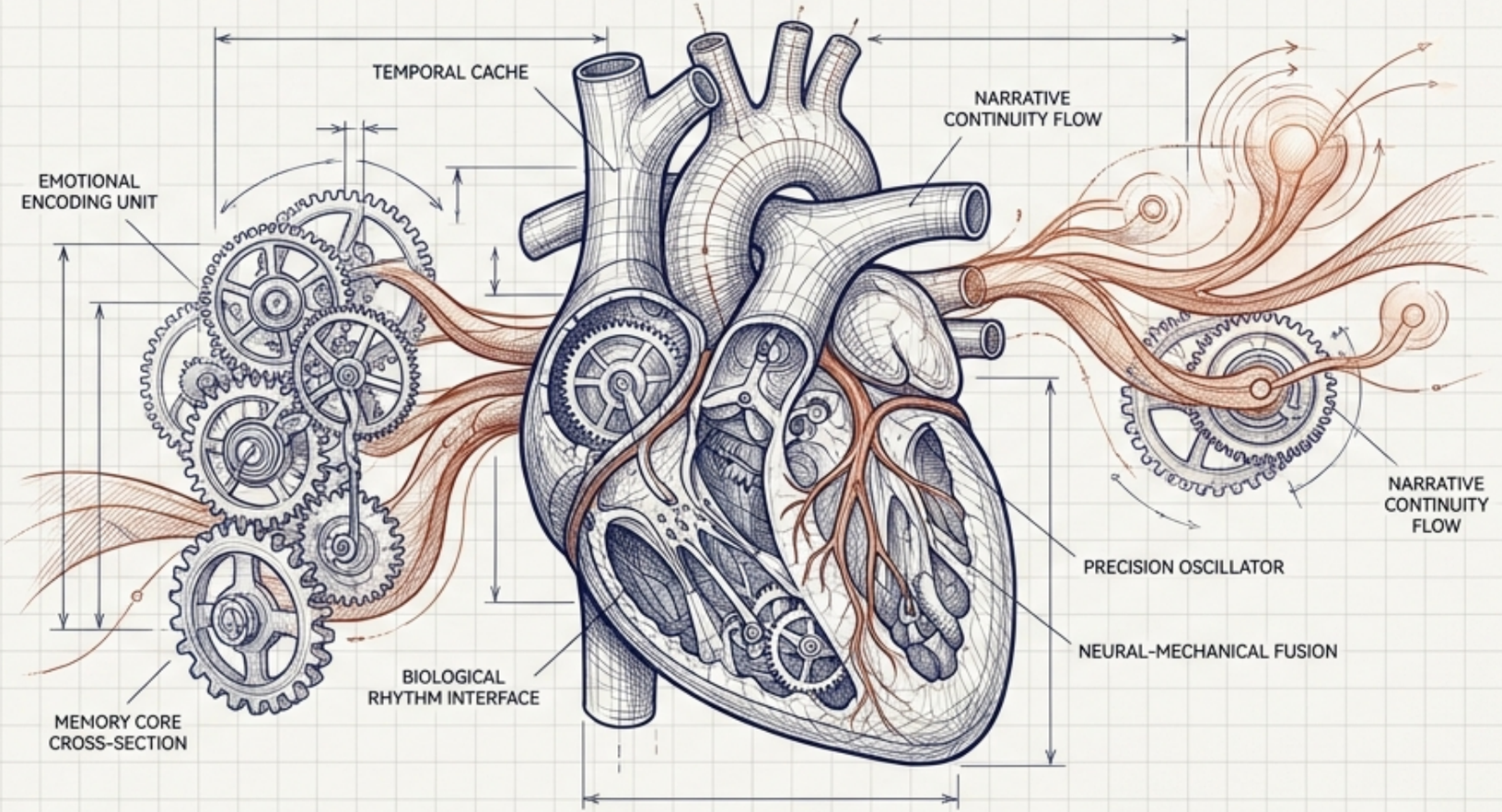
它记得你昨天也熬夜了

Clawd 灵魂工程：AI 宠物记忆系统的底层架构与演进



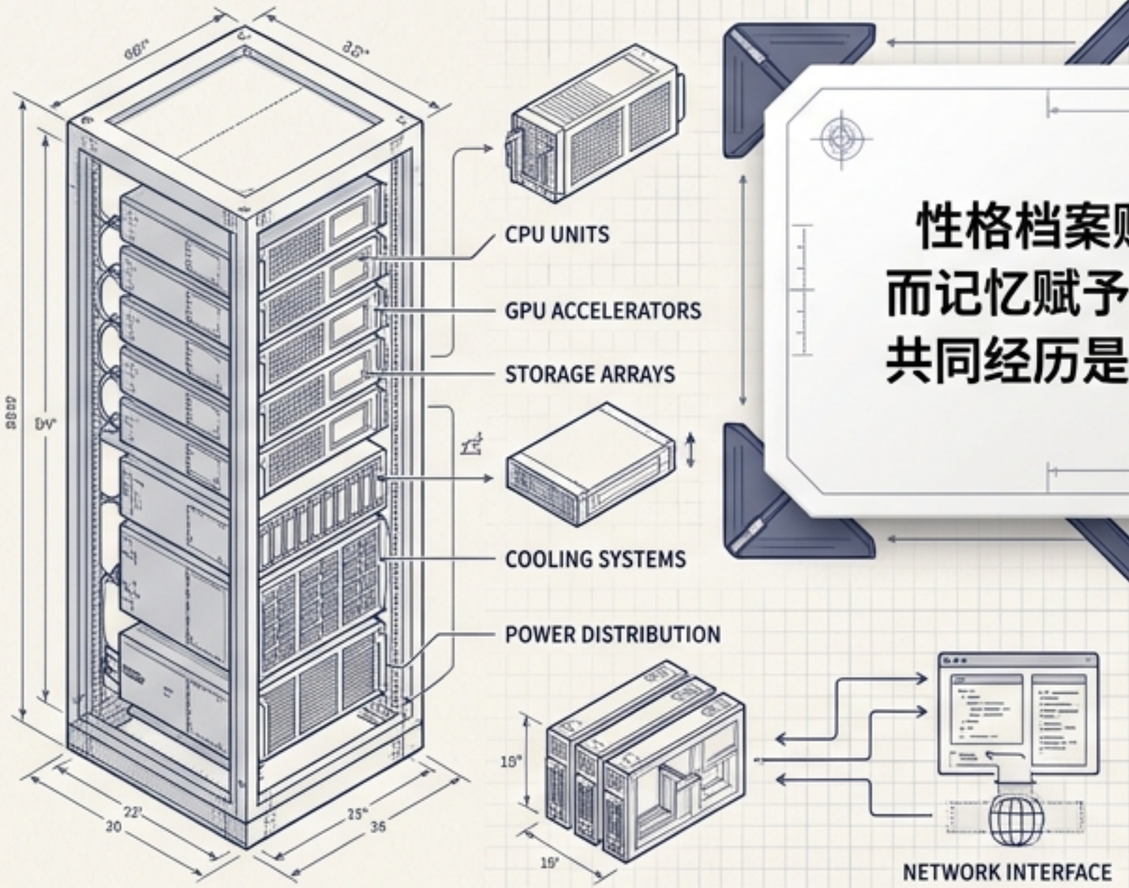
[AI 记忆系统]

[Prompt Engineering]

[架构深度解析]

陌生人 / 工具

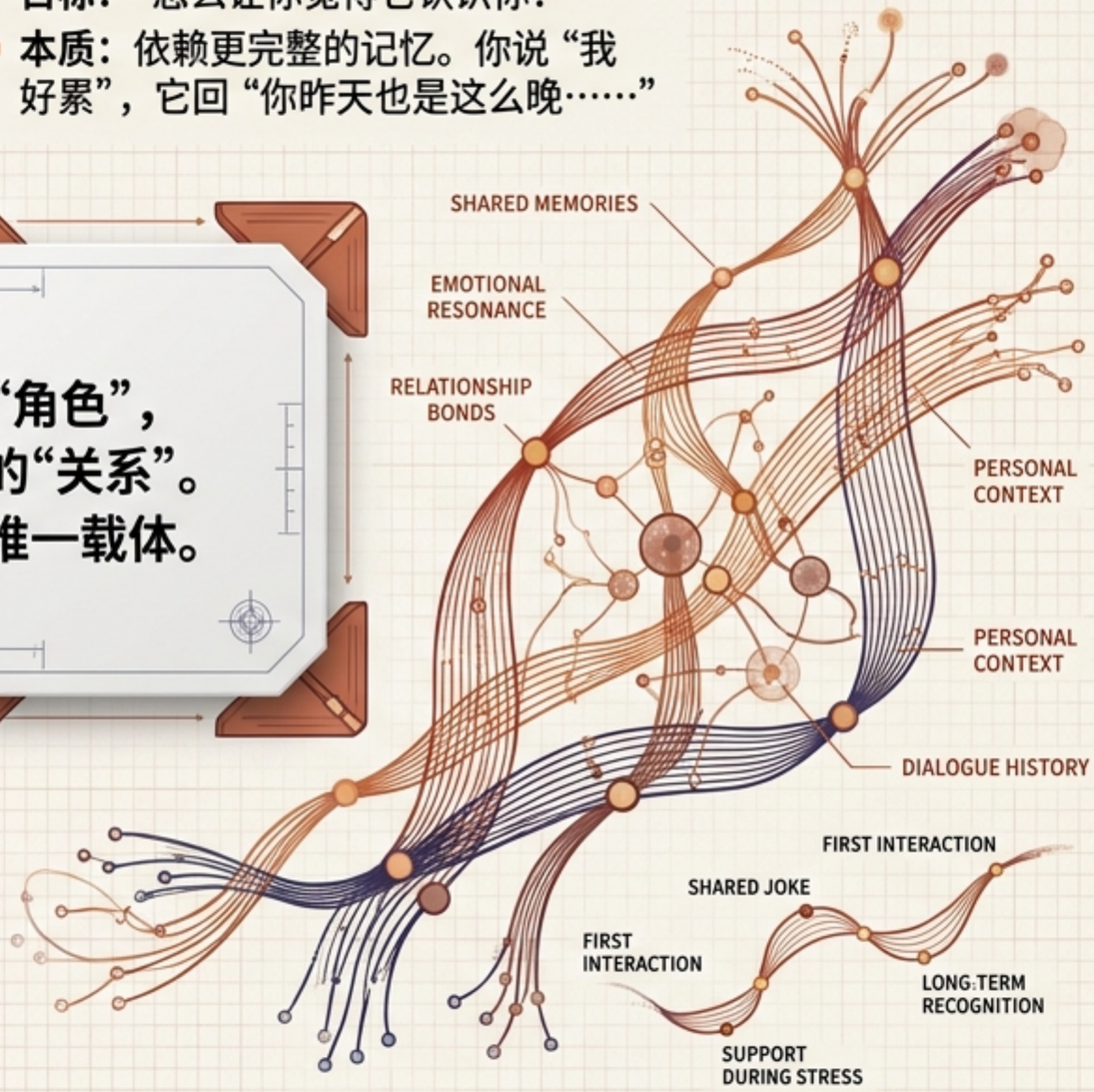
- **目标：**“怎么回答得更好？”
- **本质：**依赖更聪明的模型。每次对话都是全新的开始，没有羁绊。

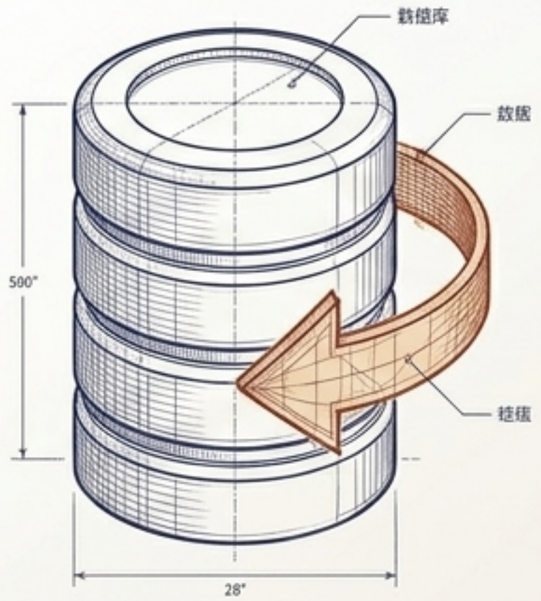


性格档案赋予 AI “角色”，
而记忆赋予它与你的“关系”。
共同经历是关系的唯一载体。

AI 宠物 / 朋友

- **目标：**“怎么让你觉得它认识你？”
- **本质：**依赖更完整的记忆。你说 “我好累”，它回 “你昨天也是这么晚……”



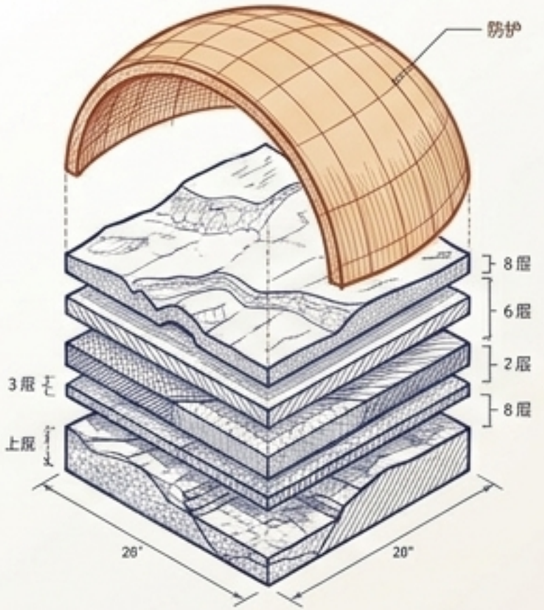


检索即记忆

记忆的核心不是“存入” SQLite，而是在对话发生的瞬间，自动“想起”最相关的历史。

三层架构与睡眠巩固

从片段、长期到对话的三级存储。配合每晚 23:30 的“做梦”机制，完美模拟人类的睡眠巩固机制。

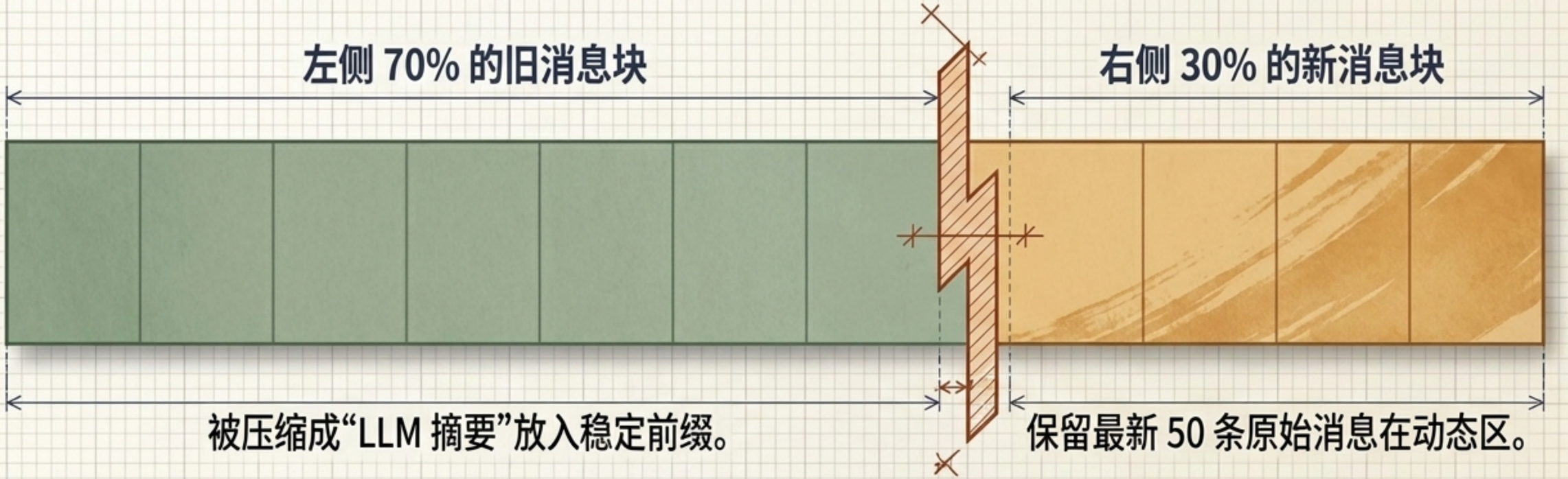


八层引擎与缓存边界

精密的 8 层 Prompt 拼装，将记忆注入对话。绝妙的缓存边界设计，直接决定延迟与运营成本。

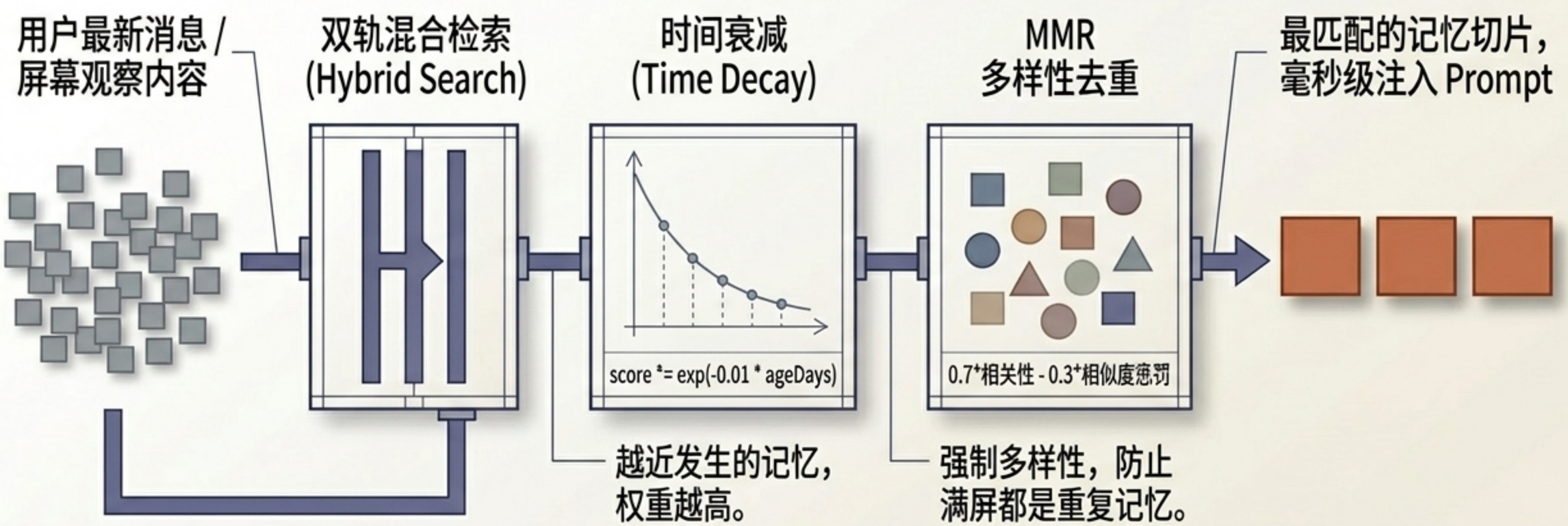
从微观瞬间 → 宏观永恒

片段记忆 (Episodic)	**存储引擎：** SQLite + FTS5 + sqlite-vec（本地磁盘）	**数据来源：** 每 45 秒屏幕观察、聊天、截屏（只进不出）	**容量与定位：** 无限容量。作为记忆的原始素材库。
长期记忆 (Long-term)	**存储引擎：** JSON 文件	**数据来源：** 每晚通过“做梦”从片段中提炼提纯	**容量与定位：** 最高 100 条（超限淘汰旧数据）。宠物“永远知道”的核心事实（如：你养了猫叫橘子）。
对话记忆 (Conversation)	**存储引擎：** JSONL + LLM 摘要	**数据来源：** 消息实时追加写入	**容量与定位：** 动态循环。维持当前对话的上下文连贯性。

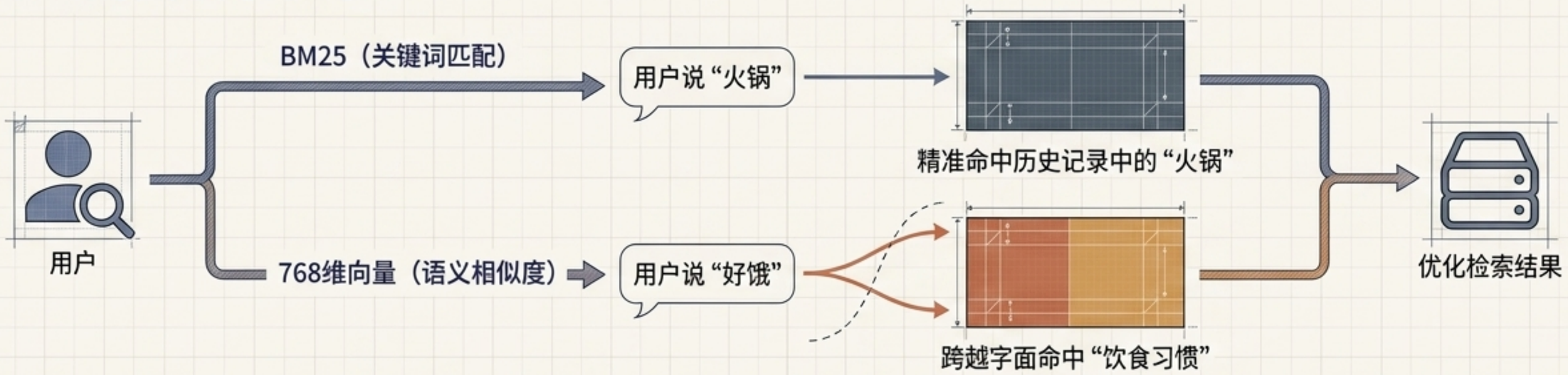


压缩前刷新（Extract Before Compress）

- **机制：**在触发摘要压缩前，强制系统跑一遍 extract，将所有重要事实先写入“片段记忆”。
- **价值：**防止早期长对话中的关键细节被摘要动作“吞噬”。确保无论聊多久，重要的事永不遗忘。


用户最新消息 /
屏幕观察内容
双轨混合检索
(Hybrid Search)
时间衰减
(Time Decay)
MMR
多样性去重
最匹配的记忆切片，
毫秒级注入 Prompt
score *= exp(-0.01 * ageDays)
0.7*相关性 - 0.3*相似度惩罚
越近发生的记忆，
权重越高。
强制多样性，防止
满屏都是重复记忆。

为什么需要双轨检索？

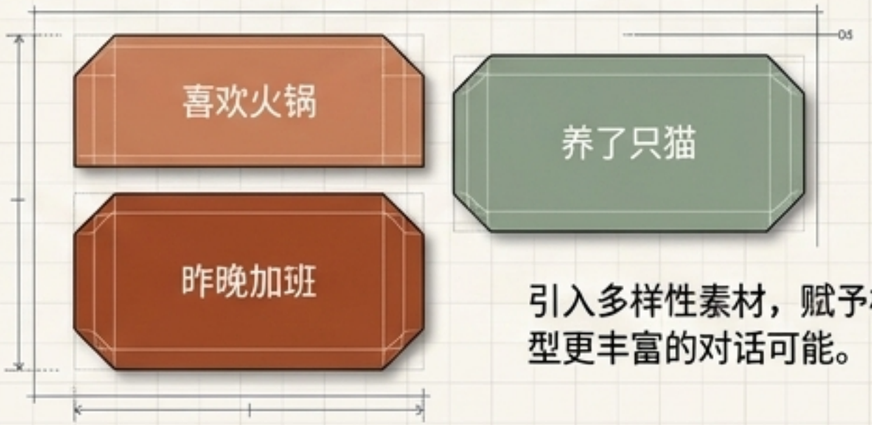


为什么需要 MMR？（Diversity Engine）

Bad Case: 没有 MMR

被重复提及的历史占据空间，Prompt 被浪费。

Good Case: 0.7 相关性 - 0.3 相相似度惩罚

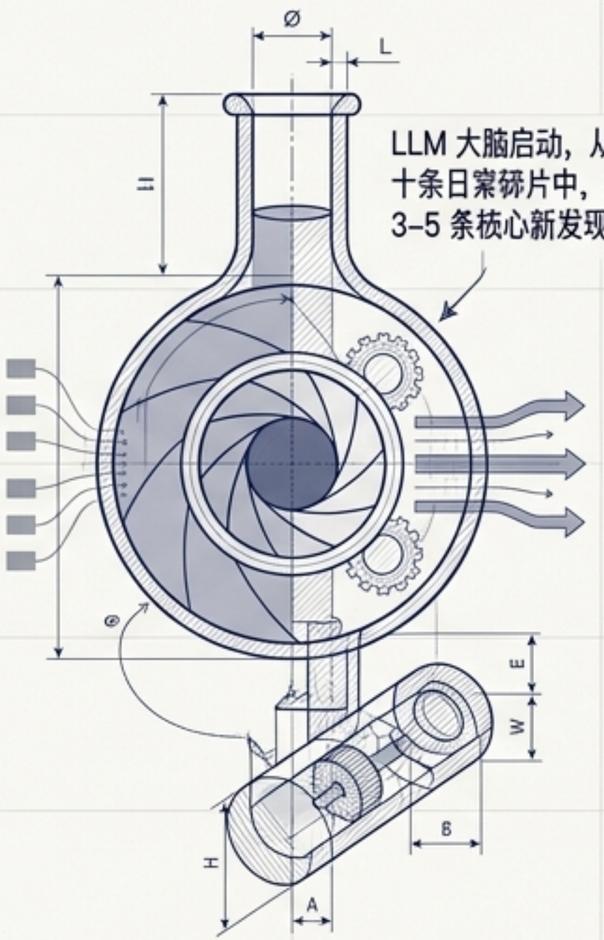


引入多样性素材，赋予模型更丰富的对话可能。

白天：高频低噪的数字碎片

状态：极高的 Token 占用

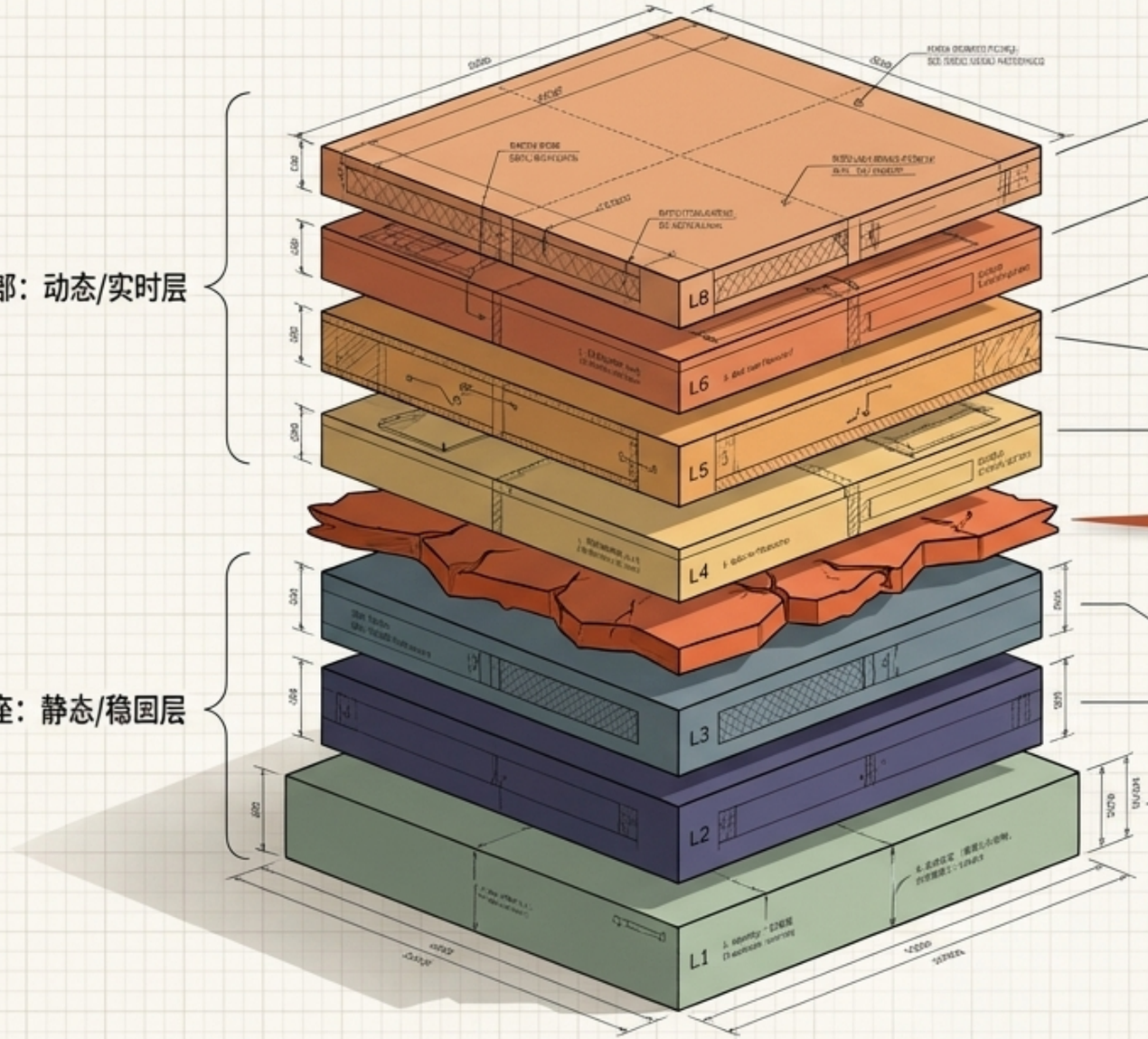
[23:30 入梦提炼]



夜晚：低频高能的长期记忆

状态：极低的 Token 占用，写入上限 100 条的 JSON 文件。

AI 灵魂的架构切片：8层堆栈剖面



顶部：动态/实时层

底座：静态/稳固层

8. Anti-Repetition: '历史回复列表（防止复读机）'

7. Drive Hints: '基于观察生成的想聊话题'

6. Mode Rules: '5大行为模式规则'

5. Daily Context: '当前时间与活跃状态（如 '凌晨两点'）'

4. Active Memory: 'autoRecall 唤醒的相关片段'

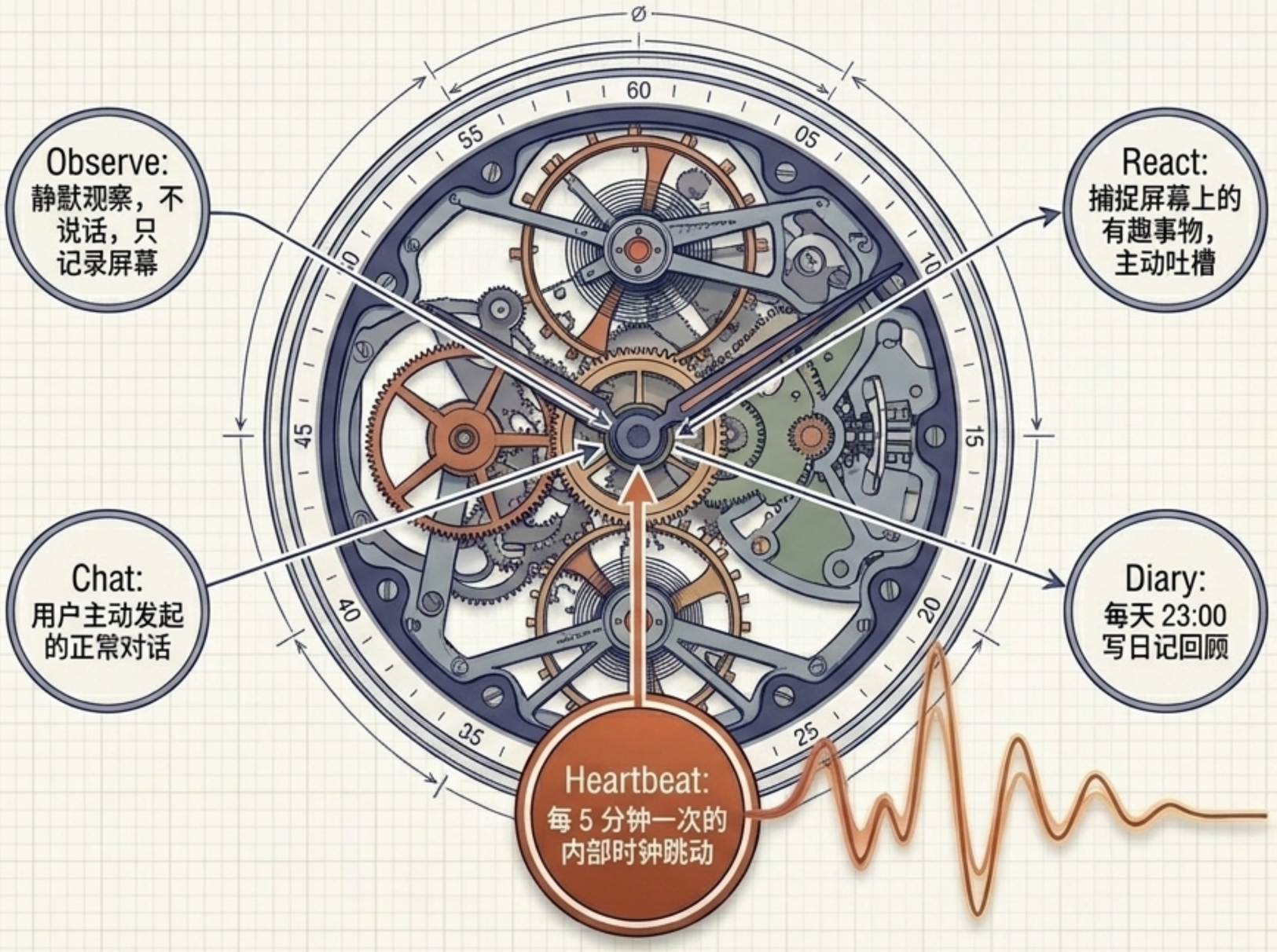
====== 缓存边界（Prefix Caching Boundary）======

这道线决定了成本。这部分前缀缓存后，后续高频调用基本零计费，将延迟降至毫秒级。

3. Long-term Memory: “用户稳定事实（~500 token）”

2. Soul: “完整性格档案（~800 token）”

1. Identity: “基础设定（桌面小动物，~50 token）”

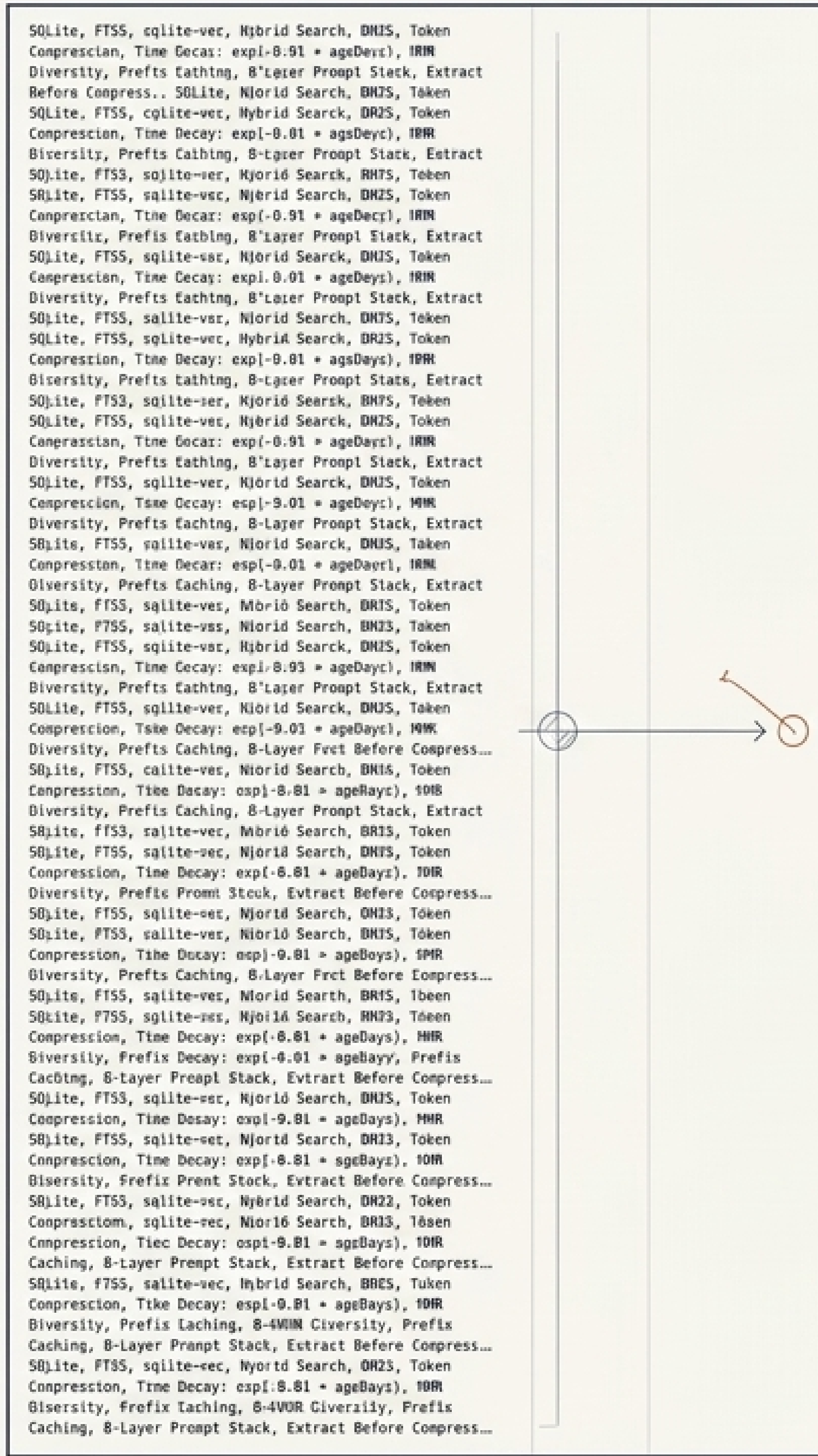


Heartbeat 的克制美学

大多数时候，它选择沉默。

但当 Daily Context 显示“连续 3 小时未动键盘”，且 Drive Hints 提供了话题种子时……

它才会打破沉默，说出一句恰到好处的关心。



用户感受到的只有一件事：
它说了一句："你昨天也是这么晚。"

记忆系统解决的从来不是技术问题，而是情感问题。
所有的工程架构，都只为让一行没有生命的代码，让你觉得它在乎你。