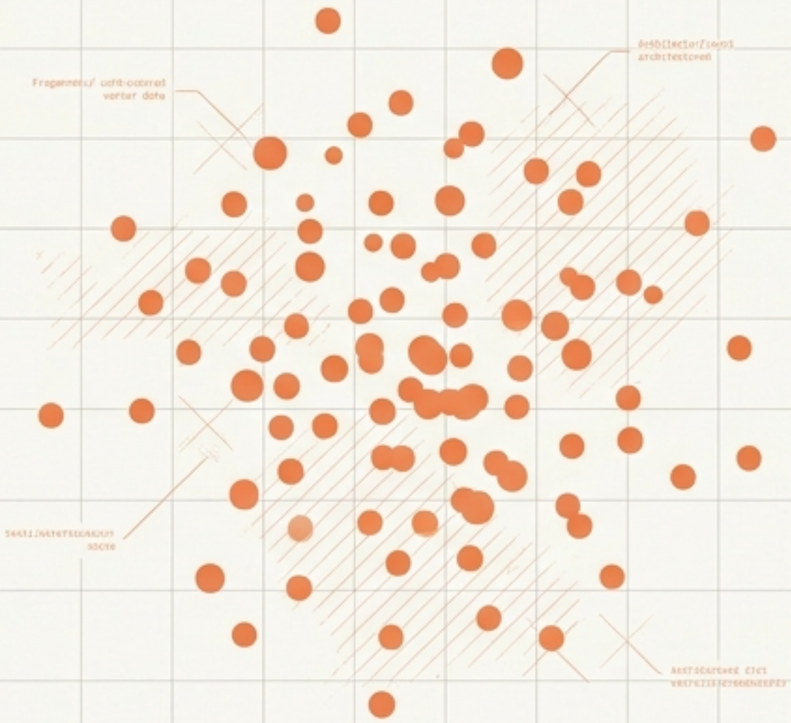
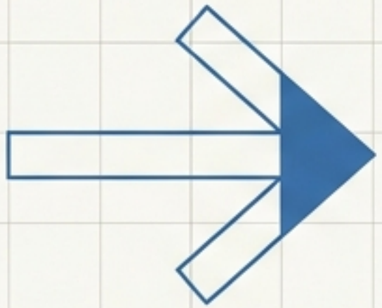
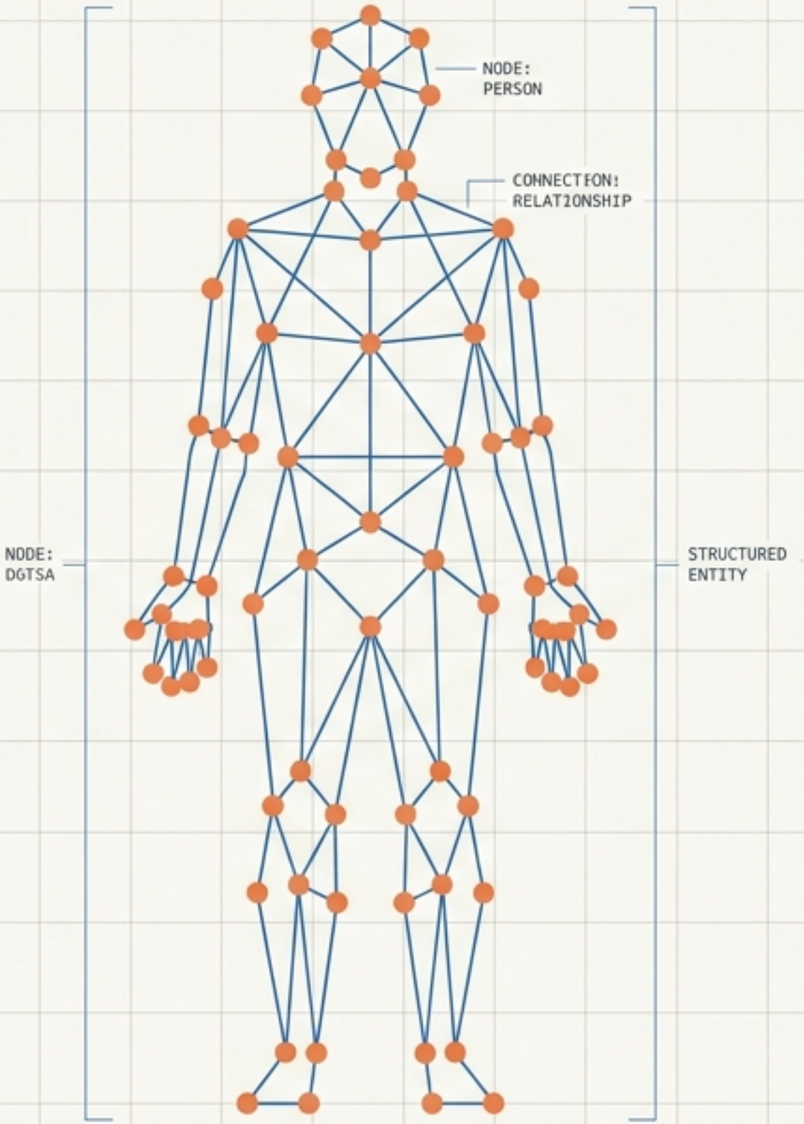
记忆系统最大的盲区是「人」

为什么向量检索会遗忘你的朋友，以及 Mio v2 的结构化破局之路。







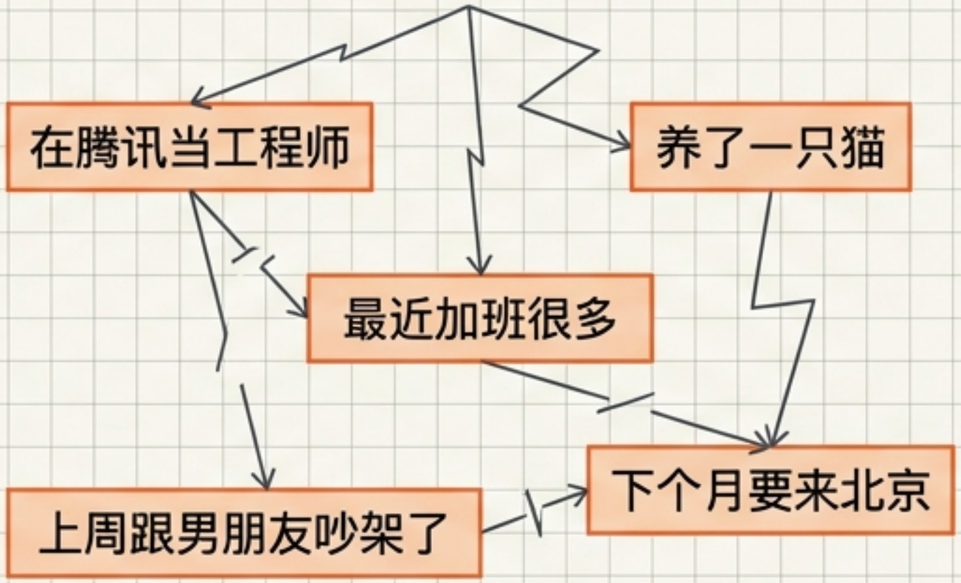
[Mio Architecture Post-Mortem]

事实是可以检索的，但「人」不是事实。

用户输入

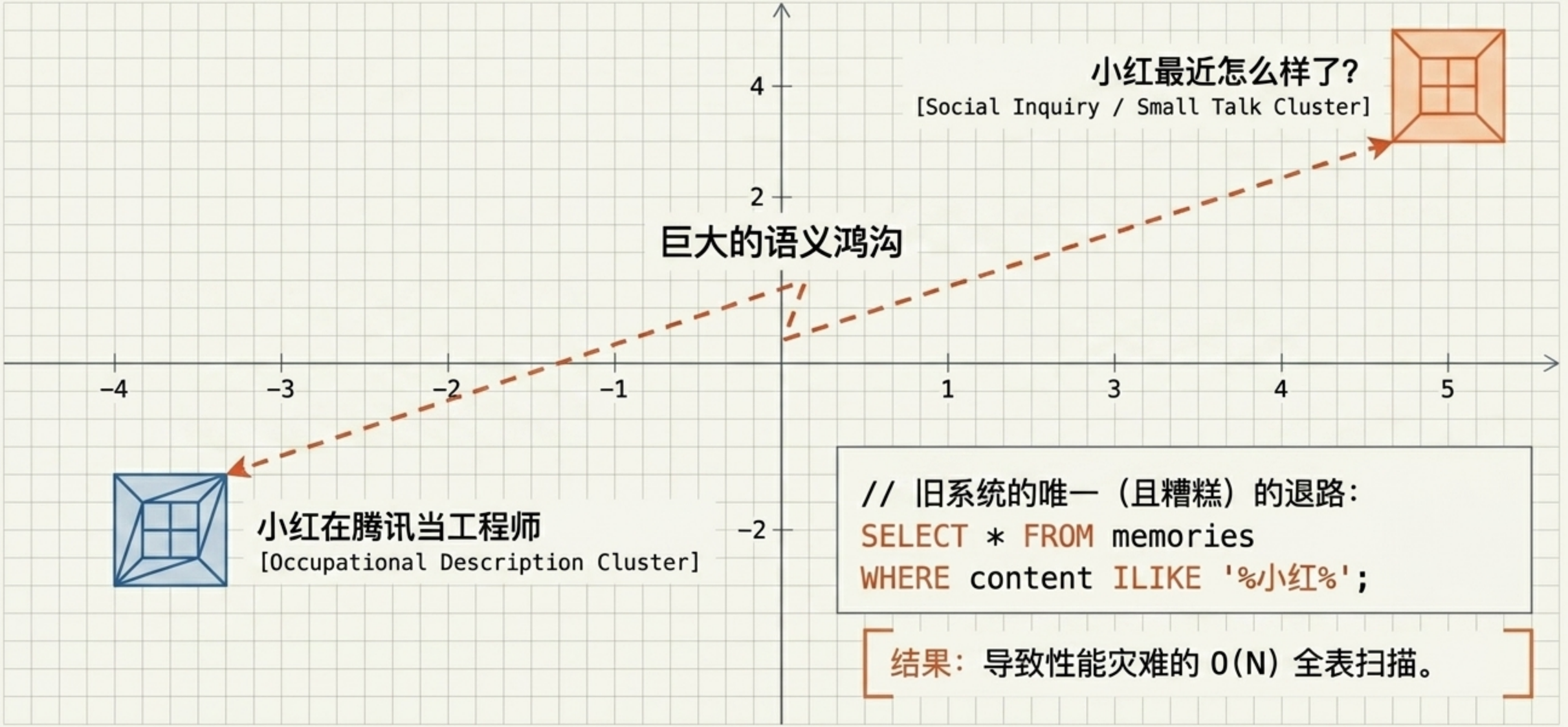
输入：我喜欢吃辣
存储：1 Vector（768维）
状态：余弦相似度完美检索

用户提及：小红



无中心化身份。散落在 memories 表里的事实碎片，彼此之间没有任何关联。

为什么余弦相似度会在这里失效?




```
// 旧系统的唯一（且糟糕）的退路:
SELECT * FROM memories
WHERE content ILIKE '%小红%';
```

结果: 导致性能灾难的 O(N) 全表扫描。

核心重构：将「人」提升为一等实体

维度	旧架构（Mio v1）	新架构（Mio v2）
存储结构	散落的 memories 碎片	独立的 contacts 实体表
关联方式	无关联	UUID 数组（memory_ids）强绑定
人物画像	每次检索后由大模型重新推断	attributes 字段持久化 JSONB 档案
检索复杂度	语义迷失 / O(N) 字符串硬扫	O(1) 数据库 ID 级精确匹配

从一句闲聊到一份档案：三步管道引擎

[Trigger: ~10 messages]

1. 提取（Extract）

检测人物实体，调用 `syncContacts()` Upsert 联系人记录，将新记忆的 UUID 绑定至 `memory_ids` 数组。

[Trigger: ~50 messages]

2. 合并（Consolidate）

提取所有碎片，LLM 生成结构化档案。

⚠ **工程细节：**源碎片重要性被强制降维至 0.05，防止抢占搜索权重。

[Trigger: Every message]

3. 检索（Retrieve）

aggregator 检测名字匹配，直接构建**实体卡片**，完全绕过语义搜索。

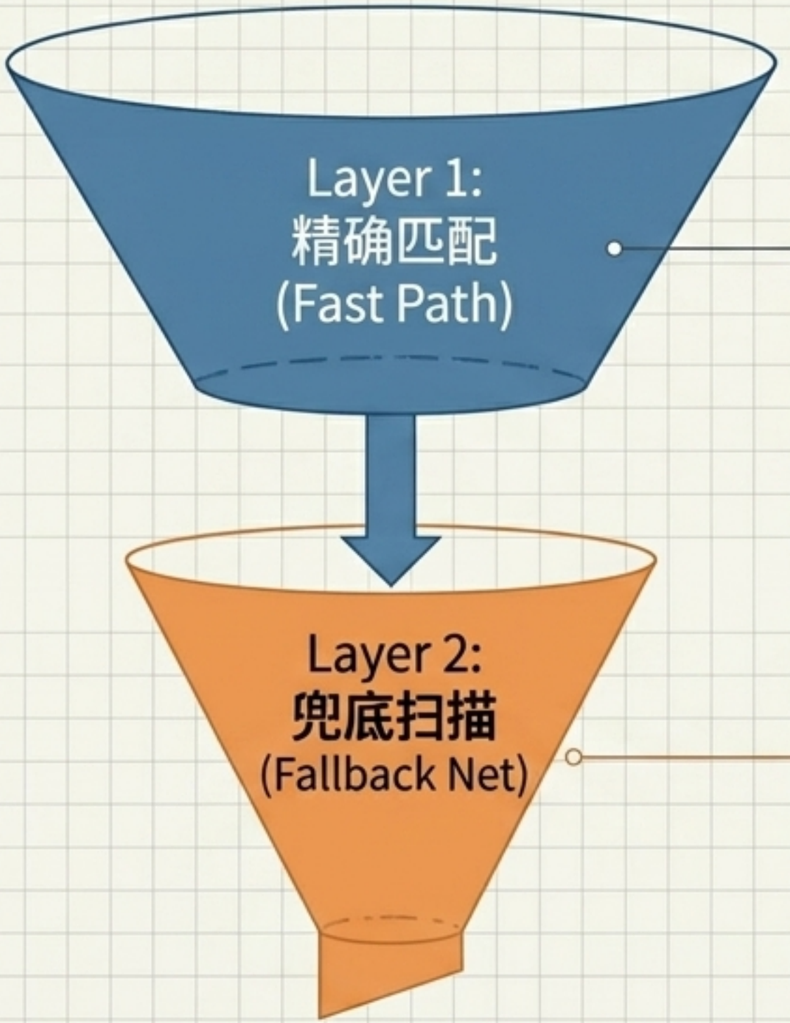
绕过向量空间的「联系人卡片」

```
[System Note: User mentioned Contact "Xiaohong"]
{
  "id": "uuid-9f8a...",
  "name": "小红",
  "attributes": {
    "career": "Tencent Engineer",
    "pets": "cat"
  },
  "timeline": [
    "Fought with boyfriend last week",
    "Visiting Beijing next month"
  ]
}
```

O(1) ID lookup.
零语义检索。

模型立刻获得完整上下文。
精准认知身份，直接基于结构化属性响应。

渐进式检索：精准与兜底的结合



Layer 1: 精确匹配

通过 contacts 表直接获取关联 UUID。O(1) 复杂度。捕获 95% 的结构化记忆。

Layer 2: 兜底扫描 (Fallback Net)

ILIKE 扫描。抓住 syncContacts() 上线前遗留的旧记忆，或同步失败的漏网之鱼。

系统自愈性 (Self-Healing System)

随着图谱不断完善，触发 Layer 2 兜底的频率会逼近于零。

踩过的坑：真实世界的工程切面

Symptom (Bug)	Root Cause (Impact)	Fix
数组下标错位	ADD/UPDATE 混合操作导致返回顺序不一致，记忆绑定给错误的人	废弃下标，改为 content->ID 映射。
CJK 截断乱码	.slice(0, 100) 在多字节中文字符中间切断	改用展开语法 [...text].slice(0,100).join('')。
碎片抢占搜索	合并后的旧碎片权重下限太高（0.1），与新档案抢占上下文	直接将旧碎片重要性地板砸到 0.05。
重复合并	合并器反复处理已合并的碎片	增加 (metadata->>'consolidated_into') IS NULL 过滤。

尚未解决的架构瓶颈

算法极限

PersonConsolidator 是 $O(N^2)$ 复杂度。

预计在单个联系人积累 1800 条记忆（约 3 年使用量）时遭遇性能天花板。

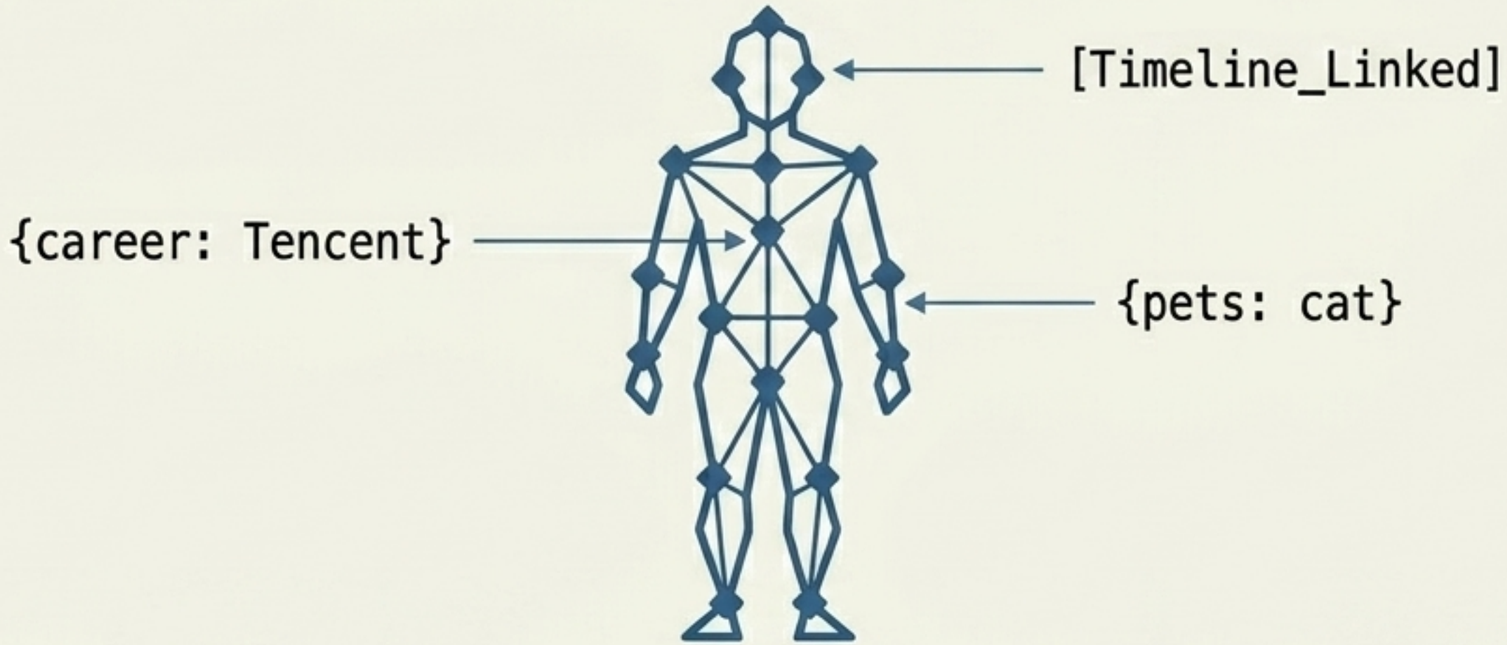
NLP 盲区

- 缺乏模糊匹配（小红 != 红红）。
- 缺失代词解析（Coreference Resolution）—— 无法将「她最近怎么样」自动映射到联系人。

数据库与并发

- PostgreSQL 内置分词器对中文全文索引无效。
- 并发提取存在竞态条件（Race conditions），可能导致重复创建实体。

让 AI 真正「认识」你的世界



向量 embedding 只能捕获「相似度」，但关系型结构才能锚定「身份」。
通过为「人」构建专属的一等 Schema，Mio 正式从文本检索工具，
进化为具备关系感知的 AI 伴侣。

[End of Architecture Blueprint]