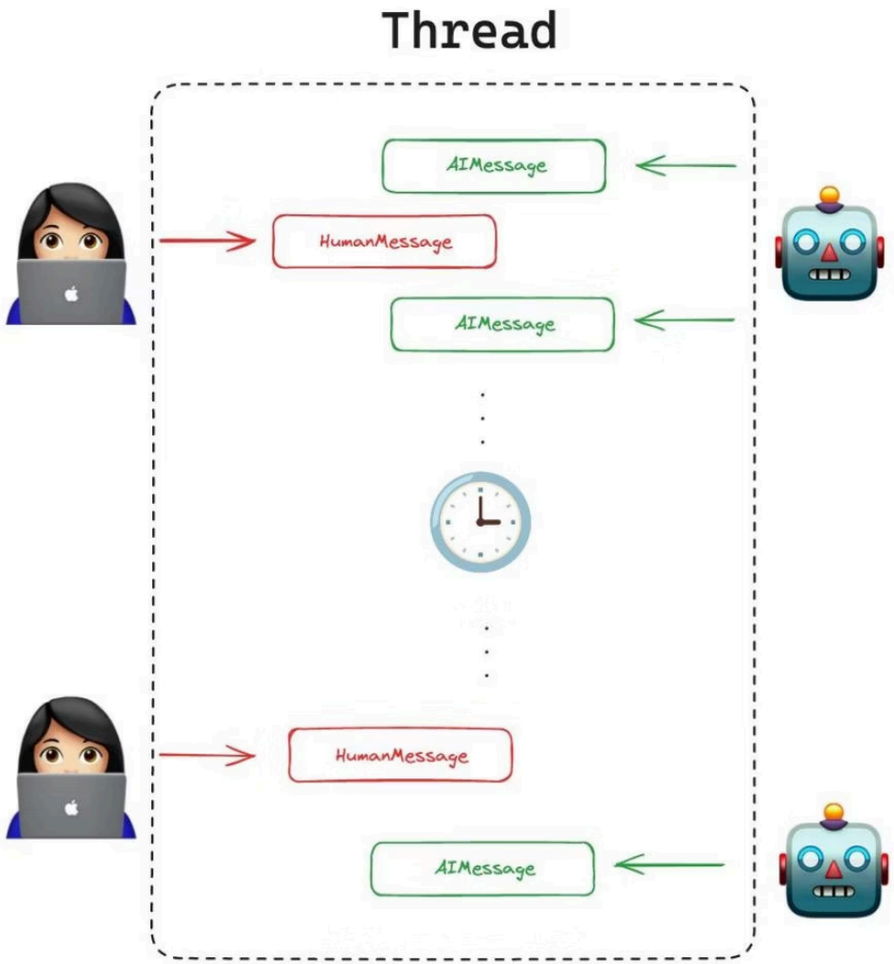
记忆：理解与应用

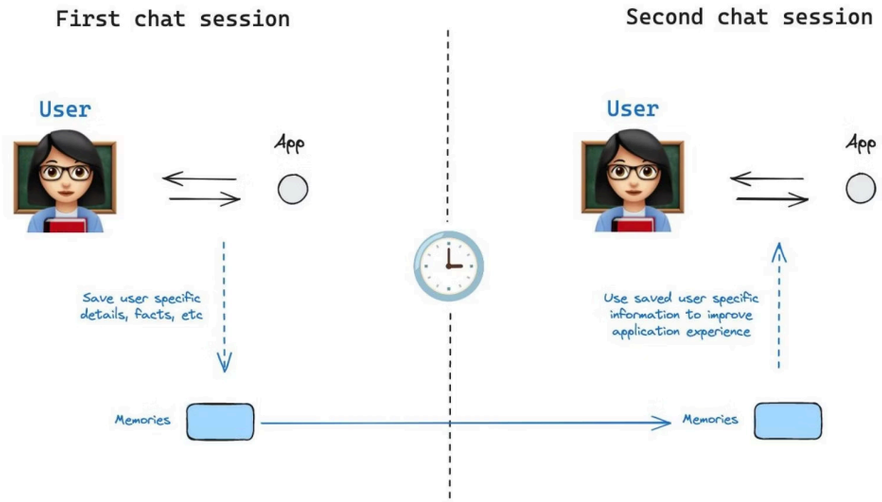
记忆是一种认知功能，它使人们能够**存储、检索和使用信息**来理解现在和未来。

会话内记忆 vs. 跨会话记忆



会话内记忆（短期）

在单个会话（线程）中保持对话历史，允许聊天中断。



跨会话记忆（长期）

记住特定用户在所有聊天会话中的信息。

记忆类型：短期与长期

1

短期记忆

范围：会话内（线程）。

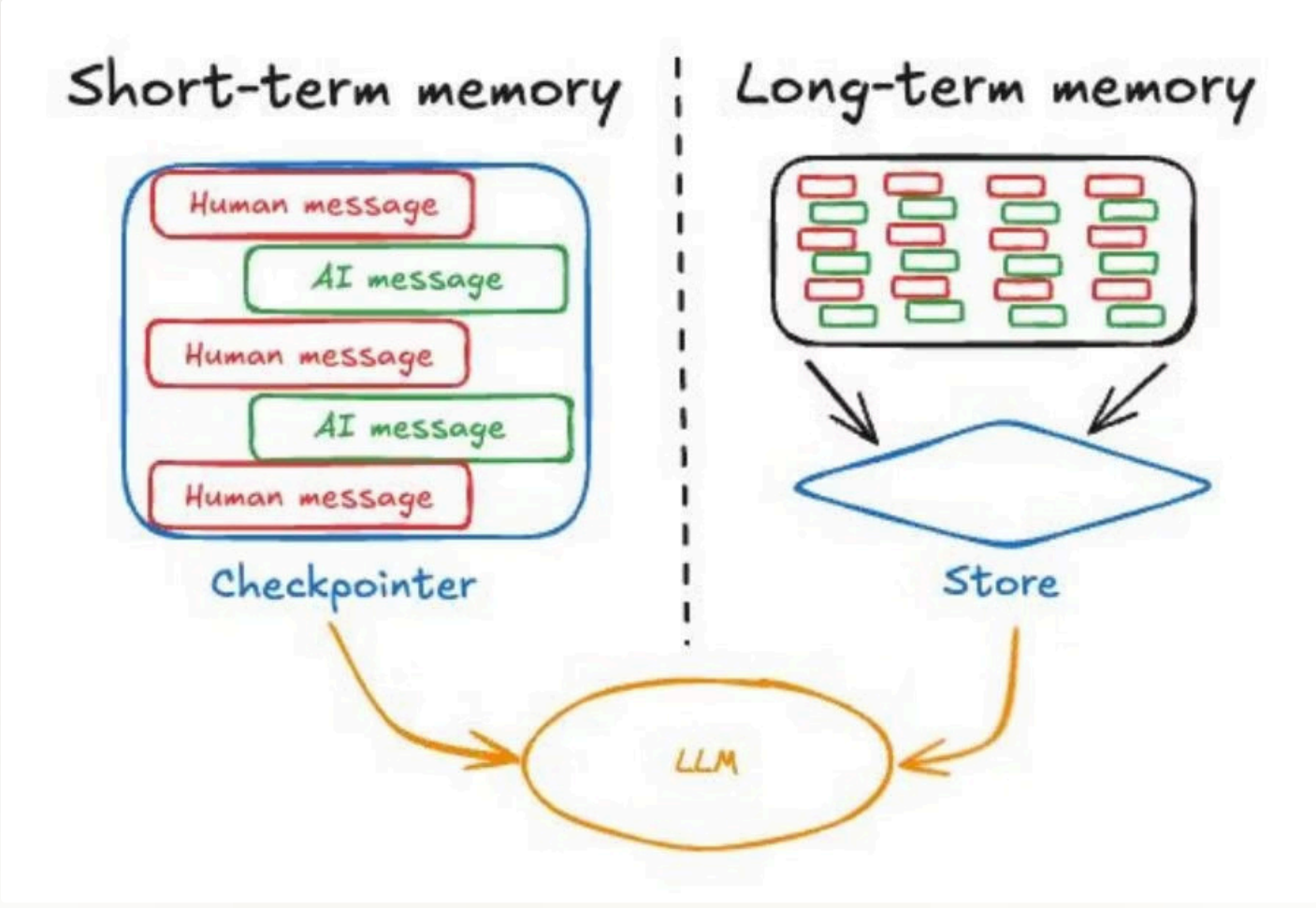
用例：保持对话历史，处理聊天中断。

2

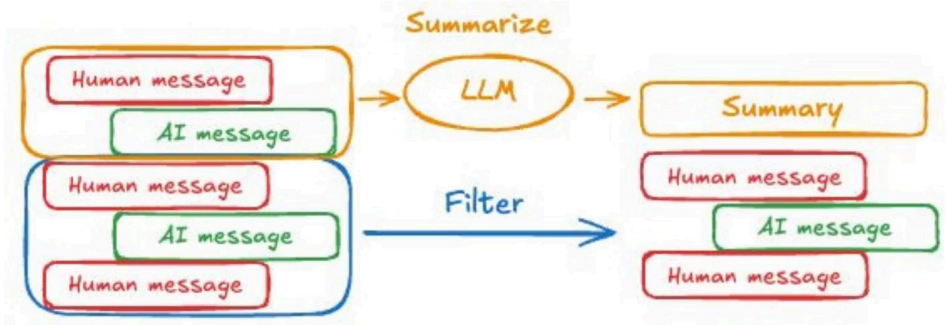
长期记忆

范围：跨会话（线程）。

用例：记住特定用户在所有聊天会话中的信息。



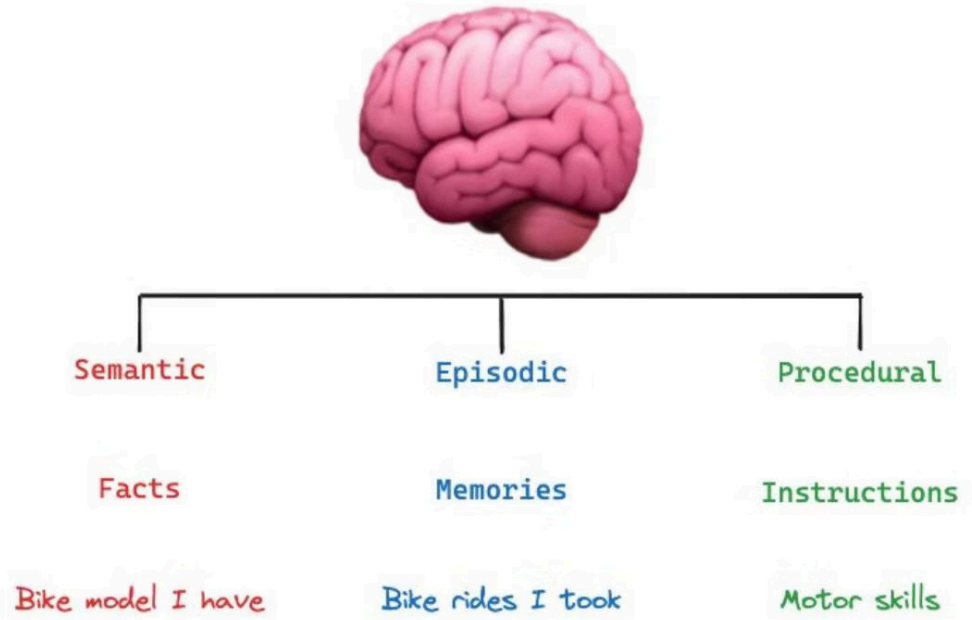
LangGraph 中的记忆应用



Checkpointer

短期记忆

用于会话内记忆，保存对话状态。



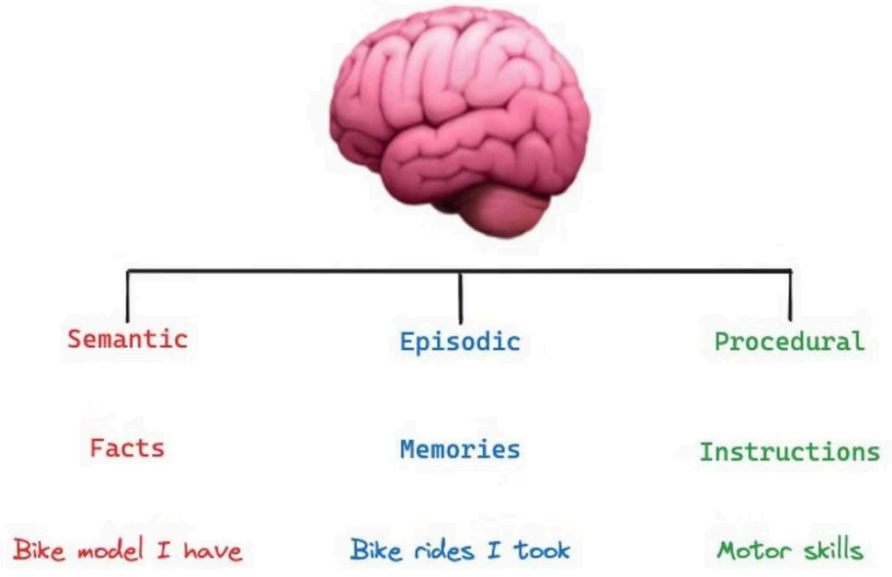
Store

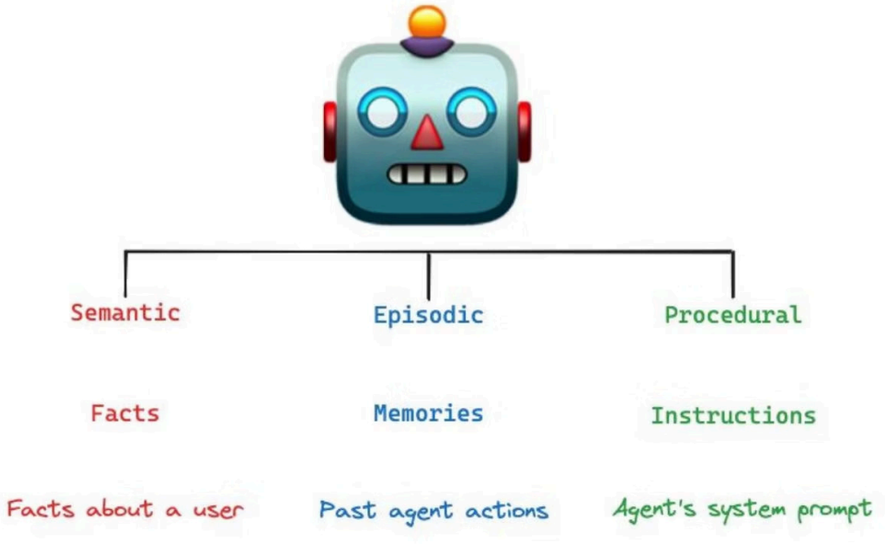
长期记忆

用于跨会话记忆，存储用户相关信息。

何时更新记忆？这取决于记忆的类型和应用场景。

长期记忆：人类类比





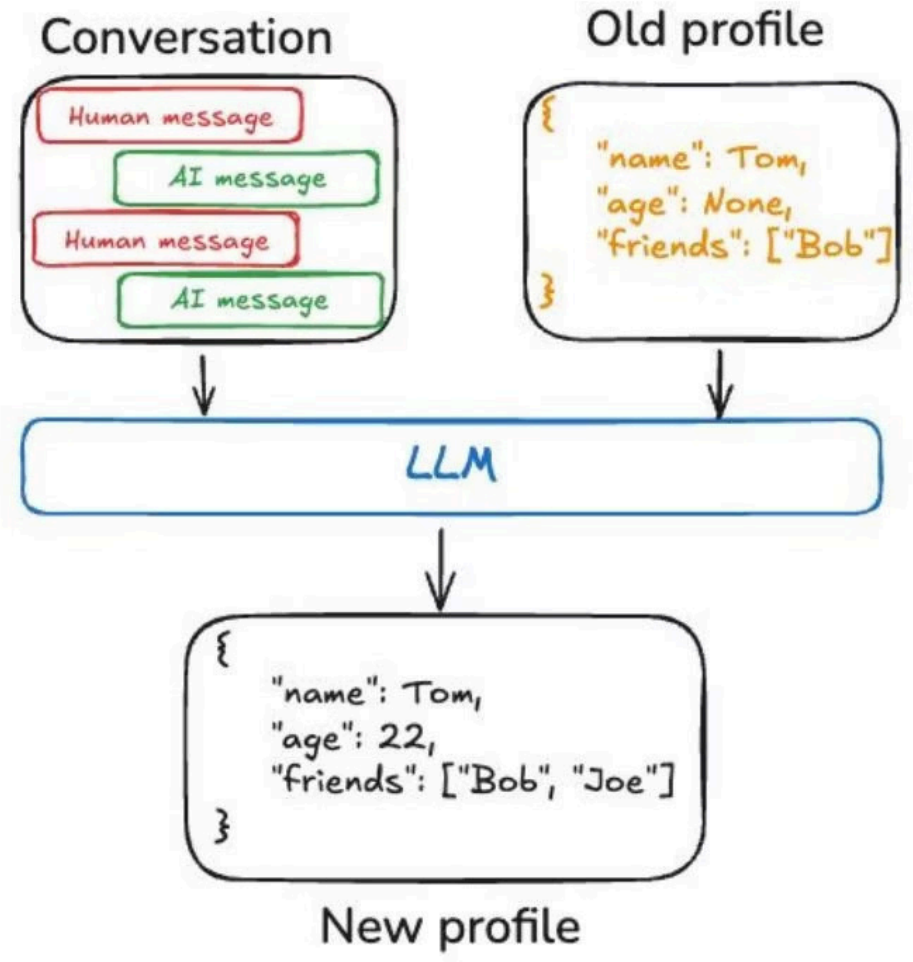
记忆类型	人类长期记忆	LangGraph 智能体长期记忆	应用示例
语义记忆（Semantic）	关于世界的事实与知识例：巴黎是法国的首都	关于用户的事实与属性例：用户喜欢喝美式咖啡	个性化交互
情景记忆（Episodic）	关于个人经历的事件例：我去年去过巴黎	智能体的交互与行动历史例：用户上次让我订机票	上下文延续与决策优化
程序性记忆（Procedural）	关于技能与动作的记忆例：骑自行车	智能体的运行指令与规则例：系统提示词、调用 API 的顺序	稳定执行与操作规范

长期记忆：语义记忆（Profile）

在人类中，**语义记忆**是我们存储事实与知识的方式；

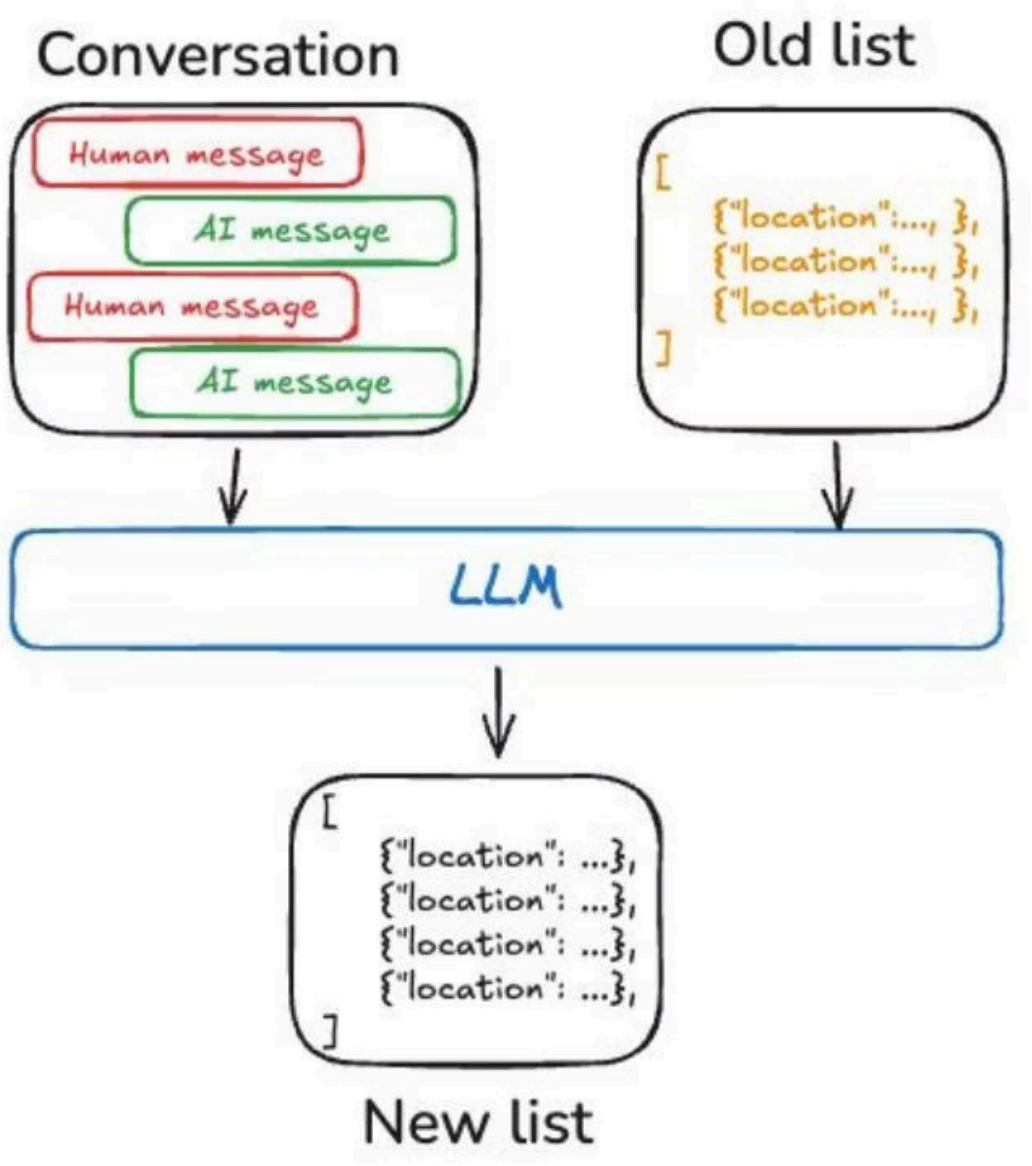
Agent如何组织事实？

Profile：单一文档表示



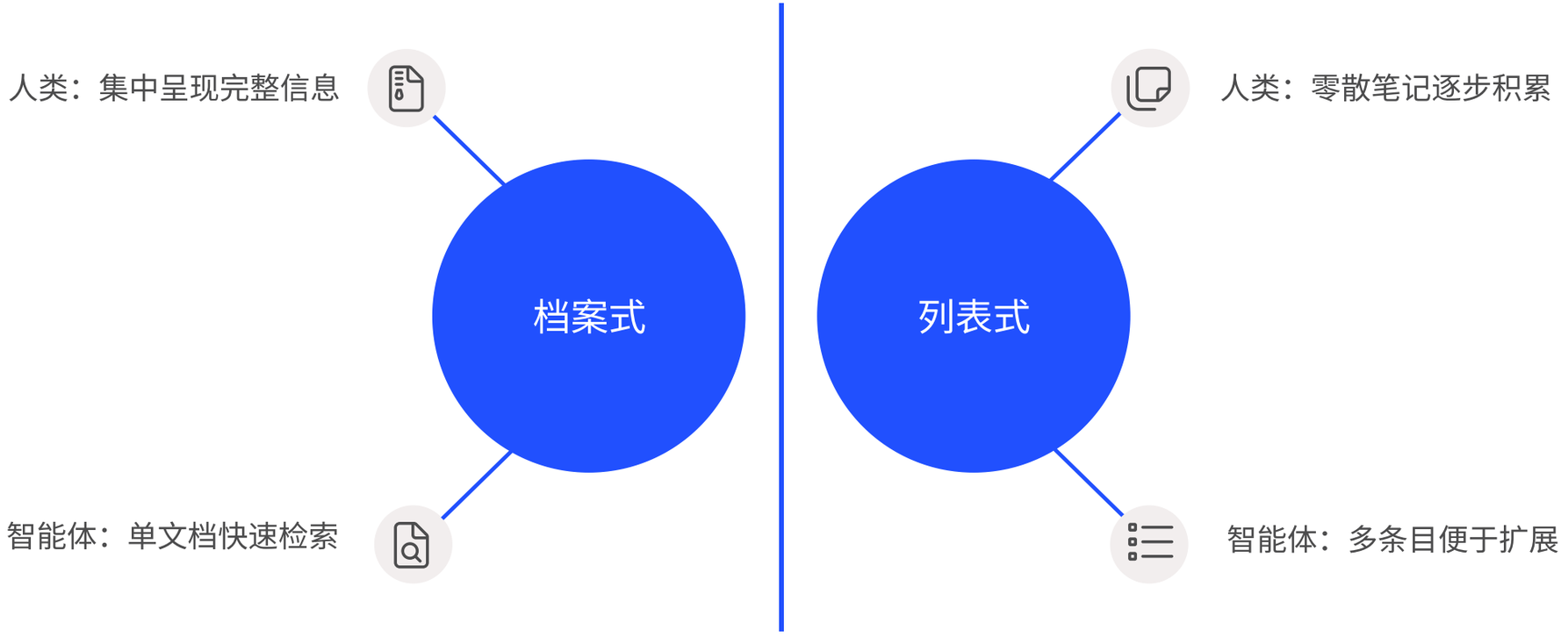
在智能体中，**Profile 式语义记忆**就像一份“集中档案”，把关于用户或对象的所有信息放在单一文档里，方便一次性检索与调用，但随着信息量增加，维护和更新会变得更加困难。

List：多个文档组成的集合



在人类中，**语义记忆**不仅包含单一档案式的知识，也表现为分散的零碎事实；在智能体中，**Collection 式语义记忆**就像“信息清单”，将事实拆分为多个文档或条目，便于逐步扩展和新增，但在检索时可能因为数量增多而变得复杂，查找效率下降。

长期记忆：语义记忆（Collection）

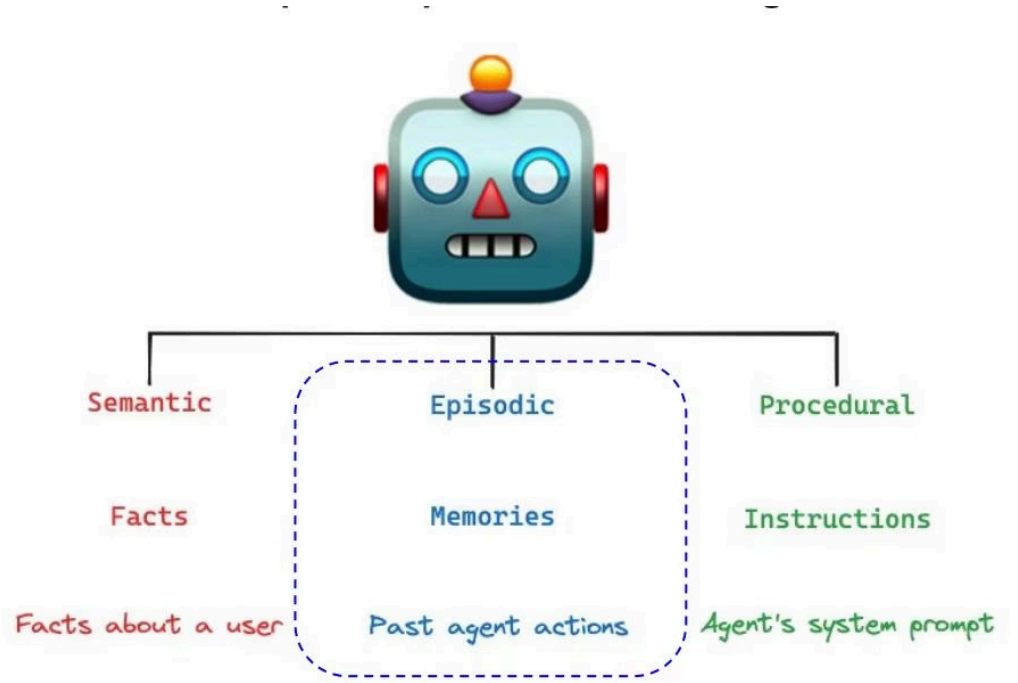


类型	描述	优点	缺点
Profile（档案式）	单一文档，集中存储所有事实	检索方便：一次性获取完整信息	随规模扩大难以维护，更新成本高
Collection（列表式）	多个文档，每个存储部分事实	易于扩展：可逐步添加新信息	检索复杂：随着文档数量增加，查找效率下降

长期记忆：情景记忆（Episodic）

在人类中，**情景记忆**让我们能够回忆起具体的经历；

在智能体中，**情景记忆**则是保存交互与推理轨迹，使模型能够“借鉴历史经验”，提升决策与上下文理解。

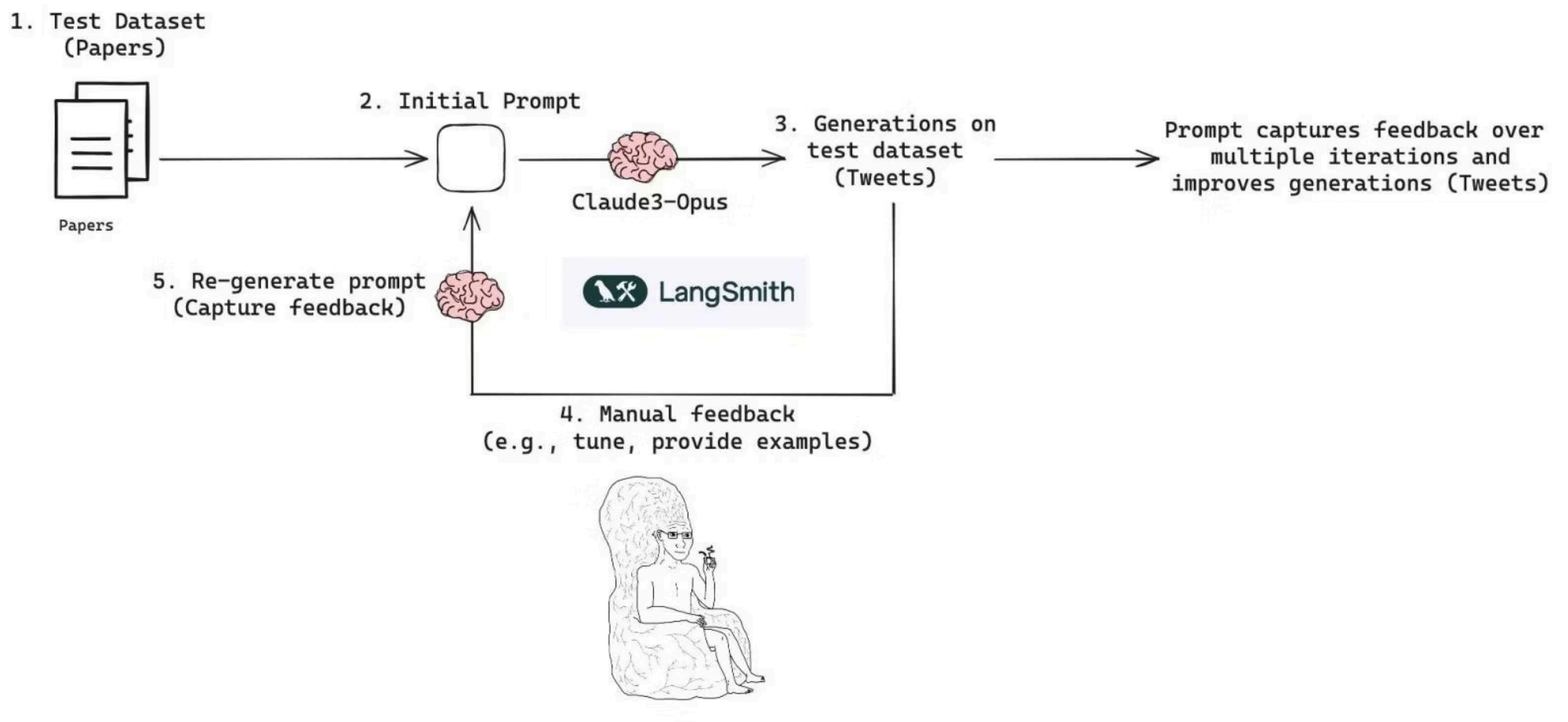


如何让智能体记住并引用过去的行动?

- 一种做法是“少量示例（few-shot prompting）”，将过去的推理轨迹作为示例提供。
- 可改善工具调用与上下文推理。

参考：[Few-shot prompting for tool calling]

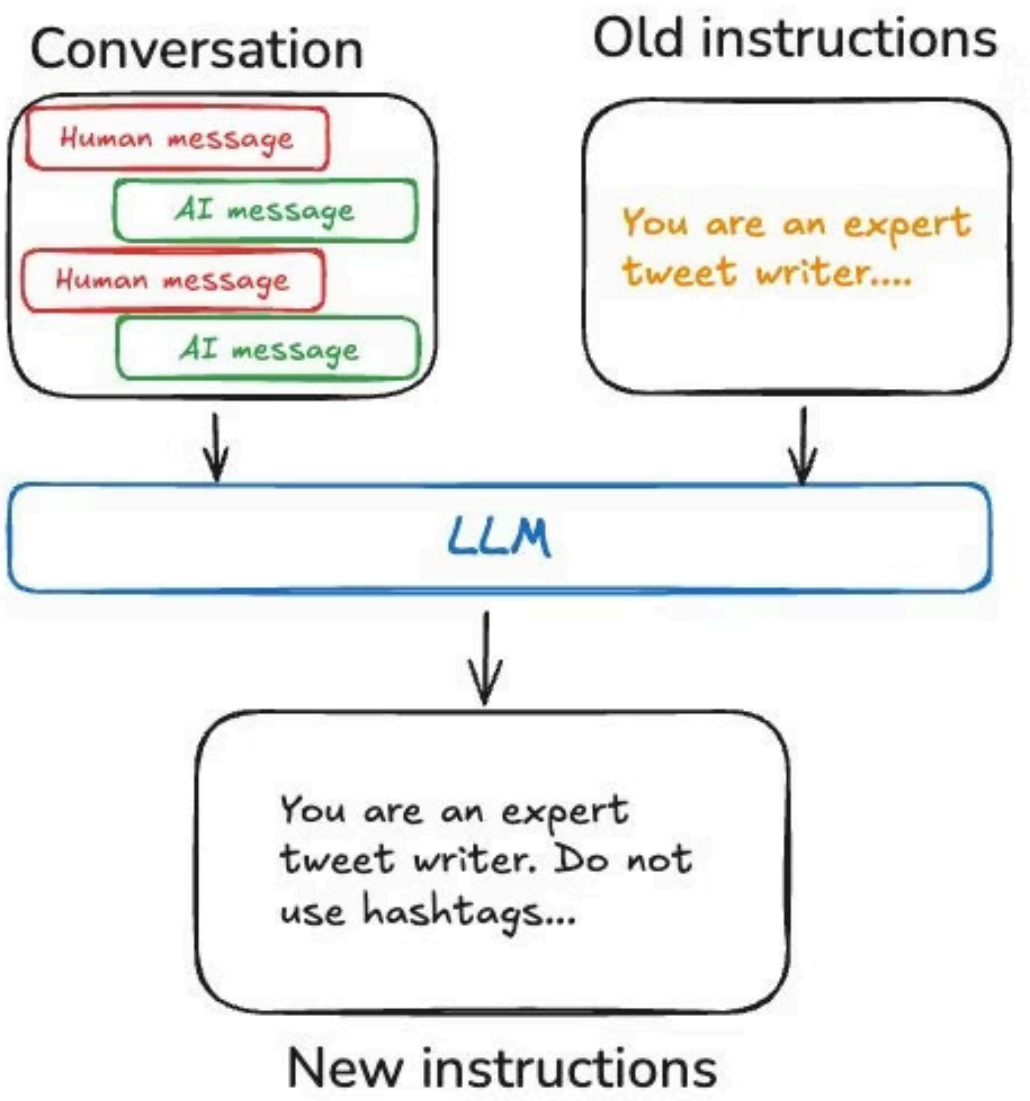
（下图） 展示了将先前推理路径存储为可复用的记忆。



长期记忆：程序性记忆（Procedural）

在人类中，**程序性记忆**是我们学习并保留的技能与动作模式；

在智能体中，**程序性记忆**则是它的“操作说明书”，通过不断更新与迭代，使其能够更稳定、更高效地完成任务。



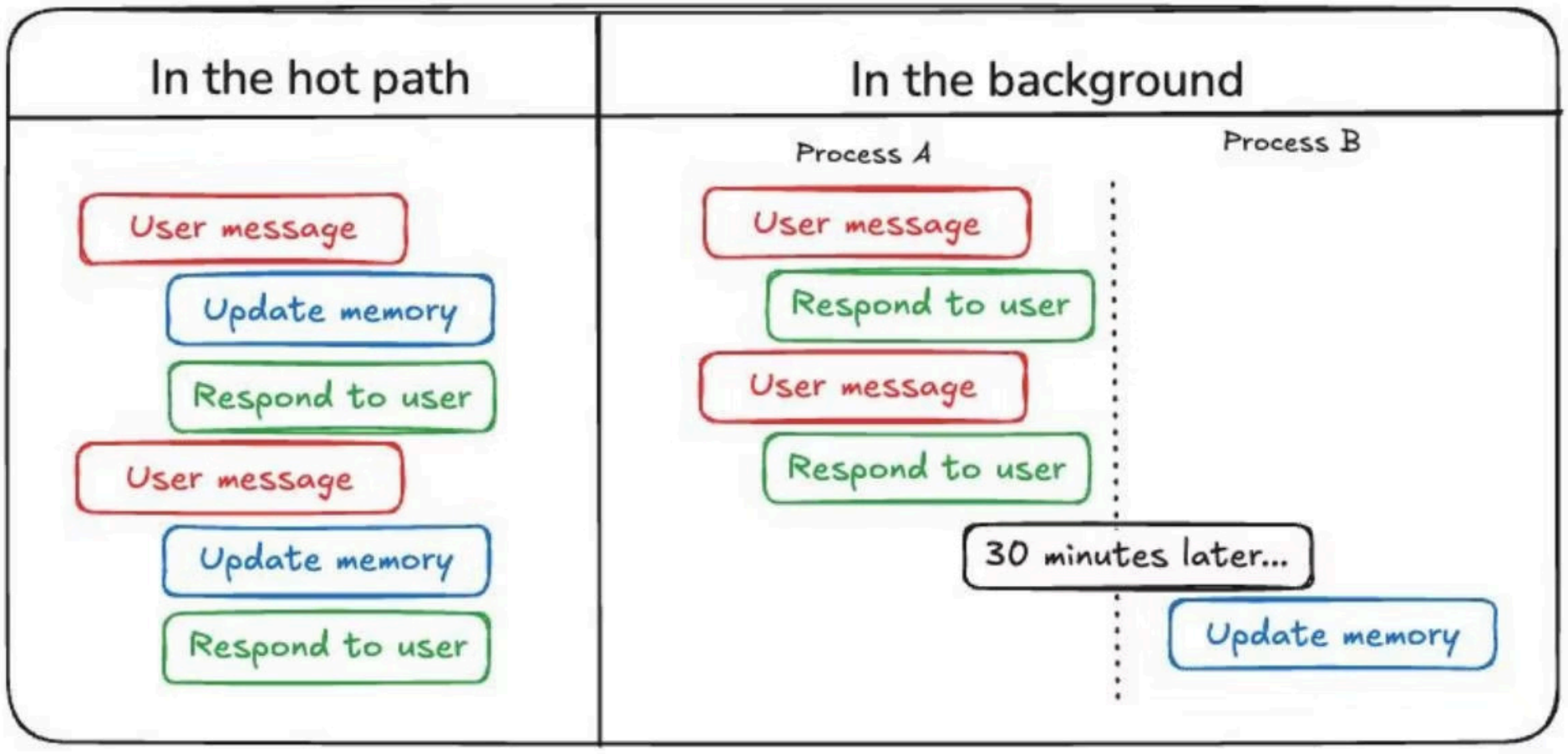
问题：如何更新智能体的“指令”？

- 程序性记忆类似于技能学习（例如如何骑自行车）。
- 在智能体中，可以通过更新“操作指南”或“行为规则”来实现。

参考：

- [《LARGE LANGUAGE MODELS ARE HUMAN-LEVEL PROMPT ENGINEERS》]

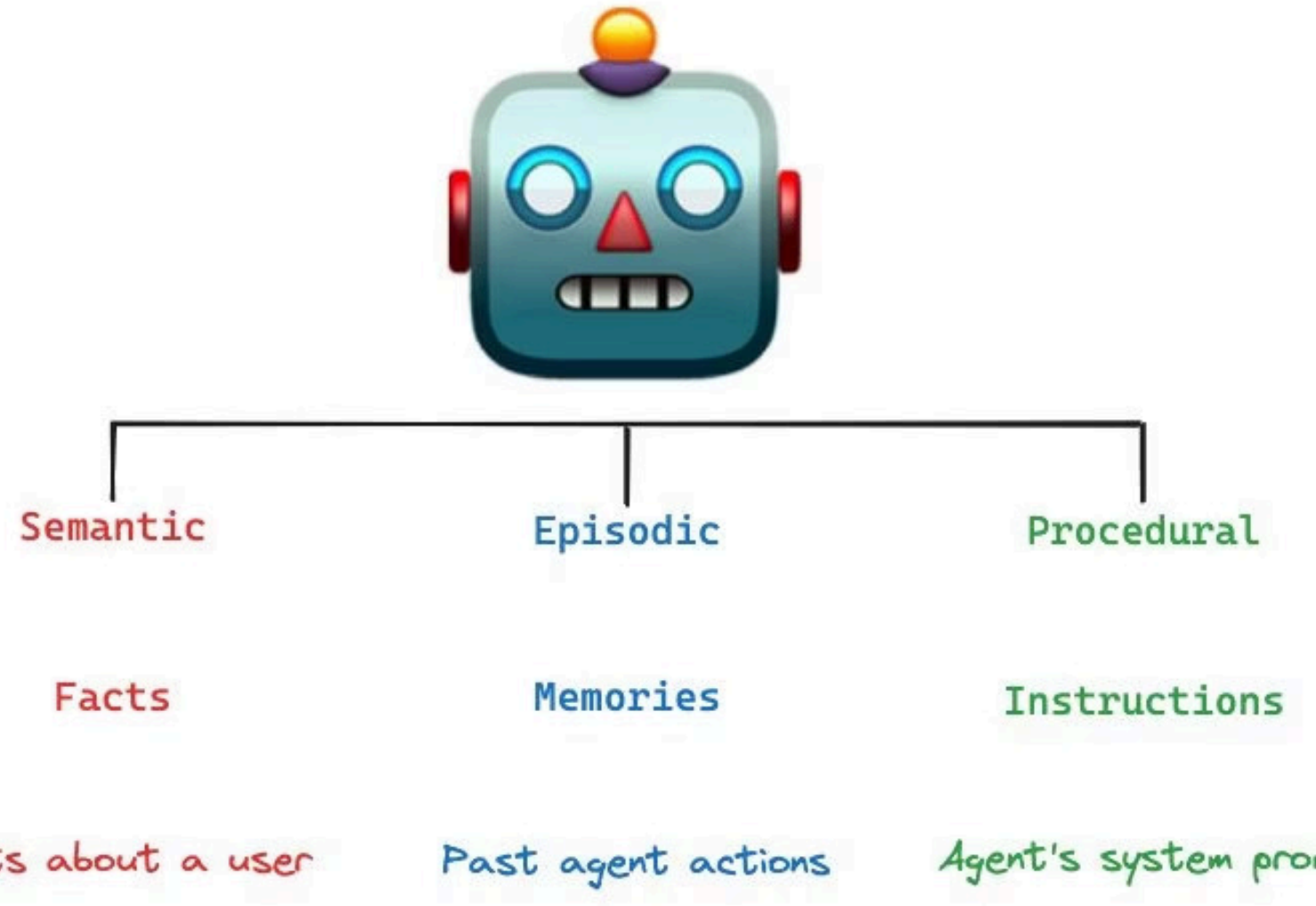
长期记忆的更新策略



更新方式	描述	优点	缺点	使用场景示例
Hot-path（实时写入）	在对话运行时直接更新（如 ChatGPT）	用户可见，实时更新	可能影响用户体验和响应延迟，导致性能下降	对话式助手（如 ChatGPT）、实时问答系统
Background（后台写入）	在单独进程中异步更新	降低性能风险	写入频率需要精心调优	长期用户建模、个性化推荐系统、后台日志分析

最终目标：智能记忆系统

构建一个能够有效存储、检索和更新记忆的智能系统，以支持更高级的AI功能。



长期记忆在智能体中的设计目标，不仅仅是“存储信息”，而是要实现 **持久、个性化与高效性** 的平衡。具体而言：

1. **持久化记忆**
   - 智能体能够在多次会话、长期使用中，持续记住用户的偏好、身份信息与交互历史。
   - 例如：用户喜欢喝美式咖啡，智能体即使在几天或几周后仍能记得并主动提供个性化建议。
2. **上下文理解**
   - 智能体能够在不同的任务与对话中，回忆并利用与当前情境相关的信息。
   - 不仅是“记住事实”，还包括“理解情境”，从而让回答更贴合用户需求。
3. **高效的更新机制**
   - 在保证性能与用户体验的前提下，智能体能够动态更新记忆。
   - 更新方式既可以是实时的（Hot-path），也可以是异步的（Background），确保在速度与稳定性之间取得平衡。
4. **个性化与适应性**
   - 智能体通过长期记忆形成“用户画像”，从而在交互中展现出个性化的风格与行为。
   - 随着时间推移，智能体能适应用户的新习惯或新偏好，避免信息过时。