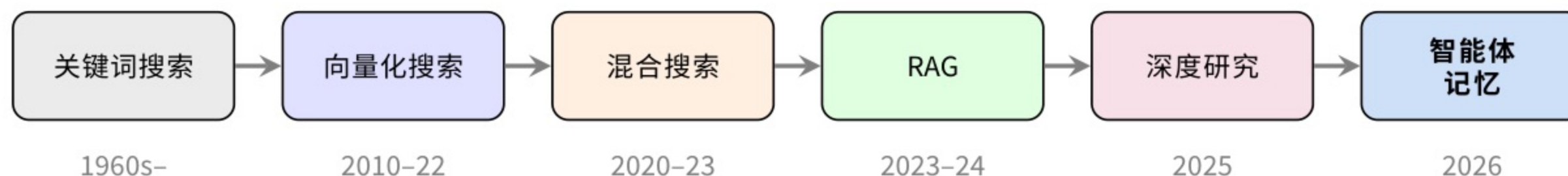


# 一个例子：“我们之前做的 911 图表发给我一下”





# AI 搜索范式的演进



## 2025 年行业趋势:

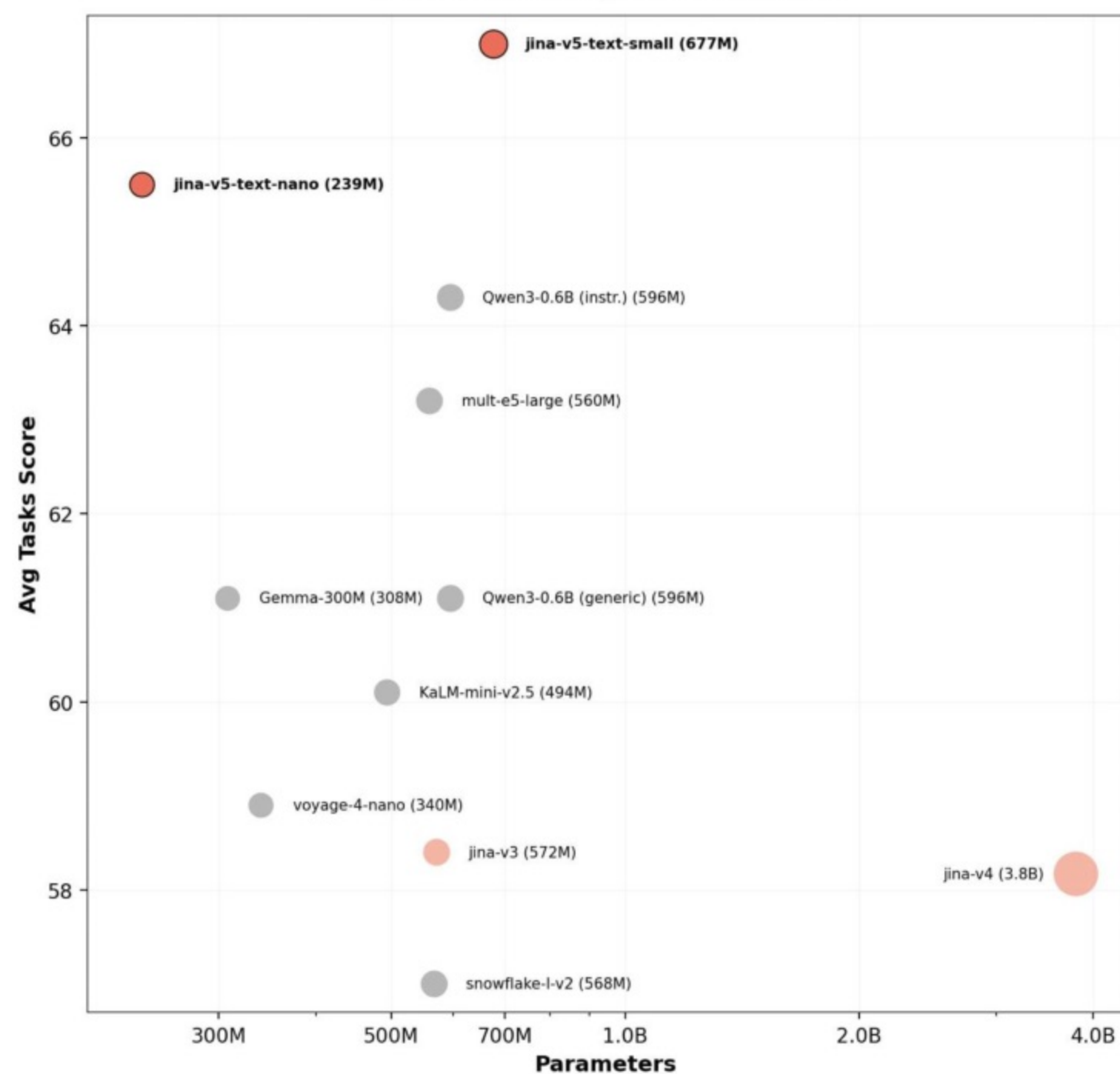
- 纯解码器架构 (Decoder-only) 主导 (Qwen3, NV-Embed)
- 小模型逆袭: <500M 参数竞争力强
- 长上下文 32K-128K 标配; ColPali 视觉检索
- Matryoshka 维度自适应; 二值量化 32× 压缩

## Jina AI 三大件:

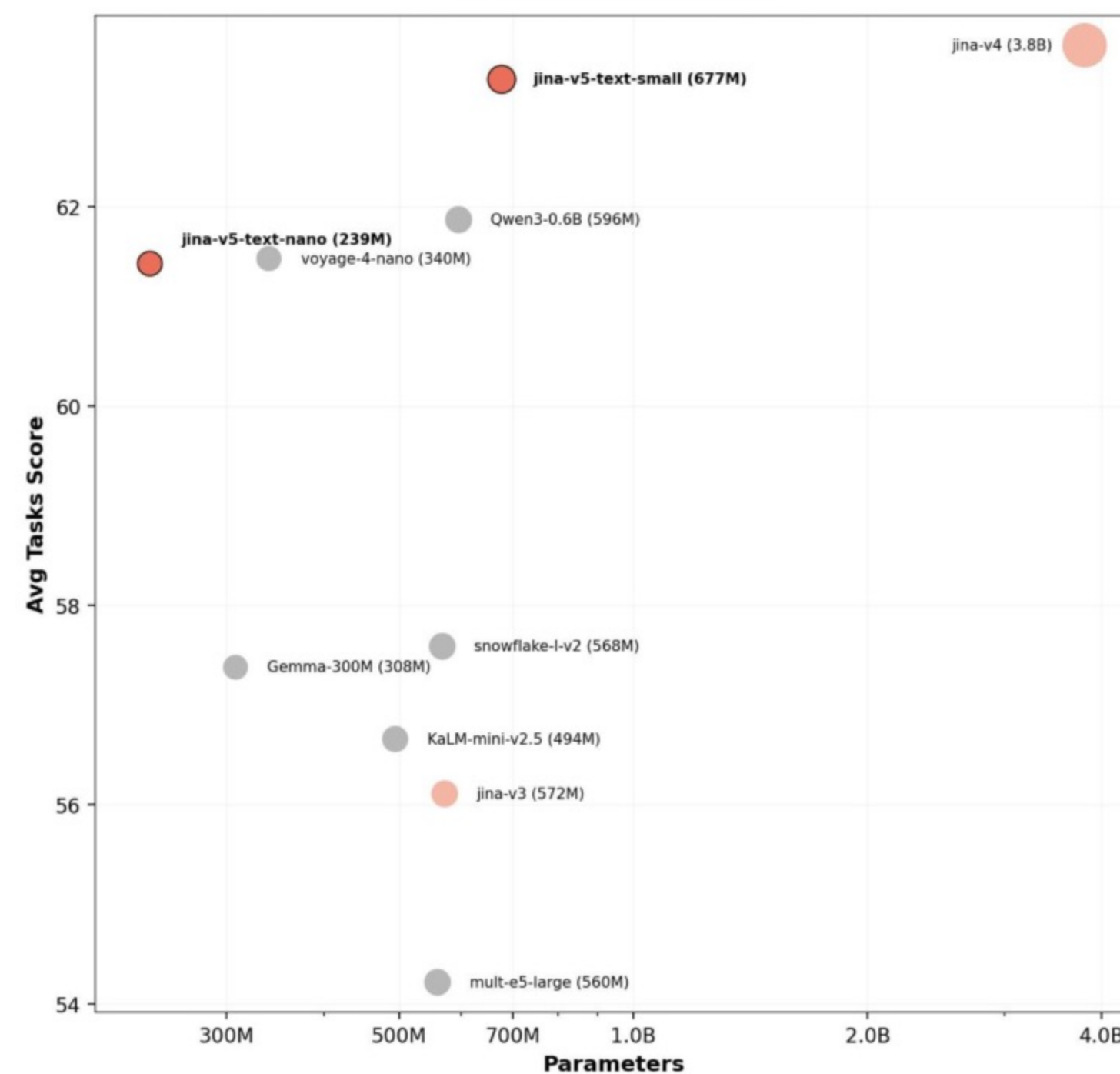
- **Embeddings:** v3 → v4 (多模态) → v5
- **Reranker:** v3 (131K 上下文) + m0 (多模态)
- **Reader:** r.jina.ai (HTML→MD/JSON)
- 全部开放权重, 多模态, 多语言
- llama.cpp, ONNX, vLLM, MLX

# jina-embeddings-v5: 最好最小的向量模型 (SIGIR 2026)

## MMTEB Multilingual Benchmark

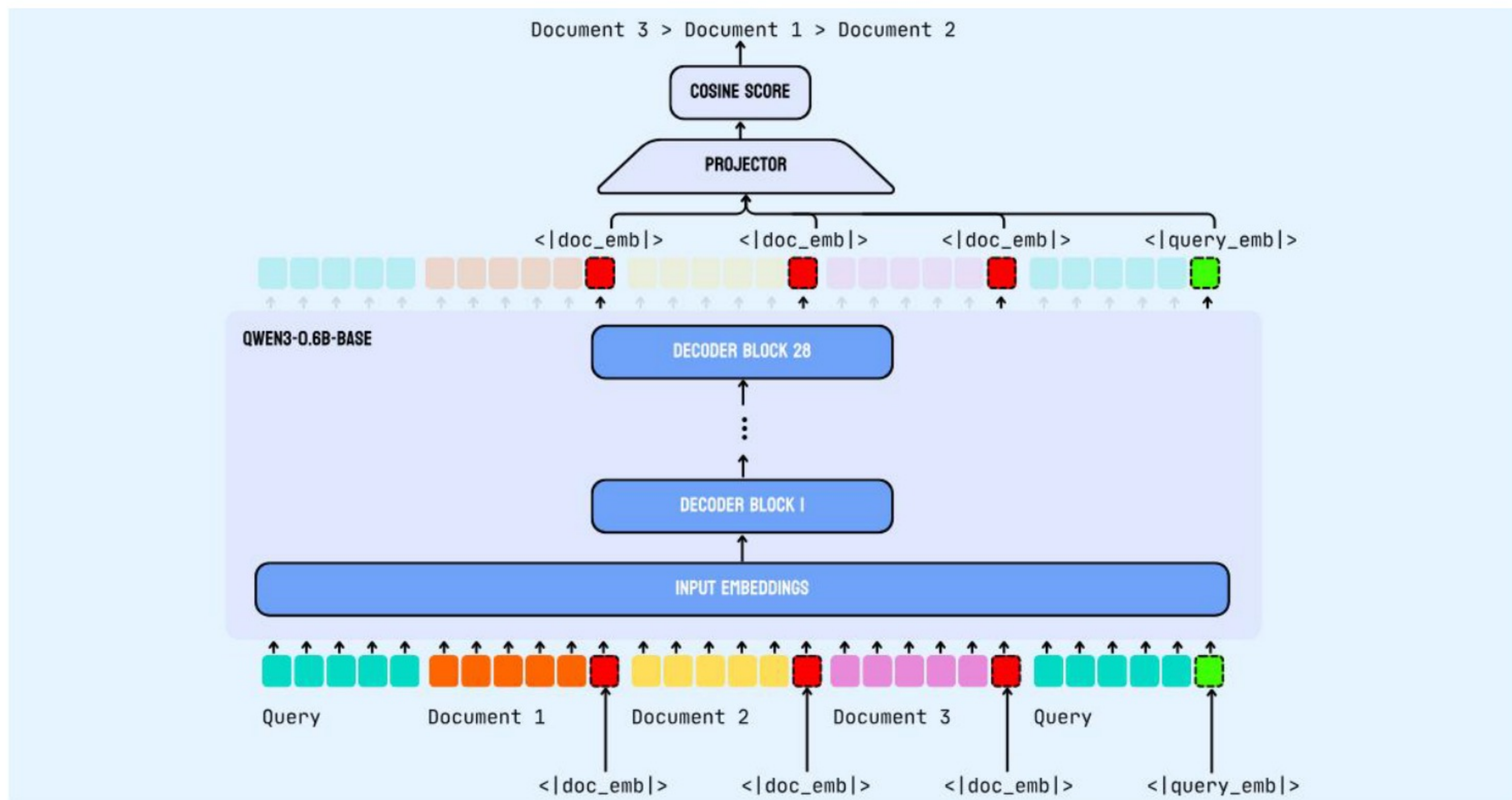


## Retrieval Benchmark



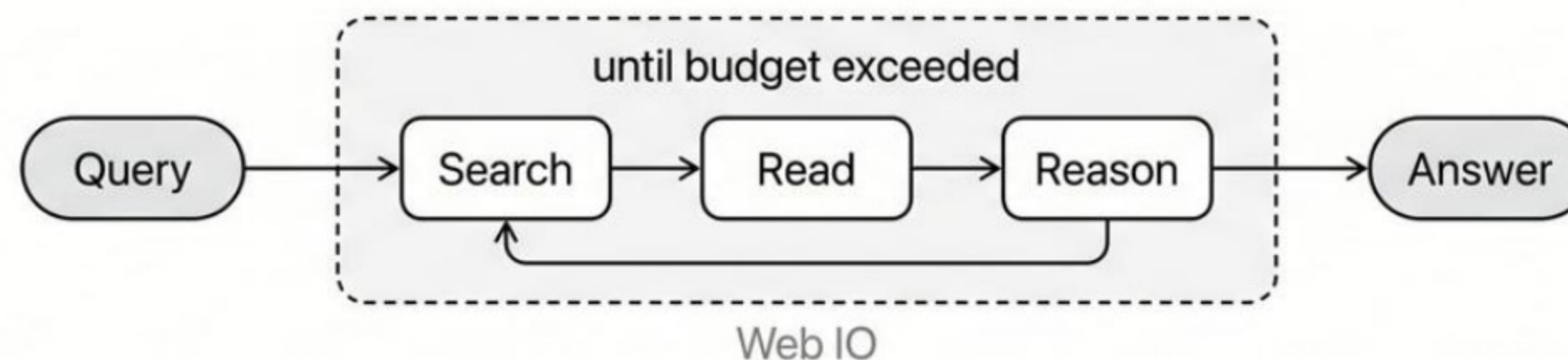


# jina-reranker-v3: 第一个长文本重排器 (AAAI 2026)

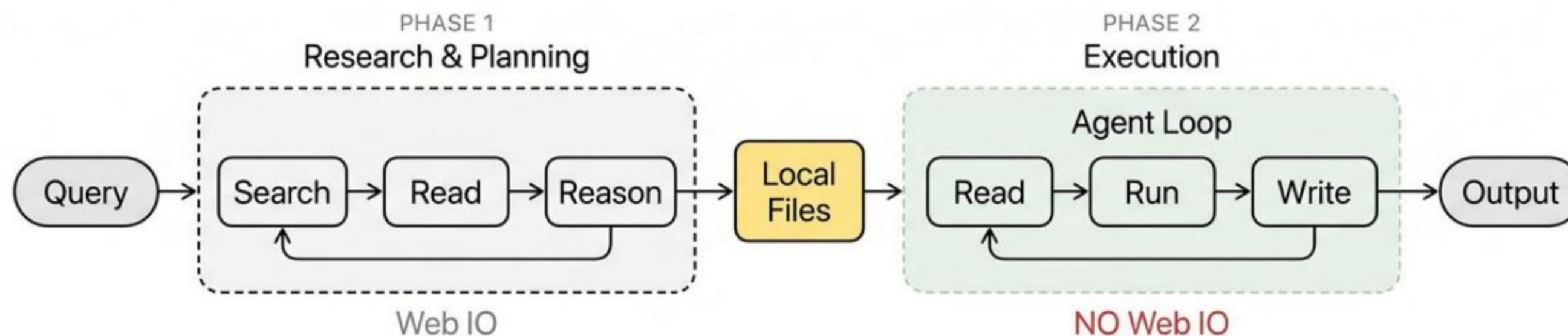


# 长任务成为 2026 年的重点

## 2025 Deep Research



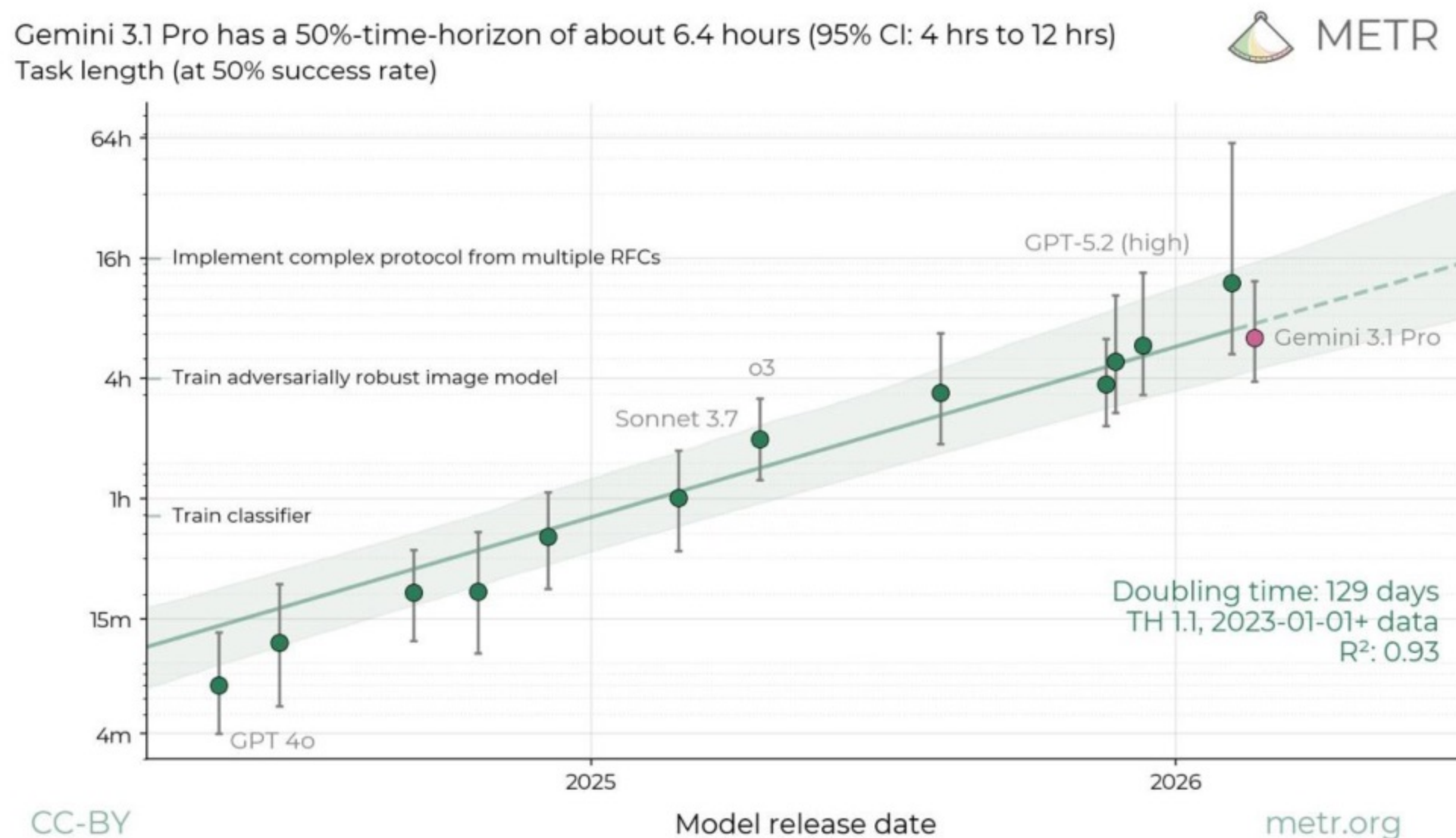
## 2026 Long-Horizon Tasks





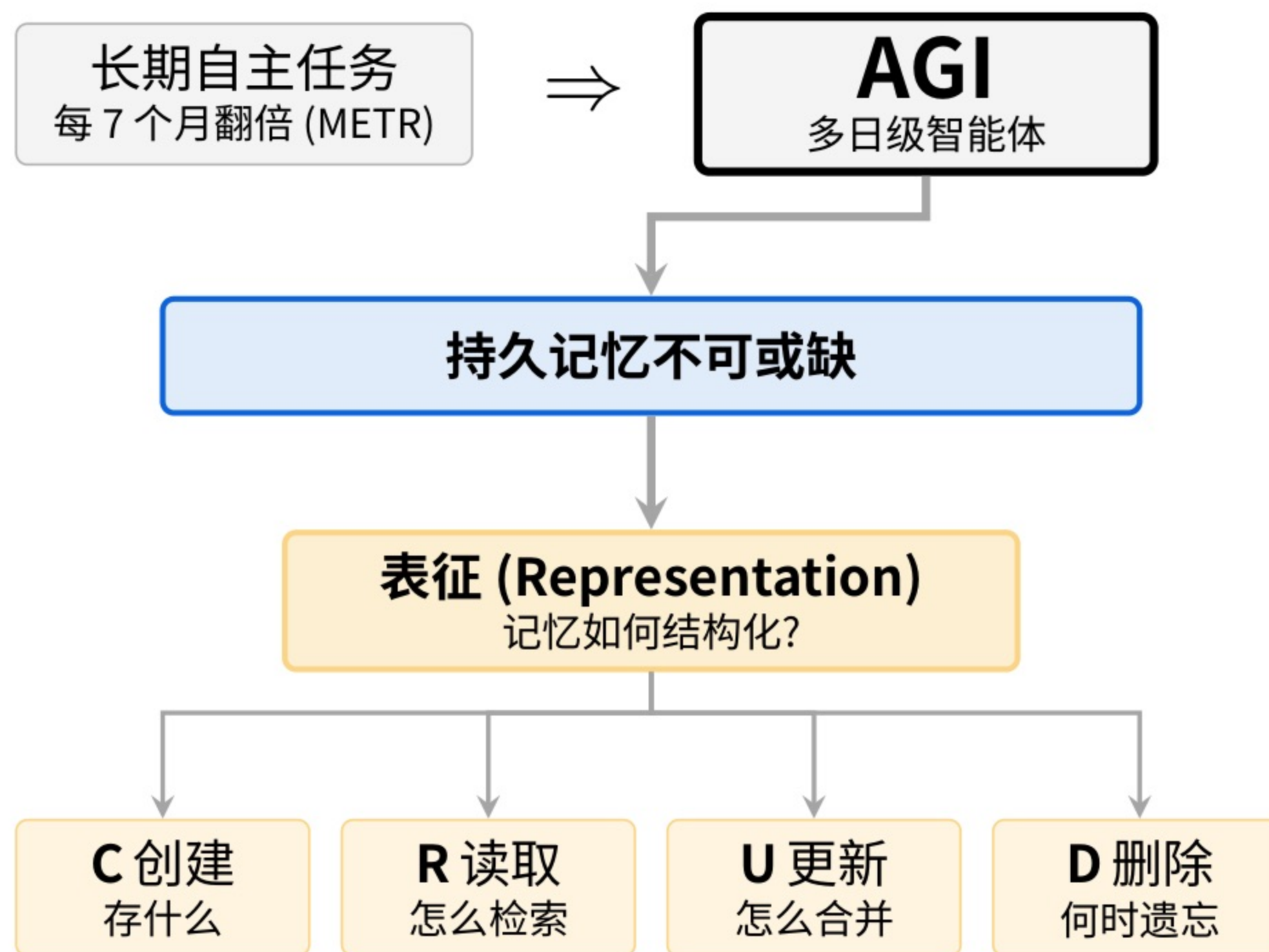
# 任务越长，记忆越重要

智能体从聊天走向长时间自主任务。METR 测量显示, 前沿模型能自主完成的任务时长每 7 个月翻一倍, 2026 年初已达数小时级别。核心矛盾: 智能体越来越强, 会话结束却失忆。



来源: METR, Task-Completion Time Horizons of Frontier AI Models, 2026.3

# 为什么记忆很重要





# 记忆问题的定义

## 核心问题

如何在不破坏对话效率的前提下，把对话中的知识变成持久、可检索的组织资产？

四个硬约束：

**不破坏效率** – 需要手动操作的方案都会失败。

**持久** – 会话历史不算，必须独立于平台生命周期。

**可检索** – 存储不是问题，检索才是。

**组织资产** – 不是个人笔记，是团队可共享的知识。

## 不解决会怎样

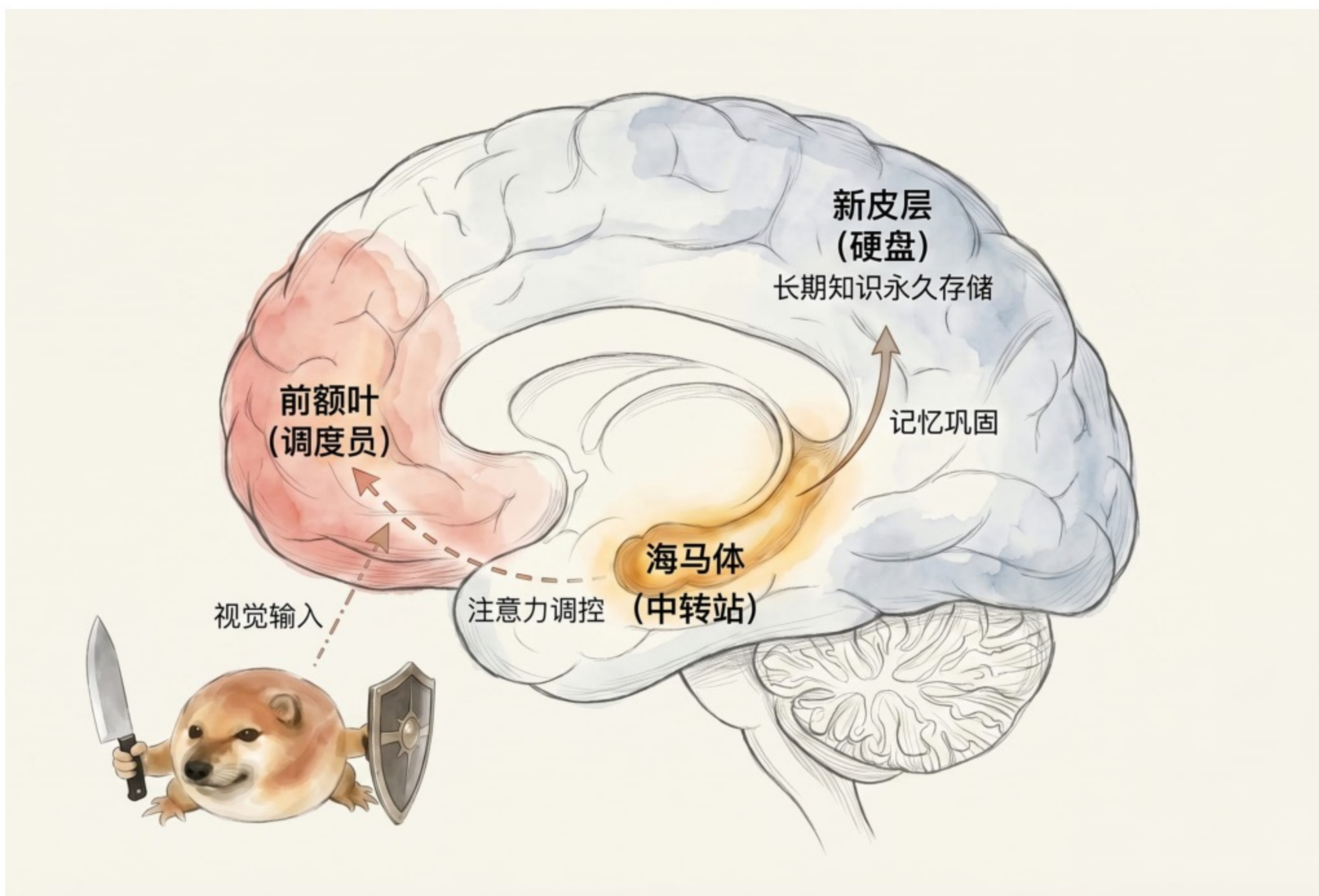
**组织失忆症** 决策推理链只存在于个人对话中，人走了推理链就永久丢失。

**决策漂移** 没有推理链锚定，技术决策无意识漂移。对话蒸发把漂移周期从「年」压到「月」。

**知识债务** 每一次未记录的决策都是知识债务。比技术债务更危险 – 你不知道丢了什么，直到急需它。



# 从生物记忆到智能体记忆



## 记忆形成流程

视觉输入  
↓  
感觉皮层（新皮层的一部分）  
↓  
前额叶判断值不值得记  
↓  
海马体编码存入  
↓  
睡眠时海马体重放  
↓  
烧录回新皮层

# 遗忘问题

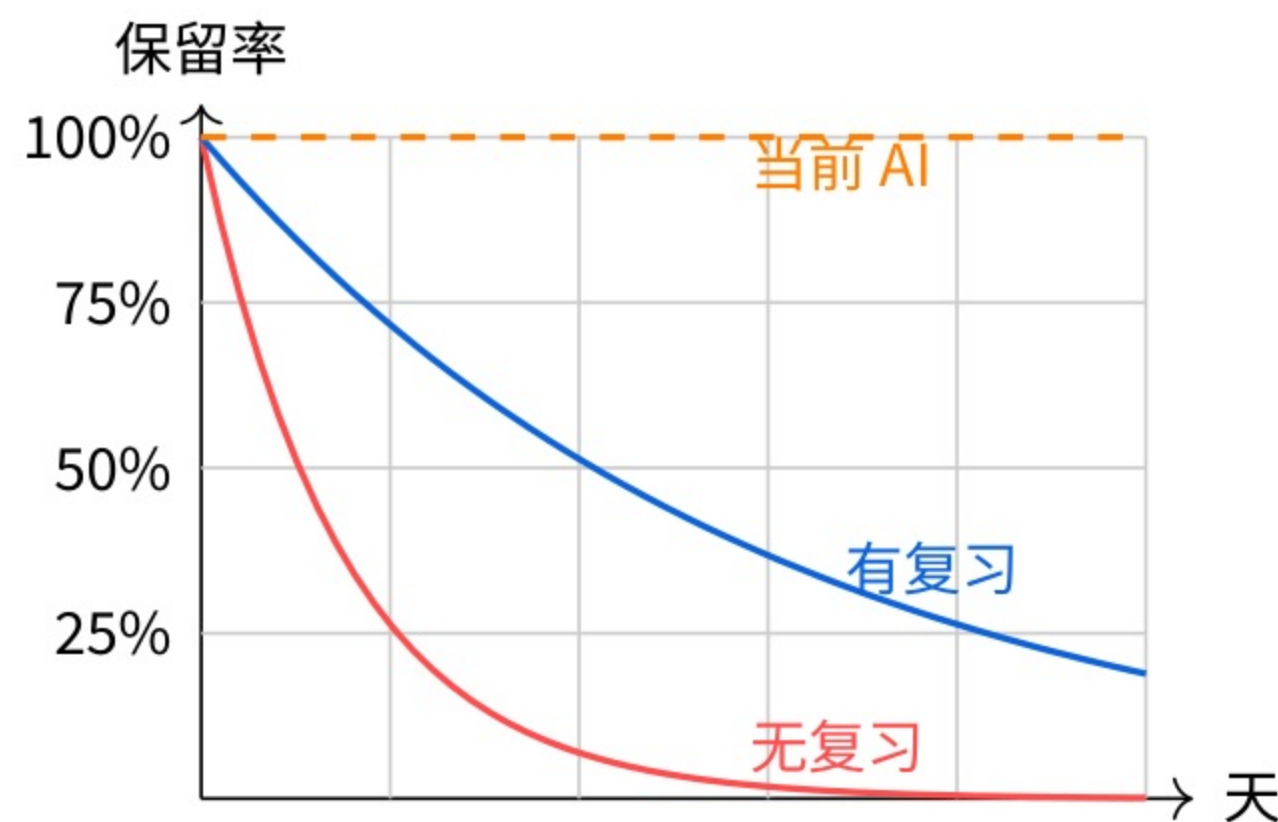
Karpathy: “所有 LLM 的个性化记忆都有一个共同问题: 两个月前的随意问题, 会被永远当作深层兴趣反复提及。” 不仅要 “记住”, 更要 “聪明地遗忘”。

## 根因分析:

- 当前智能体记忆 = 只增不减, 无衰减
- 检索权重与时间和相关性无关
- 违反生物原则: 海马体有时间依赖遗忘

**Ebbinghaus 遗忘曲线:**  $R = e^{-t/S}$

未复习的记忆指数衰减; 重复访问增强稳定性  $S$ 。  
当前绝大多数系统尚未实现任何衰减机制。





# 智能体记忆的三大缺陷

- ① **遗漏** – “明明讨论过, 系统说没有相关记忆。” 提取阶段判断 “不够重要” 而丢弃。
- ② **失真** – “记住了我说的话, 但细节不对。” 摘要压缩损失, 关键细节被简化或扭曲。
- ③ **幻觉继承** – “编造了一条我从未说过的东西。” 普通幻觉在当次对话结束; 被存入记忆的幻觉会持续污染所有后续对话。

## 后果:

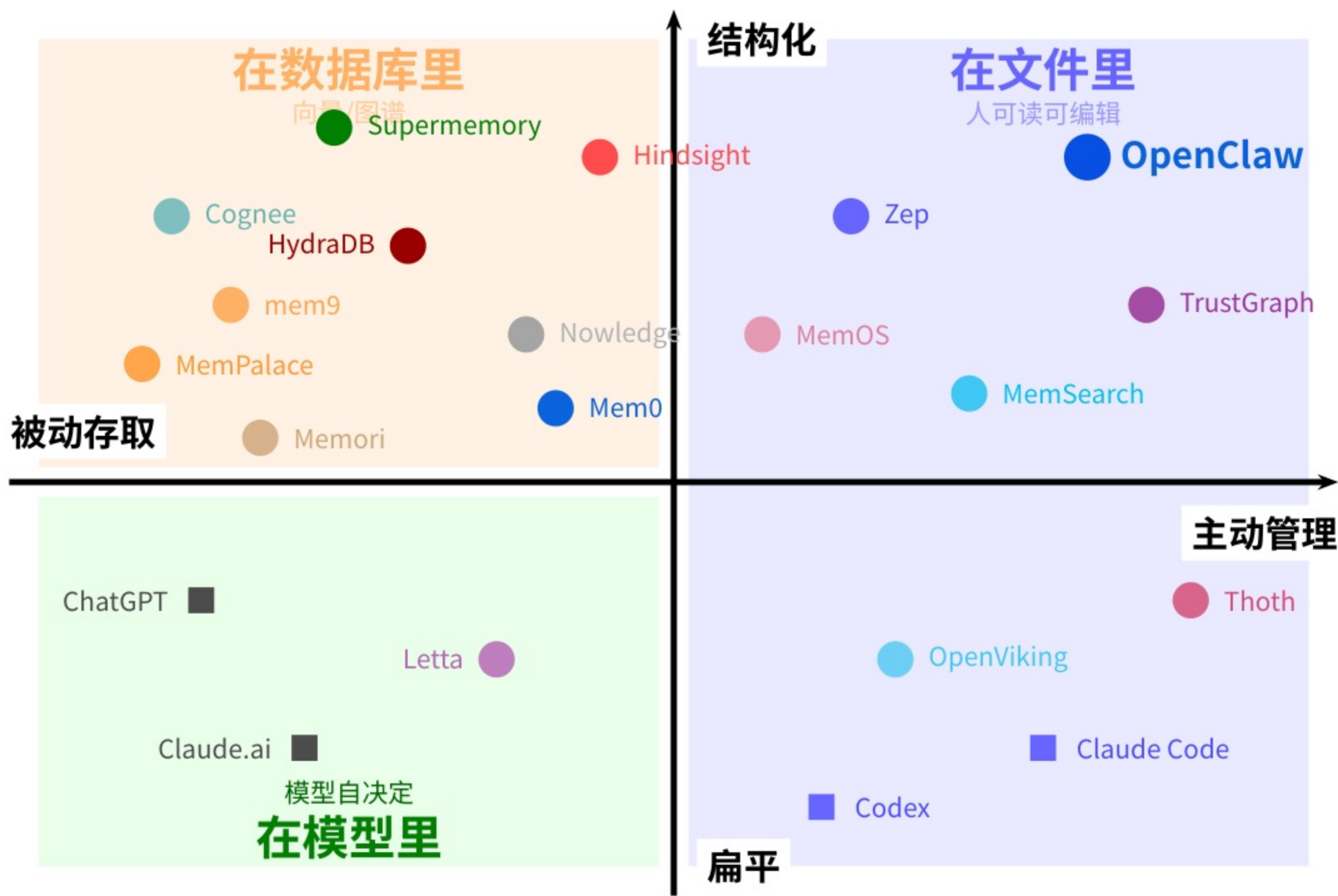
- **信任赤字** – 几次错误记忆后用户不再信任系统。验证记忆的成本 > 不用记忆的成本。
- **冷启动困境** – 记忆少时体验好 (期待低), 记忆多时体验差 (错误累积)。最佳窗口很窄, 用户必然滑出。
- **行业误导** – 错误的实现路径可能杀死正确的想法。早期 VR 头显 → “VR 不行”; 当前智能体记忆 → “AI 记忆不靠谱”?

# 智能体记忆产品总览

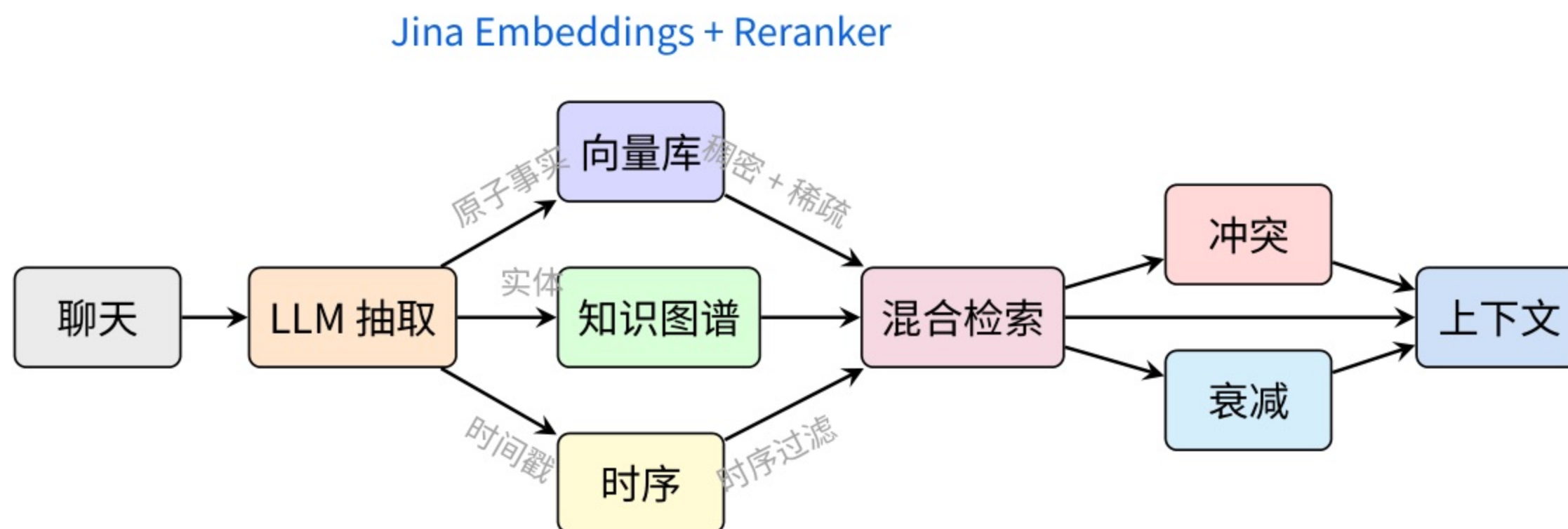
产品	公司/创始人	Stars	融资	架构	后端	LongMemEval
Mem0	Deshraj Yadav (YC)	50K	\$24M A 轮	向量 + 图谱	20+ 后端 +Neo4j	49%
Graphiti/Zep	Zep AI	24K	–	时序知识图谱	Neo4j	71.2%
Letta/MemGPT	UC Berkeley	22K	\$10M	LLM-as-OS	SQLite/PG	–
OpenViking	字节跳动/火山引擎	19K	内部	上下文数据库	火山引擎	–
Supermemory	Dhravya Shah (19 岁)	17K	\$2.6M 种子	原子事实 + 关系	Cloudflare KV	85.2%
Cognee	Cognee AI	13K	\$7.5M	ECL+Memify	LanceDB+Kuzu	–
Memori	GibsonAI (Bobur U.)	12K	–	纯 SQL	PG/SQLite	–
Nowledge	Nowledge Labs	108	–	本地个人 KG	本地 +MCP	–
Hindsight		6.4K	–	四网络 + 四路检索	图 + 向量 +BM25	<b>91.4%</b>
HydraDB	HydraDB Inc	闭源	\$6.5M	Git 式追加 KG	自建图 + 向量	90.8%
MemOS	MemTensor (上交/IAAR)	7.9K	–	记忆操作系统	Qdrant+Neo4j	75.8%
mem9	PingCAP (Ed Huang)	751	–	无状态插件	TiDB	–
平台内置记忆						
ChatGPT	OpenAI	–	–	用户事实记忆 + 历史检索	闭源	–
Claude.ai	Anthropic	–	–	项目知识库 (Knowledge Base)	闭源	–
Claude Code	Anthropic	–	–	CLAUDE.md+MEMORY.md	文件系统	–
Codex	OpenAI	–	–	文件系统 + 沙箱	文件系统	–



# 产品定位图: 记忆的 Source of Truth 在哪?

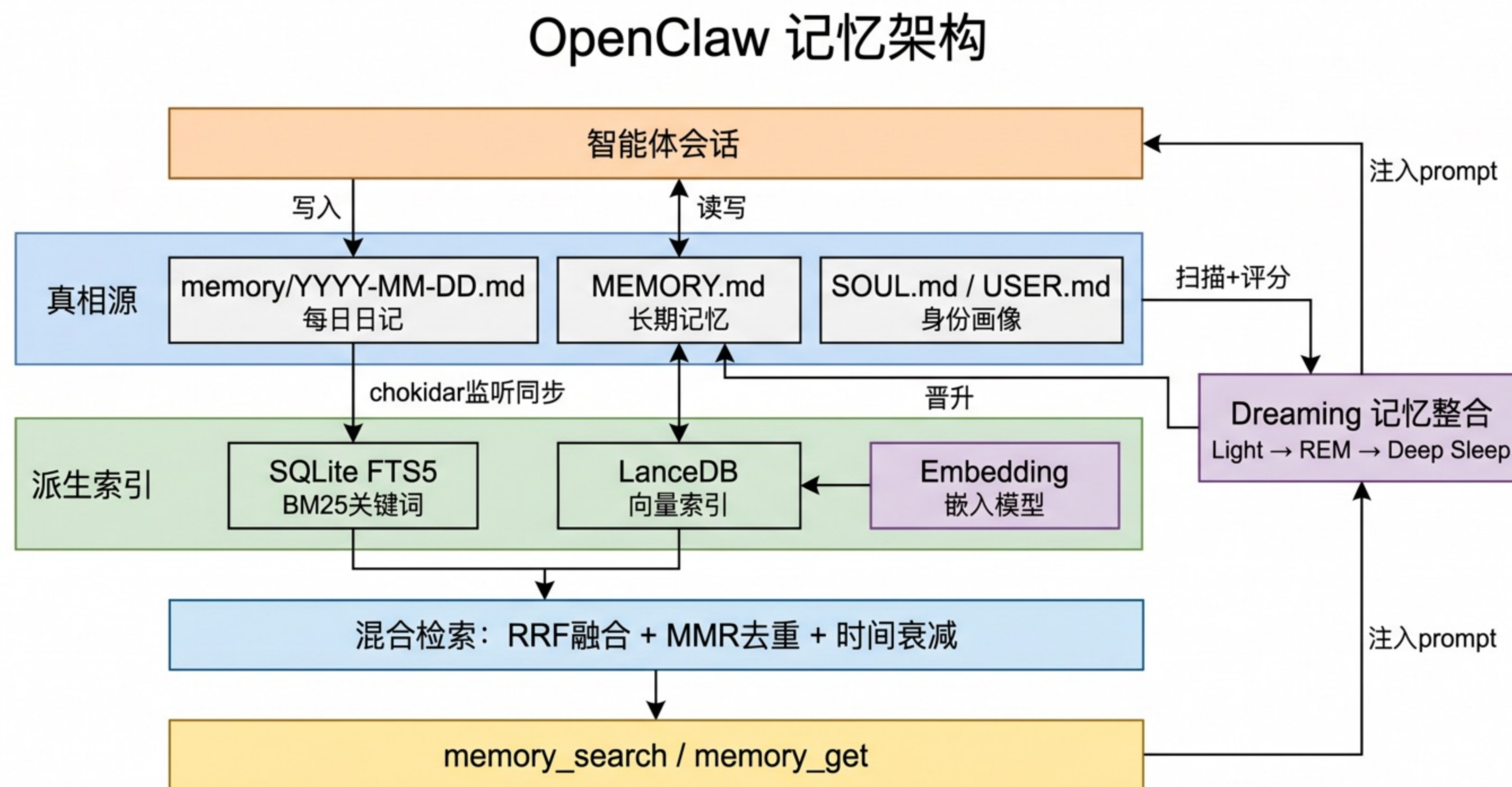


# 主流架构模式





# OpenClaw 记忆架构 (358K ★)



文件是唯一真相源。索引从文件派生，随时可删除重建。



# 智能体记忆的主要基准测试数据集

## LongMemEval (Wu et al., 2024)

500 问, 五大能力: 信息提取、多会话推理、知识更新、时序推理、安全拒答。  
LongMemEval-s 简化版。

## EverMemBench (2026.2)

首个多方协作基准, 100 万 + tokens。Alpha: oracle 下多跳仅 26%; 时序需版本语义。

## MABench (Hu et al., ICLR 2026)

四大能力: 精确检索、测试时学习、长程理解、冲突解决。含 EventQA 和事实整合数据集。

## MemBench (ACL 2025)

事实记忆 vs 反思记忆, 参与 vs 观察。Alpha: 反思记忆远难于事实提取。

## MemoryArena (2026.2)

记忆-智能体-环境循环, 评测跨会话任务完成能力。

## MemoryBench (清华, 2025.10)

首个用户反馈持续学习基准, 11 数据集。Alpha: 现有系统无法利用程序性记忆。

## LOCOMO (Snap Research)

10 段超长对话 (>300 轮), 问答 + 事件摘要 + 多跳推理。Mem0 报告 +26%。

## Letta Leaderboard

评测 LLM 自主记忆管理。文件方案 74.0% 超 Mem0 68.5%。



# 关键论文

2026 年的论文普遍认为: (1) RAG 作为 “无状态查找表” 不够, 记忆需要**状态管理**; (2) 图结构是组织记忆的优选; (3) 遗忘/衰减是开放问题; (4) 多模态和多智能体记忆是下一个前沿。

论文	arXiv	核心贡献
Memory in the Age of AI Agents	<a href="#">2512.13564</a>	47 人大型综述。三维分类: 形式 (token/参数/潜在), 功能 (事实/经验/工作), 动态 (形成/演化/检索)
Memory for Autonomous LLM Agents	<a href="#">2603.07670</a>	write-manage-read 循环形式化。5 大机制族: 上下文压缩、RAG、反思改进、层次虚拟上下文、策略学习
AriadneMem	<a href="#">2603.03290</a>	解耦两阶段: 离线构建 (熵感知门控 + 冲突感知粗化) + 在线推理 (算法桥发现)。Multi-Hop F1 +15.2%, 仅需 497 tokens
CMA (Continuum Memory)	<a href="#">2601.09913</a>	定义持续记忆架构: 持久存储、选择性保留、关联路由、时间链接、抽象整合
Graph-based Agent Memory	<a href="#">2602.05665</a>	图视角全面综述。生命周期: 抽取 → 存储 → 检索 → 演化。开源库和基准测试汇总
HydraDB / Cortex MemOS	<a href="#">2507.03724</a>	Git 式版本化时序图谱 + 滑动窗口 + 三层重排。LongMemEval-s 90.8% 记忆操作系统: MemCube 统一三类记忆 (明文/激活/参数), 生命周期调度 + 版本控制 + 权限管理。LoCoMo 75.8%, LongMemEval +40.43% vs OpenAI
Hindsight		四网络记忆分离 (事实/经验/观察/观点) + 四路并行检索 + 交叉编码器重排。LongMemEval SOTA 91.4%

# Jina 在哪一层

搜索三大件: 向量化 + 精排 + 阅读器 = 每个记忆层下面的默认搜索基础设施。

智能体框架 (OpenClaw, Letta, LangChain, CrewAI, ...)

记忆逻辑层 (事实抽取, 生命周期管理, 冲突解决)

存储 + 搜索 (**Elasticsearch**, 向量库, 图数据库, SQL)

**搜索三大件: Jina Embeddings + Reranker + Reader**

Jina + ES