

QCon  
全球软件开发大会

# 从上下文到长期记忆：大模型记忆工程的架构设计与实践

李志宇

记忆张量（上海）科技有限公司 联合创始人兼CTO



📍 北京

**QCon**

全球软件开发大会

会议时间: 4月10-12日

- 大模型赋能 AI Ops
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

**AICon**

全球人工智能开发与应用大会

会议时间: 6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

**QCon**

全球软件开发大会

会议时间: 10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

**AICon**

全球人工智能开发与应用大会

会议时间: 5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

**AICon**

全球人工智能开发与应用大会

会议时间: 8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

**AICon**

全球人工智能开发与应用大会

会议时间: 12月19-20日

- 通用大模型
- 智能硬件
- LM Ops
- 具身智能

# 极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



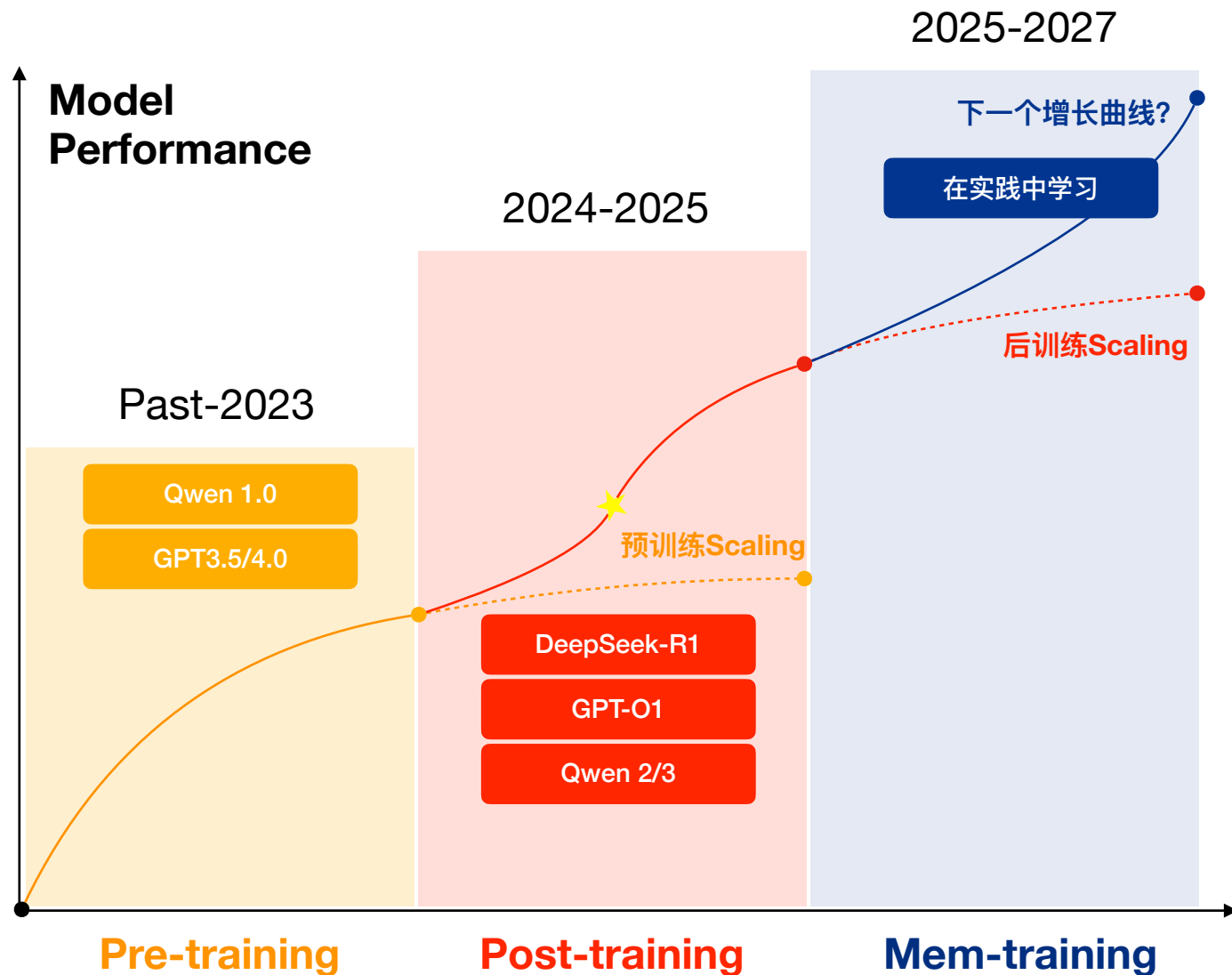
参会咨询



查看会议



# 大模型性能缩放曲线的演进历史



OpenAI

## GPT-4 更新版 (2025年4月)

“兴奋到失眠”的新功能

ChatGPT **全局记忆**

## GPT-5 (2025年8月)

**记忆能力升级**

与更多应用进行整合

## GPT-6 (2026年?)

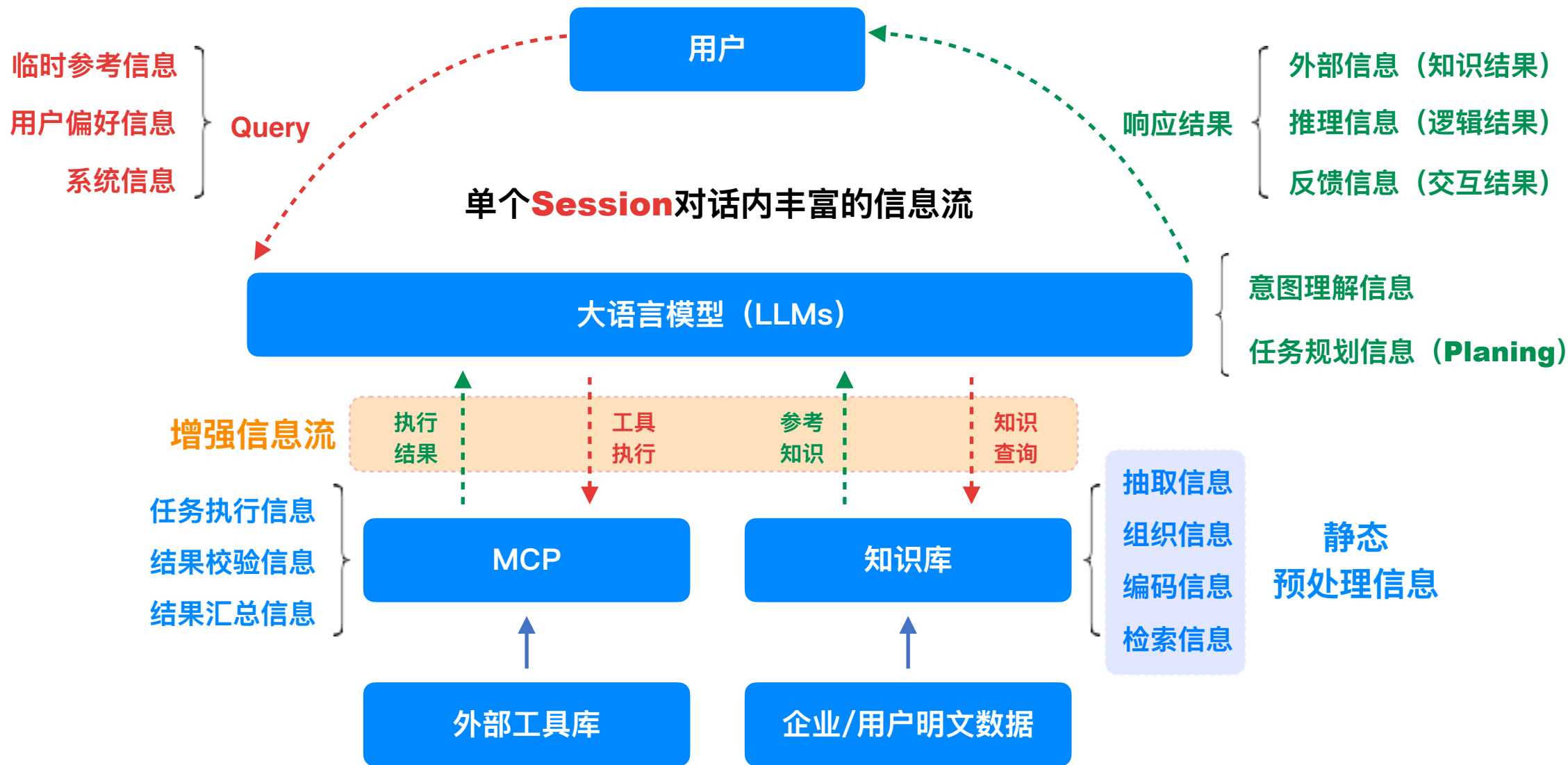
反复强调:

“People want memory”,  
记忆与**个性化**是迭代主线





# 从实践层面看记忆增强的必要性



大模型与用户日常交流过程中形成的信息流是模型持续迭代提升的最优资源！





对于单个用户的单个 **Session** 而言，需要管理：

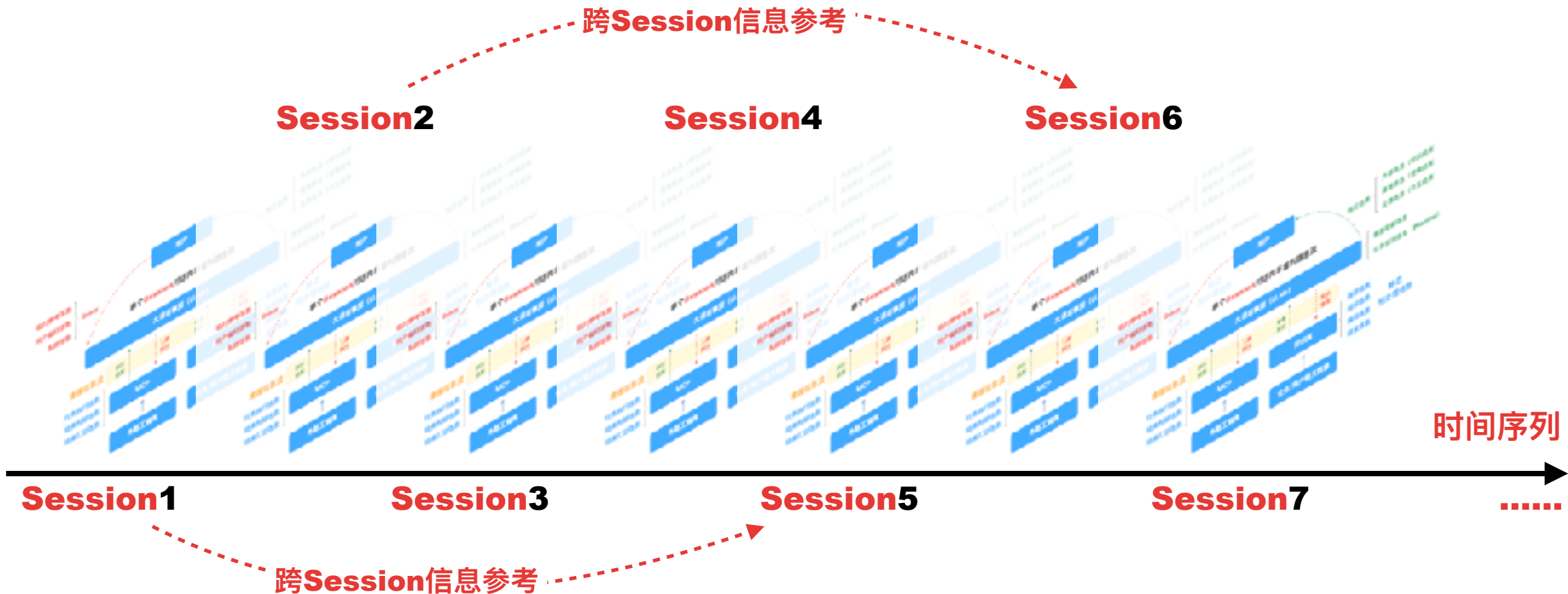
◆**动态信息**：临时参考信息、偏好信息、系统信息、MCP执行信息、任务信息、响应信息（外部、推理、反馈交互） ....

◆**静态信息**：本地知识库、云端知识库...(知识处理的完整流程框架)



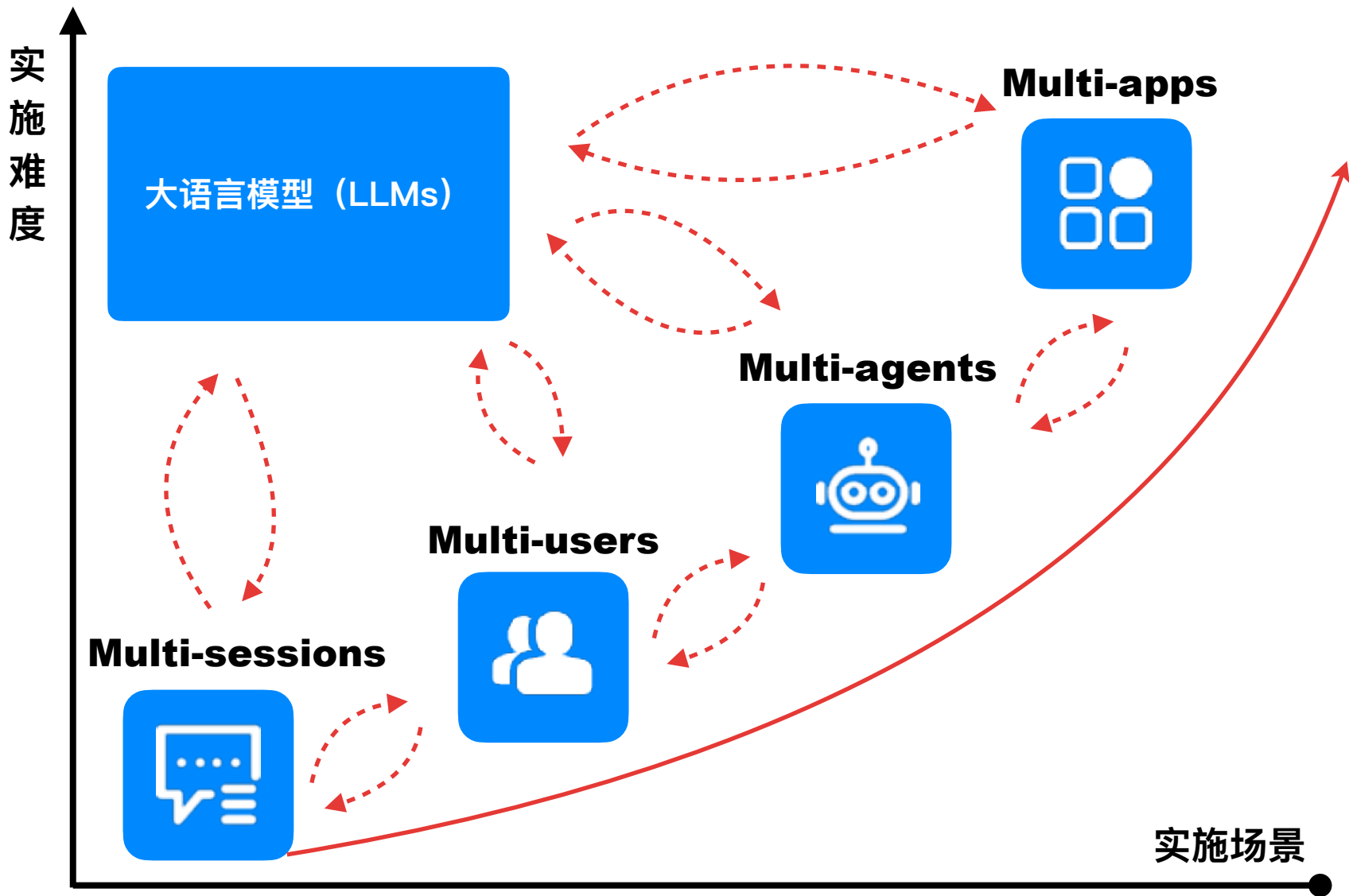
# 从实践层面看记忆增强的必要性

对于单个用户的 多个 **Session** 而言，需要管理：保障跨**Session**引用的正确性，整体信息的无歧义等

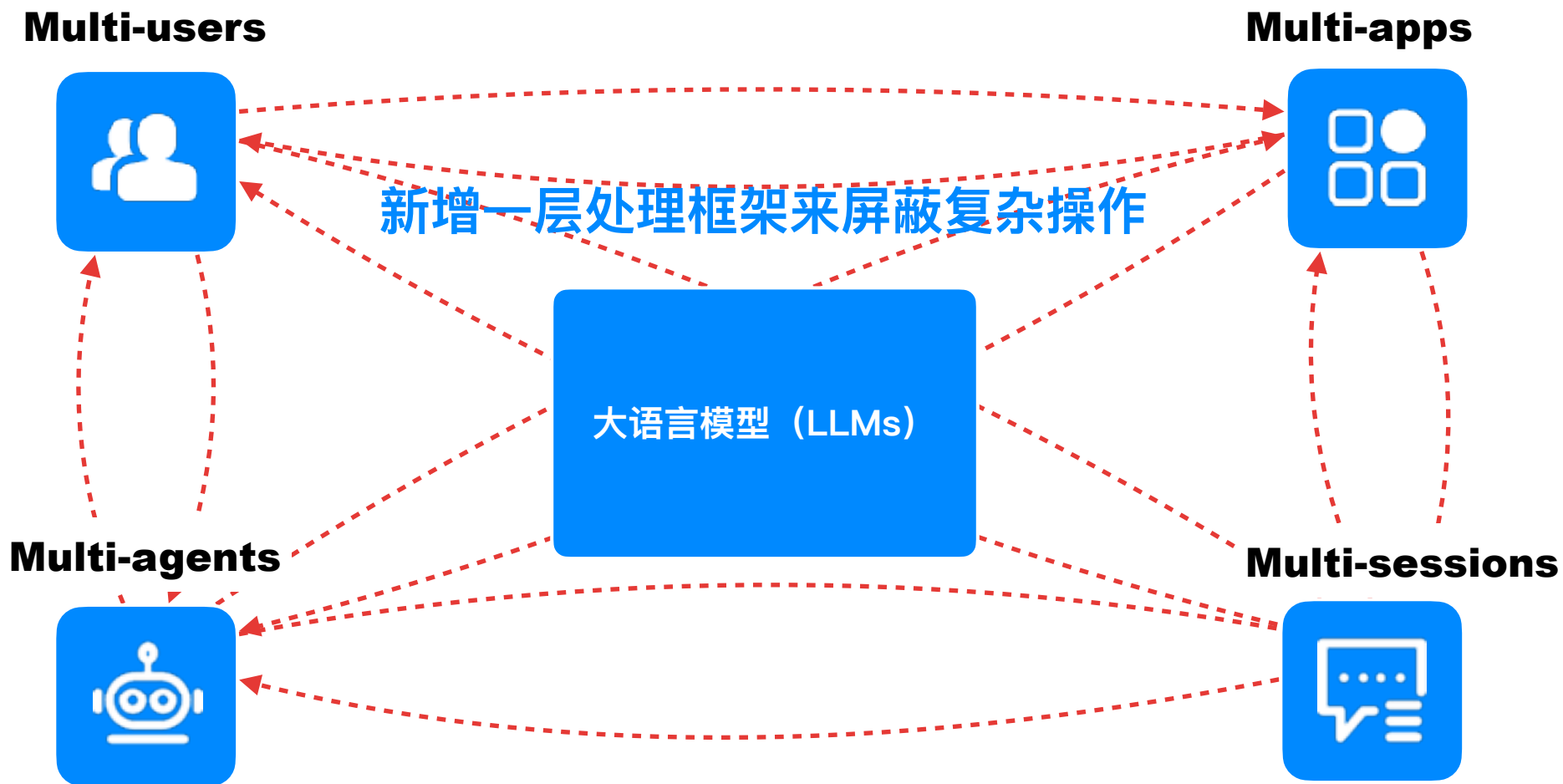




# 从实践层面看记忆增强的必要性



应用开发  
复杂度急剧增长







# 大模型 记忆增强层 的实现路径：（1）模型内生驱动的记忆增强



模型内生驱动的记忆增强：通过设计创新的基座模型架构，引入记忆增强的能力，强化模型的性能。

代表性工作	作者团队/时间	核心方案	技术特点
Memorizing Transformers	Google 2022	把局部上下文注意力和外部记忆检索融合，引入外部记忆	首次在语言模型中引入外部记忆联合模型解码
Focused Transformer	IDEAS NCBR 2023	引入对比训练，让 KV 空间更好区分上下文相关/无关信息	方法侧重训练策略改进，不改架构，可迁移到
MemoryLLM	清华大学 2024	在每层引入固定大小 memory tokens，作为可更新参数池	强调内置可更新记忆，能持续吸收新知识并抗遗忘
Memory3	记忆张量 2024	基于不同记忆类型进行分层管理和建模，缩减主干参数	首次提出记忆分层框架，对模型记忆进行分层建模
WISE	浙江大学 2024	提出双参数记忆：主记忆存预训练知识，侧记忆存编辑知识	面向 lifelong model editing进行记忆编辑
Titans	Google 2025	提出神经长时记忆模块，学习存储/遗忘	模拟人类记忆分层，支持超长上下文
MemAgent	ByteDance 2025	基于强化学习的Agents上下文外推扩充方案	侧重在短期记忆的扩充



# 大模型 记忆增强层的实现路径：（1）应用外向驱动的记忆增强



模型外向驱动的记忆增强：通过设计Prompt / Agent 流，模拟记忆过程，增强模型性能

代表性框架	时间	Slogan	技术特点	Star数量
Letta /MemGPT	2023	Create stateful AI agents that truly remember, learn, and evolve.	开源社区驱动，目标构建具备状态记忆与自我改进能力的 LLM Agent 平台，非 Production-grade.（是最早受到计算机系统启发设计记忆框架的）	18.2k
Mem0	2024	Universal memory layer for AI Agents.	纯明文记忆管理框架，较早面向应用层提供记忆管理的代表性框架。服务生产环境，强调平台化服务。	39.2k
Zep	2024	Build Agents That Recall What Matters.	强调采用 Temporal Knowledge Graph（时序知识图谱）结构来组织长期记忆	17.6k
Memobase	2025	Profile-Based Long-Term Memory for AI Applications.	强调用户画像与事件时间线的记忆系统，将会话内容抽取为结构化 profile 和事件，并关联时间戳，形成用户长期记忆	2.1k
Memories.ai	2025	Building AI to See and Remember	强调多模态记忆，把原始视频转化成可搜索、带上下文关联的数据库，支持SaaS服务	非开源
HippoRAG团队、MIRIX团队、北大、清华、交大、人大等			从 Memory面临的创新问题出发，提出了对应的解决方案	

对比项	基于模型驱动	基于应用驱动
定位	从模型底层嵌入记忆机制，改变模型本体	在应用层叠加记忆系统，管理交互与任务信息
关注点	提升模型的认知与学习能力	提升系统的连续性与个性化体验
实现方式	新型架构、新的训练策略、新的建模策略	明文存取与索引-会话/任务级状态管理
优势	记忆读取效率更高，性能上限高	记忆写入效率更高、落地快，易扩展
局限	研发成本高，落地周期长	依赖底层模型，缺乏深度学习，幻觉严重
价值导向	推动 AI 基础能力演进	驱动应用生态与商业落地

对比项	MemOS 融合范式		基于模型驱动	基于应用驱动
定位	内外记忆协同，形成系统级记忆一体化		从模型底层嵌入记忆机制，改变模型本体	在应用层叠加记忆系统，管理交互与任务信息
关注点	兼顾认知深度与应用广度		提升模型的认知与学习能力	提升系统的连续性与个性化体验
实现方式	分层协同+多触点调度		新型架构、新的训练策略、新的建模策略	明文存取与索引-会话/任务级状态管理
优势	读写效率全局最优		记忆读取效率更高，性能上限高	记忆写入效率更高、落地快，易扩展
局限	设计难度高，开发+理论双重要求		研发成本高，落地周期长	依赖底层模型，缺乏深度学习，幻觉严重
价值导向	构建系统级可持续演进的个性化记忆		推动 AI 基础能力演进	驱动应用生态与商业落地

模型驱动决定上限，应用驱动决定下限！ 需要从系统层面结合两者！



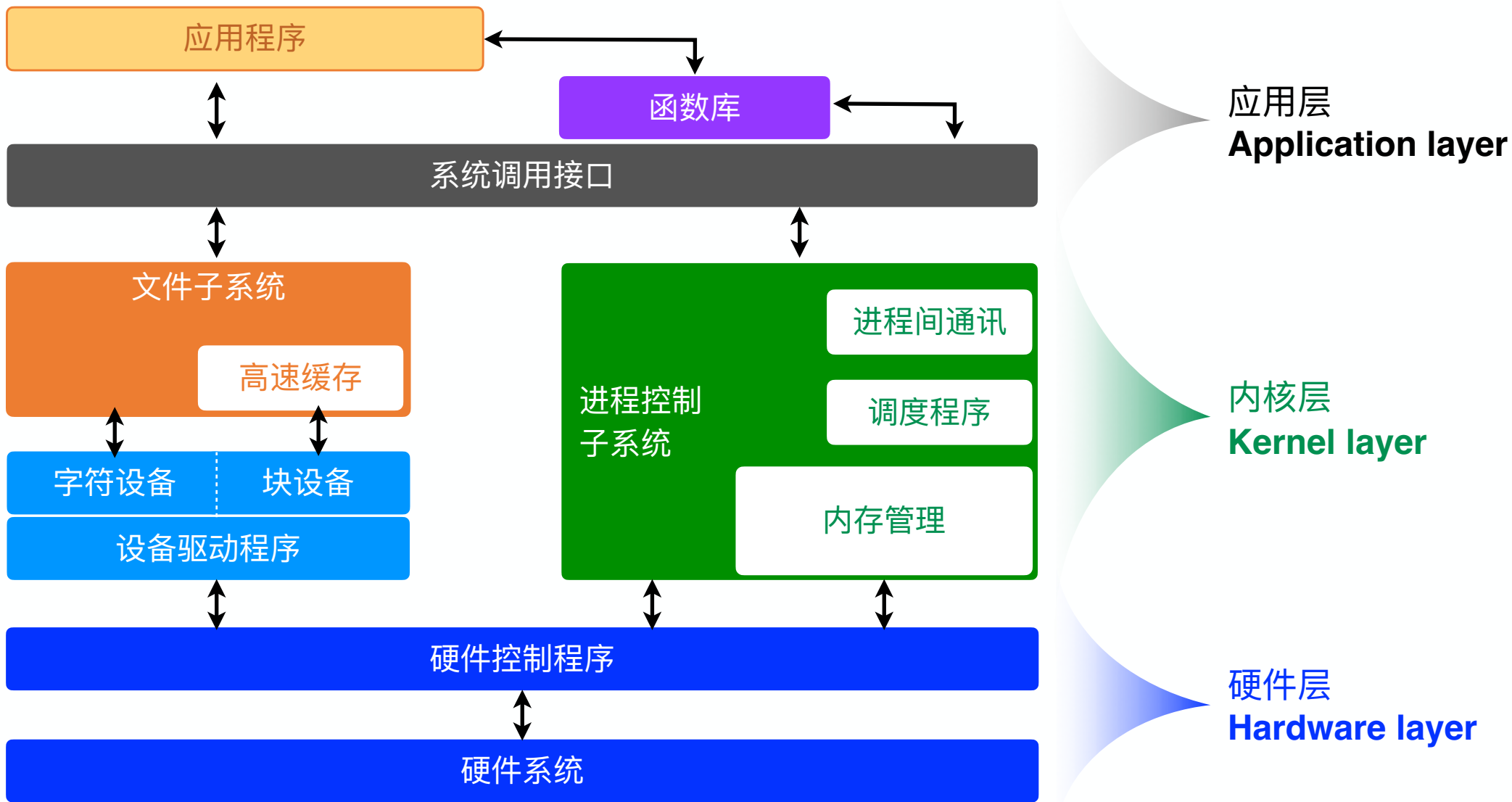
## 记忆系统的核心功能点

不仅依赖更大的模型与干净的数据，更需要一套完整的 记忆操作系统功能链路 —— 让 AI 能记住、能组织、能调用、能更新、还能共享。

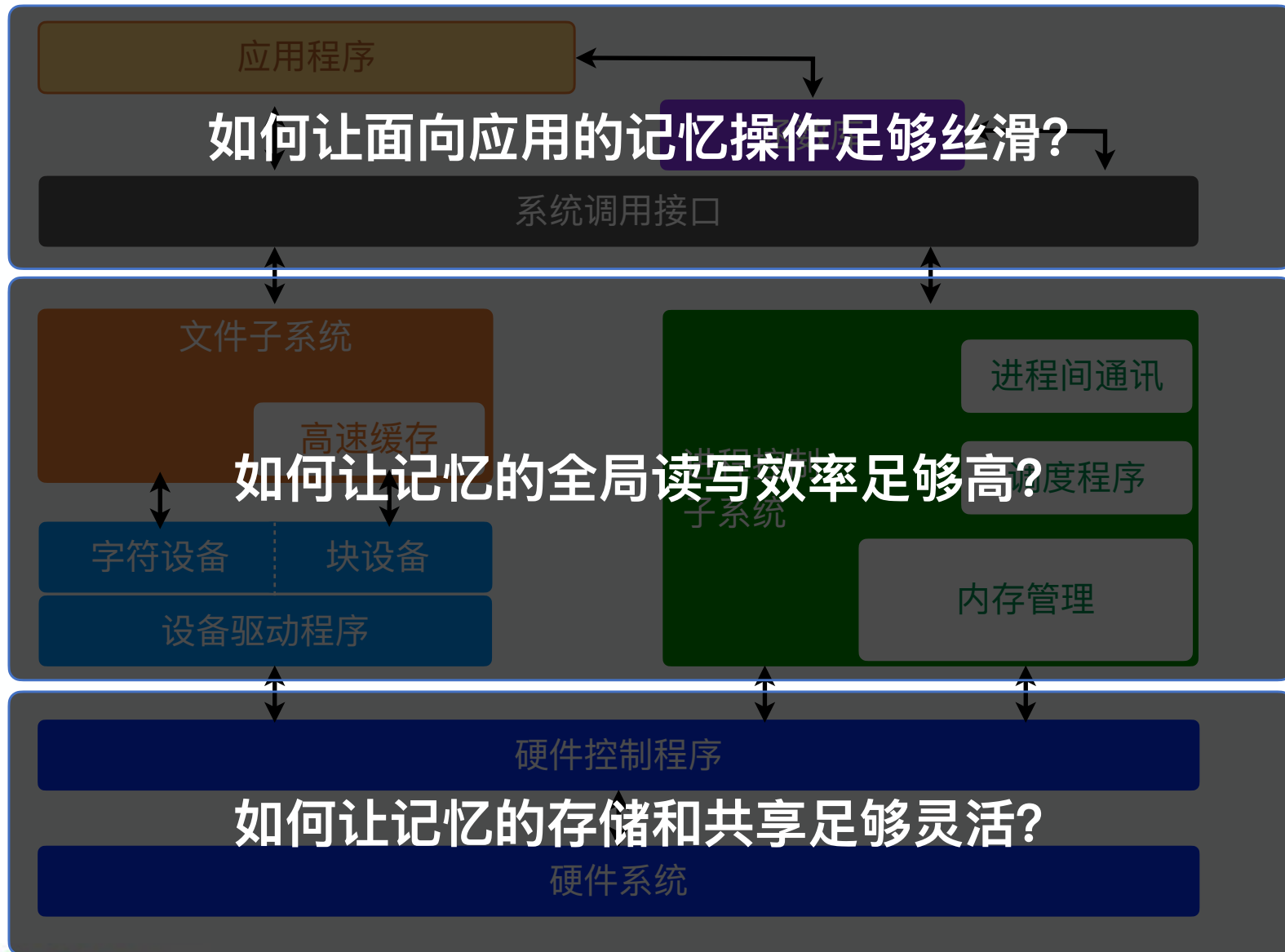




# MemOS 的核心设计思路：从 OS 到 记忆操作系统 (MemOS)



# MemOS 的核心设计思路：从 OS 到 记忆操作系统 (MemOS)



应用层  
Application layer

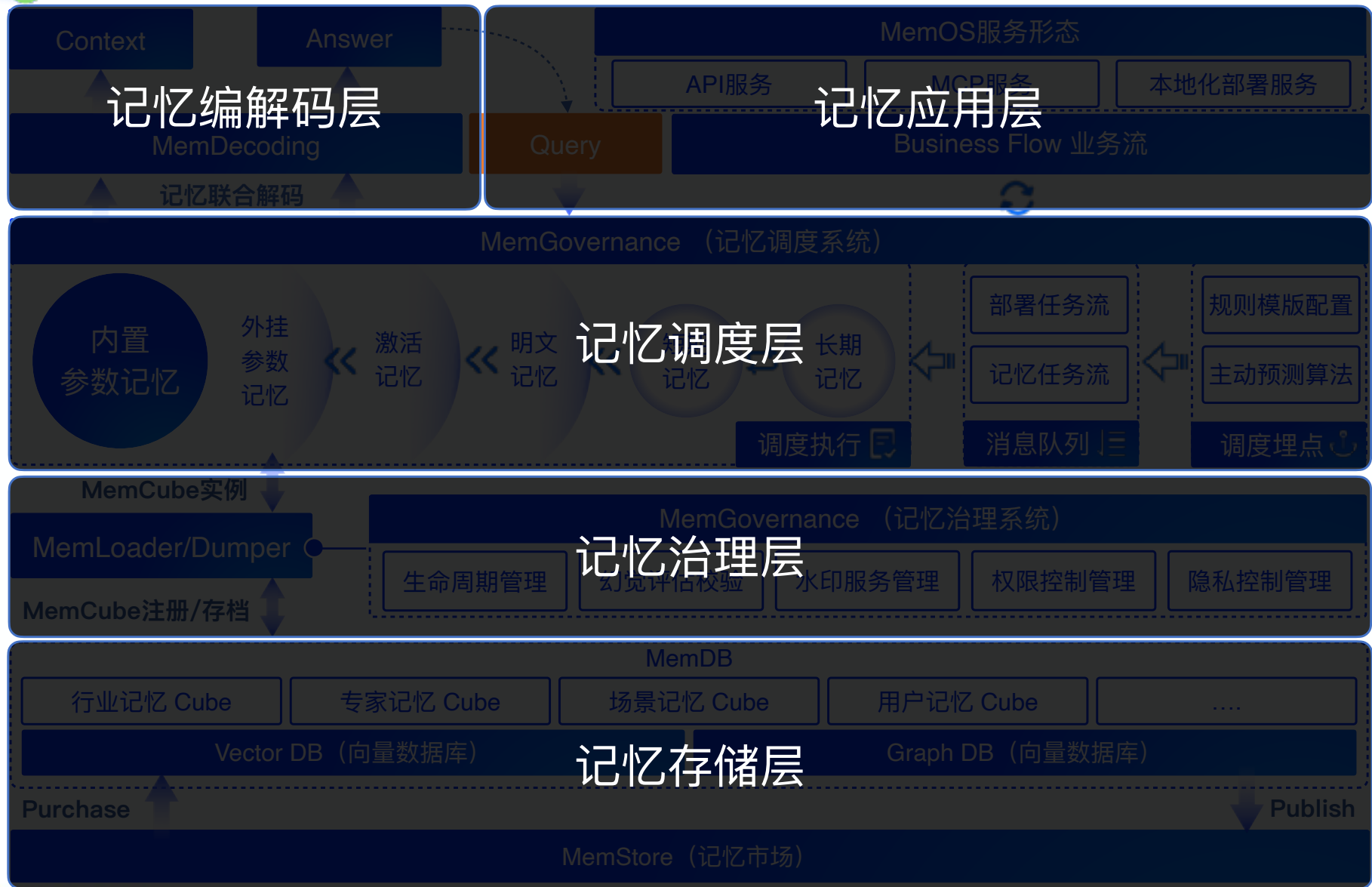
内核层  
Kernel layer

硬件层  
Hardware layer





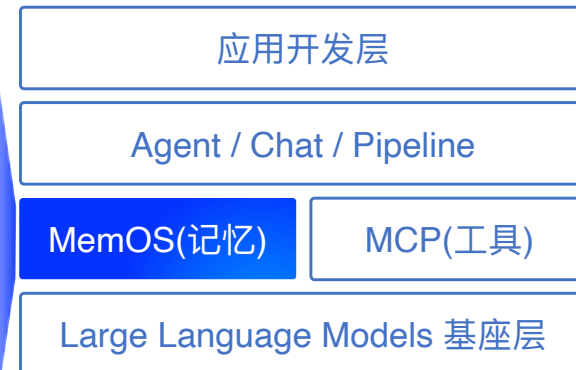
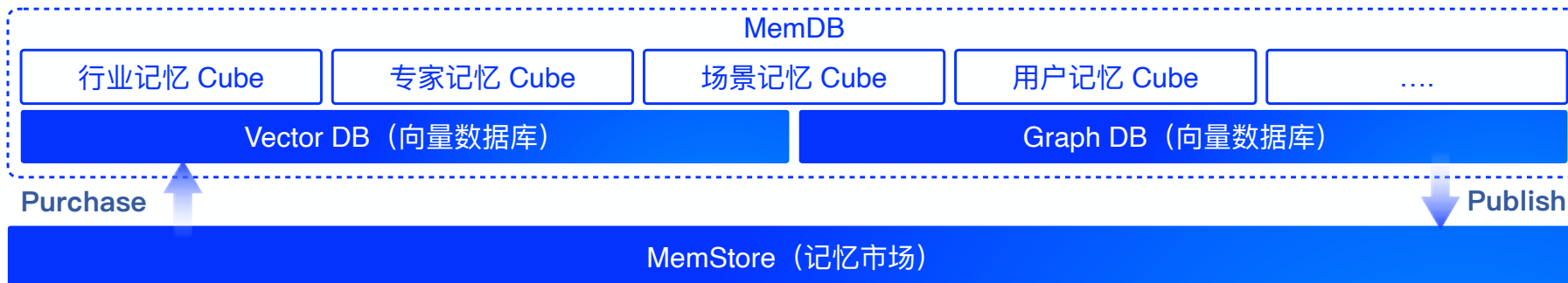
# MemOS的系统框架



## 记忆存储层

MemCube: 可独立打包的最小记忆单元

MemStore: 可交易的记忆市场平台





## 记忆治理层

MemHaluEval: 面向记忆系统的幻觉评估框架

MemControl: Agent驱动的记忆权限管理器



## 记忆调度层

MemTrigger: 可配置的多粒度调度触发模版

记忆分层: 融合场景的记忆分布建模机制



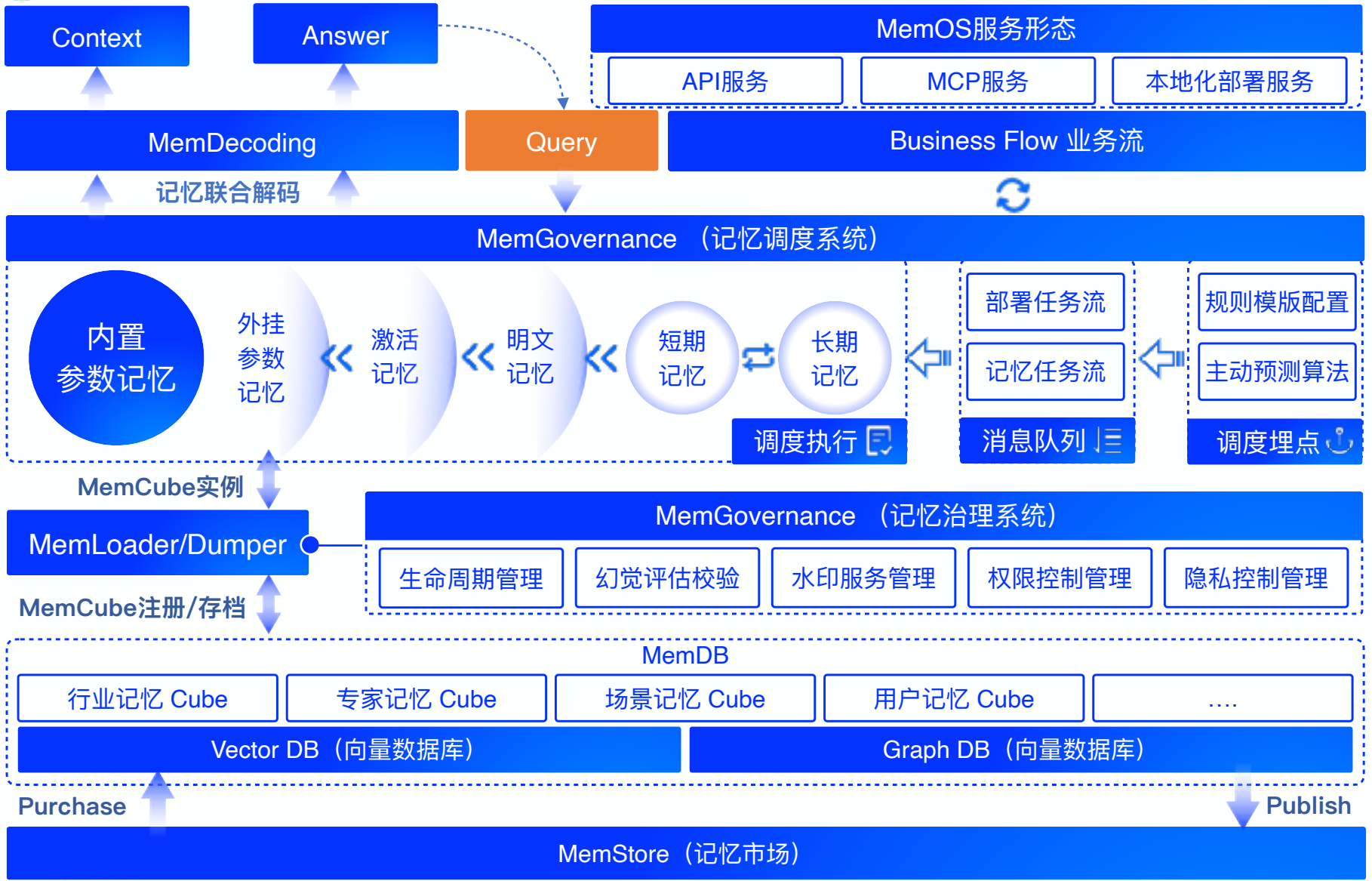
# MemOS的系统框架

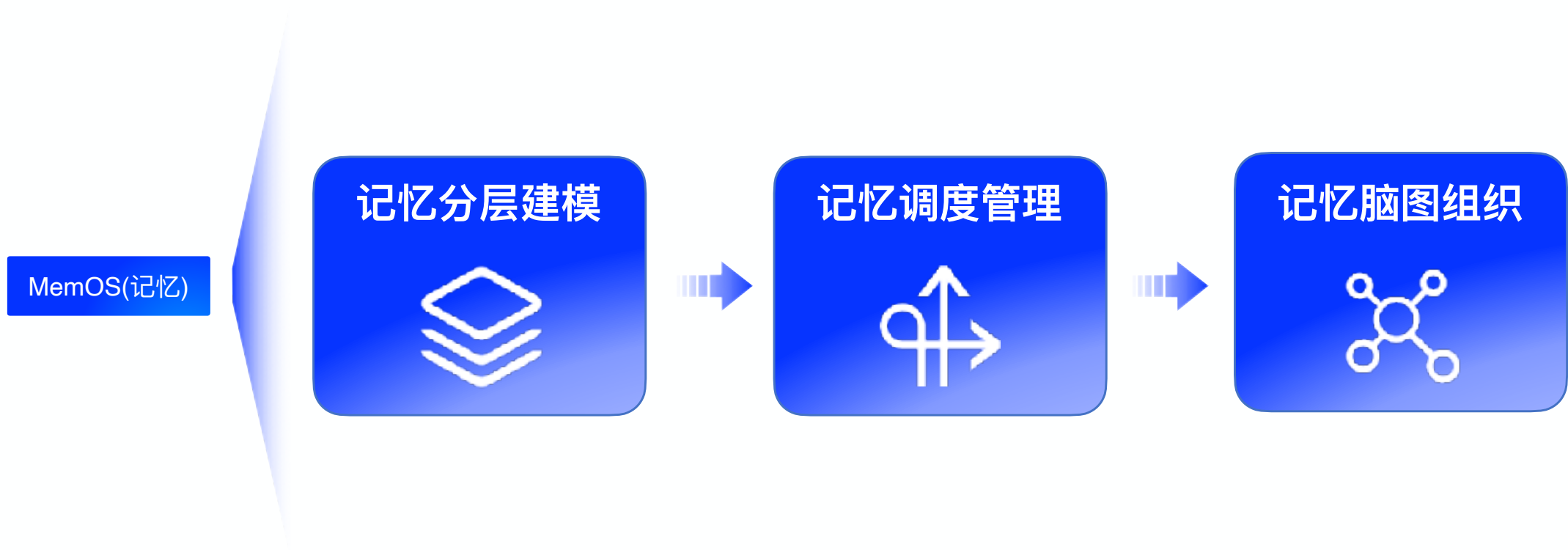


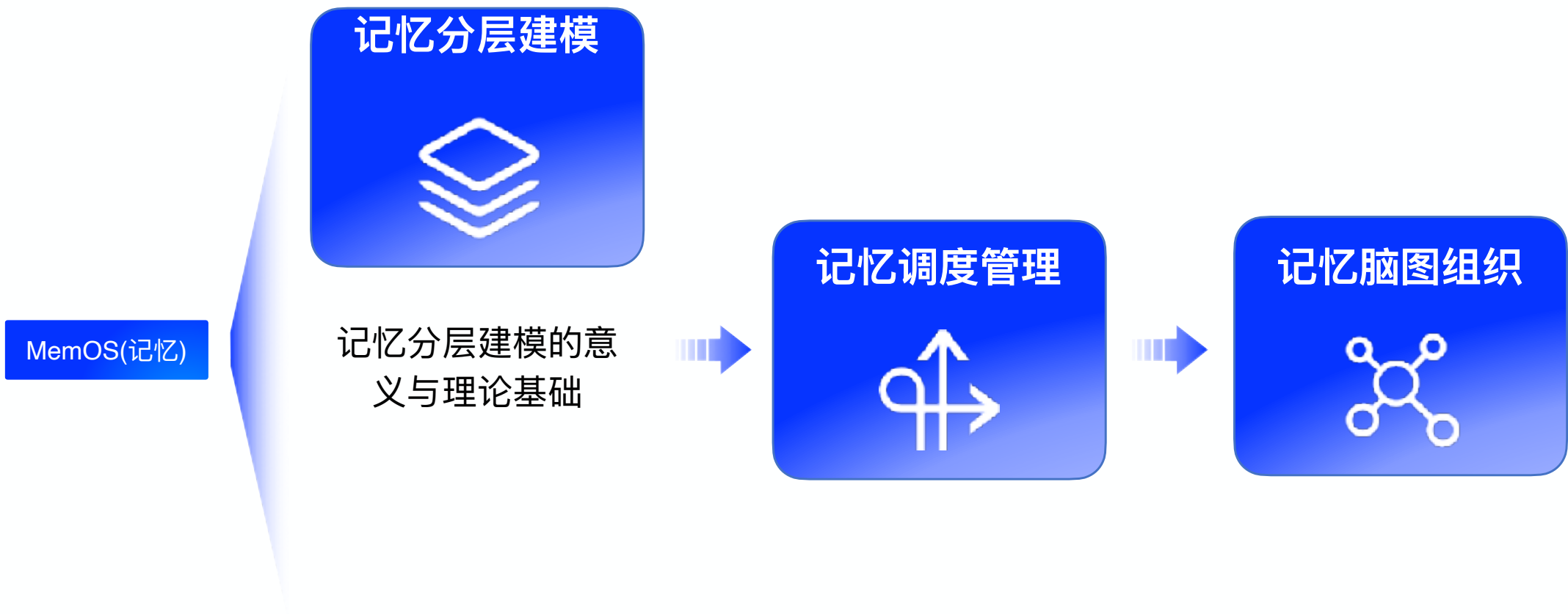
## 应用与编解码

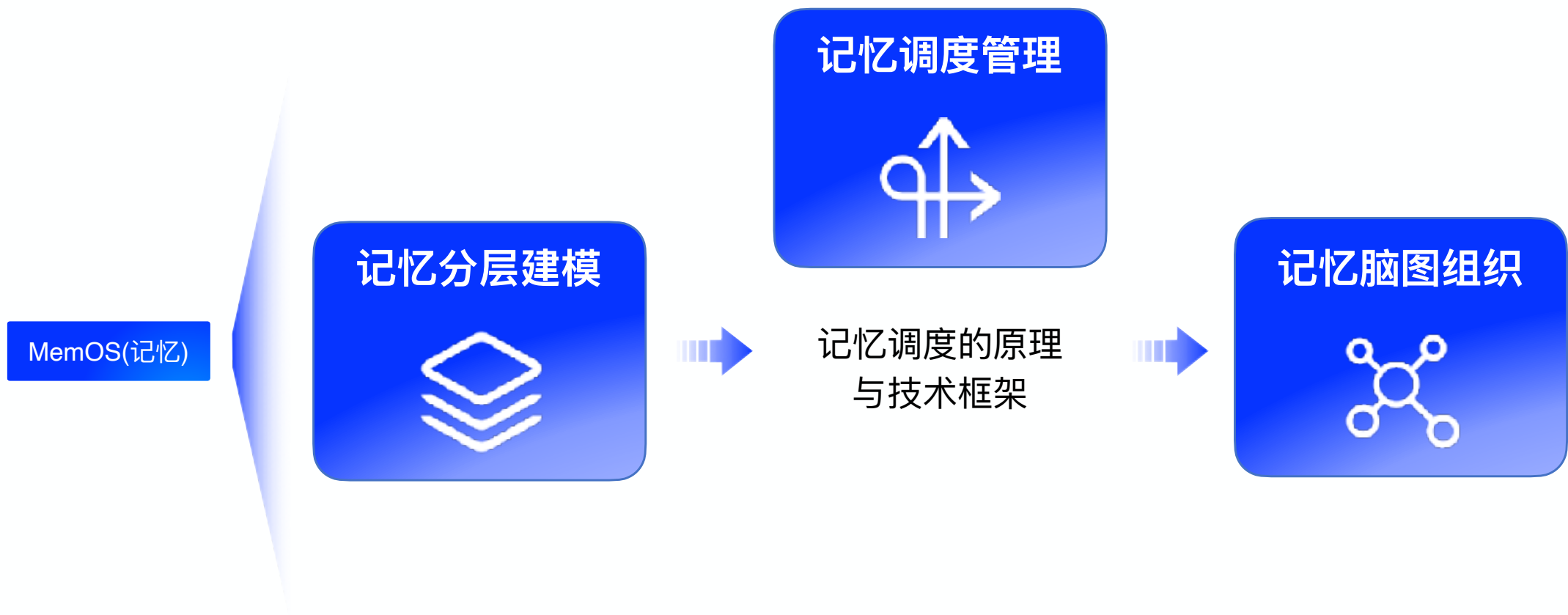


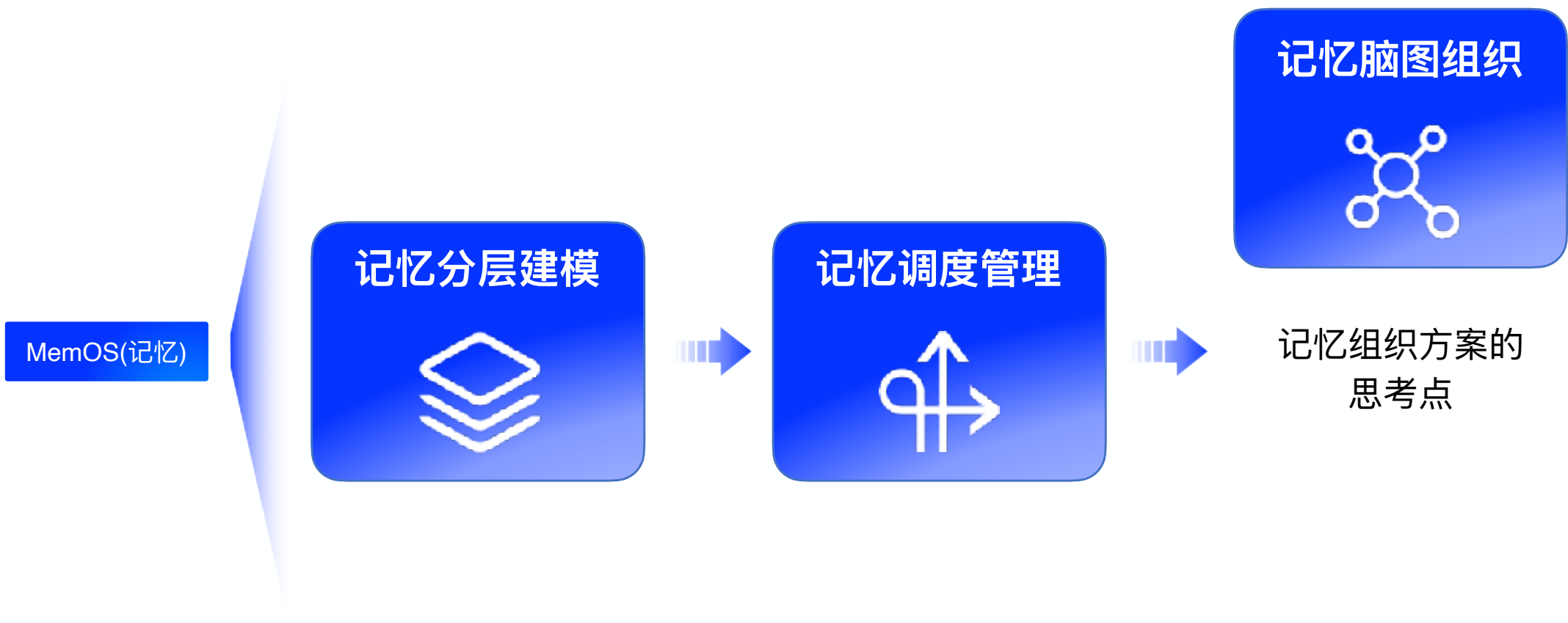
# MemOS的系统框架





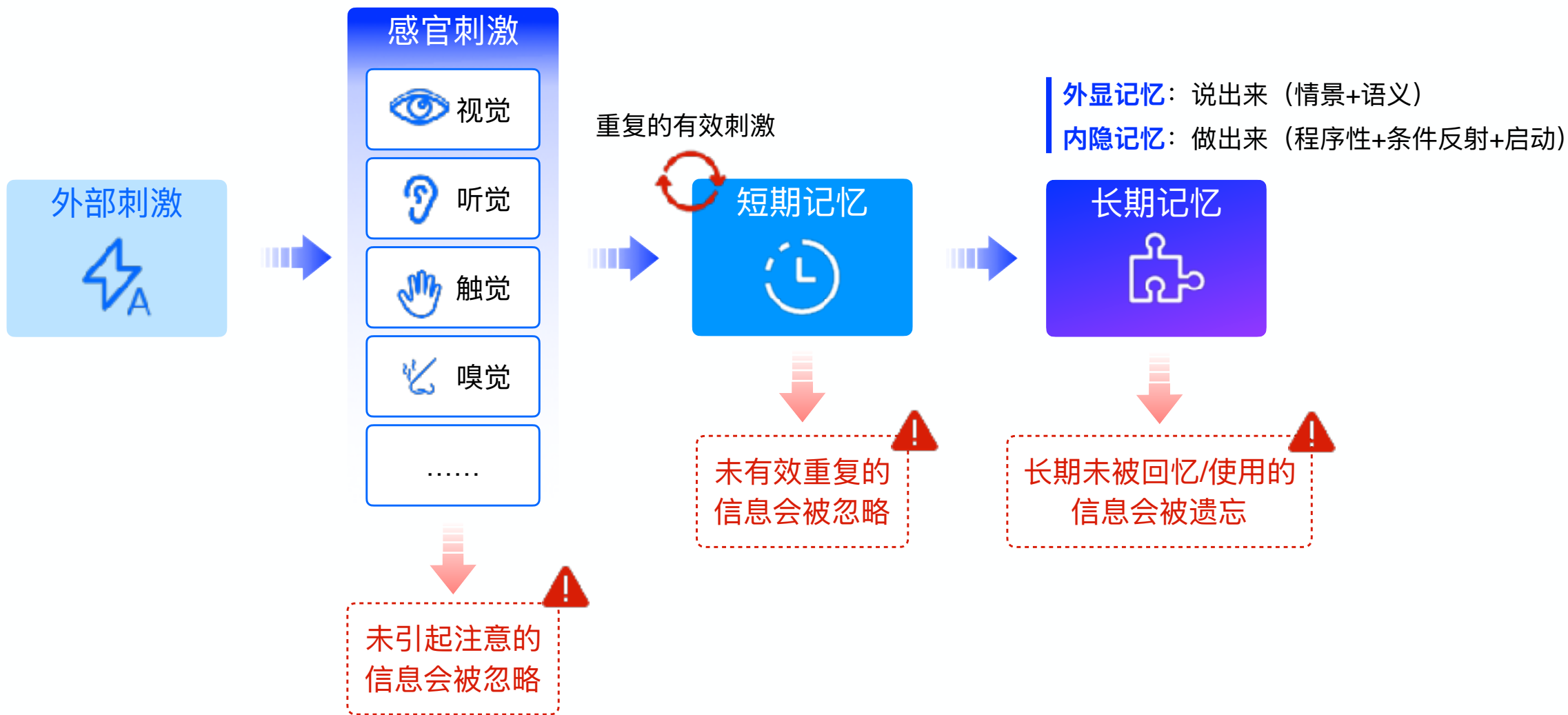






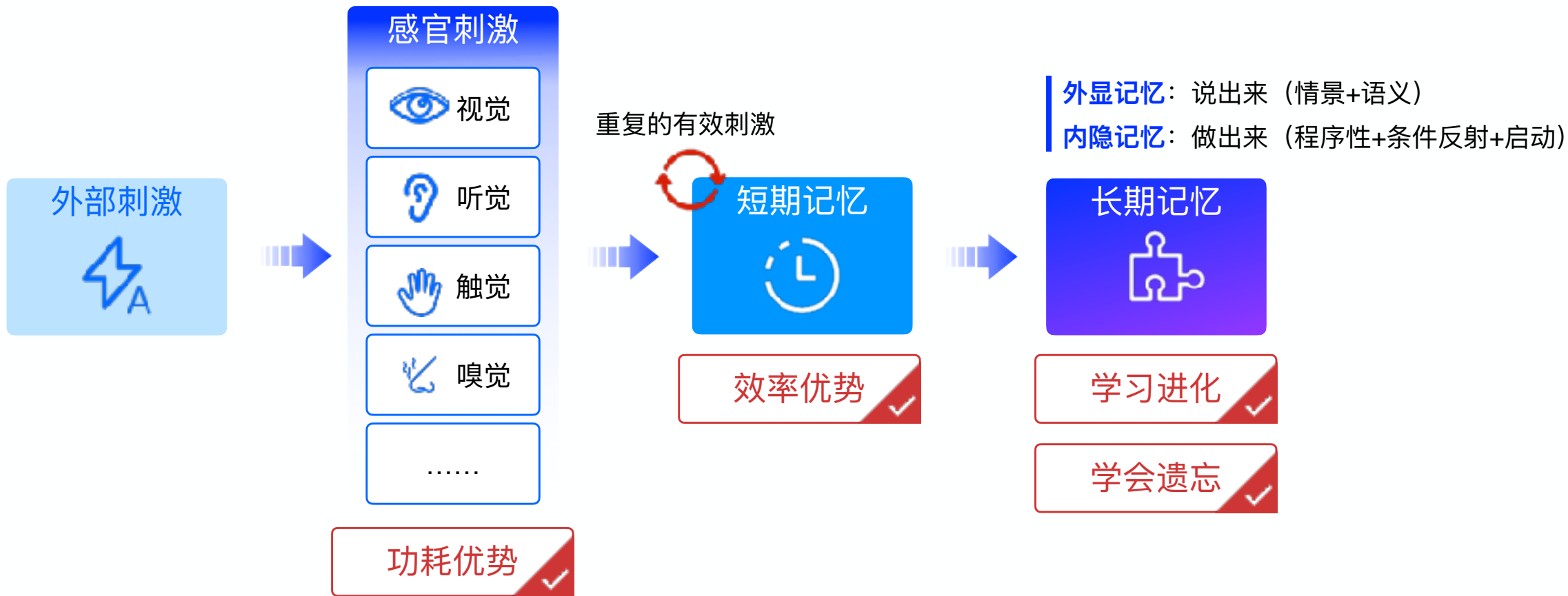


# MemOS的核心机制—记忆分层建模





# MemOS的核心机制—记忆分层建模



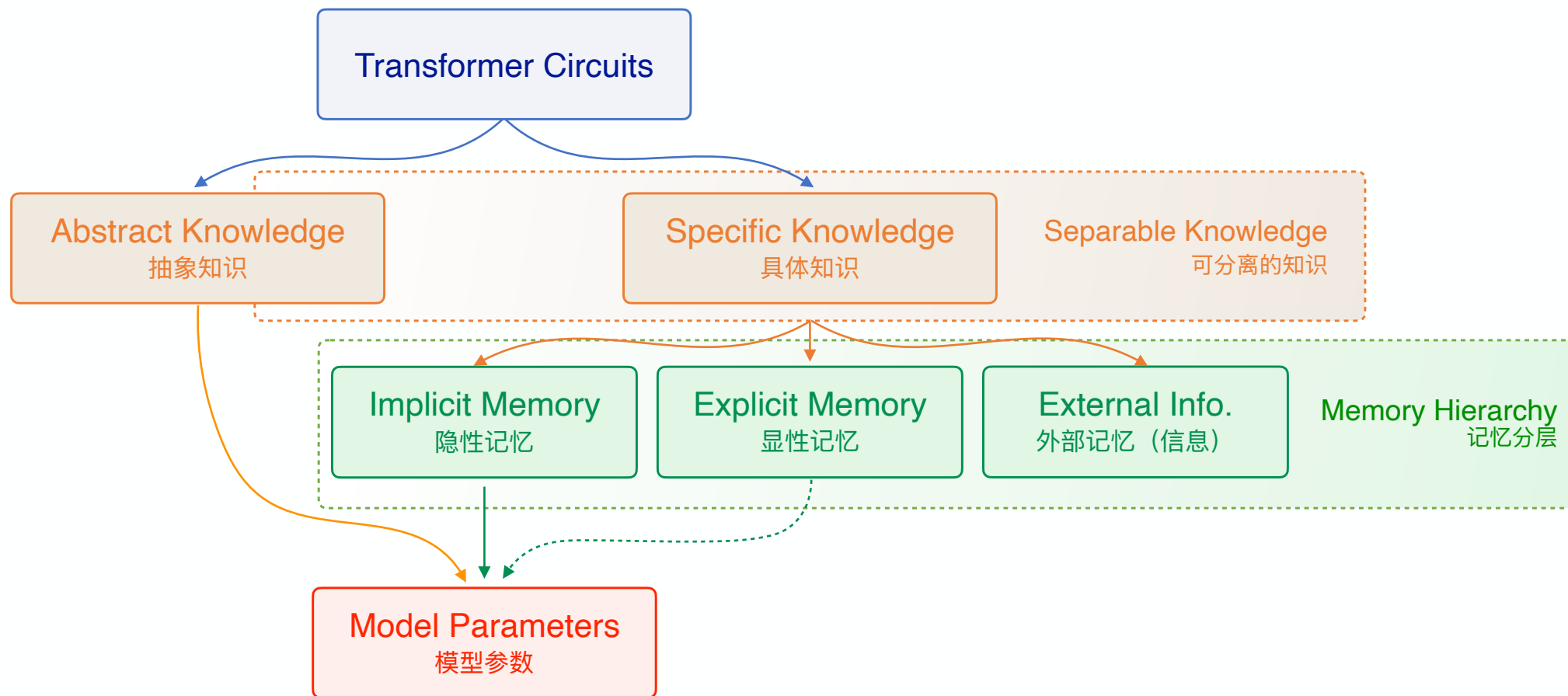
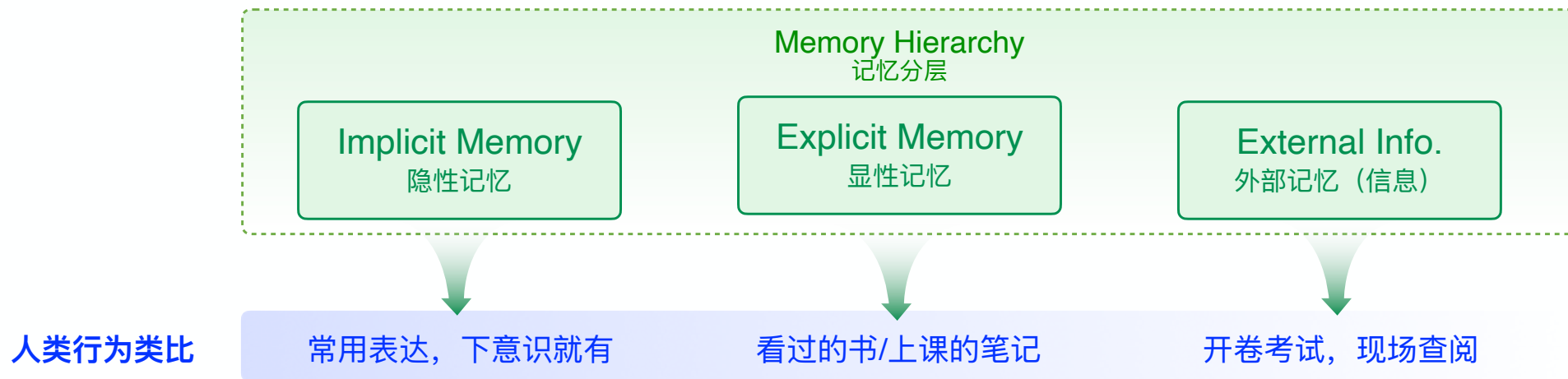
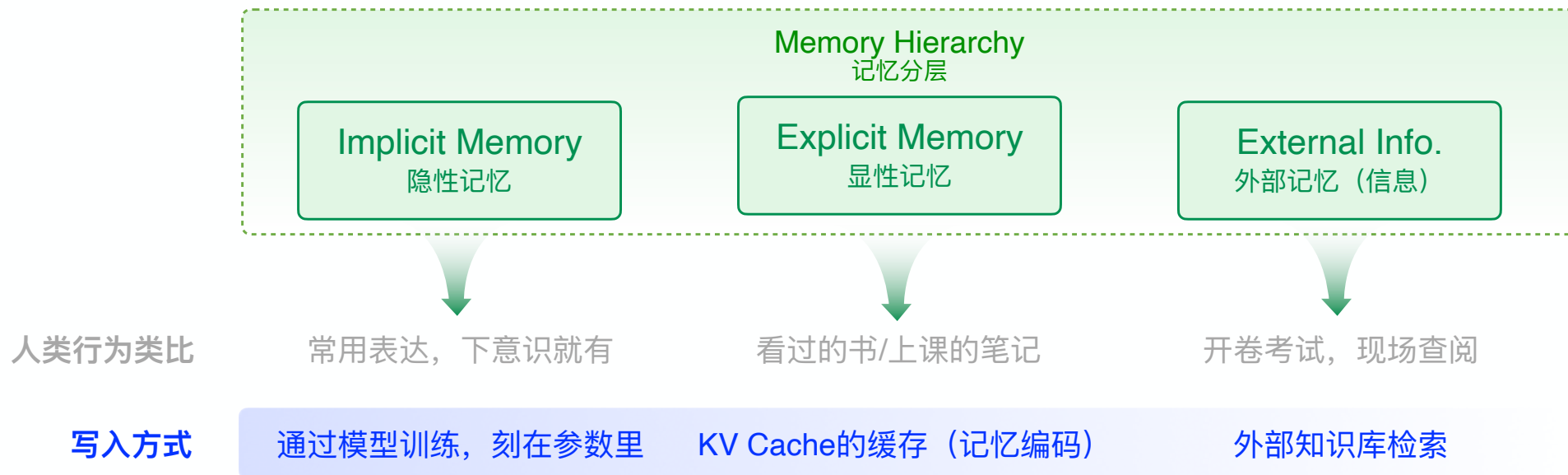
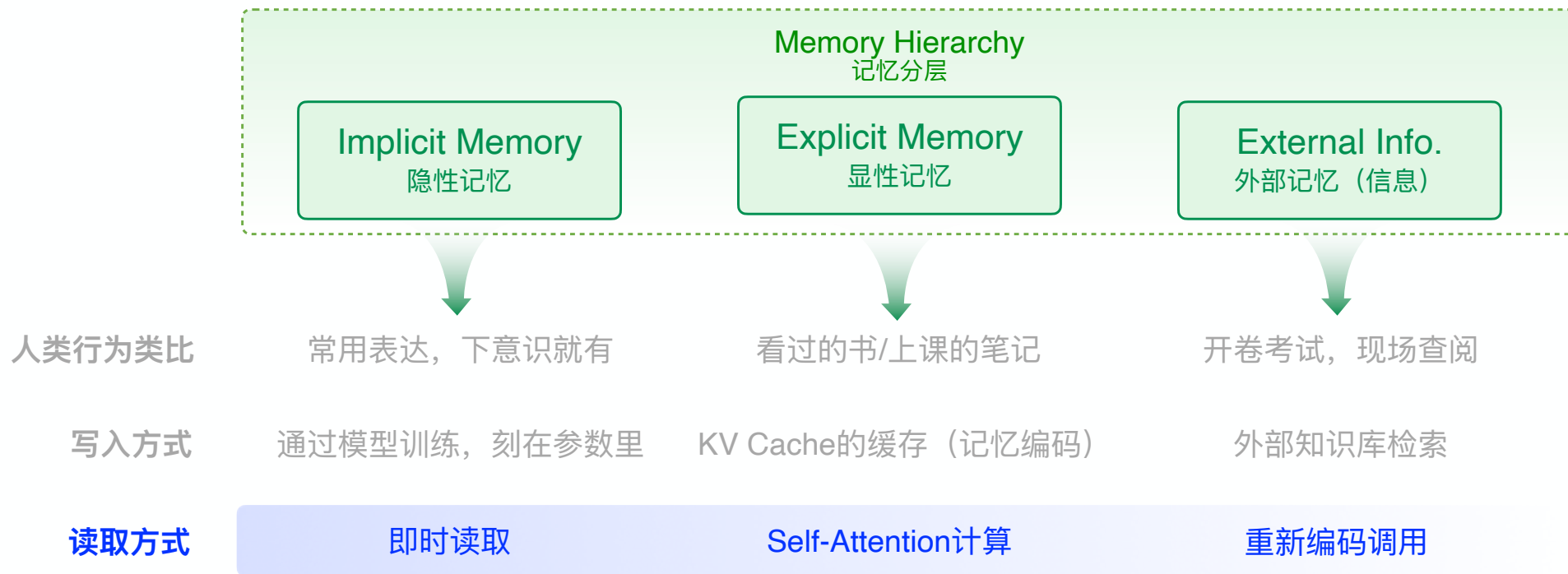
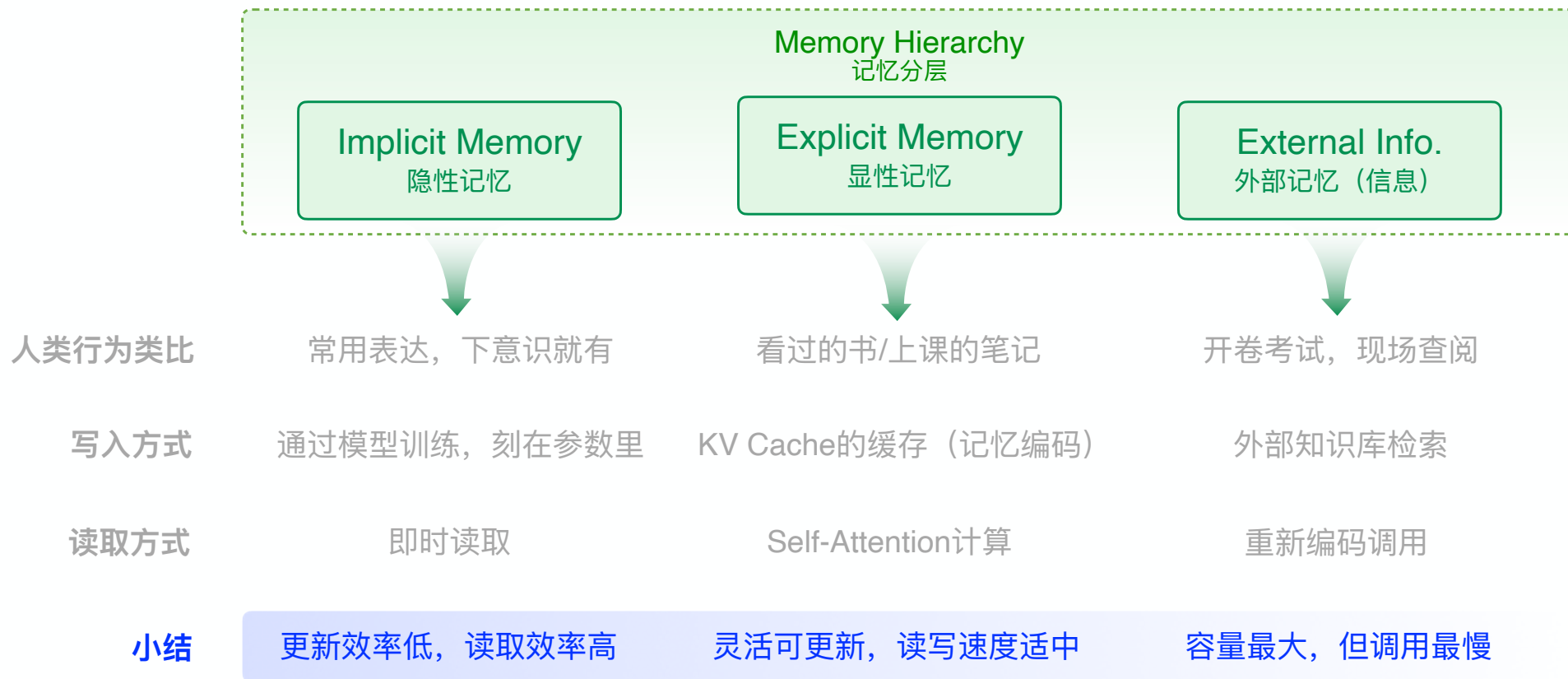


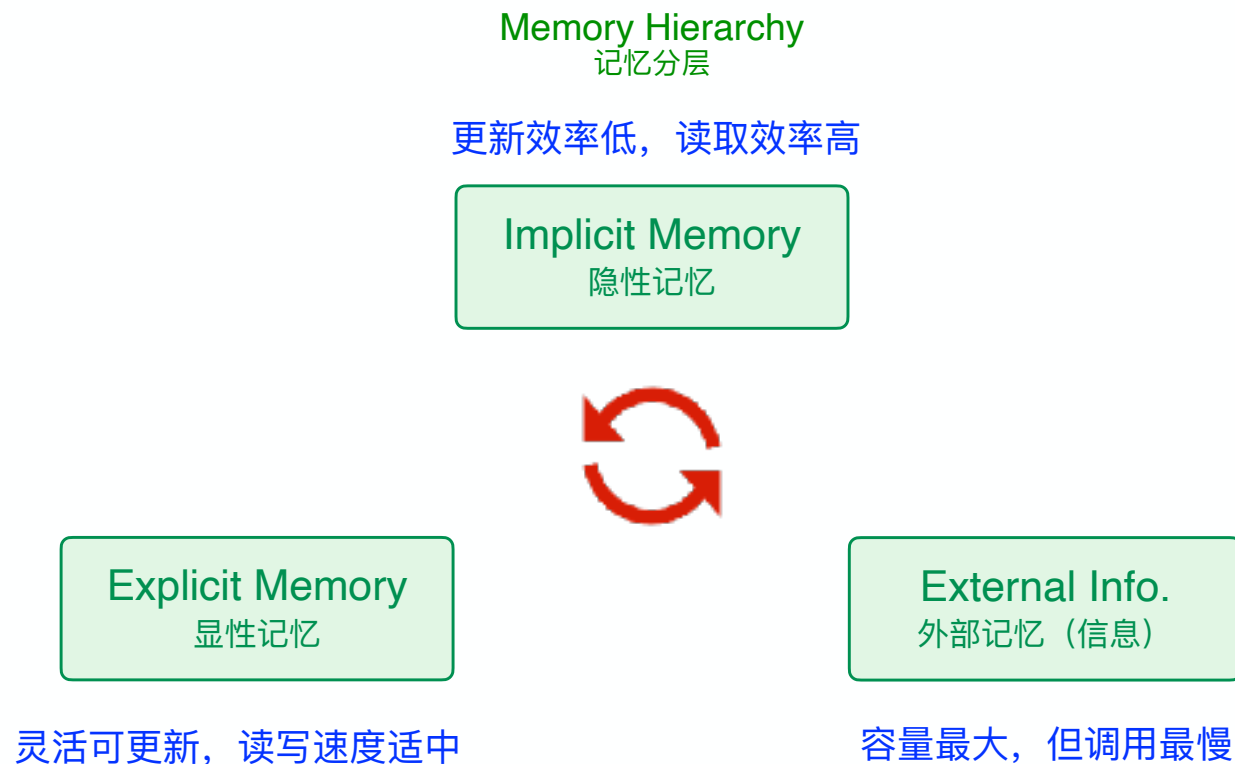
图 1: 大模型记忆分层理论 (源自Memory3论文) [1]

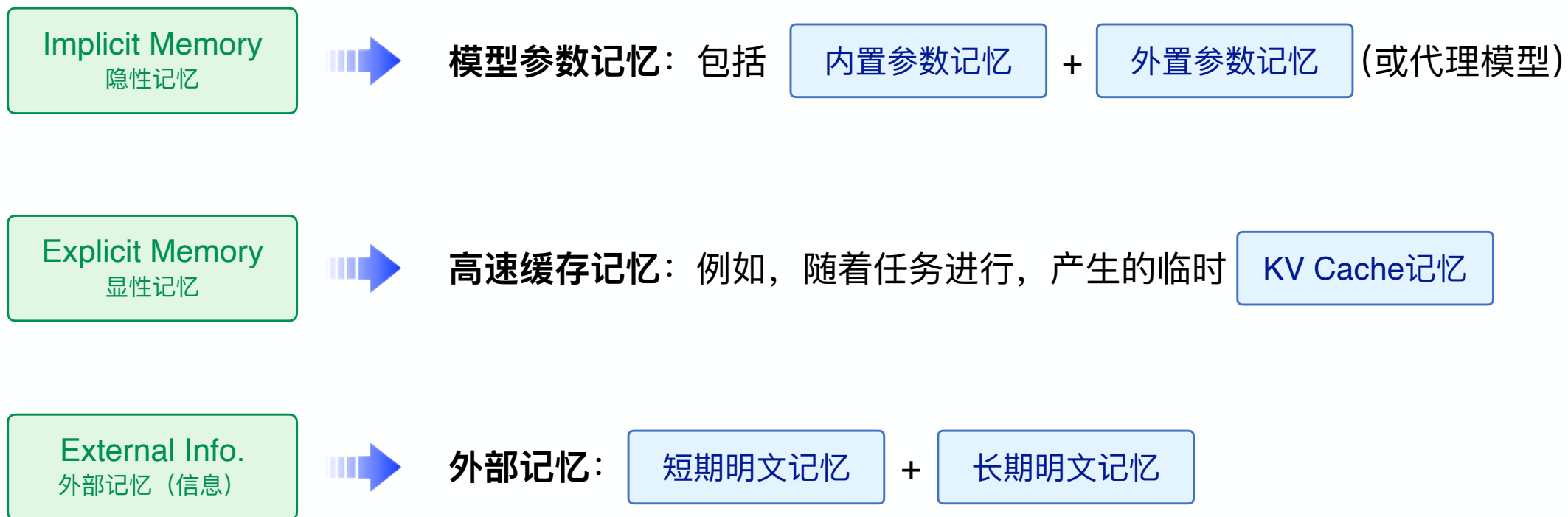






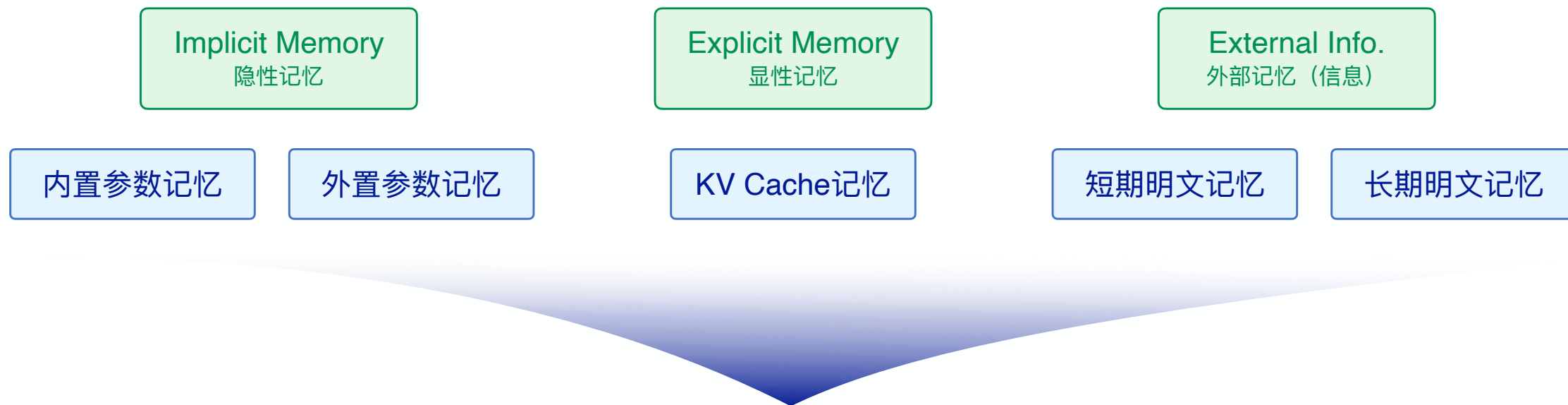








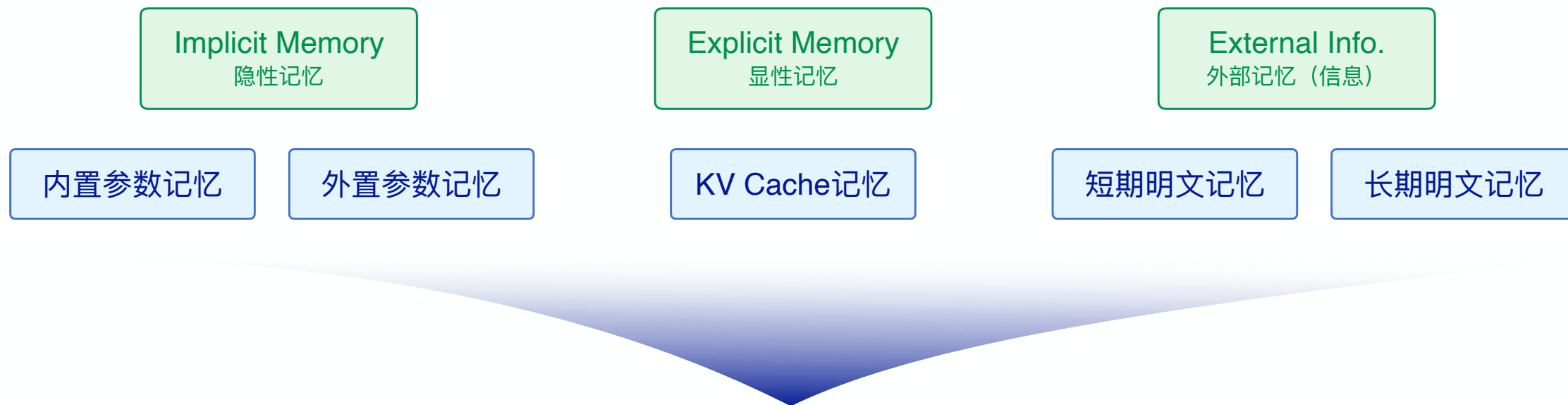
构建  
调度  
使用



对话压缩/向量存储 ◀ ★ 构建

上下文堆积 ◀ ★ 调度

事实校验、主体校验、权限校验 ◀ ★ ★ ★ ★ ★ 使用



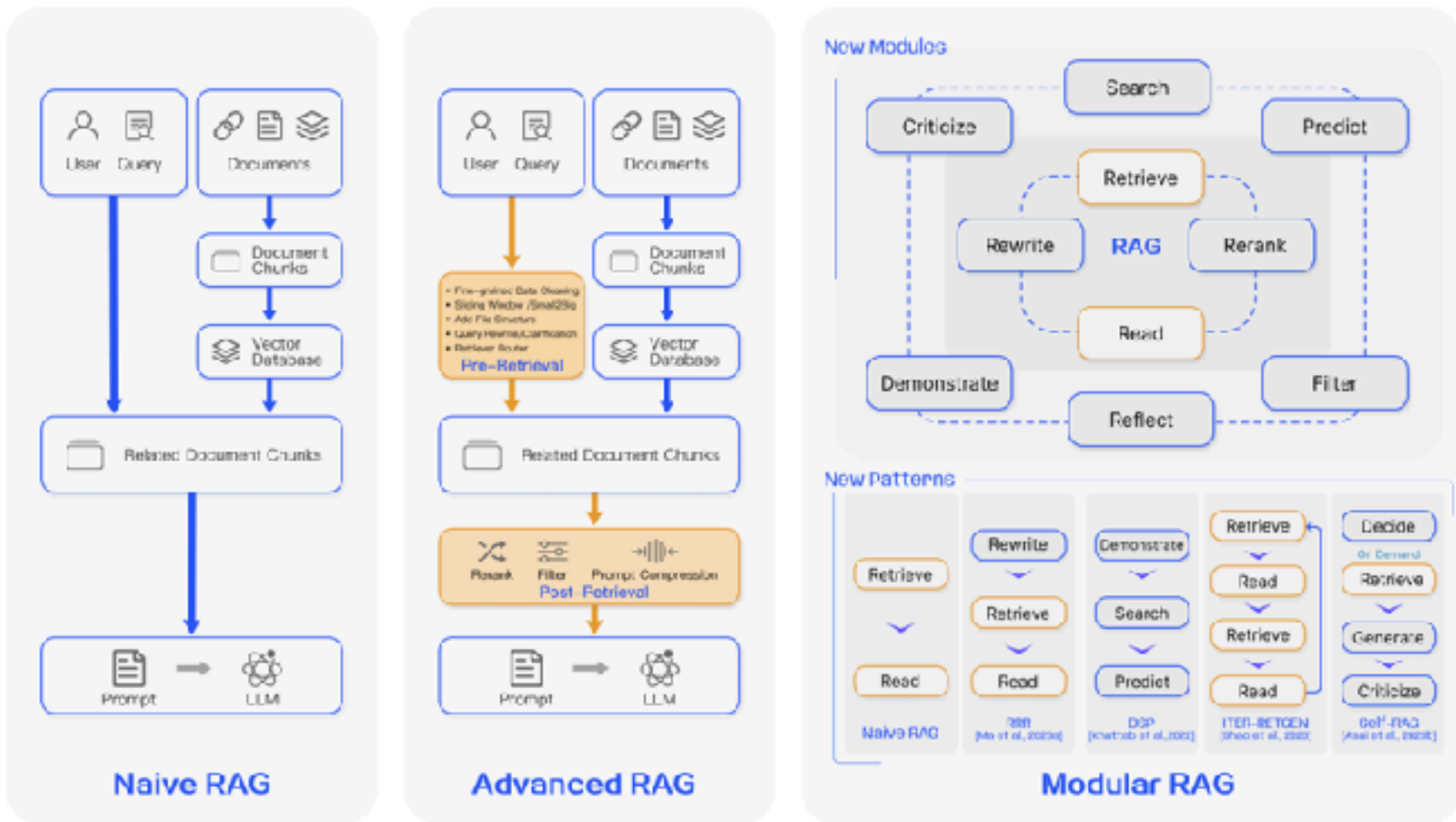
对话压缩/向量存储 ◀ ★ **构建** ★ ★ ★ ★ ▶ 脑图记忆组织/抽取、图+向量化存储

上下文堆积 ◀ ★ **调度** ★ ★ ★ ▶ 主动预测，将记忆放在最合适的位置

事实校验、主体校验、权限校验 ◀ ★ ★ ★ ★ ★ ★ **使用** ★ ▶ 场景自动识别，记忆编排框架



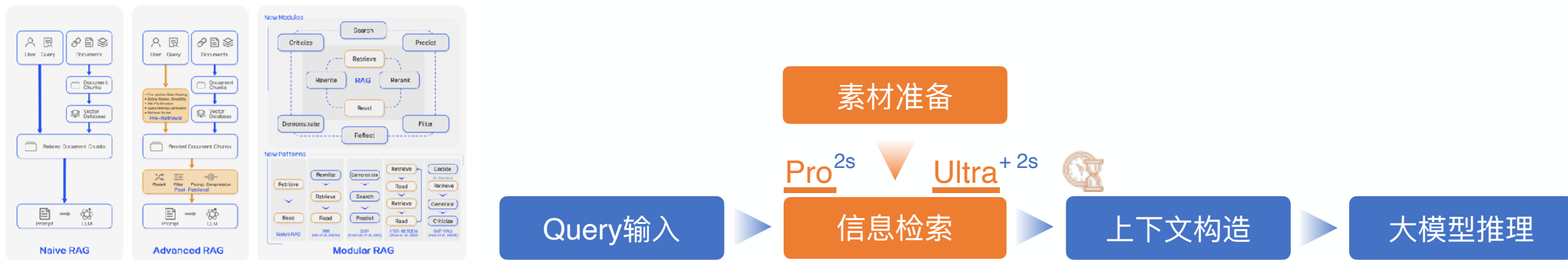
## MemOS的核心机制二：记忆调度管理 ■ 将记忆放在最合适的位置



\* RAG检索增强的典型范式（被动式检索）<sup>[1]</sup>

# 记忆调度建模：从被动式检索到主动式生成

\* RAG检索增强的典型范式（被动式检索）<sup>[1]</sup>：在用户提出完整请求时被动触发检索模块，是一种典型的阻断检索

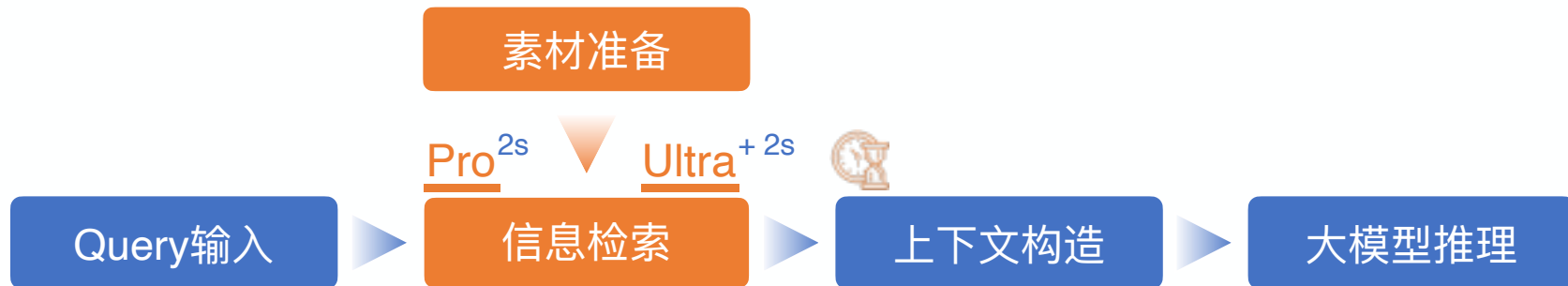


\* RAG检索增强的典型范式（被动式检索）<sup>[1]</sup>：在用户提出完整请求时被动触发检索模块，是一种典型的阻断检索



1 高延迟，复杂模型无法开展

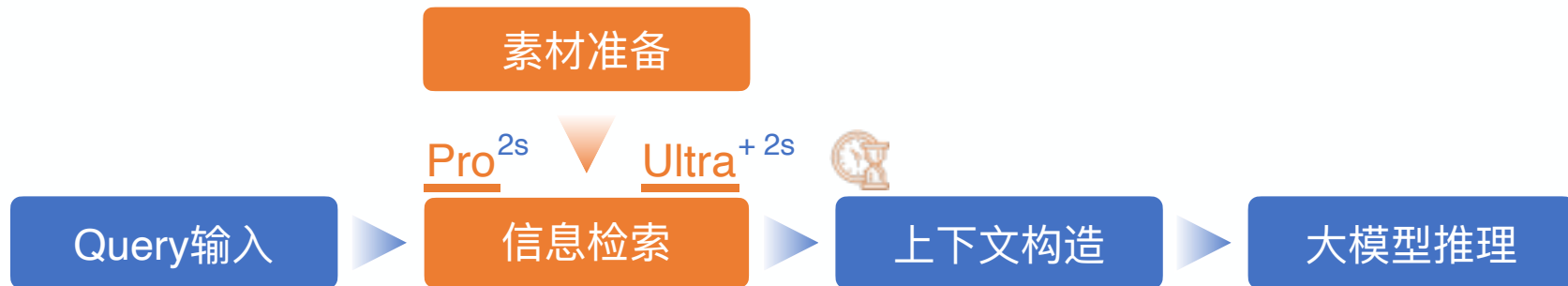
\* RAG检索增强的典型范式（被动式检索）<sup>[1]</sup>：在用户提出完整请求时被动触发检索模块，是一种典型的阻断检索



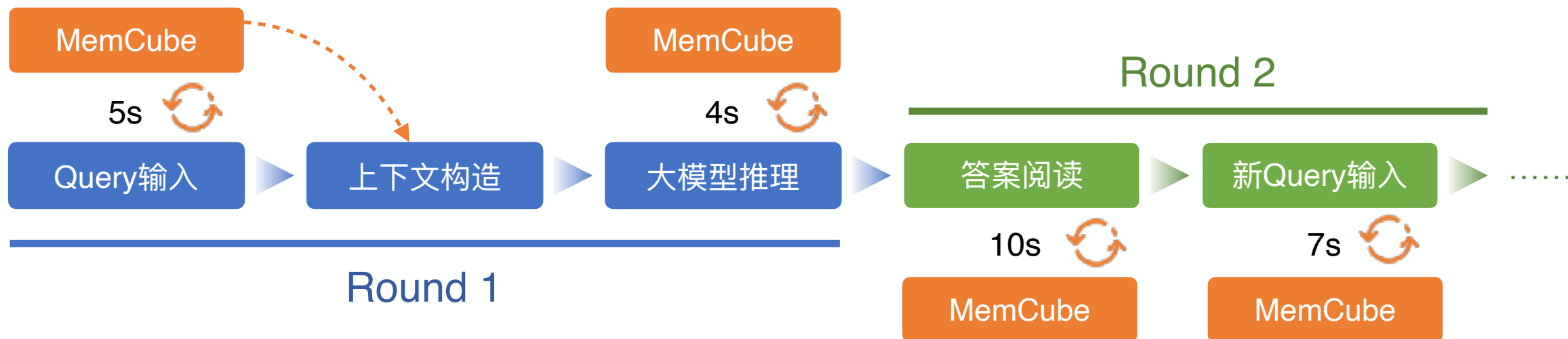
- 1 高延迟，复杂模型无法开展
- 2 碎片化，缺乏跨块整合能力

# 记忆调度建模：从被动式检索到主动式生成

\* RAG检索增强的典型范式（被动式检索）<sup>[1]</sup>：在用户提出完整请求时被动触发检索模块，是一种典型的阻断检索



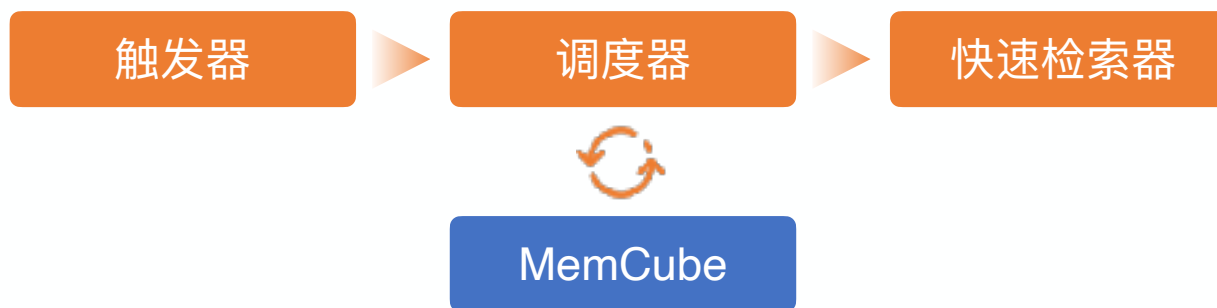
- 1 高延迟，复杂模型无法开展
- 2 碎片化，缺乏跨块整合能力
- 3 高成本，每次都需重新检索



在应用场景需要进行记忆管理的位置进行埋点，满足条件后出发调度机制，包含各类触发模版和 Action List.

根据触发器发送的调度信息，进行记忆的复杂预测、检索和上下文准备，提高下次检索的缓存命中率。

使用更加简单、快速的传统检索方案进行额外的、少量的信息补充，以保障时间效率为前提。（可选步骤）





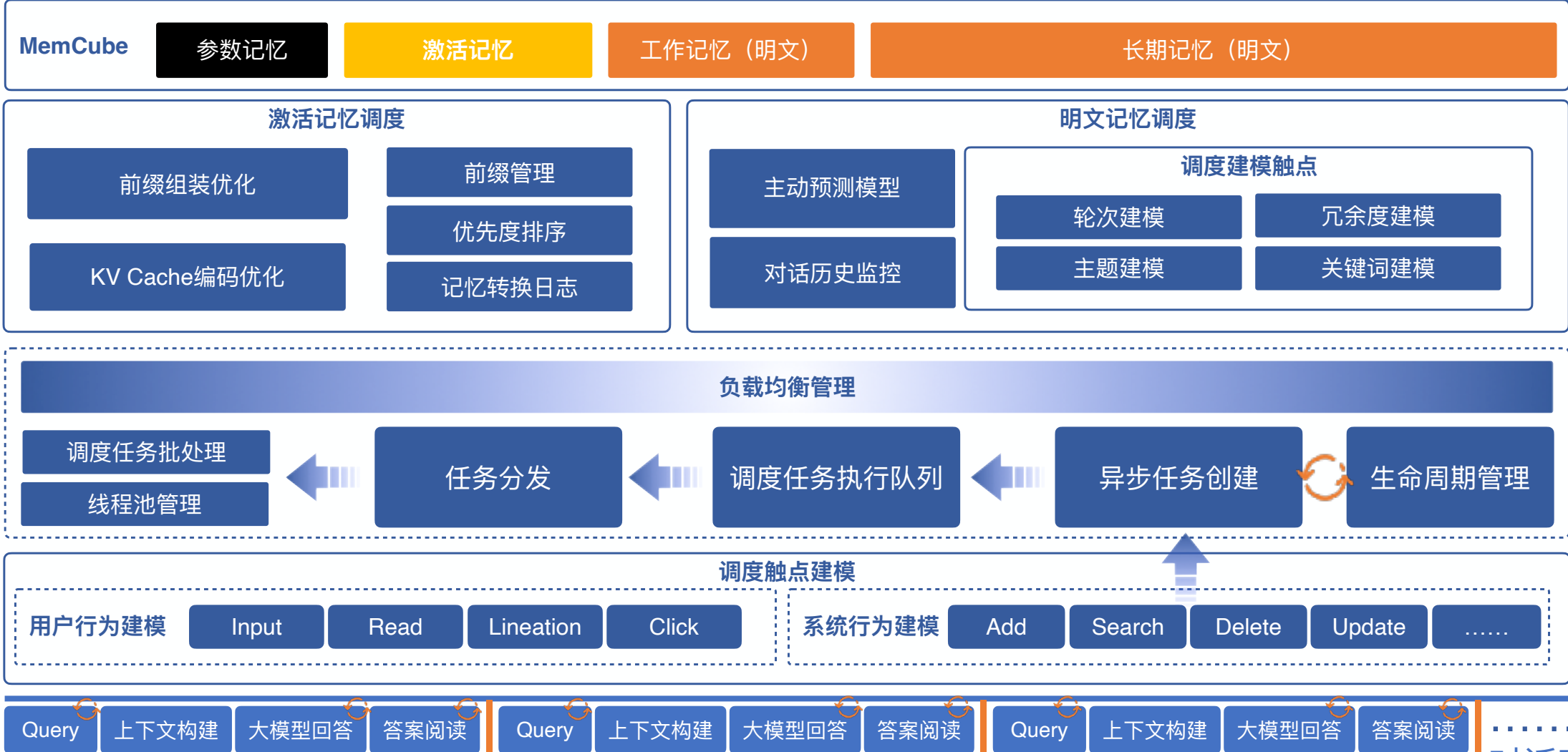
# MemOS记忆调度框架

记忆体

调度执行层

调度管理层

调度触发层



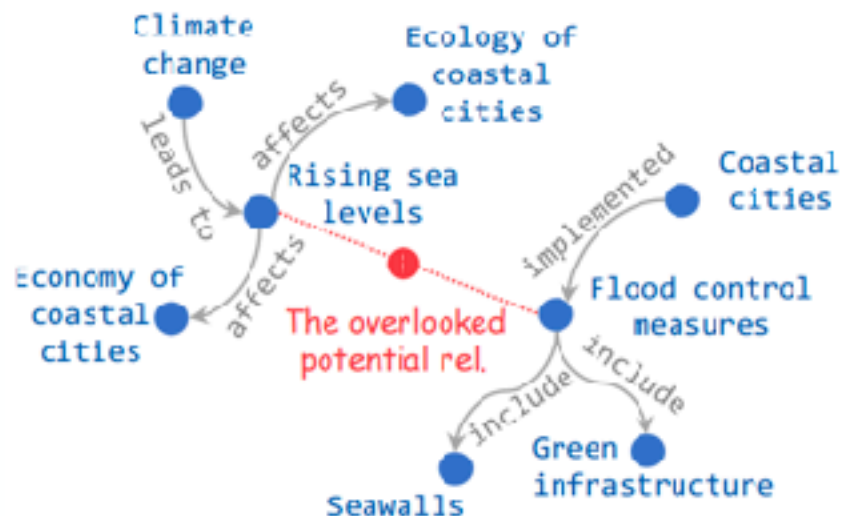
## MemOS的核心机制三：记忆脑图组织与检索

Climate change leads to rising sea levels, affecting the ecology and economy of coastal cities.

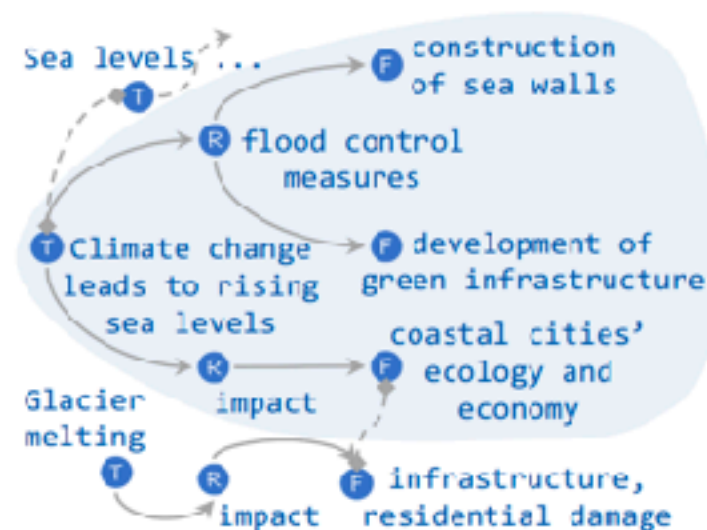
Similarity Rel.

Coastal cities have implemented flood control measures, including the construction of seawalls and green infrastructure, to address the ongoing rise in sea levels.

Chunk-based



KnowledgeGraph-based



Xmind-based

灵活高效，窗口内信息丢失少

低压缩比

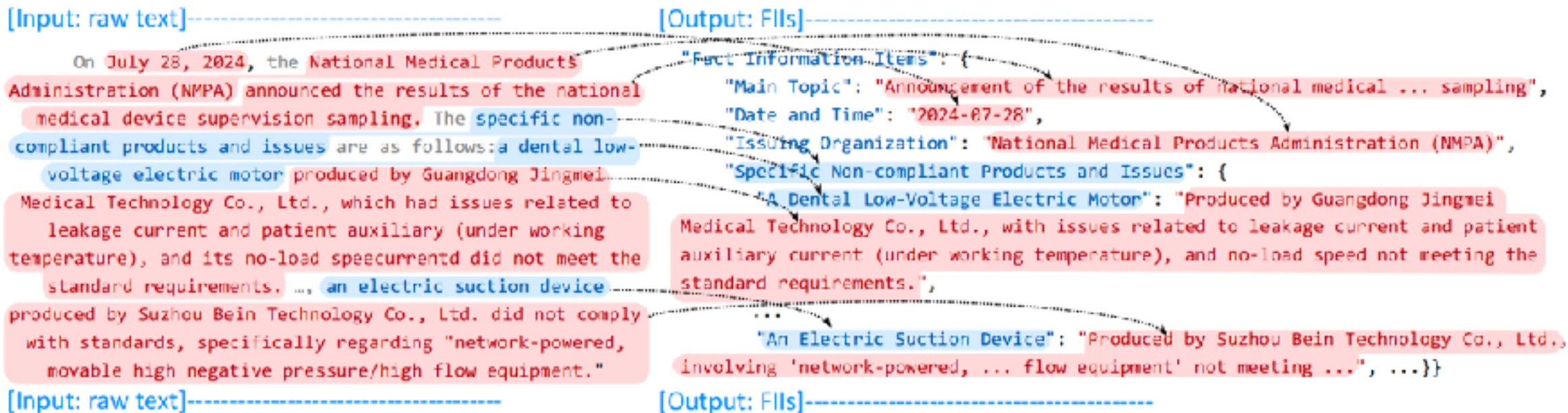
可推理，容易校验标注

高压缩比

易关联，灵活度高，主动记忆

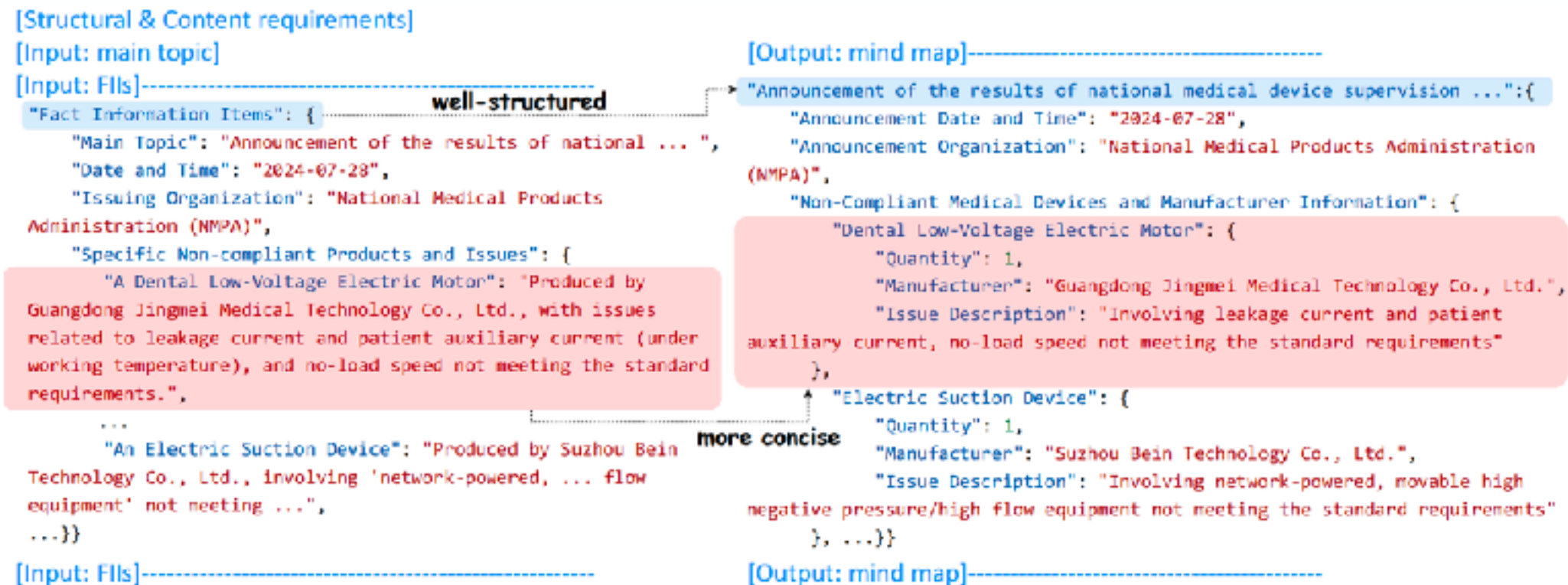
适中压缩比

**主动记忆**：是指大模型在处理输入时，不再仅仅依赖用户检索触发的被动式切片，而是能够 **主动分析对话或文档内容的语义结构**，并基于任务目标，对其中需要长期保留的信息进行拆分、筛选、归纳与组织，从而形成高效的思维导图。



关键步骤一：抽取逻辑分析（形成记忆COT过程）

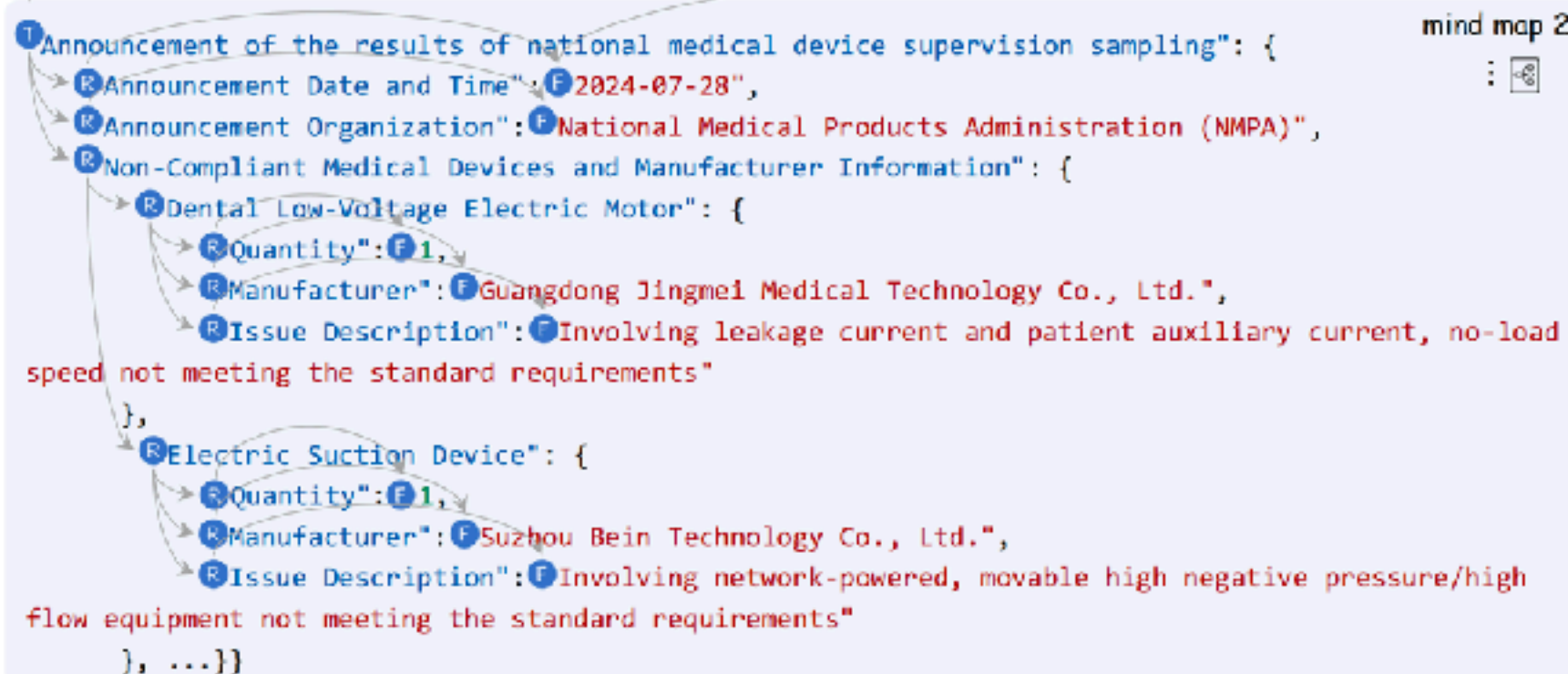
**主动记忆：**是指大模型在处理输入时，不再仅仅依赖用户检索触发的被动式切片，而是能够 **主动分析对话或文档内容的语义结构**，并基于任务目标，对其中需要长期保留的信息进行拆分、筛选、归纳与组织，从而形成高效的思维导图。



## 关键步骤二：二次校验与关联性边重构

## MemOS的核心机制三：记忆脑图组织与检索

topic node: T route node: R fact node: F fact path: T → R\* → F mind map: T → R → F super node: S



### 丰富的检索特性:

- 主题-路由-事实 路径
- 关键词检索
- 跨Session推理
- 时序节点（版本管理）



# MemOS的整体性能表现



## 效果对比 | 基于LoCoMo数据集的实验性能对比

### MemOS评估结果



#### 任务准确率提升

在LoCoMo数据集上, MemOS在四类核心任务中的平均准确率较OpenAI 的全局记忆方案提升

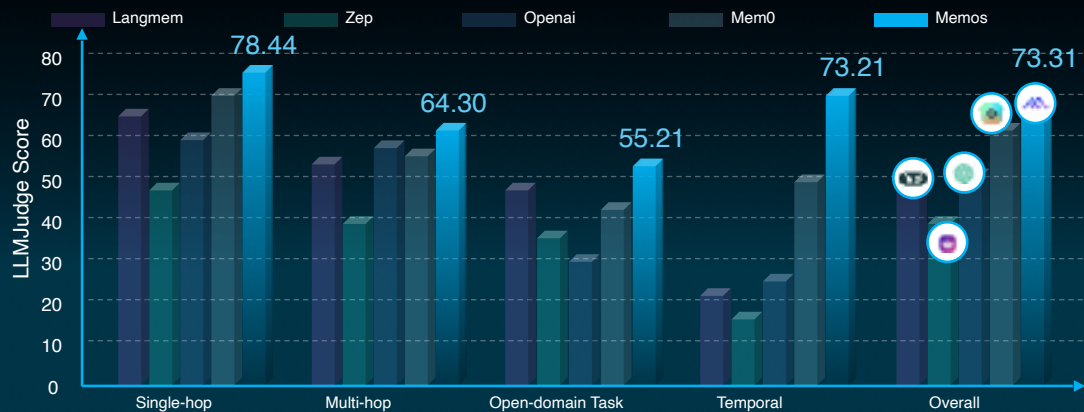
39%



#### 成本优化

相比OpenAI, MemOS在Token消耗上显著降低61%, 能够用更少的检索Token达到同样任务效果

果



## 效果对比 | 基于LongMemEval数据集的实验性能对比

### MemOS评估结果



#### 任务准确率提升

在LongMemEval数据集上, MemOS相比业内的记忆模型如Mem0、Zep等均具有明显性能

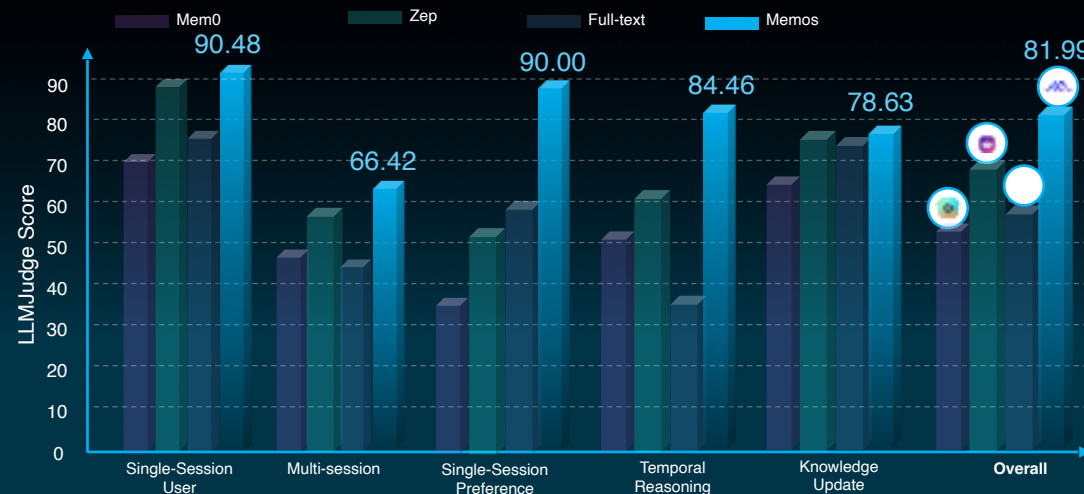
优势。



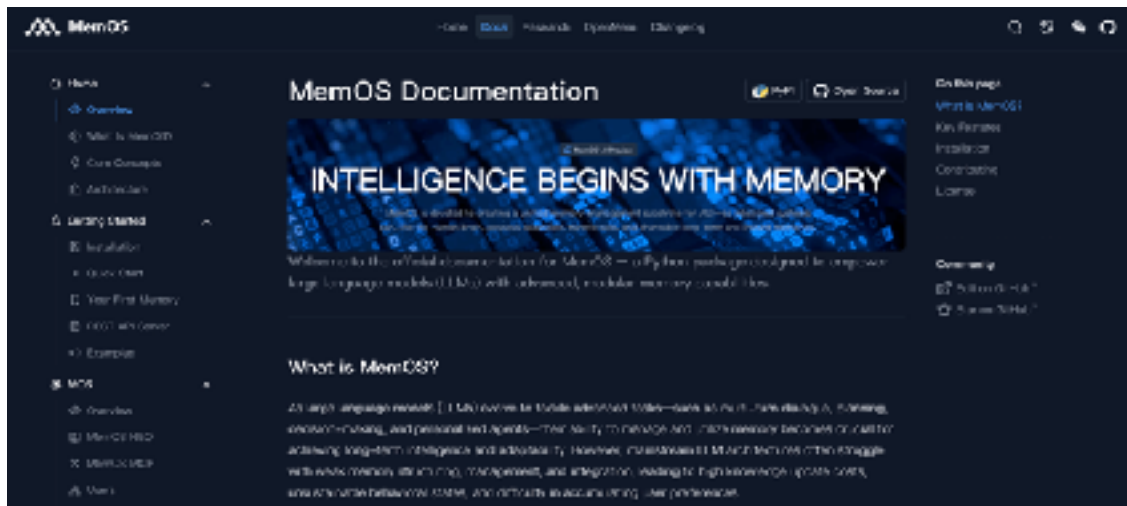
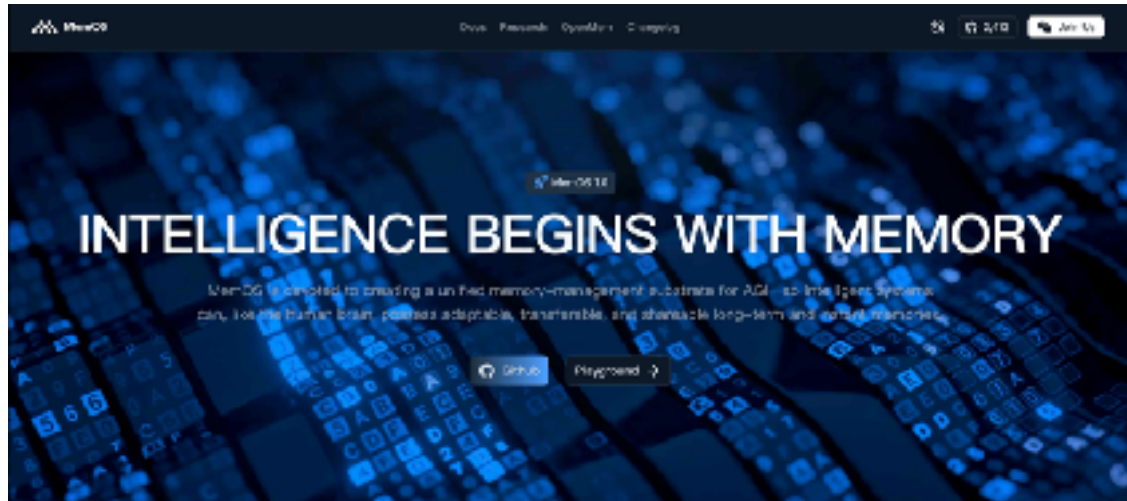
#### 成本优化

相比与记忆框架相比, MemOS通过精准的检索召回, 大幅降低解码所需填充的上下文内容40%

以上。



# MemOS 的开源框架与OpenMem社区



项目地址: <https://memos.openmem.net/>

仓库地址: <https://github.com/MemTensor/MemOS>

开源许可: Apache 2.0 License

Star数: 2.5k

开发者社群: 超1600+人

生产可用: 已支持 Playground & API (0925上线)

**OpenMem社区:** 汇集来自上海交大、同济、北大、中科大、人大、北航、天津大学等众多高校的科研团队, 以及多个工业界机构。

欢迎开发者/研究者加入, 共建记忆生态!





## 智能投顾

痛点：Agent无法“记住”用户的风险偏好与历史对话，每次交互都从零开始，建议缺乏连续性

### MemOS 应用

- 跨会话持续保存客户画像、风险等级、持仓、历史建议等信息
- 智能体能基于记忆生成符合其投资风格与历史行为的建议
- 提供“长期陪伴式”服务，让投顾从单次回复转向持续理解

投资顾问服务半径扩大约20%



## 工业运维

痛点：大量“老师傅经验”分散在各类文档与记录中，未形成知识化沉淀，Agent难以继承，诊断重复、效率受限。

### MemOS 应用

- 将“老师傅”的诊断经验、维修方案与案例结果转化为记忆
- Agent 可自动召回相似案例，复用过往解决路径，快速定位问题
- 新人也能即时调用资深经验库

平均诊断响应时间缩短约30%



## 酒店商户服务

痛点：某TOP在线旅游服务平台研发的酒店商户服务Agent，Bad Case 主要源于检索未命中而非知识缺失——人工修正内容无法快速更新到知识库；传统 ES 堆叠方案维护成本高

### MemOS 应用

- 以记忆机制存储人工反馈的正确答案

一教就会，Bad Case 不再复现



## 科研助手

痛点：某科研Agent聚焦“读/算/写”一体化科研流程，需要持续理解论文、实验与笔记间的上下文关系。

### MemOS 应用

- 将论文、实验参数、笔记等结构化存入记忆
- Agent 可跨会话召回历史研究内容，实现连续理解
- 支撑多轮科研对话与工具调用，提升任务连贯性

已有2w+用户，日均调用超 4000 次



# One More Thing

# Text2Memory <sup>NEW</sup> 1.0

面向记忆工程自动化编排的操作语言

An Operational Language for Automated Memory Engineering



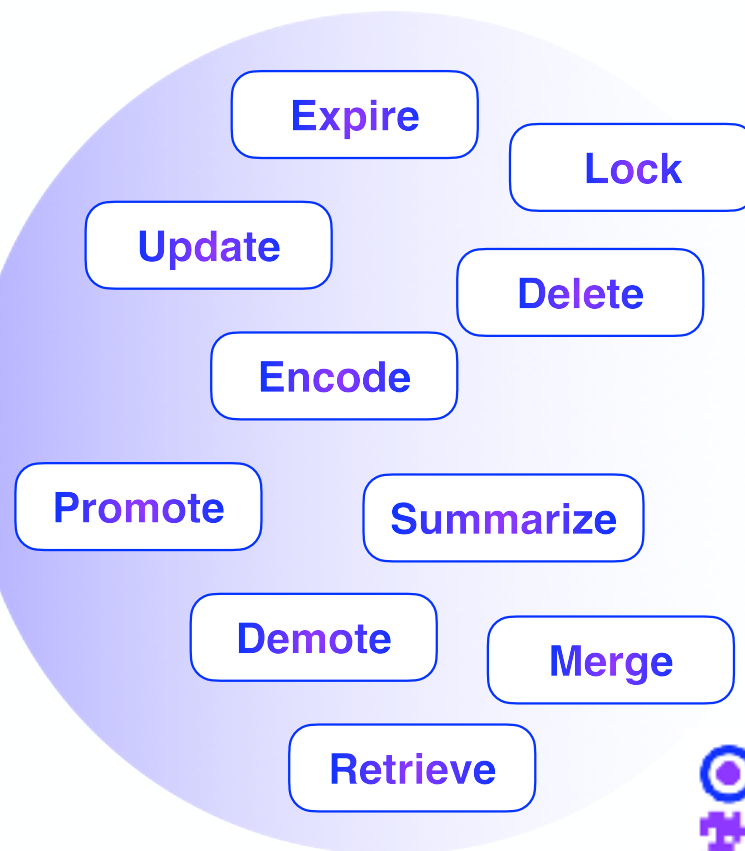


One More Thing:

# Text2Memory <sup>NEW</sup> 1.0

面向记忆工程自动化编排的操作语言

An Operational Language for Automated Memory Engineering



📍 北京

**QCon**

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AI Ops
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

**AICon**

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

**QCon**

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

**AICon**

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

**AICon**

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

**AICon**

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LM Ops
- 具身智能

# 极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

# THANKS

智能始于记忆  
张量链接未来

INTELLIGENCE BEGINS WITH MEMORY

