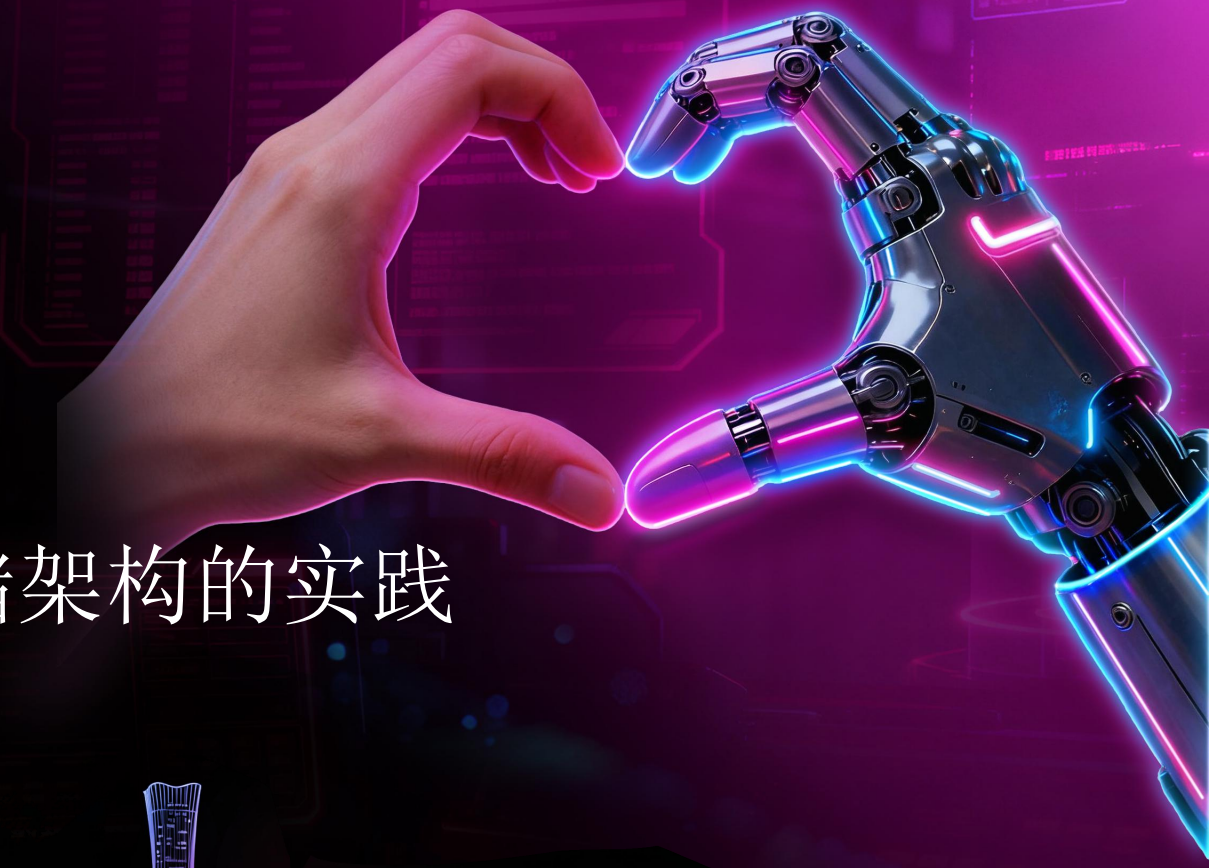
开源社 kaiyuanshe

AI Infra
CCF开源发展技术委员会

COSCon'25
第十届中国开源年会

众智开源 | Open Source, Open Intelligence



构建智能AI记忆系统：从认知科学到数据库混合存储架构的实践

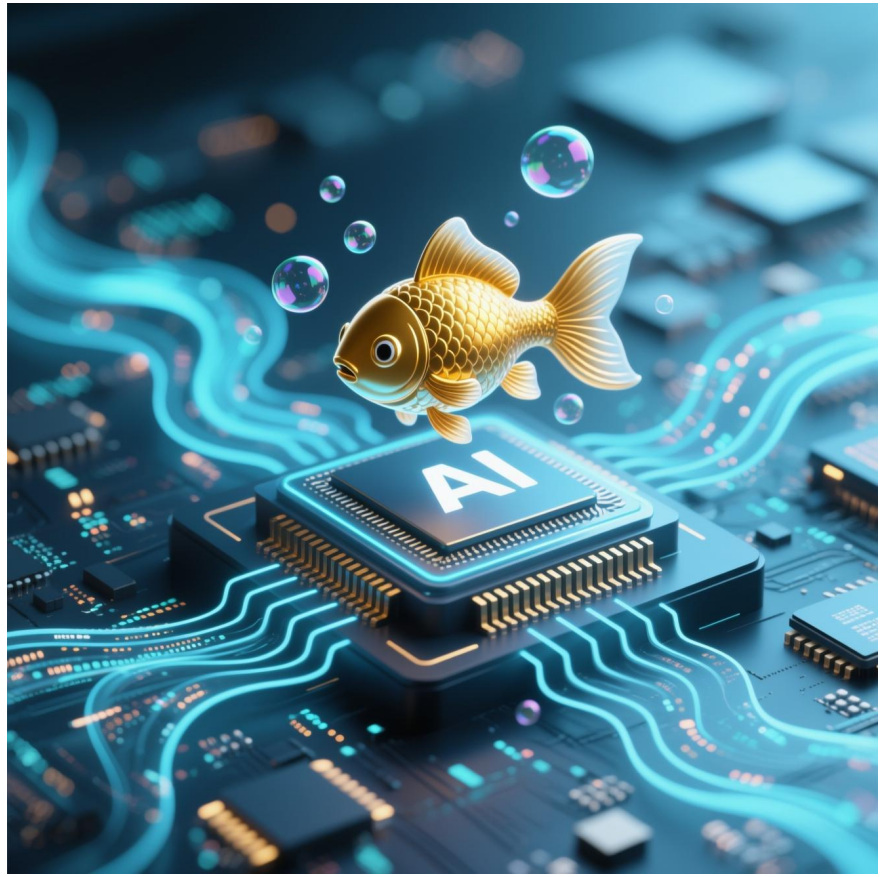
汤庆(OceanBase 技术专家)

CONTENTS

目录

PART 01 AI 的“健忘症”

AI的“金鱼记忆”

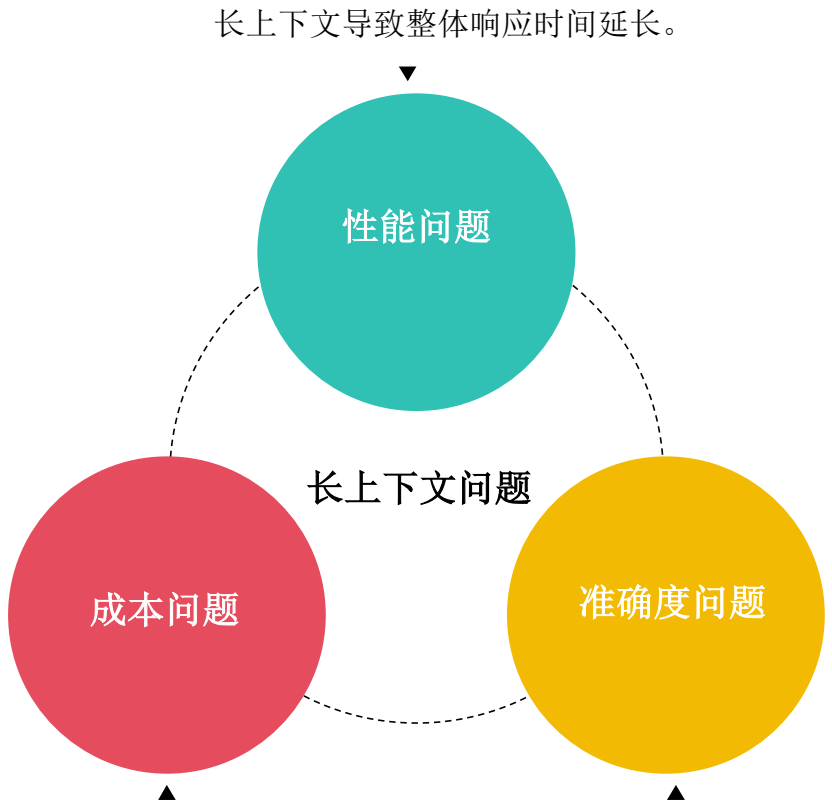


“金鱼记忆”

①上下文割裂：无法维持连续认知
②Multi_agent/Multi_user上下文无法复用”

怎么办？？？扩展上下文？？？

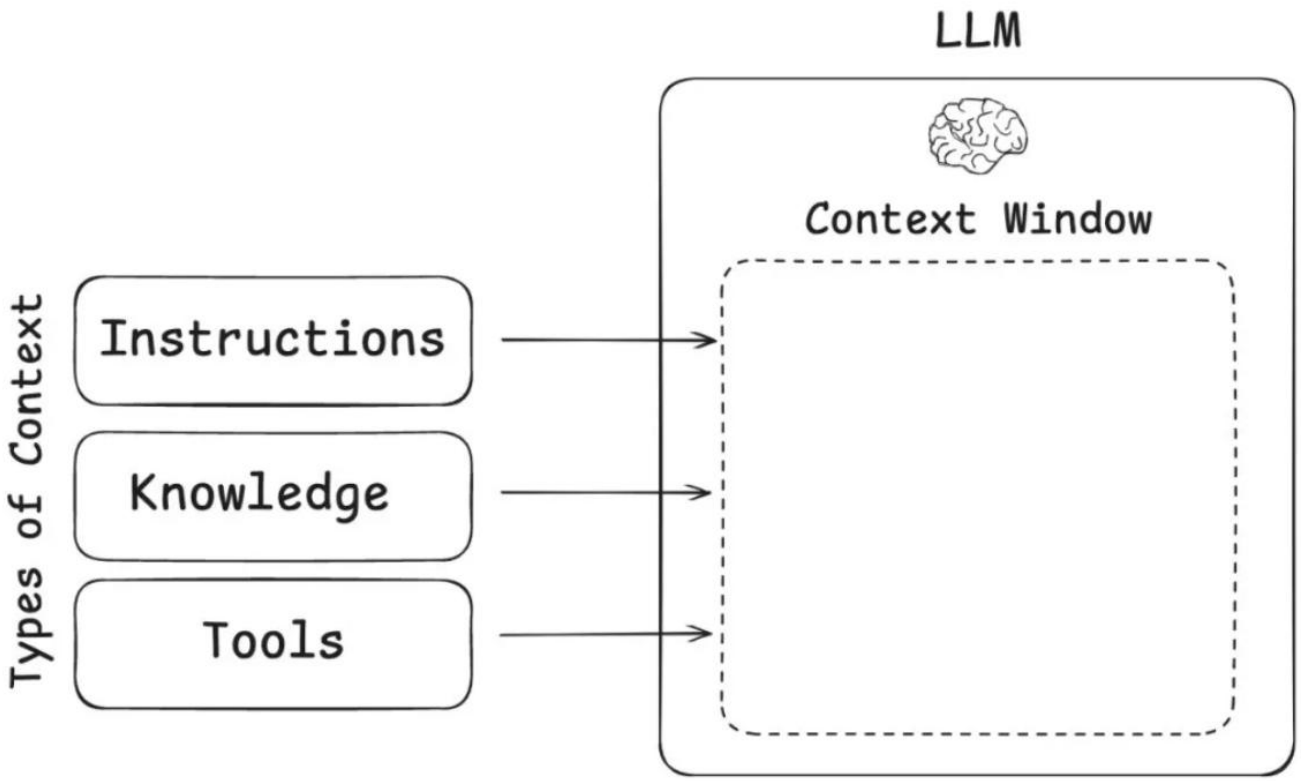
长上下文带来的问题



随着对话历史增长，模型推理速度急剧下降。同时Token用量呈线性增长，导致成本大幅上升。

长上下文导致关键信息因注意力资源有限而被弱化，重要指令易被淹没。

Context Engineering(上下文工程)



上下文工程的目标就在于在有限的窗口内找到对于完成任务有用的上下文信息

PART 02 powermem + seekdb 解决方案

PowerMem 是什么？

Your AI-Powered Long-Term Memory — Accurate, Agile, Affordable.

Accurate

+ 48.77%

More accurate vs. OpenAI Memory
78.70 % VS 52.9 %

Agile

91.83%

Lower selective retrieval p95 latency vs. full-context
1.44 s vs. 17.12 s

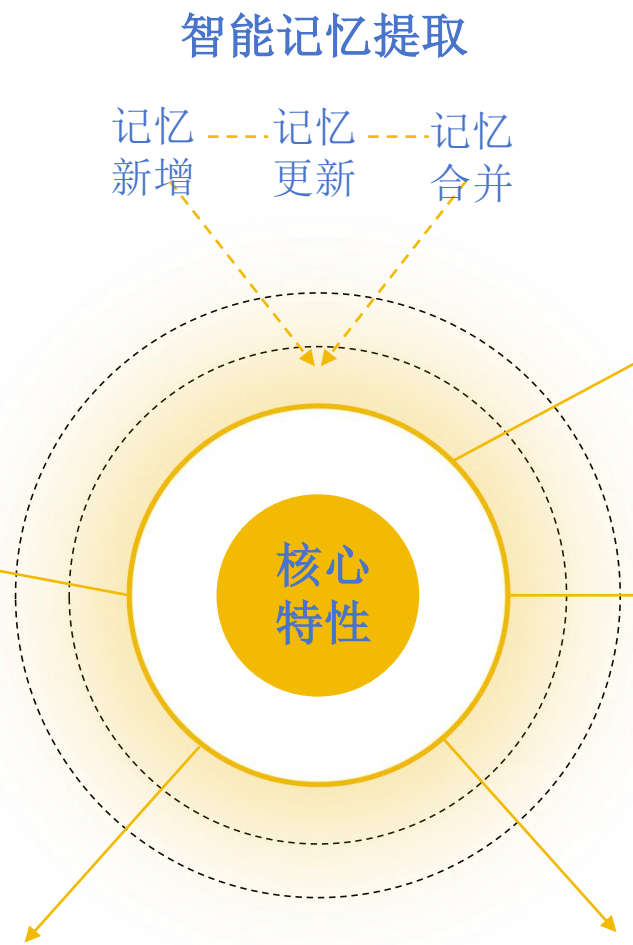
Affordable

96.53%

Token cost savings vs. full-context
0.9k VS 26k

基于 Apache2.0 开源: http://github.com/oceanbase/powermem

Powermem 核心特性



艾宾浩斯遗忘曲线

基于认知科学的记忆遗忘规律，自动计算记忆保留率并实现时间衰减加权，优先返回最近且相关的记忆，让 AI 系统像人类一样自然"遗忘"过时信息

开发者友好

python SDK/MCP 方式接入

Multi-Agent 支持

为每个智能体提供独立的记忆空间，支持跨智能体记忆共享和协作，通过作用域控制实现灵活的权限管理

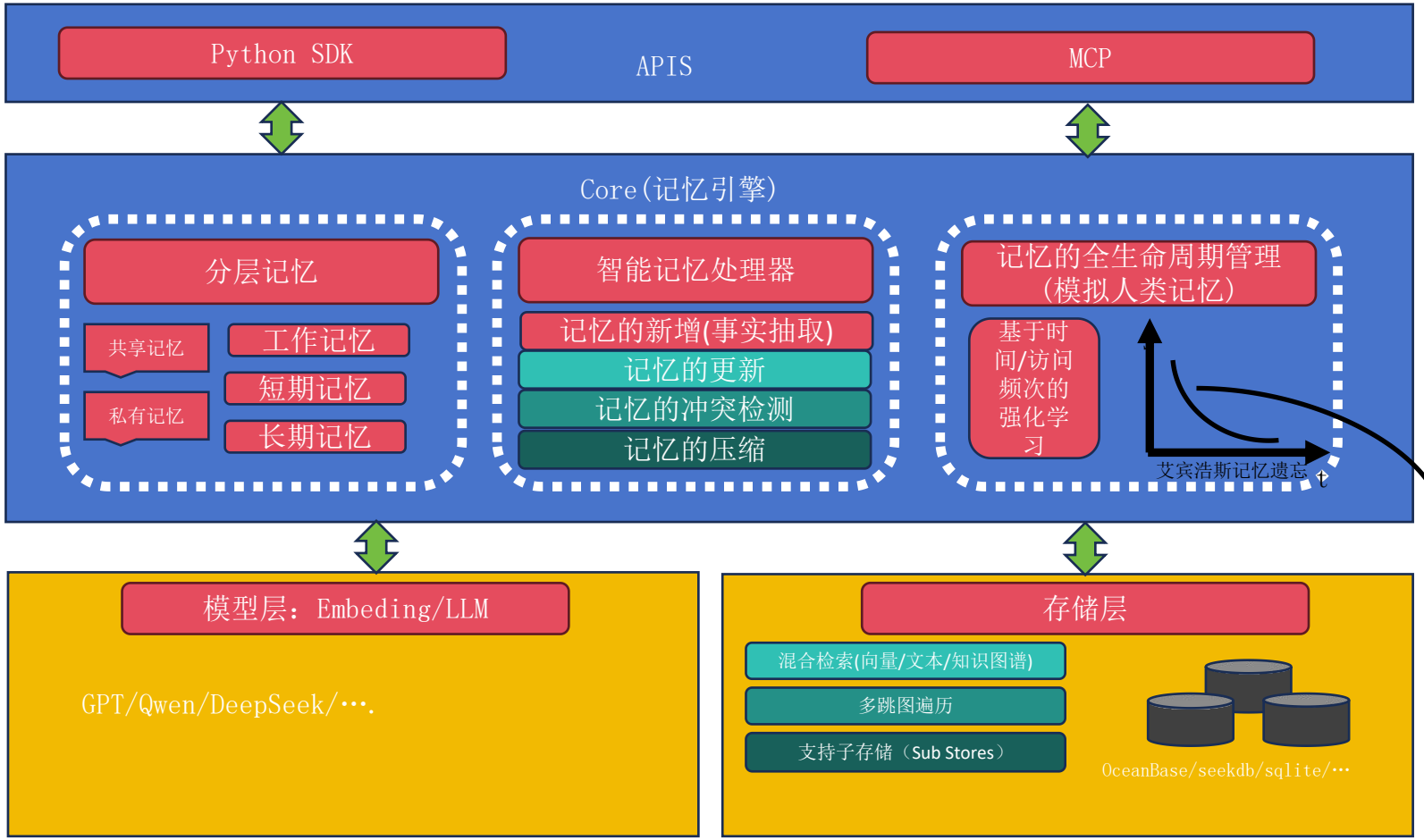
多模态

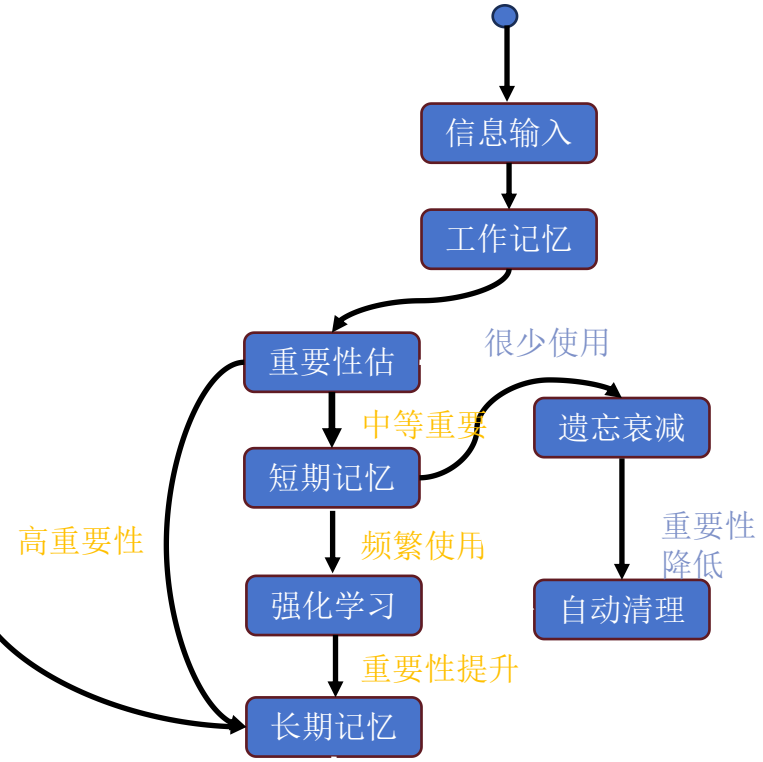
自动将图像和音频转换为文本描述并存储，支持多模态混合内容（文本+图像+音频）的检索

混合检索/多路召回

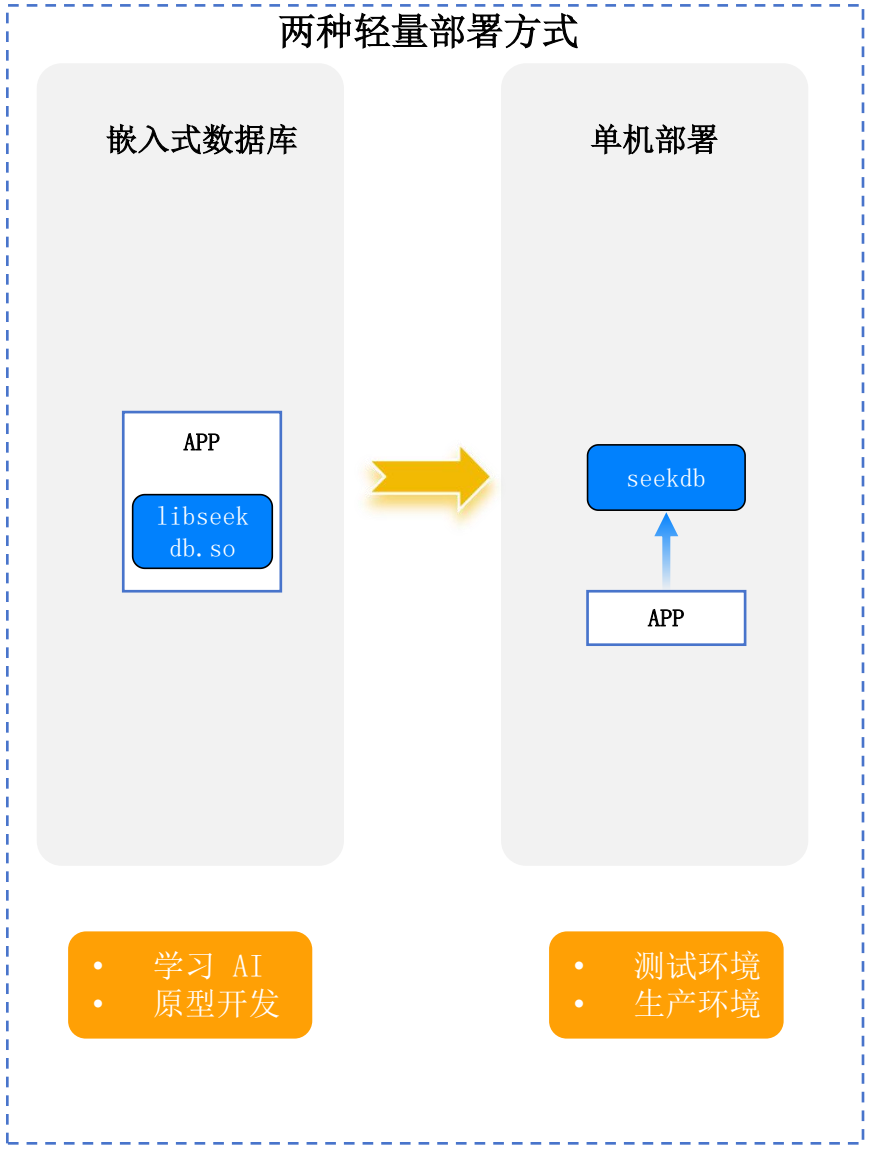
融合向量检索、全文搜索和图检索的多路召回能力，通过 LLM 构建知识图谱并支持多跳图遍历，精准检索复杂的记忆关联关系

powermem 架构图





OceanBase seekdb：AI 原生混合合搜索数据库



1. **开源开放**：Apache 2.0 开源协议
2. **快速构建**
   - SDK 易学易用，开箱即用，极速开发，SDK
   - 支持 1C2G 小规格，
3. **混合搜索**
   - 高性能向量搜索、全文搜索、混合搜索
   - 支持向量索引、全文索引混合多路召回
   - 查询结果重排序支持权重、RRF、大模型排序
4. **多模数据**
   - 结构化、半结构化、非结构化数据
   - 关系表、向量、文本、JSON、GIS等多模数据
5. **AI 内置**
   - DBMS_AI_SERVICE 包管理大模型服务
   - AI_EMBED, AI_COMPLETE, AI_RERANK等函数
6. **SQL Inside**
   - 数据实时写入，实时可查
   - 兼容 MySQL 生态
7. **兼容 OB**：应用可平滑迁移到 OceanBase

OceanBase seekdb 混搜架构

统一应用接口

- 基于 SQL 的支持多模数据的统一查询语言
- 面向开发者更加友好的 Python SDK

支持混合负载的多模计算层

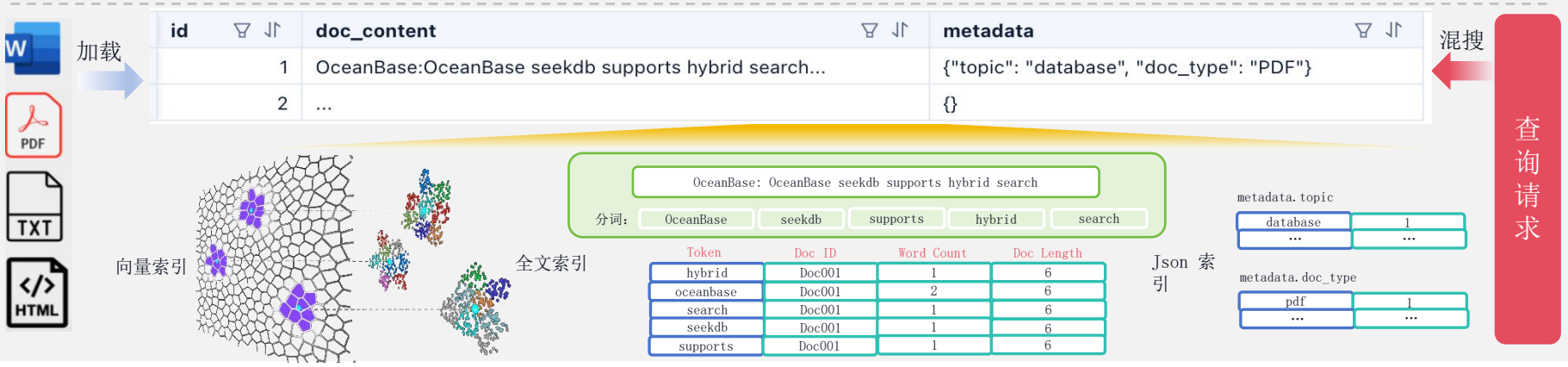
- 混合搜索
- AI 函数
- ACID 事务
- 混合负载自适应执行
- 混合负载查询优化
- 灵活 UDF

多模数据层

- 关系表
- 向量
- 文本
- JSON
- GIS
- 数组/位图...

多模索引层

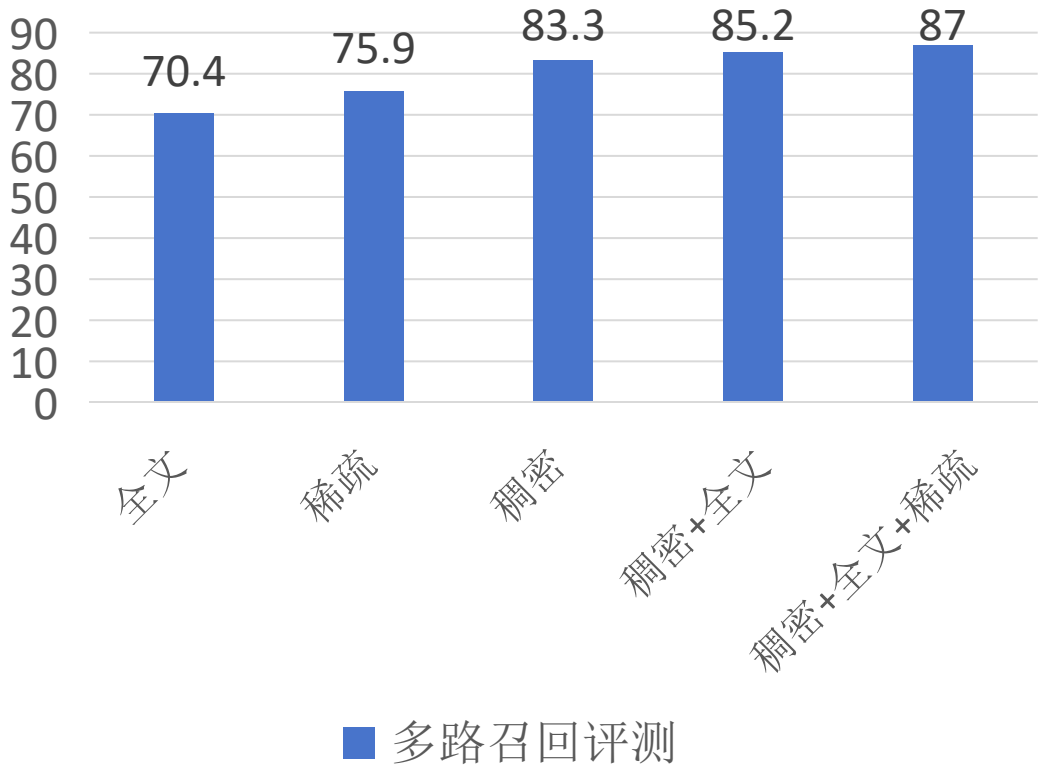
- 向量索引
- 全文索引
- 混合索引
- JSON 索引
- 二级索引
- GIS 索引

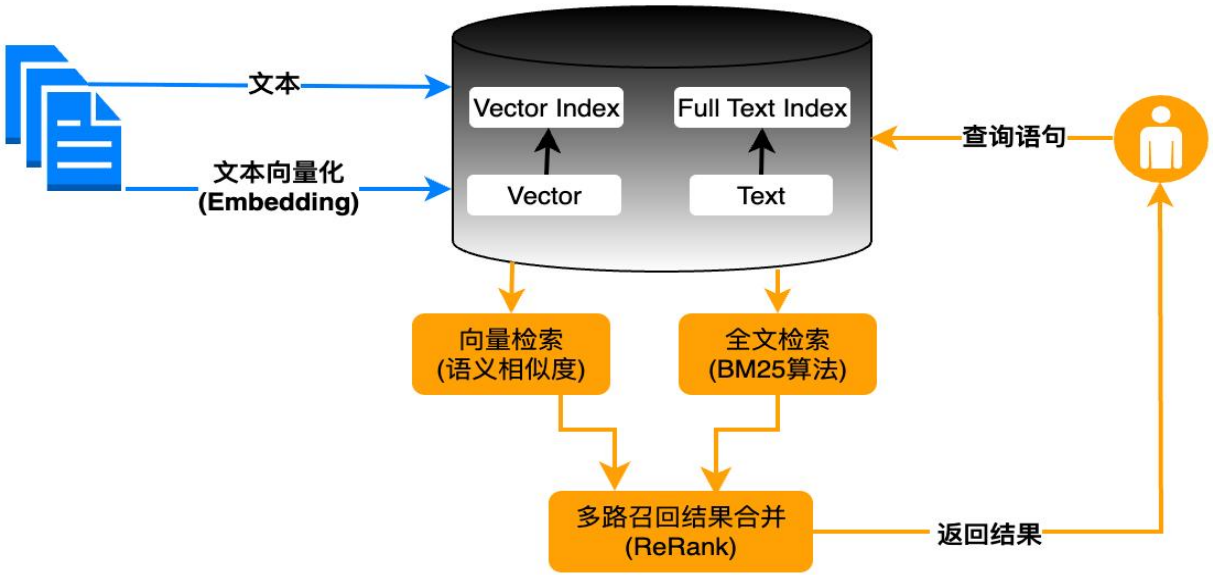


部署模式

- 嵌入式模式
- 服务器模式

混合搜索：更全面的数据关联、更准确的意图命中

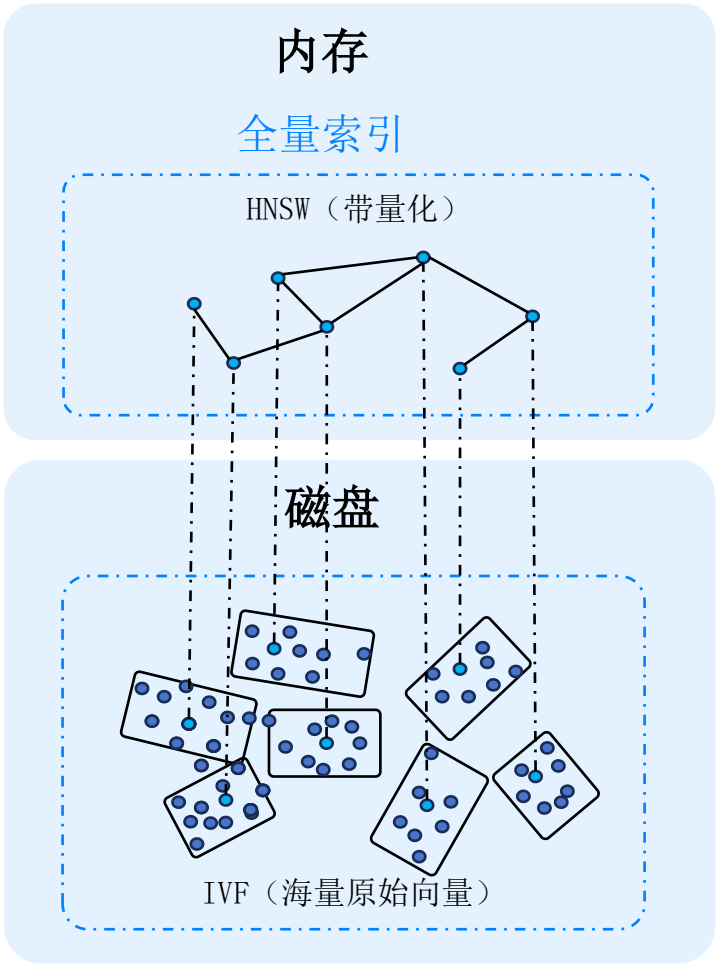




兼具向量搜索的准确性和索引存储的低成本

内存 + 磁盘混合 + 分布

= 超大规模向量低成本管理



极致量化：召回、性能 & 成本，全都要

同等召回率&性能

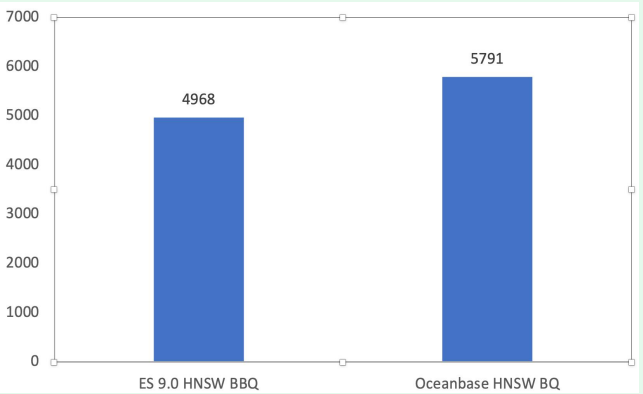
成本较HNSW降低95%

HNSW	HNSW + BQ
1.2TB 内存	58.6GB 内存

向量数量：2亿
向量维度：OpenAI 1536维

同等召回率&同等成本

性能超ES 9.0 BBQ 16%



向量数据集：VectorDBBench中OpenAI 1536维50K数据集
机器环境：8C64G
召回率：0.95

PART 03 典型应用场景

如何上手

STEP 01

安装

Python SDK/MCP server

极其轻量

一条命令搞定

```
pip install powermem
```

STEP 02

使用

add/search/update/delete/reset

方便易用

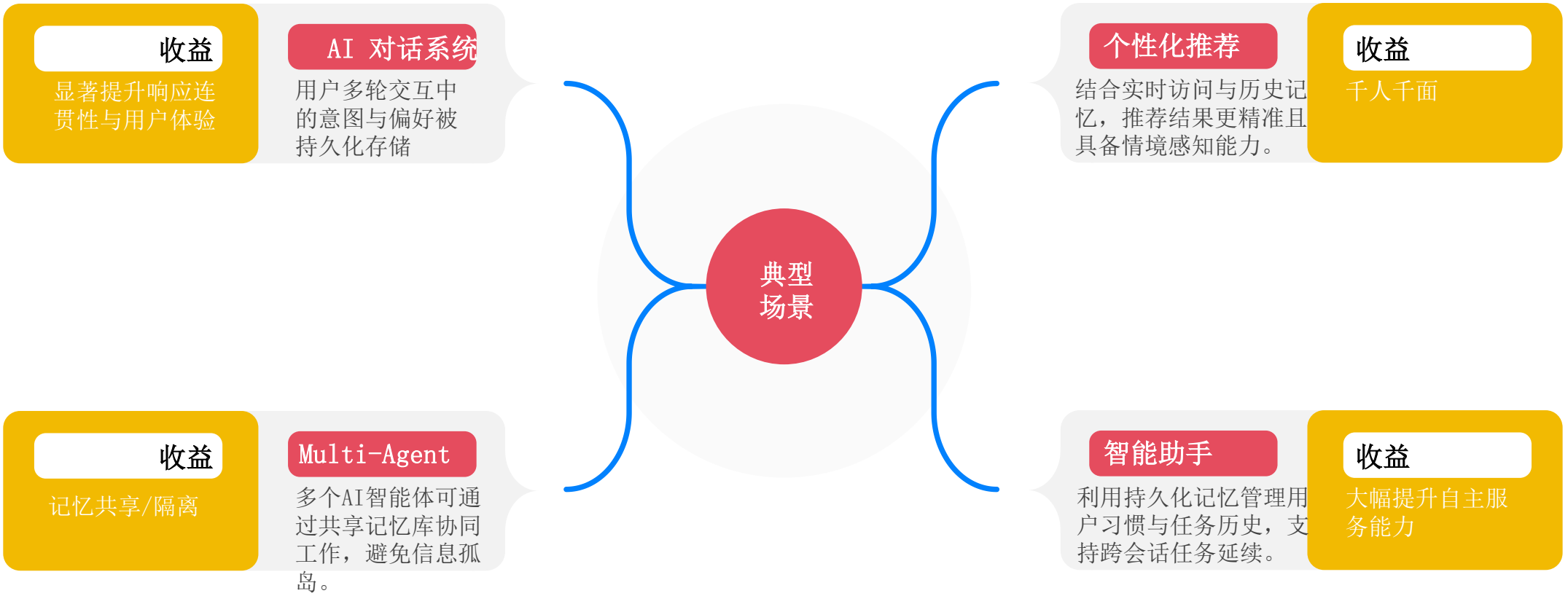
支持同步和异步两种操作

```
from powermem import Memory, auto_config

memory = Memory(config=auto_config())
memory.add("用户喜欢喝咖啡", user_id="user123")
results = memory.search("用户偏好", user_id="user123")
```

典型应用场景

开源社 kaiyuanshe



PART 04 总结与展望

总结：PowerMem + seekdb 的解决方案

省token，高性能

LOCOMO Benchmark 开源 SOTA

省token 96.53%	高召回 78.70	生态兼容
相对于 full-context Token减少 96%	LOCOMO Benchmark Overall 78.70	兼容mem0 接口

基于 seekdb 实现

深度优化、功能强大

混合检索引擎：自动结合标量过滤进行稠密/稀疏向量、全文 混合检索

分层记忆架构：自定义分层管理，独立优化，为不同层选择 更合适 的索引

智能记忆管理：结合遗忘曲线自动降级低频记忆，强化重点记忆，检索 更精确

多模态记忆融合：支持多模态模型，融合文本、语音、图像，更贴近 物理世界

丰富的应用场景

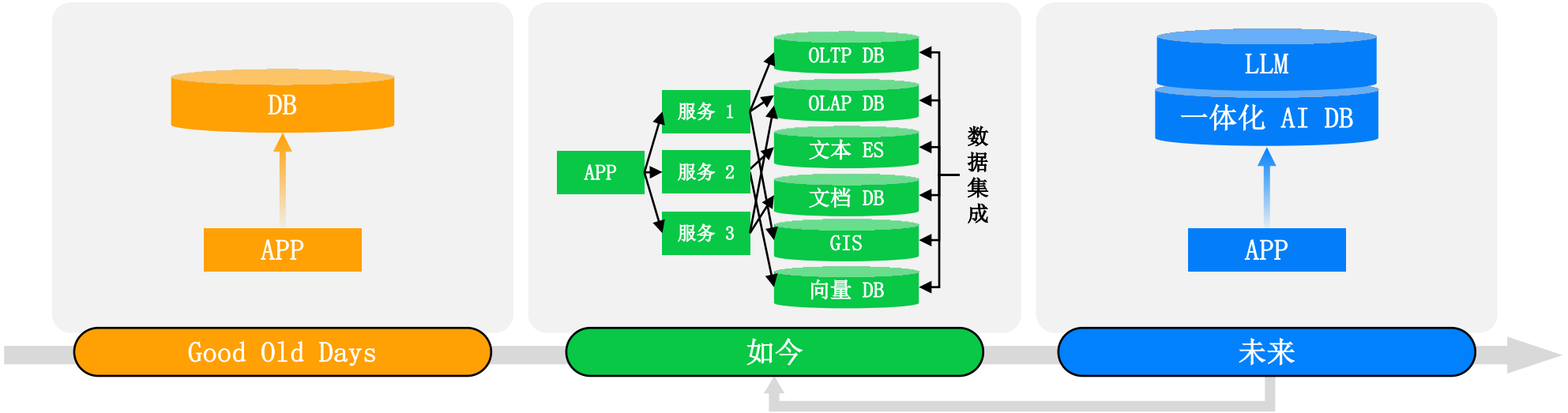
开源开发，生态共赢

智能体（Dify）
- 单智能体长期记忆
- 多智能体协同记忆

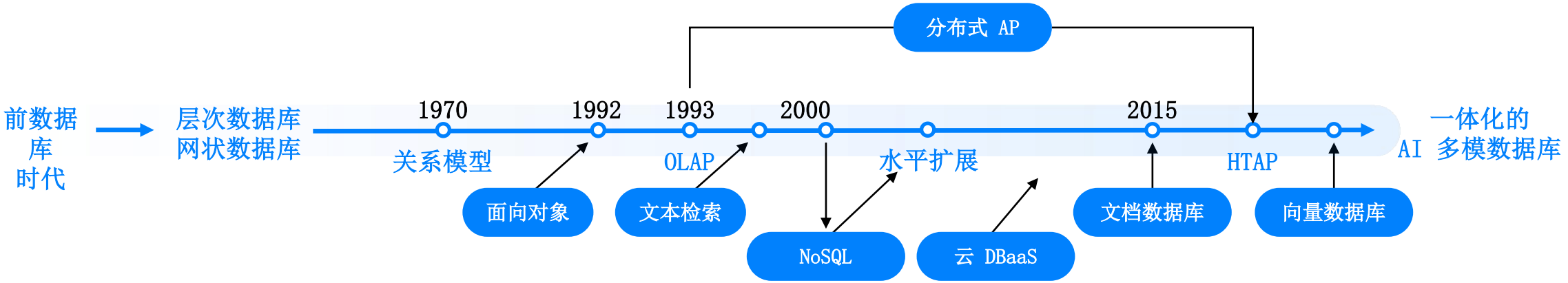
AI coding（Qoder）
- 个人编程习惯
- 项目编程规范

千人千面
- 情感化交互
- 场景化交互
- 个性化推荐
- ……

展望一：融合支持 AI 负载的一体化多模数据库是未来方向



需求：1. 构建 AI应用 所依赖的数据服务过于复杂. 2. 大模型的智能依然依赖于外部数据



数据库与 AI 融合是必然也是最优解：应用使用数据的方式是一体化的，不是割裂的；关系模型的声明式语言使得一体化更具优势

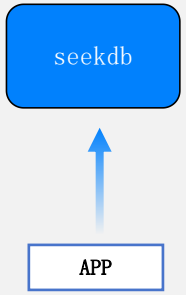
展望二：数据基座弹性架构满足 AI 应用快速迭代的需求

OceanBase seekdb 与 OceanBase 的 API 兼容

嵌入式

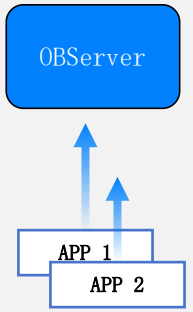
- 个人开发者
- 原型开发

单机部署



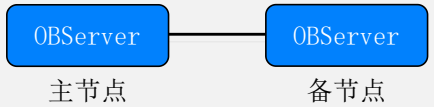
- 小规模工作负载
- 测试和生产环境

单机部署

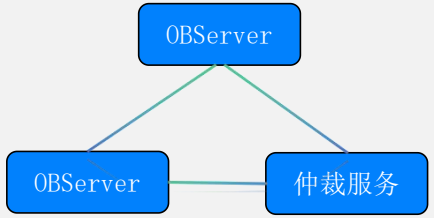


- 小规模工作负载
- 测试和生产环境
- 多租户

单机主备库

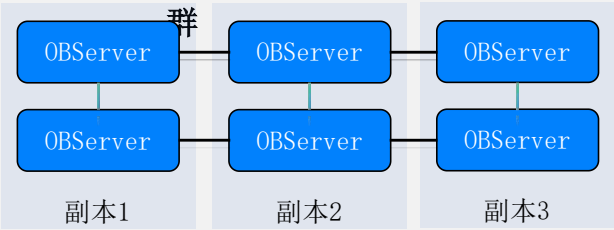


三副本 或 2F1A



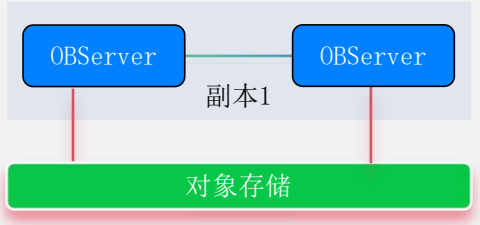
- 中小规模工作负载
- 容灾高可用

无共享的分布式集群



- 大规模工作负载
- 关键业务负载
- 高性能高并发

基于对象存储的存算分离集群



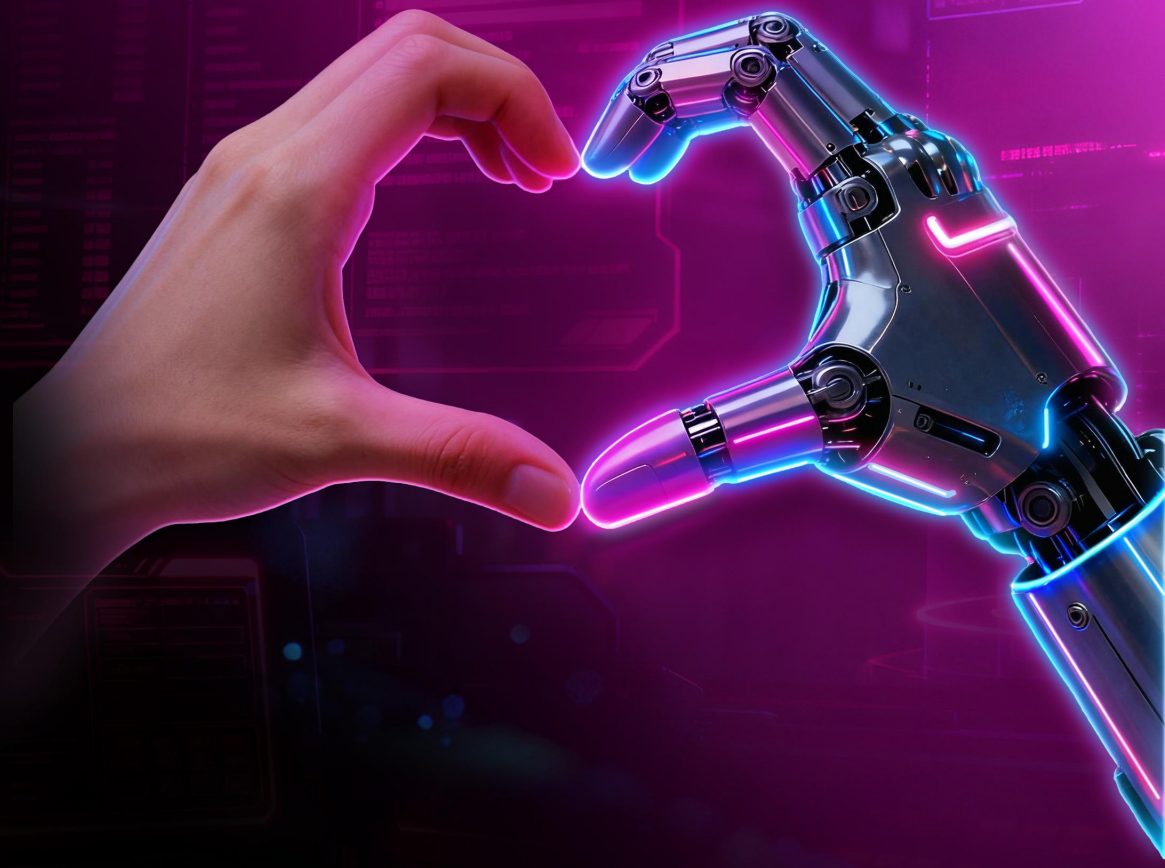
- 大规模工作负载
- 更高性价比

开源社 kaiyuanshe

AI Infra
CCF开源发展技术委员会

COSCon'25
第十届中国开源年会

众智开源 | Open Source, Open Intelligence



Thanks

汤庆

Wechat: OBCE888