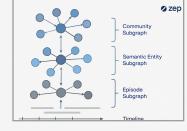


# Graph RAG: Unleashing the Power of



## Knowledge Graphs with LLM

# ZEP: 一种用于代理记忆的 时间知识图谱架构

解析AI代理的长期记忆解决方案

LLM

Zep AI团队

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, Daniel Chalef

# 问题背景

## ! LLM代理的记忆限制

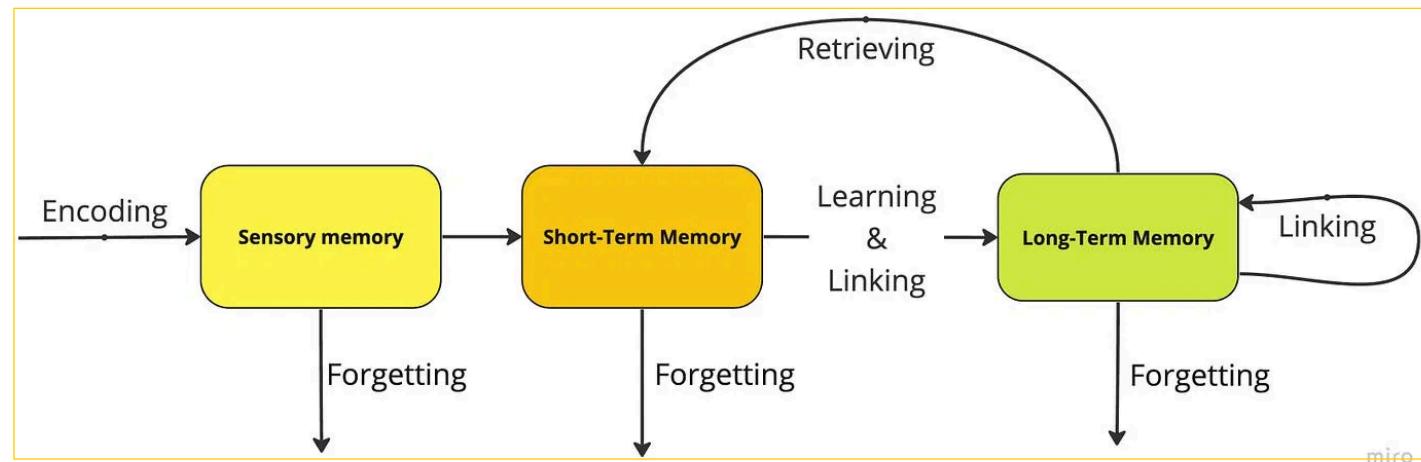
大型语言模型（LLMs）的上下文窗口有限，无法有效利用长期记忆，导致对话连贯性和知识保留能力受限。

## ! 现有RAG方法的局限性

传统检索增强生成（RAG）框架主要依赖静态文档检索，难以处理动态变化的信息和复杂的时间关系。

## ! 企业应用的特殊需求

企业级应用需要从多样化来源（对话历史、业务数据等）动态整合知识，并保持历史关系的完整性。



# Zep解决方案概述

## ✓ Graphiti：时间感知知识图谱引擎

Zep的核心组件，一个动态、时间感知的知识图谱引擎，能够表示复杂、不断演变的世界。

## ✓ 动态合成多源数据

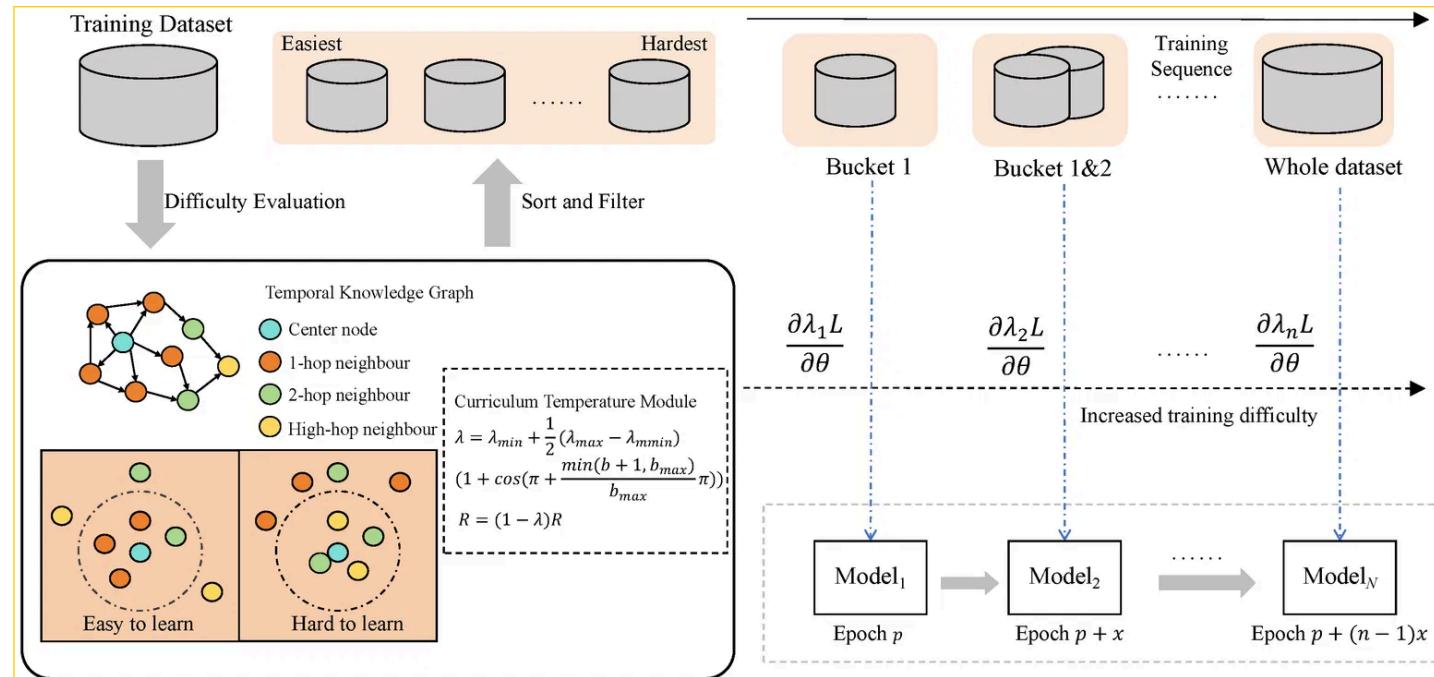
无缝整合非结构化对话数据和结构化业务数据，提供全面的上下文理解。

## ✓ 时间线维护

以非损失方式动态更新知识图谱，维护事实和关系的有效期时间线，支持时间推理。

## ✓ 生产级性能

专注于准确性、延迟和可扩展性，在DMR和LongMemEval基准测试中表现优异。



# 知识图谱架构

## 三层子图结构

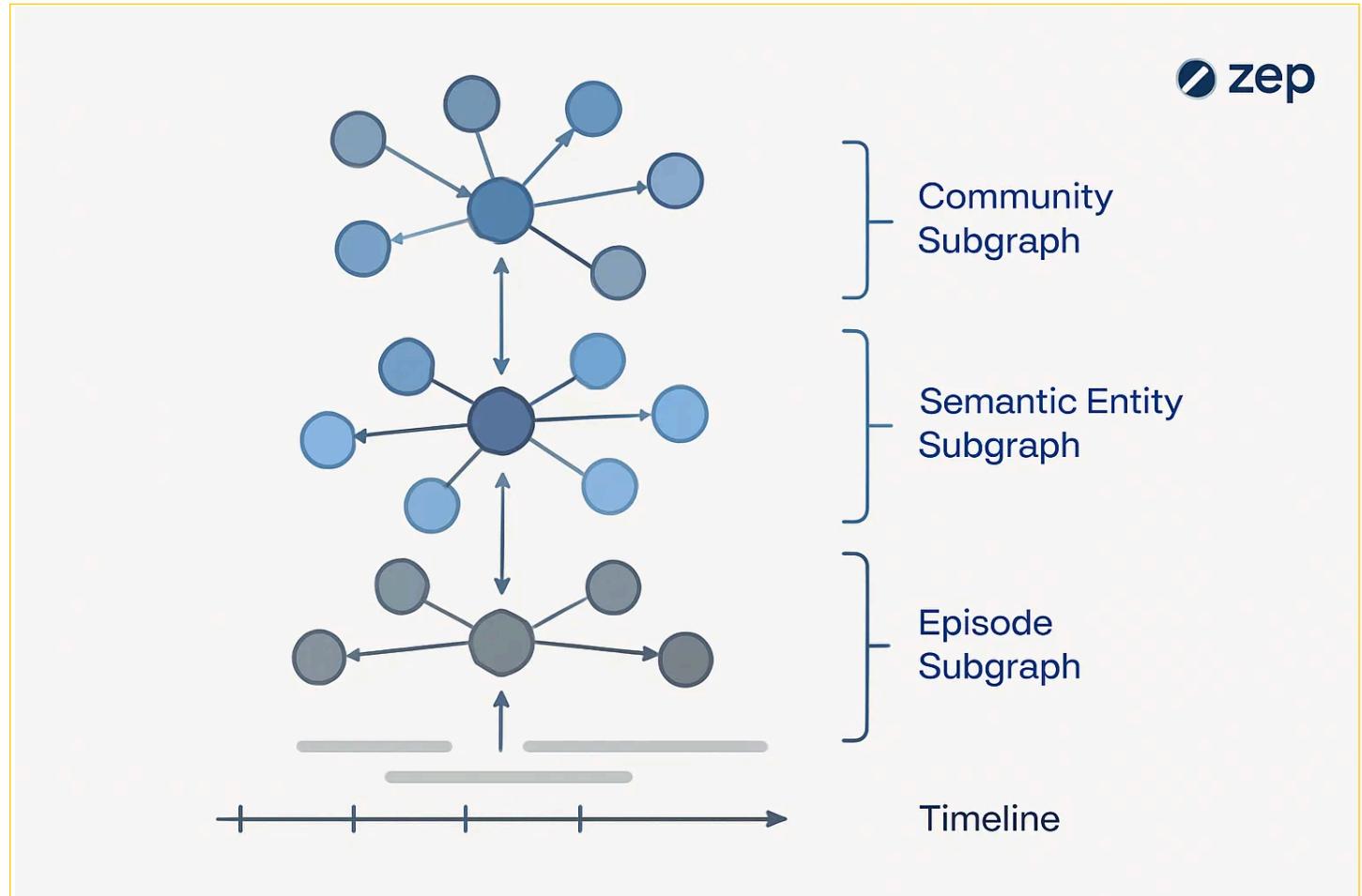
- **情景子图 (Ge)**：包含原始输入数据（消息、文本、JSON）
- **语义实体子图 (Gs)**：表示从情景中提取的实体及其关系
- **社区子图 (Gc)**：表示强连接实体的集群及其高级摘要

## 双时间模型

- **时间线 T**：事件的编年顺序
- **时间线 T'**：Zep数据摄取的事务顺序

## 人类记忆模型对应

模拟人类记忆系统中的情景记忆（特定事件）和语义记忆（概念关联）。



# 核心技术：情景处理

## 情景类型

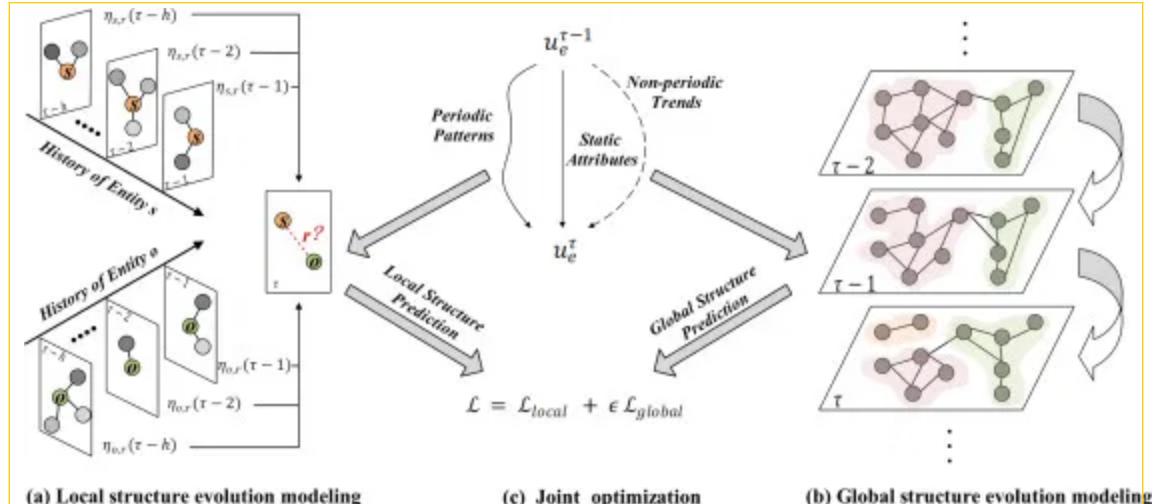
- 消息：对话内容与发言者信息
- 文本：结构化文档内容
- JSON：结构化业务数据

## 时间戳处理

每个情景包含参考时间戳 $t_{ref}$ ，用于准确提取和表示绝对时间戳（如"2023年6月1日"）和相对时间戳（如"两周前"、"下周四"）。

## 双向索引与无损存储

情景边(Ee)连接情景与提取的实体节点，维护双向索引以实现前向和后向遍历，确保语义信息可追溯到源头。



```
// 情景数据示例 { "type": "message", "content": "我下周四要去北京出差",  
"actor": "用户", "timestamp": "2023-06-01T10:30:00Z" }
```

# 核心技术：实体与事实

## 实体提取与解析流程

提取实体 → 生成嵌入 → 实体解析 → 图谱整合

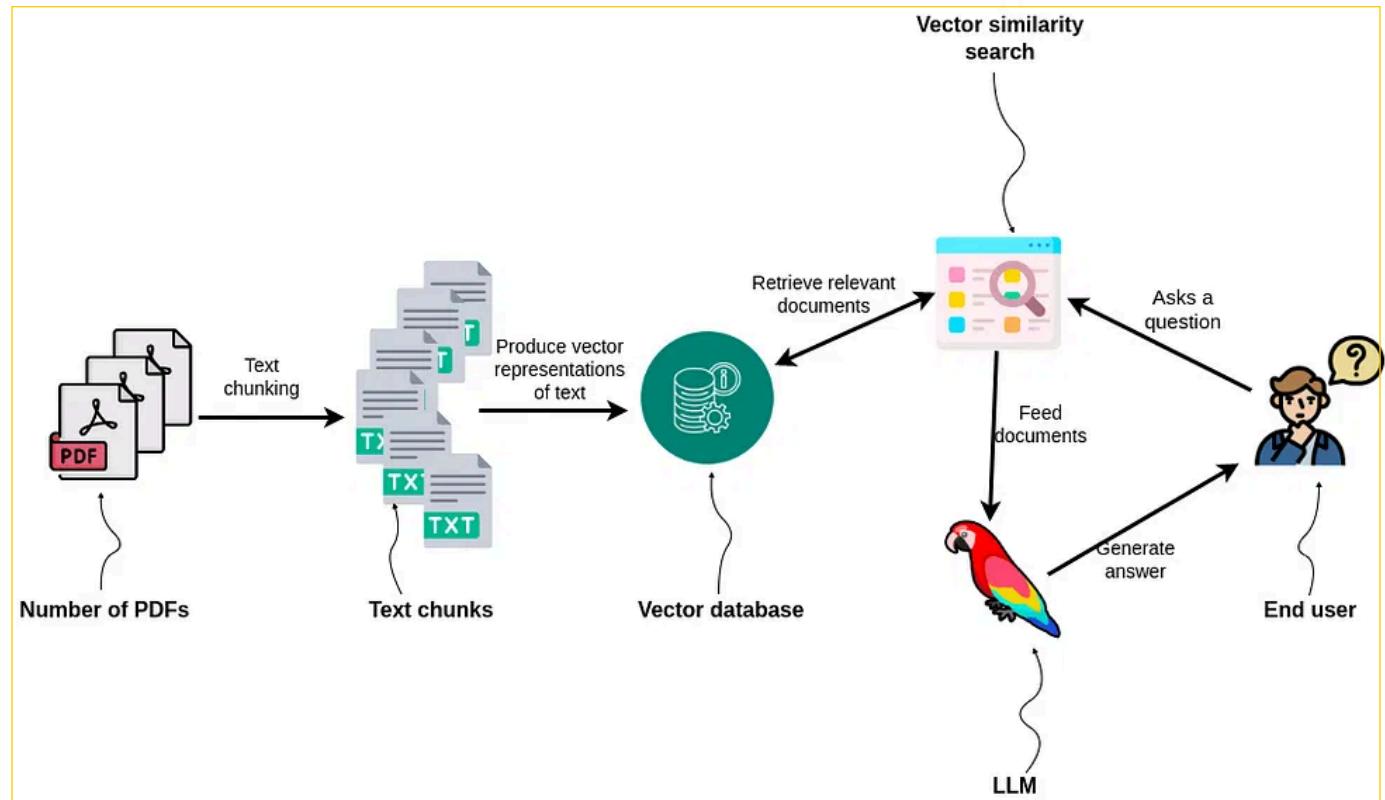
使用反思技术（reflexion）最小化幻觉并增强提取覆盖率，实体名称嵌入到1024维向量空间。

## 事实提取与超边实现

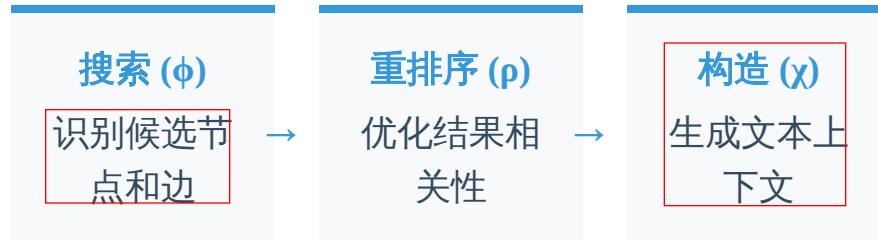
同一事实可在不同实体间多次提取，通过超边实现复杂多实体事实的建模，并使用相似的解析流程进行边去重。

## 时间提取与边失效机制

- 四个时间戳： $t'_{\text{created}}$ 、 $t'_{\text{expired}} \in T'$ （系统时间）和  $t'_{\text{valid}}$ 、 $t'_{\text{invalid}} \in T$ （事实有效期）
- 边失效：新边可使相关现有边失效，设置其  $t'_{\text{invalid}}$  为新边的  $t'_{\text{valid}}$



# 记忆检索系统



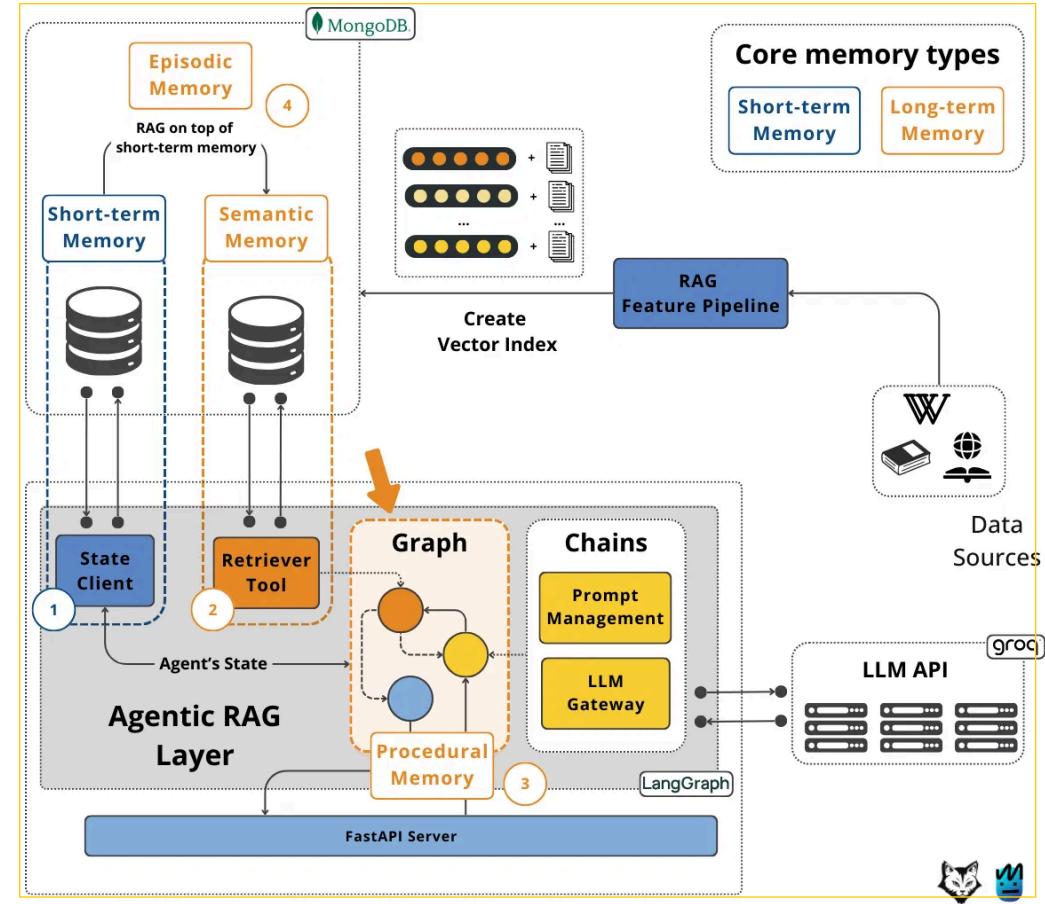
$$f(\alpha) = \chi(\rho(\phi(\alpha))) = \beta$$

## 多种搜索方法

- 余弦语义相似度搜索 ( $\phi_{cos}$ )：捕获语义相似性
- Okapi BM25全文搜索 ( $\phi_{bm25}$ )：识别词语相似性
- 广度优先搜索 ( $\phi_{bfs}$ )：揭示上下文相似性

## 重排序策略

- RRF和MMR：传统重排序方法
- 情景提及重排序器：基于实体/事实提及频率
- 节点距离重排序器：基于图距离
- 交叉编码器：使用LLM生成相关性分数



# 性能评估

## DMR基准测试

在Deep Memory Retrieval基准测试中，Zep达到94.8%的准确率（使用gpt-4-turbo），超过MemGPT的93.4%。

## LongMemEval基准测试

在更具挑战性的LongMemEval基准测试中，Zep使用gpt-4o达到71.2%的准确率，比基线提高了18.5%。

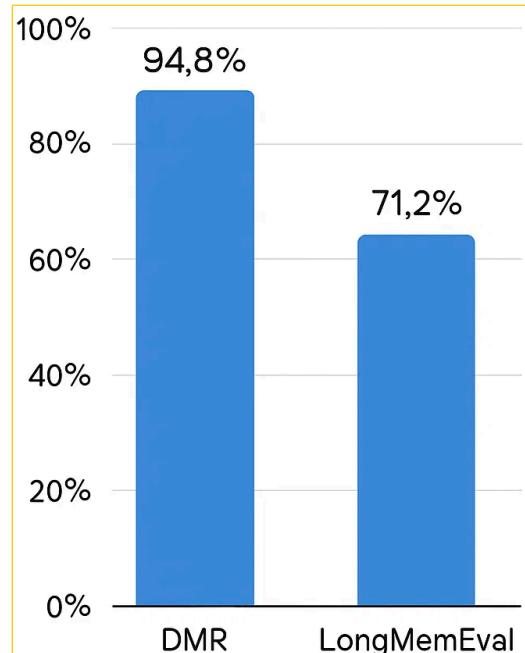
## 延迟改进

Zep将响应延迟从基线的28.9秒减少到2.58秒，降低了约90%，同时保持更高的准确率。

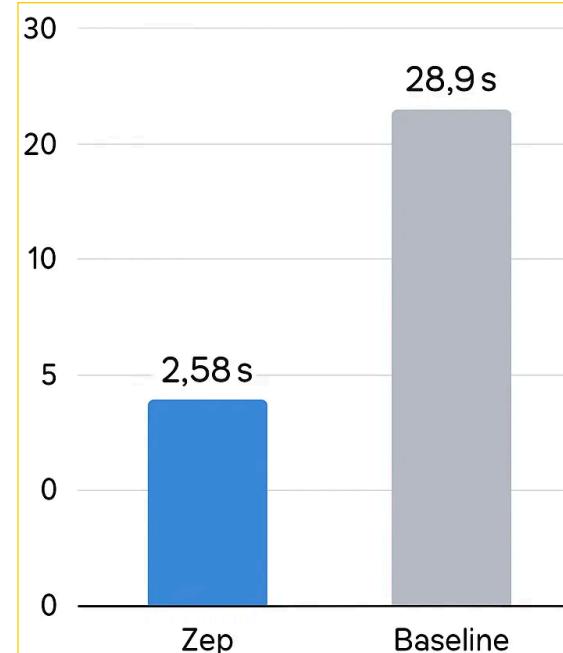
## 复杂任务表现

在复杂问题类型（如单会话偏好、多会话和时间推理）中表现出最显著的提升，特别是与更强大的模型结合时。

Accuracy Comparison



Latency Comparison



问题类型	基线 (gpt-4o)	Zep (gpt-4o)	提升
单会话偏好	20.0%	56.7%	+184%
时间推理	45.1%	62.4%	+38.4%
多会话	44.3%	57.9%	+30.7%
单会话用户	81.4%	92.9%	+14.1%

# 应用场景与未来展望

## ■ 企业级应用场景

客户服务代理

业务智能助手

销售支持系统

知识管理平台

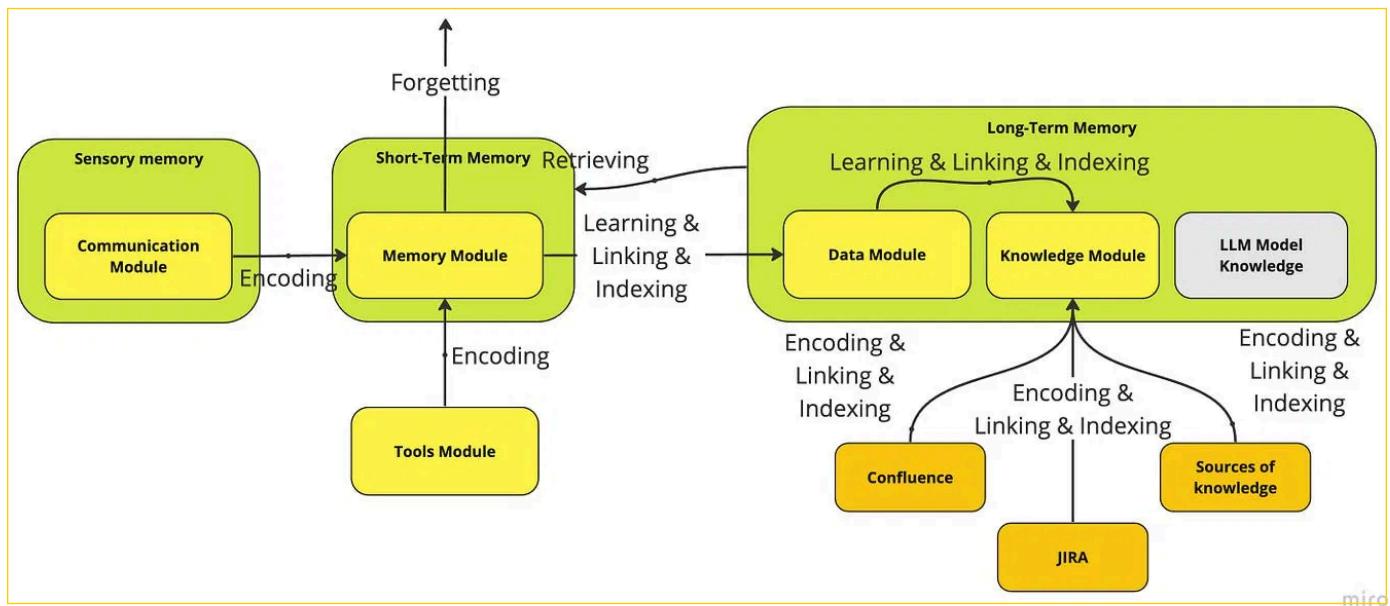
## 研究方向

- **模型微调**：为Graphiti提示词微调专用模型，提高准确率并降低成本和延迟
- **领域本体**：探索特定领域知识图谱本体在Graphiti框架中的应用
- **多模态集成**：扩展到图像、音频等多模态数据的处理和记忆

## ■ 集成可能性

与其他GraphRAG方法（如AriGraph

GraphRAG、LightRAG等）的集成，结合各自优势创建更强大的记忆系统。



# Graph RAG: Unleashing the Power of

## 总结与参考

### 核心创新点

## Knowledge Graph

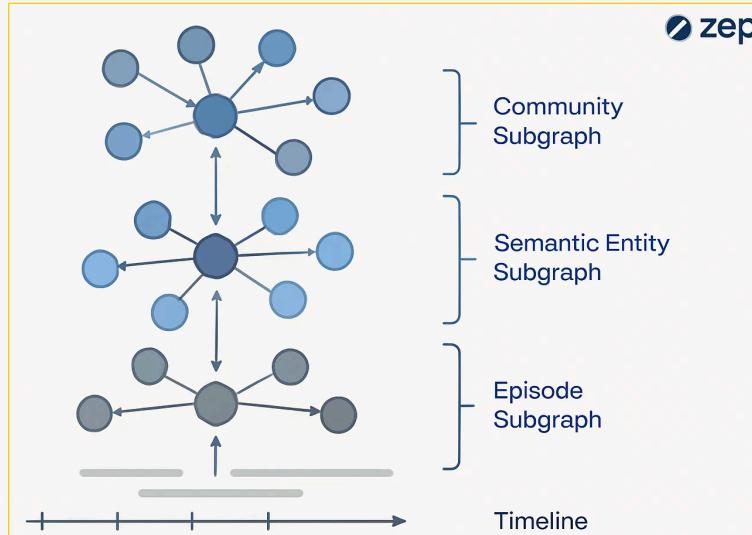
时间感知知识图谱：维护事实和关系的有效期  
时间线

三层子图结构：情景、语义实体和社区层次

动态更新机制：非损失方式整合新信息

### 关键技术优势

- 优异的准确性：在DMR和LongMemEval基准测试中表现卓越
- 显著的延迟改进：响应时间减少约90%
- 复杂任务处理能力：在跨会话信息合成和长期上下文维护方面表现突出



### 主要参考文献

- [1] Zep: Long-term memory for AI agents. <https://www.getzep.com>
- [2] Graphiti: Temporal knowledge graphs for agentic applications. <https://github.com/getzep/graphiti>
- [3] Packer, C. et al. Memgpt: Towards llms as operating systems, 2024.
- [4] Edge, D. et al. From local to global: A graph rag approach to query-focused summarization, 2024.
- [5] Wu, D. et al. Longmemeval: Benchmarking chat assistants on long-term interactive memory, 2024.

联系方式：[info@getzep.com](mailto:info@getzep.com) | [github.com/getzep/graphiti](https://github.com/getzep/graphiti)