

基于 RAG/Agent 技术的混元大模型业务落地实践

演讲嘉宾：赵喜生，腾讯机器学习平部

CONTENTS

- ① 腾讯大语言模型应用场景
- ② RAG技术原理及优化实践
- ③ GraphRAG在角色扮演场景中的应用
- ④ Agent技术原理和应用

1 第一节

腾讯大语言模型应用场景

腾讯大模型应用场景

内容生成

文本创作
内容扩写
文案生成
定制翻译
辅助评论
文案润色
素材生成
输入联想

内容理解

文本审核
图文匹配
实体提取
恶意判断
标签提取
诈骗识别
文本摘要
文本分类

智能客服

知识问答
问题推荐
用户引导
情绪理解
文档提取
交互式任务
智能客情

开发Copilot

代码评审
低代码生成
自动化测试
代码生成
Text2SQL
自动补全
优化建议
代码解读

角色扮演

角色扮演
数字人
情感陪伴
游戏NPC
剧情演绎
游戏会话



视频号

腾讯体育

腾讯新闻



王者荣耀

腾讯游戏
Tencent Games

腾讯地图



企业微信

QQ

腾讯文档

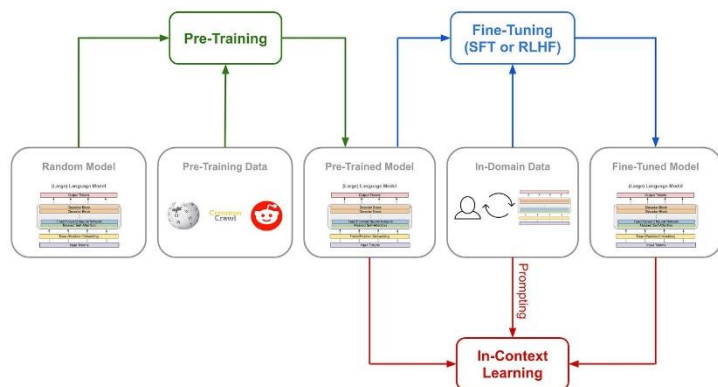
腾讯云

腾讯广告

腾讯会议
Tencent Meeting

大模型应用技术

SFT



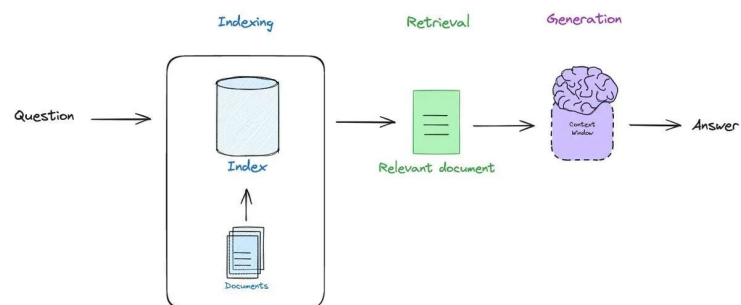
Pros

- 特定领域微调后效果较好
- 端到端生成结果

Cons

- 需要构建标签数据
- 训练和使用成本较高
- 存在大模型幻觉

RAG



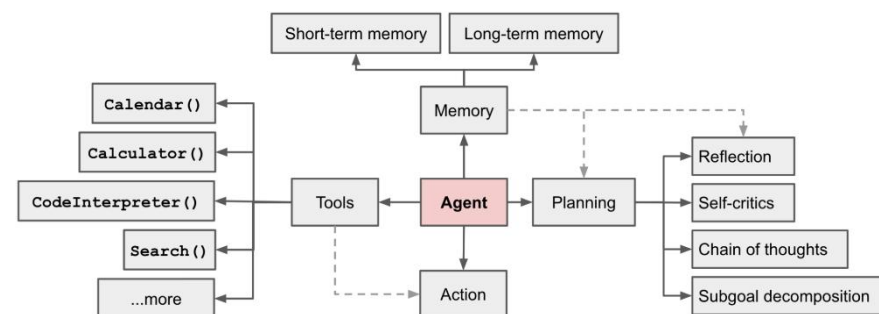
Pros

- 外部知识检索，实时性强，幻觉低
- 过程透明，可解释性强

Cons

- 生成效果依赖于知识库质量和召回准确率
- 延时高

Agent



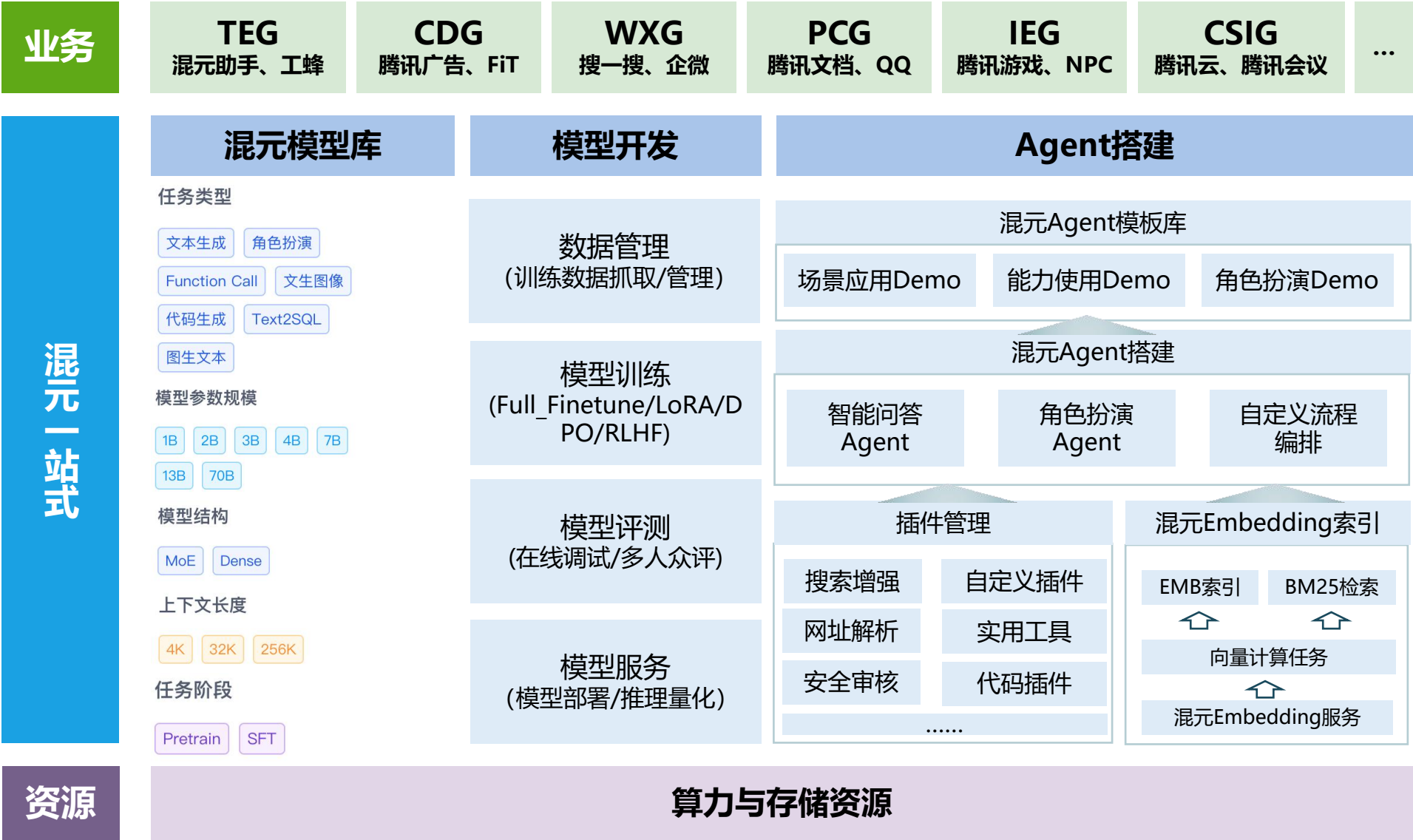
Pros

- 具有自主性、交互性、适应性和推理能力
- 可处理复杂任务，支持多模态任务处理

Cons

- 结果不稳定
- 模型训练优化成本高

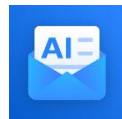
混元一站式大模型应用解决方案



2 第二章节

RAG技术原理及优化实践

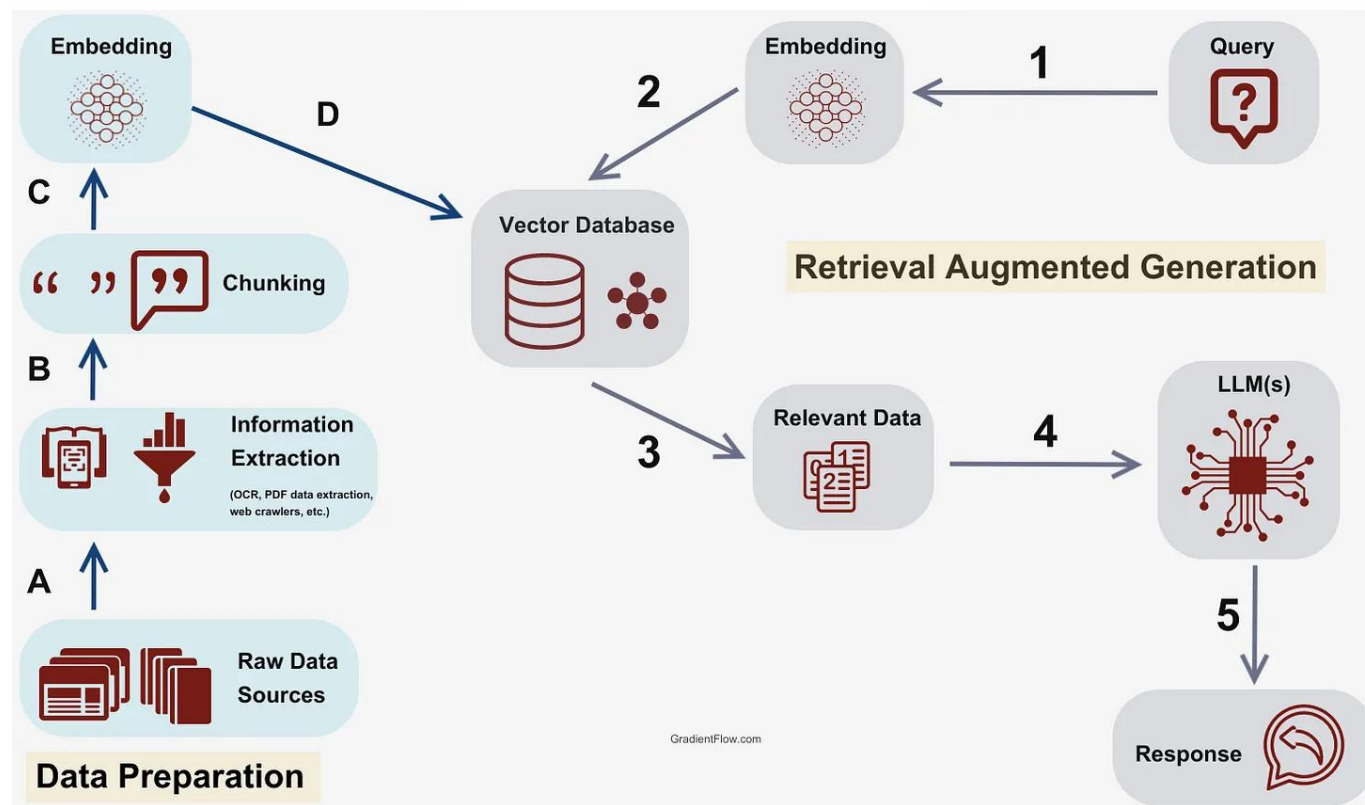
RAG技术介绍



- 参考外部知识
- 基于事实无幻觉
- 数据更新及时
- 回复具有可解释性
- 安全和隐私

数据准备
知识库构建

知识召回
生成增强



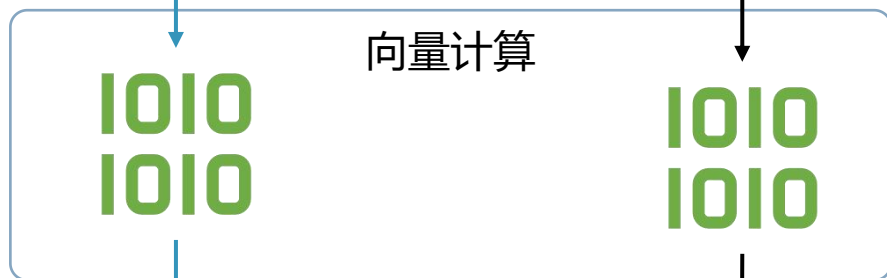
RAG技术原理

QA Knowledge Base

Q	Doc
腾讯成立于哪一年?	腾讯是中国领先的互联网科技公司，成立于1998年，总部位于深圳。旗下拥有众多知名产品，如社交应用微信和QQ，以及游戏、广告、金融科技、云计算等业务。腾讯致力于通过技术创新，为用户提供更好的产品和服务，同时也推动了整个中国互联网行业的发展。
腾讯总部在哪里?	
.....	
腾讯旗下的主要产品有哪些?	

Q 腾讯公司CEO是谁?

Query



Vector Database



Recalled Docs



LLM

RAG应用关键挑战

多格式内容提取

- 文档格式种类多样
- 布局格式复杂
- 公式、表格识别和提取
- 文本、图片、表格嵌套

文档切分

- 不同格式的文本文档采用何种切分方式
- 如何确保文档切片的语义完整性

知识库构建

- 在没有QA问答对或只有少量QA问答对的情况下，如何帮助扩充知识库
- 如何确保知识更新的时效性

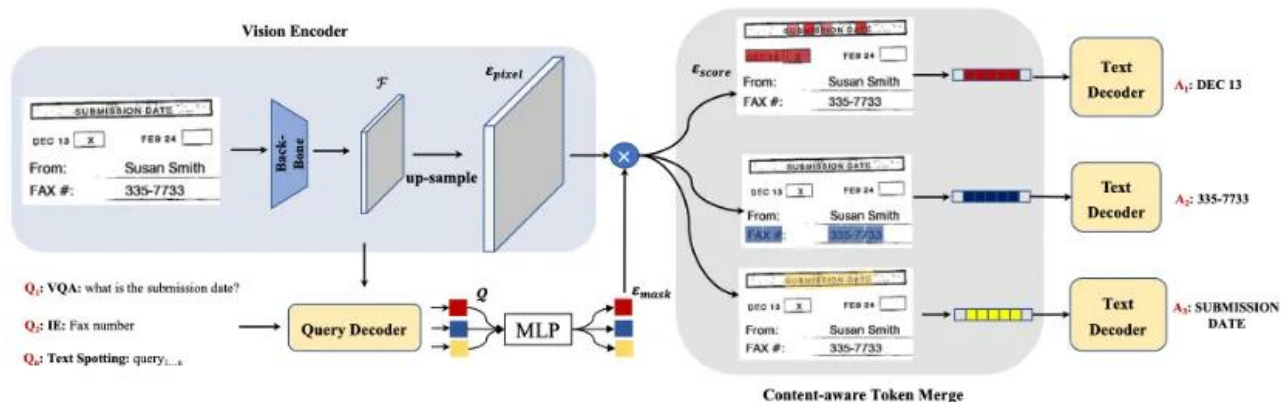
文档召回

- 如何保证召回内容的相关性
- 如何融合多种召回方式召回的文档

内容生成

- 如何确保模型跟随指令
- 解决数据隐私、安全问题
- 如何提升模型回答的效果
- 如何让模型具备领域知识

文档解析



输入图片

Markdown渲染结果

支持元素多样

能够解析文档中的段落、表格、公式、图像、标题、页眉、页脚等多种版面元素

应用场景广泛

覆盖论文、杂志、说明书、试卷、海报、PPT等众多场景

算法能力全面

支持版面分析、元素排序、文字识别、表格识别、公式识别等多种算法能力。

模型准确率高

基于优图自研新一代多模态文档解析大模型，各版面元素解析准确率达到88%，效果业内领先。

文档切分

固定长度文本切分

按照固定的文本长度切分文本，不同分块之间可以有固定长度重叠内容

Markdown标题切分

根据Markdown原生标题分割文本内容，将相同标题级别的文本片段切分在同一个chunk中

递归文本切分

按照分割字符递归切分，递归分块使用一组分隔符以分层和迭代的方式将输入文本分成更小的块

中文语义切分

使用模型文档根据语义分割模型来切分文档

离线知识扩充技术

DocQAGenerator

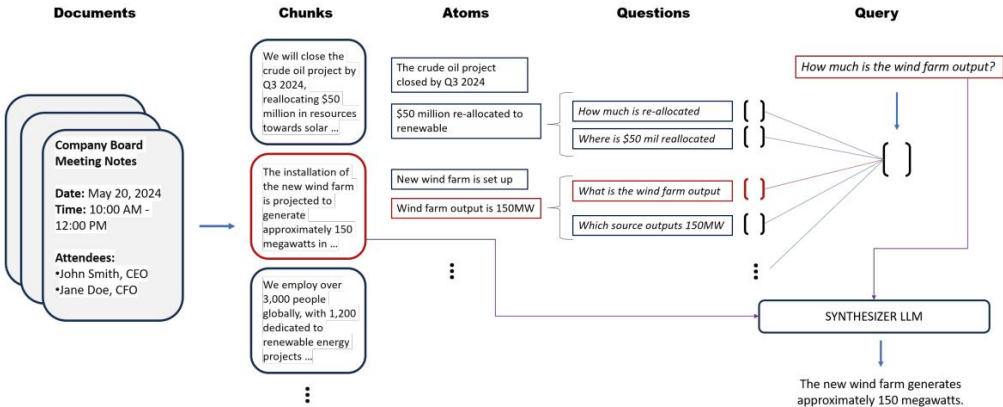
直接基于原始文档生成可能的QA对

AugmentedQuestionGenerator

在用户提供了<Question, Context>对的情况下，
基于当前问题和上下文，为上下文生成更多可能的
用户问题

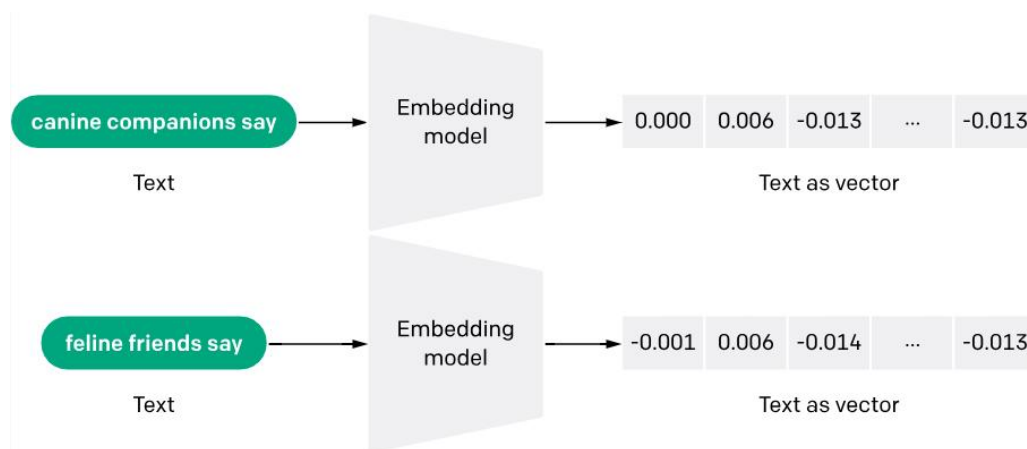
AtomicUnitsQAGenerator

Atomic Units QA的方法首先对原始文本进行分块，
然后将块分解为原子陈述，再针对这些原子（以块
为上下文）生成一组合成问题

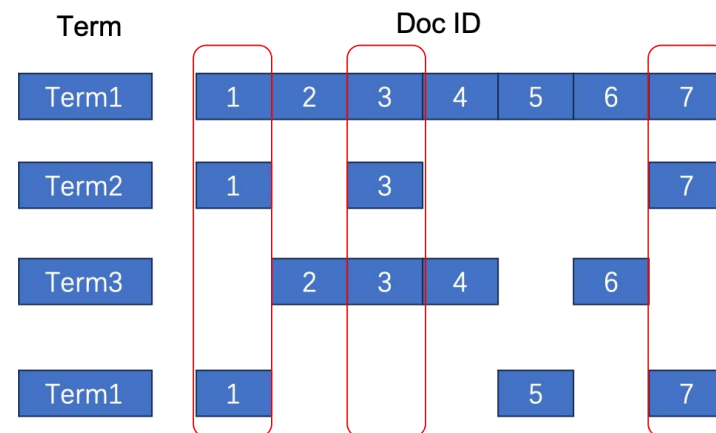


Question-Based Retrieval using Atomic Units for Enterprise RAG

索引召回



Embedding

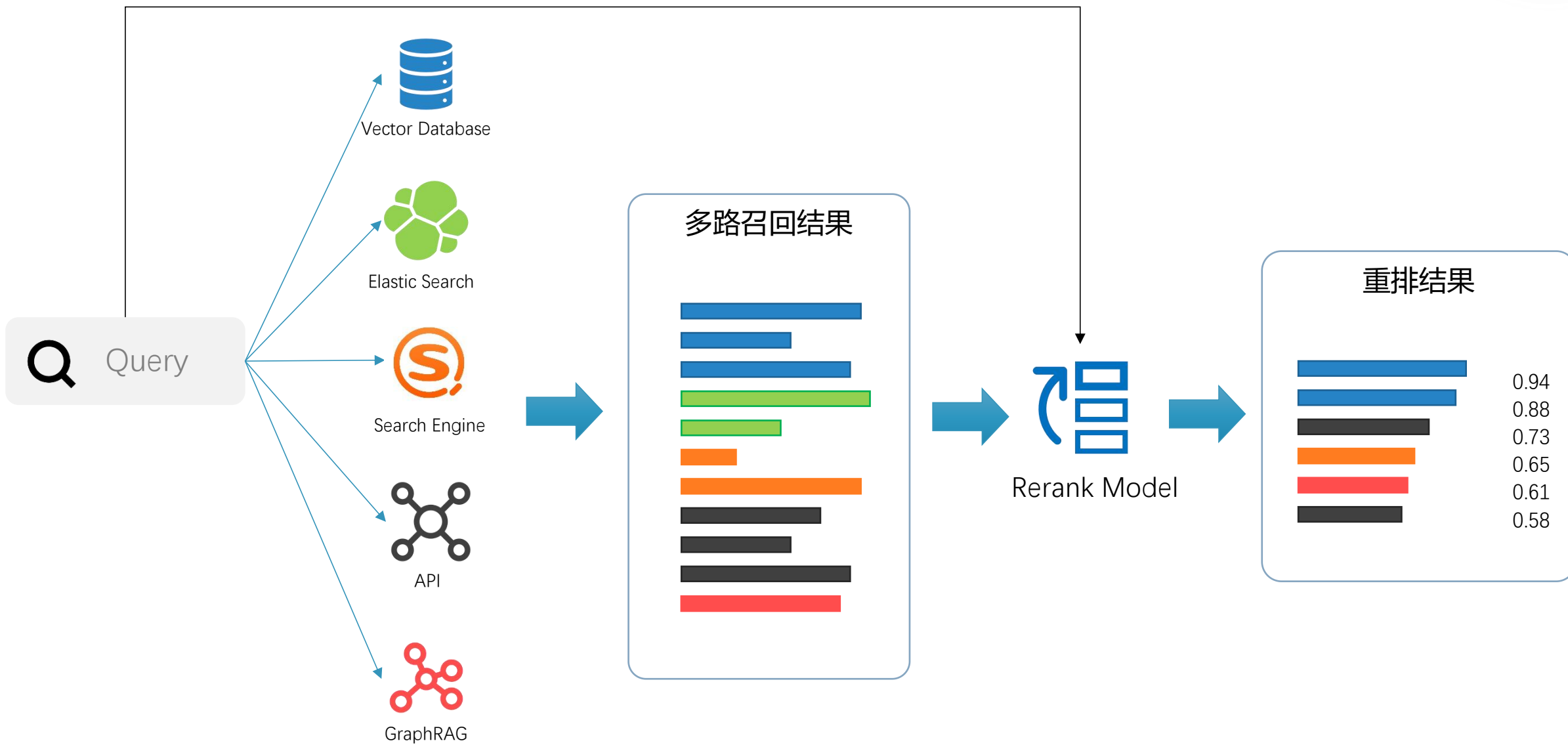


BM 25

Approximate Nearest Neighbor

- BruteForce算法是最直接的搜索方法，它通过计算查询向量与数据库中每个向量之间的距离来找到最近邻，因此可以得到精确的结果，适合于较小的数据集计算；
- HNSW算法是一种基于图结构的近似最近邻搜索方法，通过局部性原理和层次化结构来加速搜索过程，适用于大规模数据集且对搜索效率有较高要求的场景。

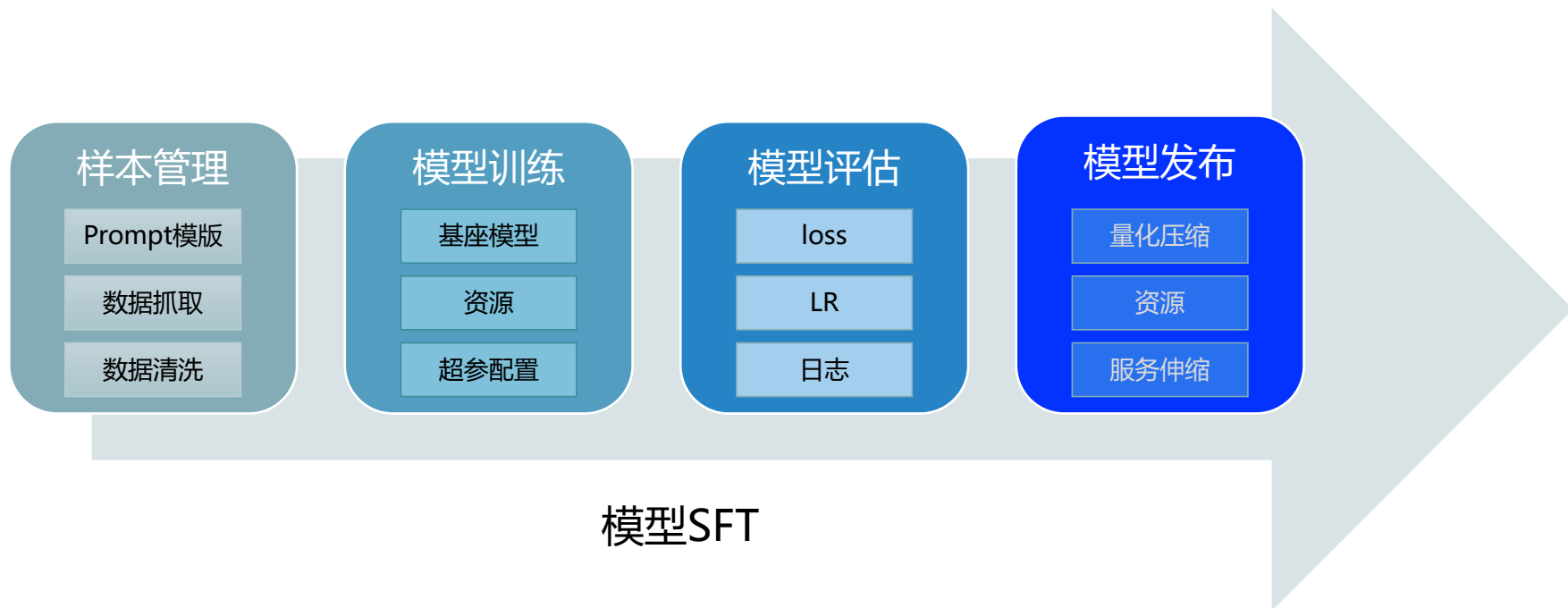
多路召回



知识生成



Prompt工程



模型SFT

3 第三章节

GraphRAG在混元大模型中的应用

RAG局限



孙悟空的如意金箍棒的由来

RAG局限

- 忽视关系
- 冗余信息
- 缺乏全局信息
- 长文逻辑推理能力



太上老君
九转镔铁炼制

大禹治水时
测江河深浅

大禹治水后
东海定海神针

东海龙王
镇压海眼

GraphRAG: 基于图的检索增强方法

GraphRAG 框架:

- G-Indexing: 基于输入语料构建知识图谱与摘要等并存储于图数据库
- G-Retrieval: 使用问题对图数据库进行检索, 获得相关上下文
 - Local: 检索与提问相关的实体、关系、社区等信息, 生成上下文
 - Global: 检索全部实体, 社区信息, 生成上下文
- G-Generation: 问题 + 上下文 => 回答
 - Local: 直接生成回答
 - Global: 需经过reduce阶段, 生成回答

上下文理解

- 更完整的上下文信息, 帮助理解复杂查询

信息整合能力

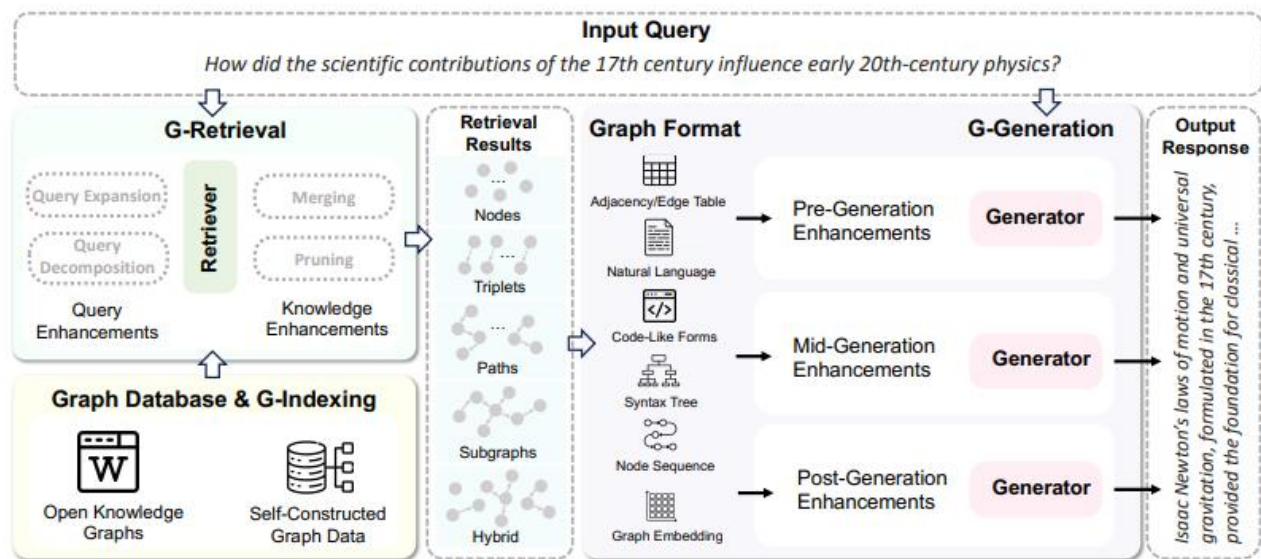
- 整合异构、相互关联的信息

知识推理能力

- 利用知识图谱进行知识推理

可解释性

- 方便信息追溯



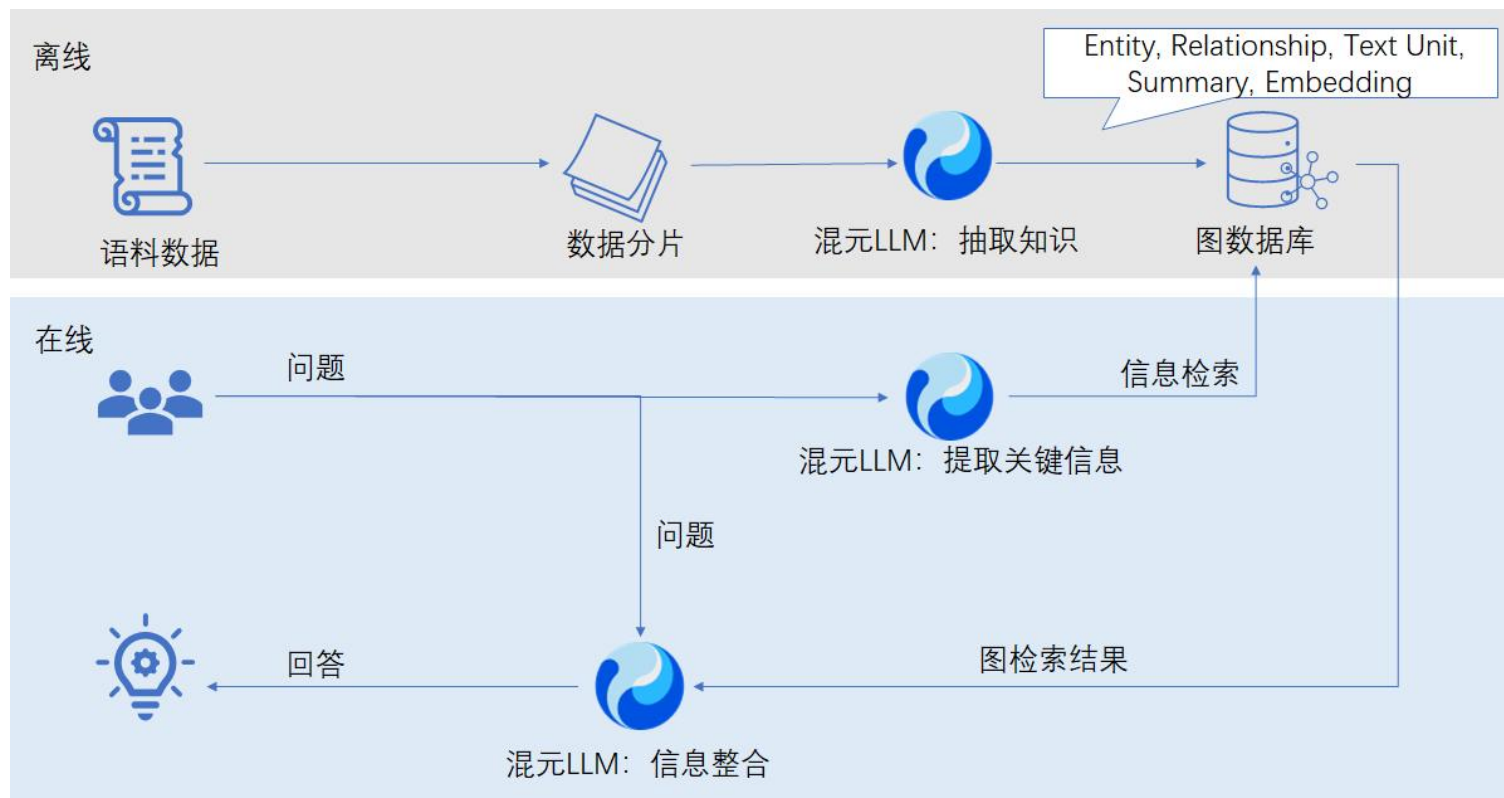
GraphRAG特点总结

	传统RAG	GraphRAG
数据结构和组织方式	非结构化文本	非结构化数据
组织方式	向量数据库	利用知识图谱，将非结构化数据组织成结构化的点和边 + 向量
检索方法	向量检索	利用图结构和语义聚类进行检索
上下文理解	有限	通过知识图谱提供更加丰富的上下文信息，使LLM更能理解复杂查询
信息整合能力	处理复杂、多方面信息使存在局限	更好的整合异构、相互关联的信息
推理能力	依赖文本匹配，较弱	利用知识图谱具备一定的推理能力
实现复杂度	容易	需要构建知识图谱，复杂度较高
可解释性	相对较弱	通过知识图谱提供良好的可解释性，方便追溯信息来源
适用场景	一般性问题和信息检索	需要全面理解大量数据集和复杂文档的场景

基于混元LLM构建图检索增强框架

基于混元LLM构建图检索增强框架

- 离线：通过离线训练充分学习语料知识并存储于图数据库
- 在线：对问题提取关键信息并向图数据库检索，结果经过混元大模型整合信息输出回答



GraphRAG在混元LLM落地框架

角色扮演场景介绍

大模型角色扮演：

基于大语言模型出色的语义理解能力和自然语言的交互方式，通过对角色的人设设定，让大语言模型在人设的合理基础上可以与用户对话交流、并完成用户的其它指定任务

场景特色：

- 需要具备出色的**长文本**理解，总结能力
- 存在多种角色设定，每种角色对话需要**符合角色特点**
- 对话存在**互动性**，**推理性**等

常见问题：

- 上下文限制
- 人称与角色不对应
- 角色关系识别不准确
- 角色回答不贴人设



易错点问题

人称问题

你我他的身份**认知错乱**和**幻觉**问题：
你是谁？你喜欢吃什么？
我是谁？我喜欢吃什么？
他是谁？（上文中没提到的他）

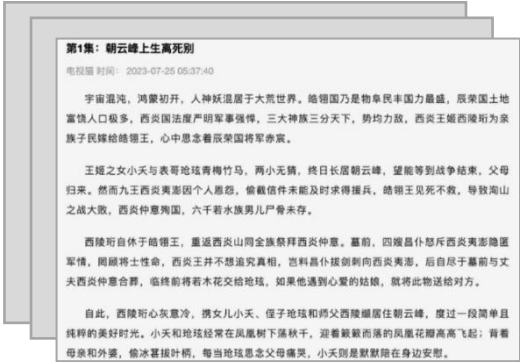
人物关系

角色之间的关系，容易陷入“**逆转诅咒**”（本质原因在于模型是概率推理），需遍历问法：
例子：小天和西陵珩是母女
对小天的问题：
西陵珩是谁？
你母亲是谁？
你是谁的女儿？
你和西陵珩什么关系？

长相思: G-Indexing



《长相思》剧本



文本分片



Prompt:

-目标-

从给定文本中识别出这些类型的所有实体以及已识别实体之间的所有关系。

-步骤-

1. 识别所有实体。对于每个已识别的实体，提取以下信息：

- 实体名称: 实体名称
- 实体类型: 以下类型之一: [人名、地点、组织、事件]
- 实体描述: 实体属性和活动的综合描述将每个实体

格式化为 ("**实体**"<|><**实体名称**><|><**实体类型**><|><**实体描述**>)

2. 从步骤 1 中识别的实体中，识别所有 *明显相关* 的 (源实体、目标实体) 对。对于每对相关实体，提取以下信息：

- source_entity: 源实体的名称，如步骤 1 中所述
- target_entity: 目标实体的名称，如步骤 1 中所述
- relationship_description: 源实体和目标实体相互关联的原因
- relationship_strength: 表示源实体和目标实体之间关系强度的数字分数，将每个关系格式化为

("relationship"<|><source_entity><|><target_entity><|><relationship_description><|><relationship_strength>)

3. 以中文形式返回输出，作为步骤 1 和 2 中确定的所有实体和关系的单个列表。使用 **###** 作为列表分隔符。

4. 完成后，输出 <|任务完成|>

prompt



混元大模型

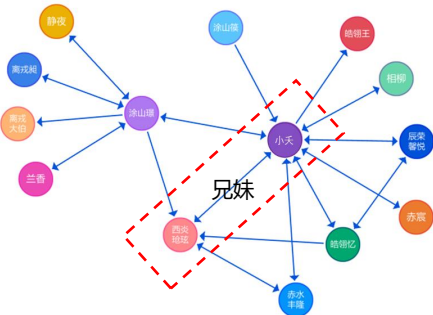
生成知识库

知识库

Entities

- **小天|人名**|小天是故事中的关键女性角色，她是西陵珩和赤宸的女儿。小天不仅擅长箭术和毒术，还是一位医术高超的女医师，拥有自己的医馆。.....
- **清水镇|地点**|.....
- **西陵氏|组织**|.....
- **五王之乱|事件**|.....

Relationships



关系: 小天与西陵珩是一对兄妹，他们之间有着亲密的血缘关系。.....

Community Reports

西炎珩珩和皓翎玖瑶之间的关系复杂而深刻。西炎珩珩对皓翎玖瑶表现出情感依赖，而皓翎玖瑶则寻求独立。尽管如此，两人之间共享了重要的时刻，讨论了个人感受和政治责任。.....

Embedding

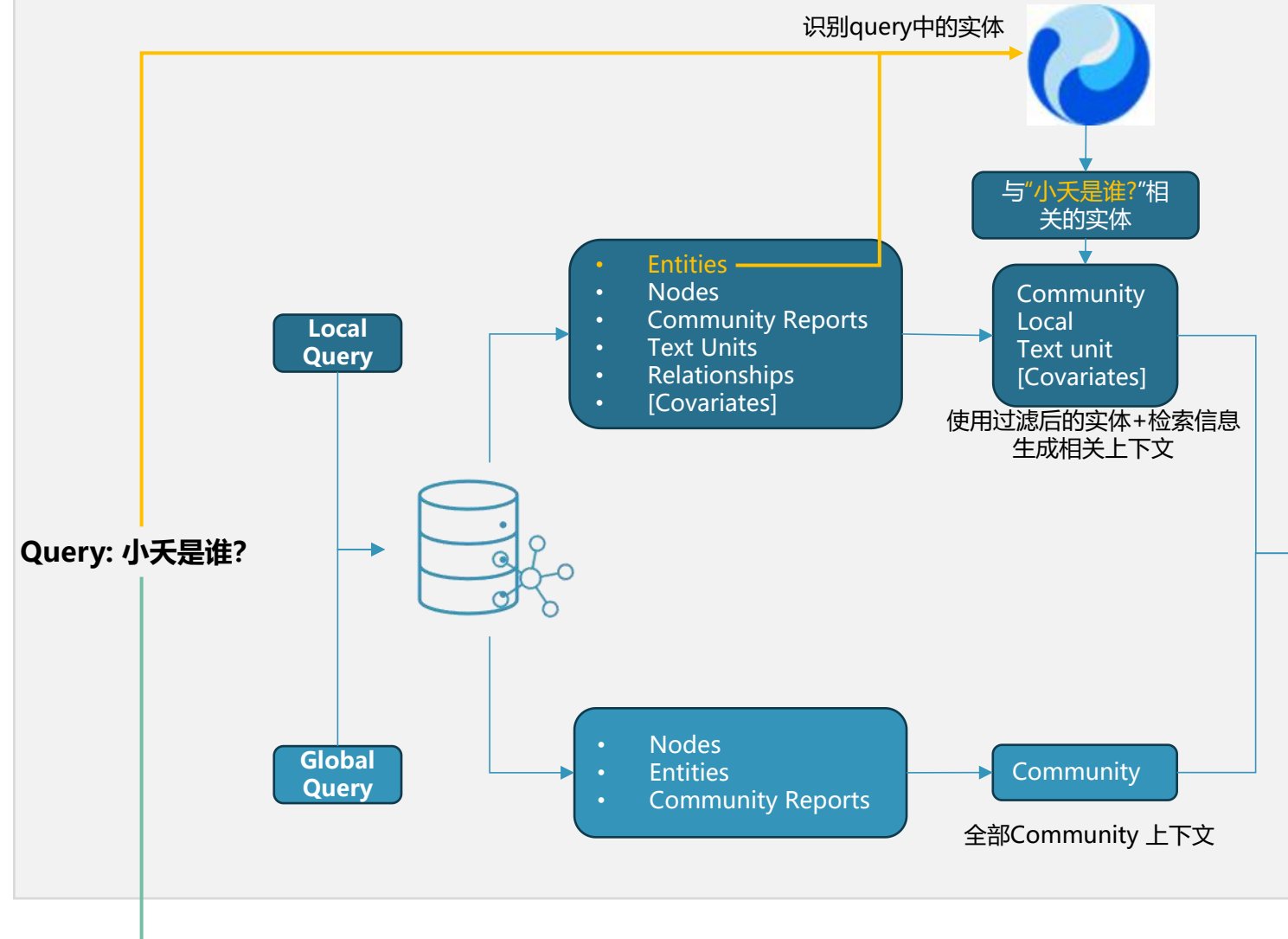
0.6 0.3 0.1 - - - -

0.8 0.5 0.3 - - - -

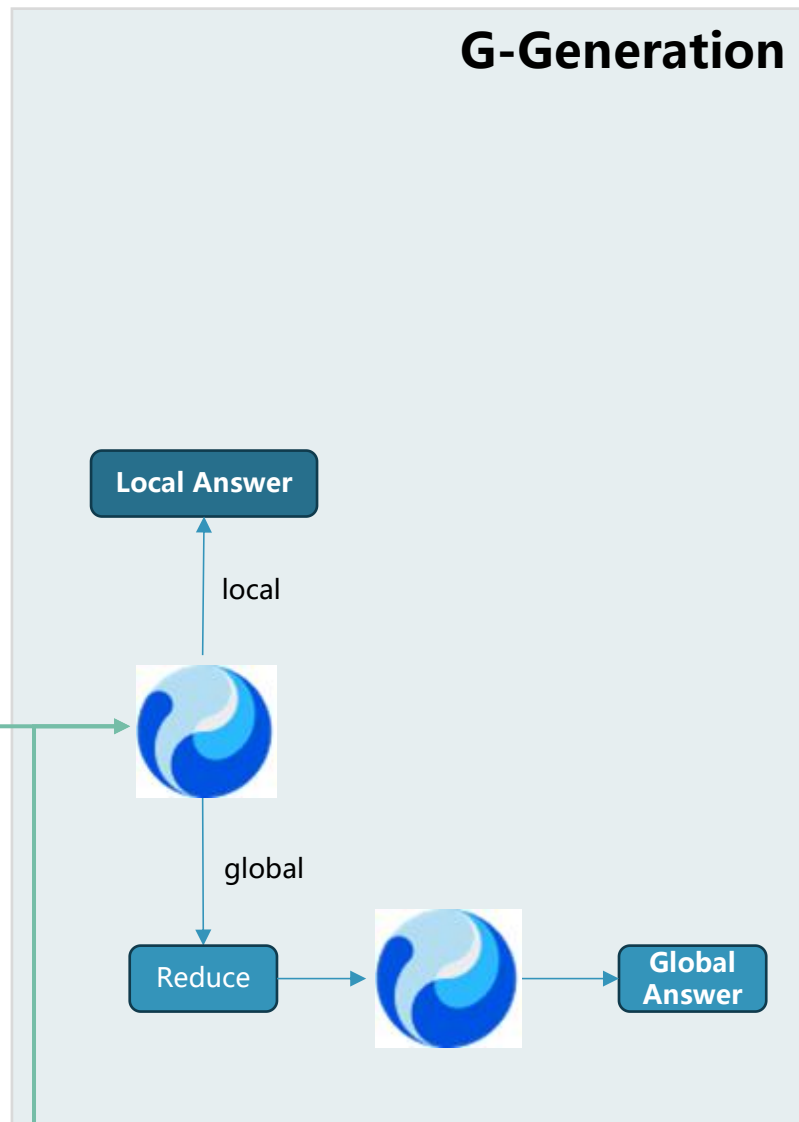
0.4 0.2 0.9 - - - -

长相思: G-Retrieval & G-Generation

G-Retrieval



G-Generation



GraphRAG-长相思角色扮演示例

你的母亲是谁？



讲一下你的故事



我的母亲是****西陵珩**** [Data: Entities (48)]



我是皓翎玖瑶，我的故事是一段跨越了身份与情感纠葛的旅程。

****身世之谜****

我，小天，实际上是西陵珩与赤宸之女，被王母抚养长大。我的身世秘密曾经被揭穿，从皓翎大王姬变成了西陵氏的大小姐。这一转变让我必须面对许多人的恨意，同时也学会了如何坚强地活下去 [Data: Entities (48)]。

****情感纠葛****

在我的生命中，涂山璟是一个重要的存在。尽管他已与防风意映订婚，但我对他的感情深刻而复杂。璟与我有着深厚的感情，他的离去成为了我生命中无法承受之痛 [Data: Entities (27, 48)]。此外，我与相柳之间也存在着复杂的感情纠葛，我们的关系特殊而微妙 [Data: Entities (48)]。

****家族与社会的期待****

作为皓翎大王姬，我被许多人所期待，包括我的妹妹阿念。她对我既嫉妒又羡慕，因为她认为如果拥有同样的身份，她会比我更强大。但即便如此，我始终追求的是自我价值的实现和真爱的寻觅 [Data: Relationships (49); Reports (0)]。

****自我成长与探索****

我不仅渴望安定生活，也不断追求个人的成长。我在神农山等待璟的消息，在轵邑的陋巷开设小医馆，用医者之心行医救人。我的生活充满了对未知的好奇和对自我能力的探索 [Data: Entities (48)]。

****结局的抉择****

面对爱情与家族的压力，我曾感到痛苦和迷茫。但我明白，无论身世如何变迁，我都必须坚持自己的选择，勇敢地面对生活的挑战 [Data: Entities (48)]。

我的故事是一段关于爱、勇气和自我发现的旅程，它教会了我即使面对再大的困难，也要坚持自己的内心。

4 第四章

Agent技术原理和应用

Agent应用场景



请根据天气情况建议下周深圳的旅行安排，预算在一万元以内。

基于天气情况，查询户外的旅游产品

THOUGHT

Reason about the next action to take

ACTION

Decide on a action to take

OBSERVATION

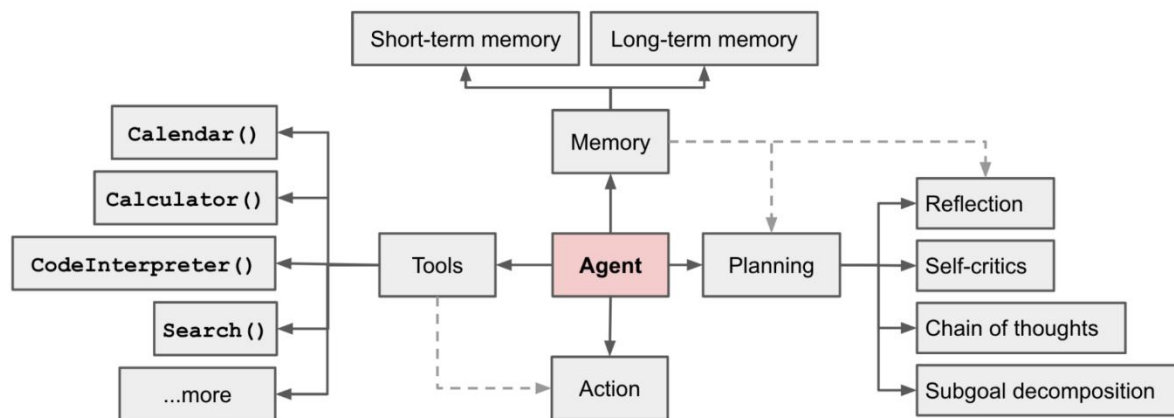
Execute action and feedback observation

执行旅游产品查询



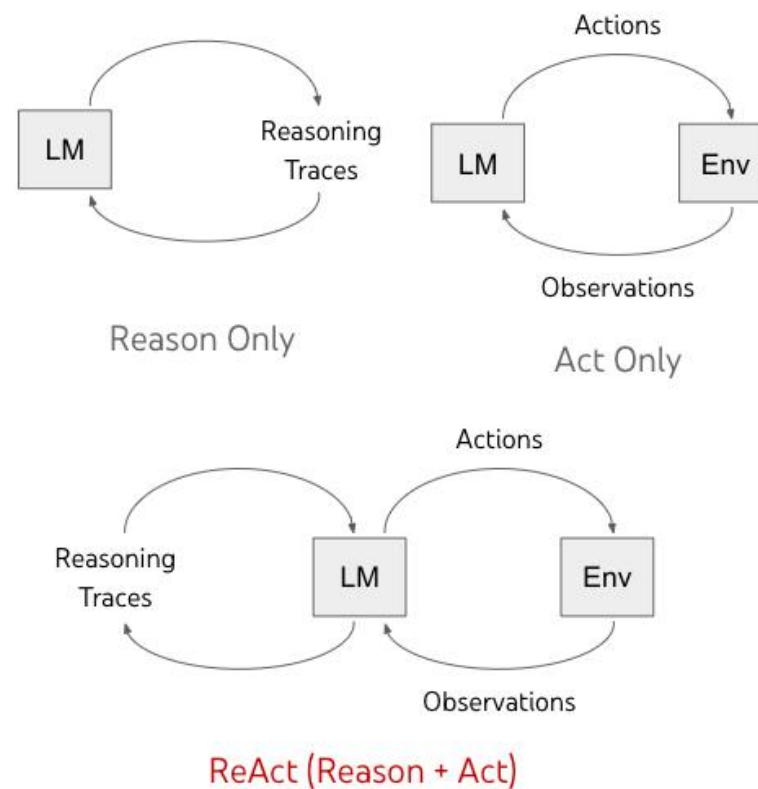
下周深圳天气晴朗，温度适宜，微风，适合户外活动

Agent技术原理

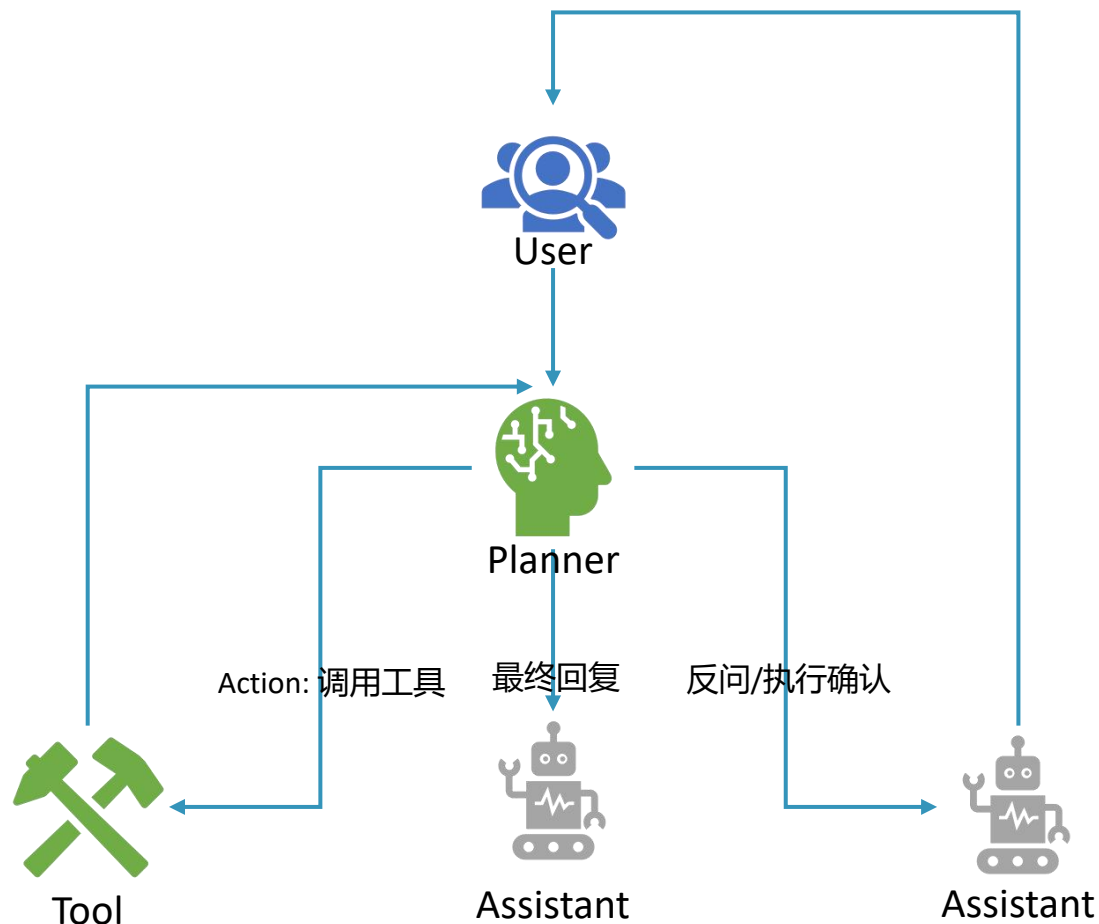


LLM agents are advanced AI systems designed for creating complex text that needs sequential reasoning. They can **think** ahead, **remember** past conversations, and **use** different **tools** to adjust their responses based on the situation and style needed.

<https://www.superannotate.com/blog/llm-agents>



Agent在混元中的实现



```
[系统信息]=""  
当前场景下有四个角色，分别是用户、Planner、Tool和AI Agent助手。  
用户：提出任务，或根据AI Agent助手的回复进行回答。输入格式参考[用户输入格式]。  
Planner：判断当前任务完成情况，并给出思考过程、逻辑分析、计划和要执行的动作。输出格式参考[Planner输出格式]。  
Tool：执行Planner输出的动作，返回执行结果。输出格式参考[Tool输入格式]。  
AI Agent助手：根据Planner输出的思考过程、逻辑分析、计划和要执行的动作决定如何回复。输出格式参考[AI Agent助手输出格式]。  
  
每一轮对话，你需要扮演Planner或AI Agent助手，遵循以下要求扮演不同的角色：  
如果上一轮角色为用户、Tool，则请你扮演Planner给出思考、分析、计划和要执行的动作，回复要符合AI Agent助手定义。  
如果上一轮角色为Planner，且输出的动作为READY_ANS、ASK_USER_FOR_HELP、ASK_USER_TO_CONFIRM_EXECUTION中的任意一个，则请你扮演AI Agent助手，回复要符合AI Agent助手定义。  
""  
  
[Planner输出格式]=""  
Planner:  
<Plan> 根据用户任务、AI Agent助手定义、对话历史和当前环境，给出计划方案</Plan>  
<Thought> 解决用户任务时进行的内部思考过程和逻辑分析</Thought>  
<Action> 参考[Action类型]，给出要执行的动作</Action>  
""  
  
[Action类型]=""  
情况1：如果你认为用户任务需要使用[工具列表]中的一个工具来完成，则给出工具名称，并根据环境信息填写输入参数，选择的工具都存储在一个list[]中，  
每个工具使用json格式，包含name、arguments两个字段，分别填写工具名称和输入参数。  
情况2：如果你认为已经可以完成用户任务了，则输出READY_ANS。  
情况3：如果你认为需要寻求用户帮助，则输出ASK_USER_FOR_HELP。  
情况4：如果你认为需要让用户确认是否执行工具，则输出ASK_USER_TO_CONFIRM_EXECUTION。  
""  
  
[Tool输入格式]=""  
Tool:  
<Observation> 工具的执行结果</Observation>  
""
```

Agent应用案例





面向生态合作伙伴交付运维的支撑平台，包括交付作业流程/线下项目变更等场景



当前任务的下一步计划是什么？



 FunctionCall 

插件名称	插件描述
获取工作群聊天记录	通过任务单ID，可以获取该任务对应的工作群的讨论内容，以此发现可能的风险、进展、问题等信息。注意要不要让时间范围过小，推荐至少一周作为时间范围。
获取任务详情	通过任务单ID获取任务单的详细信息，包括订单信息、项目信息、单据信息、人员信息等
获取任务日志	获取人工上传的不同类型的任务日志，这些日志包括进展、风险、计划、日志时间等信息。

企业 / 任务

迁移

任务进展

Copilot

访问记录



2024-09-26 18:47:52

您好！我是安兔Copilot，您身边的任务百事通，请问现在有什么可以帮助您的吗？

请写一份关于当前任务的周报

当前任务的基本情况是什么？

当前任务进展怎么样了？

当前任务是否有什么风险？

当前任务下一步计划是什么？

团队在任务里做了哪些事？



新开会话

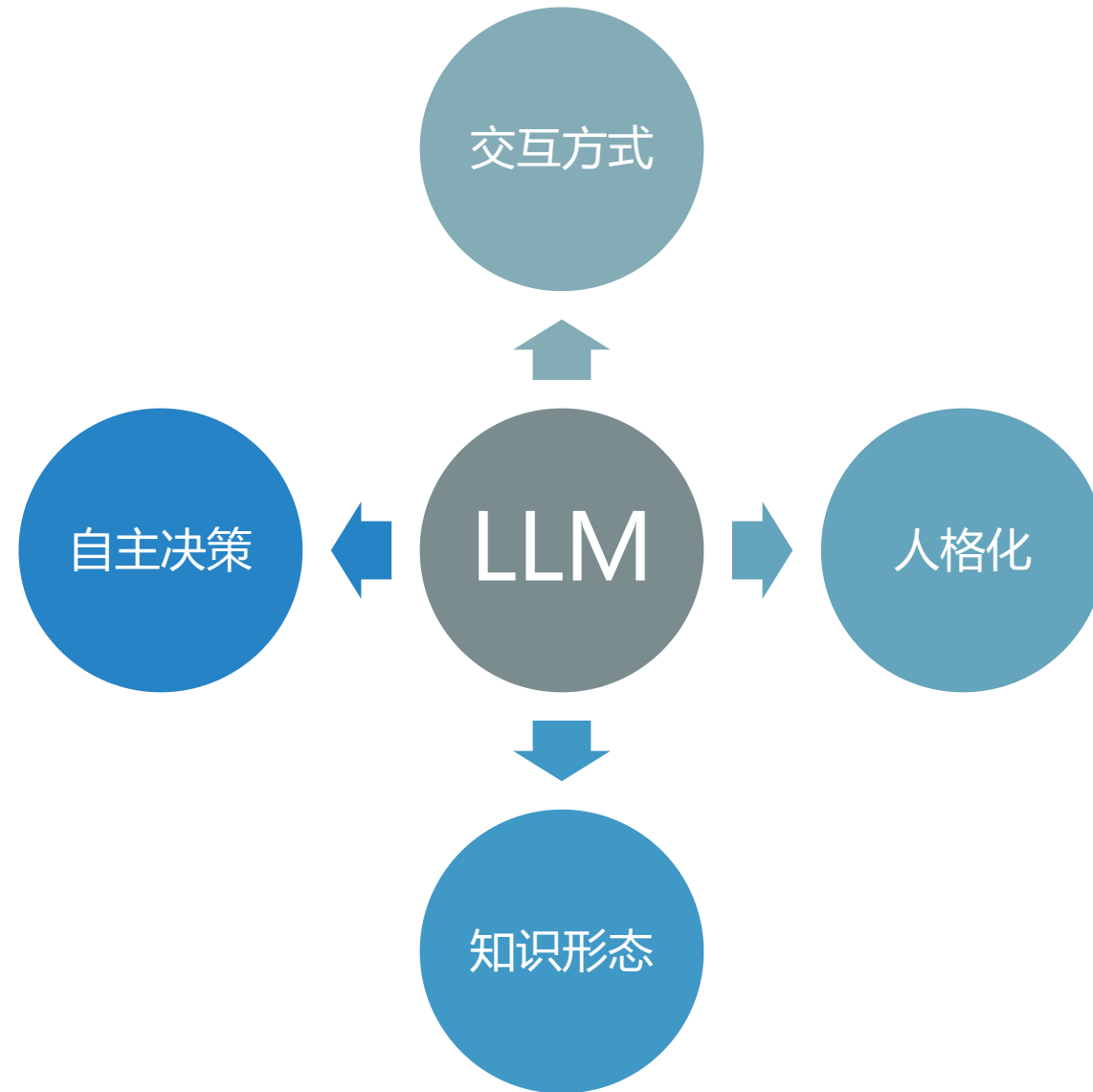


常见问题

这个任务您有什么想了解的，快来问我~



总结



THANKS