大模型在智能问答领域的技术实践

柴春燕 / 字节跳动

DataFunSummit # 2024

- 智能问答场景痛点
- 大模型技术方案选型
- 技术架构设计
- 技术挑战与优化策略

01

智能问答场景痛点

智能问答场景痛点

1 理解能力局限

难以准确理解用户的复杂查询或隐含意图，导致回答不准确或不相关。

2 上下文维持困难

在多轮对话中，系统难以维持对话上下文，导致回答可能与用户之前的问题或陈述脱节。

3 个性化服务不足

缺乏根据用户历史行为或偏好提供个性化回答的能力。

4 用户意图识别不准确

难以准确识别用户的真正意图，尤其是在用户表达模糊或多义的情况下。

5 交互自然性不足

缺乏自然流畅性，难以模仿人类对话的自然节奏和风格。

6 多任务处理能力弱

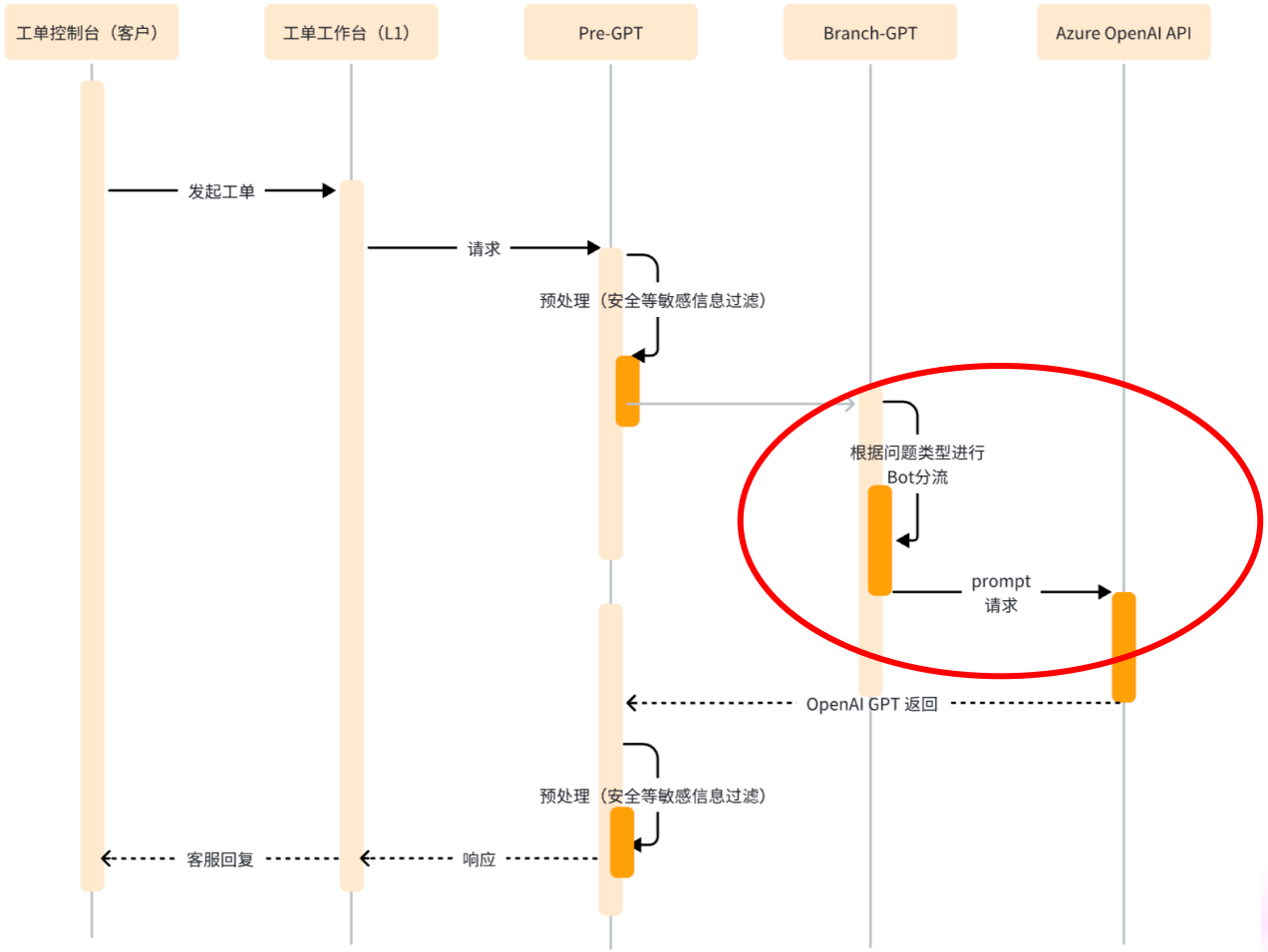
同时处理多个任务或问题时，可能会降低回答的准确性和效率。

02

大模型技术方案选型

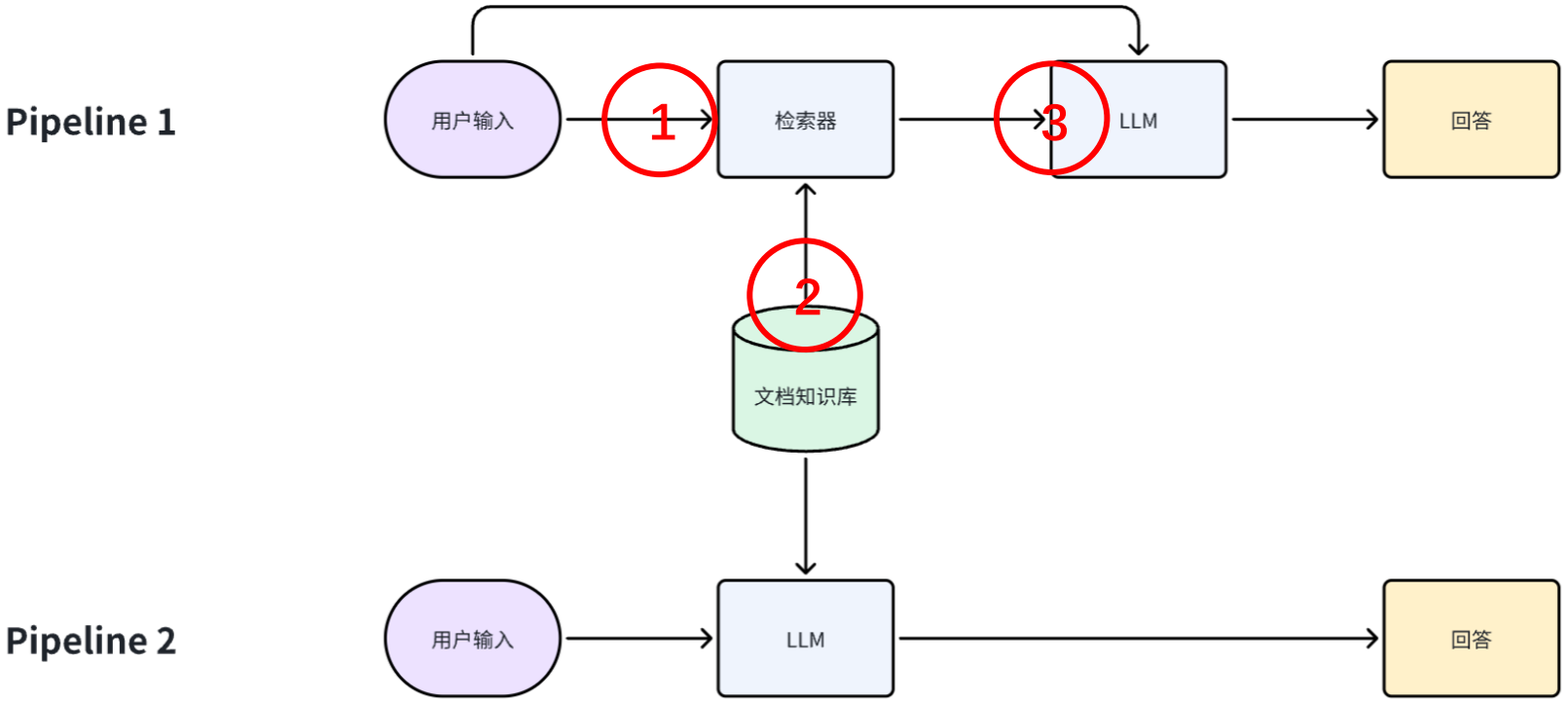
方案一：构建 Prompt 生成问题回复

通过构造 prompt 模板，进行用户意图识别/问题分类（功能模块准入/准出）



方案二：RAG（检索增强生成）

通过从外部知识源动态检索信息，并使用检索到的数据作为组织答案的参考，显著提高响应的准确性和相关性，有效解决LLMs中存在的幻觉问题



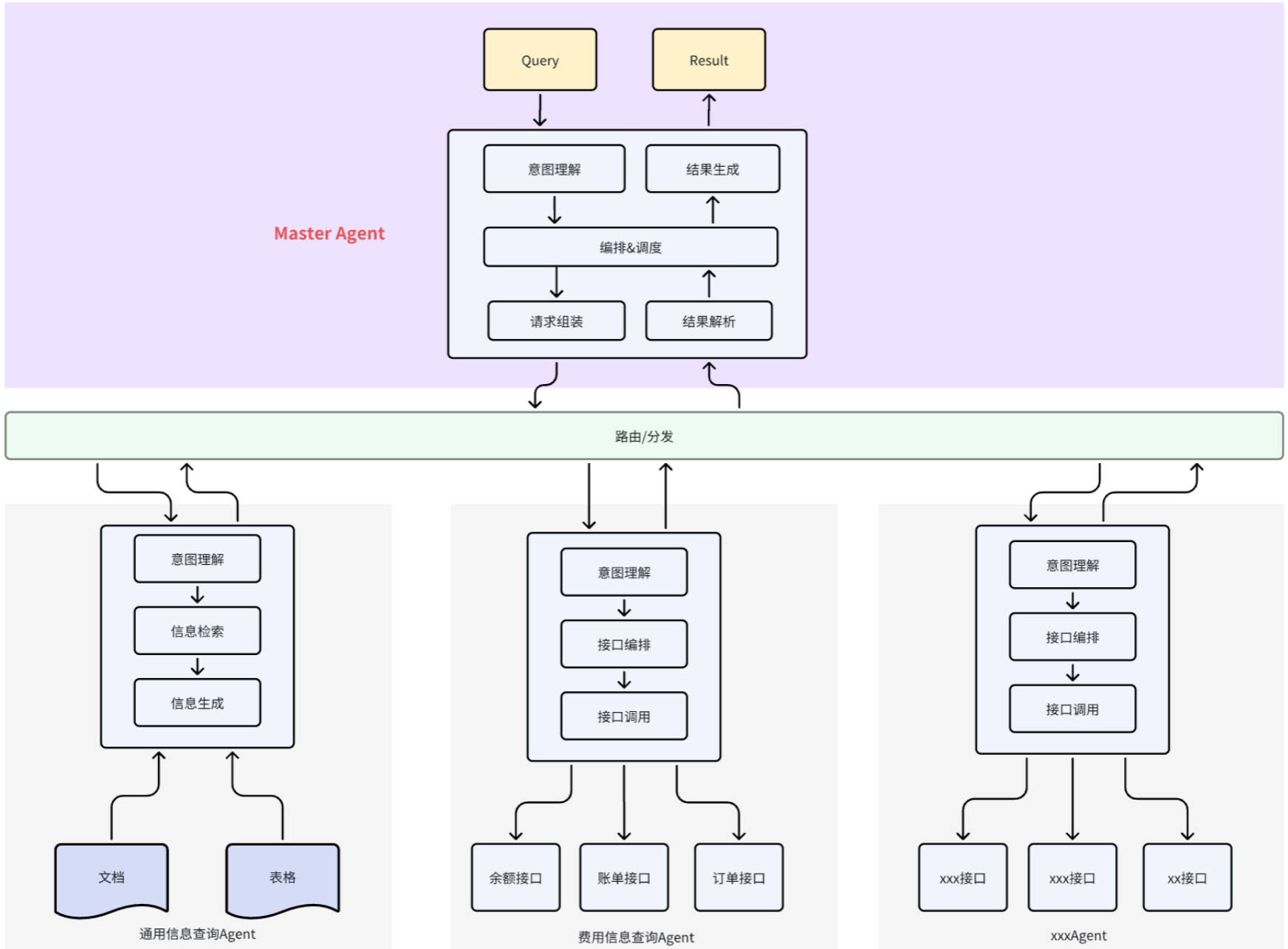
03

技术架构设计

Multiple-Agent + Master&Slave架构

- Master Agent：理解用户的意图，根据用户的意图去编排&调度Slave Agent
- Slave Agent：提供特定的能力，供Master调用。Slave Agent内部可以是单完成

Master-Slave架构


Query
Result
Master Agent
意图理解
结果生成
编排&调度
请求组装
结果解析
路由/分发
意图理解
信息检索
信息生成
文档
表格
通用信息查询Agent
意图理解
接口编排
接口调用
余额接口
账单接口
订单接口
费用信息查询Agent
意图理解
接口编排
接口调用
xxx接口
xxx接口
xx接口
xxxAgent

04

技术挑战与优化策略

RAG（Retrieval-Augmented Generation）检索增强生成

RAG 工作原理：

- **检索阶段**：当模型接收到一个查询（如问题或提示）时，首先在一个大型的文档集合中检索相关的文档或信息片段。这些文档通常存储在一个专门的数据库中，如 Wikipedia 或其他专业知识库。检索是基于相似性度量完成的，目的是找到与查询最相关的信息。
- **生成阶段**：模型使用检索到的文档作为额外的上下文信息来生成回答或内容。这一步通常是由一个序列到序列（seq2seq）模型完成的，如变换器（Transformer）模型。

RAG 方案优势：

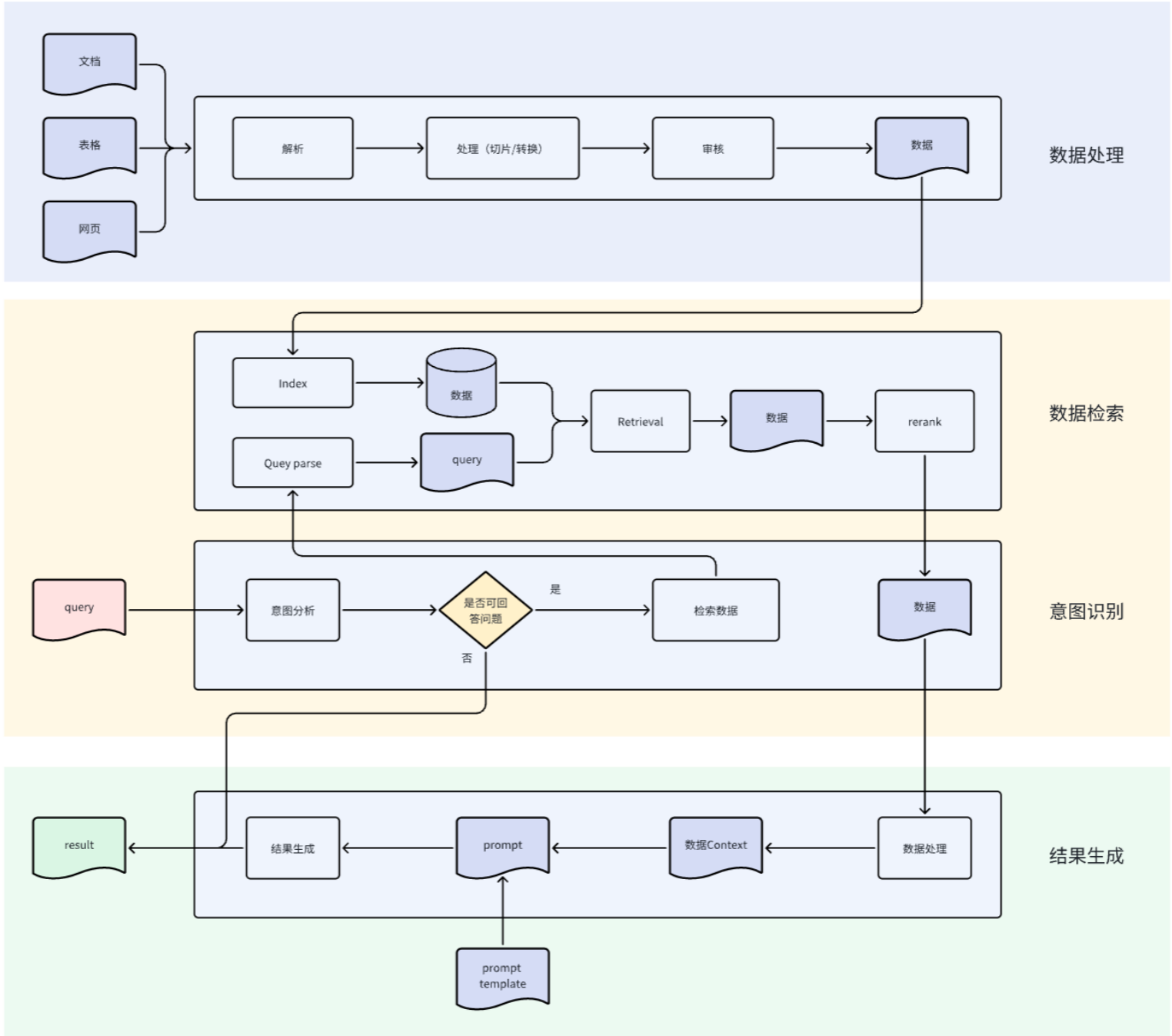
1）提高准确性和相关性：通过将检索到的信息纳入生成过程，RAG 能够生成更准确、更相关的回答。

2）知识更新：与传统的预训练语言模型不同，RAG 可以通过更新其检索数据库来接入最新的信息，而无需重新训练整个模型。

RAG 方案核心：

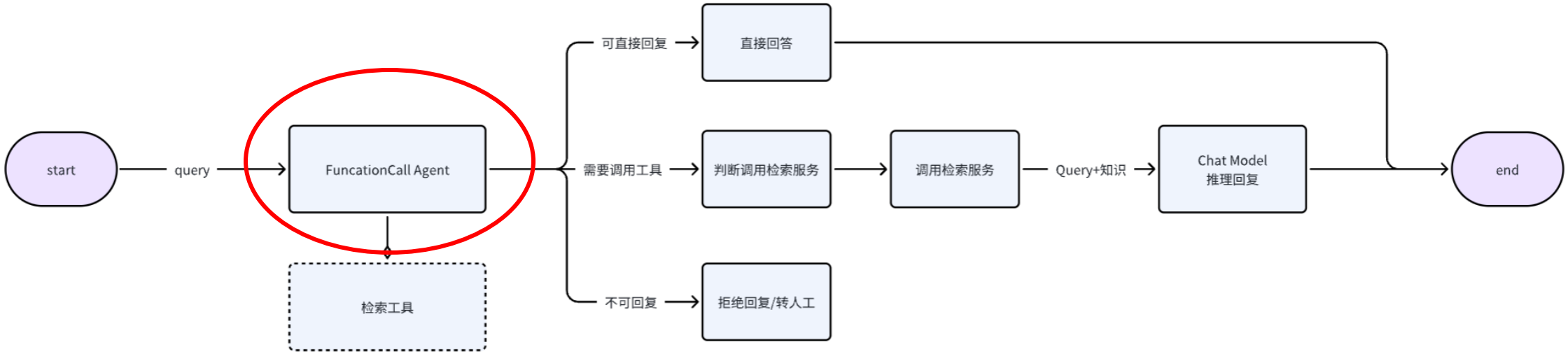
1）检索阶段做到比较高的 topk 召回率

2）LLM 支持比较大的 context window， 并能从较多相关信息中总结出正确答案


文档
表格
网页
解析
处理（切片/转换）
审核
数据
数据处理
Index
数据
Quey parse
query
Retrieval
数据
rerank
数据检索
query
意图分析
是否可回答问题
是
否
检索数据
数据
意图识别
result
结果生成
prompt
数据Context
数据处理
结果生成
prompt template

回答能力增强

使用functioncall的逻辑，将检索服务当作大模型的一个tool，让大模型自己判断是否该调用检索



service能力增强

采用基于规则和统计的两路service识别。

基于规则：

针对每个service，人工总结了若干keywords，当用户问题包含某个keywords，就将其对应的service加入到service_list中

```
{
    "service": "备案",
    "service_code": "beian",
    "keywords": [
        "备案",
        "ICP"
    ]
},
{
    "service": "费用中心",
    "service_code": "billing",
    "keywords": [
        "账单",
        "费用",
        "明细数据",
        "明细",
        "账单明细",
        "退订",
        "资源包",
        "合同",
        "发票",
        "计费",
        "退款",
        "汇款",
        "到账"
    ]
},{
```

service能力增强

基于统计：

faq看作产品特征数据集，拿query召回10条faq，归并10条faq中出现的service，若有某个service出现的次数超过特定阈值，我们就认为该query与该service是相关的，进而将该service加入到service_list中

```
1 {'ID': '345', 'Uint64ID': 297795584092, 'ServiceCode': 'beian', 'Question': '已收到接入备案初审通过短信，请问短信中所提示的备案负责人是主体负责人还是网站负责人啊？', 'Answer': '以原服务商备案成功的信息为基准，若在火山备案时，填写的主体信息发生变更，主体负责人需要完成核验；若网站信息发生变更，服务负责人需要完成核验--没有回答问题', 'Tags': None, 'Score': 0.7633594870567322}]
```

```
1 {'ID': '248', 'Uint64ID': 244779581532, 'ServiceCode': 'beian', 'Question': '域名已备案成功，若备案服务器不再使用，域名是否可正常解析？', 'Answer': '若在火山备案的服务器不再使用，可以指向其他的火山云资源，但若监管部门核查备案IP与实际解析IP不一致，则需提交备案变更IP或将域名指向备案IP    ----答案需修改下', 'Tags': None, 'Score': 0.8042183518409729}
```

```
1 {'ID': '524', 'Uint64ID': 397032816732, 'ServiceCode': 'beian', 'Question': '备案需要准备哪些资料？', 'Answer': '请结合您的备案性质（个人或企业）、备案省份以及备案类型，参考准备备案资料的内容，准备对应资料。 ----内容需完善，可贴链接', 'Tags': None, 'Score': 0.7477016448974609},
```

THANKS

DataFunSummit # 2024

DataFun.