B站基于LLM的大数据智能体实践

郭跃鹏 /大数据架构师

张巍 /计算平台Leader

DataFunSummit # 2024

- 背景介绍
- 架构介绍
- 技术落地
- 挑战展望

01

背景介绍

B站大数据架构介绍

平台架构

1. 计算平台
2. 湖仓一体
3. 存储集成
4. 资源调度
5. 平台工程

计算集群

任务量

- 27万/日 离线任务
- 2万/日 Kyuubi 任务
- 7千/日 实时任务

咨询量

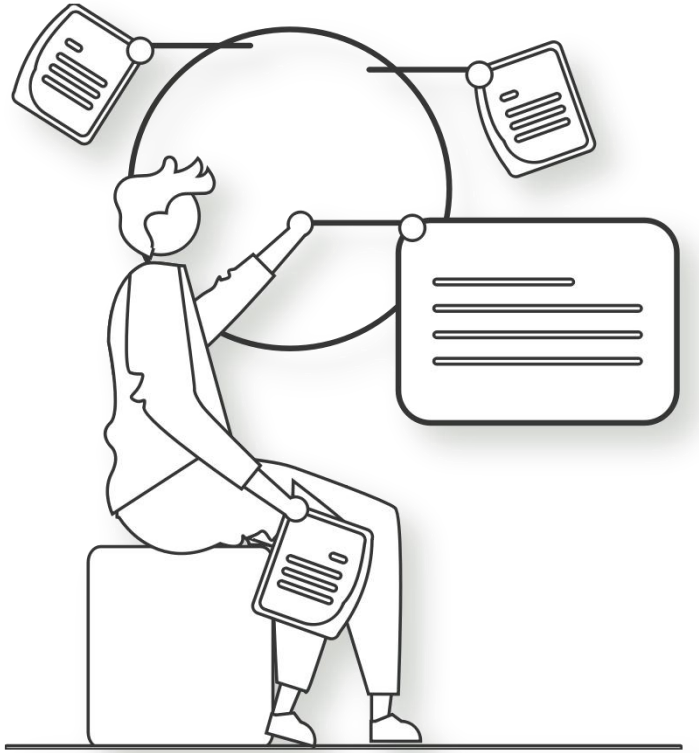
- 上千条/周
- 3人天/周/方向

离线计算引擎

- 为什么任务失败
- 为什么任务变慢

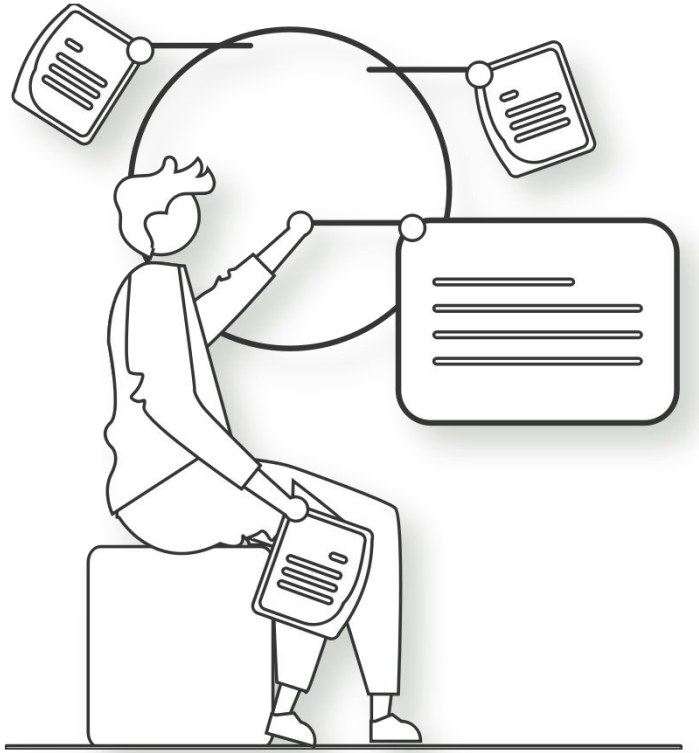
任务失败

- 系统内核缺陷
- 依赖组件问题
- 数据质量问题



任务变慢

- 硬件老化引起
- 资源调度相关
- 数据分布问题



用户提问的习惯

- 1. 抛链接+问题

- 2. 截图和链接+问题

智能小助手

- 私域咨询问题
- 诊断问题
- 时光机
- 其他Agent

02

架构介绍

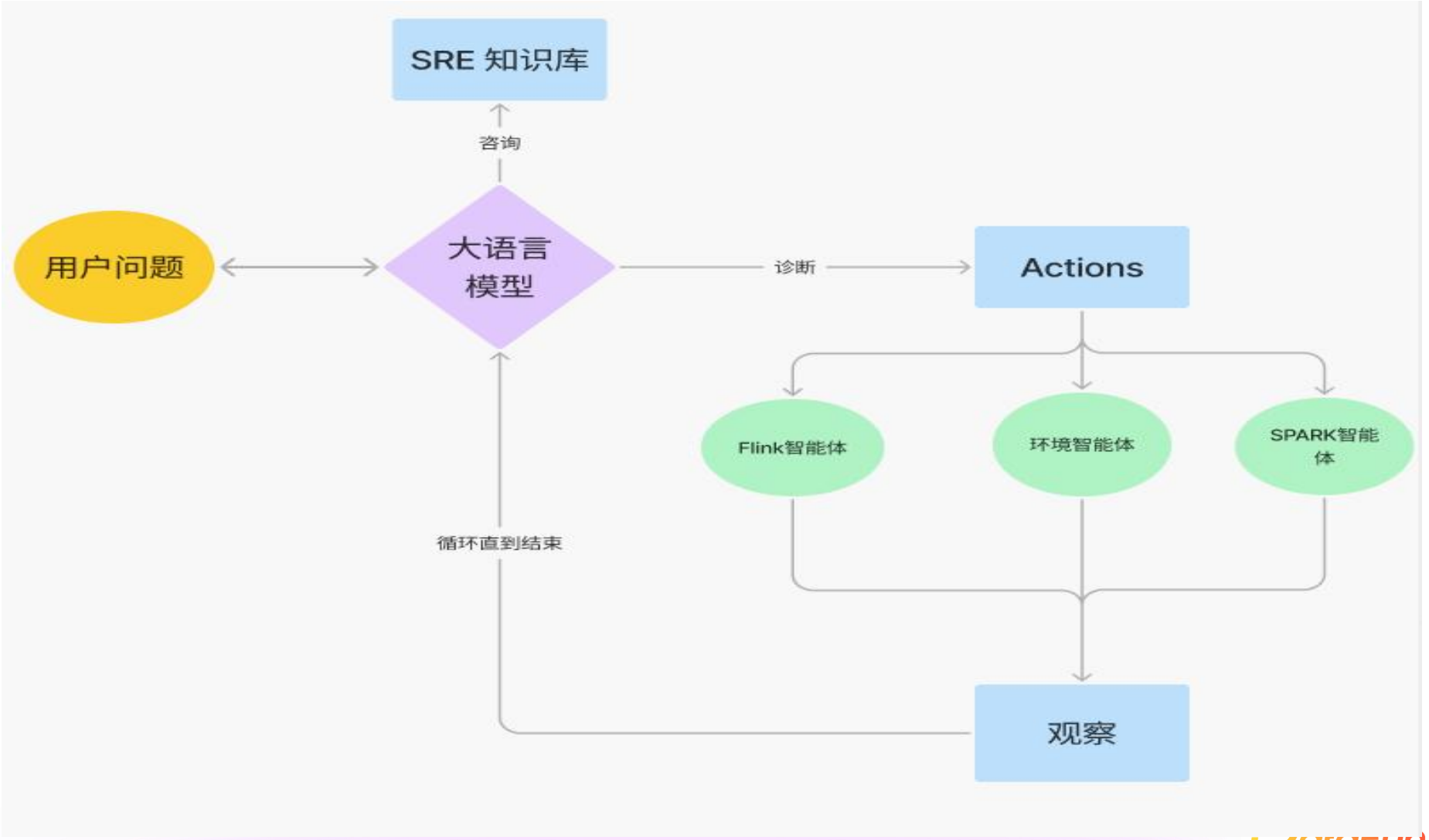
架构与挑战

架构介绍

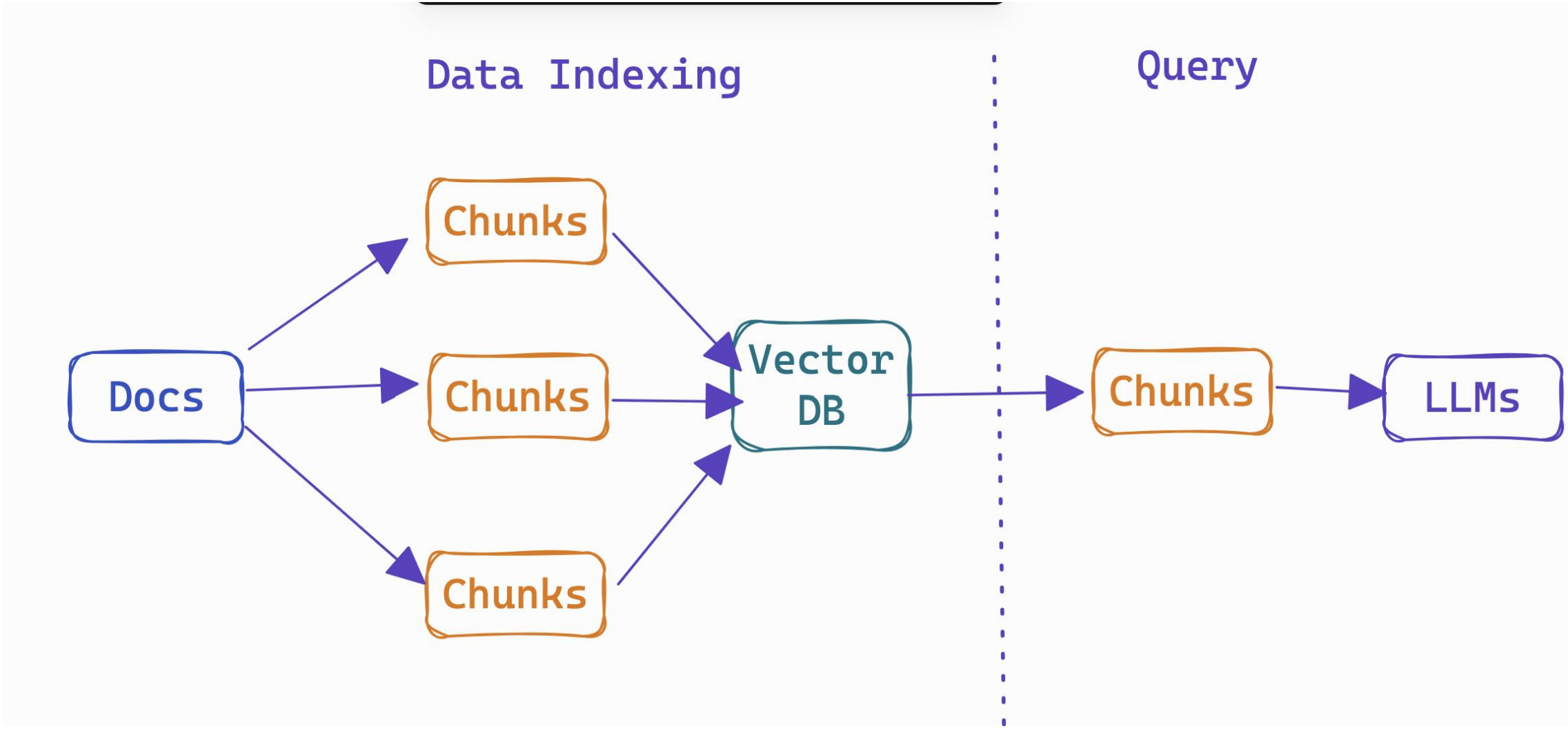
-

挑战

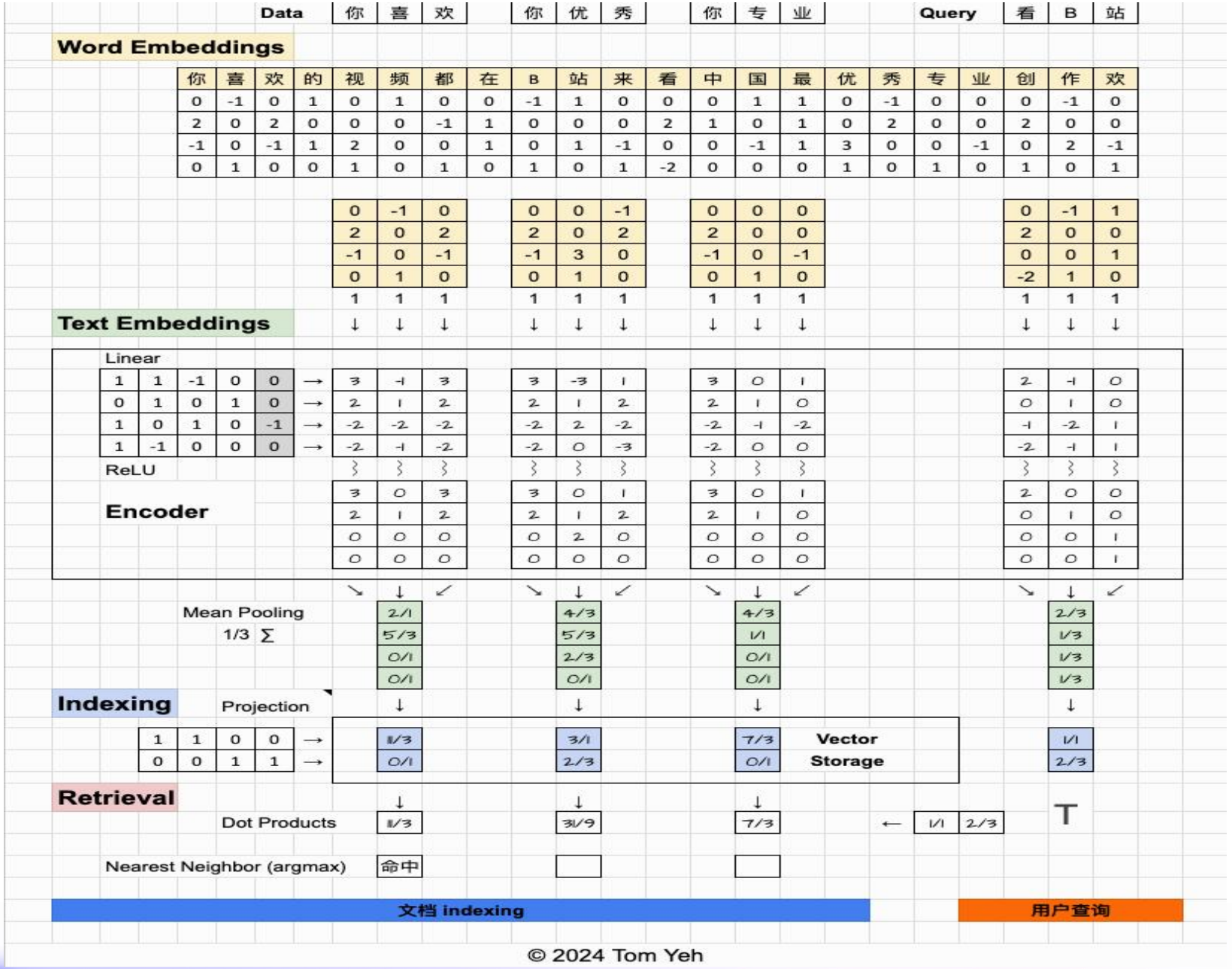
- precision问题
- recall问题



知识库

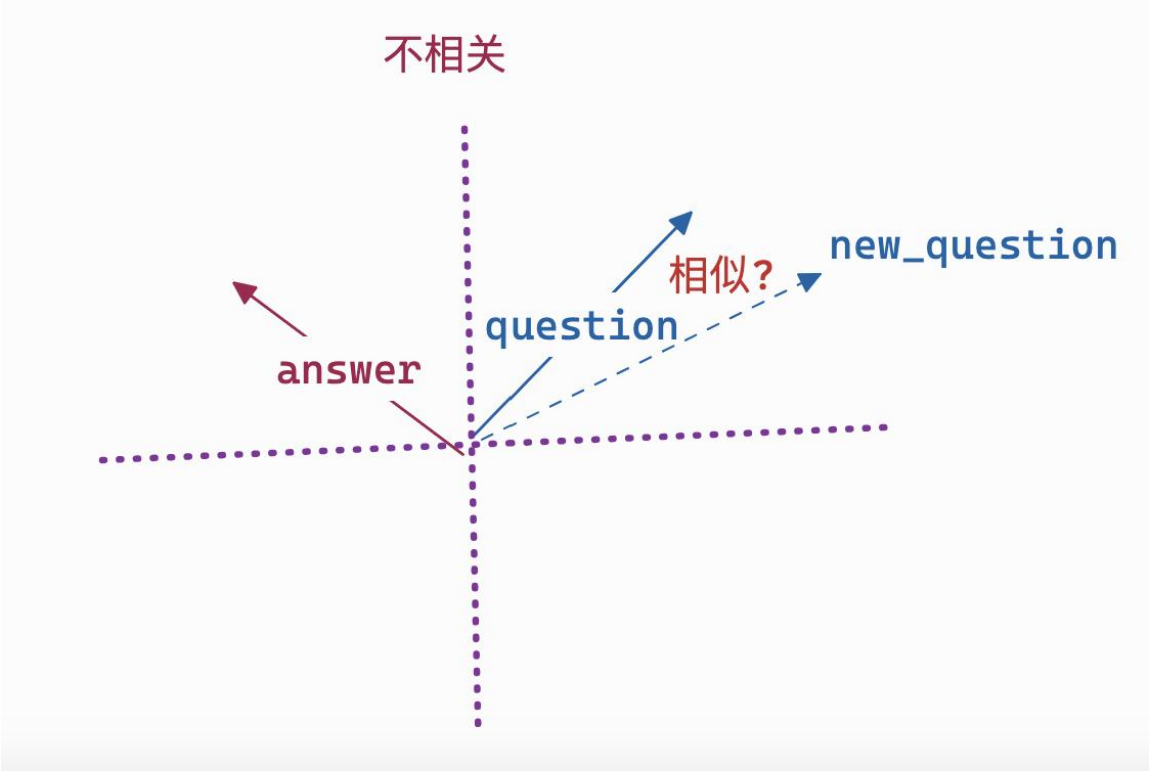

Data Indexing
Query
Docs
Chunks
Chunks
Chunks
Vector DB
Chunks
LLMs

Retrieval基础



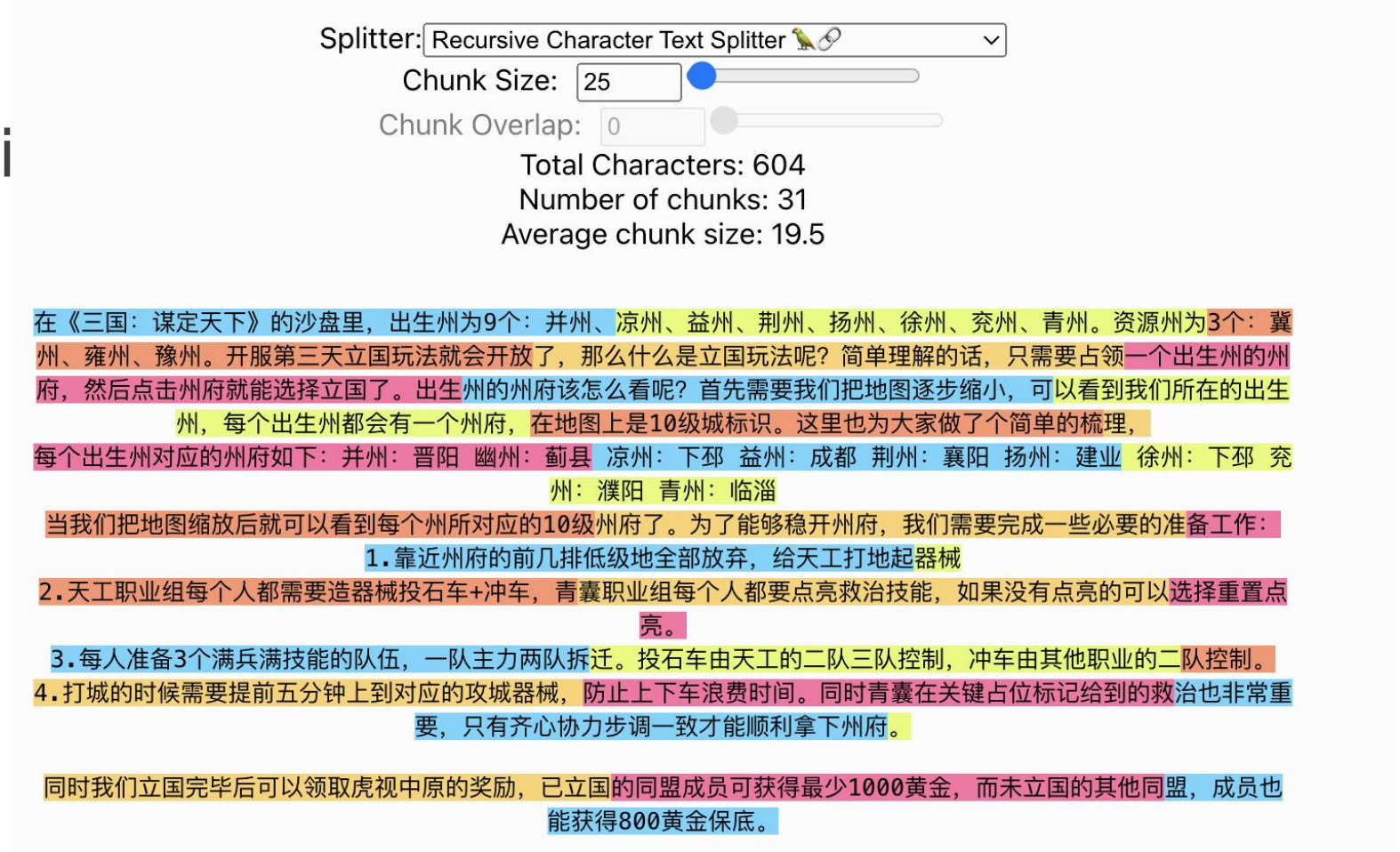
Precision问题：建议只embedding 问题

不要将问题和答案一起embedding

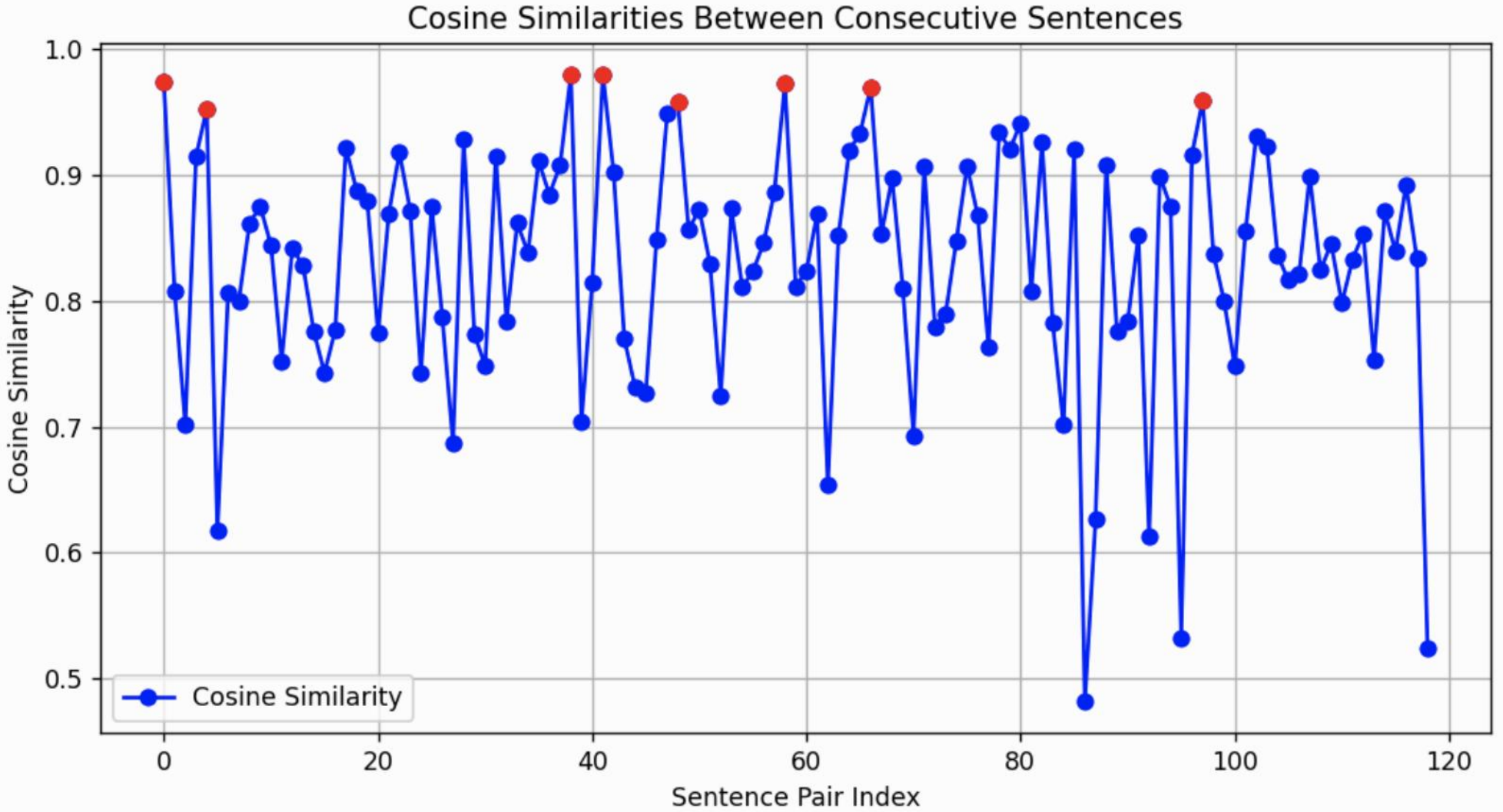


Precision问题: 建议基于语义做chunk

- 字符级别的 splitting
- 递归字符级别的 splitti
- 句子级别 splitting
- 语义级别 splitting
- Agentic splitting



Precision问题：建议基于语义做chunk



Recall问题: 建议对知识库做元数据过滤

Metadata Filtering

```
{
    "id":     ',
    "engine": "SPARK",
    "component": "HMS",
    "sub_component": "CONSTRAINT",
    "oncall": "离线小助手",
    "question": " Internal error processing add_partitions_req",
    "unique_keywords": "add_partitions_req",
    "answer": "hive metastore 请求异常，请联系离线小助手",
    "explanation": ""
},
```

Recall问题：建议top-k之后rerank

Rerank & Filter

- LLMRerank
- CohereRerank
- bge-reranker-v2-m3

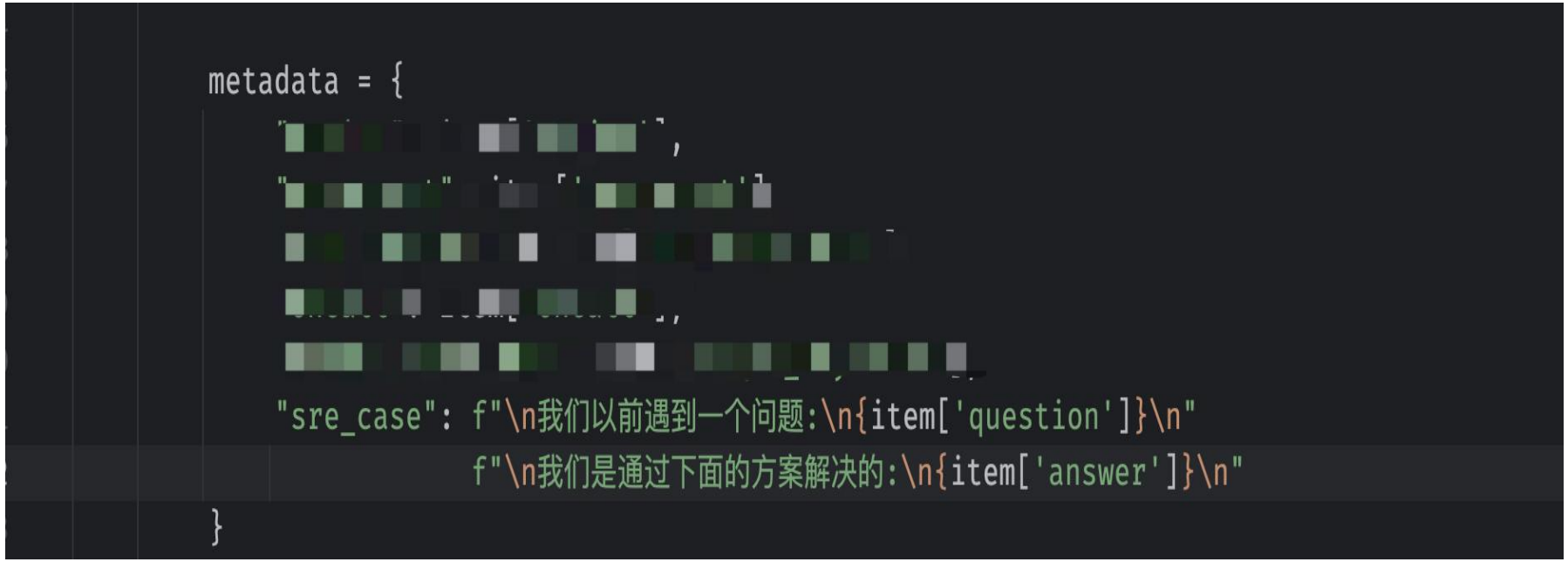
```python
def _postprocess_nodes(
        self,
        nodes: List[NodeWithScore],
        query_bundle: Optional[QueryBundle] = None,
) -> List[NodeWithScore]:
    """Postprocess nodes."""

    new_nodes = [node for node in nodes if node.score > 0]

    return new_nodes
```

提示词

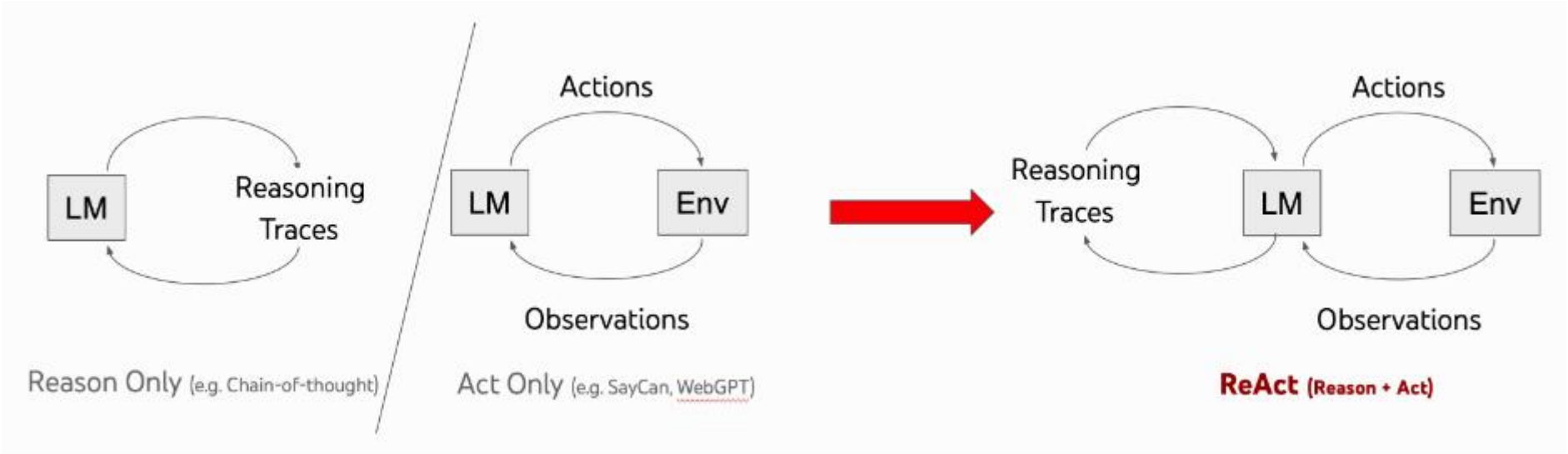
https://platform.openai.com/docs/guides/prompt-engineering



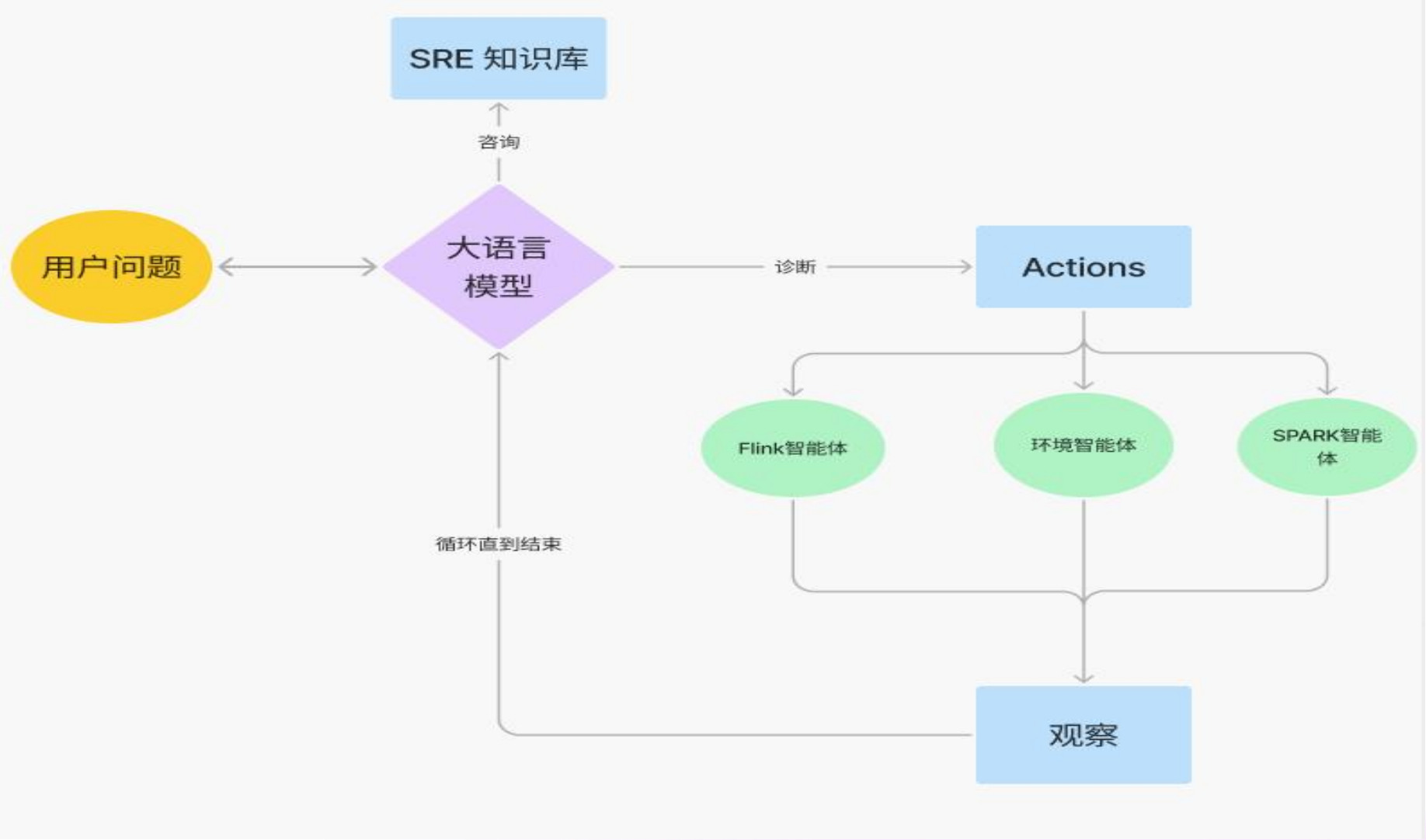
2024 Agent 原生应用之元年

- 自动化企业的workflow
- agent模式
  + llm as router
  + state management
  + agents

ReAct


LM
Reasoning
Traces
Reason Only (e.g. Chain-of-thought)
Actions
LM
Env
Observations
Act Only (e.g. SayCan, WebGPT)
Reasoning
Traces
Actions
LM
Env
Observations
ReAct (Reason + Act)

ReAct

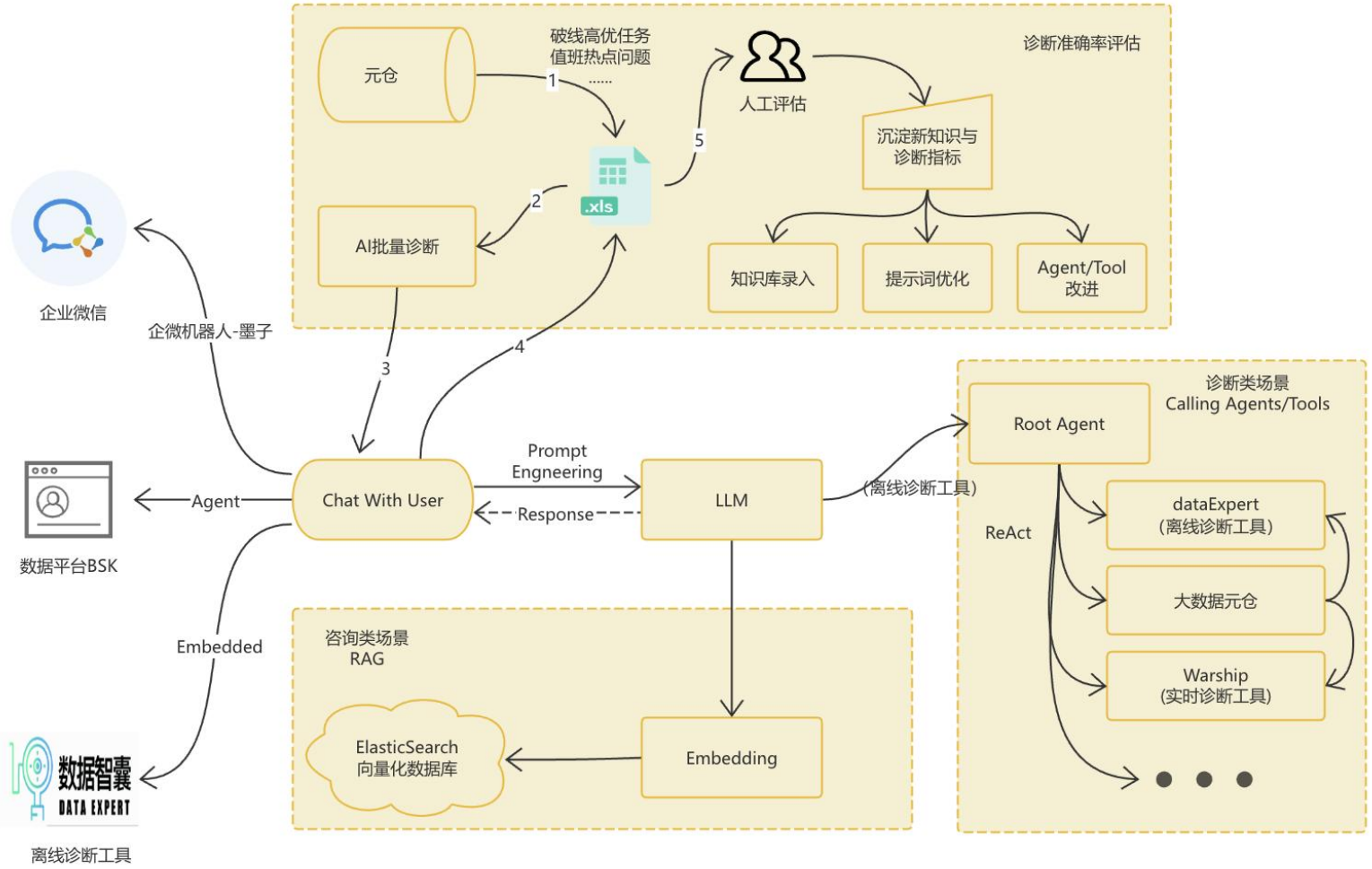

SRE 知识库
咨询
用户问题
大语言模型
诊断
Actions
Flink智能体
环境智能体
SPARK智能体
循环直到结束
观察

03

技术落地

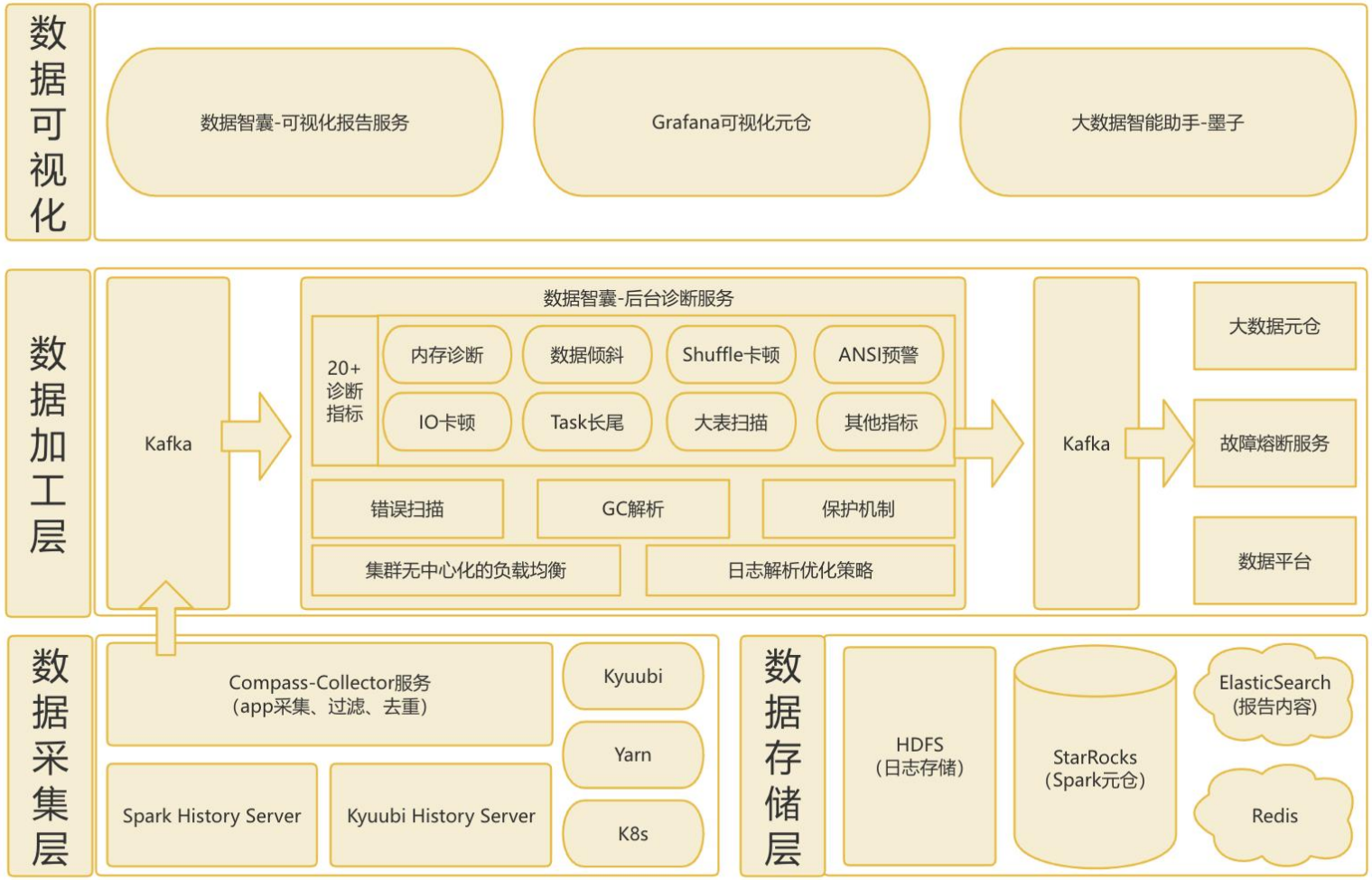
架构设计

- 元仓
- 诊断
- Rag
- 评估



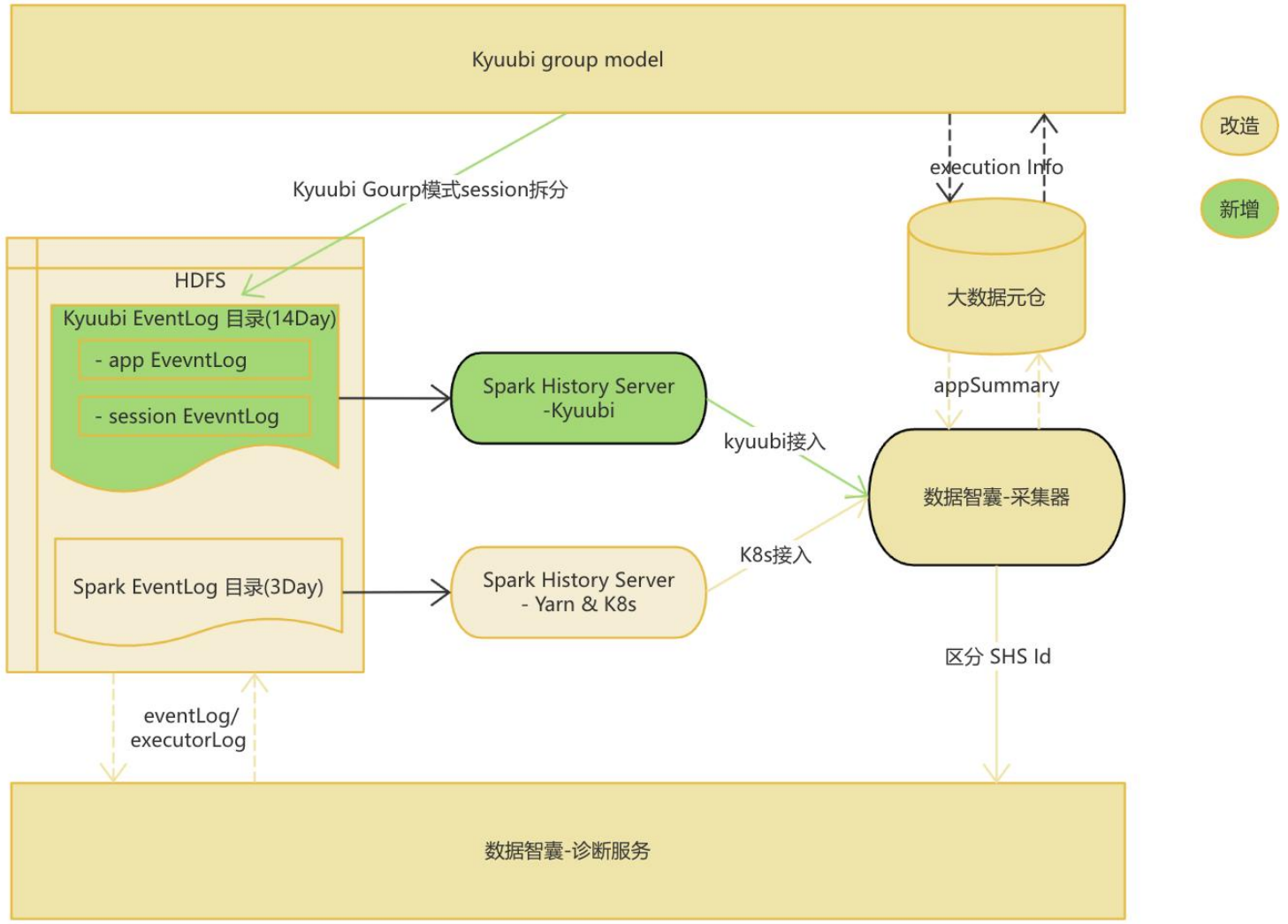
离线诊断的架构设计

- 27万/天诊断任务
- 3千万日志/天
- Kyuubi/K8s/Yarn



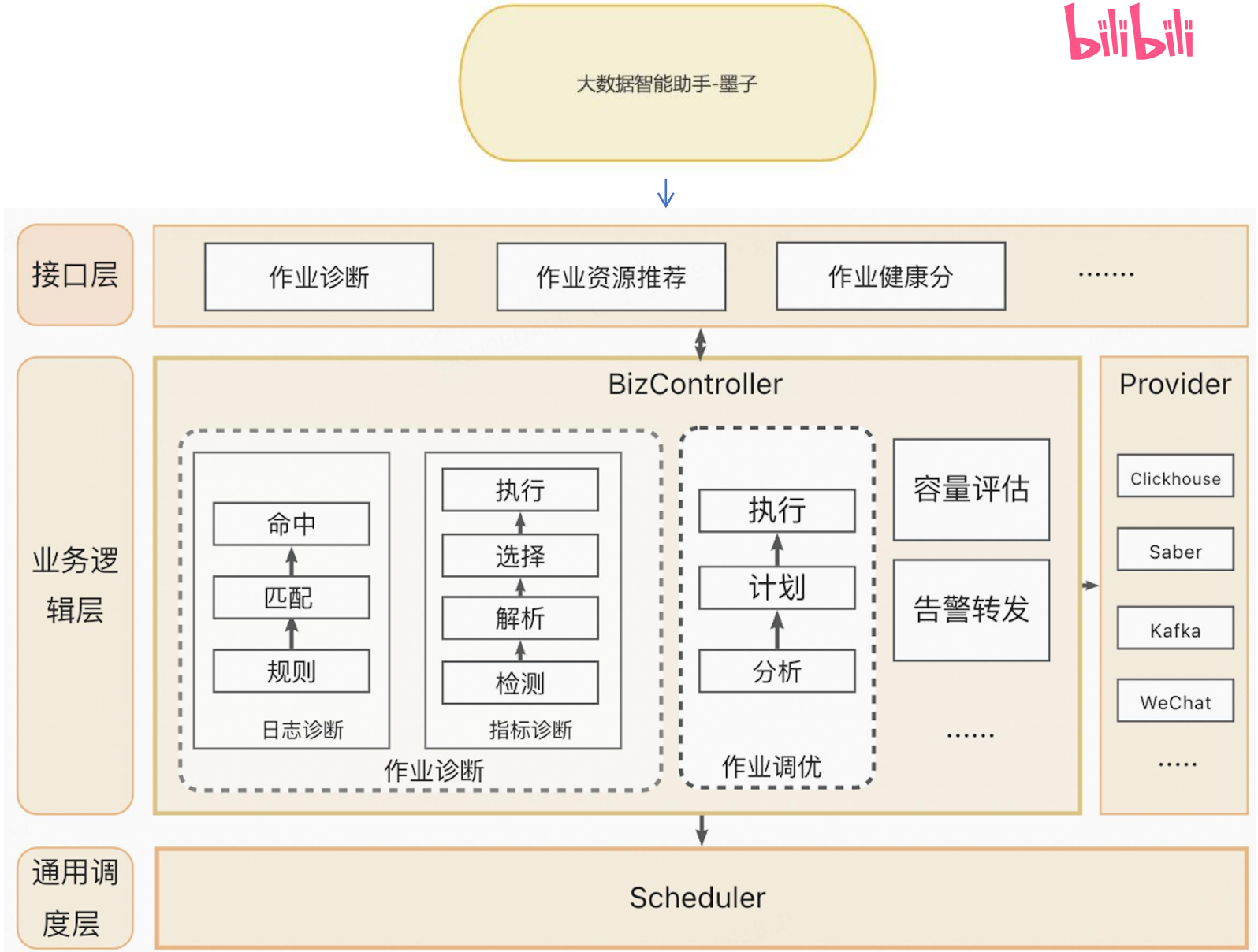
Kyuubi 诊断

- 历史画像
- Session 拆分



实时诊断的架构设计

- 实时诊断
- 自动扩缩



智能小助手服务人群

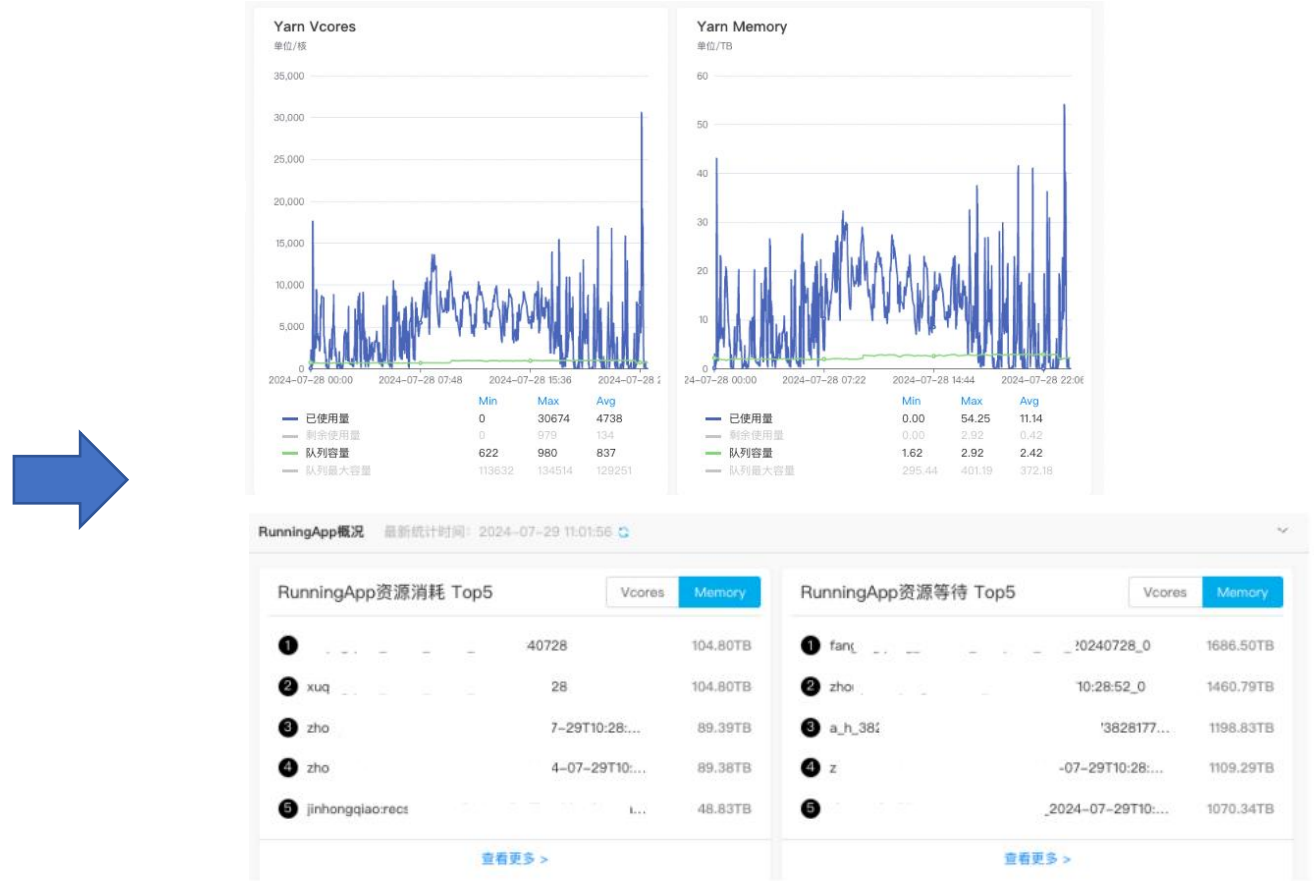
- SRE工程师
- 组件运维人员
- 数仓专业用户
- 非相关领域用户

SRE 用户

- SRE 工程师从大量的指标中判断中脱离出来，通过询问大模型直接给出机器潜在的问题。

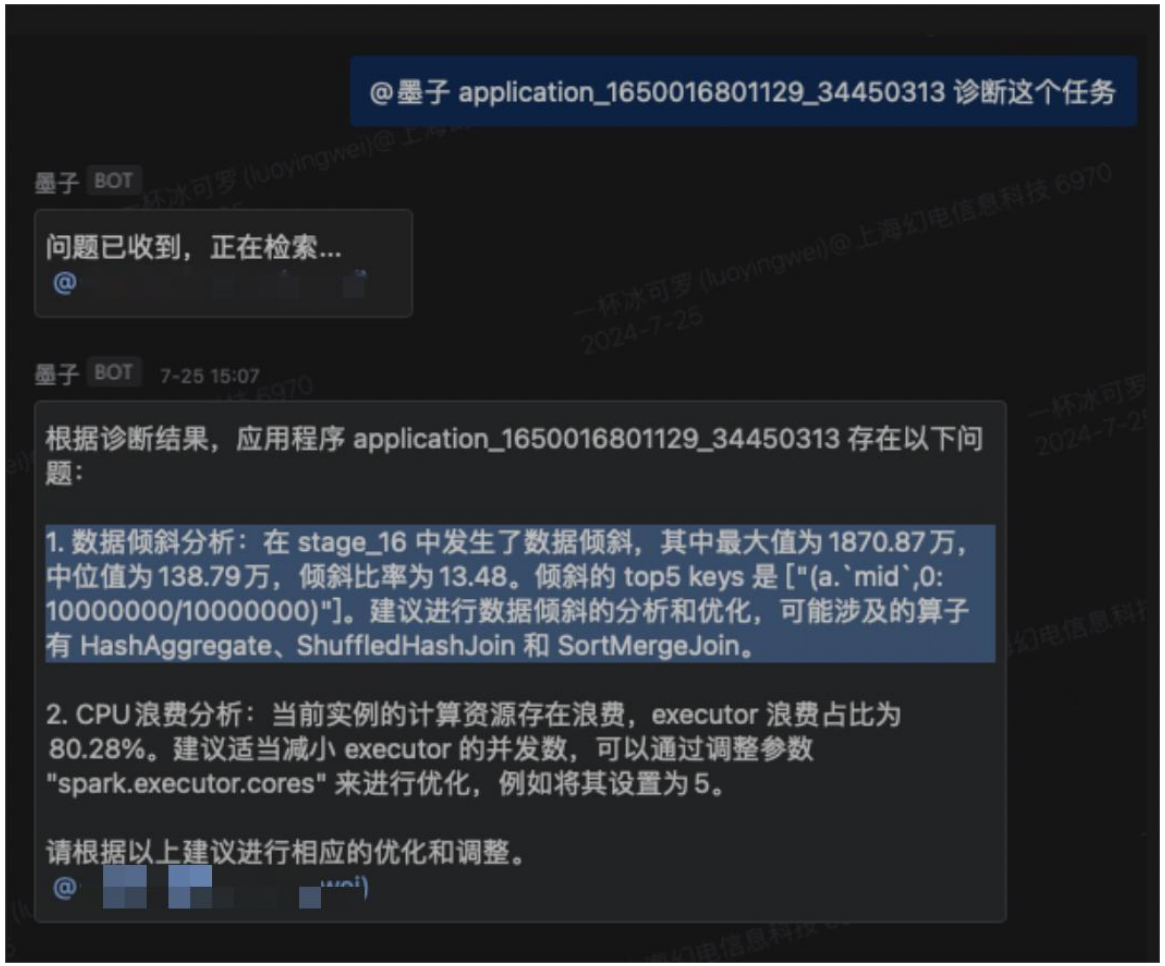
组件运维人员

- 当发现任务慢，组件运维人员可以通过询问墨子，诊断出作业运行慢是因为资源调度出现拥塞，并能否通过可视化的看到资源拥塞时序变化



数仓专业用户

- 当任务变慢的时候用户可以询问智能体，智能体发现作业存在数据倾斜，并且将倾斜的key，和key对应的值展示给用户，帮助用户快速定位数据质量问题。



非相关领域用户

- 有些刚涉及相关领域的用户对于底层相关的专业知识也可以提供问答

04

挑战展望

挑战

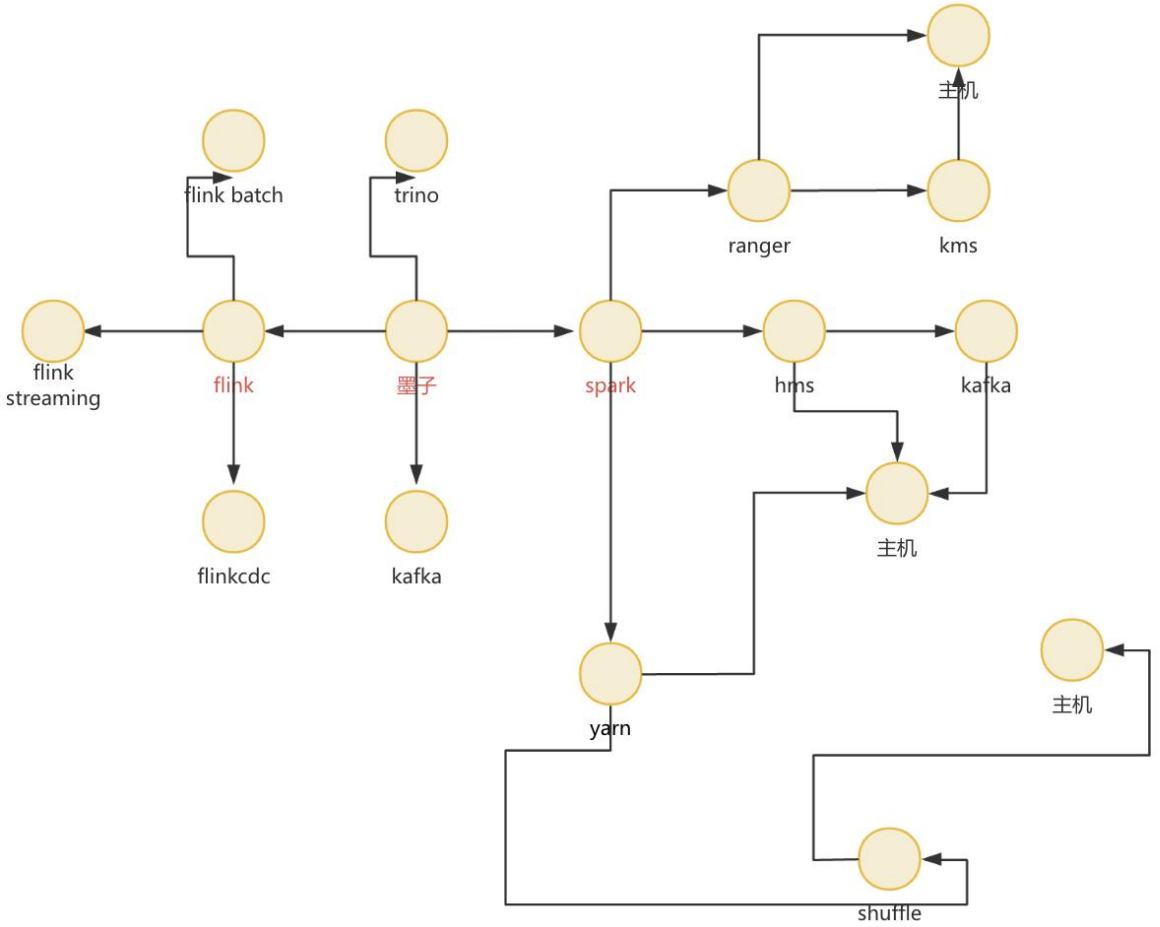
- 问答过程中准确度
- 数据质量
- 用户问题复杂多样

展望

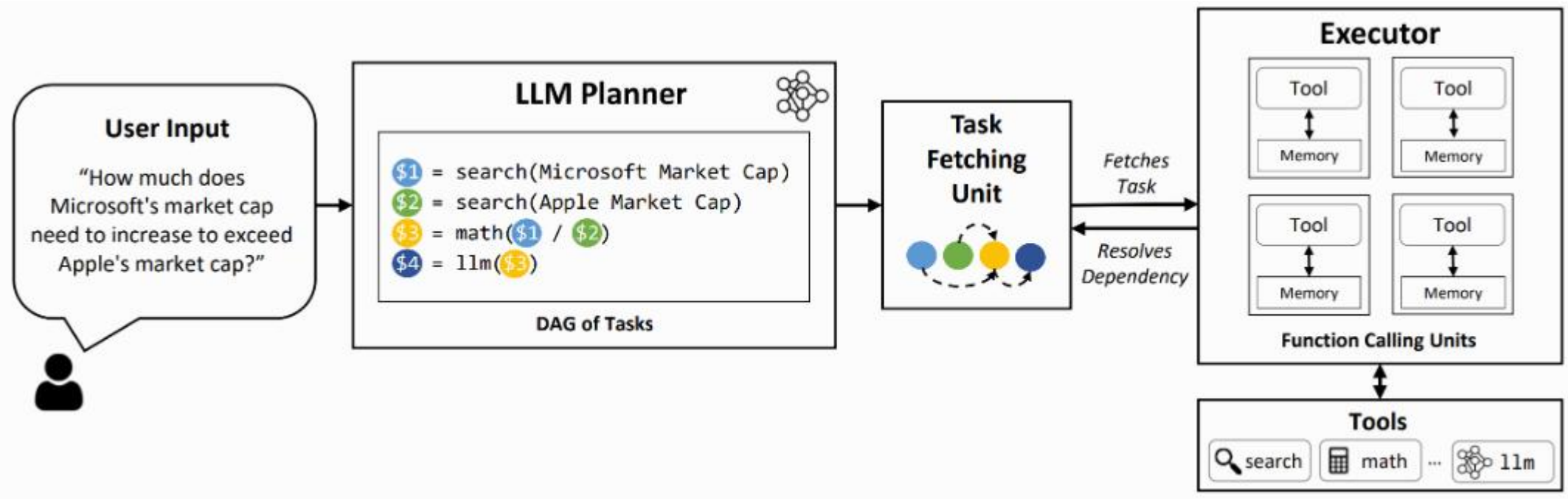
- 多专家系统
- 减少推理latency
- 提升产品体验

MutiAgent

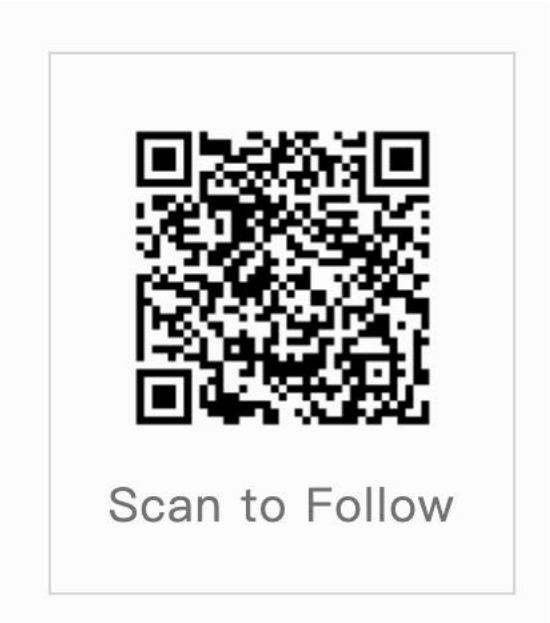
- 更多的组件
- 构建思维链
- 故障半径侦测



LLMCompiler



欢迎关注哔哩哔哩技术



THANKS

DataFunSummit # 2024