

SRE-Copilot



LLM & Aiops 在字节跳动的探索与实践

字节跳动--王宁

SRE-Copilot

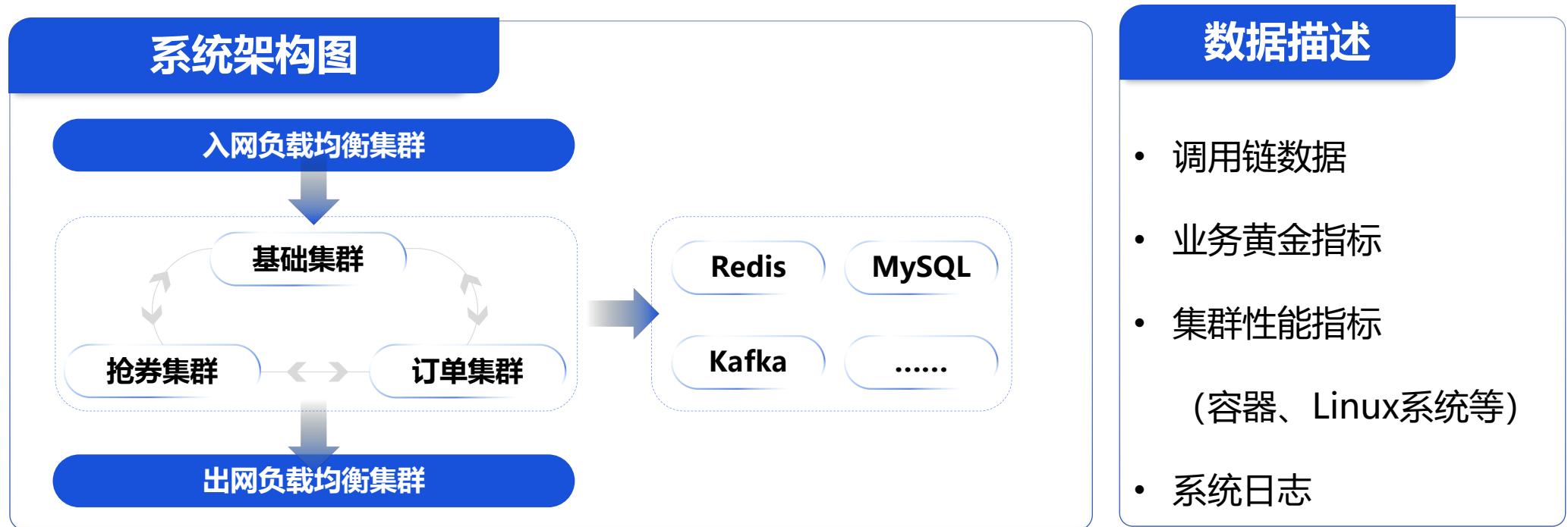
- AIOps 挑战赛方案
- SRE-Copilot架构
- SRE-Copilot技术细节
- 在字节跳动的探索实践



第一章节

赛题及数据分析

- 基于建行稳定性系统模拟建行生活类APP的真实环境
- 希望解决企业运维团队面对规模庞大、结构复杂、动态变化的运维数据需要解决的一系列挑战



SRE-Copilot 优势



痛点

系统复杂，海量数据

数据无标注，训练成本高

接入，维护成本高

无法解决未知故障

交互使用成本高



传统AIOps

单个运维专家，甚至单个团队难以掌握上下游全部知识，技术细节。也难以处理全部告警/异常

大部分异常检测算法需要依赖标注，无监督算法能力一般。根因诊断算法更加依赖标注。专家经验很难编码成算法模型

维护门槛高，要完全理解每一个复杂模型，客户的数据和系统都是私域的，需为客户现场定制与优化，增加了接入成本。调整或接入新数据要重新训练

无法推理未知故障

交互复杂，需要严格传递参数等



SRE-Copilot

大模型几乎能学习人类全部知识。通过Multi-Agent以及知识库可以无限扩展

把专家经验转化为故障表现，让模型推断，无需训练

通过“混合专家模型”的集成学习概念，只需关注组件与模型，客户自己的模型/逻辑也可以像乐高积木一样轻松接入，灵活调整。甚至框架自己可以动态编排

LLM已经出现了涌现和推理能力，基于自己的通用知识，并且可以不断学习领域知识进行推理，似乎是解决新故障根因定位的最佳选择

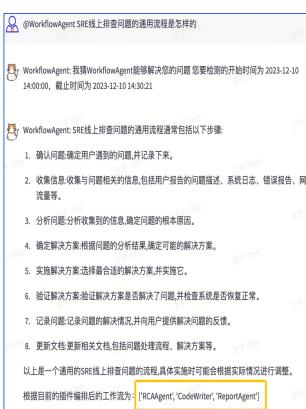
自然语言交互，更加智能，可以开放给更多用户

SRE-Copilot: 基于function-call的多功能实现

LLM优势：收敛模型统一入口，理解、拆分用户意图，编排调用不同工具，实现多场景复杂智能运维，并持续学习迭代

故障分类

故障类别	解释说明
流量异常	用户入流流量突增异常，一般建议进行限流
磁盘空间不足	磁盘被写满，一般建议清理冗余文件
CPU利用率过高	无状态服务的CPU利用率突增，一般建议重启处理
JAVA进程GC异常	GC次数过多，导致程序性能下降或崩溃，一般建议调整GC参数



ReportAgent: 告警总结：

一、各类告警的开始时间

1. 性能指标故障最早的开始时间：2023-09-14 00:23:00
2. 调用链异常：无法确定，因为日志异常信息中并未提及具体时间。
3. 日志异常：无法确定具体时间，因为日志异常信息中并未提及具体时间。

二、故障集群列表，返回结果中请对故障集群进行去重及排序

1. 性能指标故障：nginx_01、Weblogic_07、Weblogic_09
2. 调用链异常：未提及具体集群
3. 日志异常：未提及具体集群

三、各类告警中的异常具体表现

1. 性能指标故障：nginx_01、Weblogic_07、Weblogic_09的磁盘空间不足，指标异常表现为/home磁盘使用率过高。
2. 调用链异常：大量调用链中存在点丢失问题错误，占比超过0.5491，同时伴有少量数据库写入错误，违反数据库唯一约束。
3. 日志异常：未提及具体异常表现。

四、此次故障的根因是什么，请结合异常情况以及根因定位结果进行总结，给出最有可能的故障根因

1. 性能指标故障：磁盘空间不足，可能是由于系统负载过高、硬件故障或软件错误导致的。建议对磁盘空间进行合理分配，优化系统性能，同时检查硬件设备状态和软件配置，以尽快解决问题。
2. 调用链异常：Web服务在集群内的调用失败，可能是由于网络问题、服务间通信异常、服务进程异常等原因导致。同时，也有可能与数据库的约束冲突有关，例如表结构不匹配、数据重复等问题。
3. 日志异常：未提及具体根因，建议检查日志异常信息，以便更好地判断故障原因。

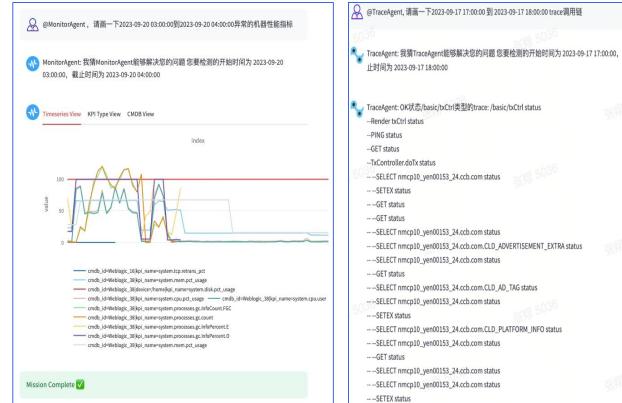
五、推荐采取的措施和未来改进的方向

1. 性能指标故障：对磁盘空间进行合理分配，优化系统性能，同时检查硬件设备状态和软件配置，以尽快解决问题。
2. 调用链异常：对Web服务进行排查，找出具体的调用异常点，同时对数据库进行相应的调整和优化，以确保系统的稳定运行。
3. 日志异常：检查日志异常信息，以便更好地判断故障原因。未来改进的方向包括完善日志记录机制，提高日志分析能力，以便更好地发现和解决潜在问题。

工作流生成（运维计划） 故障自愈+代码生成

故障报告

运维可视化



@QAProxy: 你知道aiops挑战赛的主办方是谁吗

QAProxy: 我猜QAProxy能够解决您的问题 您要检测的开始时间为 2023-11-19 17:10:12， 截止时间为 2023-11-19 17:10:12

QAProxy: 很抱歉，我没有找到关于AIOPS挑战赛主办方的具体信息。

@QAProxy: 你知道aiops挑战赛的主办方是谁吗，请参考local的数据源

QAProxy: 我猜QAProxy能够解决您的问题 您要检测的开始时间为 2023-11-19 16:10:44， 截止时间为 2023-11-19 17:10:44

QAProxy: 是的，我知道。本届CCF国际AIOPS挑战赛由中国计算机学会（CCF）、清华大学、中国建设银行股份有限公司联合主办；CCF互联网专委会、清华大学计算机科学与技术系、中国建设银行股份有限公司运营数据中心、北京必示科技有限公司承办。

知识库问答

比赛效果展示

采用与建行云堡垒机同一套代码复现 相近配置（8c16g显卡Nvidia Tesla-v100）模型chatglm3-6b



第二章节

AI-Agent相关概念：Tool calling

Function calling是可靠地将LLMs连接到外部工具以实现有效的工具使用和与外部API的交互的能力。



```
messages = [
    {
        "role": "user",
        "content": "What is the weather like in London?"
    }
]
```

```
tools = [
    {
        "type": "function",
        "function": {
            "name": "get_current_weather",
            "description": "Get the current weather in a given location",
            "parameters": {
                "type": "object",
                "properties": {
                    "location": {
                        "type": "string",
                        "description": "The city and state, e.g. San Francisco, CA",
                    },
                    "unit": {
                        "type": "string",
                        "enum": ["celsius", "fahrenheit"]
                    },
                    "required": ["location"]
                }
            }
        }
    }
]
```

```
ChatCompletionMessage(content=None, role='assistant', function_call=None, tool_calls=[ChatCompletionMessageToolCall(id='...', function=Function(arguments='{"location":"London","unit":"celsius"}', name='get_current_weather'), type='function')])
```

AI-Agent相关概念：RAG

检索增强生成 (Retrieval-Augmented Generation, 又称RAG) 通过检索LLMs之外的数据源来支持其生成答案。RAG=搜索+LLM提示，根据用户的查询要求，LLMs会使用搜索算法从外部数据源获取上下文信息，最后，查询和检索到的上下文合成后送入到LLM的提示中。



私域知识

“公司服务器的数量，线上的利用率是多少”
“当前工单状态是什么”



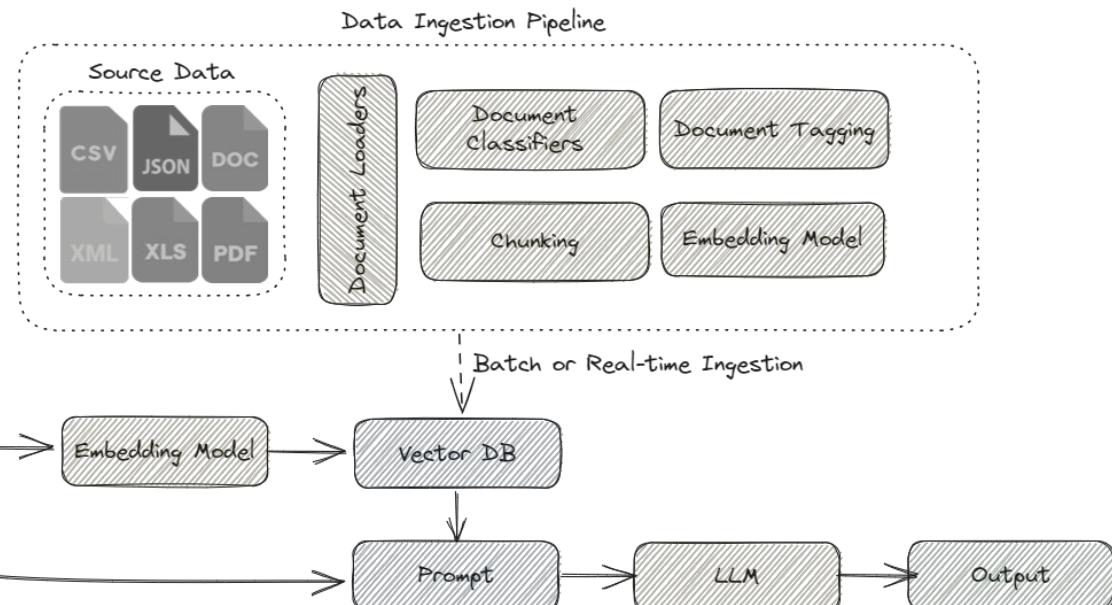
新知识问题

“今天有哪些故障？”
“最新电影的主演是谁”



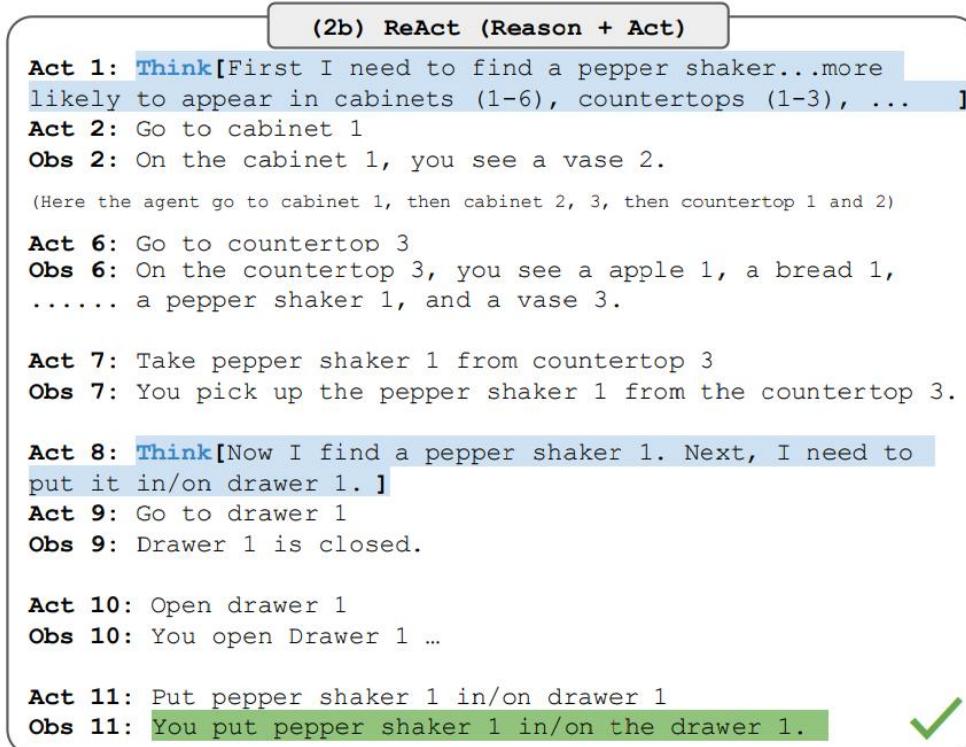
长尾问题

“使用XX语言在某环境如何连接XX数据库”

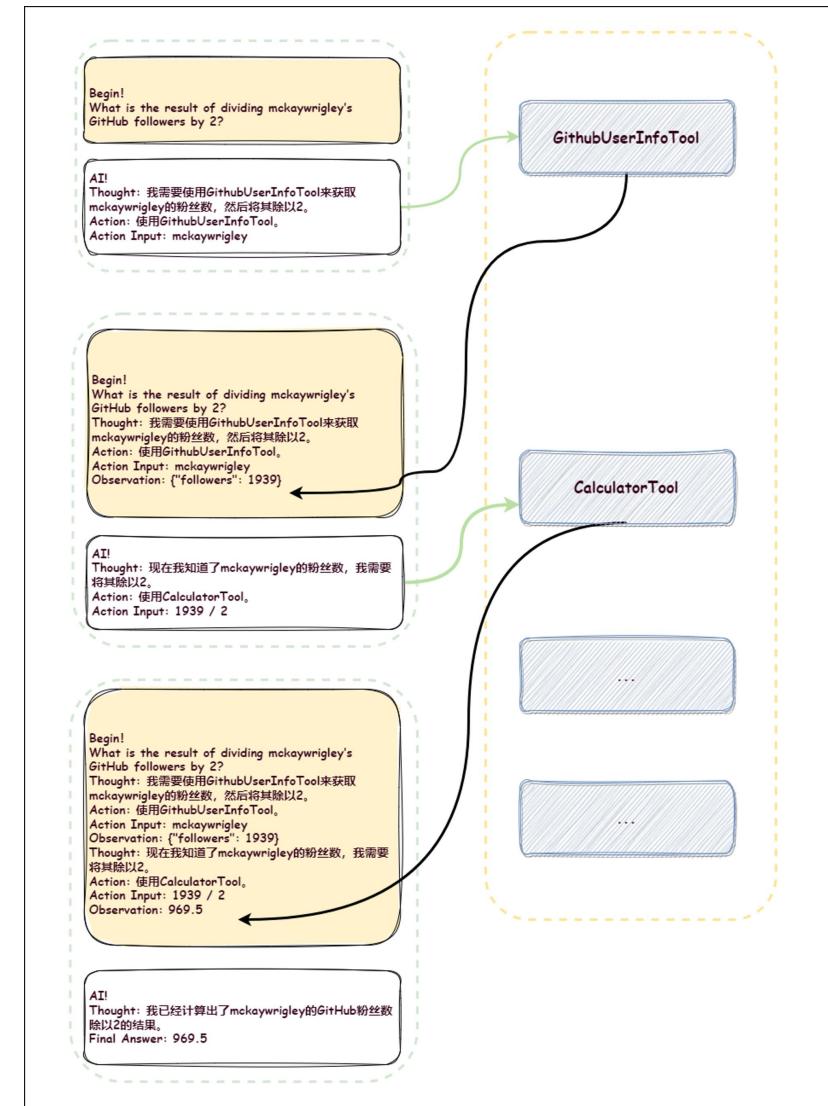


AI-Agent相关概念：Reason + Act

ReAct是一种LLM提示和结果处理方法，它结合了推理、行动规划和知识来源的整合，使LLM超越其语言模型，在其预测中使用来自现实世界的信息。ReAct是推理和行动的结合。



当任务是多步，或者任务复杂结果充满不确定时，react表现更好

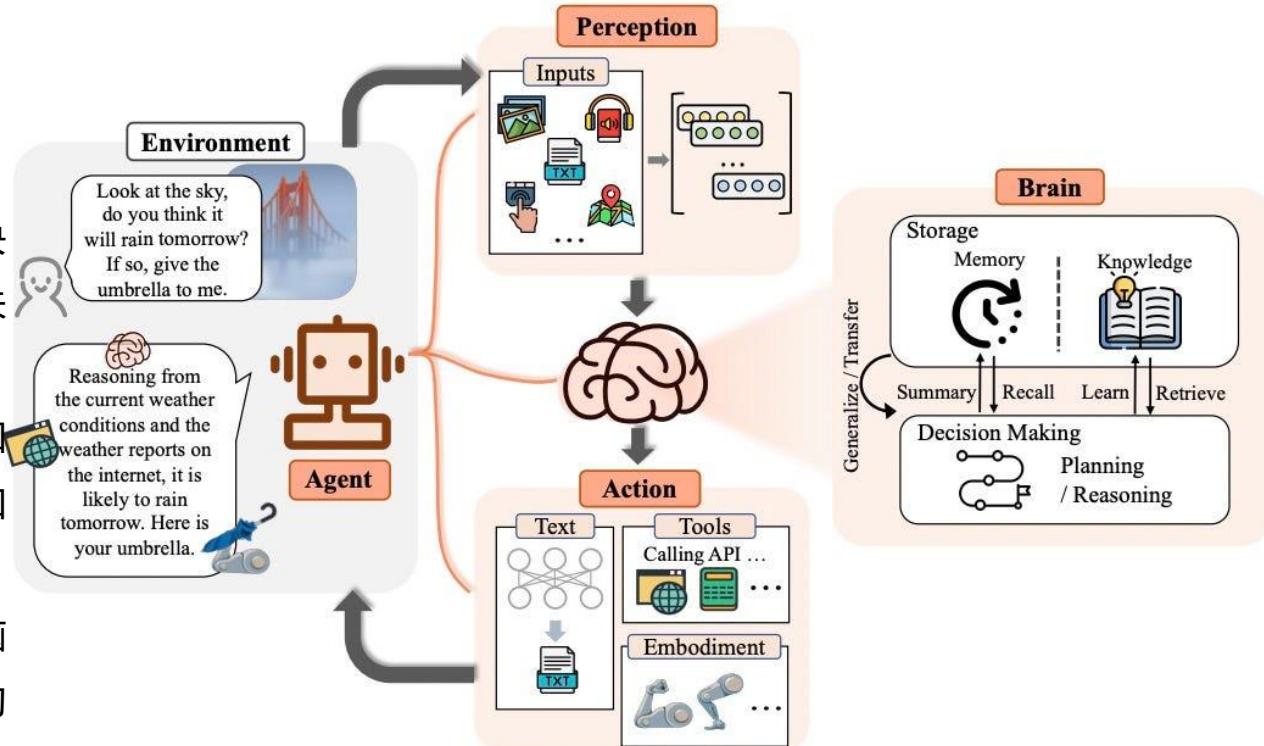


AI-Agent相关概念—Agent智能体

代理（Agent）指能自主感知环境并采取行动实现目标的智能体。

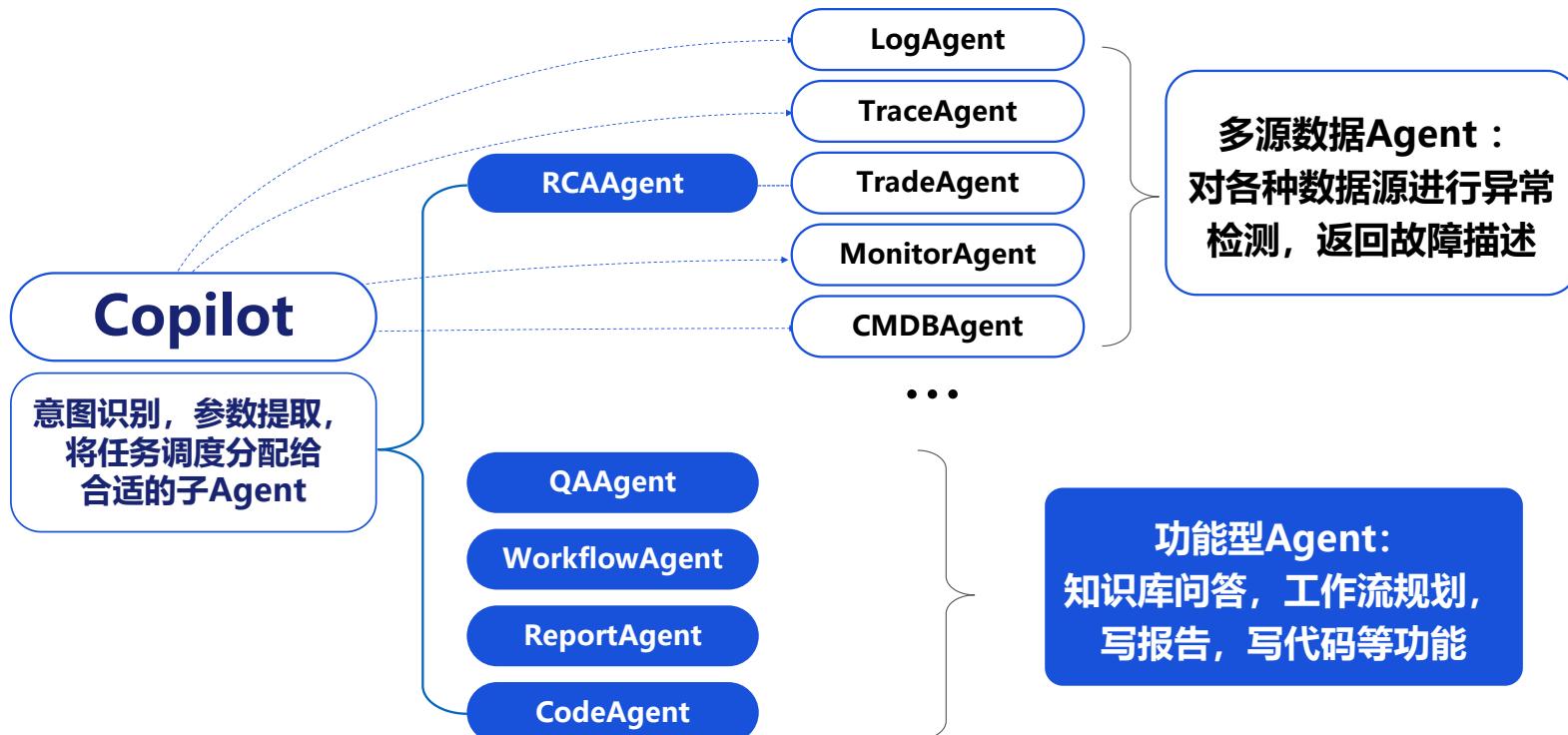
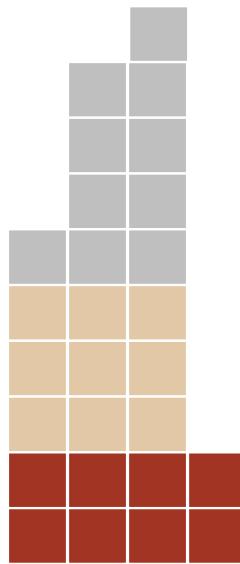
AI代理的整体框架由三个关键部分组成：大脑、感知和行动。

- 大脑：大脑主要由一个大型语言模型组成，不仅存储知识和记忆，还承担信息处理和决策功能，并能呈现推理和规划过程以处理未知任务。
- 感知：感知模块的核心目的是将代理的感知空间从纯文本领域扩展到包括文本、听觉和视觉模态。
- 行动：在代理的构建中，行动模块接收大脑模块发送的行动序列，并执行与环境交互的行动。



SRE-Copilot架构简介

SRE-Copilot是基于LLM的多场景智能运维框架，支持Multi-Agent协作与动态编排，具备计划、记忆、反思、推理与ReAct等能力，为SRE提供智能化服务。



参考GPT的思想，通过集成学习多个专业的LLM的agent组成强大的混合专家(*MoE, Mixture of Experts*)系统。

第三章节

主要运维能力：异常检测

——基于ReAct框架与思维链CoT的Multi-Agent编排

角色定义

Copilot主持人: 解析用户需求，制定运维Plan，安排不同Agent工作（如根因定位交给RCA）

多数据源Agent: 分别负责不同模态的数据，选择合适的算法进行异常检测与检索

RCAAgent: 收集其他Agent检测到的异常信息与链路、配置信息，进行根因定位

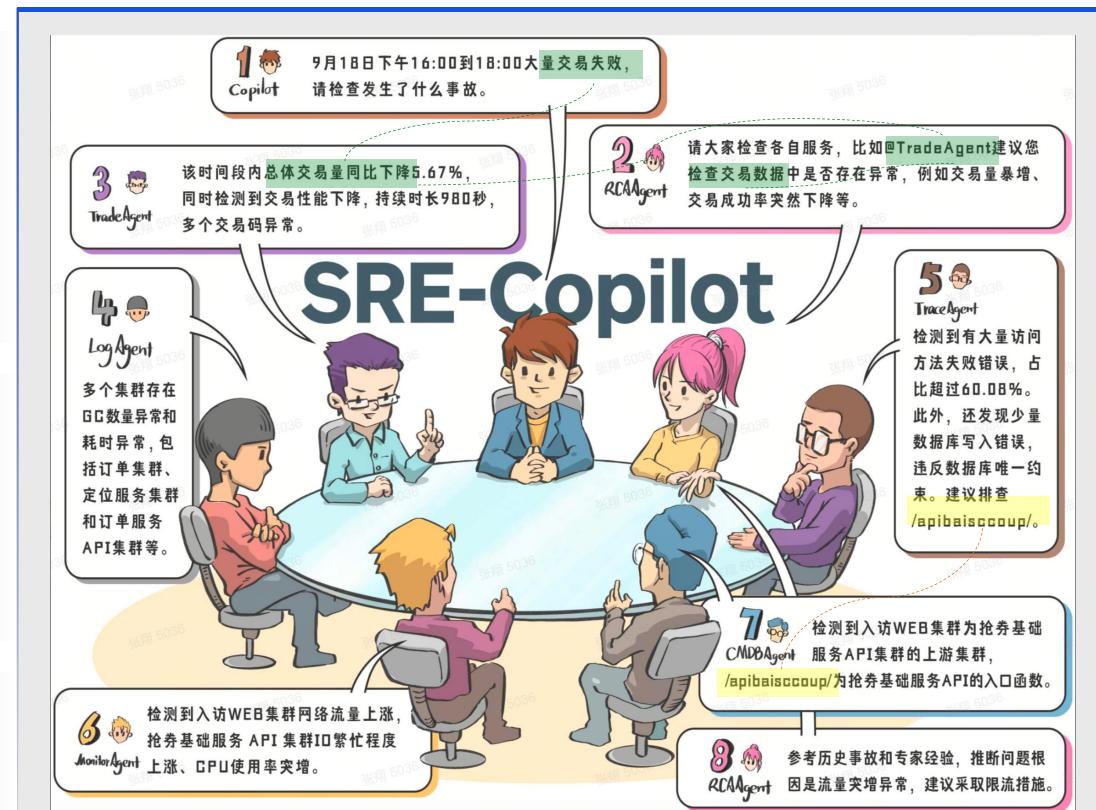
Multi-Agent编排

ReAct包括推理Reasoning和行动Action，推理帮助模型生成、追踪和更新计划并处理异常，行动允许模型与外部环境交互以获取更多信息Observation，提升准确率与适应性

每个Agent均根据检测到的异常动态编排，RCAAgent负责收敛协作轮次，并根据其他Agent反馈决定下一步分析与下钻的方向

优势

模拟真实云平台跨组件协同定位，可以采用多个agent替代多个组件运维团队



主要运维能力：根因定位

——基于RAG增强的推理与反思

本次故障持续10分钟，CPU飙升，内存打满，接口出现大量失败.....



向量数据库

专家经验：内存打满后服务一般不可用，此时应重启.....

历史故障：2023年11月5日，xx接口不可用，持续半小时，CPU飙升.....

专家经验：.....

历史故障：.....

Knowledge 本次故障.....

专家经验：.....

专家经验：.....

历史故障：.....

Question



大语言模型

根因：内存打满
建议：重启

应用专家经验

- RAG检索增强：检索相关/支撑文档，作为上下文和原始提示词组合，再提交给LLM使用。LLM的知识是静态的，且会有“幻觉”现象，RAG让LLM不用重新训练就能获取最新相关信息或者私域知识
- 我们用RAG检索最符合本次故障的Top N专家经验，交由LLM进行根因定位

学习新的故障

- 每次诊断结果会加入模型记忆，再次诊断时对最近专家经验与诊断结果进行推理，让模型获得持续学习的能力
- 反思 Reflexion：通过让模型进行自我评估和自我反省，能够进一步提升模型推理根因的准确率30%以上

推理未知故障

- 对于未发生过的，也没有专家经验的新故障，SRE-Copilot也会根据自己的知识，尝试推理故障根因
- 本次比赛本地化部署的模型为6B的小模型，但我们仍然在一些新故障上发现模型的推理能力，比如在没有经验和历史的情况下，正确诊断出磁盘写满/GC等故障

第四章节

基于LLM的RCA-Agent构建

我们优先选择了在根因诊断的场景中探索Agent的模式，期望能够在日常的故障定位，oncall等场景降低SRE同学的负担

01 知识库的构建

02 基础工具的构建

03 核心工具：LLM根因推理

04 工作流的构建

05 Agent的使用场景

The screenshot displays a software interface with a sidebar containing various management and monitoring tools. The main area is divided into three sections:

- Tools (工具插件):** A table listing diagnostic items and their descriptions, such as '变更事件查询' (Change Event Query) and '指标异常检测' (Metric Abnormal Detection). The table includes columns for Diagnosis Item Type, Last Updated, and Updated User.
- Workflows (工作流):** A table listing processes with details like 'Copilot Available' status, last updated time, and updated user. Processes include 'wangning_test', 'chat-ops', and 'tce-rca-flow'.
- Knowledge (知识):** A section showing a message: '开发中，预计24Q1上线' (Under development, expected to be released in Q1 2024).

知识库的构建

知识库当前包含以下三个部分，后续我们还会持续引入用户文档，历史oncall等不同类型的数据

1. 排障的专家经验

- 这里是只针对根因定位场景，业务同学可以将他们的经验积累&管理起来。
- 我们定义专家经验是一组故障根因，故障表现，以及故障止损措施的组合。以便大模型去推理故障

The screenshot shows a table titled '故障诊断经验' (Fault Diagnosis Experience) with the following data:

故障名称	故障根源	根因描述	故障止损措施	操作
Pod pending	Spacex	大量的pod异常，存在多个space集群侧变更单，如编排发布，sops的事件等	RollBack	编辑 删除
Pod pending	IPv6	单个pod报错异常，且存在ipm变更，且该osm存在IPv6布配问题，要求osm重新一致	call business	编辑 删除
Pod pending	TBD	多个pod异常，集群可用资源变少	oncall	编辑 删除
Pod pending	resource	多个osm异常，集群可用资源变少	add resource	编辑 删除
Pod pending	inf/data	异常osm包含inf/day字段	nothing	编辑 删除

共5条 [1](#) / [10 / page](#)

2. 故障场景的SOP文档

- 通过sop文档的形式，希望能提供给组件同学更加灵活的知识管理方式
- 当前由于大模型的能力局限，我们通过这种半规范的文档，将指标/诊断项，诊断流程等内容管理维护起来

The screenshot shows a table titled '文档列表' (Document List) with the following data:

标题	描述	文档源	URI	创建人	创建时间	操作
pod pending的SOP	集群发生pod pending告警时排障流程	飞书文档	SQqda07y089cxXnBcBk3nGf	王宁	2024-04-07 11:43:10 UTC+8	编辑 预览

共1条 [1](#) / [10 / page](#)

3. 历史的故障信息

- 每一次的历史故障，会被记录下来，用来给组件同学“训练”/打标模型。

The screenshot shows a table titled '故障管理' (Fault Management) with the following data:

执行ID	流程名称	流程输入	步骤名称	输出结果	建议输出结果	原始输出	步骤输入	执行时间	操作
495740	wanging_test	{"l": "171176916..."}	lm_infer_4	准确识别为 TBO, ...	{"human_res": "res..."}	2024-03-30 11:27:49	故障归因		
492995	wanging_test	{"l": "171168334..."}	tox_platform_query_2	pm存在变更工单, ...	{"pipeline_res": "res..."}	2024-03-29 14:23:56	故障归因		
489465	wanging_test	{"l": "171161377..."}	anomaly_detection_0	集群pod pending检测...	{"human_res": "detal..."}	2024-03-28 16:16:23	故障归因		
495746	wanging_test	{"l": "171176916..."}	anomaly_detection_0	集群pod pending检测...	{"human_res": "detal..."}	2024-03-31 11:27:49	故障归因		
492544	wanging_test	{"l": "171168323..."}	anomaly_detection_0	集群pod pending检测...	{"human_res": "detal..."}	2024-03-29 11:34:56	故障归因		
492544	wanging_test	{"l": "171168323..."}	space_change_query_1	集群削没有变更	{"human_res": "res..."}	2024-03-29 11:34:56	故障归因		

基础工具的构建

参考openai的tools/gpts的接入方式，我们将运维场景的指标和其他基础工具管理起来。基础工具包含几类，例如指标通用的异常检测，变更事件查询，组件自定义的检测项。

1.一些集群诊断场景的指标 2.自定义的检测项

用户实际部署的工具包含通用指标的异常检测，变更事件查询，自然语言的意图理解，大语言模型的根因推理

Metrics指标						
metrics指标名	监控部署区域	指标描述	核心tags	异常时下钻	结果描述	指标链接
add_on_controller_prod.clusterer_metrics.pod_count	China-North	集群中 podpending 的数量	<ul style="list-style-type: none"> reason=ResourcePending namespace=default status=Pending cluster=\${cluster} node_level=\${nodeLevel} psm=* 	psm	<p>反映集群当前的Pending Pod数量，常态下应 < 10</p> <p>指标上升：当前集群的Pending Pod增加，需要尽快排查和干预</p> <p>指标下降：当前集群的Pending Pod降低，需要持续关注，直至降低为0</p> <p>智能诊断需求：需要能够分析造成指标上涨的具体PSM，基于分析出的PSM进一步分析该PSM Pending的原因（通过抽样部分Pod，分析其Pod Event）</p>	https://metrics/max:add_on_clusterer_metrics.pod_count
Influxdb指标						
influxdb指标	指标描述	核心tags	异常时下钻	结果描述	dbname	InfluxQL
escloud_es_pod_cpu_metric	CPU耗使用率	node	node	<ul style="list-style-type: none"> 指标上连续5分钟超过85%：CPU使用率过高 智能诊断需求：返回使用率过高的node列表，进一步排查差机器问题还是业务问题 	xxx	<pre>SELECT max("value") FROM "escloud_es_pod" /"\$escluster" AND \$timeFilter GROUP BY time(1m)</pre>
诊断项名称	接口url	接口描述	参数	返回结果	调用示例	数据处理方法
集群状态查询	https://xxx.byted.org/api/v1/cluster/	查询集群状态	请求方式: get body: 无 鉴权: 无 params: {"cluster": "\${cluster_name}", "time_range": "2024-01-09%2017:20:00,2024-01-09%2017:40:00"} {"	<pre>"success": true, "data": { "cluster": "xxx", "region": "cn-beijing", "client_status": [{ "status": "NO", "name": "延迟过大", "result": "延迟过大，延迟为2.8s", "reason": "" }] }</pre>	curl -location https://xxx.byted.org/api/v1/cluster/ ?cluster=\${cluster_name}&time_range=2024-01-09%2017:20:00,2024-01-09%2017:40:00"	<p>需要返回“client_status”列表中的所有状态为“NO”的条目，“result”字段作为当前集群异常表现</p> <pre>let raw_res = res.Result().JSONBody(); let res = filter(raw_res, .status == "NO"); map(res, .result)</pre>

诊断项列表						
ID	诊断项名称	描述	诊断项类型	创建时间	更新时间	更新用户
79	llm_infer	大语言模型推理根因	HTTP请求	2024-03-20 14:42:32	2024-03-20 16:45:33	
78	tce_platform_query	tce平台变更查询	HTTP请求	2024-03-20 11:59:30	2024-03-20 15:51:02	
77	spacex_change_query	查询spacex平台的变更	HTTP请求	2024-03-20 10:39:32	2024-03-20 16:29:06	
46	bke_prober	写入过程状态	规则组	2024-01-31 22:16:52	2024-01-31 22:16:52	
16	chat_ops	自然语言诊断任务指令机器人	HTTP请求	2024-01-04 02:47:58	2024-01-17 14:35:57	
15	tce_dignosis_rule_set	TCE常用诊断规则	规则组	2024-01-04 02:29:03	2024-01-04 02:29:03	
14	event_fetch	指标异常相关的变更事件获取	HTTP请求	2024-01-04 02:22:33	2024-01-04 02:22:33	
13	anomaly_detection	指标管理配置的指标异常检测结果获取	HTTP请求	2024-01-04 02:16:30	2024-03-20 15:01:48	

核心工具：LLM根因推理

- 相较于传统的根因定位/故障分类
----将异常时刻的时序信息等进行编码聚类，在向量空间里面求距离和相似度进行分类。
- 我们尝试对异常时刻的信息映射到自然语言描述，利用大语言模型的能力进行分类

The screenshot shows a 'Fault Management' interface with a 'Fault Diagnosis Experience' tab selected. It lists five faults:

Fault Name	Root Cause	Description
Pod pending	SpaceX	大量psm异常, 存在多个spacex集群变更工单, 如编排发布, sops的事件等
Pod pending	IPV6	单个psm明显异常, 且存在psm变更, 且该psm存在ipv6等配置问题, 要求psm前后一致
Pod pending	TBD	多个psm异常, 集群可用资源变多
Pod pending	resource	多个psm异常, 集群可用资源变少
Pod pending	int.data	异常psm包含inf.dayu字段

Below the table, a 'Diagnostic Output Summary' section contains a box with text about multiple psm anomalies and IPv6 configuration issues.

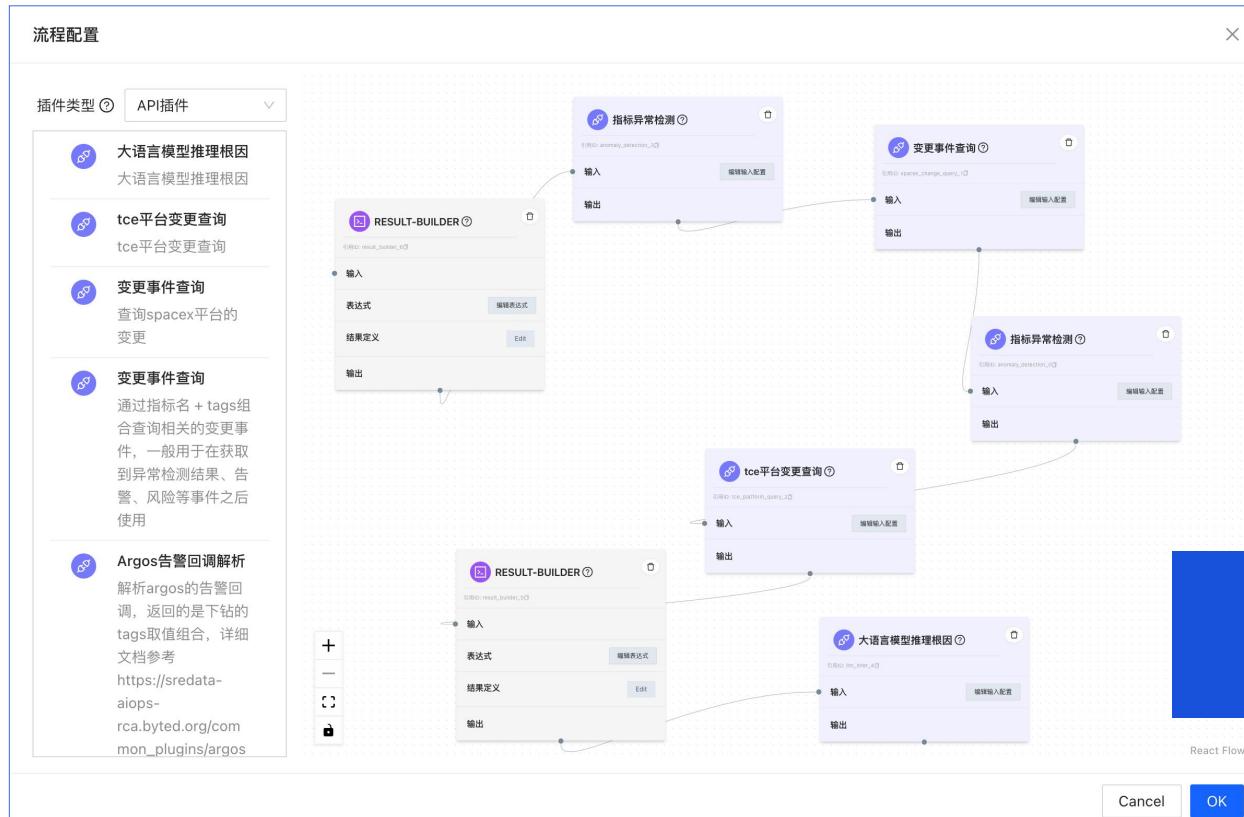
At the bottom, there is a note: '显示故障记录和专家经验整合的例子' (Show an example of integrating diagnostic records and expert experience).

The screenshot shows an 'LLM Diagnostic Prompt Template' interface. It includes sections for 'Expert Experience and Historical Faults', 'Current System Status', and a summary box. Below it is an 'LLM Intelligent Diagnosis Output' section with a summary of the current system status.

At the bottom, there is a note: '用户可以调整专家经验, 结果, prompt模板等, 对模型进行测试' (Users can adjust expert experience, results, prompt templates, etc., to test the model).

工作流的构建

根据上述提到的基础工具，用户可以自由组合这个工作流



- 目前仍然依赖用户自己配置同时在探索让模型自己从用户的SOP文档中生成工作流
- 未来期望可以利用ReAct的思路动态编排这些诊断项



Agent的应用与调试



通过引入了LLM，改变了传统的交互方式，甚至可以通过模糊地提问，识别意图和参数。并将结果优化展示后返回



当前的一些使用场景1

1. 在群里/私聊机器人，用自然语言提问，会匹配你想要的工作流



当前的一些使用场景2

2. 事故拉群等场景自动匹配工作流，对聊天记录里的关键信息提取，进行诊断

The screenshots illustrate the diagnostic capabilities of the Agent. The left screenshot shows a general diagnosis for a cluster, listing various metrics and their anomalies. The right screenshot shows a more detailed diagnosis for a specific pod pending issue, providing a detailed log and suggesting corrective actions.

第五章节

未来展望

未来需要增强的几个方向



1.更智能的大模型

随着大语言模型的发展，可以期待越来越智能的通用大模型

在意图理解，动态决策，根因推理上准确率与效率会越来越高。



2.更强大的专业模型

在更加细分的方向，也有越来越多的专业模型

例如时序大模型，日志大模型等，这些专业模型可以当作工具一样集成到Agent的工具集中



3.更灵活的Agent框架

随着模型能力的提高，也能支持更多轮的对话

以及动态根据人类的反馈建议等作出决策。同时能让多人，多Agent进行合作。



4.更丰富的场景

不断探索Agent在运维领域的应用场景，早期可以像“实习生”一样，完成一些基础的工作，查看监控，处理数据等；

再进化到做一些基础的分析，执行简单决策的“初级员工”；到最后执行复杂分析任务，做高级决策的“资深专家”