

SOP × 动态Prompt × 裁判Agent—— 新一代大模型根因定位框架

锐捷网络股份有限公司 郭逸浩

主办单位：中国计算机学会（CCF）

承办单位：中国计算机学会互联网专委会、中国科学院计算机网络信息中心、中国移动研究院、清华大学

协办单位：华为2012实验室、阿里云、中兴通讯、中国移动九天团队、南开大学、西安电子科技大学、清华大学计算机科学与技术系、神州灵云

目录 CONTENTS

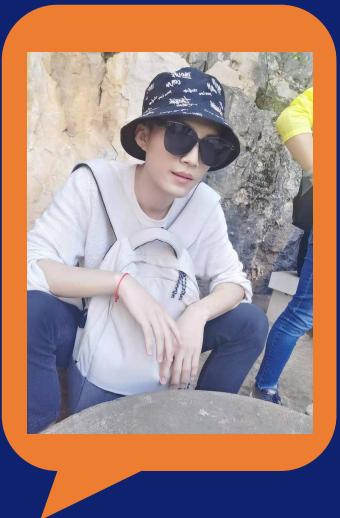
- 第一节 团队介绍
- 第二节 算法精度
- 第三节 选题分析
- 第四章节 解决方案
- 第五章节 创新性与实用性
- 第六章节 总结展望

第一章节

团队介绍



钟兆伟



郭逸浩



陈俊鹏



张纬峰



卢增通

锐捷网络，成立于2003年，行业领先的ICT基础设施及解决方案提供商。SBG数据智能组是面向中小企业网络售前地勘，售中配置，售后运维的全链路AI算法研发团队。以工商价值场景为导向，提供贯穿网络全生命周期的智能化能力，致力于打造新一代自智网络，探索以智能体驱动的自主运维模式，加速从人工管控走向智能自治，让网络真正做到好用易用。

第二章节

算法精度



本方案在提供的微服务架构数据集上实现了组件回答准确率达60%+（共400个任务）。

队伍	初赛分数	复赛分数			分差	初赛	复赛	稳定性
NO1	71.97	74.20	73.71	73.94	0.49	1	1	2
我们是有技术的	51.95	49.81	49.57	49.53	0.26	2	4	1
NO3	51.33	55.06	54.21	54.54	0.85	3	3	3
NO4	50.68	55.36	53.72	54.77	1.64	4	2	5
NO5	50.32	47.09	47.93	50.59	3.50	5	5	6
NO6	47.82	38.08	37.93	37.42	0.66	7	12	4

第三章节

选题分析

选题分析 - 背景与挑战

AIOps | 2025 CCF国际AIOps挑战
赛 2025 CCF International AIOps Challenge

在如今AIOps场景中，故障诊断方法常常面临以下核心挑战：

1、数据分散且多，分析主线丢失

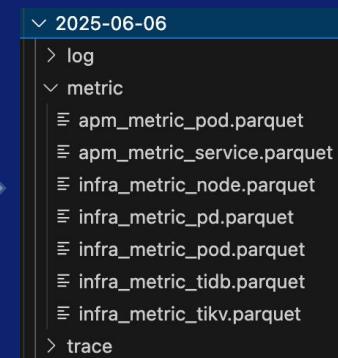
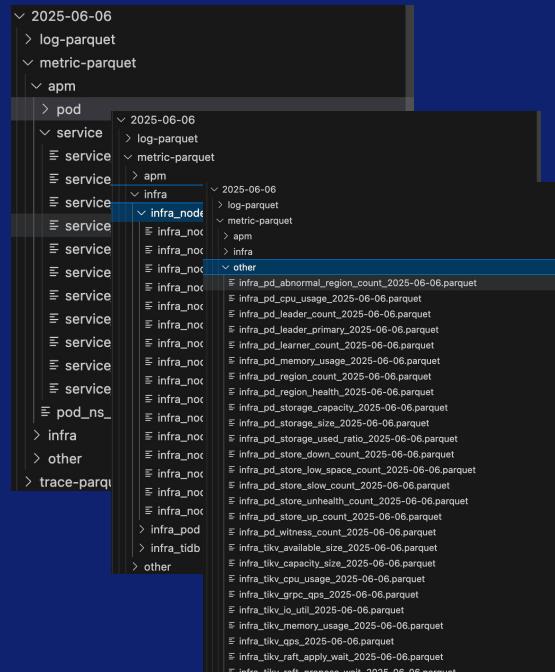
2、分析结果不具备可解释性

3、分析链路长，容错率低

4、结论不稳定，难以复现

5、指标繁多，难以捕捉关键信息

6、.....



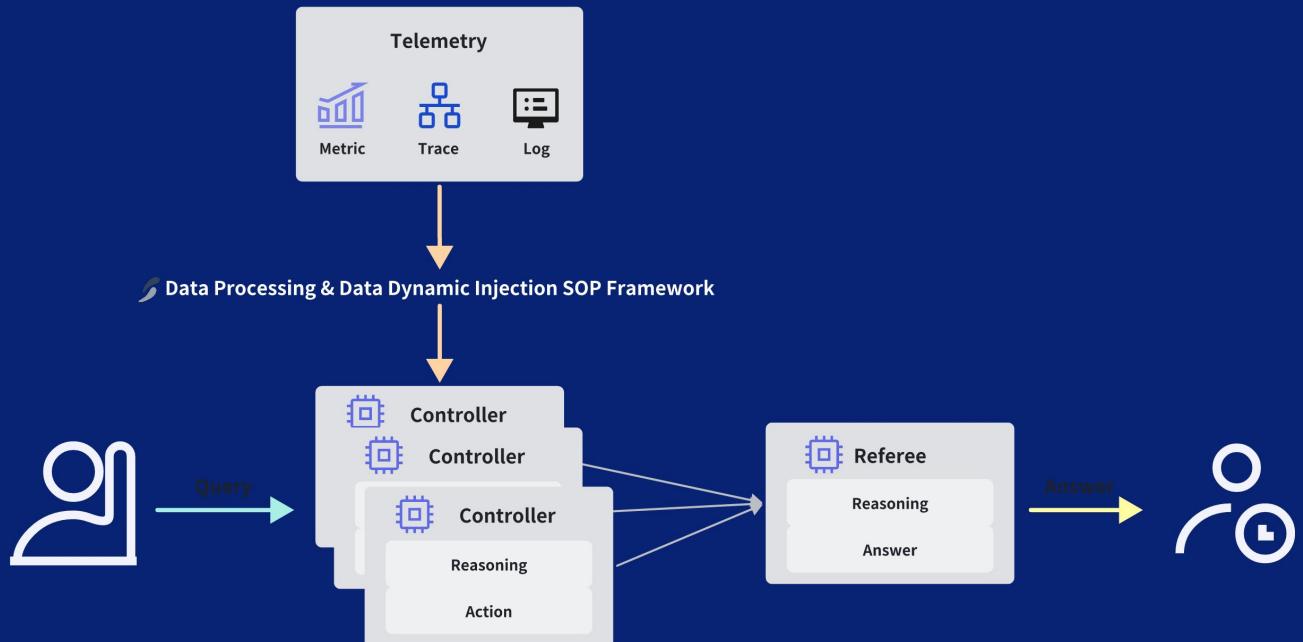
目标：构建一套全新的智能根因定位框架，旨在通过优化算法设计和分析流程，解决数据孤岛问题、简化推理链路、解决结论不稳定等核心矛盾。

第四章节

解决方案

4.1 解决方案 - 整体框架简介

本方案针对微服务环境中的数据分散、分析流程复杂、结论不稳定等问题，设计了一套高效、稳定、可解释的故障诊断框架，通过**数据预处理**、**动态 Prompt生成**、**系统图谱构建**、**标准化诊断流程**及**裁判Agent机制**五大模块，全面提升诊断效率、精度与稳定性。



总体方案架构简图

- 1. 数据预处理：**通过**三元组架构**[对象类型]-[组件名称]-[指标名称]，将海量数据压缩为多个核心指标文件；整合 APM、Infra、日志和调用链数据，构建跨源标准化数据集。
- 2. 动态 Prompt生成：**以故障时间窗口确定数据，筛选显著异常点，生成组件级多维特征，并输出**结构化提示模板**，确保信息完整且高效适配模型上下文。
- 3. 系统图谱构建：**构建完整调用链拓扑，确保模型在复杂依赖中聚焦关键路径。
- 4. 标准化诊断流程：**基于机器与应用性能分析、日志解析、调用链追踪，联合决策四阶推理链，构建符合运维直觉的**SOP 流程**，快速锁定单一故障源。
- 5. 裁判Agent机制：**通过**三重保障设计**（并行诊断、裁判架构、拓扑验证），解决模型幻觉和结果波动问题，确保结论一致性与稳定性。

4.2 解决方案 – 数据预处理

解决模型在海量数据中迷失主线的关键问题。



1. 三元组数据聚合：通过**三元组架构**[对象类型]-[组件名称]-[指标名称]，将海量数据压缩为多个核心指标文件；整合 APM、Infra、日志和调用链数据，构建跨源标准化数据集。示例：pod.emailservice-0.pod_fs_writes_bytes 精确描述特定 Pod 的磁盘写操作。

2. 跨源数据整合：

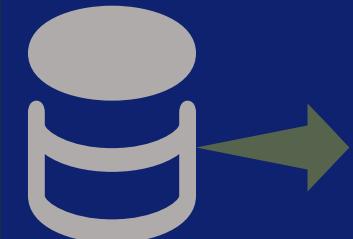
APM/Infra层：建立七类标准化指标文件，覆盖从服务实例到物理节点的多层次性能数据。

日志/调用链：关键字段提取，如状态码、下游服务、容器标识等。

4.3 解决方案 - 遥测数据动态嵌入

亮点1：遥测数据动态嵌入

- 注入核心数据，解决上下文过长导致的模型性能衰减、难以捕捉关键信息、输出结果不稳定等问题。



预处理后的多模态数据

多模态数据动态处理流程

APM/Infra数据

- 加载问题对应日期的所有APM & Infra指标数据
- 执行异常检测算法
- 计算当前故障时间段内单个组件对应单个指标的数据总数
- 根据上一步计算并聚合所有数据得到每个'组件-指标'在故障时间段内的异常次数、异常次数占比、平均值和最大值等特征

日志数据

- 加载问题对应日期所有的日志数据
- 筛选出故障时间段所有发生异常的日志数据
- 根据算法对错误信息进行截取与聚合，保留聚合得到的异常次数

调用链数据

- 加载问题对应日期所有的调用链数据
- 筛选出故障时间段所有发生异常的调用链数据
- 根据算法对错误信息进行截取与聚合，保留聚合得到的异常次数

1. RTT data for all pods, where:

- ****rrt_mean****: Represents the average latency of the current component during the failure period.
- ****rrt_max****: Represents the maximum latency of the current component during the failure period.
- ****rrt_p95_threshold****: Represents the anomaly threshold calculated using P95 (you don't need to know th

The data details are as follows:

cmbd_id	anomalies_count	rrt_mean	rrt_max	rrt_p95_threshold	proportion_of_anomalies
pod.checkoutservic-2	21	1235465.708	4151997.5	378623.905	0.913
pod.shippingservice-1	10	385601.755	1372842.0	1651.505	0.435

12. Abnormal log data during the failure period:

k8_pod	error_message	anomalies_counts
frontend-0	failed to complete the order, desc=context canceled	5
frontend-0	failed to complete the order, desc=shipping error	4
frontend-0	failed to complete the order, desc=shipping quote failure	26
frontend-1	failed to complete the order, desc=context canceled	3
frontend-1	failed to complete the order, desc=shipping quote failure	30
frontend-2	failed to complete the order, desc=context canceled	5
frontend-2	failed to complete the order, desc=shipping error	1

13. Abnormal trace data during the failure period:

cmbd_id	status_message	server	nodeName	anomalies_counts	mean_duration
checkoutservic-2	connection error: desc = "transport: Error while d	ShippingService	aioips-k8s-06	55	1.999933e+07
checkoutservic-2	connection error: desc = "transport: Error while d	ShippingService	aioips-k8s-06	13	5.551001e+07
checkoutservic-2	shipping error: shipment failed: rpc error: code = Checkou	Service	aioips-k8s-06	7	4.604563e+07
checkoutservic-2	shipping quote failure: failed to get shipping quo	CheckoutService	aioips-k8s-06	90	2.493828e+07
checkoutservic-2	timed out waiting for server handshake	ShippingService	aioips-k8s-06	29	2.000075e+07
frontend-0	HTTP status code: 500	10.233.89.225	aioips-k8s-03	35	2.792071e+07
frontend-0	context canceled	CheckoutService	aioips-k8s-03	5	5.99992e+07
frontend-0	shipping error: shipment failed: rpc error: code = Checkou	Service	aioips-k8s-03	4	3.892617e+07
frontend-0	shipping quote failure: failed to get shipping quo	CheckoutService	aioips-k8s-03	26	2.005642e+07
frontend-1	HTTP status code: 500	10.233.89.225	aioips-k8s-07	33	2.369229e+07

4.3 解决方案 - 遥测数据动态嵌入

```
2025-08-26 19:08:29.487 | INFO    | __main__:single_exc:972 - Currently processing Task 0, with Task ID 74a44ae7-81.  
2025-08-26 19:08:29.487 | INFO    | __main__:single_exc:973 - objective: Please analyze the abnormal event between 2025-06-06 01:10:04 and 2025-06-06 01:33:04 and provide the root cause.  
2025-08-26 19:08:30.620 | INFO    | __main__:cal_all_p95:702 - Complete the loading of abnormal data from APM and Infra.  
We have provided a total of 15 data items, including APM metrics and Infra metrics. The components involved include node, pod, service, TiDB, TiKV, and PD (TiDB, TiKV, and PD are all database components).  
Anomaly information during the failure period for all pods and services under APM metrics.
```

Explanation of data fields for pods and services under APM metrics:

- ****cmbd_id**:** Represents 'component type.component name' (the part before the '.' indicates whether the component is a pod or a service, and the part after the '.' specifies the name of the component).
- ****anomalies_count**:** Represents the number of anomalies that occurred during the failure period (the number of times the threshold was exceeded).
- ****proportion_of_anomalies**:** Represents the proportion of anomalous data during the failure period relative to all data within the failure time window (e.g., if there is 1 data point per minute, the failure period lasts 20 minutes, and there are 10 anomalies, the proportion of anomalies would be 0.5).

Next, we will present all processed data during the failure period:

1. RTT data for all pods, where:

- ****rrt_mean**:** Represents the average latency of the current component during the failure period.
- ****rrt_max**:** Represents the maximum latency of the current component during the failure period.
- ****rrt_p95_threshold**:** Represents the anomaly threshold calculated using P95 (you don't need to know the detailed calculation process).

The data details are as follows:

cmbd_id	anomalies_count	rrt_mean	rrt_max	rrt_p95_threshold	proportion_of_anomalies
pod.checkoutservice-2	21	1235465.708	4151997.5	378623.905	0.913
pod.shippingservice-1	10	385601.755	1372842.0	1651.505	0.435

2. Error data for all pods:

- ****error_mean**:** Represents the average number of errors for the current component during the failure period.
- ****error_max**:** Represents the maximum number of errors for the current component during the failure period.
- ****error_p95_threshold**:** Represents the anomaly threshold calculated using P95.

The data details are as follows:

cmbd_id	anomalies_count	error_mean	error_max	error_p95_threshold	proportion_of_anomalies
pod.cartservice-0	7	10.286	12	0.0	0.304
pod.shippingservice-1	8	12.000	12	0.0	0.348
pod.shippingservice-2	16	12.000	12	0.0	0.696

3. Timeout data for all pods (the field meanings are similar to the RTT and error data for pods):

The data details are as follows:

Empty DataFrame
Columns: [cmbd_id, anomalies_count, timeout_mean, timeout_max, timeout_p95_threshold, proportion_of_anomalies]
Index: []

4. RTT data for all services (the field meanings are similar to the RTT and error data for pods):

The data details are as follows:

cmbd_id	anomalies_count	rrt_mean	rrt_max	rrt_p95_threshold	proportion_of_anomalies
service.checkoutservice	21	1235465.708	4151997.5	378623.905	0.913

5. Error data for all services (the field meanings are similar to the RTT and error data for pods):

The data details are as follows:

cmbd_id	anomalies_count	error_mean	error_max	error_p95_threshold	proportion_of_anomalies
service.cartservice	8	13.5	24	0.0	0.348
service.shippingservice	10	14.4	24	0.0	0.435

4.3 解决方案 - 遥测数据动态嵌入

12. Abnormal log data during the failure period:

k8_pod	error_message	anomalies_counts
frontend-0	failed to complete the order, desc=context canceled	5
frontend-0	failed to complete the order, desc=shipping error	4
frontend-0	failed to complete the order, desc=shipping quote failure	26
frontend-1	failed to complete the order, desc=context canceled	3
frontend-1	failed to complete the order, desc=shipping quote failure	30
frontend-2	failed to complete the order, desc=context canceled	5
frontend-2	failed to complete the order, desc=shipping error	1
frontend-2	failed to complete the order, desc=shipping quote failure	23

Data field descriptions:

- Logs: The Log Data is used for subsequent text-based search and log analysis. The fields and their descriptions are as follows:
 - k8_pod: The name of the Pod.
 - error_message: Records detailed information about the error.
 - anomalies_counts: The total number of exceptions in the current group after aggregating the log data with exceptions by the columns ['cmbd_id', 'error_message'].

13. Abnormal trace data during the failure period:

cmbd_id	status_message	server	nodeName	anomalies_counts	mean_duration
checkoutservice-2	connection error: desc = "transport: Error while d ShippingService	aiops-k8s-06		55	1.999933e+07
checkoutservice-2	context canceled	ShippingService	aiops-k8s-06	13	5.551001e+07
checkoutservice-2	shipping error: shipment failed: rpc error: code = CheckoutService	aiops-k8s-06		7	4.604563e+07
checkoutservice-2	shipping quote failure: failed to get shipping quo	ChecoutService	aiops-k8s-06	90	2.493828e+07
checkoutservice-2	timed out waiting for server handshake	ShippingService	aiops-k8s-06	29	2.000075e+07
frontend-0	HTTP status code: 500 10.233.89.225	aiops-k8s-03		35	2.792071e+07
frontend-0	context canceled	ChecoutService	aiops-k8s-03	5	5.999992e+07
frontend-0	shipping error: shipment failed: rpc error: code = CheckoutService	aiops-k8s-03		4	3.892617e+07
frontend-0	shipping quote failure: failed to get shipping quo	ChecoutService	aiops-k8s-03	26	2.005642e+07
frontend-1	HTTP status code: 500 10.233.89.225	aiops-k8s-07		33	2.369229e+07
frontend-1	context canceled	ChecoutService	aiops-k8s-07	3	5.999952e+07
frontend-1	shipping quote failure: failed to get shipping quo	ChecoutService	aiops-k8s-07	30	2.005998e+07
frontend-2	HTTP status code: 500 10.233.89.225	aiops-k8s-04		29	2.786278e+07
frontend-2	context canceled	ChecoutService	aiops-k8s-04	5	6.000005e+07
frontend-2	shipping error: shipment failed: rpc error: code = CheckoutService	aiops-k8s-04		1	4.662797e+07
frontend-2	shipping quote failure: failed to get shipping quo	ChecoutService	aiops-k8s-04	23	2.005871e+07

Data field descriptions:

- Trace Information: Trace data provides detailed information about each span (sub-call) during microservice interactions, helping to identify bottlenecks and anomalies in cross-service request chains. The main fields and their meanings are as follows:

- cmbd_id: Component name.

- cmbd_id: The name of the current pod.

- status_message: Displays the details of the error.

- mean_duration: Indicates the average duration of the current call chain group.

- server: Indicates the downstream service called by the current node.

- nodeName: Indicates which node the current pod is deployed on.

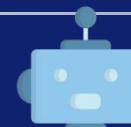
- anomalies_counts: The total number of exceptions in the current group after aggregating the traces with exceptions by the columns ['cmbd_id', 'status_message', 'server', 'nodeName'].

- Note: When analyzing this part of the data, you should focus more on the fields `status_message` and `server`, rather than placing too much emphasis on the `cmbd_id` field. This is because failure information often appears in the `status_message` field, while the faulty component is usually indicated in the `server` field.

4.4 解决方案 - 标准化根因诊断流程 (SOP)

AIOps | 2025 CCF国际AIOps挑战
赛 25 CCF International AIOps Challenge

亮点2: 标准化根因诊断流程



步骤1: APM/Infra指标分析

shippingservice-1 异常数量较多，平均延迟 1235465.7 微秒，最大延迟 4151997.5 微秒，异常占比达 91.3%。

shippingservice-1 和 **shippingservice-2** 均具有较高的错误数。

步骤2: 解析日志数据

frontend-0、**frontend-1** 和 **frontend-2** 都存在与“运费报价失败”和“上下文取消”相关的错误。其中最常见的错误是“订单未完成，原因：运费报价失败”。

步骤3: 解析调用链数据

1. **checkoutservice-2** 存在多种异常，状态信息包括“连接错误”“上下文取消”“运费错误”和“运费报价失败”。下游服务为 **ShippingService**，节点为 **aiops-k8s-06**。

2. **frontend-0**、**frontend-1** 和 **frontend-2** 存在类似错误，下游服务为 **CheckoutService**，节点分别为 **aiops-k8s-03**、**aiops-k8s-07** 和 **aiops-k8s-04**。

步骤4: 联合log和trace以及APM/Infra指标数据共同定位故障

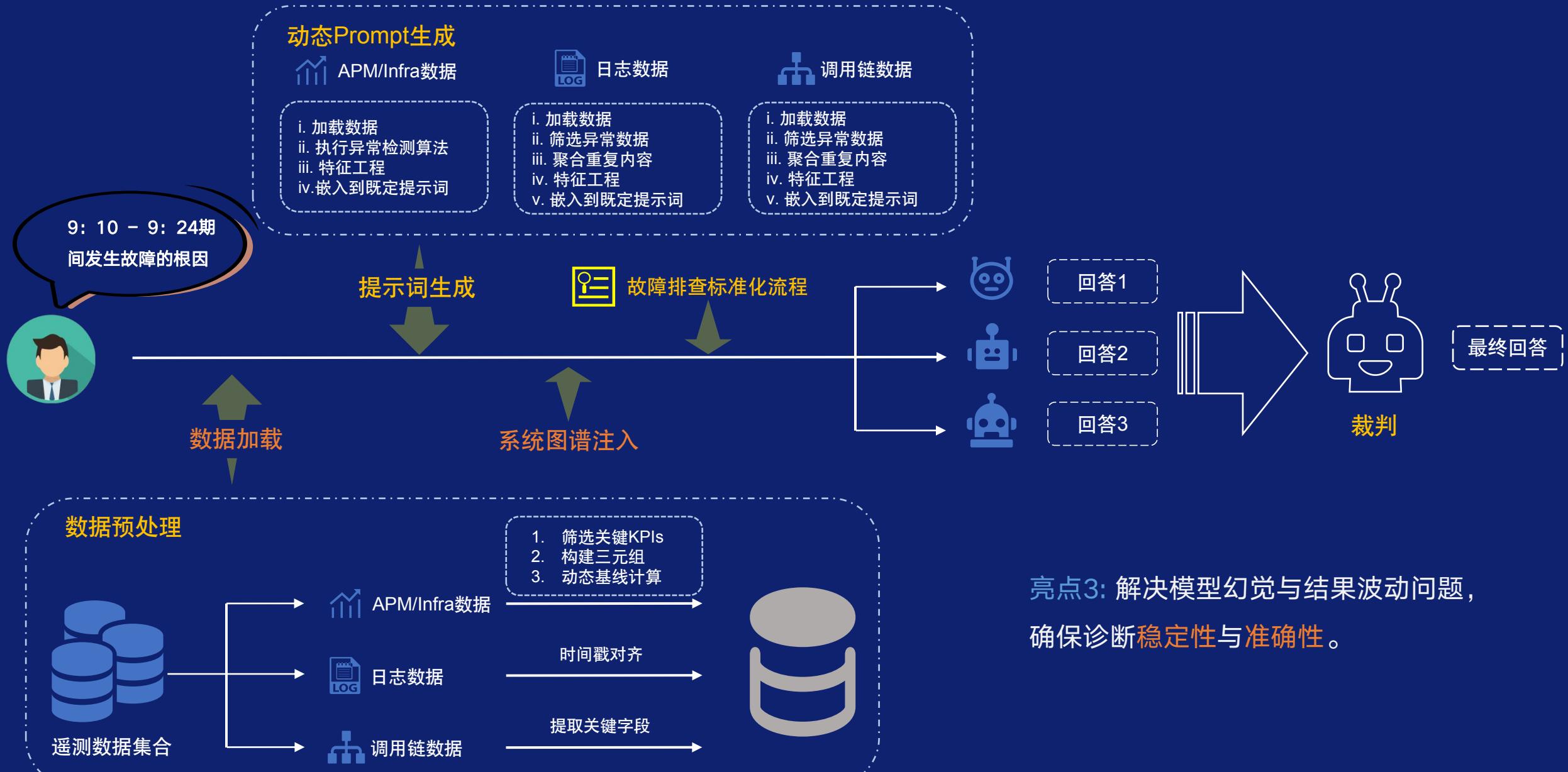
1. **下游组件识别**：主要问题是“运费报价失败”和“上下文取消”错误，源自 **frontend** 调用 **checkoutservice** 和 **shippingservice**。

2. **节点分析**：**aiops-k8s-06** 和 **aiops-k8s-07** 节点在内存和文件系统使用上有异常，但可能只是更深层问题的表现。

3. **服务分析**：**shippingservice** 和 **checkoutservice** 是主要相关服务。

4. **实例分析**：**checkoutservice-2**、**shippingservice-1** 和 **shippingservice-2** 异常最多。由于 **shippingservice** 导致 **checkoutservice** 失败，更可能是根本原因。

4.5 解决方案 - 裁判机制



4.4 解决方案 - 标准化 SOP 诊断流程

```
2025-08-26 19:54:35.105 | INFO    | __main__:single_exc:972 - Currently processing Task 0, with Task ID 74a44ae7-81.  
2025-08-26 19:54:35.105 | INFO    | __main__:single_exc:973 - objective: Please analyze the abnormal event between 2025-06-06 01:10:04 and 2025-06-06 01:33:04 and provide the root cause.  
2025-08-26 19:54:36.131 | INFO    | __main__:cal_all_p95:702 - Complete the loading of abnormal data from APM and Infra.  
To diagnose the root cause of the failure, I will follow the standardized root cause identification workflow (SOP) and analyze the provided data step by step. Here's a detailed breakdown of the analysis:
```

Step 1: Analyze APM Metrics and Infra Metrics

```
#### APM Metrics  
- **RTT Data for Pods:**  
- `checkoutservice-2` has a high number of anomalies (21) with a very high average latency (`rrt_mean = 1235465.708 ms`) and maximum latency (`rrt_max = 4151997.5 ms`). The proportion of anomalies is 0.913.  
- `shippingservice-1` also has a significant number of anomalies (10) with a high average latency (`rrt_mean = 385601.755 ms`) and maximum latency (`rrt_max = 1372842.0 ms`). The proportion of anomalies is 0.435.  
  
- **Error Data for Pods:**  
- `cartservice-0` has 7 anomalies with an average error count of 10.286 and a maximum of 12.  
- `shippingservice-1` has 8 anomalies with an average error count of 12 and a maximum of 12.  
- `shippingservice-2` has 16 anomalies with an average error count of 12 and a maximum of 12.  
  
- **RTT Data for Services:**  
- `checkoutservice` has 21 anomalies with a very high average latency (`rrt_mean = 1235465.708 ms`) and maximum latency (`rrt_max = 4151997.5 ms`). The proportion of anomalies is 0.913.  
  
- **Error Data for Services:**  
- `cartservice` has 8 anomalies with an average error count of 13.5 and a maximum of 24.  
- `shippingservice` has 10 anomalies with an average error count of 14.4 and a maximum of 24.  
  
#### Infra Metrics  
- **Node Metrics:**  
- `aiops-k8s-06` has multiple anomalies in `node_memory_MemAvailable_bytes` and `node_filesystem_free_bytes`.  
- `aiops-k8s-07` has anomalies in `node_memory_MemAvailable_bytes` and `node_filesystem_free_bytes`.  
  
- **TiKV Metrics:**  
- `tikv.tikv.io_util` has 9 anomalies with a mean value of 0.065 and a maximum of 0.066.
```

Step 2: Analyze Log Data

```
- **Log Data:**  
- `frontend-0`, `frontend-1`, and `frontend-2` all have errors related to "shipping quote failure" and "context canceled". The most frequent error is "failed to complete the order, desc=shipping quote failure".
```

Step 3: Analyze Trace Data

```
- **Trace Data:**  
- `checkoutservice-2` has multiple anomalies with status messages like "connection error", "context canceled", "shipping error", and "shipping quote failure". The downstream service is `ShippingService` and the node is `aiops-k8s-06`.  
- `frontend-0`, `frontend-1`, and `frontend-2` have similar errors, with the downstream service being `CheckoutService` and nodes `aiops-k8s-03`, `aiops-k8s-07`, and `aiops-k8s-04` respectively.
```

Step 4: Combine Logs, Traces, APM, and Infra Metrics Data to Pinpoint Catastrophic Failures

```
#### Analysis:  
1. **Downstream Component Identification:**  
- The logs and traces indicate that the primary issue is related to "shipping quote failure" and "context canceled" errors. These errors are consistently reported by the `frontend` services, which call the `checkoutservice` and `shippingservice`.  
- The `checkoutservice-2` pod has a very high number of RTT anomalies and a high proportion of anomalies, indicating severe performance issues.  
- The `shippingservice-1` and `shippingservice-2` pods also show significant anomalies in both RTT and error metrics.  
  
2. **Node-Level Analysis:**  
- `aiops-k8s-06` and `aiops-k8s-07` have multiple infra metric anomalies, particularly in memory and filesystem usage. However, these anomalies alone do not necessarily indicate a root cause, as they could be symptoms of a deeper issue.  
  
3. **Service-Level Analysis:**  
- The `shippingservice` and `checkoutservice` are the primary downstream services involved in the errors. The `shippingservice` is called by the `checkoutservice`, and the `checkoutservice` is called by the `frontend`.  
  
4. **Pod-Level Analysis:**  
- `checkoutservice-2` and `shippingservice-1` and `shippingservice-2` are the specific pods showing the highest number of anomalies. However, since the `shippingservice` is the direct downstream service causing the `checkoutservice` to fail, it is more likely to be the root cause.  
  
#### Conclusion:  
- **Root Cause Component:** `shippingservice`  
- **Root Cause Reason:** High RTT and high error rate, leading to "shipping quote failure" and "context canceled" errors.
```

The `shippingservice` is the most downstream component causing the observed failures in the `checkoutservice` and `frontend` services. The high RTT and error rates in the `shippingservice` pods (`shippingservice-1` and `shippingservice-2`) are the primary indicators of the root cause.

实际案例展示：模型推理可视化

第五章节

创新性与实用性



01

高准确率

本方案在提供的微服务架构数据集上实现了**组件**回答准确率达**60%+**（共400个任务）。

02

高推理速度

仅需**两步**即可完成根因定位，平均每条故障数据的检测耗时仅用**1分钟**。

03

高稳定性

结合定制化的**SOP流程&裁判机制**，模型推理的稳定性得到了极大的提升，最大波动仅**0.28**（49.81、49.57、49.53），幻觉问题明显减小。

04

高扩展性

本方案采用**模块化设计**，能够根据需求快速适配新场景。

05

可解释

在第一步推理结束后生成完整的**推理日志**，让结果可追溯。

06

低资源消耗

分析链路短，大幅降低了资源消耗问题。

第六章节

总结展望

本方案针对微服务环境中的数据分散、分析流程复杂、结论不稳定等问题，设计了一套高效、稳定、可解释的故障诊断框架，通过**数据预处理**、**动态 Prompt生成**、**系统图谱构建**、**标准化诊断流程及裁判机制**五大模块，全面提升诊断效率、精度与稳定性。本方案在提供的微服务架构数据集上实现了**组件回答准确率达60%+**（共400个任务）。





THANKS

主办单位：中国计算机学会（CCF）

承办单位：中国计算机学会互联网专委会、中国科学院计算机网络信息中心、中国移动研究院、清华大学

协办单位：华为2012实验室、阿里云、中兴通讯、中国移动九天团队、南开大学、西安电子科技大学、清华大学计算机科学与技术系、神州灵云