

# Multi-Agent仿生会诊 低成本、高收益根因分析框架实现

刘嘉明 中数通信息有限公司

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

第一章节 团队介绍

第二章节 核心方案实现

第三章节 实践成效与展望

# 第一节 团队介绍

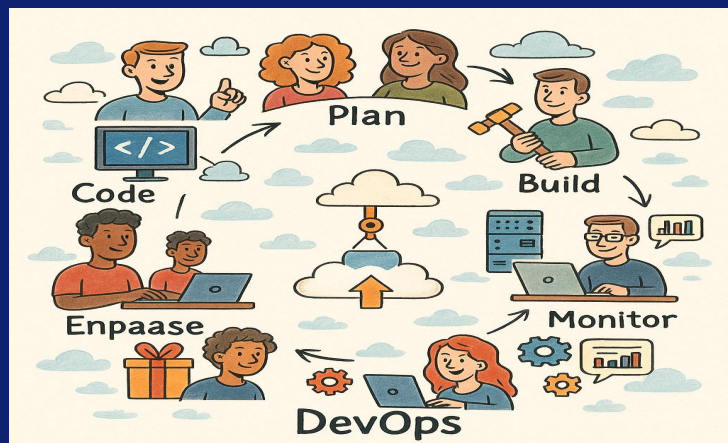
# 团队简介：运维主导、AI 辅助

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



中数通信息技术有限公司

SAP中国区**唯一**云服务运营与托管方

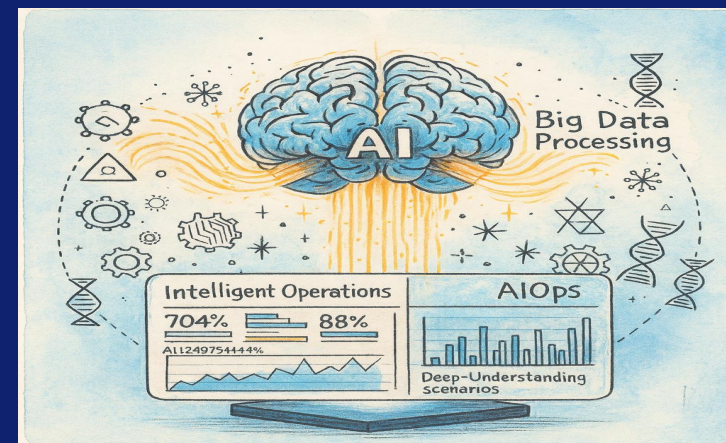


高效协同的**DevOps**模式



**多元互补**的专业岗位

AI工程师、运维专家、数据处理工程师



**跨领域**的知识融合

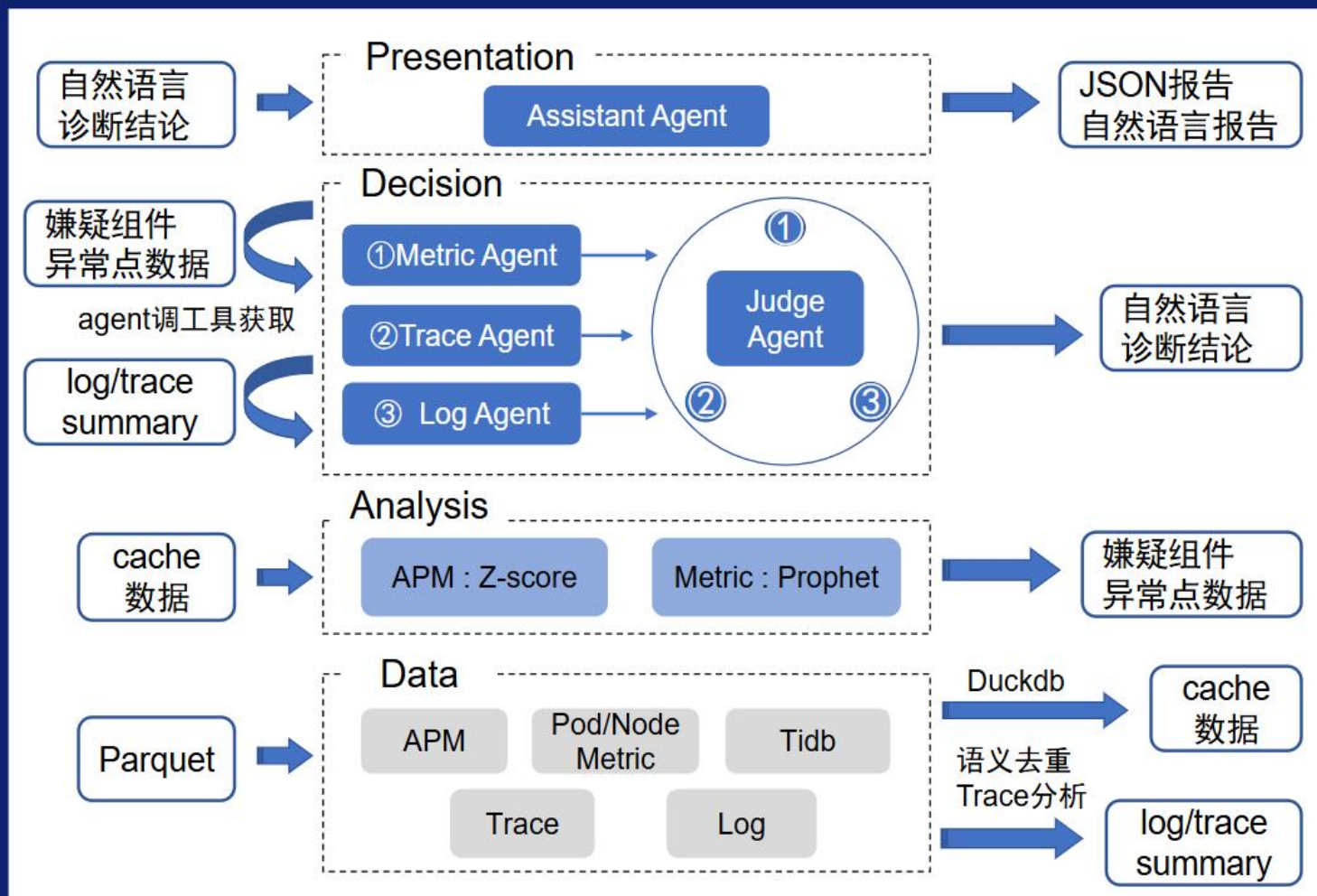
人工智能、智能运维、大数据处理

## 第二章节

# 核心方案实现



# 主流架构微创新：数据驱动的多智能体、分阶段协同架构



## 多智能体专家会诊

- 面向agent编程范式
- 多维讨论



## 分阶段诊断流水线

- 数据与业务逻辑分离
- 解耦复杂诊断任务



## 多模态数据证据链

- 深度融合四类核心数据
- 可交叉验证的完整证据链



## LLM友好的语义降噪

- 日志“语义去重”
- Trace“摘要化”处理

- 1、多维注入，让Agent发挥更稳定
- 2、整数化处理让大模型“看”得更准
- 3、Prophet实现有规律、周期不定异常数据高效检测
- 4、“体检单式”数据整合实现高效初诊
- 5、免维护、自适应的日志收敛，抑制日志风暴
- 6、面向 LLM 编程，由“对象”到“摘要”

# 微创新 1：多维注入，让Agent发挥更稳定

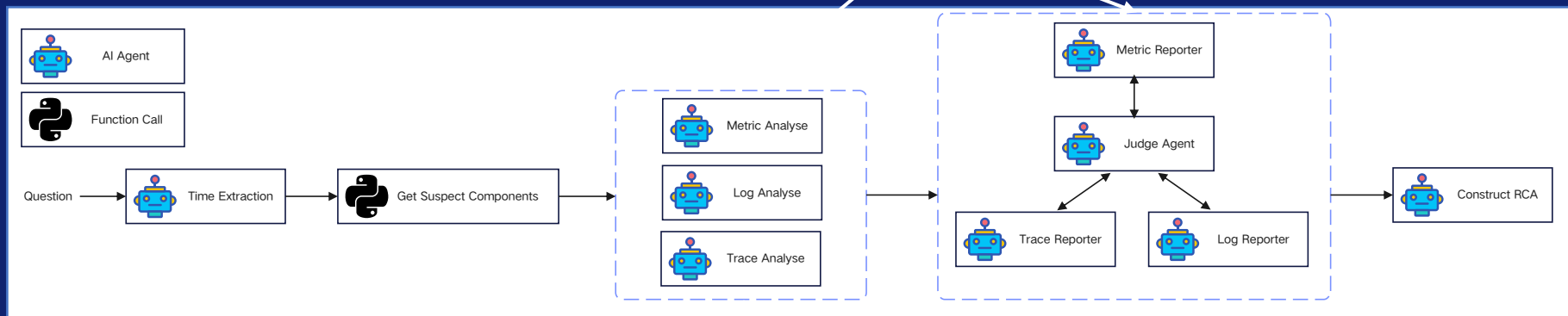
➤ **挑战：**通用模型的运维专业知识深度不足、信息的简单罗列导致模型决策困难、模型的思考路径具有随机性导致输出不稳定

## 推理约束

设计基于SOP的结构化Prompt，为Agent的分析过程提供一套标准的逻辑框架。

## 优化收益

诊断准确率提升21%、构建高置信度的推理链、确保诊断结果具有稳定性和一致性。



## 知识注入

利用知识库(RAG)，为Agent注入SRE专家知识（指标含义、故障模式等）和系统背景知识。

## 决策加权

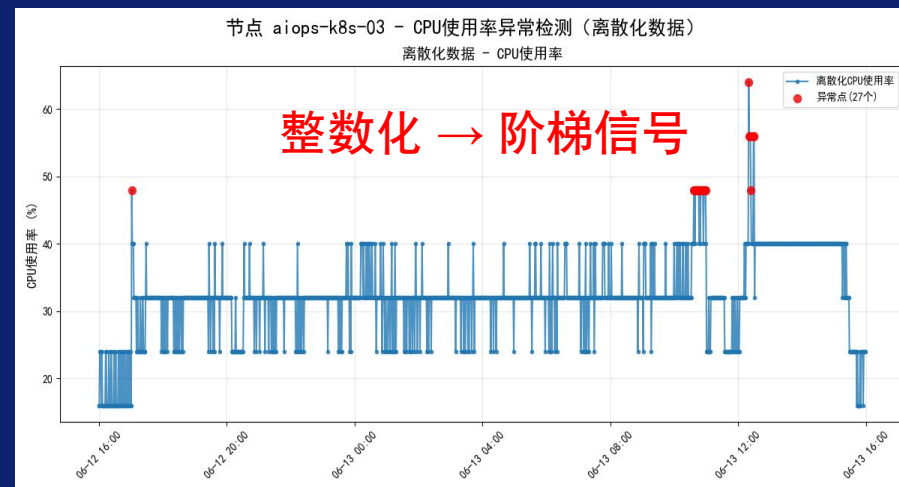
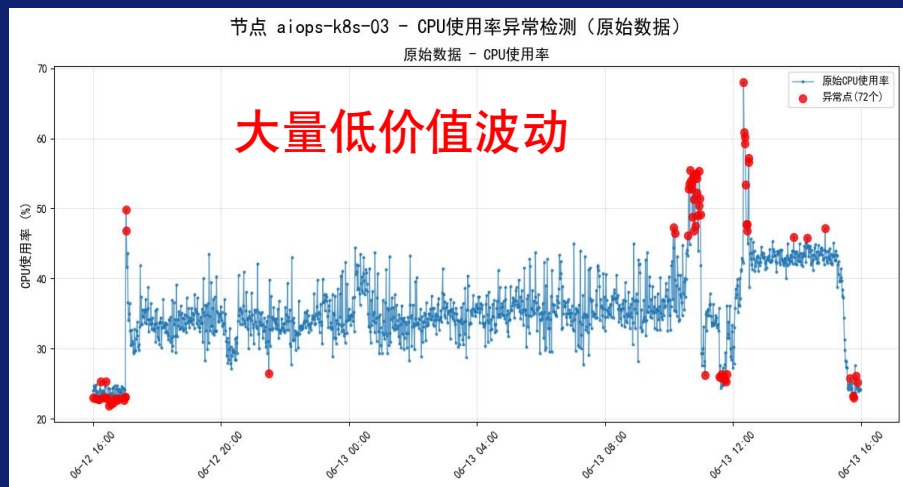
设计置信度评分机制，Agent在提交证据时，会附带量化的置信度，为决策提供权重依据。



# 微创新 2：整数化处理让大模型“看”得更准

## ➤ 挑战：

Metric浮点数据存在噪声和波动，导致检测误报率高、阈值判断模糊，难以进行有效决策。



## ➤ 创新点：

低价值连续波动指标



整数化处理



连续波动压平为稳健阶梯信号

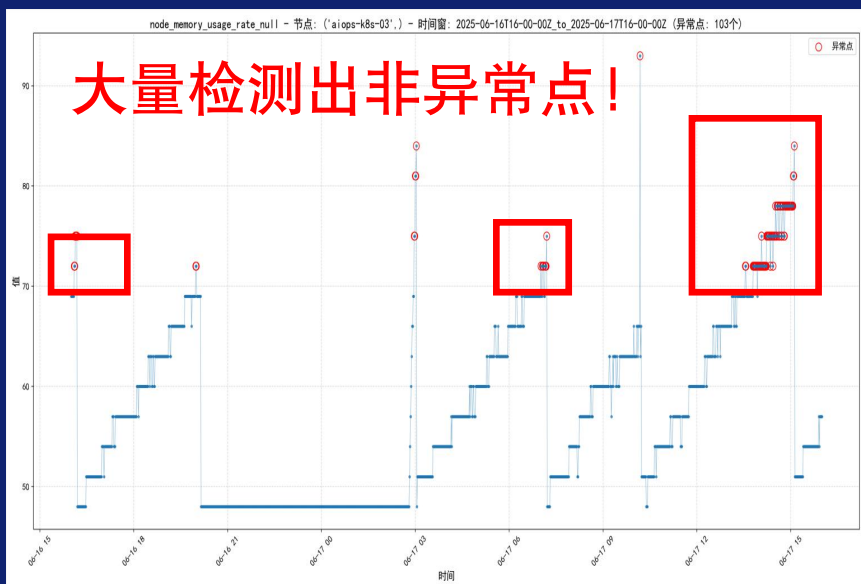
## ➤ 优化收益：

成功过滤了超过 40% 的数据噪声，将异常点检测的误报率降低了60%。

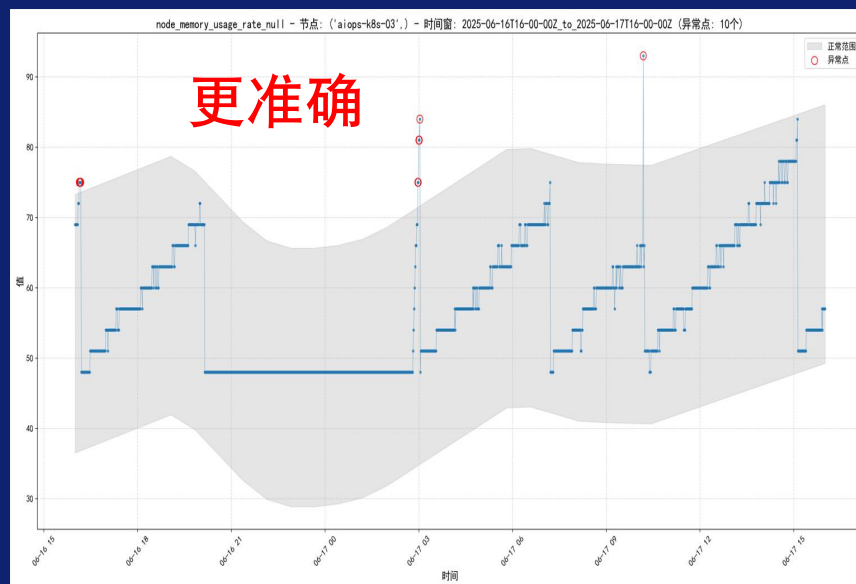
# 微创新 3: Prophet实现有规律、周期不定异常数据高效检测

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

- **挑战:** 部分数据存在固定变化趋势，但周期难以捕捉



Isolation Forest



Prophet

- **创新点:** 摒弃为不同数据匹配不同算法的孤岛式方案，使用Prophet构建**统一检测框架**。
- **优化收益:** 模型开发周期从平均1天**缩短至1小时**，检测速度**提升3倍**的同时，异常检测的**准确率提升了15%**。

# 微创新 4：“体检单式”数据整合实现高效初诊

## 挑战：

原始数据JSON/CSV繁杂冗余，大模型难以准确地抓取要点，处理效率低下且容易产生误判。

	检验项目	结果	参考值
L	二氧化碳结合力	24	21-28
L	磷	1.42	0.81-1.55
	镁	1.0	0.6-1.2
	葡萄糖	3.5	↓ 3.6-6.11
	总胆固醇	5.91	↑ 3.10-5.69
	甘油三脂	1.90	↑ 0.41-1.88
	高密度脂蛋白	2.22	↑ 1.16-1.82
	低密度脂蛋白	2.53	2.10-3.10



```
# pod异常数据
time_hm      pod      run_on_k8s_node  value  lag_value  diff      kpi_type
14:04      adservice-0      aiops-k8s-03      0.0      6.0      -6.0      client_error
14:04      checkoutservice-2      aiops-k8s-06      260.0      406.0      -146.0      request_count
14:04      productcatalogservice-2      aiops-k8s-03      2888.0      3155.0      -267.0      response_count
14:04      adservice-0      aiops-k8s-03      0.0      6.0      -6.0      error_count
14:04      checkoutservice-2      aiops-k8s-06      259.0      406.0      -147.0      response_count
14:04      frontend-0      aiops-k8s-03      1522.0      1624.0      -102.0      request_count
14:04      frontend-0      aiops-k8s-03      1518.0      1622.0      -104.0      response_count
14:04      productcatalogservice-2      aiops-k8s-03      3400.0      3712.0      -312.0      request_count
14:05      productcatalogservice-2      aiops-k8s-03      3802.0      3400.0      402.0      request_count
14:05      productcatalogservice-2      aiops-k8s-03      3231.0      2888.0      343.0      response_count
14:05      frontend-2      aiops-k8s-04      1196.0      968.0      228.0      request_count
14:05      frontend-2      aiops-k8s-04      1196.0      968.0      228.0      response_count
14:05      currencyservice-0      aiops-k8s-05      377.0      273.0      104.0      response_count
14:06      checkoutservice-2      aiops-k8s-06      396.0      270.0      126.0      response_count
14:06      currencyservice-0      aiops-k8s-05      186.0      377.0      -191.0      response_count
14:06      productcatalogservice-2      aiops-k8s-03      3922.0      3802.0      120.0      request_count
14:06      frontend-2      aiops-k8s-04      848.0      1196.0      -348.0      request_count
14:06      checkoutservice-2      aiops-k8s-06      386.0      270.0      116.0      request_count
14:06      currenc
14:06      currenc
14:06      currenc
14:06      currenc
.....

# node异常数据
time_hm      k8s_node  value  lag_value  diff      kpi_type
14:04      aiops-k8s-01      96.0      6.0      90.0      node_cpu_usage_rate
14:05      aiops-k8s-05      24.0      63.0      -39.0      node_memory_usage_rate

# service异常数据
time_hm      a
14:04      checkout
14:04      frontend 3086.0      3254.0      -168.0      request_count
14:04      frontend 3082.0      3256.0      -174.0      response_count
14:04      checkoutservice 260.0      406.0      -146.0      request_count
14:04      adservice 0.0      6.0      -6.0      error_count
.....

# tidb未找到异常数据
no suspect records
# tikv未找到异常数据
no suspect records
# pd未找到异常数据
no suspect records
# trace异常数据
no suspect records
```

## 优化效益：

- ◆ 诊断准确率提升 21%（从 17% 提升至 38%）。
- ◆ 平均推理耗时减少 42%（从 14.3 秒降至 8.3 秒）。
- ◆ 错误定位范围收敛：模型给出的可疑组件数由平均 10 个减少至 5 个。

微创新 5：免维护、自适应的日志收敛，抑制日志风暴

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

➤ 挑战：各服务日志格式异构、海量重复、信噪比极低

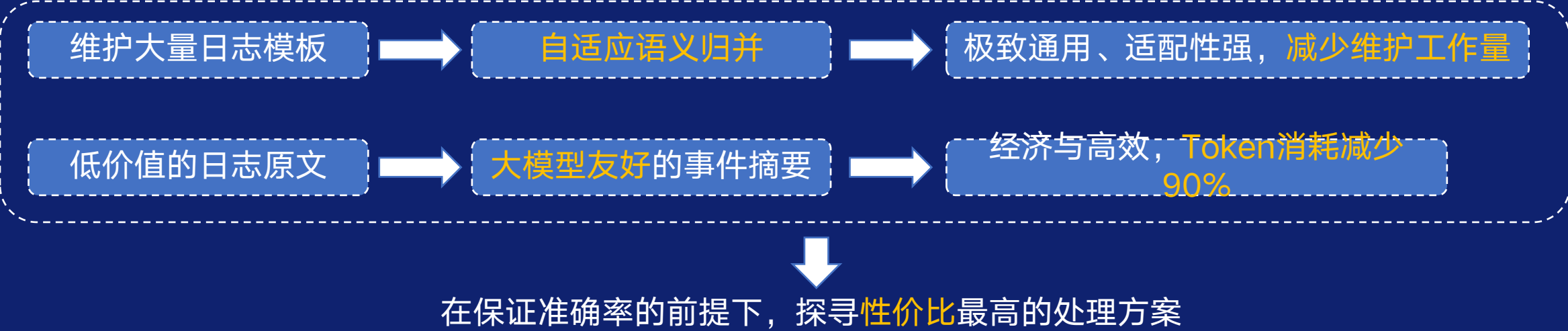
日志收敛

Service	Language
<a href="#">frontend</a>	Go
<a href="#">cartservice</a>	C#
<a href="#">productcatalogservice</a>	Go
<a href="#">currencyservice</a>	Node.js
<a href="#">paymentservice</a>	Node.js
<a href="#">shippingservice</a>	Go
<a href="#">emailservice</a>	Python
<a href="#">checkoutservice</a>	Go
<a href="#">recommendationservice</a>	Python
<a href="#">adservice</a>	Java

```
{
  "k8_namespace": "hipstershop",
  "k8_node_name": "aiops-k8s-03",
  "k8_pod": "frontend-0",
  "logs": [
    {
      "timestamp": "2025-06-06T03:10:14.267Z",
      "message": "{\"http.req.id\":\"181d15f3-a374-44f9-a6f1-f398d5a30ee0\",\"http.req.method\":\"GET\"}"
    },
    {
      "timestamp": "2025-06-06T03:10:14.268Z",
      "message": "{\"currency\":\"JPY\",\"http.req.id\":\"181d15f3-a374-44f9-a6f1-f398d5a30ee0\",\"http.req.method\":\"GET\"}"
    },
    {
      "timestamp": "2025-06-06T03:10:14.349Z",
      "message": "{\"http.req.id\":\"181d15f3-a374-44f9-a6f1-f398d5a30ee0\",\"http.req.method\":\"GET\"}"
    },
    {
      "timestamp": "2025-06-06T03:10:14.545Z",
      "message": "{\"http.req.id\":\"365a3e40-e460-4dfb-b403-da53dc09422e\",\"http.req.method\":\"GET\"}"
    },
    {
      "timestamp": "2025-06-06T03:10:14.546Z",
      "message": "{\"http.req.id\":\"365a3e40-e460-4dfb-b403-da53dc09422e\",\"http.req.method\":\"GET\"}"
    }
  ]
}
```



```
{
  "component": "frontend-0",
  "k8_node_name": "aiops-k8s-04",
  "time_window": "2025-06-19T03:10:13Z to 2025-06-19T03:22:13Z",
  "event_summary": [
    {
      "event": "{\"http.req.id\":\"181d15f3-a374-44f9-a6f1-f398d5a30ee0\",\"http.req.method\":\"GET\"}",
      "first_occurrence": "2025-06-19T03:10:14.036Z",
      "occurrence_count": 501
    },
    {
      "event": "{\"http.req.id\":\"59c51d4e-9572-4417-a249-080309913857\",\"http.req.method\":\"POST\"}",
      "first_occurrence": "2025-06-19T03:10:14.036Z",
      "occurrence_count": 201
    },
    {
      "event": "{\"error\":\"failed to complete the order: rpc error: code = Internal desc = cart failure\"}",
      "first_occurrence": "2025-06-19T03:10:16.367Z",
      "occurrence_count": 11
    }
  ]
}
```

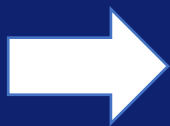


# 微创新 6：面向 LLM 编程，由“对象”到“摘要”

## ➤ 挑战：

原始数据的高Token消耗与低信息密度，不适合大模型的摘要式认知模式。

```
{
  "traceID": "be122c05acb4bdce54f6fff69185faaa",
  "spanID": "07195e77384aca05",
  "flags": 1.0,
  "operationName": "hipstershop.ProductCatalogService/GetProduct",
  "references": [
    {
      "refType": "CHILD_OF",
      "spanID": "4c72196f3436c187",
      "traceID": "be122c05acb4bdce54f6fff69185faaa"
    }
  ],
  "startTime": 1749139200377313,
  "startTimeMillis": 1749139200377,
  "duration": 14346,
  "tags": [
    {
      "key": "rpc.system",
      "type": "string",
      "value": "grpc"
    },
    ...
  ],
  "logs": [
    {
      "fields": [
        {
          "key": "message.type",
          "type": "string",
          "value": "SENT"
        },
        ...
      ],
      ...
    },
    {
      "timestamp": 1749139200391654
    }
  ],
  "process": {
    "serviceName": "frontend",
    "tags": [
      {
        "key": "node_name",
        "type": "string",
        "value": "aiops-k8s-04"
      },
      ...
    ]
  }
}
```



[#time] #service execute #operationName on  
#pod@#namespace , duration #time, called  
#callFunction Error: #errorMessage.



[2025-06-08T09:25:49.386]frontend(service) execute  
hipstershop.Frontend/Recv. on frontend-1@aiops-  
k8s-07 , duration 0.197s, called frontend.hipstershop  
Error: HTTP status code: 500.

## ➤ 优化效益：

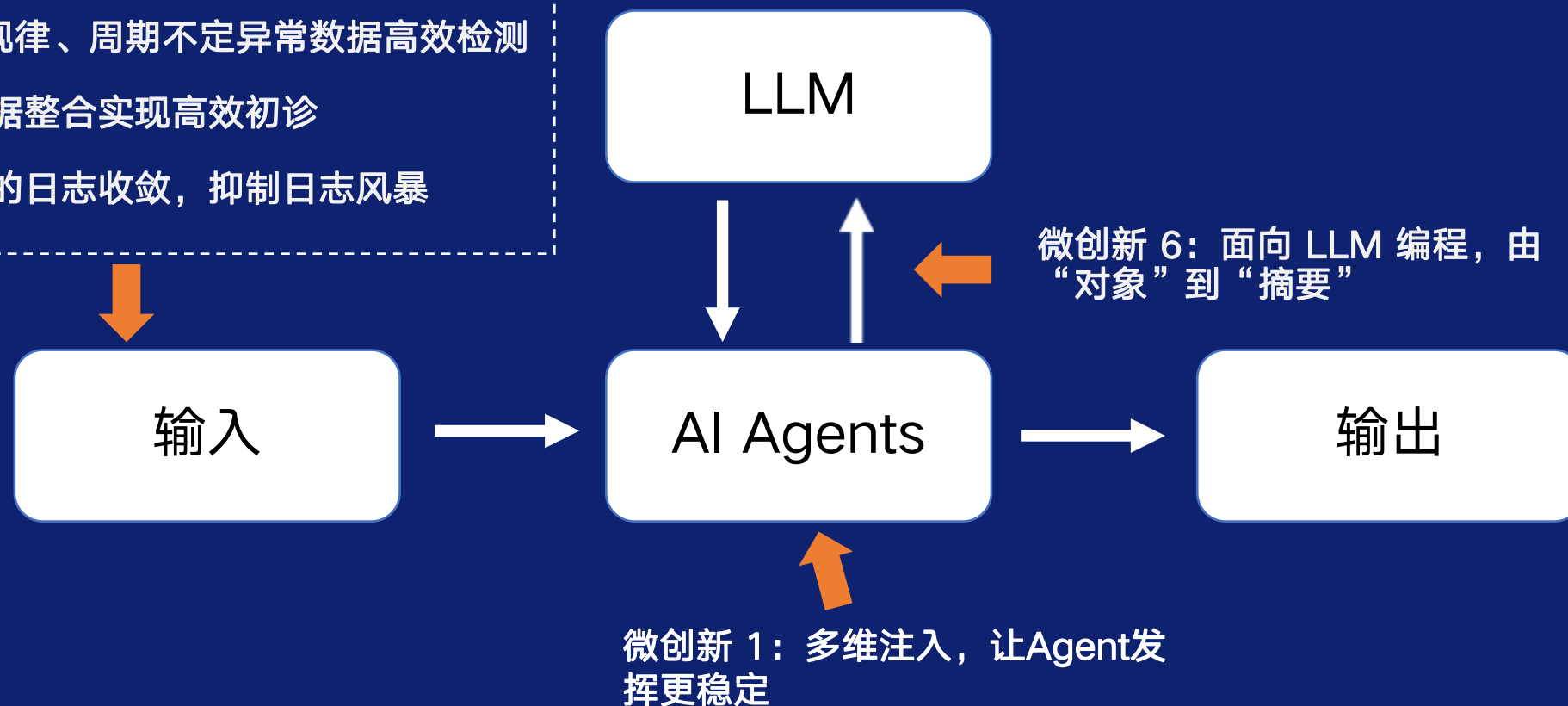
- ◆ 输入给 LLM 的 tokens 数量大幅下降，推理速度显著提升
- ◆ LLM 的注意力更集中在“有证据的可疑点”，准确率上升
- ◆ 一次上下文即可完成处理，降低幻觉和避免上下文腐蚀

## 微创新 2：整数化处理让大模型“看”得更准

### 微创新 3: Prophet实现有规律、周期不定异常数据高效检测

#### 微创新 4：“体检单式”数据整合实现高效初诊

### 微创新 5：免维护、自适应的日志收敛，抑制日志风暴

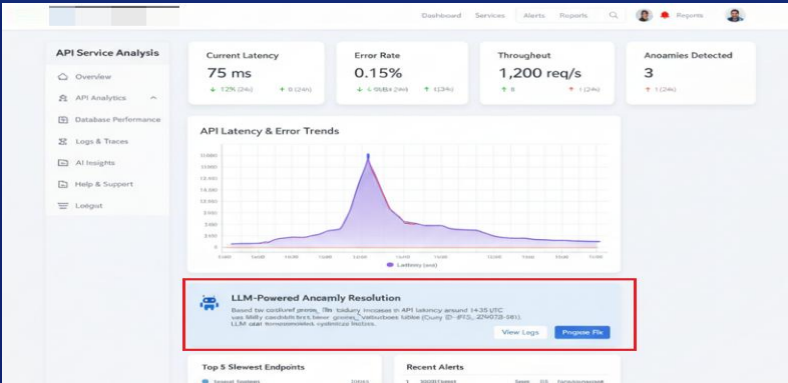




## 第三章节

# 实践成效与展望

➤ 立足当下：本方案已落地内部运维，辅助人工判断，实现了高效排查与快速定位。



应用：分析核心服务的接口延迟、错误率等模式异常。  
输出：LLM自动关联分析并输出根因判断，定位根因为“下游数据库慢查询”，缩短了故障定位与恢复时间。

➤ 跨界赋能：可复用的诊断“元流程”。只需更换特定领域的数据、工具和专家知识，就能将AIOps的成功经验快速复制到多个领域。

扩展领域	不变核心	可拔插的组件
企业经营分析	①多模态时序数据 ②通用时序分析算法库 ③五阶段智能诊断流水线 ④多智能体协作会诊机制	注入Prompt-->财务、业务、数据专家领域知识
数字营销活动复盘		注入Prompt-->渠道特性、漏斗分析经验

# 展望未来：构建平台的“大脑、记忆与感官”

强大的根因诊断工具



自主学习与全面洞察的认知智能系统



构建更聪明的大脑

构建多模型混合路由调度，在诊断流程不同阶段调用最优模型（如高数学能力的DeepSeek、强逻辑推理的Qwen），实现诊断精度与成本控制的最佳平衡。



打造会学习的记忆

引入基于RAG的动态诊断知识库，将每次成功的诊断报告结构化并自动入库，使智能体能从历史故障中学习，形成“诊断-沉淀-反哺”的自优化闭环。



赋予更立体的感官

增加多模态数据融合分析机制，让智能体在诊断时能结合监控截图、系统架构图等视觉信息进行综合推理，实现数字、文本与视觉证据的协同佐证，显著提升诊断的置信度。

OpenAIOps AIOPS | 2025 CCF国际AIOps挑战赛  
2025 CCF International AIOps Challenge

# 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