项目立项报告 · 2026年Q1

AIOps 智能运维平台项目立项

基于 Prometheus + Agent 协同的"感知 → 响应 → 决策 → 恢复"智能闭环

行业背景与市场机遇

企业运维正面临前所未有的复杂性挑战

1. **告警风暴问题严峻**

   企业平均每天产生数千条监控告警，运维团队 70% 以上时间耗费在告警分类与噪音过滤上，真正有效处理时间严重不足。

2. **MTTR 居高不下**

   行业调查显示，企业关键系统故障平均恢复时间超过 4 小时，每小时宕机损失可达数十万至数百万元。

3. **商业价值验证与自主可控**

   市场领导者 PagerDuty 的成功案例证明了事件管理平台在提升运维效率方面的巨大潜力，本项目旨在打造自主可控的同类解决方案。

4. **国产化替代需求迫切**

   在数据安全合规与降本增效双重压力下，企业对自主可控的 AIOps 解决方案需求强烈。

项目研究方向

构建"感知 → 响应 → 决策 → 恢复"AIOps 智能闭环

研究方向	核心技术	对标商业产品	预期突破
智能感知层	Prometheus + Alertmanager	Datadog APM	多维指标实时采集与异常检测
事件编排层	Agent 协同工作模式	自主智能事件管理	告警去重、分组与上下文关联
智能诊断层	Agent LLM + RAG 知识库	PagerDuty AIOps	根因分析（RCA）自动生成
自愈执行层	Ansible / K8s API	市场领先产品	故障自动修复与弹性扩容

研究核心问题：

- 如何利用 Agent 将非结构化告警信息转化为可执行的故障处置方案?
- 如何构建高质量的运维知识库（SOP）以支撑 RAG 精准检索?
- 如何在保障安全边界的前提下实现故障自愈的自动化执行?

项目整体架构

四层技术栈协同驱动智能运维闭环

感知层

- 实时监控 K8s、数据库及应用核心指标
- 触发阈值时执行告警路由与抑制规则
- 向下游推送结构化告警数据

PROMETHEUS / ALERTMANAGER

流转层

- 执行告警去重、智能分组与抑制
- 记录完整事件上下文与生命周期时间线
- Agent 协同处理告警并触发诊断流程

AGENT 协同 / 事件管理

大脑层

- Agent 调用 Tools 查询实时指标与日志
- Agent 检索 RAG 知识库，对比历史 SOP
- Agent 生成根因分析报告（RCA）与建议

AGENT 智能体 / RAG 知识库

执行层

- 将分析结果推送至即时通讯工具
- 调用 Ansible 或 K8s API 执行自愈操作
- 实现重启、扩容、流量切换等闭环响应

ANSIBLE / K8S API

核心商业价值

对标 PagerDuty，打造企业级 AIOps 核心竞争力

能力维度	传统运维模式	PagerDuty（市场标杆）	本项目目标
告警降噪	人工筛选，效率低	AI 自动去重分组	LLM 语义级降噪
故障诊断	依赖专家经验	事件时间线 + 上下文	RAG+LLM 自动 RCA
响应速度	MTTR > 4小时	MTTR 降低 60%	目标 MTTR < 30分钟
知识沉淀	Wiki 文档，难检索	Runbook 自动化	向量知识库，RAG 精准召回
自愈能力	手动执行脚本	市场领先产品	K8s/Ansible 自动执行
数据主权	数据上传第三方	SaaS 托管	完全私有化部署

效率提升

显著减少 L1/L2 运维人力投入，MTTR 从小时级压缩至分钟级，系统可用性目标 99.9%+

安全合规

完全私有化部署，满足金融、政务等行业数据安全要求，核心技术自主可控

可扩展性

模块化架构支持持续演进，支持接入更多数据源与执行引擎，适配多种云环境

项目团队与人员分工

精简核心团队，确保高效交付

角色	成员	核心职责	技术要求
项目负责人（PM）	党泽荣	项目整体规划、进度把控、跨团队协调	运维/研发背景，5年以上项目管理经验
AIOps 平台工程师	王伟, 赵光飞	监控体系部署、后端 API 开发、Webhook 集成、自动化执行模块开发	K8s, Prometheus, Python/Go, Ansible, API 开发
AI Agent 工程师	毛越民	Agent 编排、RAG 知识库构建、Prompt 工程优化	LangChain/Agent 框架, 向量数据库, RAG 技术
运维领域专家	罗辉	SOP 知识库整理、故障案例沉淀、RCA 模板设计、测试验证	丰富的运维故障处理经验、测试技能

总团队规模： **5人**

核心成员： **党泽荣, 王伟, 赵光飞, 毛越民, 罗辉**

第一阶段规划（Month 1-3）

夯实基础，打通核心链路

Month 1：基础设施搭建

- 部署 Prometheus + Alertmanager 监控体系
- 完成 Grafana 可视化仪表盘搭建，建立基线告警规则
- 搭建 Agent 协同事件管理模块，打通监控 → Agent 链路

Month 2：Agent 智能体基础版

- 部署 Agent 框架，完成基础 LLM 接入
- 设计并实现第一版告警诊断 Agent
- 构建初始运维知识库，完成 RAG 基础检索验证

Month 3：集成联调与验证

- 完成四层架构端到端集成联调
- 设计 10+ 个典型故障场景，执行故障注入测试
- 验证 RCA 报告生成质量，与人工诊断结果对比

第一阶段交付物：

可运行的 MVP 系统 + 测试报告 + 优化路线图

第二阶段规划（Month 4-6）

深化智能，实现自愈闭环

Month 4：Agent 智能诊断能力增强

- 优化 Agent 诊断逻辑，引入多轮对话诊断
- 扩充运维知识库，提升 RAG 召回精度
- 接入 ELK 日志查询，实现联合诊断

Month 5：自愈执行层开发

- 开发自动化执行引擎，基于 Agent 决策触发
- 实现 K8s 层面自愈操作：Pod 重启、HPA 扩容
- 建立执行安全边界，实现关键操作人工审批

Month 6：生产就绪与商业化

- 完成生产环境压力测试与安全审计
- 编写完整部署文档、运维手册与用户指南
- 制作产品 Demo 视频，准备商业化推广材料

第二阶段交付物：

生产级 AIOps 平台 + 完整文档 + 商业化材料

技术风险与应对策略

主动识别风险，制定预案确保项目成功

风险类型	具体风险	影响	概率	应对策略
技术风险	LLM 幻觉导致错误诊断	高	中	引入人工审核机制，高风险操作强制人工确认
技术风险	RAG 召回精度不足	中	中	持续优化 Embedding 模型，建立知识库评估体系
数据风险	运维 SOP 文档缺失/质量差	高	高	第一阶段专项投入知识库建设，引入专家审核
集成风险	第三方系统 API 变更	中	低	抽象适配层设计，降低与具体产品的耦合度
安全风险	自动化执行误操作	高	低	严格的执行权限分级，关键操作双人审批
组织风险	运维团队对 AI 决策信任度低	中	中	渐进式推广，先建议后自动化，用数据建立信任

总体风险评级：中等可控，通过分阶段交付和充分测试可有效降低风险

成功标准与关键指标

用数据衡量项目成果，确保目标可量化

指标类别	关键指标 (KSI)	当前基准 (BASELINE)	目标值 (TARGET)
效率指标	MTTR（平均恢复时间）	> 240 分钟	< 30 分钟
效率指标	告警处理时间（人均/天）	4 小时	< 1 小时
质量指标	告警误报率	> 40%	< 15%
质量指标	RCA 报告准确率	N/A（人工）	> 80%
自动化	故障自动恢复率	0%	> 60%
知识库	RAG 检索召回准确率	N/A	> 85%
可用性	系统自身可用性	N/A	> 99.9%

MONTH 3 验收节点

MVP 系统上线

完成 10+ 典型故障场景验证，MTTR 改善幅度 > 50%

MONTH 6 验收节点

生产系统正式上线

低风险故障自愈率 > 60%，全部 KPI 指标达标

总结与下一步行动

项目价值清晰，呼吁快速推进立项决策

技术先进性

融合 Agent + RAG + 自动化执行，代表 AIOps 技术前沿方向。

商业可行性

对标 PagerDuty 验证的成熟商业模式，具备显著的降本增效潜力。

战略意义

构建企业自主可控的智能运维核心能力，支撑数字化转型。

可扩展性

模块化架构支持持续演进，可扩展至安全运营等更多场景。

立即需要的决策

1

批准立项

正式启动项目，完成团队组建（目标：本月内）

2

指定负责人

确认项目 PM 及核心技术负责人

3

环境准备

协调 K8s 测试集群与 LLM API 资源

Q & A