IT 运维大模型(HAIOps)2026 年研发计划

全网监控中心第三党支部

2026-04-27

一、IT 运维大模型问题反思 2

1.问题 2

2.挑战 2

二、2026 年工作计划 2

1.运维人员的一天 2

1.1 开工-开启工作任务清单 2

1.2 流程作业-批不完、提不完也派不完的工单 4

1.3 运维作业-生产主机上 Coding 的艺术 5

1.4 异常处置 6

1.4.1 根因定位-在海量数据上抽丝剥茧 6

1.4.2 根因分析-求助开发人员揭示异常产生的本质 8

1.4.3 操作稽核-确保 AI 每次决策都安全可信 9

1.5 模型适配-信创数据库生态构建 10

2.AI+运维能力图谱 11

3.里程碑计划 15

一、IT 运维大模型问题反思

1. 问题

- **AI 赋能不足**：已建 146 个运维场景，多为简单自动化智能体，未形成异常分析闭环，一线使用意愿低。
- **产品力不足**：缺乏可全国推广的通用标杆场景，省侧部署成本高、难以快速复制。

2. 挑战

- **大模型适配难**：运维场景个性化强、容错要求高，与大模型通用性存在矛盾，输出针对性弱、人员信任度低。
- **高价值场景挖掘难**：高价值的运维场景需要基于 AI 能力特点、运维专家经验、运维场景特性等协同挖掘，而专家经验的数字化、模型化转化尚未形成标准化路径的挑战造成复杂运维场景解决方案挖掘困难。
- **结构化运维知识建模挑战**：系统数据和运维知识未完成结构化建模与沉淀，数据分散、语义缺失等挑战造成 AI 在进行异常分析时无法获取问题根因异常信息，大模型应用只能围绕简单的自动化操作展开，无法发挥大模型的应用价值。

二、2026 年工作计划

为推动运维大模型赋能生产，聚焦 AI + 运维应用、客户需求与标杆企业布局开展行业对标，明确发展机遇。当前客户核心关注异常处理，对模型的可信度、实用性、易用性与复用性需求迫切。**因此 2026 年将对标国内外优质产品，围绕日常运维提效与异常故障闭环解决两大方向，聚焦集中化运维的 AI + 运维可用性问题，打造高价值标杆产品。**

1. 运维人员的一天



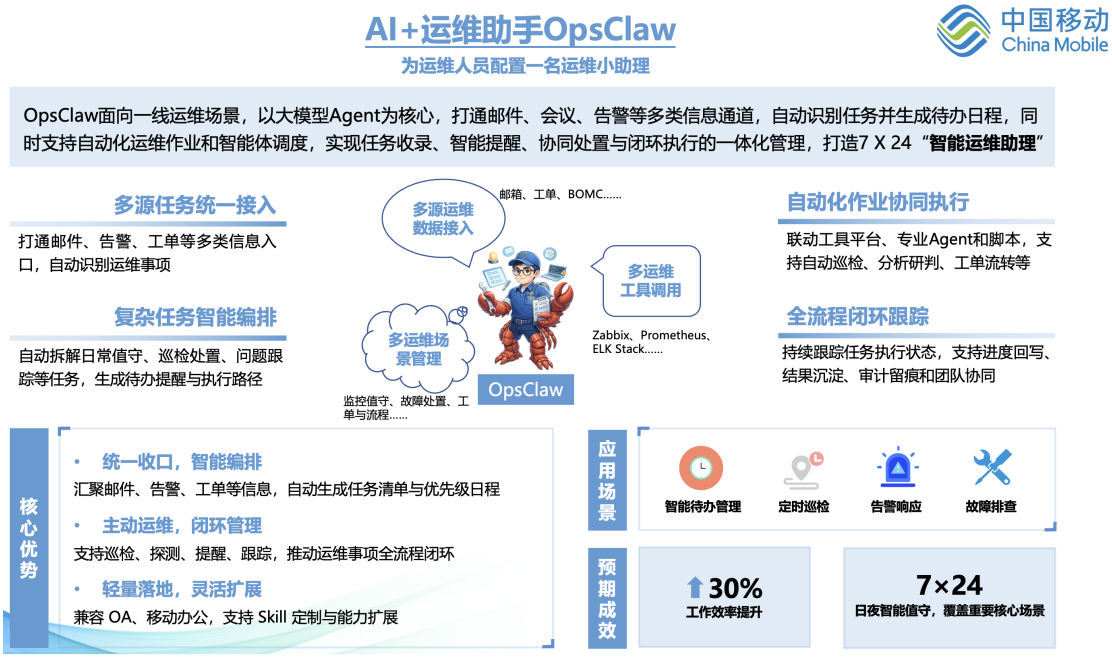
增加对比图

1.1 开工-开启工作任务清单

工作描述

- 通过邮箱、运维管理系统等渠道，梳理安全漏洞处置、审计、材料编制、日常巡检等各类待办事务，明确各项工作的紧急程度、完成时限、前置条件及对接责任人，形成任务代办清单。
- 登录运维管理系统开展日常系统巡检工作，分析巡检报告，排查系统潜在风险与安全隐患。
- 接收告警信息后，主动核查业务运行状态，确认相关服务是否正常稳定运行。

解决方案：AI+运维助手(OpsClaw)



- 运维日程管理：登录 OpsClaw 查看经 AI 汇聚后的待办事项，AI 依据任务紧急程度、协同实施难度自动完成优先级排序，并通过实时提醒机制保障运维人员高效处置，杜绝工作遗漏。
- 定时任务：自动进行巡检工作，点击查看智能分析报告，所有隐患一目了然；
  - 如需新增定时巡检任务，仅需通过指令向 OpsClaw 发起需求即可；
  - 针对复杂巡检流程规则，可依托 Skill 模块自定义巡检任务逻辑与报告解读逻辑。
- 主动查询、探测：告警信息来了，让 OpsClaw 去查询相关服务信息、发起主动探测；
  - 针对多智能体分散管理问题，可通过 OpsClaw 实现统一调度，避免多平台重复登录操作，提升运维协同效率。

业界对标

- 钉钉 AI 助理
- Motion 智能助理

- Anthropic Manage Agent

1.2 流程作业-批不完、提不完也派不完的工单

工作描述

- 批单
  - 稽核变更工单中测试案例是否完整、跨文档信息是否一致、签字审批是否合规、关联工单是否有效、图文是否一致等
  - 稽核开通工单开通需求合理、操作范围和申请权限是否一致等
- 提单：基于任务要求确定工单类型及填写字段；
- 派单：基于工单内容确定处理业务小组及处理人；

解决方案：AI+智能工单管理引擎(OpsFlow)



- 工单稽核：工单设计者可以基于稽核引擎，自主配置工单稽核规则，实现对工单内容、工单附件、关联工单的自动稽核，运维人员再也不用在复杂的附件文档、资料中比对信息，AI 直接稽核出工单问题点，并以高、中、低三类风险呈现。
- 智能填单
  - 基于用户需求智能推荐工单类型，用户通过对话形式完成工单字段填写
  - 填好的工单主动发起稽核检查，自动发现工单填写问题及时修改；
- 智能派单：根据工单内容自动转派到处理人，加速工单处理流程

业界对标

- 嘉为蓝鲸智能工单

BOMC 变更工单场景已在全网、市服、计清、平能、管信、政企六个部门上线使用，月均 311 单，稽核准确率 90.3%；正在对接广东数据修改单、河北客诉单场景；

1.3 运维作业-生产主机上 Coding 的艺术

工作描述

- 主机作业：需熟练掌握各类 Linux 操作命令，在生产主机上完成 Shell 命令的编写与执行，保障生产环境相关操作的顺利开展。
  - 若对命令记忆不熟练，可通过两台设备配合操作：一台设备连接公网查询相关命令，对照查询结果逐字符在生产终端中输入执行；
  - 针对复杂流程操作，需先在离线模拟环境中完成脚本编写与全面测试，测试通过后，通过 FTP 工具将脚本上传至生产主机执行；若执行过程中出现错误，需返回模拟环境重新调试测试，直至脚本无误后，再次上传至生产主机执行。
- 数据分析：需熟练掌握各数据库的模型结构，编写 SQL 语句完成数据库信息的查询与数据修改操作，确保数据处理的准确性与合规性。
  - 若对数据库模型或 SQL 语句编写不熟练，需求在测试主机上完成 SQL 语句的编写与测试，测试通过后通过 FTP 工具上传至生产主机执行；若执行过程中出现错误，需重新梳理操作流程，在测试主机上排查问题、优化语句，直至测试通过后，再次执行上传操作。

解决方案：AI+运维命令控制台(OpsCLI)



- 不会的命令大模型自动生成，再复杂的命令，也无需担心，生成准确率≥93%
- 大模型自动检查 shell，错误的 shell 命令自动检查修正；
- 长流程脚本、sql，大模型自动生成并在沙箱环境中完成验证，下发主机执行；
- 运维专家复杂分析操控逻辑沉淀为 Skill，实现复杂任务自动分析处理；
- 安全
  - 敏感操作自动屏蔽，即便是隐藏在脚本中，大模型依然可以识别；
  - 终端回显敏感数据，大模型基于金库规则自动脱敏；
- 与 4A 集成实现触达运维人员的统一 AI 门户
- 基于 xterm.js 插件开发可以适配公司和业界所有的 webshell 终端

业界对标

- DataDog 智能终端
- Google Gemini-CLI
- Github Copilot-CLI

已经完成一阶段开发目标，操作生成准确率达 97.7%，无论是外协厂家和局方人员都很喜欢这个产品。

1.4 异常处置

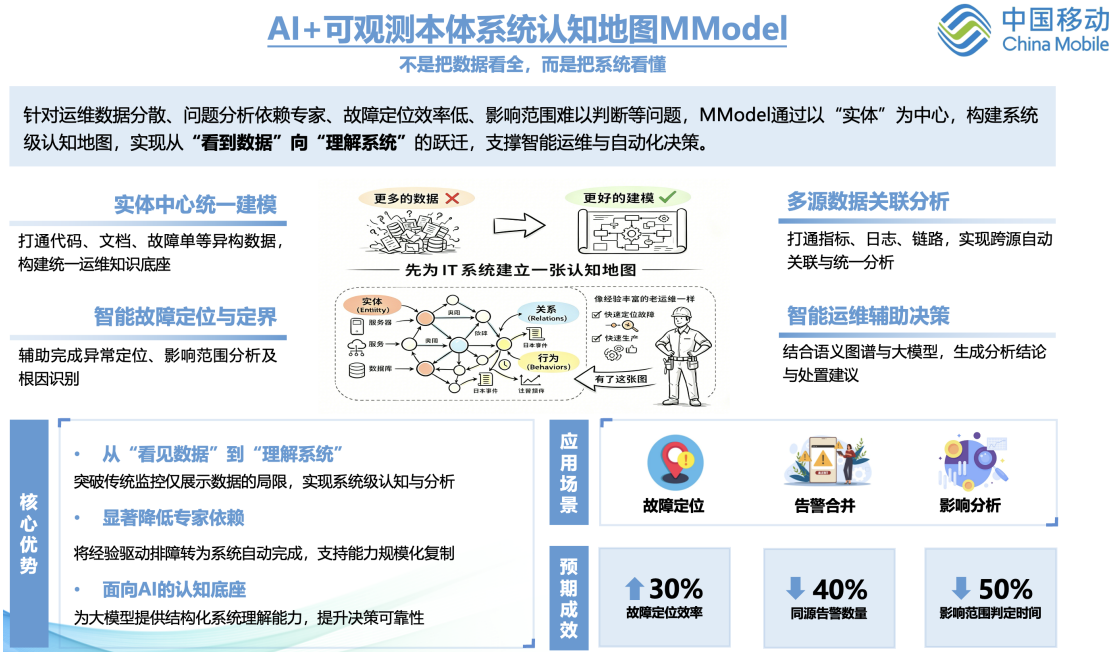
1.4.1 根因定位-在海量数据上抽丝剥茧

工作描述

收到故障告警，运维专家需登录不同的系统，基于自己对系统业务的调用逻

辑理解，从不同的系统中串联相关的 metric、log、trace 数据，在告警噪声中还原异常链路日志，寻找关键故障信息。

解决方案：AI+可观测本体系统认知地图（MModel）



基于 IaaS、PaaS、SaaS 各层采集的指标（Metrics）、日志（Logs）、调用链（Traces）、事件（Events）等可观测数据，以本体实体、实体关系、观测数据抽象、数据归属关系及模型为核心研究对象，构建基于可观测数据的语义层，为业务系统运维团队绘制一张可理解、可查询、可推理的系统“认知地图”，提供了一套通用的可观测"交互语言"，让人、程序和 AI 都能够理解和分析可观测数据。

解决传统监控系统中指标、日志、链路数据分散在不同系统、查询方式割裂、依赖人工拼接的问题，实现跨数据源的统一语义建模与统一查询能力。把当前“到处找数据”的监控方式，升级为“围绕一个业务对象统一查看”的方式。

- 数据互联：通过关联关系构建数据图谱和实体图谱
  - 日志与指标通过实体关联，实现日志异常与指标异常的关联分析。
  - 链路与日志通过 TraceID 关联，实现分布式追踪与详细日志的融合。
  - 事件与实体关联，实现告警事件与系统状态的统一视图。
- 服务拓扑自动发现：服务间调用关系的自动发现和拓扑构建
  - 调用关系发现：基于链路追踪数据自动识别服务间调用关系。
  - 动态拓扑更新：实时更新服务拓扑，反映系统架构变化。
  - 多层次拓扑：支持服务级、实例级、基础设施级等不同粒度的拓扑视图。
- 智能故障根因定位：基于实体关系图的智能故障分析能力

- 多维度关联分析：结合指标异常、日志错误、链路异常进行综合分析。
- 时空关联分析：基于时间和空间维度分析故障传播路径。

业界对标

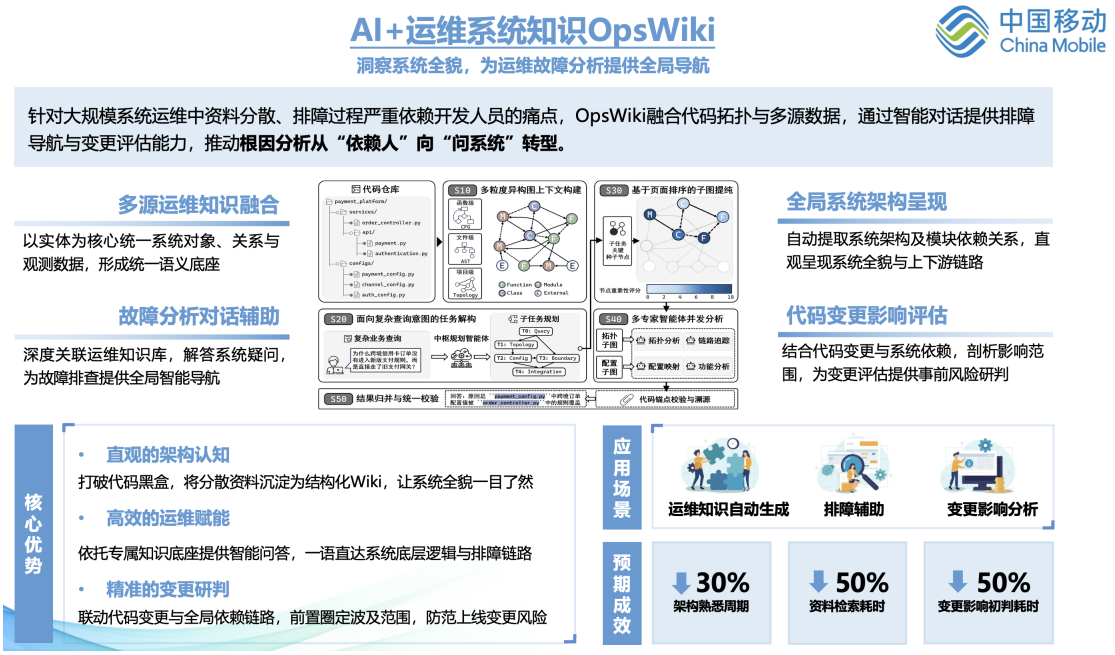
- 阿里云云监控 2.0（UModel）
- AWS DevOps Agent

1.4.2 根因分析-求助开发人员揭示异常产生的本质

工作描述

由于运维人员缺乏对系统架构、关键业务时序、内部代码逻辑的了解，一旦定位完成故障定位，找到故障关键错误日志。运维只能进行重启、切换、熔断的紧急操作。故障根因必须叫上开发人员协同分析，才能给出最终故障解决方案。

解决方案：AI+运维系统知识（OpsWiki）



使用大模型的系统代码进行结构化语义解析，构建全局统一的代码拓扑网络。在此基础上，深度结合系统部署配置、现网运维手册等，构建跨代码仓库的系统联动与映射关系，实现系统架构、关键业务时序、模块依赖关系的系统理解能力。为运维人员和 LLM 驱动的智能应用提供系统理解能力，解决目前故障定位依赖开发人员，LLM 异常分析缺乏代码层的语义知识的问题。

- Ask AI：运维人员可以通过对话的形式从系统实时代码中获取系统的系统架构、模型说明、关键业务时序逻辑、功能实现的运维知识，辅助用户了解系统知识，进行故障根因分析
- 系统变更影响分析：当业务代码迭代、系统部署配置信息发生变更时，系统

能自动识别差异，并沿着代码逻辑进行推演，定界出改动会波及的其他联动仓库、模块和功能边界。

业界对标

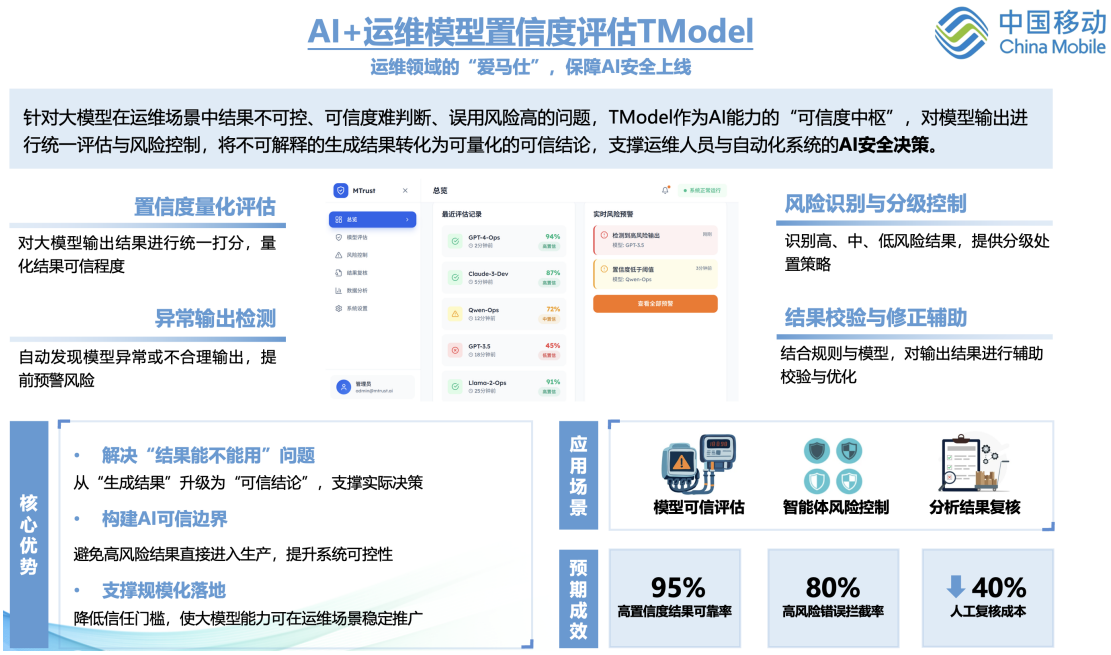
- DeepWiki

1.4.3 操作稽核-确保 AI 每次决策都安全可信

工作描述

运维人员对通用大模型给出的解决方案操作命令极其谨慎，仔细推敲大模型给出的每一个操作建议，防止误操作。

解决方案：AI+运维模型置信度评估(TModel)



运维大模型置信度评估能力旨在为运维场景下的大模型交互输出提供可信度量化与决策支持，核心表现为"一次 LLM 调用 = 一个答案 + 一个'可信程度'"的范式。通过结合大模型输出自带的 logprob 和运维领域的经验认知构建置信度评分机制，为每个模型输出附加可量化的可信度指标，使运维人员能够基于可信程度自动决定如何使用该答案，置信度高的结果可直接应用于运维决策，置信度中的结果可触发二次验证或人工完善，置信度结果低的结果则被自动过滤或标记为不可用。

- 智能体/Claw 运维操作：对模型即将执行的操作给出置信度，引导用户二次验证低置信度操作动作，降低误操作概率

- 运维模型分析结果：对运维模型分析的结果进行置信度评估，辅助专家快速识别分析结果可靠性。例如，工单智能稽核可靠性评估，运维告警分析可靠性评估。

业界对标

- Probable

1.5 模型适配-信创数据库生态构建

工作描述

数据库运维人员日常运维工作需要登录数据库主机，通过一些列的工具命令完成作运维作业。去年项目已在磐维数据库上完成核心能力验证，实现典型故障场景的结构化排障路径生成，及自主定位和处置。

解决方案：AI+信创数据库排障大模型(DModel)



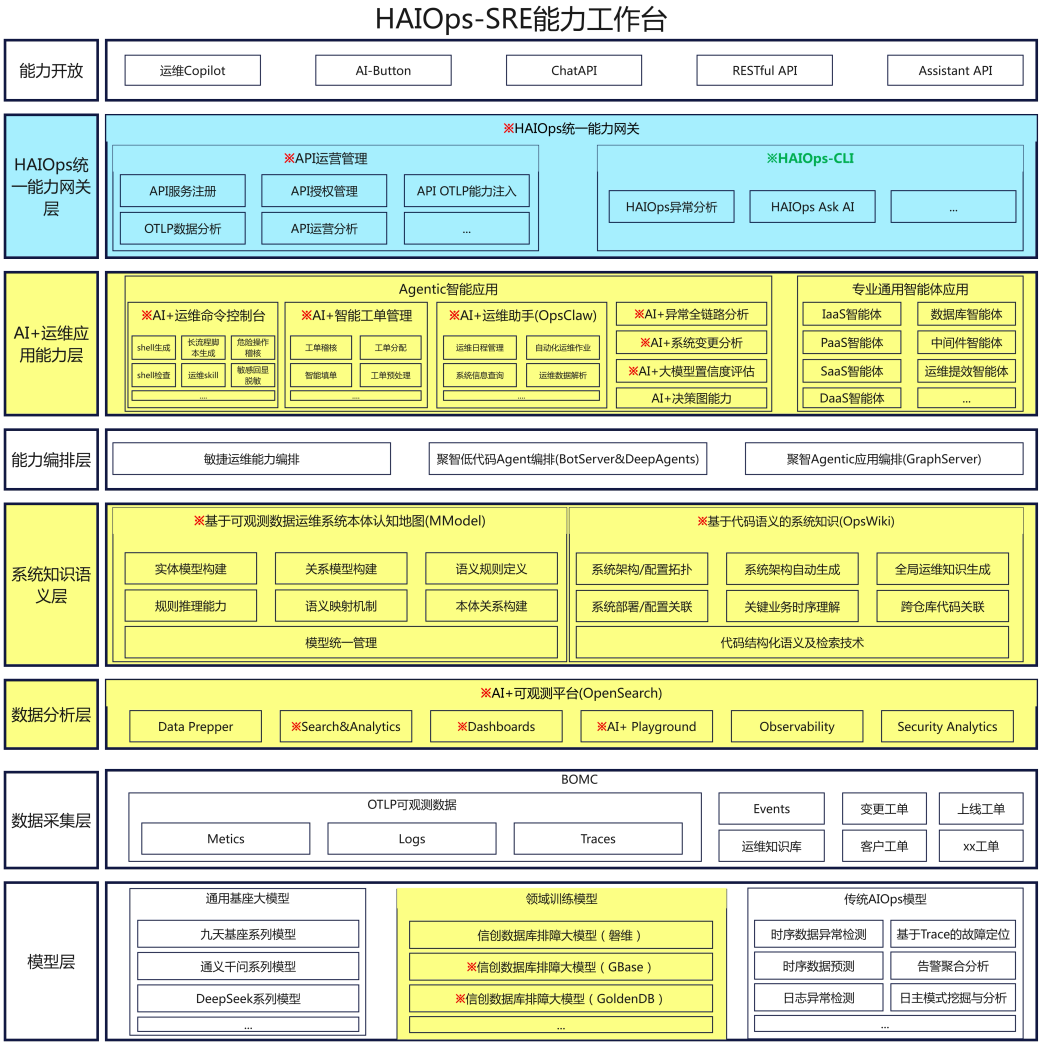
今年重点从“单一数据库能力验证”升级为“跨数据库体系推广能力”，解决传统排障模型高度依赖具体数据库实现、难以复制推广的问题。通过抽象通用排障逻辑与统一运维图结构，将数据库相关能力从“具体产品适配”升级为“方法论级复用”，实现：

- 将磐维排障模型能力推广至 GBase、GoldenDB 等信创数据库
- 抽象数据库无关的通用诊断逻辑
- 支持快速适配不同数据库架构差异
- 降低数据库替换与迁移过程中的运维风险

业界对标

- Datadog SRE
- 微软 Azure SQL Copilot
- 字节跳动-DB 智能助手（DBCopilot）
- 贝联珠贯 SREAgent

2. AI+运维能力图谱



26 年 IT 运维大模型团队瞄准解决一线运维人员觉得简单智能体不好用、用处不不大的问题，深挖问题根因和一线运维人员的使用场景，制定了 26 年以 Agentic 应用研发为主线的研发方向。

1. **为了解决简单智能体无法理解系统知识，无法感知、推理系统异常根因，研发系统知识语义层：**参考阿里云 UModel 构建了**基于可观测数据运维系统本体认知地图(MModel)**，为大模型进行异常分析时精确提供可观测数据，使得大

模型可以按图索骥。参考 DeepWiki，**基于系统代码自动生成系统架构、关键业务时序等系统知识(OpsWiki)**，解决故障分析依赖系统专家、开发人员的问题。有了这层系统知识语义层能力后，大模型针对运维问题再也不是盲人摸象，可以更好的进行系统异常分析、系统变更分析。

2. **针对一些运维人员在生产环境无法使用大模型服务，运维人员编写 shell 命令、长流程脚本麻烦的痛点，研发 AI+运维命令控制台(OpsCLI)**：无需在被控主机上安装任何服务，用户就可以基于自然语言生成操作 shell 命令、长流程脚本，对终端错误信息进行分析，将自己的日常操作沉淀为 skill 智能调用。而且可以实现一直困扰安全的敏感操作稽核以及终端回显数据的金库脱敏。
3. **针对运维人员日常大批量的运维工单处理的痛点，研发发了 AI+运维工单能力(OpsFlow)**：目前运维人员可以自主以自然语言的方式编排稽核规则，实现工单数据的多模态稽核、跨文档稽核等稽核能力。

4.针对运维人员作业入口多、自动化巡检、主动探测缺乏便捷通道，运维工作繁杂的问题，研发 AI+运维助手(OpsClaw)：实现运维日常管理、自动化运维作业、运维数据查询等通用运维助手。

围绕运维 Agentic 场景应用标品的构建，形成 HAIOps-SRE 能力工作台。这次在开发这些应用能力的时候，团队完全采用 W3C 的 OTLP 标准来开发所有的服务，如此就可以将所有服务的可观测数据自动串联在一起。之后就可以构建一个类似于 Claude Code 的 CLI 智能体 **HAIOps-CLI**，实现对 SRE 工作台自身的异常智能分析，**探索业界前沿的 AgenticOps 自治系统运维实践。**

场景维度	能力名称	运维痛点	价值	应用场景	对标爆品
日常运维作业提效	AI+运维命令控制台(OpsCLI)	1. 日常运维作业终端操作无法便捷使用大模型，命令、脚本编写、日志分析效率低； 2. 无法在终端中使用运维工作中积累的专家经验和系统分析策略；	1. 建立触达一线生产运维人员的统一AI门户； 2. 实现shell生成、shell检查、长流程分析脚本自动生成提升运维效率； 3. 实现敏感操作过滤、生产终端敏感信息脱敏等生产安全控制	1. 4A智能软堡垒 2. 运维日常生产终端作业	DataDog智能终端； Google Gemini-CLI； Github Copilot-CLI；
	AI+智能工单管理(OpsFlow)	工单种类多、流程复杂，填写/分派/流转效率低，占一线运维工作比重大；	1. 实现基于用户个性化需求的自动工单稽核，提升工单稽核的效率和准确率； 2. 实现工单智能填写、推荐、分派、流转，赋能工单全流程，提升处理效率；	1. 工单稽核 2. 工单填写 3. 工单转派	嘉为蓝鲸智能工单
	AI+运维助手(OpsClaw)	1. 运维任务分散在多平台，无智能提醒与协同，工作效率低、易遗漏； 2. 运维自动化作业、信息查询需要登录不同系统，流程繁琐； 3. 自动化智能体需要专门对话通道，不便使用；	1. 建立运维系统数字分身，实现系统信息查询、数据分析、主动探测、定时任务、调度专业智能体统一入口，提升运维效率； 2. 统一管理运维人员作业任务，自动准备作业资料和日程智能体管理，提升运维效率；	1. 运维日程管理 2. 系统运维信息查询、分析 3. 定时任务、主动探测作业	钉钉AI助理； Motion智能助理； Anthropic Manage Agent；
系统语义知	AI+可观测	1. 寻找故障异常的真实异常、错误日志是整	实现以实体为中心的统一观测，一键查看系统全量	1.数据互联	阿里云云监控2.0

识建立	本体系统认知地图(MModel)	个故障定位的难点，仅靠监控告警信息和故障描述难以提供有效的故障分析信息； 2. 传统监控观测数据分散、查询割裂，人、AI、智能体都无法理解系统结构；	观测数据，为智能体提供内生系统认知；结合Metrics/Logs/Traces 三维数据，实现故障定位和根因推荐。	2.服务拓扑自动发现 3.智能故障根因定位	（UModel）； AWS DevOps Agent；
	AI+运维系统知识(OpsWiki)	运维系统维护依赖开发专家，无结构化知识沉淀，故障分析无全局导航	自动生成运维系统 Wiki 文档与架构图，支持对话式故障分析，降低对专家的依赖	1.系统知识 Ask AI 2.系统变更影响分析	DeepWiki
已建能力优化	AI+模型置信度能力评估(TModel)	运维大模型输出无置信度评估，一线人员对结果缺乏信任，易引发生产风险	实现模型输出置信度精准计算，降低错误使用风险，推动大模型安全落地	1.智能体/Claw 运维操作 2.运维模型分析结果	Probable
	AI+信创数据库排障大模型(DModel)	去年基于磐维数据的信创数据库排障大模型能力向其他数据库适配迁移	重点从“单一数据库能力验证”升级为“跨数据库体系推广能力”，解决传统排障模型高度依赖具体数据库实现、难以复制推广的问题。通过抽象通用排障逻辑与统一运维图结构，将数据库相关能力从“具体产品适配”升级为“方法论级复用”	1.将磐维排障模型能力推广至GBase、GoldenDB 等信创数据库	Datadog SRE； 微软 Azure SQL Copilot； 字节-DBCopilot； 贝联珠贯 SREAgent

3. 里程碑计划

序号	场景名称	难易程度	优先级	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
1	AI+运维助手(OpsClaw)	★	P2	需求细化、技术预研		技术原型+产品设计		产品开发	上线集成	运营优化					
2	AI+智能工单管理(OpsFlow)	★★	P0	关键场景能力开发优化		工单规则引擎、原子AI能力研发		稽核场景编排能力开发				运营优化			
3	AI+运维命令控制台(OpsCLI)	★★	P0	技术原型+产品设计		产品开发+算法优化		上线集成	运营优化						
4	AI+可观测本体系统认知地图(MModel)	★★★★	P0	基于本体论的可观测性系统开发			技术原型+产品设计		产品开发+算法优化			上线集成			运营优化
5	AI+运维系统知识(OpsWiki)	★★★★★	P2	需求细化、技术预研			技术原型+产品设计		产品开发+算法优化			上线集成		运营优化	
6	AI+运维模型置信度评估(TModel)	★★★★	P2	需求细化、技术预研			技术原型+产品设计		算法开发、模型训练			上线集成		运营优化	
7	AI+信创数据库排障大模型(DModel)	★★★	P1	需求细化、知识整理			技术原型+产品设计		产品开发+算法优化			上线集成		运营优化	