

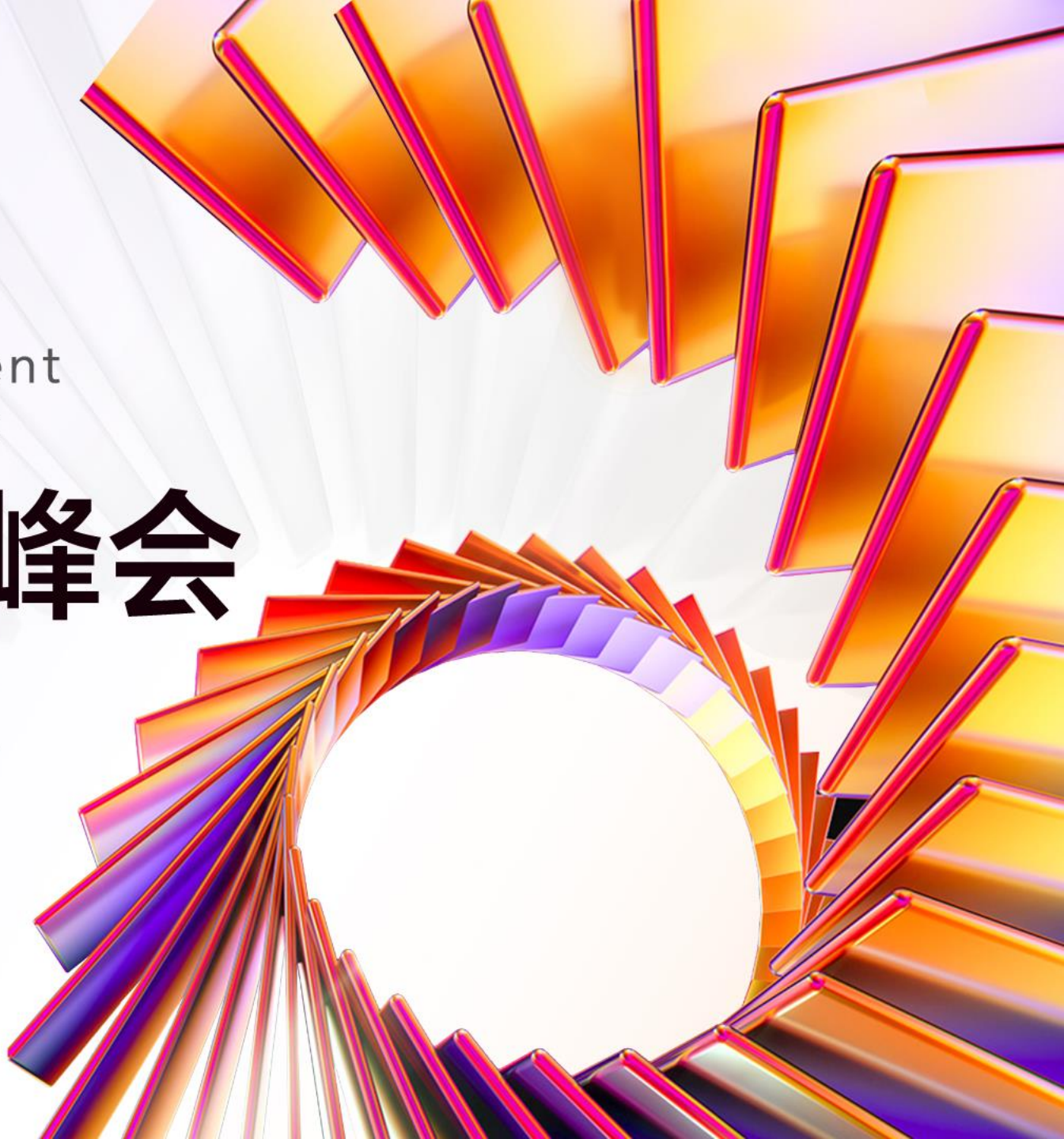


第8届 AI+ Development Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发

11月14-15日 | 深圳





2026 AI+ PRODUCT INNOVATION SUMMIT

01.16-17 · ShangHai

AI+产品创新峰会



Track 1: AI 产品战略与创新设计

从0到1的AI原生产品构建

论坛1: AI时代的用户洞察与需求发现

论坛2: AI原生产品战略与商业模式重构

论坛3: Agentic AI产品创新与交互设计

2-hour Speech: 回归本质



用户洞察的第一性

——2小时思维与方法论工作坊

在数字爆炸、AI迅速发展的时代，
仍然考验“看见”的“同理心”



Track 2: AI 产品开发与工程实践

从1到10的工程化落地实践

论坛1: 面向Agent智能体的产品开发

论坛2: 具身智能与AI硬件产品

论坛3: AI产品出海与本地化开发

Panel 1: 出海前瞻



“出海避坑地图”圆桌对话

——不止于翻译:

AI时代的出海新范式



Track 3: AI 产品运营与智能演化

从10到100的AI产品运营

论坛1: AI赋能产品运营与增长黑客

论坛2: AI产品的数据飞轮与智能演化

论坛3: 行业爆款AI产品案例拆解

Panel 2: 失败复盘



为什么很多AI产品“叫好不叫座”?

——从伪需求到真价值:

AI产品商业化落地的关键挑战

智能重构产品 数据驱动增长
Reinventing Products with Intelligence, Driven by Data

80+

业界大咖

汇聚产品大咖的真知灼见

40+

创新案例

开拓全新AI+产品视角

10+

分论坛

涵盖产品到商业增长的全链路

800+

听众规模

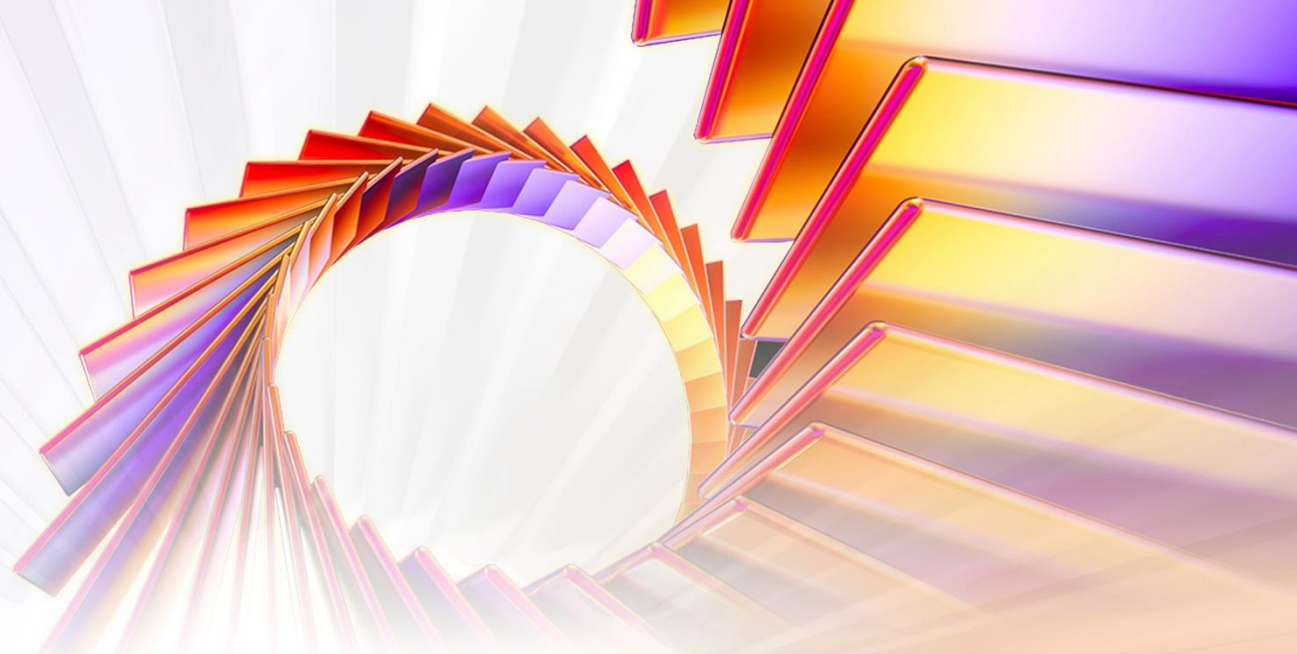
共同探索产品的智能重构



扫码查看会议详情

AI+DD 8th 2025 | 11月14-15日 | 深圳

第8届 AI+ Development
Digital Summit
AI+研发数字峰会
拥抱AI 重塑研发



智算集群故障诊断算法研究与实践

陈文潇 | 华为技术有限公司



陈文潇

华为天才少年 智能体软件专家

华为天才少年，智能体软件专家，清华大学NetMan实验室博士，一直从事AIOps，人工智能，网络自动运维等工作。在WWW、INFOCOM、FSE等国内外学术会议上发表演讲，担任SIGMETRICS, AAAI, WWW等多个国际会议期刊审稿人。担任openFuyao核心技术负责人，与科大讯飞、蚂蚁集团、清华大学共同打造昇腾生态竞争力。

目录

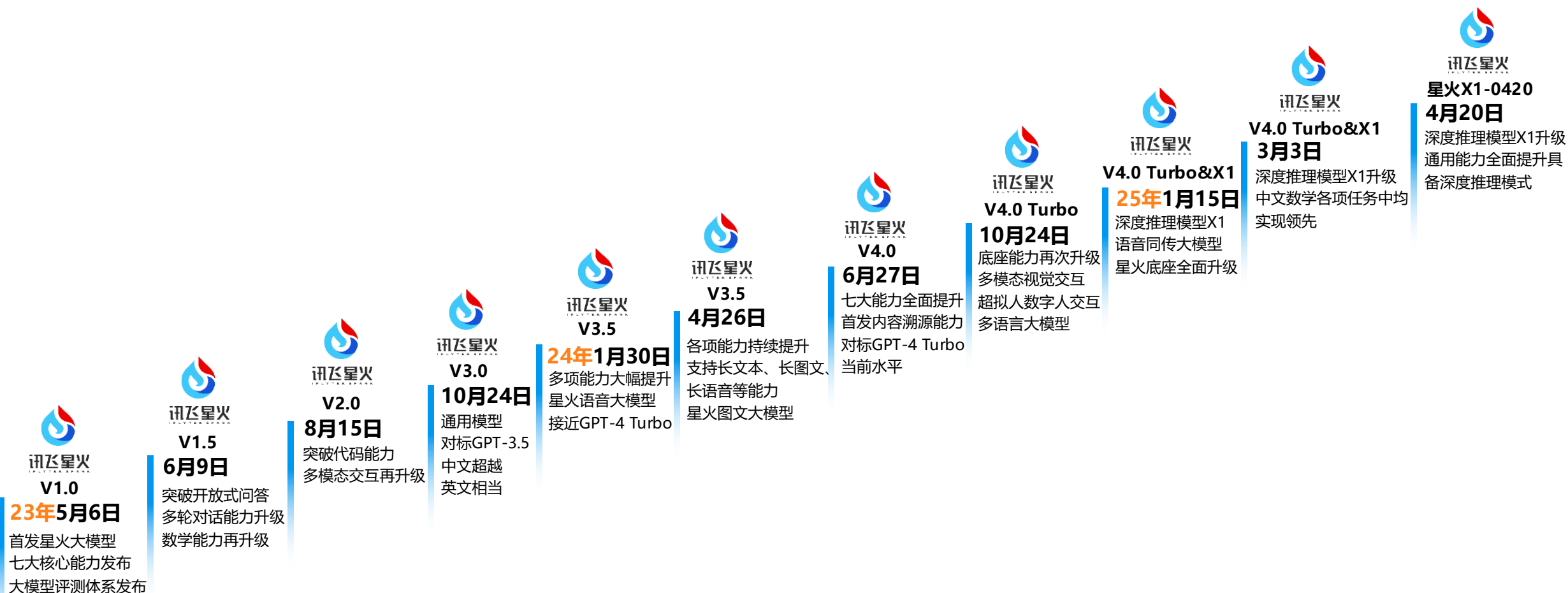
CONTENTS

- I. 背景
- II. 问题/痛点
- III. 解决思路/整体方案
- IV. 具体实现/技术实践
- V. 总结与展望

PART 01

背景：讯飞大模型的发展

从0到1，讯飞星火大模型实现从快速追赶到自主创新



讯飞持续探索全栈国产化无人区

讯飞星火实现了训练和推理的全国产化，星火大模型持续引领国产平台发展

“飞星一号”平台2024年全年平均使用率达到95%

500+次

解决基础软硬件问题

30+项

新增框架和平台特性

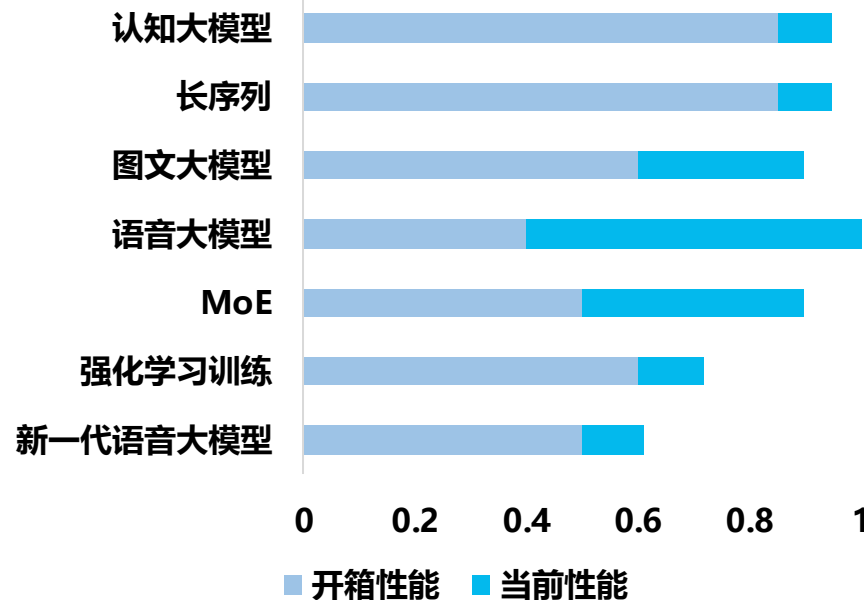
150+个

基础、通信和融合算子优化

75%

模型训练适配优化平台耗时优化

2024年“飞星一号”集群
优化过程中创新积累



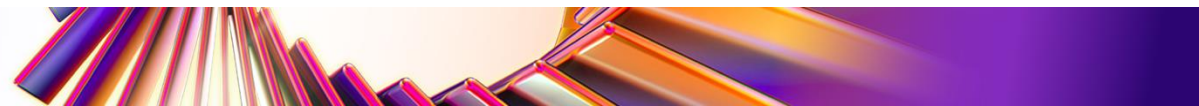
基于昇腾的大模型
训练效率持续优化

讯飞&华为联创探索“开车换车轮”
规模无关的断点续训新方案
可支持集群规模再次倍增

智算集群规模的
再次跃迁

“飞星二号”达到万P

未来基于国产算力的大模型自主技术创新，探索新模型新算法的持续适配及智算集群规模的再次跃迁



PART 02

痛点：智算集群维护

大模型的能力及效果受多种因素影响，集群算力是驱动大模型创新的基础

数据集质量

更好的数据质量更长的训练时间

新训练范式

强化学习决定模型对齐程度

Agent

决定使用工具、上下文记忆能力

多模态

决定模型功能丰富度

MOE

决定特定复杂任务处理能力

“算力规模 $\approx$ 参数量 * 数据量 / 训练时长”

集群大算力是支撑更高质量大模型创新落地的最关键基础

算力需求与数据量成正比

多模态数据训练需求是文本数据的320倍

算力需求与模型参数成正比

百亿百卡、千亿千卡

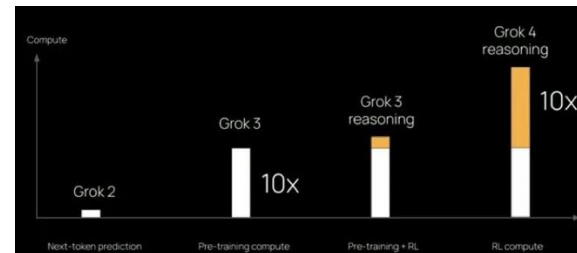
算力需求与序列长度成正比

32K+序列的模型已商用，1M+序列已出现

精度需求要匹配合适的浮点运算

综合使用FP32/BF16/FP16以快速模型收敛

Grok4算力分配：通过RL压强投入获得模型能力大幅提升

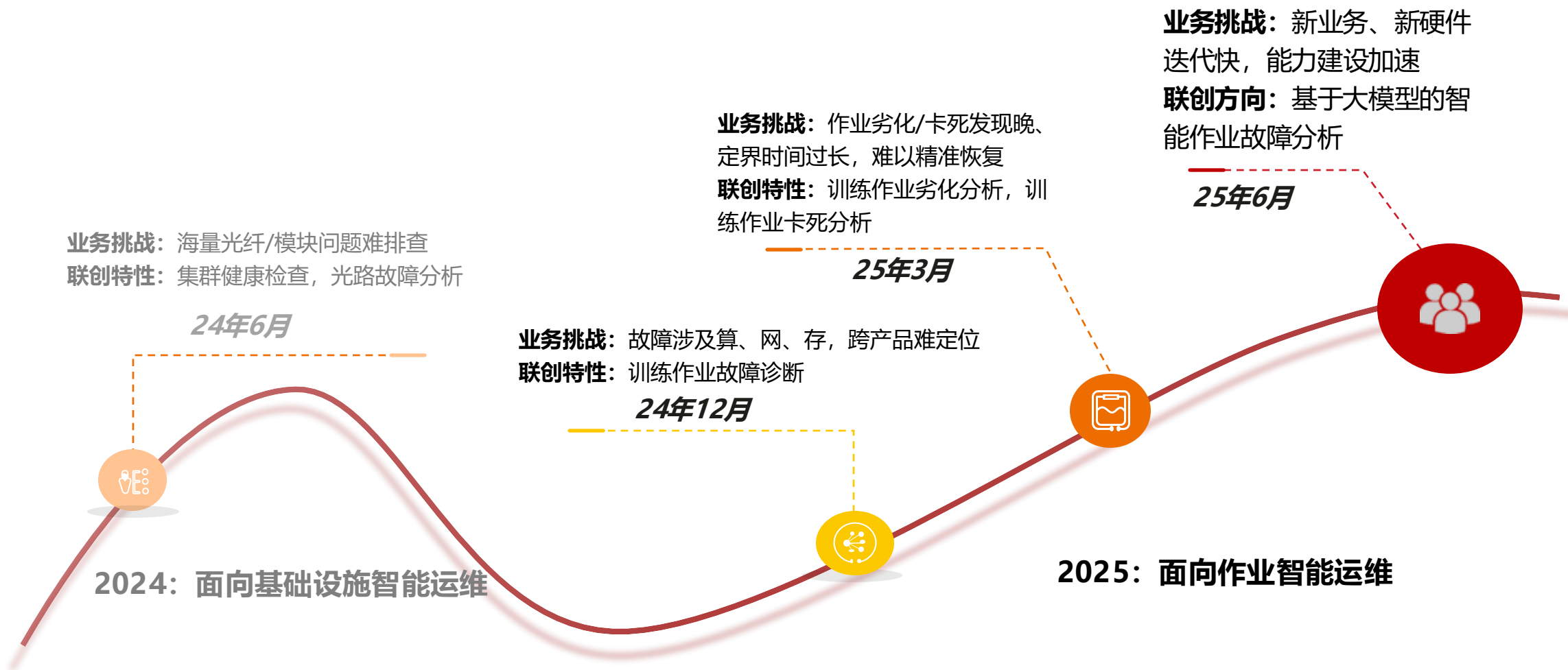


国内TOP企业计算规模较国外存在近10倍差距

集群规模
(万)



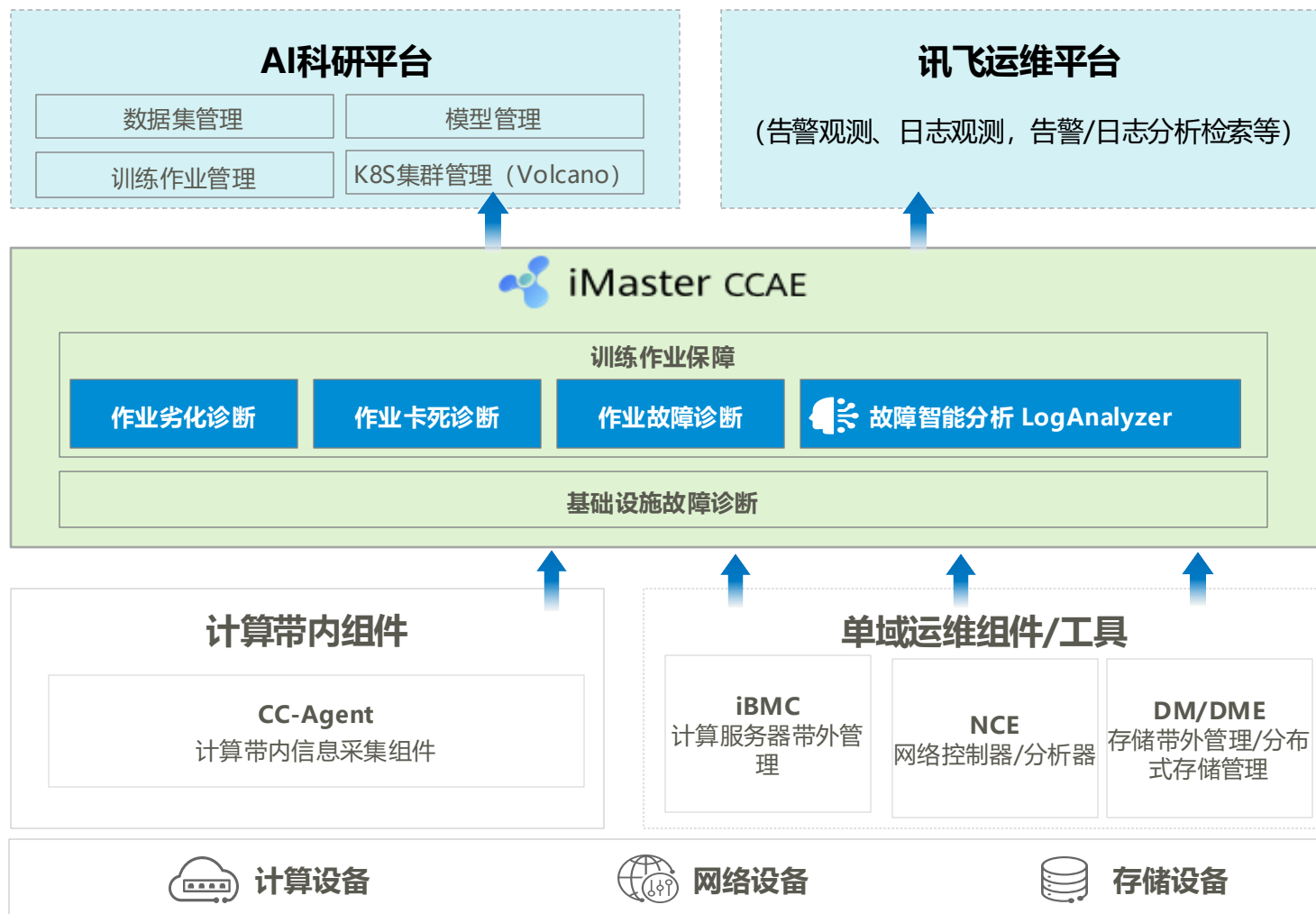
► 从0到1，讯飞星火大模型实现从快速追赶到自主创新



PART 03

整体方案：运维智能化

▶ 将华为CCAIE融入集群日常运维流程，提升运维智能化水平



飞星一号：高频次使用CCAIE特性

训练作业保障:

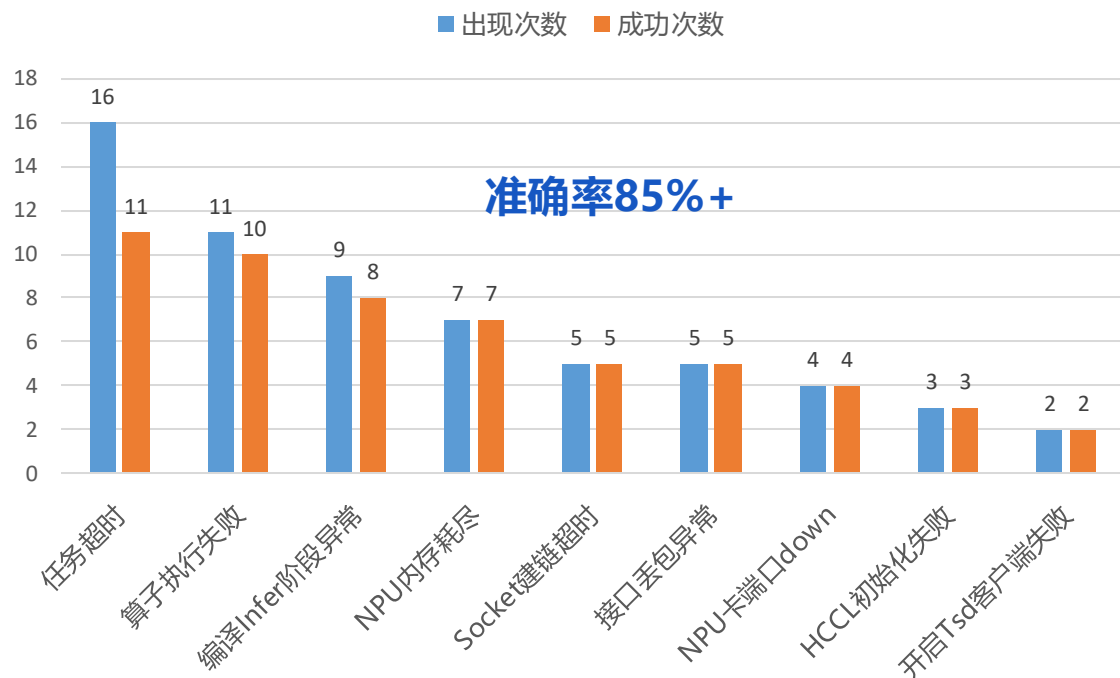
作业故障诊断，月均自动调度次数150+，
准确率85%+

基础设施故障诊断:

覆盖算网存基础设施故障，累计识别关键
故障180+

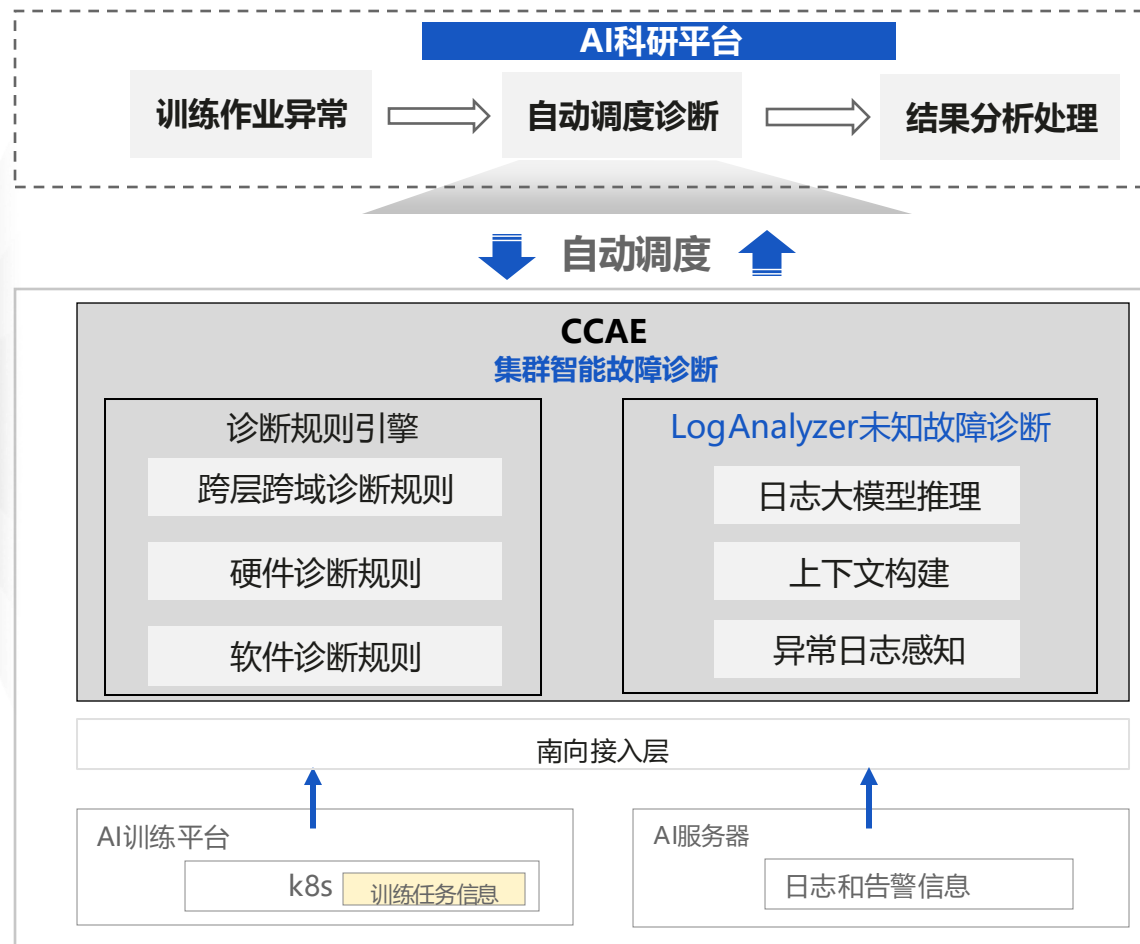
故障模式库已覆盖场景诊断准确率85%+，需要加强未知故障定位准确率

现网故障诊断TOP问题及诊断情况

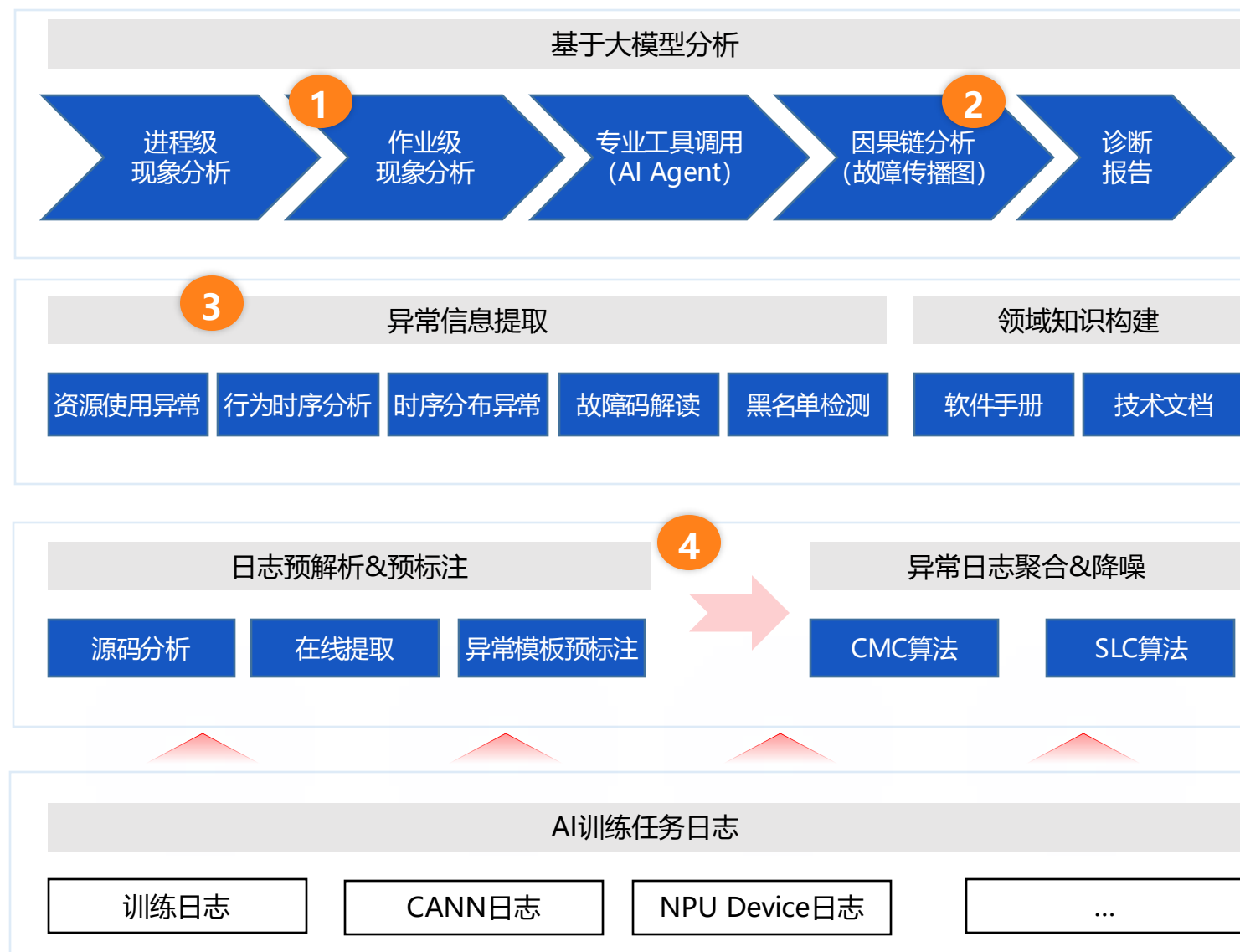


- ✓ 已知故障无法覆盖所有问题：随着新硬件上市、新业务应用，故障模式库逐步增长，基线-配套固定3个月周期，故障覆盖率呈上下波动特征。
- ✓ 未覆盖故障分析依赖各域专家手工分析，故障分析需要天级-周级不等

需要用大模型手段解决未覆盖故障的分析



▶ 首次引入基于日志大模型底座的AI辅助分析引擎



1 作业级故障分析技术

- 结合提取的异常信息和CANN、NPU等领域知识，生成训练进程故障时间线和故障现象
- 汇总进程级现象，结合HCCL通信知识分析导致集群作业中断的故障类别，生成作业级别的故障现象

2 故障传播链分析技术

- 结合作业、进程级别的异常事件信息，构筑基于大模型的故障传播链技术

3 异常信息提取技术

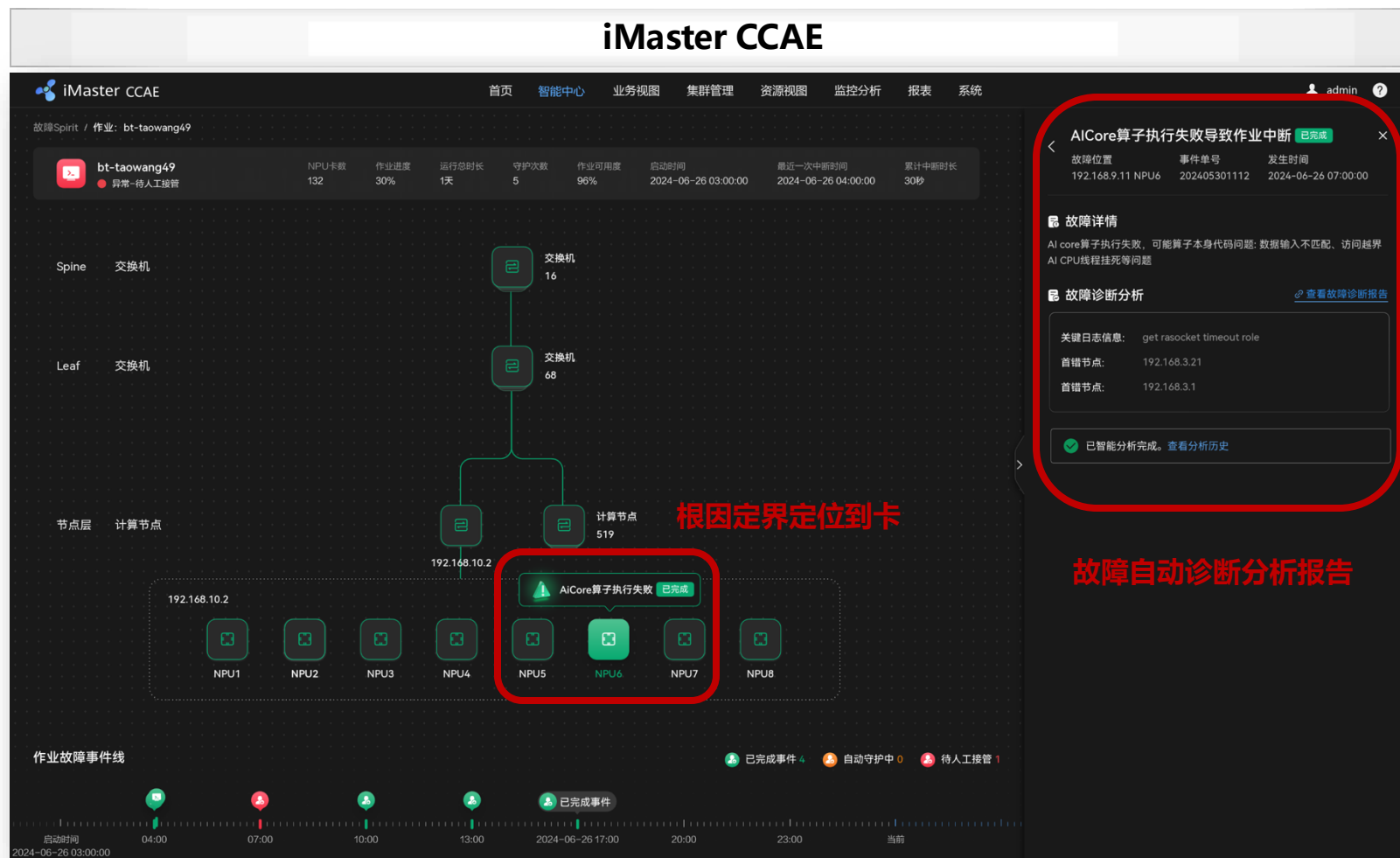
- 多种手段提取各节点的日志异常信息，生成大模型的有效输入数据

4 异常日志识别技术

- 日志预解析&预标注，跳过正常日志分析
- 异常日志聚合，降低入库数据量



客户视角目标态：故障自动感知定界，支撑端到端无感闭环



➤ **目标用户：** 客户运维、客户算法工程师、华为驻场人员

➤ **使用场景：** 作业中断时，自动触发诊断定界，10min内给出定界报告，一键下发进行故障隔离恢复

➤ **能力目标：** 算网存全场景的精准定界

- 计算侧定界到节点和NPU卡
- 网络/存储侧定界到域



PART 04

技术实践：LogAnalyzer

LogAnalyzer基于大模型全面分析全域日志，显著提升故障定位效率

挑战

- 某业务团队在运行2千多卡的MoE长文本微调训练任务，每运行2个小时，偶发中断，阻塞任务训练
- 37G运行日志，没有典型错误日志，在千卡（测试）环境上，多次跑任务验证，问题不复现

方案



LogAnalyzer方案



成果

涉及多个模块的疑难故障，LogAnalyzer能够全面分析，显著降低定位时长：天级 → 分钟级

LogAnalyzer构建集合通信域透视图，提升通信域初始化超时问题诊断效率

挑战

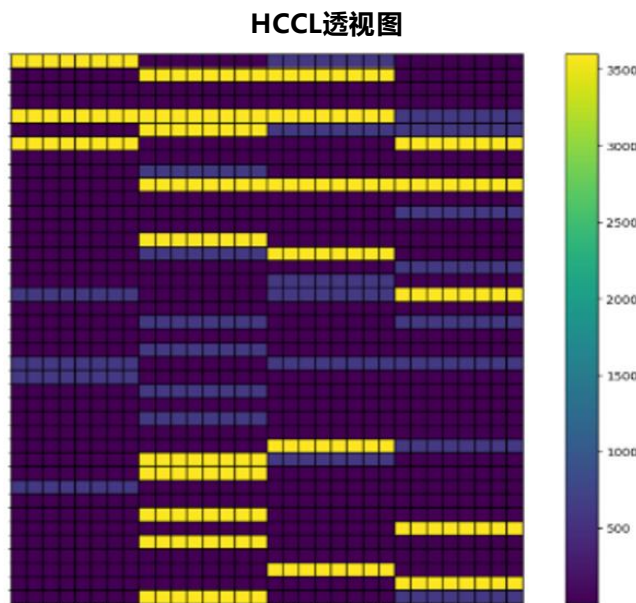
生产环境大规模集群，模型训练多个任务发生偶现卡死，影响模型正常训练，降低集群资源利用率，高度关注

方案

1. 快恢措施：将HCCL执行时间延长，先消减



2. 根治措施：HCCL透视图 + 运维专家，找到根因



昇腾运维专家



【分析过程】

在黄色节点上，现场结合训练脚本的日志发现，rank=4的算子，在执行时卡住了，同时在plog中没有发现通信域创建日志，此故障为通信域创建未下发，导致超时。

【问题根因】

通信域创建指令未能成功下发至计算卡导致超时中断

【修复措施】

1、Rank4卡隔离；2、减少数据量

- 黄色：未下发通信域初始化
- 蓝色：通信域初始化超时(默认600s)

成果

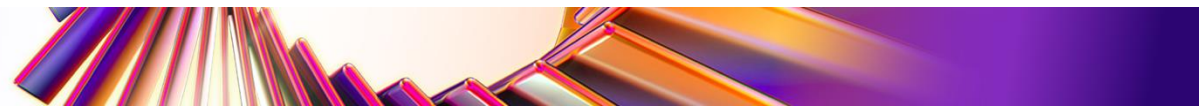
针对集合通信库业界难题，LogAnalyzer通过HCCL透视图+大模型，**HCCL相关问题，诊断效率提升2倍**

► LogAnalyzer上线使用4个月，未知故障诊断达到预期效果

4月~7月期间， LogAnalyzer有效故障诊断任务数72个， 离线诊断覆盖率100%（有日志数据即可诊断覆盖）；

诊断正确63个， 整体准确率：87.5%； 其中未知故障51个， 诊断正确43个； 未知故障准确率84.3%

问题分类	问题类别	数量	典型故障
未知故障诊断	1、跨栈定位类	合计：30 准确率：80%	异常表现在HCCL， 但根因可能是用户脚本， PTA、 CANN、 RTS等多模块耦合导致， 典型故障包括： ➢ CANN软件Bug导致通信/建链超时 ➢ 硬件问题导致算子超时
	2、环境配置类	合计：13 准确率：92.3%	训练环境和软件依赖导致作业无法拉起， 根因发散无法详尽穷举， 典型故障如： ➢ 训练通信环境配置错误 ➢ 软件版本依赖残余进程占用 ➢ ...
	3、硬件类故障	合计：8 准确率：87.5%	训练任务因传输链路故障， NPU 故障导致的任务中断或者超时， 典型故障包括： ➢ AICORE TIMEOUT导致算子执行超时 ➢ Kernel无效内存访问
已知故障优化	4、用户脚本及任务初始化类	合计：15 准确率：93.3%	训练任务初始化失败， 根因与用户代码和依赖库相关： ➢ IndexError、配置参数错误等用户脚本类问题 ➢ GE初始化失败， TBE依赖项缺失 ➢ 残余进程占用导致任务拉起失败
	5、资源占用类	合计：6 准确率：100%	由于存储、内存、计算资源冲突或分配问题导致， 包含： ➢ 磁盘配额满问题 ➢ OOM问题， 指向片上内存内存分配



技术途径：基于专业模型压缩故障上下文，融合算网存上下文自主诊断

理念：大模型作为诊断的“大脑”（负责推理、分析），调用各种专业工具和专用模型的“双手”（执行特定任务），从而解决复杂故障定位问题。

智算故障诊断专业算法方案设计



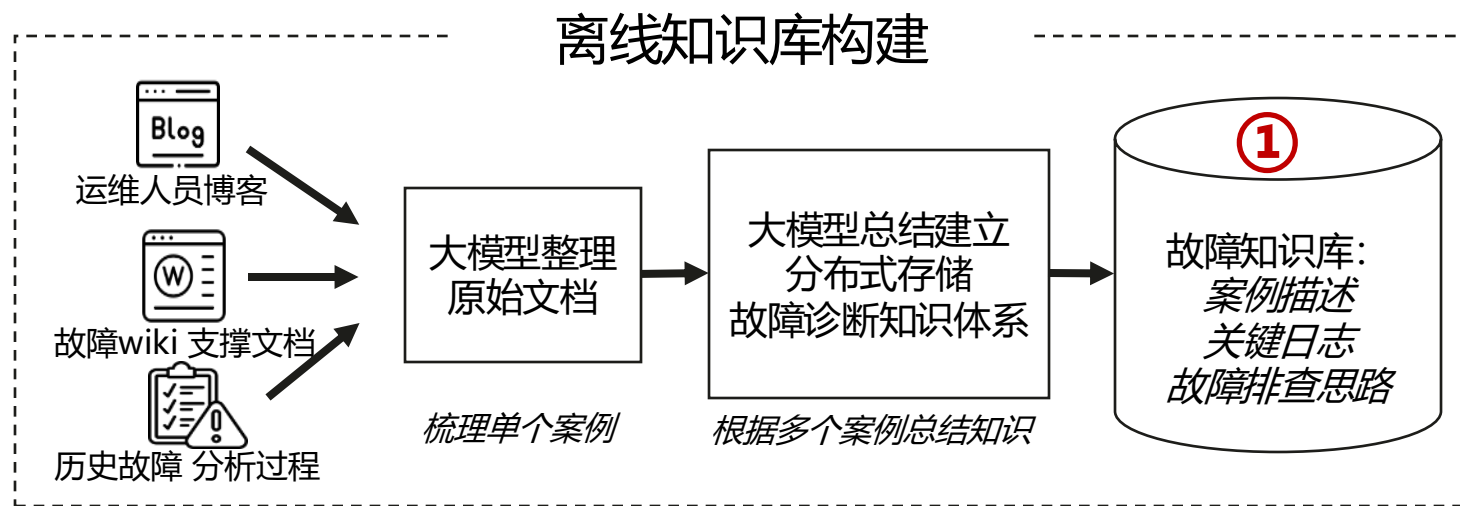
根因推导：多种报错类型间建立传播关系

- **根因识别：**通过Trainlog信息模型分析为资源问题
- **根因传播：**磁盘满引发关键资源竞争，导致All-Reduce操作期间任务超时
- **最终体现：**通信延迟持续过长，流同步失败，任务无法拉起



关键技术1：大模型自驱动构建故障诊断完备的背景知识

① 知识库构建成本太高

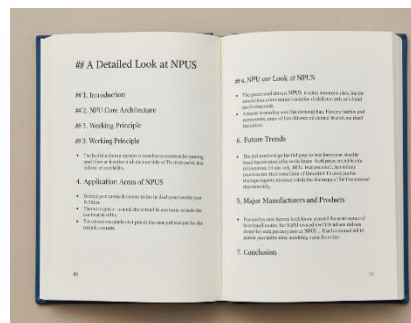


大模型自驱动知识库构建，高效赋能在线诊断

通过大模型将内部开发文档，用户支撑文档，排障案例，一线人员博客中的海量运维知识自总结为大模型易理解，易检索的文档，高效构建分布式存储领域故障知识库，克服人工构建耗时耗力的瓶颈。

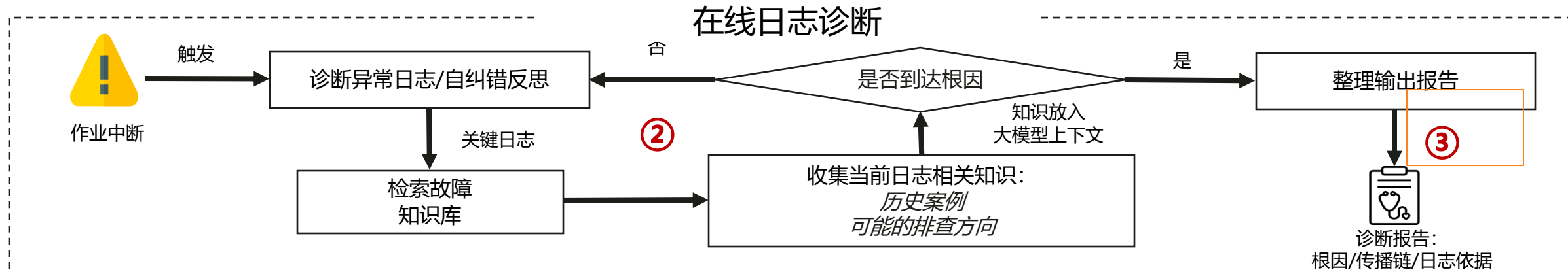


Wiki, 运维人员博客



一本LLM看得懂，能搜索的教科书

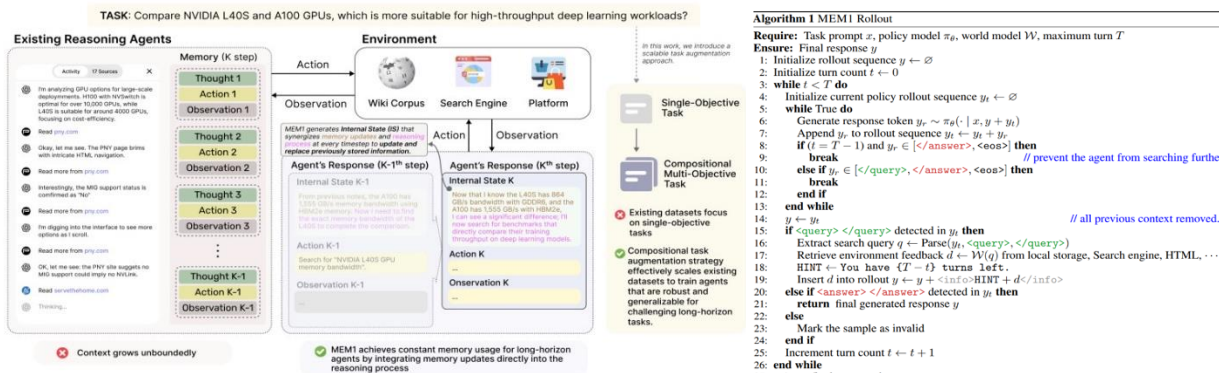
关键技术2：融合算网存上下文自主诊断



② 背景知识太多了，一次放不下

DeepResearch诊断流程，小步获取知识

借鉴DeepResearch方法提高大模型诊断准确性，在诊断过程中遇到大模型无法理解的信息时，通过检索知识库在上下文中自适应补充领域知识；获取新知识后自纠错反思，及时修正故障诊断方向，避免一错再错。



DeepResearch Agent流程

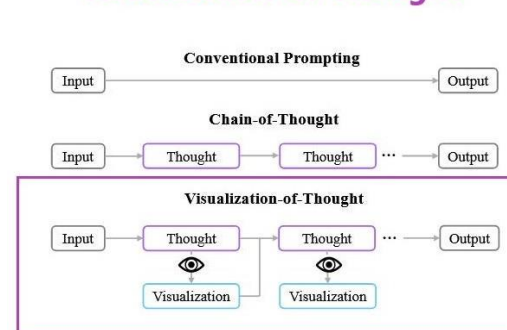
流程伪代码

③ 模型生成结果用户不相信

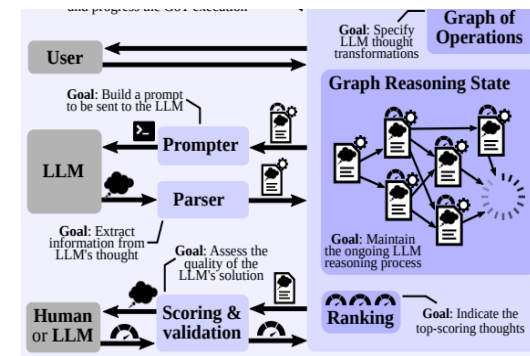
提取大模型思维过程，构建可信故障传播路径

从大模型诊断过程中的思考过程中提取关键定位步骤，反向构建故障传播链，从原始日志中反查关键定位日志证据，验证模型定位的结果，提供支撑诊断结果的日志证据链，提升结果的可信度。

Visualization-of-Thought



思维链可视化

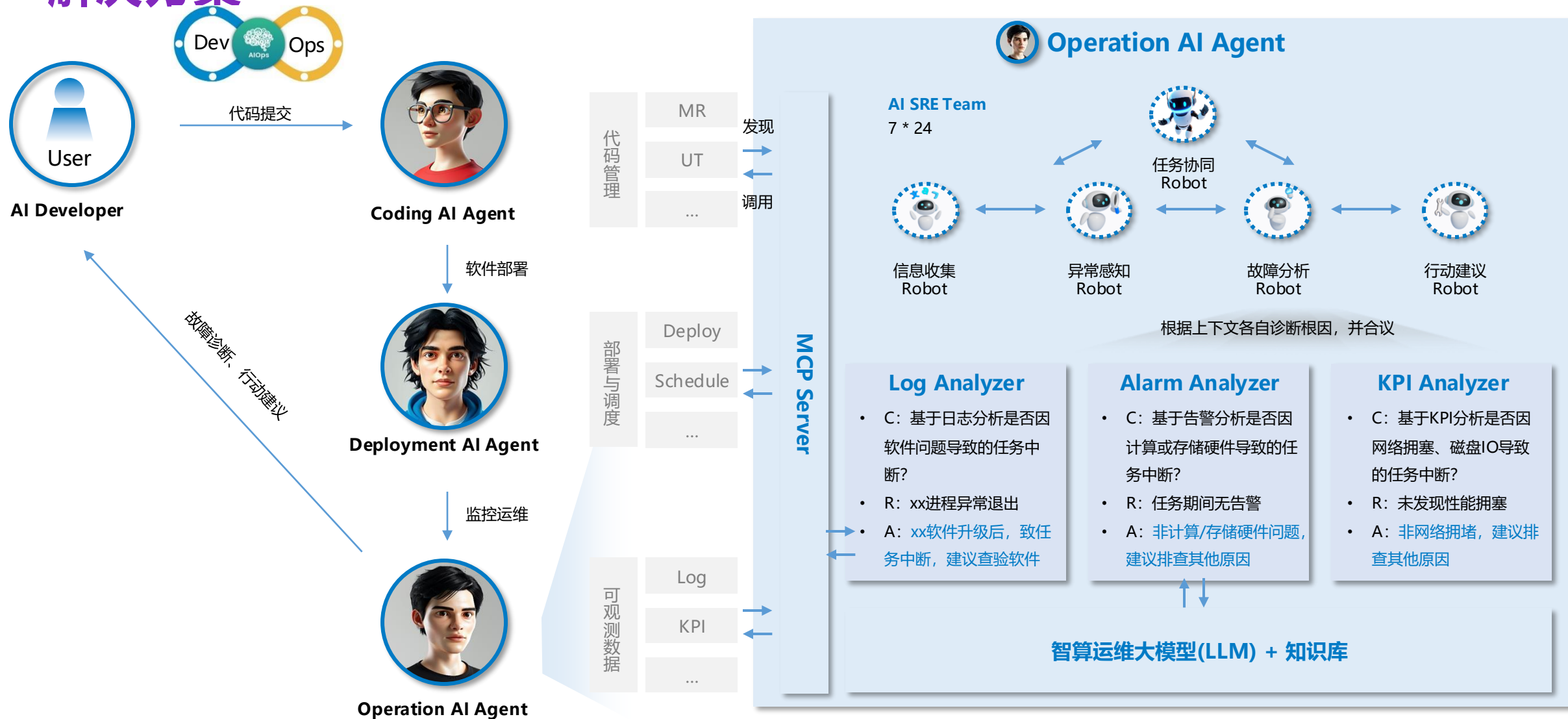


将LLM的思维转为图结构

PART 05

总结与展望：智算DevOps

联合实践 & 持续创新，打造业界领先的智算集群智能化运维解决方案



● C: Context 提示词/上下文 ● R: Response 回复 ● A: Action 建议

科技生态圈峰会 + 深度研习

——1000+ 技术团队的选择



NiDD 峰会

NiDD 峰会 上海站

AI+研发数字峰会

时间: 2026.05.22-23

NiDD 峰会 北京站

AI+研发数字峰会

时间: 2026.08.21-22

NiDD 峰会 深圳站

AI+研发数字峰会

时间: 2026.11.20-21



AiDD峰会详情

NiDD × AI+产品峰会

NiDD × 上海站

AI+产品创新峰会

时间: 2026.01.16-17

NiDD × 北京站

AI+产品创新峰会

时间: 2026.08.23-24



产品峰会详情



2026 AI+ PRODUCT INNOVATION SUMMIT

01.16-17 · ShangHai

AI+产品创新峰会



Track 1: AI 产品战略与创新设计

从0到1的AI原生产品构建

论坛1: AI时代的用户洞察与需求发现

论坛2: AI原生产品战略与商业模式重构

论坛3: Agentic AI产品创新与交互设计

2-hour Speech: 回归本质



用户洞察的第一性

——2小时思维与方法论工作坊

在数字爆炸、AI迅速发展的时代，
仍然考验“看见”的“同理心”



Track 2: AI 产品开发与工程实践

从1到10的工程化落地实践

论坛1: 面向Agent智能体的产品开发

论坛2: 具身智能与AI硬件产品

论坛3: AI产品出海与本地化开发

Panel 1: 出海前瞻



“出海避坑地图”圆桌对话

——不止于翻译:

AI时代的出海新范式



Track 3: AI 产品运营与智能演化

从10到100的AI产品运营

论坛1: AI赋能产品运营与增长黑客

论坛2: AI产品的数据飞轮与智能演化

论坛3: 行业爆款AI产品案例拆解

Panel 2: 失败复盘



为什么很多AI产品“叫好不叫座”?

——从伪需求到真价值:

AI产品商业化落地的关键挑战

智能重构产品 数据驱动增长
Reinventing Products with Intelligence, Driven by Data

80+

业界大咖

汇聚产品大咖的真知灼见

40+

创新案例

开拓全新AI+产品视角

10+

分论坛

涵盖产品到商业增长的全链路

800+

听众规模

共同探索产品的智能重构



扫码查看会议详情



第8届 AI+ 研发数字峰会
AI+ Development Digital Summit

感谢聆听!

扫码领取会议PPT资料

