

QCon
全球软件开发大会

亿级用户背后的智能诊断：多模态数据融合与实时诊断实践

徐建伟

bilibili可观测平台负责人



目录

01 从"人肉排查"到"AI诊断": bilibili 的痛点与机遇

02 智能诊断核心架构和演进思路

03 核心场景的 AI 化改造实践

04 技术演进方向与实践展望

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOps
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

01

从“人肉排查”到“AI诊断”

bilibili 的痛点与机遇

● 亿级用户背景下的故障挑战：复杂度指数级增长



当故障发生时，监控系统会几百个服务，成千上万的组件中瞬间产生海量的、多维度的原始数据

数据洪流与信息过载



规模庞大、组件繁多、交互关系错综复杂，而依赖的传递性与放大效应会导致排查困难

系统复杂度与依赖黑洞



微服务之间都是网状调用，一个服务挂了，所有依赖它的服务都可能被拖垮，形成雪崩效应

爆炸半径与隔离难度



故障发生时，多个团队（网络、基础设施、应用、数据库）在高压下如何快速协同，避免扯皮和信息混乱。

组织协同与沟通成本

● 传统线上排障的三大瓶颈

时间成本

手动排查，流程串联

拉群、查日志、口头同步
MTTR（平均恢复时间）过长，业务损失

准确率

依赖经验，易误判

拉群、查日志、口头同步
处理方案无效，甚至扩大故障

知识传承

知识隐性，难复制

“大神”的直觉和经验
人员变动导致运维能力断崖式下跌

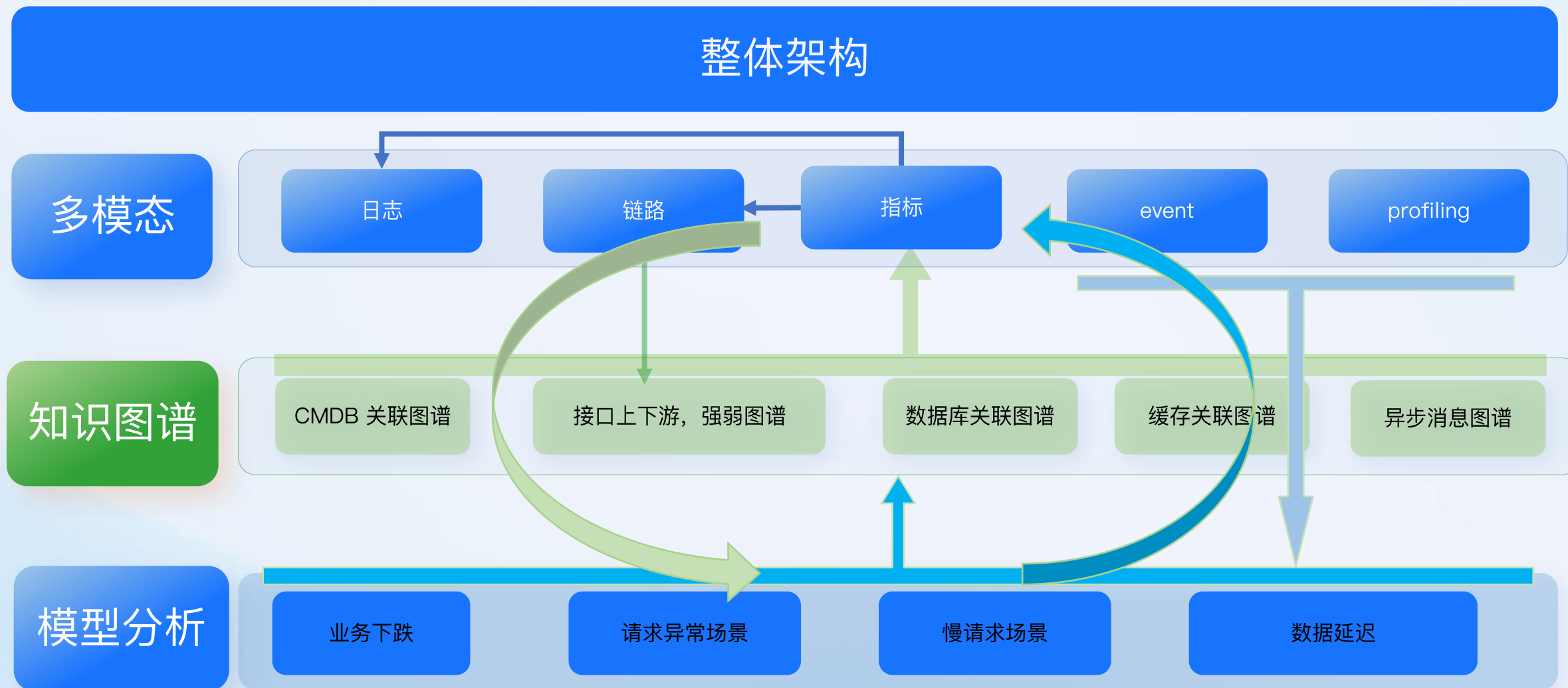
AI 技术带来的机遇：数据驱动 + 智能推理



02

根因分析核心架构和演进思路

● 场景化分析模型



多模态数据融合



知识图谱构建

多源数据集成

将来自不同源头、不同格式的监控数据（主要是指标、日志、追踪）进行采集、对齐、关联和融合，形成一个统一的、具有上下文的视图

知识存储与图谱构建

选择了为关系查询而生的图数据库作为最佳载体，将散落的知识点有机地整合成一张映射真实系统架构的语义网络

知识抽取与实体关系定义

多源异构的运维数据中，自动识别和提取出关键的、结构化的信息片段，这些信息片段将成为知识图谱中的“实体”和“属性”

知识应用与智能推理

利用已经构建好的、结构化的知识库（知识图谱），通过模拟人类专家的思维过程，对实时故障数据进行自动分析、推理和决策



大小模型分析

自然语言处理 + 时序分析的深度融合



时序算法分析

时序分析算法（如异常检测模型）检测到某个服务的P99延迟指标在10:05:00出现一个尖峰（异常点）。.



多模态数据关联与上下文构建

系统根据触发异常的时间点和实体，自动拉取相关时间段内的所有关联数据，并将其组织成一个结构化的“上下文窗口”



大模型推理与分析

准备好的多模态上下文信息，精心构造成提示词（Prompt），提交给大模型（如GPT-4, LLaMA, 或领域微调的模型）进行分析

数据提纯：打造高信噪比的诊断燃料

清洗

去重，无效值处理，噪音过滤，格式化



1

去芜存菁，得到原材料

大幅减少数据量，只保留与本次故障**强相关**的数据片段。这是降低噪音的关键一步。

对齐

时间对齐（时间同步，时间窗口划分）
实体对齐（标签统一，拓扑关联）



2

搭建骨架，建立数据之间的时空关联

对过滤后的数据进行分组、统计和模式识别，将零散的个体异常汇总成群体性规律，从而发现共性问题。

增强

富化（添加更多描述性标签）
衍生指标计算（从原始指标中计算出更有业务意义的表现力指标）
模式发现（发现频繁出现的错误模式）



3

画龙点睛，注入业务语义和上下文

将不同来源（日志、指标、追踪）和不同组件的数据进行交叉关联，为故障点补充完整的上下文，最终定位根因。这一步是“为什么”的关键。

模型迭代：动态调整推理边界

1 执行诊断与产生证据

模型基于当前的推理边界（如：只检查直接依赖的服务）对故障进行初步分析，并输出诊断结果（如：根因是服务A）和支持该结果的证据链（如：因为服务A的延迟飙升，且错误日志显示XXX）

2 结果验证与反馈获取

描述：诊断结果需要接受验证。这可以通过两种方式：

人工反馈：运维专家确认或驳回结论。

自动化验证：系统执行修复动作（如重启服务A）后，观察指标是否恢复正常。

关键：这个反馈是模型迭代的“黄金标准”。



让 AI 学会“灰度思考”

3 推理边界调整分析

根据反馈，系统会分析上次推理的不足，并决定如何调整边界

4 模型与知识库更新和重跑验证

将分析后的调整策略应用到模型和知识库中。带着更新后的、调整了推理边界的模型，对本次故障进行重新诊断，验证结果

工单即训练集：

每一次人工复盘都是模型的进化机会



发生故障与人工处置

生成高质量工单与复盘报告

报告解析与知识提炼

模型更新与进化

在此过程中，所有操作、观察点和时间线都被系统化地记录（如通过统一的运维平台），为后续复盘提供原始数据

故障解决后，负责人撰写复盘报告。这份报告是监督学习中的“标注数据”

AIOps系统利用NLP技术（如大模型）自动解析复盘报告，将其中的非结构化文本转化为结构化的知识

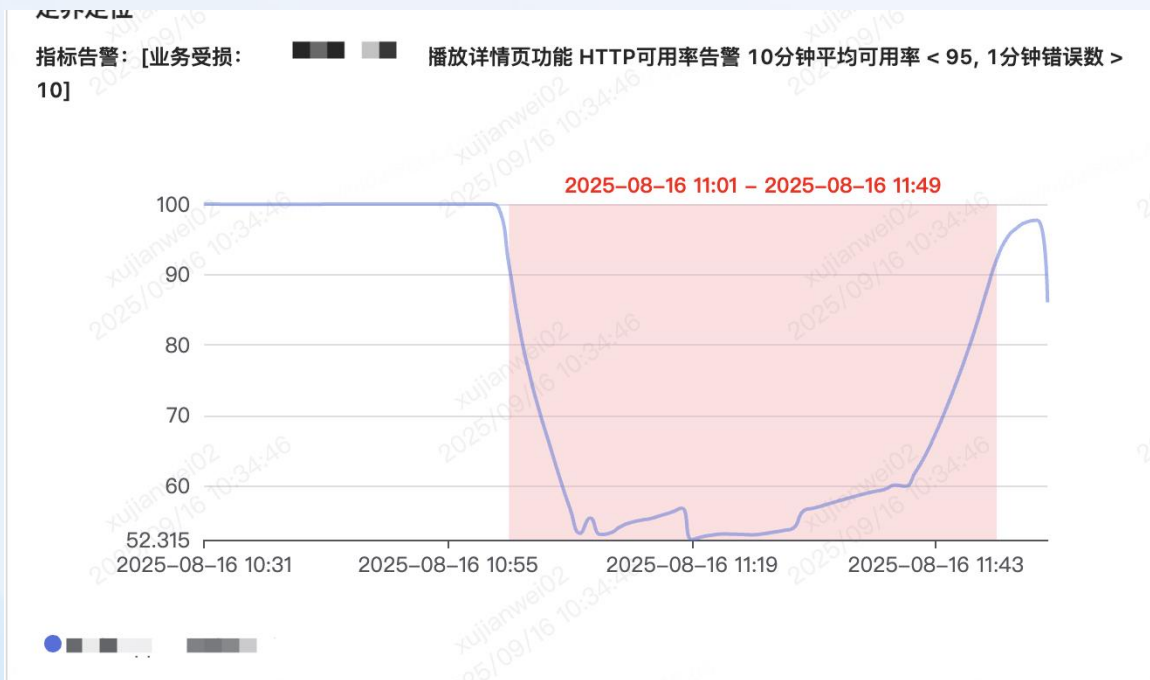
将上一步提炼出的结构化知识注入到各类模型中，模型从“新知识”中学习，实现了进化

03

核心场景的 AI 化改造实践

视频播放异常：从告警到根因定位的 3 分钟闭环

现象



定界定位

根因分析

👍 准确 🗑️ 不太准确 ⓘ

[查看更多](#)

大模型总结: 【推荐根因】告警根因大概率是下游应用 **容器资源饱和度异常** 导致服务出现问题

【相关表现】下游应用 open.bangumi.view-gateway 存在 **TCP连接数突增**、**CPU使用率突增** 以及 **容器 cpu throttle 次数突增**，且 **grpc依赖错误指标** `{'code': '-504', 'method': '/SeasonInfos'}` 突增明显，得分较高。这些异常共同指向下游应用的资源瓶颈问题。

【其他根因1】下游应用 存在依赖下游接口异常，Tag为: `{'code': '-504', 'method': 'DictContains'}` grpc依赖错误指标突增。

【其他根因2】下游应用 存在上游请求QPS突增，Tag为: `{'caller': 'GRPC请求QPS突增'}`

触发可用率告警后自动触发根因分析，30秒后推出根因分析，快速定位

场景化分析模型

领域知识库构建和因子关联

mysql

慢sql-> 扫描行数, 索引使用, 锁占用
物理机cpu、内存、io出现瓶颈或者故障

数据连接异常->连接池打满, 网络连通情况,
服务端运行健康状态

SLO

接口异常/耗时长

上游请求变化

下游异常波动

采样trace/log的
详情

容器/主机异常,
运行时异常

关联变更

缓存

慢查询/长查询

客户端检测
连接池使用情况 (连接池慢, 或者
未正确关闭)

查找/扫描记录
(识别大key, 大命令)

资源使用情况
(内存耗尽, 内存碎片等)

mq

消费者延迟或消息堆积

Producer 生产速率
服务端写入情况

Consumer的消费速率
Consumer资源变化情况

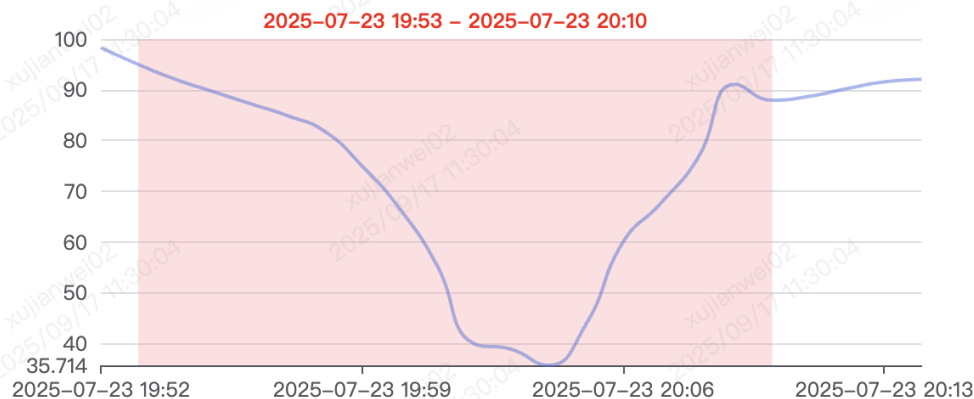
Rebalance 日志

推荐系统降级：多组件级联故障的智能溯源

现象

定界定位

指标告警：[应用] (L1) HTTP Metric的可用率告警 10分钟平均可用率 < 95, 1分钟错误数 > 10]



live.live-user.xroom-extend

错误分析

sh001: /UpdateRoomContentForAudit: -500 错误量: 108
sh001: /UpdateRoomPicForAudit: -500 错误量: 42

TOP实例分布

-2143445-66fbd48587-4f7pq: 占比: 3%
-2143445-66fbd48587-5vqn5: 占比: 2.5%

定界定位

根因分析

👍 准确 🗨 不太准确 ⓘ

查看更多

大模型总结：【推荐根因】告警根因大概率是[存在容器平台]发布单[2151823]变更]

(https://c/prod/deployment/2151823)

【相关表现】下游应用 存在服务接口异常，调用链信息显示 /live.daoanchor.v1.DaoAnchor/FetchRoomByIds 发生 rpc error: code = Unknown desc = -500, 且错误码集中在 -500, 占比约 78.5%。同时，下游应用 存在变更，且错误分布集中在可用区 sh001, 可能与变更相关。

【其他根因1】当前应用 对象可能存在依赖 Redis 异常，但未发现明显异常指标或突增。

【其他根因2】存在容器平台 (Caster) [发布单[2151819]变更](https://prod/deployment/2151819), 但未发现直接关联的异常表现。

04

技术演进方向与实践展望

场景覆盖的扩展

降低用户领域知识到根因模型的转化门槛

应用SLO

微服务可用率

对应用相关的指标，链路，日志进行分析，找到可用率下跌的原因

中间件服务端

中间件服务端

对数据库，缓存，消息中心等基础组件进行根因分析。
比如常见的慢sql问题，缓存击穿等

基础硬件和系统层

系统层或者硬件层

对网络，物理机，容器，系统层等进行根因分析，
比如节点与硬件故障，网络故障，oom，oomkiller

业务特定场景

业务场景

对业务自身的逻辑进行问题定位，尤其是一些客服召回的
比如说交易下跌，直播在线人数突降，弹幕事实同步下降

•准确率的持续提升

从问题定界向精准定根因的技术深化

演化

多根因方式定界，有一个命中就算成功

第一根因准确定界才算成功

第一根因从定界到定位

知识库

集成强弱依赖

变更明细

异常归类

第三方接口集成

通用库日志集成

推导优化

特征增强

因果关系

根因链路合并

历史数据因果特征构造

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOPs
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software

