分享主题

大模型时代的智能运维（AIOps）

裴丹 清华大学

人工智能产业链联盟

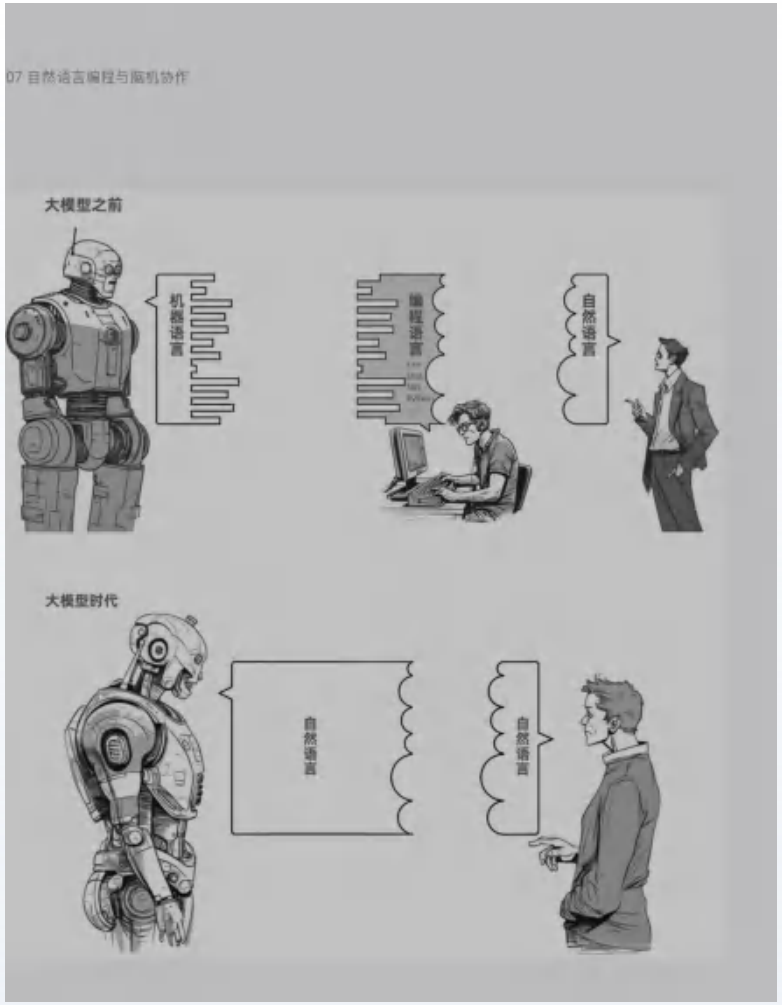
星主： AI产业链盟主

知识星球

微信扫描预览星球详情

大模型时代，运维界普遍关注的问题

在大模型时代，AIOps可以 “说人话”了



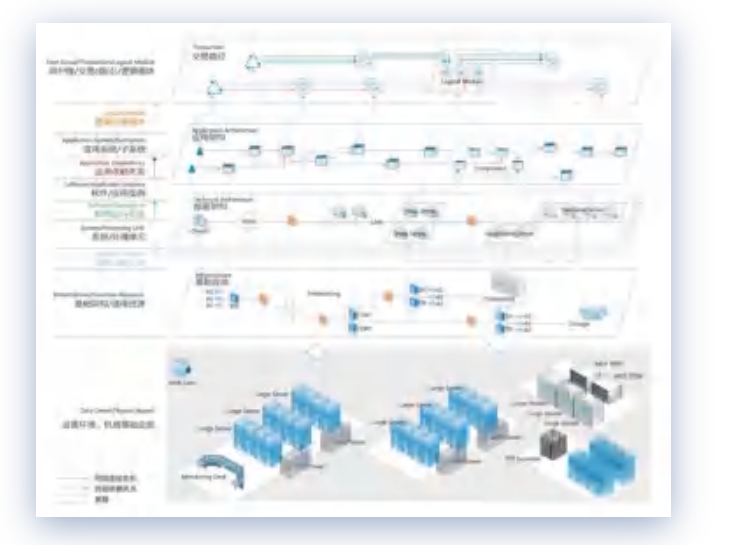
除了需要“说人话”，AIOps还需要把语言模型与已有的小模型工具有机整合

AIOps小模型数量众多

多样化的场景

- 预防
- 定位
- 发现
- 恢复
- 改进

- 多模态数据
- 实时数据
- 数据关系复杂



比尔·盖茨：AI智能体即将彻底改变我们使用计算机的方式

ifenxi

GatesNotes

LOG IN SIGN UP

THE FUTURE OF AGENTS

AI is about to completely change how you use computers

And upend the software industry.

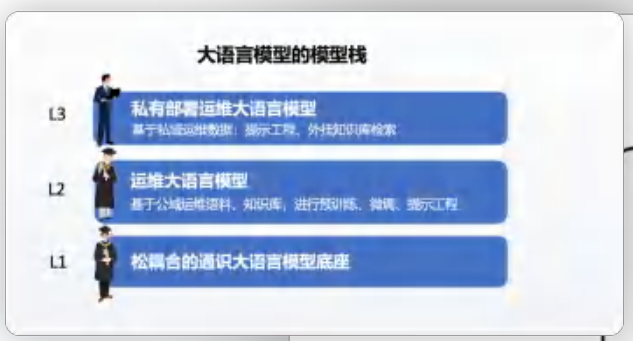
By **Bill Gates** | November 09, 2023 · 12 minute read

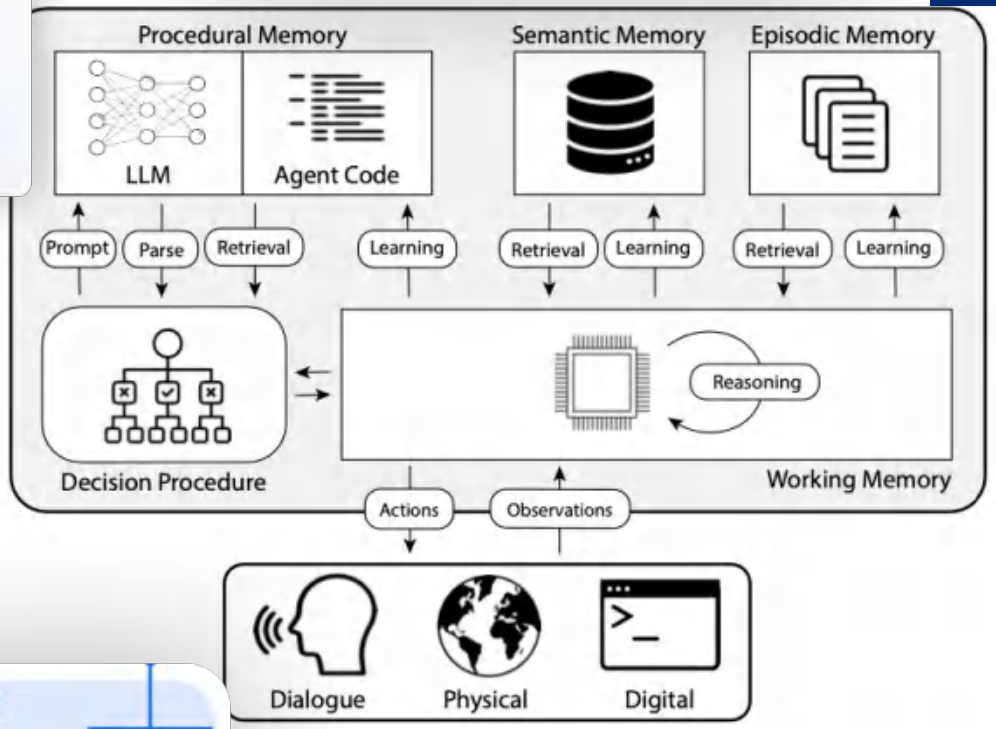
运维智能体

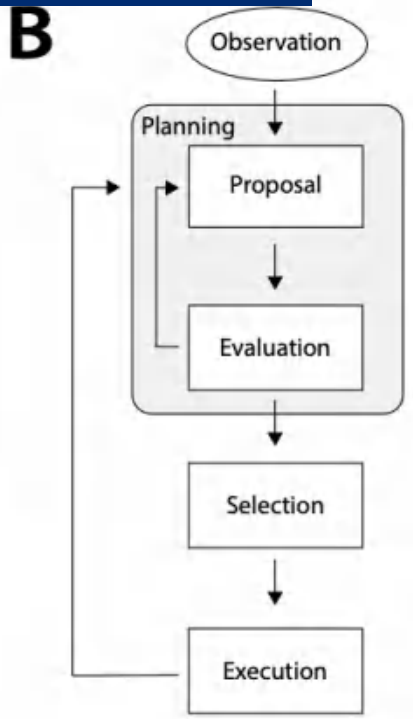
助理 教练 顾问 参谋 专家

AIOps中的智能体

历史工单、告警、操作记录、文档等



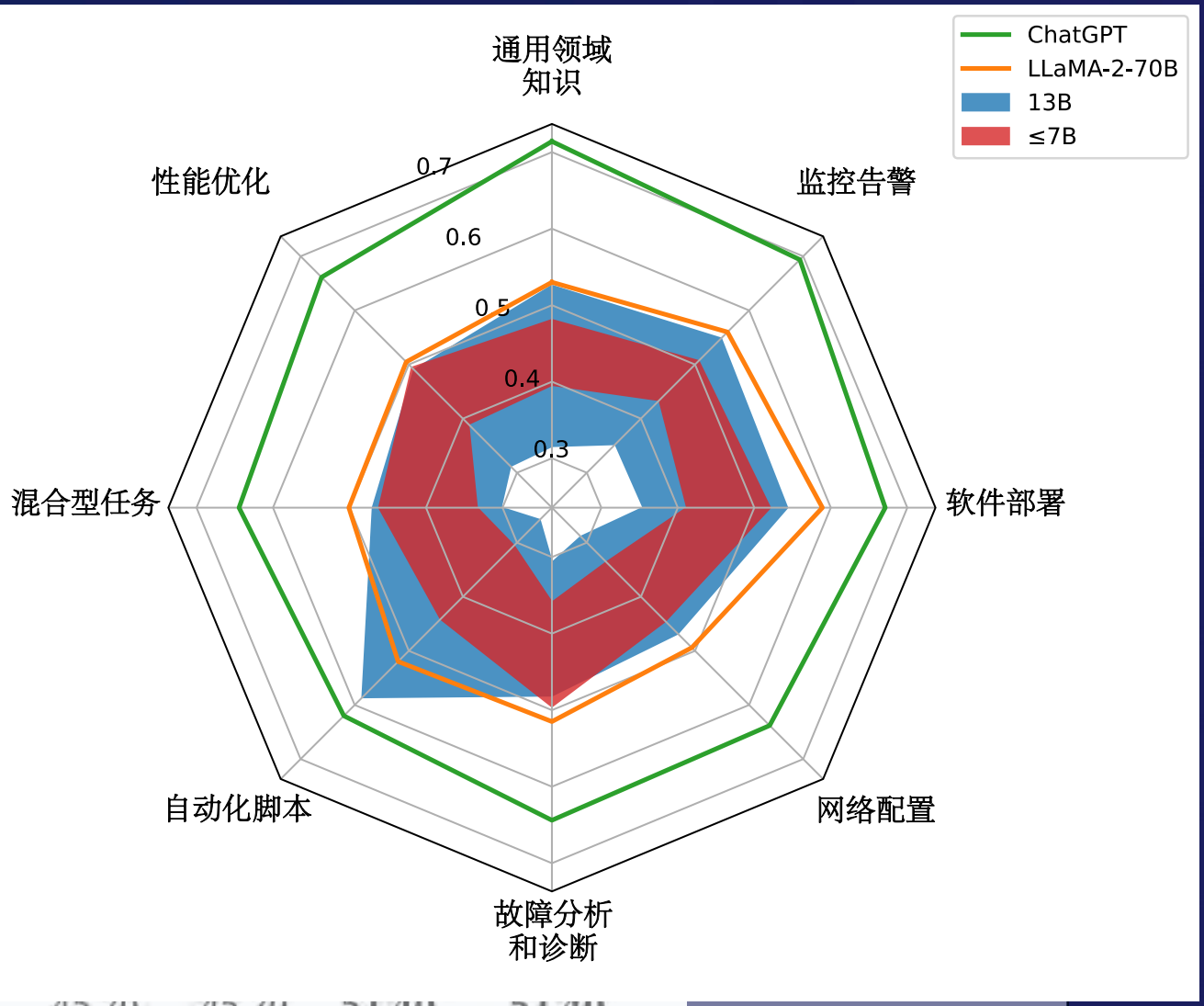




...itectures for language agents (CoALA). **A**: CoALA defines a set of interacting ...he **decision procedure** executes the agent's source code. This source code consists ... with the LLM (prompt templates and parsers), internal memories (retrieval and ...al environment (grounding). **B**: Temporally, the agent's decision procedure executes a **decision cycle** in a loop with the external environment. During each cycle, the agent uses **retrieval** and **reasoning** to plan by proposing and evaluating candidate **learning** or **grounding** actions. The best action is then selected and executed. An observation may be made, and the cycle begins again.

通识大模型在运维领域表现整体不如人意，而且参差不齐

欢迎为评测榜单贡献题目和模型：
https://opseval.cstcloud.cn/content/leaderboard



Model	Zero-shot Naive	SC	CoT	C				
GPT-4	/	/	/					
GPT-3.5-turbo	66.60	66.80	69.60					
LLaMA-2-70B	55.80	57.20	61.90					
LLaMA-2-13B	41.80	46.50	53.10					
LLaMA-2-7B	39.50	40.00	45.40					
Qwen-7B-Chat	45.90	46.00	47.30					
Baichuan2-13B-Chat	37.90	38.30	42.70					
InternLM-7B	38.70	38.70	43.90	43.90	45.20	45.20	51.40	51.40
Chinese-Alpaca-2-13B	37.70	37.70	49.70	49.70	48.60	48.60	**50.50**	**50.50**
ChatGLM2-6B	24.80	24.70	36.60	36.50	37.60	37.60	**40.50**	**40.50**
Chinese-LLaMA-2-13B	29.40	29.40	37.80	37.80	**40.40**	**40.40**	28.80	28.80
Baichuan-13B-Chat	18.30	20.40	28.60	**37.00**	24.10	26.70	18.20	17.80

第三梯队

运维领域的应用需要一个“能听懂运维语言”的大语言模型

大语言模型的模型栈

L3

私有部署运维大语言模型

基于私域运维数据：提示工程、外挂知识库检索

L2

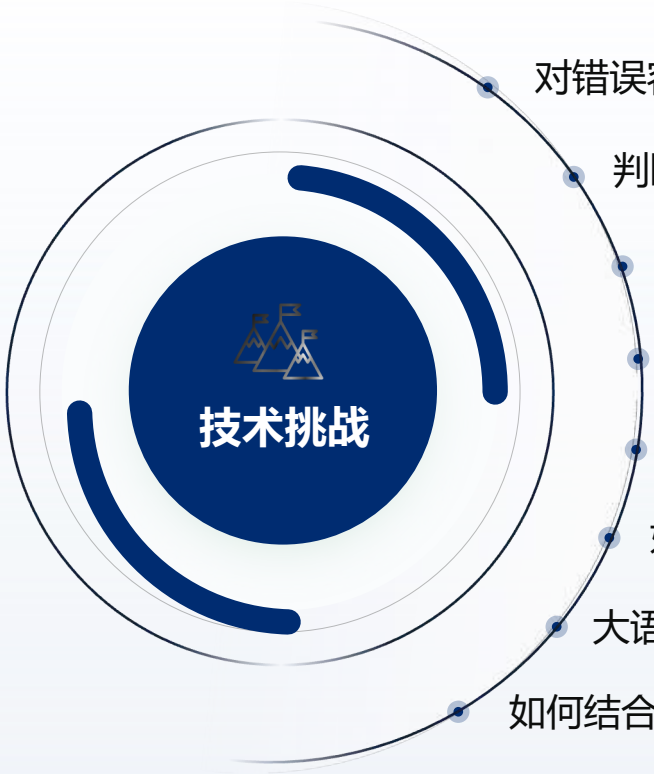
运维大语言模型

基于公域运维语料、知识库，进行预训练、微调、提示工程

L1

松耦合的通识大语言模型底座

既要避免过于乐观：大模型在运维领域普遍落地还存在不少技术挑战



- 对错误容忍度低，尽量避免幻觉
- 判断对错、标注门槛高
- 结果必须可解释性强
- 支持低开销私有部署
- 运维语料不足，私有语料质量数量更不足
- 如何融入大量存量结构化知识
- 大语言模型无法直接处理结构化、多模态、实时数据
- 如何结合大量存量AIOps、自动化运维工具

也要避免悲观：前述所有技术挑战都有技术思路可以解决

挑战	解决思路
避免幻觉	检索增强 增大显式知识占比：思维链、思维树、思维图、知识图谱 “有据可依”的生成策略提供可解释性
可解释性强	
严肃语料不足	训练策略：课程学习，由易到难
私有部署开销低	模型分层 在公域做预训练、微调、提示工程 私有部署时避免预训练、微调 通过检索方式融合本地知识库 文档、提示作为便捷的知识工程手段 通过降低模型精度降低私有部署推理开销
私域数据数量、质量不足	
底座模型选择	与开源大语言模型底座尽量解耦
大语言模型无法直接处理结构化、多模态、实时数据	多模态基础模型群； AIOps结构化大模型：左脑AIOps算法，右脑运维知识图谱
如何结合大量存量AIOps、自动化运维工具	基于智能体方式， 采用基础模型编程框架，融合小模型运维工具（插件（Plug-in） ，工具（Tool））

不求全面开花，而是小步快跑、以用促建：近期、中期、长期应用的定位

运维大语言模型带来的直接增量价值

- 眼：监控
- 手：自动化运维
- 脑：AIOps
  - 左脑：算法
  - 右脑：运维知识图谱
- 心：专家人工审核、决策
- **耳：基于运维大语言模型的输入增强**
- **嘴：基于运维大语言模型的输出增强**

近中期应用的定位

助理　教练　顾问　参谋

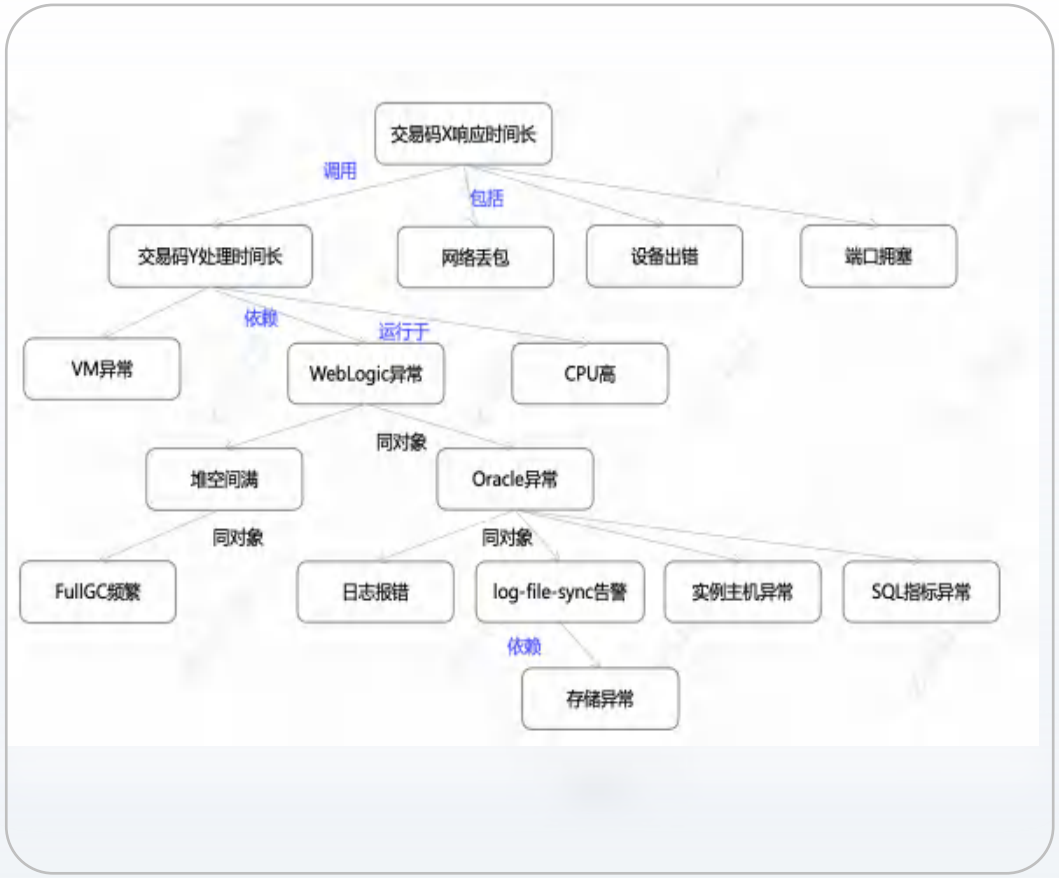
中长期应用的定位

+内部专家

运维大语言模型近期应用举例(1)：数字化运维助手

ifenxi

对结构化知识检索



问：X应用交易响应时间比平时增加1倍，可能原因是什么?

答：可能的原因包括：
1. 依赖Y应用处理时间过长
2. 中间件Weblogic问题，请检查 ...
3. 网络丢包问题，请检查相关服务间见的网络链路丢包率是否有异常
4. 第三方应用问题，请检查...

问：Oracle活跃会话数的确飙高了，请问可能的原因是什么?

答： Oracle实例B的活跃会话数的确飙高可能的原因包括：
1. log-file-syn告警
2. 实例B的主机异常
3. 有CPU占用过高的SQL

问：中间件WebLogic的确异常，请问下面怎么办?

答：中间件Weblogic异常的原因包括：
1. WebLogic堆空间满，请检查 ...
2. Oracle数据库“平均活跃会话数”指标飙高

问：查到实例B上有个SQL C占用CPU 90%，该如何处理?

答：建议杀掉该SQL C， 具体指令是 ... 。建议联系Oracle数据库管理员执行该操作。

运维大语言模型近期应用举例(2)： 私有文档问答

私有部署运维大语言模型

基于私域运维数据：提示工程、外挂知识库检索

运维大语言模型

基于公域运维语料、知识库,进行预训练、微调、提示工程

松耦合的通识大语言模型底座

检索插件

私有文档

生成技术： 检索增强

生成策略： 通过“有据可依”提供可解释性让模型能够正确的给自己生成的内容打上引用（reference），清晰地展示出每一段是来自于哪个文本

运维大语言模型近期应用举例 (3)：脚本解读 (Query2Lang)

ifenxi

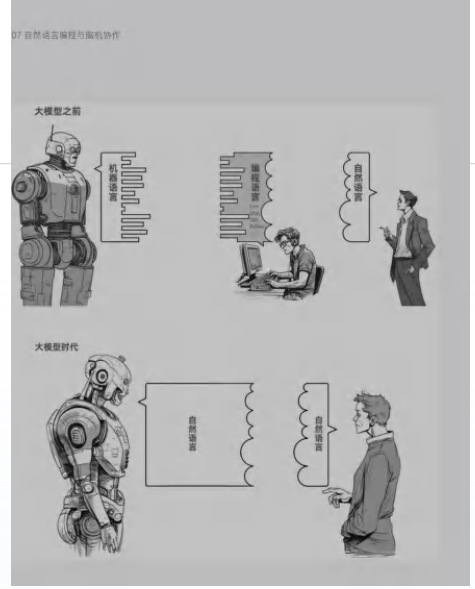
对存量脚本进行文字解读

运维大语言模型近期应用举例 (4)：数据注释（Data2Annotation）

ifenxi

对运维数据中的本体、实体、属性、字段、标签进行注释

近中期应用： Lang2Query, 为单个存量工具提供自然语言交互增强，提供意图识别、总结等能力



基础: 数据标准化、工具接口标准化

- Lang2API (自动生成API调用)
- Lang2SQL (自动生成SQL)
- Lang2GSQL (自动生成图SQL)
- Lang2SPL (自动生成日志查询语句)
- Lang2Scripts (自动生成脚本)
- Lang2Config (自动生成配置)

中长期应用：基于AIOps智能体，编排多个工具完成复杂运维任务

举例：基于大语言模型的实时故障工单自动生成

基于大语言模型，以实时日志、调用链、指标等数据为输入，结合故障检测、定位、根因分析、影响分析等AIOps工具的输出，自动生成实时故障工单。

附录 C

事故状态文档示范

莎士比亚搜索服务 新前文 ++ 过载事故：2015-10-21

事故管理网站：http://incident-management-cheat-sheet

（沟通负责人会随时更新事故概要）

摘要：莎士比亚搜索服务由于新发现的前文不在索引中而处于连锁故障状态

状态：活跃，事故编号 #465

事故处理中心：IRC #shakespeare 频道

事故处理组织架构：（参与人）

- 目前事故总负责人：jennifer
  - 运维负责人：dschrown
  - 计划负责人：jennifer
  - 沟通负责人：jennifer
- 下一个事故总负责人：待定

（沟通负责人在交接班时或者每 4 个小时更新一次）

细节状态：（最终更新时间 2015-10-21 15:28 UTC，Jennifer）

退出条件：

- 向莎士比亚搜索服务的 Search Corpus 中添加新的前文（TODO）
- 在 20 分钟内维持 SLO，可用性为 99.99%，延迟为 99%<100ms（TODO）

417

代办列表以及已提交的工单：

- 执行 MapReduce 任务，重新索引 Shakespeare corpus（DONE）
- 启用一些紧急资源来提高容量（DONE）
- 启用 flux capacitor，在集群之间均衡负载（Bug 5554823）（TODO）

事故时间线（倒序排列，时区为 UTC）：

- 2015-10-21 14:55 UTC dschrown
  - 出现大量紧急报警，全部集群出现 ManyHttp500s。
- 2015-10-21 15:01 UTC dschrown
  - 由于出现了连锁故障，声明目前进入紧急状态。
- 2015-10-21 15:04 UTC martyn
  - 从 Shakespeare-discuss@ 邮件列表中启用了新发现的前文全文。
- 2015-10-21 15:10 UTC martyn
  - 向 Shakespeare corpus 中增加新的前文，同时启动 MapReduce 索引任务。
- 2015-10-21 15:21 UTC jennifer
  - 将所有流量导向 US-A-2 资源集群，同时将其他集群下线，以便让这些集群从连锁故障中恢复，同时启动重索引任务。
  - MapReduce 索引任务完成，导致 Bigtable 复制响应在集群。
- 2015-10-21 15:28 UTC jennifer
  - 全球服务容量提升为 2 倍。

418 | 附录C 事故状态文档示范

大模型在AIOps领域的应用落地路径

避免过于乐观：运维大模型仍面临不少挑战

杜绝幻觉、可解释性强、私有部署开销低、私有语料质量数据均不足、融合存量知识、工具、多模态数据、通识大语言模型底座不易选择

避免过于悲观： 挑战都可解

模型分层：通识大语言模型、运维大语言模型、私有部署运维大语言模型

区分、整合非结构化大模型与结构化大模型

关键组件

- 运维大语言模型
- 多模态基础模型群
- 结构化大模型：左脑AIOps算法、右脑运维图谱
- 通过检索融合本地知识库
- 智能体&基础模型编程框架

运维大语言模型是核心基础

- 检索增强、有据可依
- 课程学习、由易到难
- 知识工程：从文档到知识图谱，增加显示知识
- 检索本地知识库、降低模型精度
- 与通识大语言模型底座尽量解耦

应用及路径建议：小步快跑、以用促建

应用的定位：从助手、教练、顾问、参谋到内部专家

近期	近中期	中长期
数字化运维助手，私有运维文档问答，运维脚本解读，运维数据注释	为单个运维工具提供自然语言交互增强	基于智能体，编排多个工具完成更复杂运维任务

谨慎乐观

大势所趋、前景可期、机遇与挑战并存、协同创新、以用促建

AI人工智能产业链联盟

#每日为你摘取最重要的商业新闻#

更新 · 更快 · 更精彩

- Zero
  AI人工智能产业链联盟创始人
  河北清华发展研究院智能机器人中心运营经理
- base:北京
- 主要业务:AI商业化答疑及课程应用场景探索，各类AI产品学习手册，答疑及课程

欢迎扫码交流

提供：学习手册/工具/资源链接/商业化案例/行业报告/行业最新资讯及动态

ifenxi

专 注 数 字 化