

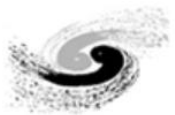


大语言模型在台山中微子实验值班系统中的研究与应用

报告人: 裴文溪

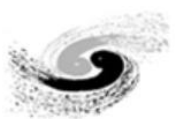
2024年07月16日

山东 青岛



主要内容

- ◆ 1、大语言模型的具体应用
- ◆ 2、台山中微子实验（TAO）数据获取系统（DAQ）的值班系统
- ◆ 3、大语言模型和DAQ值班系统的结合
- ◆ 4、总结



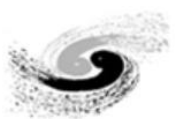
大语言模型 (Large Language Model, LLM)

- ✓ 提升自然语言处理的能力和效果
- ✓ 推动人工智能技术在文本处理领域的发展和应用
- ✓ 为各种应用场景提供更加智能、自然的文本处理解决方案



应用场景：高能物理领域如何？





AccGPT——CERN的聊天机器人

现在的优势:

关注检索增强生成 (RAG)

- ✓ 专业领域问询较为精准
- ✓ 协助用户实现操作查询
- ✓ 实现编码提醒工作

AccGPT的UI界面



ATLAS high precision measurement paper released in September about which force?

The ATLAS high precision measurement paper released in September is about the strong force.

Further readings:
<https://home.cern/news/news/physics/atlas-measures-strength-strong-force-record-precision>

AccGPT的问答实例

Origin country of CERN's ESA astronaut Slawosz Uznanski is Poland.

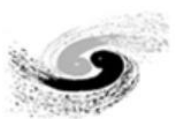
CERN app for CERN map, restaurant menus and phonebook

The CERN Campus app provides a wide range of information, accessible via your CERN login, including

Launched this year, which space telescope relies on CERN computing infrastructure to investigate the effects of dark matter and dark energy in the Universe??

The space telescope that relies on CERN computing infrastructure to investigate the effects of dark matter and dark energy in the Universe is Euclid.

Further readings:
<https://home.cern/news/news/experiments/cern-tech-help-investigate-dark-universe>



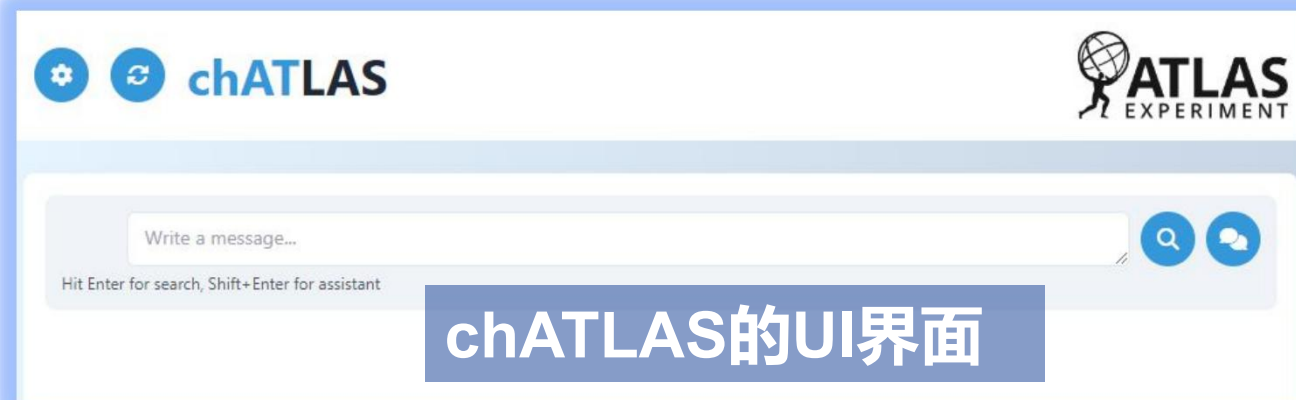
LLM 在高能物理领域的应用

chATLAS ——ATLAS合作组的AI助手

现在的优势:

关注RAG所需的向量知识库

- ✓ 整理**实验**相关信息
- ✓ 保证信息的私密性和安全性
- ✓ 持续更新知识库



Name	Size
hn-atlas-lar-electronic-calibration.txt	304.5 MB
hn-atlas-triggerReleaseValidation.txt	84.6 MB
hn-atlas-TDAQCommissioning.txt	81.1 MB
hn-atlas-tile-performance.txt	68.8 MB
hn-atlas-dist-analysis-help.txt	39.2 MB
hn-atlas-muonSWBugMonitor.txt	26.6 MB
hn-atlas-data-quality-operations.txt	18.1 MB
hn-atlas-offlineSWHelp.txt	17.4 MB
hn-atlas-PATHelp.txt	17.2 MB
hn-atlas-SITInternal.txt	8.2 MB
hn-atlas-exotics-wg.txt	6.9 MB
hn-atlas-jet-etmiss-wg.txt	6.5 MB

chATLAS知识库内容

748	How can I download a file with Rucio ?	To download a file with Rucio, you can use the following command: "" Srucio do
749	Are you working?	The provided context does not contain specific information about whether I am
750	How can I store my analysis data at BNL ?	You can store your analysis data at BNL by utilizing the BNL-OSG2_GROUPTAPE
751	how can	I'm sorry, but the question seems to be incomplete. Please provide more inform
752	how can I add rooftop	I'm sorry, but the provided context from the ATLAS documentation and wiki pag
753	how can I add a poi in rooftop	I'm sorry, but the given context does not provide information related to adding
754		What analysis was performed on the 2010 data in the ATLAS experiment, and wh
755	How much data did we take in the year 2015?	Based on the available context, the document does not provide information abo
756	What isolation working points are available for single lepton tri...	The available isolation working points for single lepton triggers include: - FixedC
757	how to create an elementlink to an electron	To create an ElementLink to an electron, you can follow the example provided in
758	What is the Zprime particle in BSM theory?	The Zprime particle in BSM (Beyond the Standard Model) theory is a hypothetica
759	In a Tag and Probe trigger, how is the topocut (e.g. 0bDRAB) us...	The information provided in the available context does not specifically address h
760		If the probe leg fails all hypothesis steps in a Tag and Probe trigger, it means tha
761		In a Tag and Probe trigger scenario, if the probe leg fails all hypothesis steps, the
762		What analysis was performed on the 2010 data with 33pb ⁻¹ , and what were the
763	What are the main features of Atlas Inner Detector ?	The main features of the ATLAS Inner Detector include: - Improved geometry for
764	What are coherent prescale sets?	Coherent prescale sets are groups of triggers within the ATLAS experiment that s
765	Describe coherent prescale sets	Coherent prescale sets, also referred to as prescaler groups, were introduced in t
766	How long would it take to generate 1 million full simulation ev...	Based on the provided context, it is not possible to determine how long it would
767	How long does the ATLAS fast chain event simulation take, per ...	Based on the provided context, the document mentions that the ATLAS fast chai
768	How long does a full delphes simulation of one full pileup even...	Based on the provided context, it is not specified how long a full Delphes simula
769	How long does a full delphes simulation of one full pileup even...	Based on the available context, I cannot provide a specific answer to your questi
770	For a run of the ATLAS experiment, how can I check if a defect ...	To check if a defect assigned to a run for data quality purposes is tolerable or int

支撑资料来源: Daniel Thomas Murnane.ChATLAS: developing an AI assistant for the ATLAS collaboration.IML Machine Learning Working Group



ATLAS和CERN都关注的（检索增强生成）RAG到底是什么？

GLM-3

ROS节点data_aggregator报告channel_2没有数据

当使用ROS (Robot Operating System) 时, 如果在节点中报告“channel_2没有数据”, 这通常意味着该节点正在订阅的话题中没有接收到预期的话题数据。为了解决这个问题, 你可以按照以下步骤进行排查:

1. 确认话题存在: 确保你期望的数据话题确实在发布。可以在ROS的监控终端 rqt_graph 中查看话题的网络结构, 或者使用 rostopic list 命令列出当前运行的所有话题。
2. 检查话题名称: 确认你在订阅节点中指定的话题名称是否正确, 包括大小写和拼写。
3. 检查节点连接: 使用 rqt_graph 或 topic_tools 包中的 /rosout 来检查节点的连接情况, 确保你的订阅节点能够连接到发布话题的节点。
4. 检查话题类型: 确保订阅节点的消息处理程序与发布的话题类型相匹配。
5. 检查话题发布顺序: 确保在订阅节点中, 你期望接收的话题是在其他话题之前发布的。你可以尝试调整话题的发布顺序。
6. 增加调试输出: 在订阅节点中, 你可以使用 rostopic echo 命令来查看接收到的话题数据。你还可以使用 rostopic pub 命令来手动发布话题数据, 以测试订阅节点是否能够接收到数据。
7. 检查话题权限: 确保你的订阅节点有权限接收该话题的数据。在某些情况下, 权限问题可能导致节点无法接收数据。
8. 检查网络问题: 如果节点不在同一个计算机上, 确保网络连接没有问题。
9. 查看错误日志: 在终端中使用 rostopic hz /channel_2 (假设这是你的话题名称) 查看话题的发布频率, 或者使用 roscore 命令启动ROS核心服务时, 会有关于节点和话题的错误日志输出, 仔细阅读这些信息可以帮助定位问题。

如果上述步骤不能解决问题, 可能需要具体查看你的节点代码和系统配置, 进一步分析问题所在。在这种情况下, 分享具体的节点代码和环境配置可能会有所帮助。

☆ 🔄 📄 📌

如何检查节点是否正确连接到话题?

如何在订阅节点中增加调试输出?

如何确保我的订阅节点有权接收话题数据?

原生的LLM

Chatbot

DAQ专家模拟提问内容

ROS节点 data_aggregator 报告 channel_2 没有数据

根据给出的上下文信息, 故障情况是ROS节点data_aggregator报告channel_2没有数据。根据经验, 这可能是因为网络问题或交换机问题导致的。首先需要检查网络连接是否稳定, 以及交换机的配置信息是否正确。如果问题仍然存在, 建议重新配置或重启GCU (通用驱动程序)。

输入文本并回车, 或上传本地文件

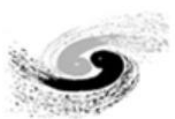
文件

RAG精准检索模拟的正确答案

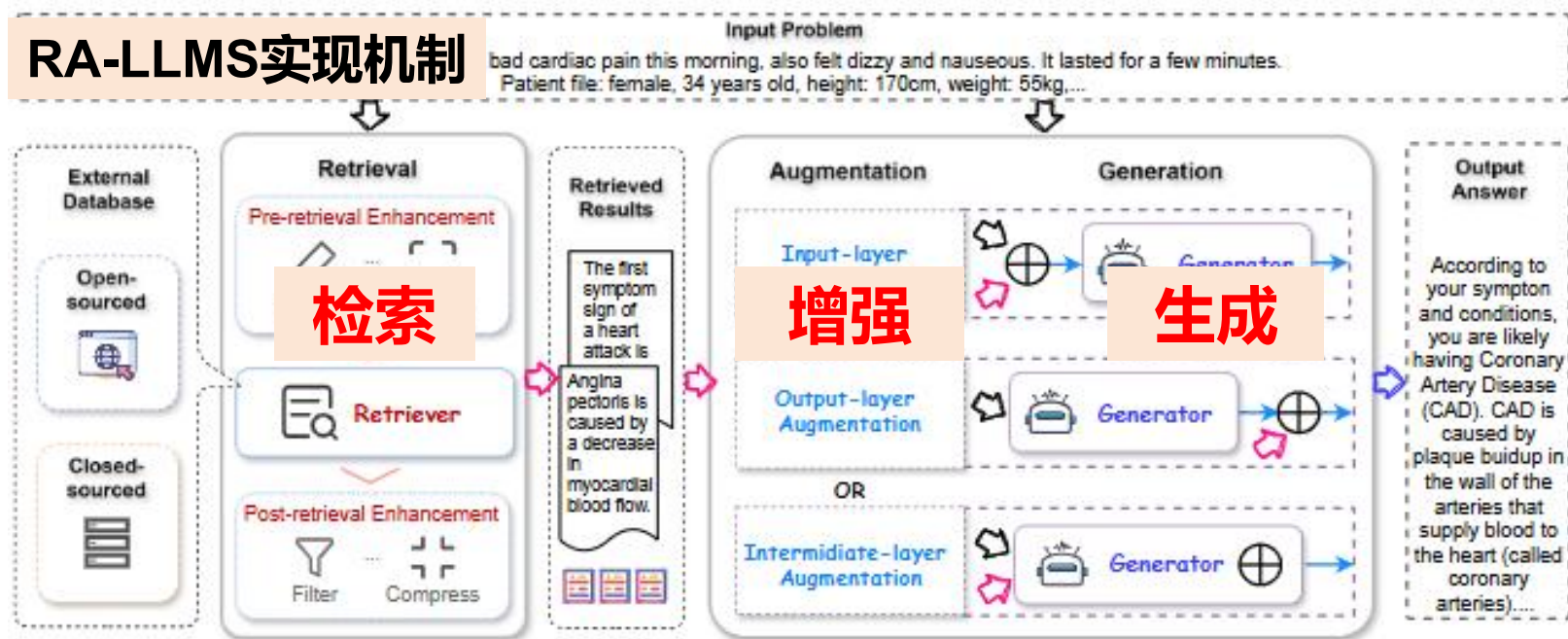
提问: ROS节点data_aggregator报告channel_2没有数据 (模拟)

回答: 首先检查网络配置、建议重新配置或者是重启GCU (模拟)

引入RAG链路的LLM



RA-LLMS (检索增强大语言模型)



上图相关内容来源: <https://arxiv.org/abs/2405.06211>

LLMs局限性:

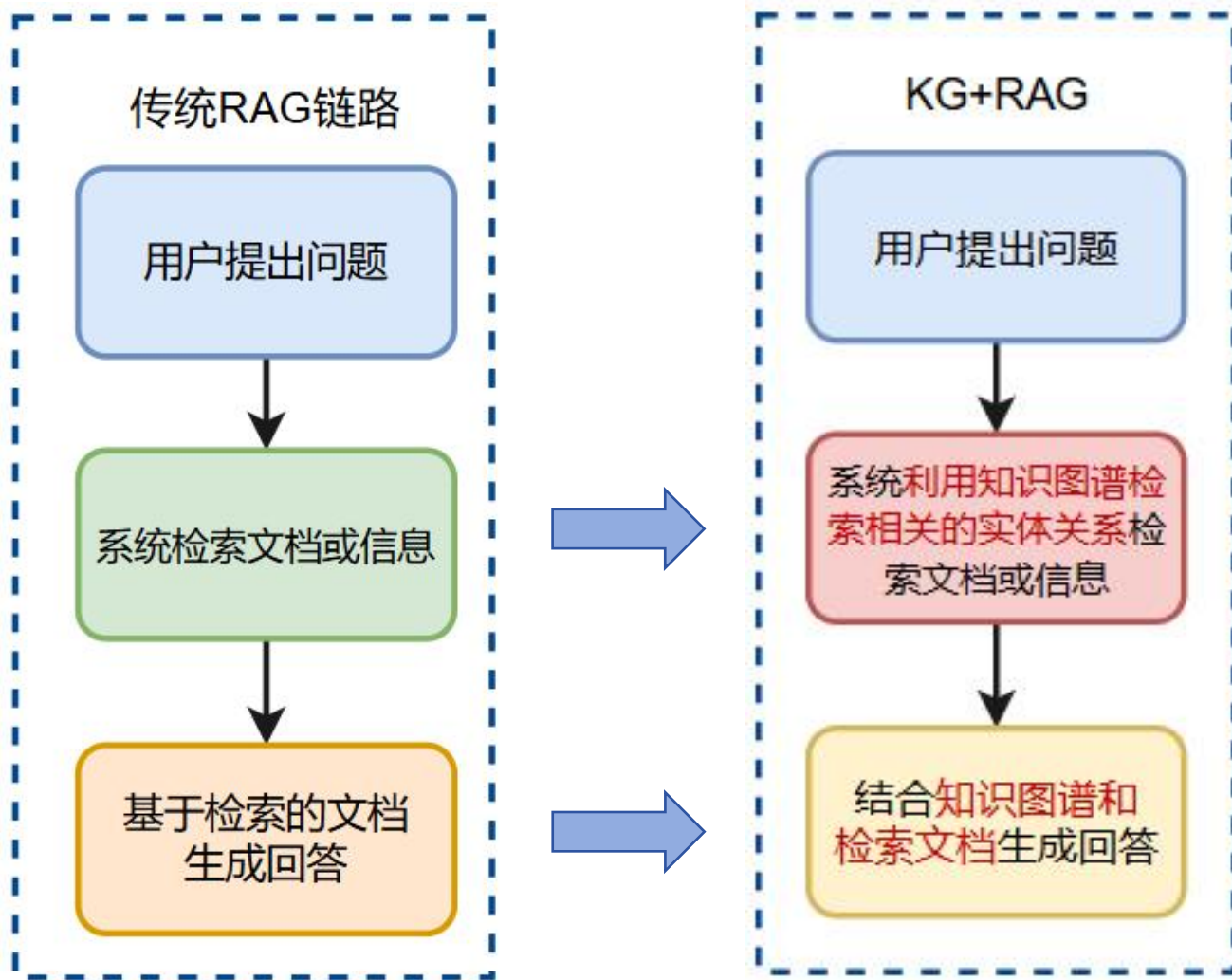
- 1.大语言模型的“幻觉”
- 2.过时的“知识”
3. ...

RA-LLMS的优势:

- 1.缓解LLMs的局限性



结合知识图谱 (KG) 的RAG的回答流程:



微软推出了GraphRAG: 【开源】

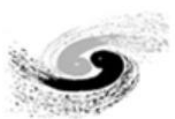
<https://github.com/microsoft/graphrag>

可能出现的问题:

- 检索不准确
- 回答不灵活缺乏深度

改进后的效果:

- 减少幻觉
- 增强语义效果
- 增强鲁棒性

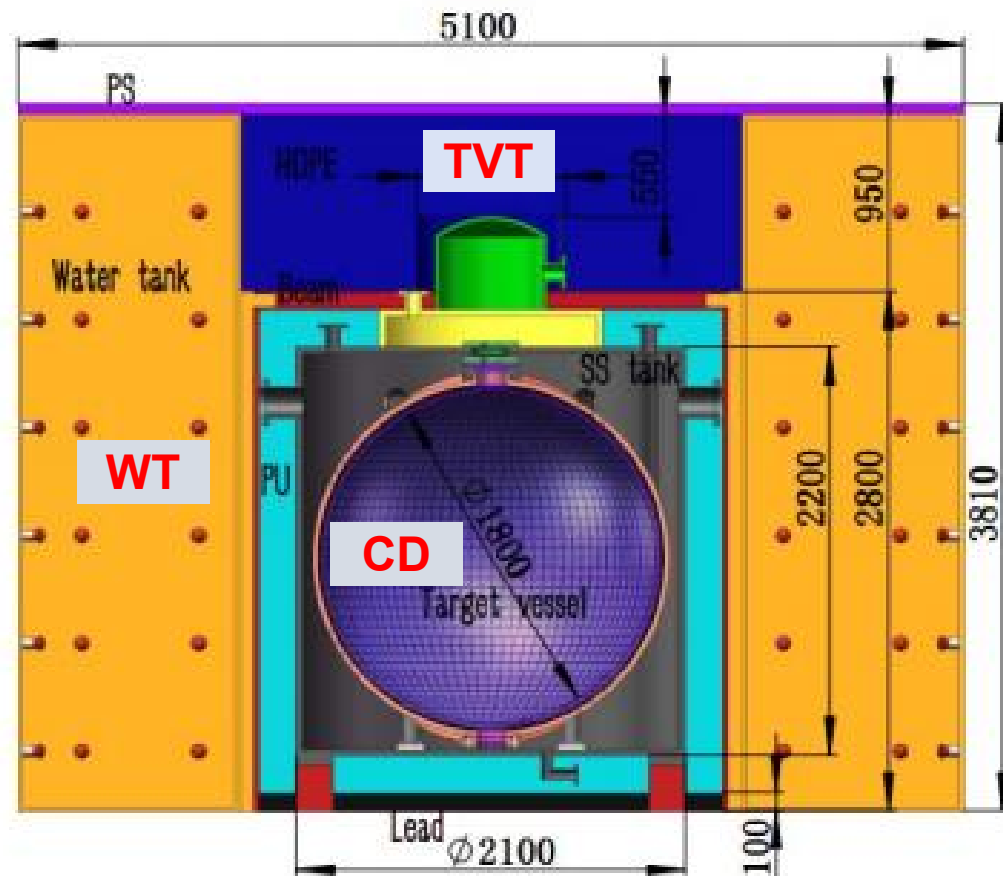


台山中微子实验简介

台山中微子实验 (Taishan Antineutrino Observatory)

—— JUNO的子实验

- 测量高能量分辨的反应堆中微子能谱
- 为江门中微子实验**提供参考能谱**
- 为核数据库提供检验基准





TAO数据获取系统 (DAQ) 介绍

- 包含**数据流软件**和**在线软件**
- 包含运行控制/信息共享等任务
- 需要长期稳定的运行和持续存储

可支持DAQ联调、取数过程

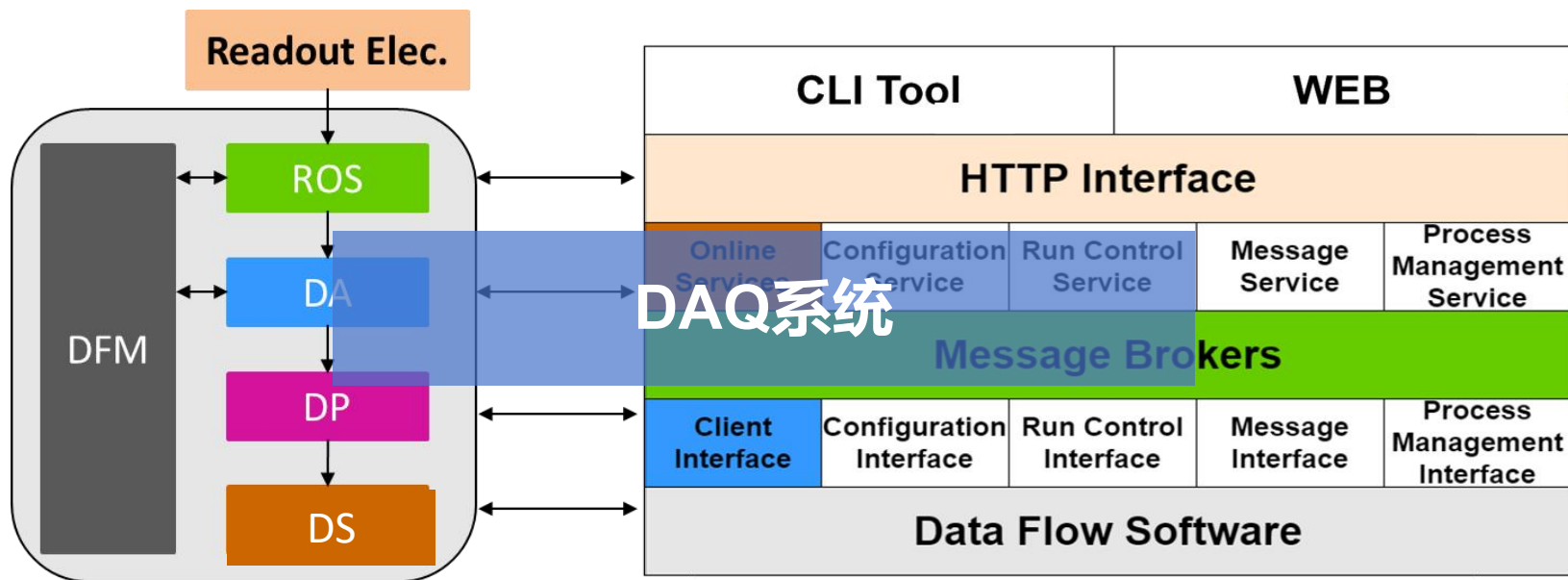
运行维护的服务

值班系统[初步实现功能]

- ✓ 实现**远程**在线**监控**实验数据
- ✓ 实现实验故障及**异常**自动**告警**
- ✓ 整理**专家经验问答对**
- ✓ **智能化地分析**各类异常



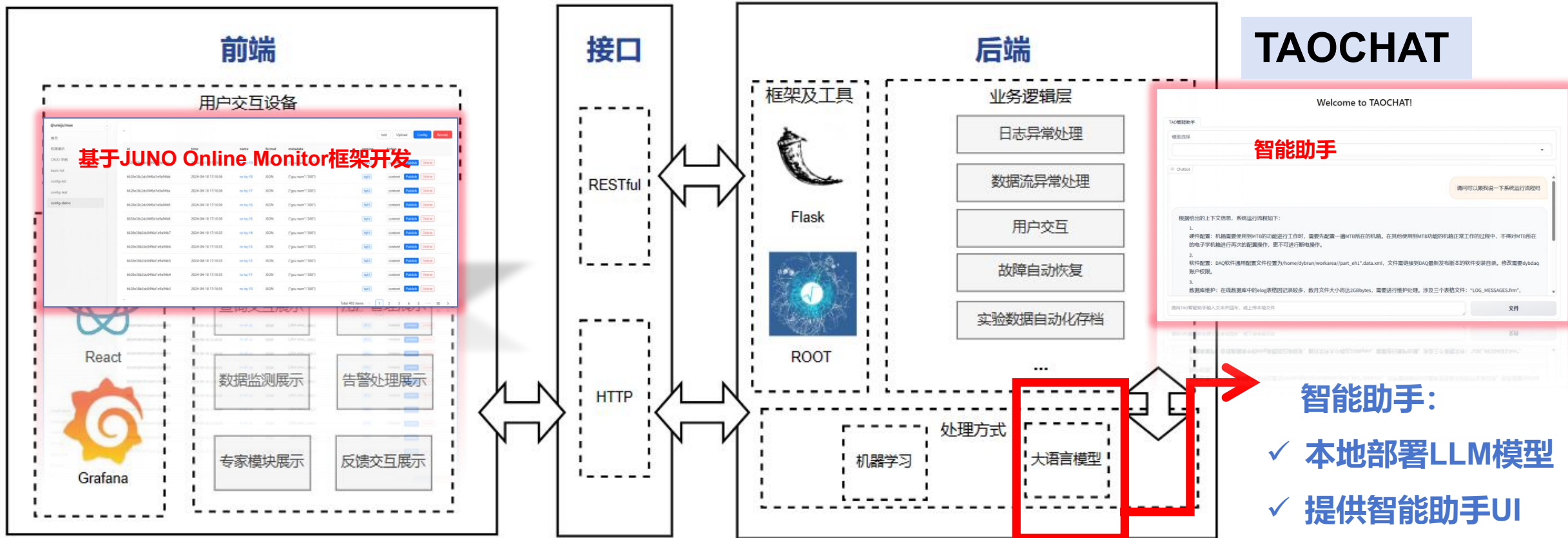
提供Web平台实现上述服务





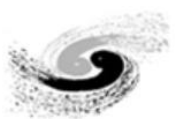
TAO DAQ 值班智能助手 (TAOCHAT) 的设计

TAO DAQ值班系统架构

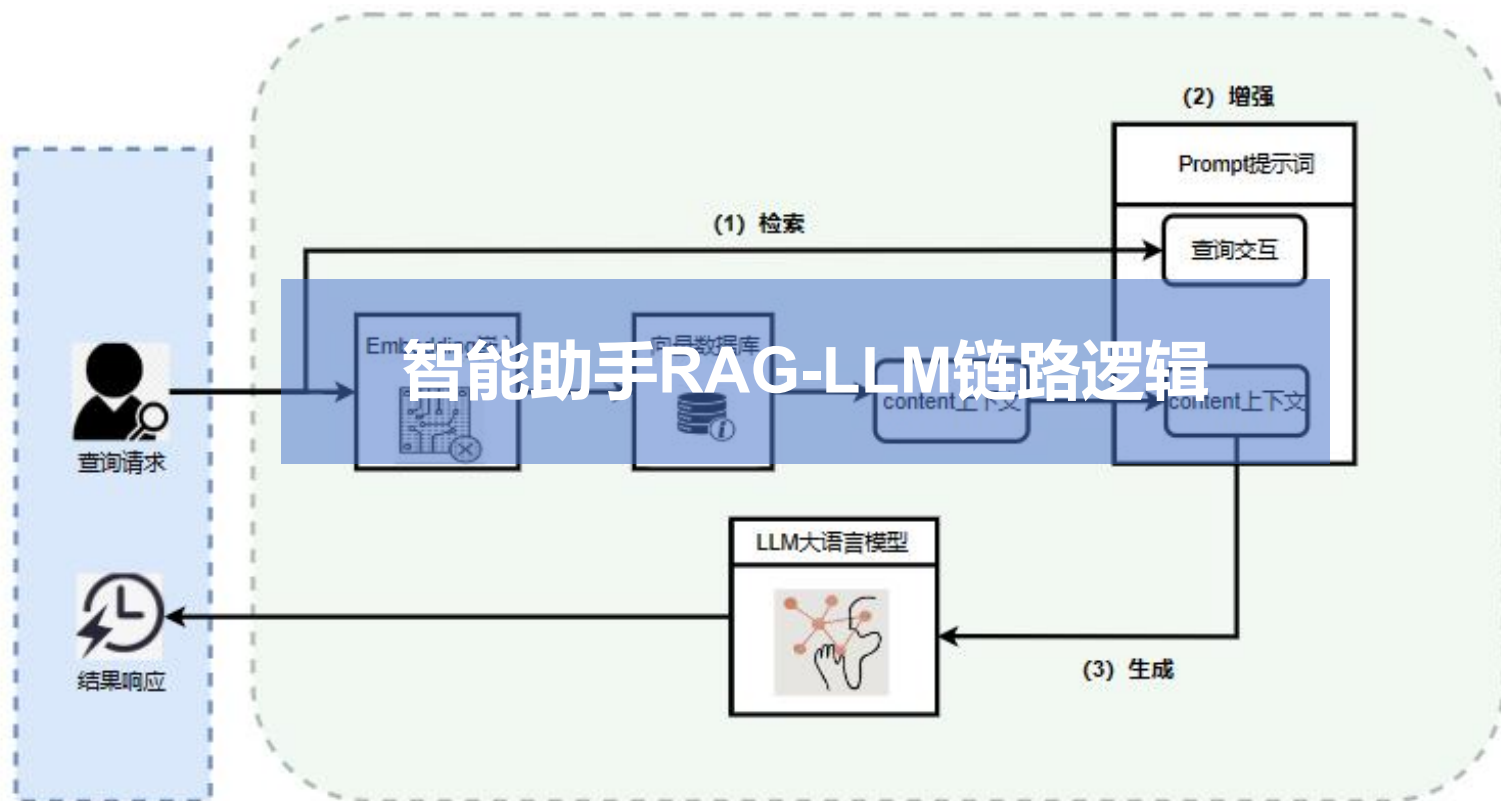


✓ 提供Web平台进行实时信息监测——“信息”传递出系统情况

- 智能助手:**
- ✓ 本地部署LLM模型
 - ✓ 提供智能助手UI
 - ✓ 内嵌私域知识库



TAOCHAT——本地部署LLM及实现RAG链路



TAOCHAT:

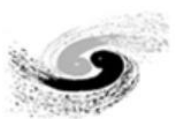
- 实现本地部署: Chatglm3-6B
- 语义向量模型: bge-large-zh-v1.5
- 向量数据库: Chroma

目前实现功能:

- 实现问询UI界面及较为精准的问询
- 私域知识库的后台增删管理
- 自动化整理问询记录

后续优化:

- 提高检索准确性
- 实现知识库可视化管理方式



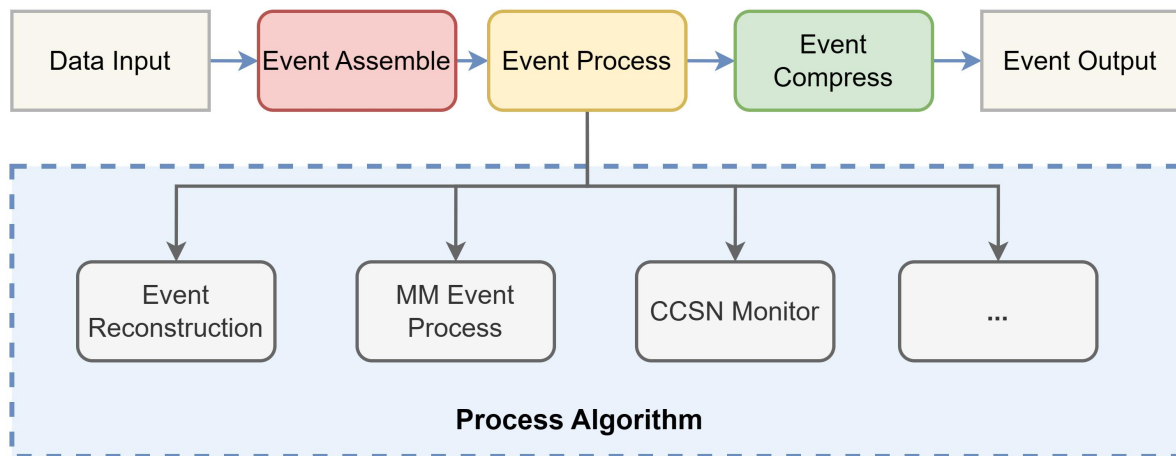
LLM结合KG方式增强回答——故障根因分析

DAQ具有故障之间的相关节点关系

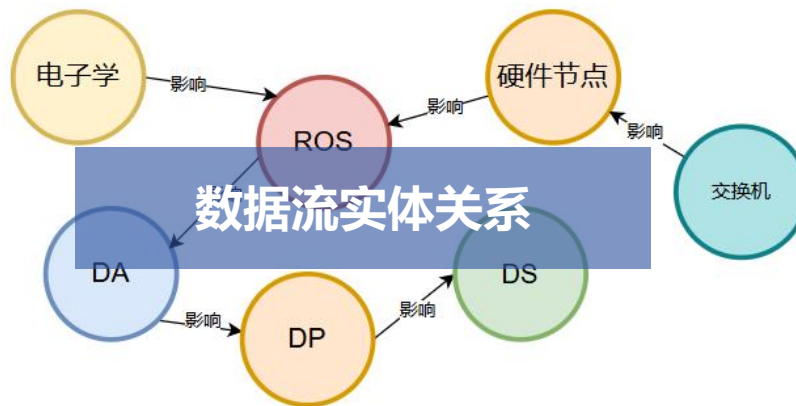
DAQ软件内含很多实体关系

诸多节点: ROS、DA、DP、DS

节点关系: 上下节点影响、反映上级问题

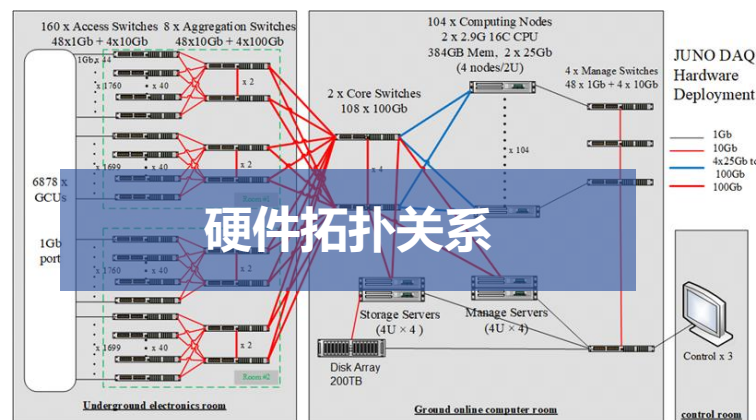
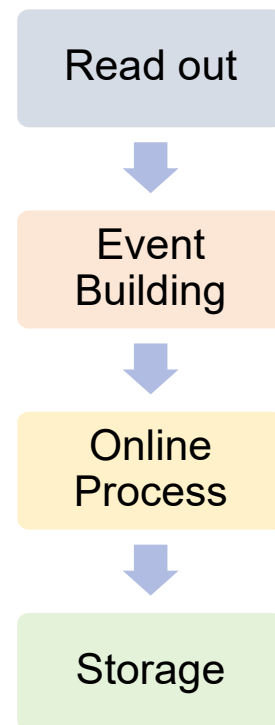


Radar框架



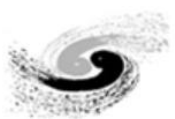
数据流实体关系

数据流



硬件拓扑关系

实体关系



TAOCHAT——功能介绍

目前通过gradio实现TAOCHAT的UI界面



可以选择不同的模型进行判断

提问与DAQ相关内容

实现私域知识较为精准的回答

此处上传问题和知识库文件

目前已嵌入TAO值班Web平台
后续面向合作组开放

结果和私域名词密切相关



TAOCHAT——RAG外嵌知识库

TAO值班知识库内容:

- DAQ常用值班信息
- DAQ专家整理的值班手册
- DAQ专家标注的既往值班故障解决办法
- ...

正式运行后知识库会动态更新补充...

私域值班信息内容片段

DAQ 系统运行与维护手册

V1.0, 2012.8

1 硬件安装与维护

DAQ 所有设备均可保持长期运行, 除非设备有故障或损坏, 一般不需要额外维护。硬件的按照和维护, 详见各设备厂家说明书, 本节只描述这些设备和配置的基本使用方法。

服务器和存储设备需要每年购买了现场保修服务, 设备序列号和服务期限详见硬件和网络网页: <http://dayabay.ihep.ac.cn/twiki/bin/view/Internal/OnsiteNetwork>。

1.1 系统启动与关闭顺序

启动顺序:

- 1) 磁盘阵列加电;
- 2) 加电启动 daqs1 管理服务器, 至出现登陆界面;
- 3) 启动其它所有计算机。

准确问答效果

Chatbot

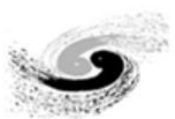
设备序列号和服务期限硬件和网络网页是什么

设备序列号和服务期限可以在硬件和网络网页上找到, 具体网址为: <http://dayabay.ihep.ac.cn/twiki/bin/view/Internal/OnsiteNetwork>。在网页上, 可以找到相应的设备信息, 包括设备序列号和服务期限。

模糊问答效果

保修服务的网页是什么

保修服务的网页是: <http://dayabay.ihep.ac.cn/twiki/bin/view/Internal/OnsiteNetwork>。



TAOCHAT——知识库管理及检索

➤ 调研知识库检索方式：【后续实现】

➤ 知识库管理：持续迭代升级

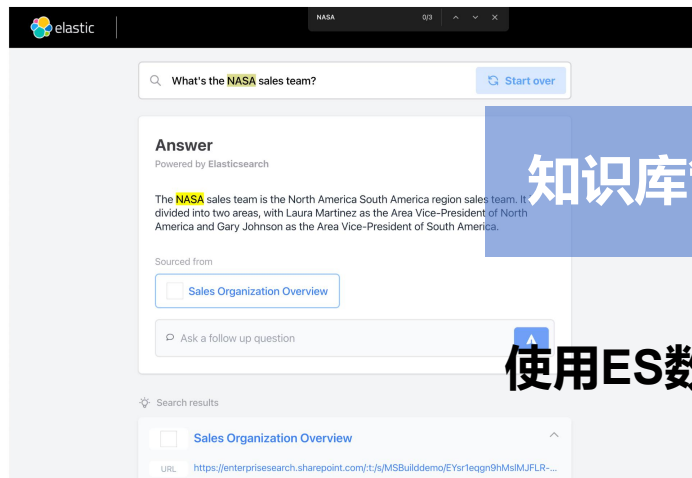
目前初步实现：

- TAOCHAT可以在后端增删数据文档
- 支持格式pdf.csv.json



后续实现：

- 对知识库的管理实现页面可视化
- 支持更多格式



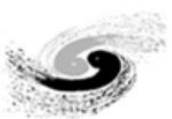
知识库管理及检索可视化方式

使用ES数据库结合RAG进行向量搜索

➤ 初步测试通过提示词实现问询结果来源

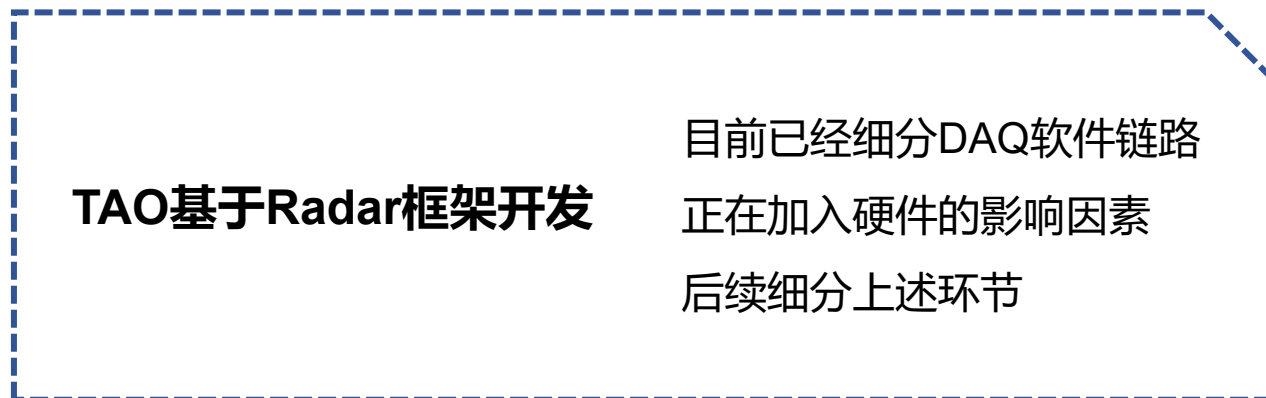
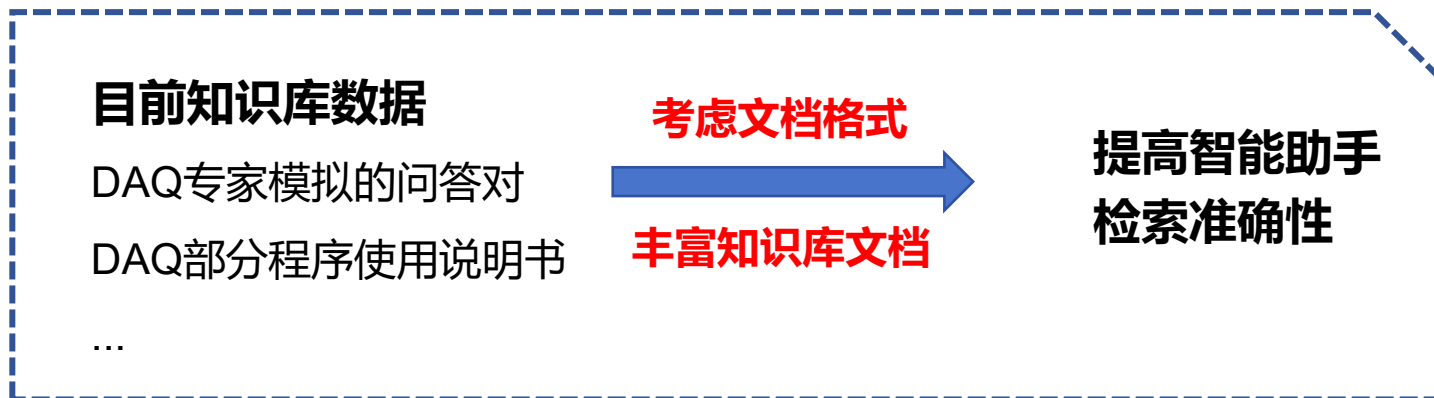


通过关键词提示问题回答来源



TAOCHAT优化的方向

- 知识库数据集优化【正在进行】
- 提升RAG链路准确性【待开发】
- 增加知识库管理级可视化方式权限【待开发】
- 细化故障分析的实体关系【正在进行】





大语言模型可高效应用于各领域的部分任务

基于上述研究情况**实现TAOCHAT的基础功能**

可实现DAQ联调取数的**运行维护**

实现LLM能力与**实验**的结合运用

目前是**初步实现**阶段，后续会继续**优化升级**

该功能可推广至各个实验及各类场景



感谢各位老师，请批评指正

特别感谢：张正德老师在系统研发中给予的帮助