



十二月研究生季度考核

报告人：胡黄腾

指导老师：朱科军、季筱璐

专业：计算机技术

课题组：触发与数据获取组

时间：2024年12月27日

主要内容

- ◆ BES III 智能数据平台RAG链路开发与优化
- ◆ BES III 智能数据平台硬件部署方案的研究
- ◆ 其他工作
- ◆ 后续工作



工作需求与设计

BESⅢ运控数据管理问题

数据的高度分散性
缺乏有效的质量检查机制



- 难以迅速定位问题根本原因
- 存储空间浪费和排查问题低效

目标:

- 大幅减轻值班人员压力
- 减少运行维护专家工作负荷
- 提高运行效率和数据质量



数据异常检测

- ✓ 机器学习训练模型
- ✓ 阈值评估



日志问题分析

- ✓ RAG+LLM文本分析
- ✓ 值班人员查询交互

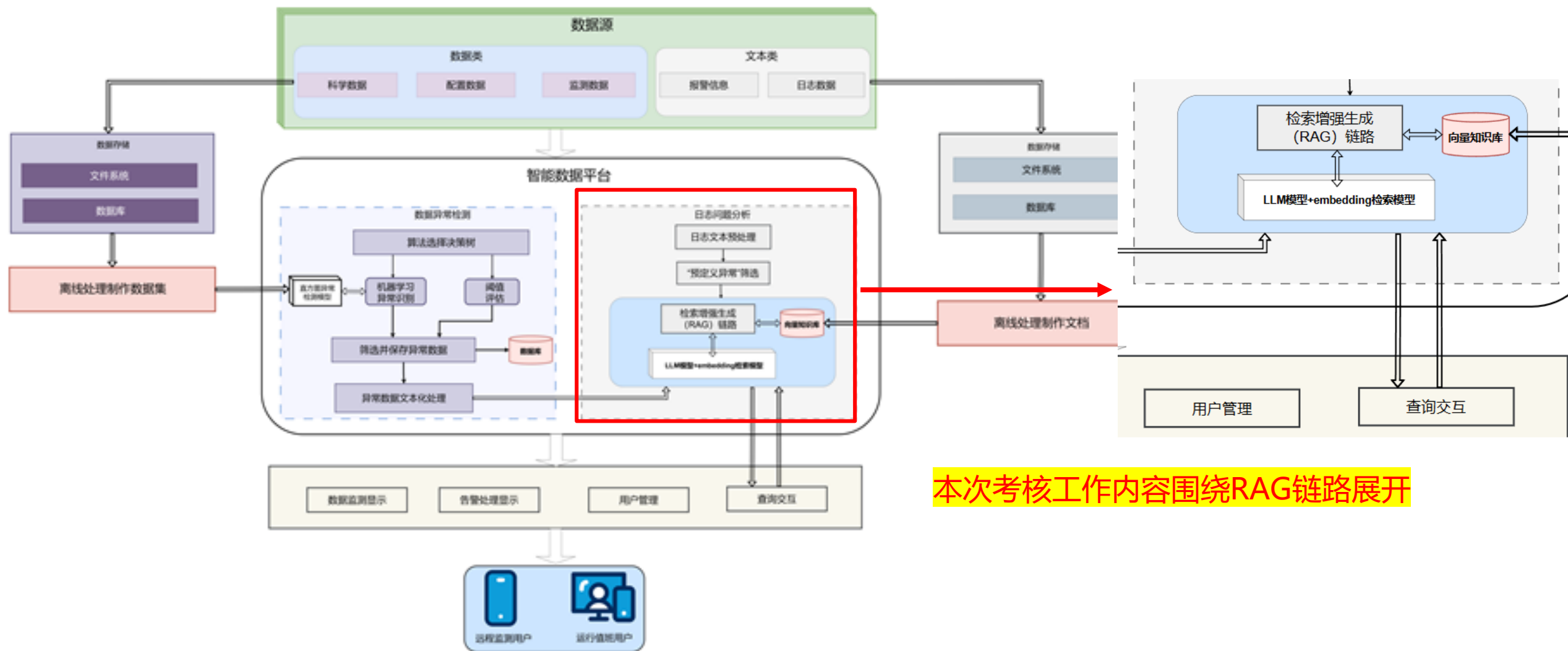


BESⅢ智能数据平台



工作设计架构

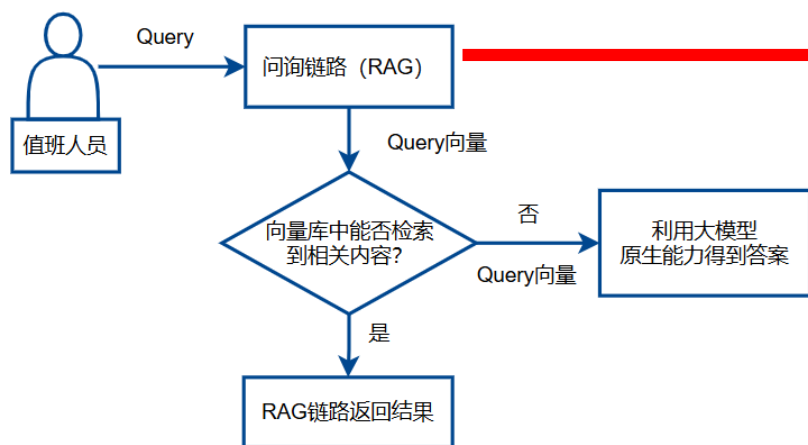
BESIII智能数据平台架构图



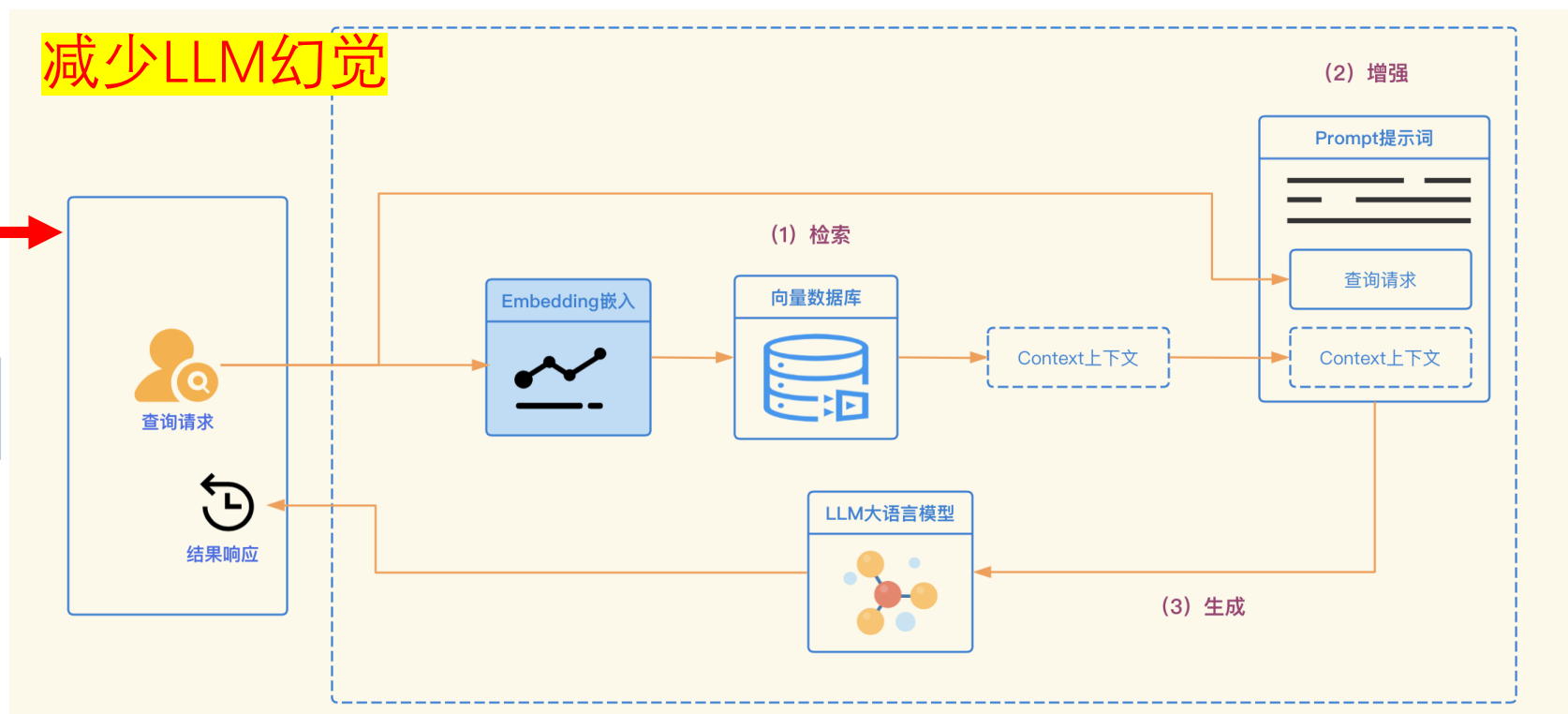


工作内容

➤ 用户查询交互流程图

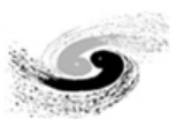


➤ 问询链路（RAG）实现



利用检索私域数据库内容提供给LLM获得相关答案

问询链路RAG进行实现



功能实现——RAG检索内容溯源

➤ RAG系统内容溯源缺失的挑战

- 模型生成的内容来源无法追踪，可能包含错误或不恰当信息
- 无内容出处，用户对内容产生怀疑

➤ RAG系统内容溯源的价值

- 对返回信息给出来源链接
- 快速定位、验证信息真实度

RAG文本对话

模型选择

Chatbot

检索内容溯源功能实现

ATLAS数据获取系统是欧洲核子中心大型强子对撞机（LHC）上重要探测器之一的ATLAS实验所使用的系统。该系统是为了应对高亮度LHC（HL-LHC）的高亮度实验条件而进行升级的。

它包括以下几个主要功能模块：

1. 读出单元：由前端链路交换（Front-End Link Exchange，简称FELIX）和数据处理器组成。
2. 数据流模块：负责数据的传输和处理。
3. 网络连接：使用千兆网络作为数据处理器和探测器控制系统（Detector Control System，简称DCS）之间的接口。
4. 通用在线软件框架：支持整个数据获取过程的管理和控制。

其中，FELIX用于提供与探测器电子学系统的定制化接口，并通过千兆网络充当数据处理器和探测器控制系统的接口。数据处理器则通过千兆网从FELIX接收数据，并在将数据发送至数据流系统之前对其进行特定处理，如格式化和监控。

文档来源：

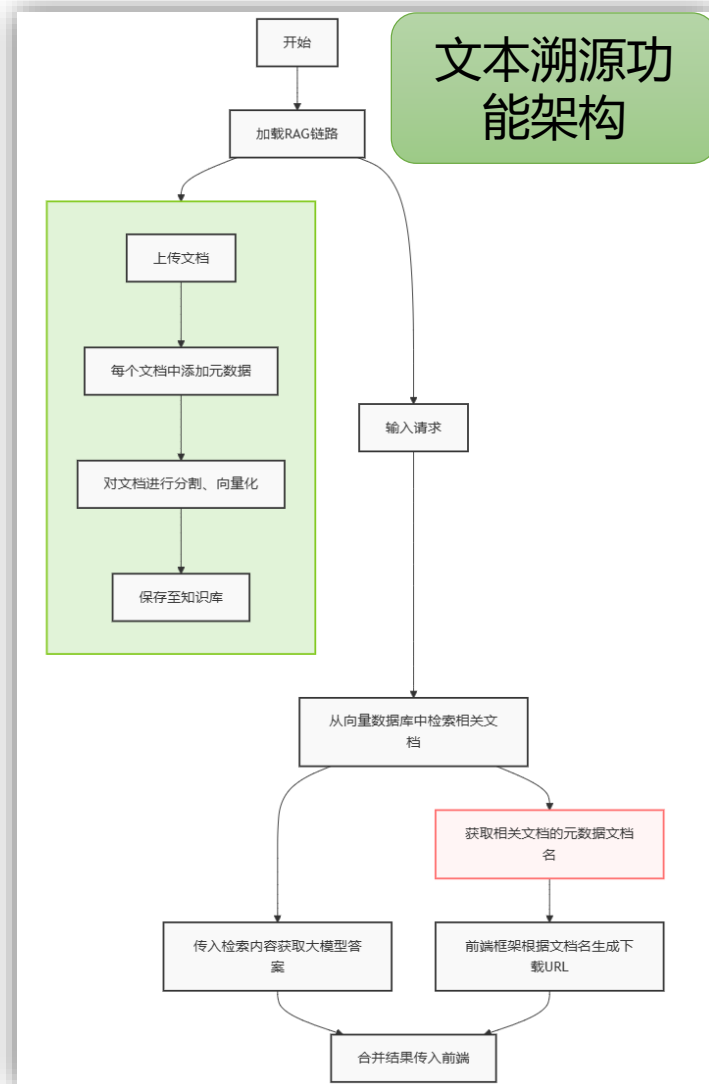
- 硅像素径迹探测器原型机数据获取系统的研制.pdf

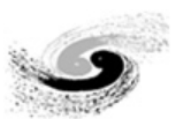
硅像素径迹探测器原型机数据获取系统的研制.pdf
[打开文件](#)

面向CEPC顶点探测器预研的读出电子学关键技术研究(2...
[打开文件](#)

第3章 关系数据管理与分析V2.0.pdf
[打开文件](#)

[查看更多](#)





功能实现——RAG链路文件管理页面

➤ RAG文件管理页面实现目的

- 减少人工多渠道维护文档带来的**重复和混乱**
- 一站式操作界面让管理者快速处理文档
- 对文件内容进行**及时更新**，有助于 RAG 检索的结果准确性

➤ RAG文件管理页面实现前后对比

对比项	实现前	实现后
管理方式	无统一的前端页面，依赖后端脚本手动操作	采用 Streamlit 统一前端界面， 集中化管理文件
操作便捷度	上传/删除文档需要命令行或后台操作，复杂度高	上传/删除/编辑 在前端页面一键完成，大幅降低使用门槛
文件格式支持	仅支持PDF文件格式，扩展性不足	多文件格式上传解析 （PDF、Docx、xlsx），适应不同文档场景
实时编辑能力	无可视化编辑文档内容的功能，需要离线处理	文档内容编辑（完善中）可直接在页面上编辑核心文本， 大幅缩短迭代周期

➤ RAG文件管理页面实现内容

文件管理

上传文件

选择要上传的文件

Drag and drop file here
Limit 10GB per file

Browse files

服务器用户手册.doc 148.5KB

X

上传文件

文件上传成功: 服务器用户手册.doc

文件管理

上传文件

选择要上传的文件

Drag and drop file here
Limit 200MB per file

Browse files

专家经验对.pdf 77.2KB

X

上传文件

删除文件

选择文件

专家经验对.pdf

删除文件

编辑文档

选择要编辑的文件

专家经验对.xlsx

加载文档

文档部分 1

试执行 daq_rst。
多次同步失败
多次同步失败。解决办法：检查黄色网线连接状态是否正常。
稳定运行时如何监测时钟同步

文档部分 2

DAQ 链路不通
数据采集链路问题。解决办法：执行 ping 测试，检查交换机端口的指示灯，确保网络连接正常。
GCU 和 BEC 同步失败

文档部分 3

GCU 和 BEC 同步失败
可以暂时去掉 GCU。
IOC 配置失败
IOC 配置问题。解决办法：确保使用正确的高压配置文件。检查 GUI 路径和操作步骤。

多文件格式支持

实现可视化编辑文档



功能实现——RAG 回答判断处理机制

大模型幻觉问题

- 大语言模型（LLM）当缺乏充足上下文，容易出现“幻觉”回答
- 缺乏可靠依据，降低用户的信任度

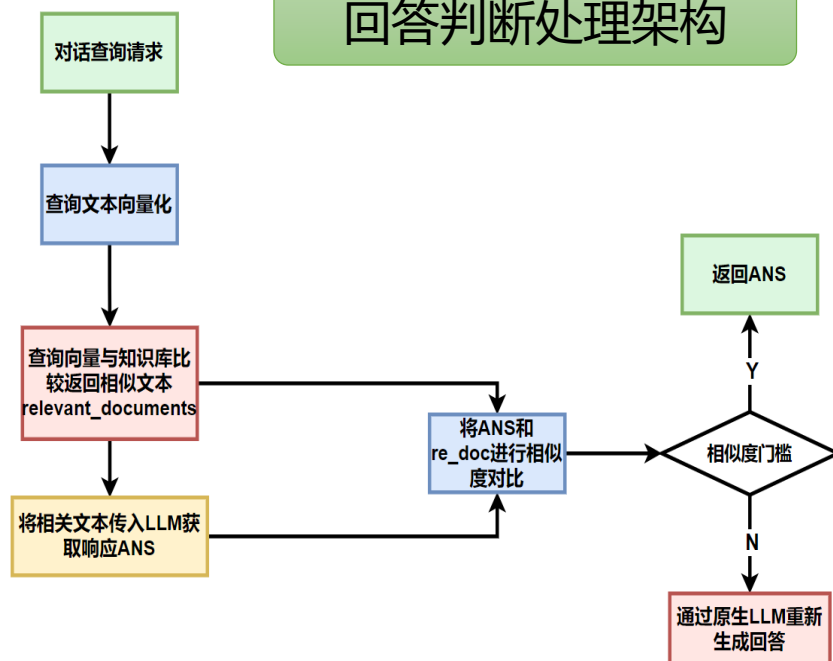
解决思路：答案相似度判定

- RAG检索内容与LLM的回答进行相似度比较
- 设定阈值区分是否根据文档内容回答

功能价值

- 标记出原生 LLM 自行推断的场景，减少用户对幻觉答案的误解
- 提升用户对系统回答的透明度与信任度

回答判断处理架构

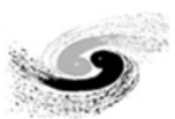


区分答案来源

根据文档内容无法回答，我将给出一个基于LLM的回答：
调整鼠标DPI（每英寸点数）可以让你在使用鼠标时更舒适和高效。以下是Windows系统中调整鼠标DPI的方法：
对于Windows系统：
1. 打开设置：点击左下角的开始按钮，然后选择“设置”图标（齿轮形状）。
2. 进入设备设置：在设置窗口中，找到并点击“设备”。

标记原生LLM

减少大模型幻觉提升可信度



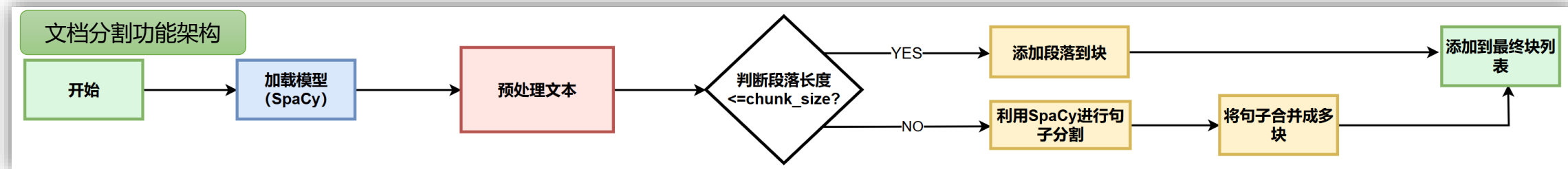
RAG 链路优化处理：提升检索能力与上下文质量

➤ RAG链路性能优化目的

- 提高对不同类型文档的检索和分割处理能力
- 提供更高质量的上下文给大模型，减少幻觉、提高可信度

➤ 优化点一：数据切分策略改进

- 原有问题：对**英文文档处理效果较差**，影响检索与上下文匹配
- 解决方案：采用**语义分割**进行文本分割
- 效果：在检索时，段落的完整性与上下文连贯性更佳，进一步**提高回答准确度**



优化数据切分提高上下文质量



RAG 链路优化处理：提升检索能力与上下文质量

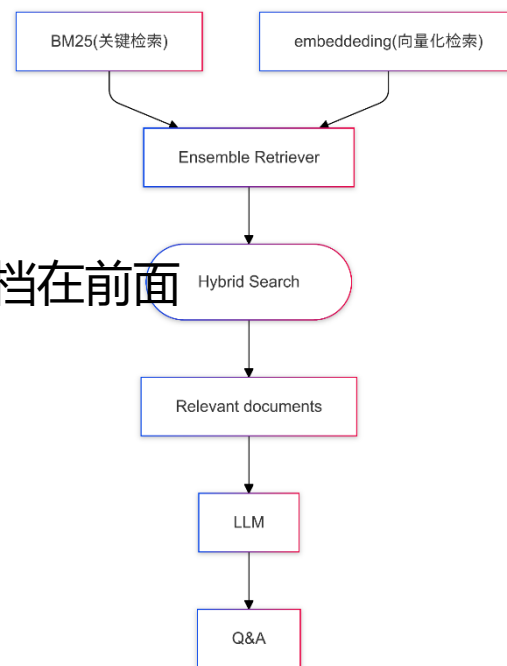
➤ 优化点二：增加混合检索（BM25+Embedding）进行多路召回

- 原有缺陷：采用单一检索不够精准或缺失
- 解决方案：BM25+ Embedding 混合检索，从而提高检索质量
- 效果：（进行中）

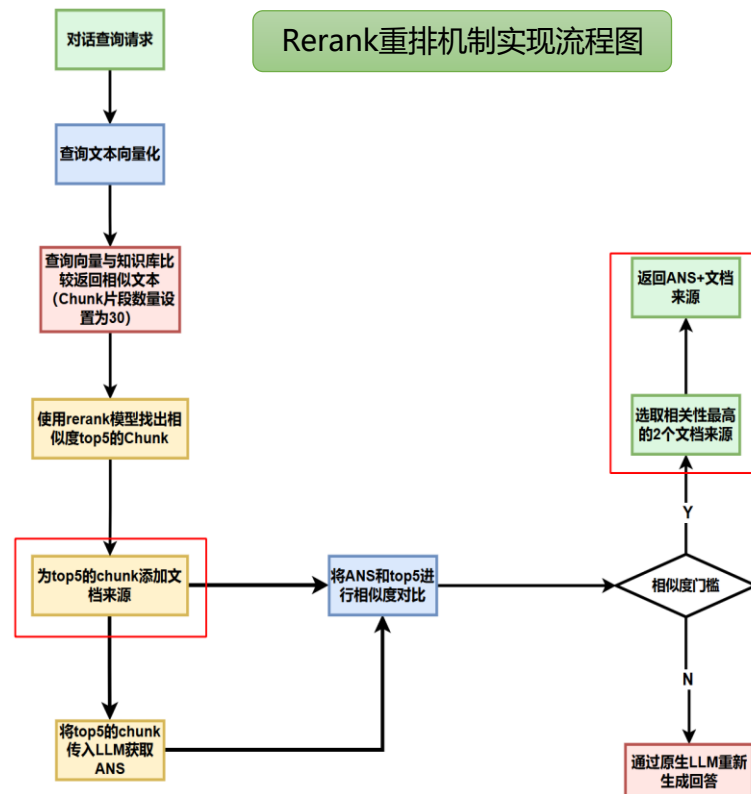
➤ 优化点三：采用Rerank对检索内容重排

- 原有问题：检索到无关内容太多，影响LLM输出
- 解决方案：采用Rerank模型重排，确保最相关文档在前面
- 效果：综合性能提升百分之10

混合检索实现流程图



Rerank重排机制实现流程图



提升检索检索能力和上下文质量



BES II智能数据平台硬件部署方案的研究

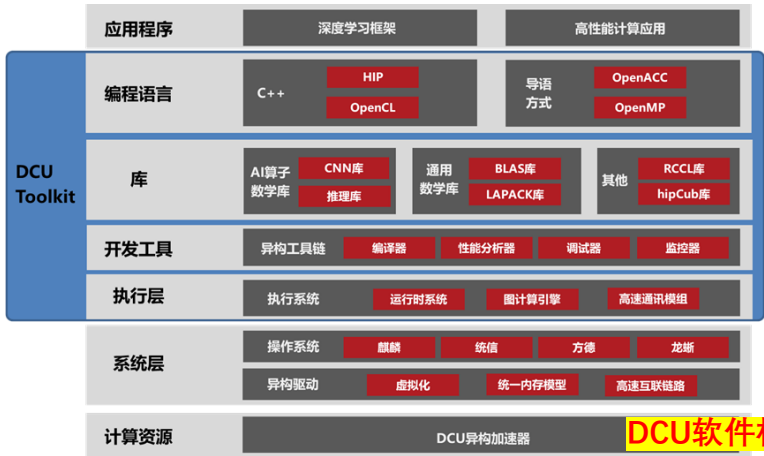
特点对比	优点	缺点
海光DCU	CUDA生态兼容，HPC支持强，工具链成熟	面向 AI/深度学习的优化较弱，HIP 虽然兼容 CUDA，但与原生CUDA 性能仍有差距
寒武纪GPU	AI推理和训练优化强，提供底层算子开发工具	CUDA兼容性差，HPC支持较弱
昇腾GPU	AI推理和训练优化强，硬件优化好，推理性能高	非AI场景支持弱，工具链学习成本高
摩尔线程GPU	CUDA兼容性高，高性能计算与图形渲染兼顾	AI优化能力不足，分布式训练和大规模数据集处理表现一般

DCU Z100与NVIDIA V100的主要性能参数

GPU		DCU	V100
核心频率		1443Mhz	1380Mhz
运算性能	FP64	5.9 TFLOPS	7.0 TFLOPS
	FP32	11.8 TFLOPS	14.0 TFLOPS
	FP16	23.6 TFLOPS	28.0 TFLOPS
显存	容量	16GB	32GB
	带宽	800GB/s	900GB/s
	位宽	4096bit	4096bit
	类型	HBM2	HBM2

基于曙光平台的海光DCU

- 移植CUDA代码，部署JUNO波形重建测试代码
- 成功运行，结果正确
- 初步尝试国产异构平台在DAQ在线处理中的应用



DCU软件栈

	Z100L (ms)	A30 (ms)	对比
数据加载	0.123	0.095	129.4%
计算执行	629.054	163.177	385.5%

NVIDIA A30与DCU Z100运行JUNO波形重建



其他工作

- 参与高能所机器学习会议并作报告——TAOCHAT智能助手
- GPU服务器维护
- 协助谱仪机房维护
- 协助江门现场工作

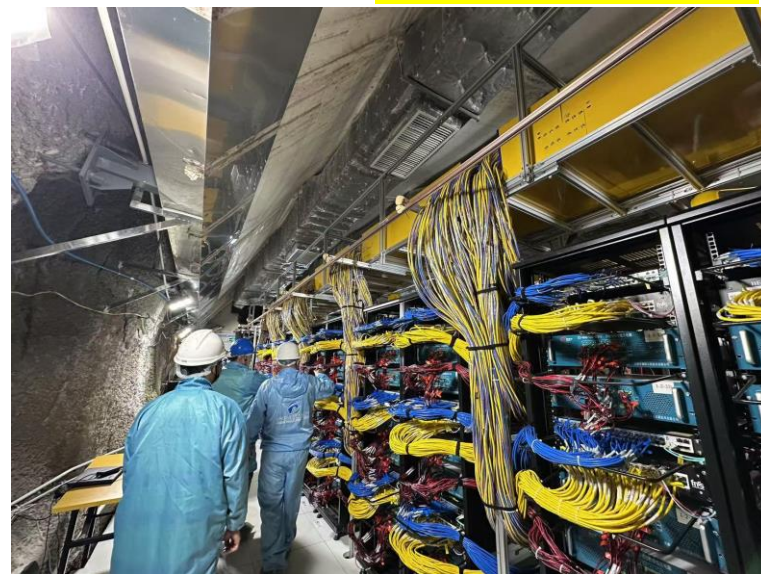
高能所机器学习会议



谱仪机房维护



协助江门现场工作





下一步计划

- 利用机器学习实现BESⅢ智能数据平台异常检测
- 利用RAG链路完成BESⅢ智能数据平台日志数据分析
- 完成其他工作

THANKS