OpenAIOps

AIOPS | 2024 CCF国际AIOps挑战赛
2024 CCF International AIOps Challenge

北京航空航天大學
BEIHANG UNIVERSITY

EasyRAG: 面向AIOps的高效RAG框架

主办单位：中国计算机学会（CCF）
承办单位：中国计算机学会互联网专委会、清华大学、中国科学院计算机网络信息中心
协办单位：OpenAIOps社区、中兴通讯、北京智谱华章科技有限公司、清华大学计算机科学与技术系、中南大学、北京必示科技有限公司
特别感谢：ModelScope社区、南开大学软件学院

队伍：搜的都队
主讲人：冯张驰
指导老师：张日崇教授

OpenAIOps AIOPS | 2024 CCF国际AIOps挑战赛
2024 CCF International AIOps Challenge

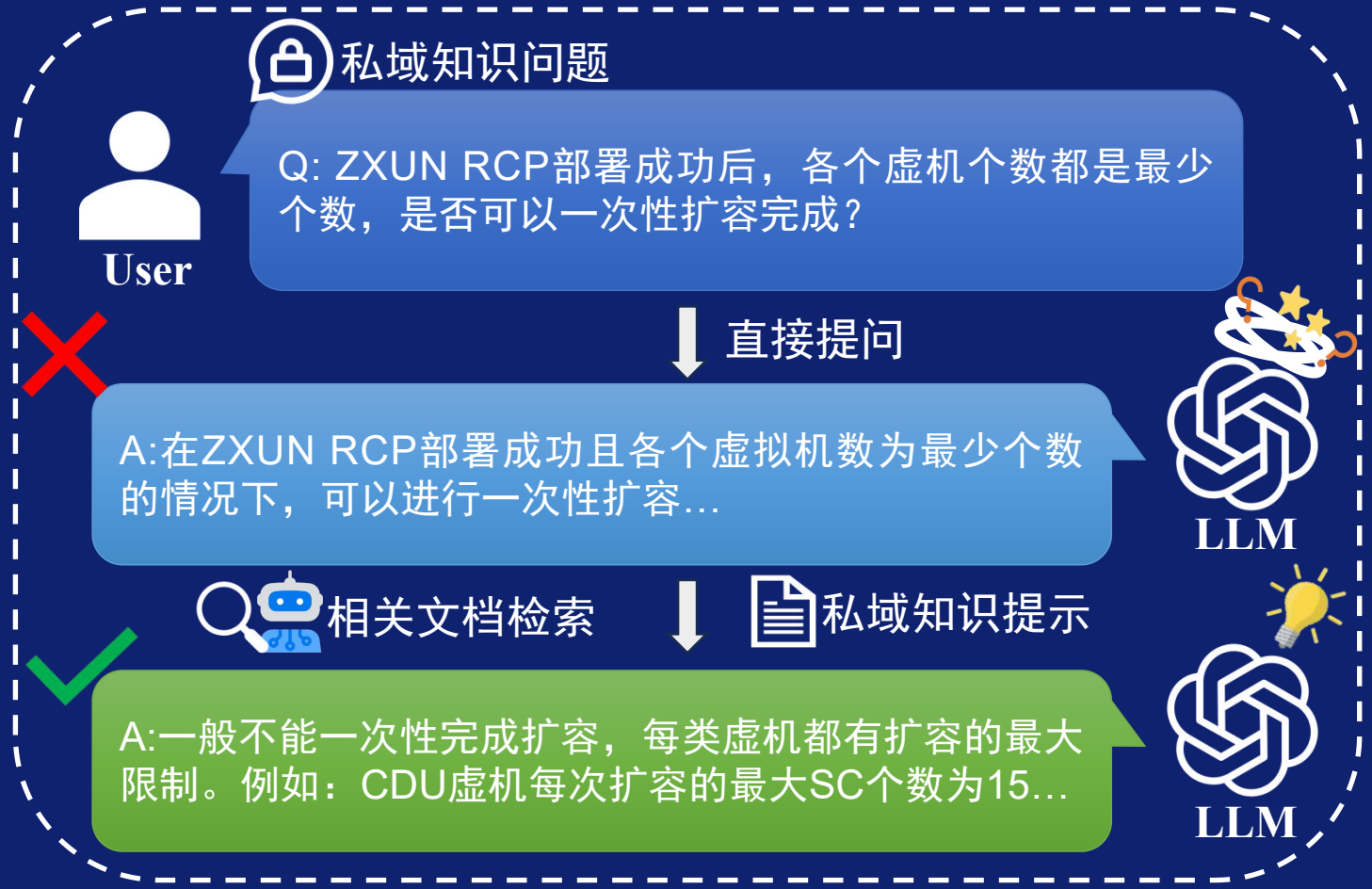
目录
CONTENTS

第一章节

概述

背景

- 通用大语言模型的设计基于广泛的公开数据，尽管其在常规任务中表现优异，但在垂直领域的直接应用中仍然面临**领域知识缺失**和**私域数据整合**的瓶颈。
- 在本次比赛中我们选择**赛道二**，模拟特定场景下没有自己微调模型能力的运维场景。



我们的思考：How to Make it Easy for RAG?

我们想要解决的问题：

1. 检索准确性：
   - 召回文档对问题回答的帮助程度。
2. 效率问题：
   - RAG流程中推理时延。
3. 实用性评估：
   - 框架的领域间迁移、部署难度。

我们的目标：

- 不微调任何模型的情况下，实现准确、高效、实用的RAG。

概述 What we do to Make it Easy:

为了更准确：
为了更高效：
为了更实用：

Chunking

- Multimodality Handling
- Chunking setting
  - Chunk size
  - Sliding windows
- Chunk Metadata
  - Knowledge Path
  - File Path
  - Image Source

Embedding

- Embedding model
  - gte-Qwen
  - bge-small-zh
  - bce-base

Retrieval

- Original Query
  - BM25s
  - Dense-Embedder
- Query Rewriting
  - Query Expansion
  - HyDE
- Rank Fusion
  - Reciprocal Rank Fusion
  - Naive Merging

Reranking

- Reranking Model
- Re-ranker Early Exit

Prompting

- Normal QA Prompt
- Cot Prompt
- Struct Prompt
- CO-STAR Prompt

Summary

- Context Compression
  - BM25-based
  - LLM-based

LLM

Deployment

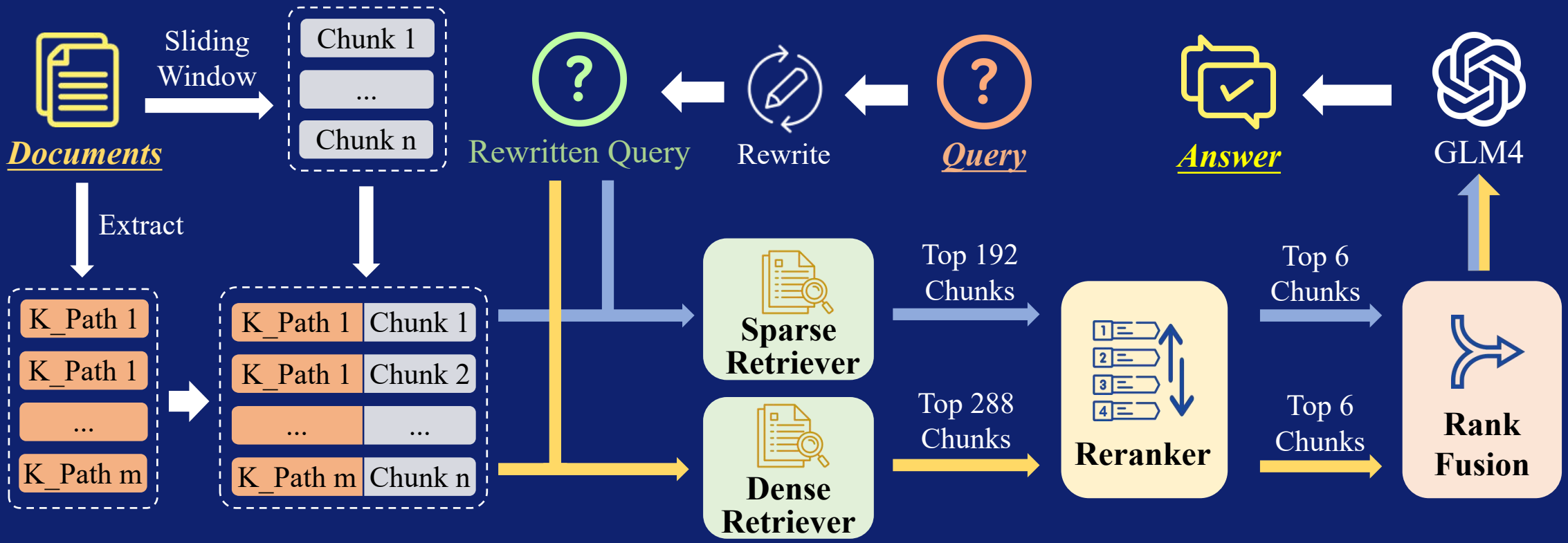
- Multiple-thread Processing
- Web-UI
- Batch Processing

第二章节

准确性

初赛Pipeline

初赛算法精度：84.38（初赛第一名）



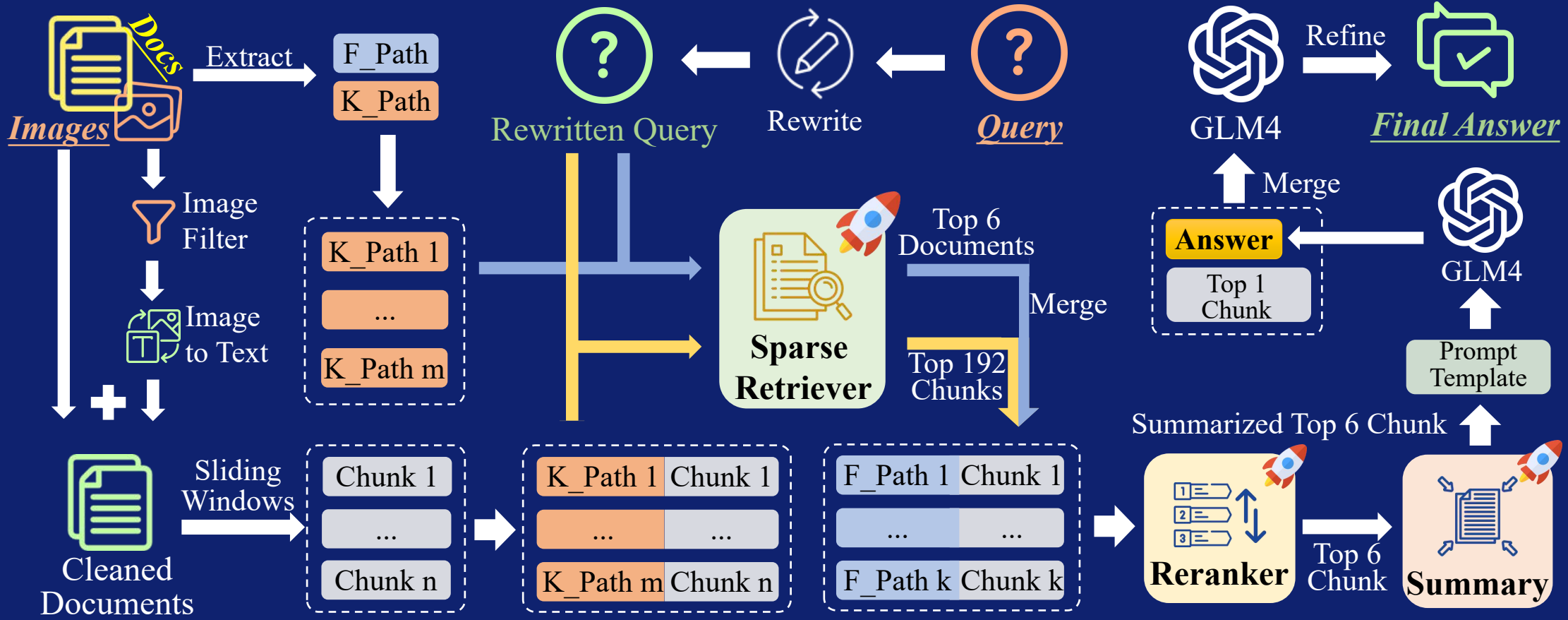
F_Path (File Path)：./emsplus/documents/软件安装/topics/版本文件准备.html

K_Path (Knowledge Path)：-emsplus-安装与调测-软件安装-安装准备-版本文件准备

复赛Pipeline

复赛算法精度：96.65（赛道二第二名）、82.92（B榜第五名）

复赛算法开销：单进程每条Query 原始16.0s 加速后12.5s （加速22%）



F_Path (File Path)：./emsplus/documents/软件安装/topics/版本文件准备.html

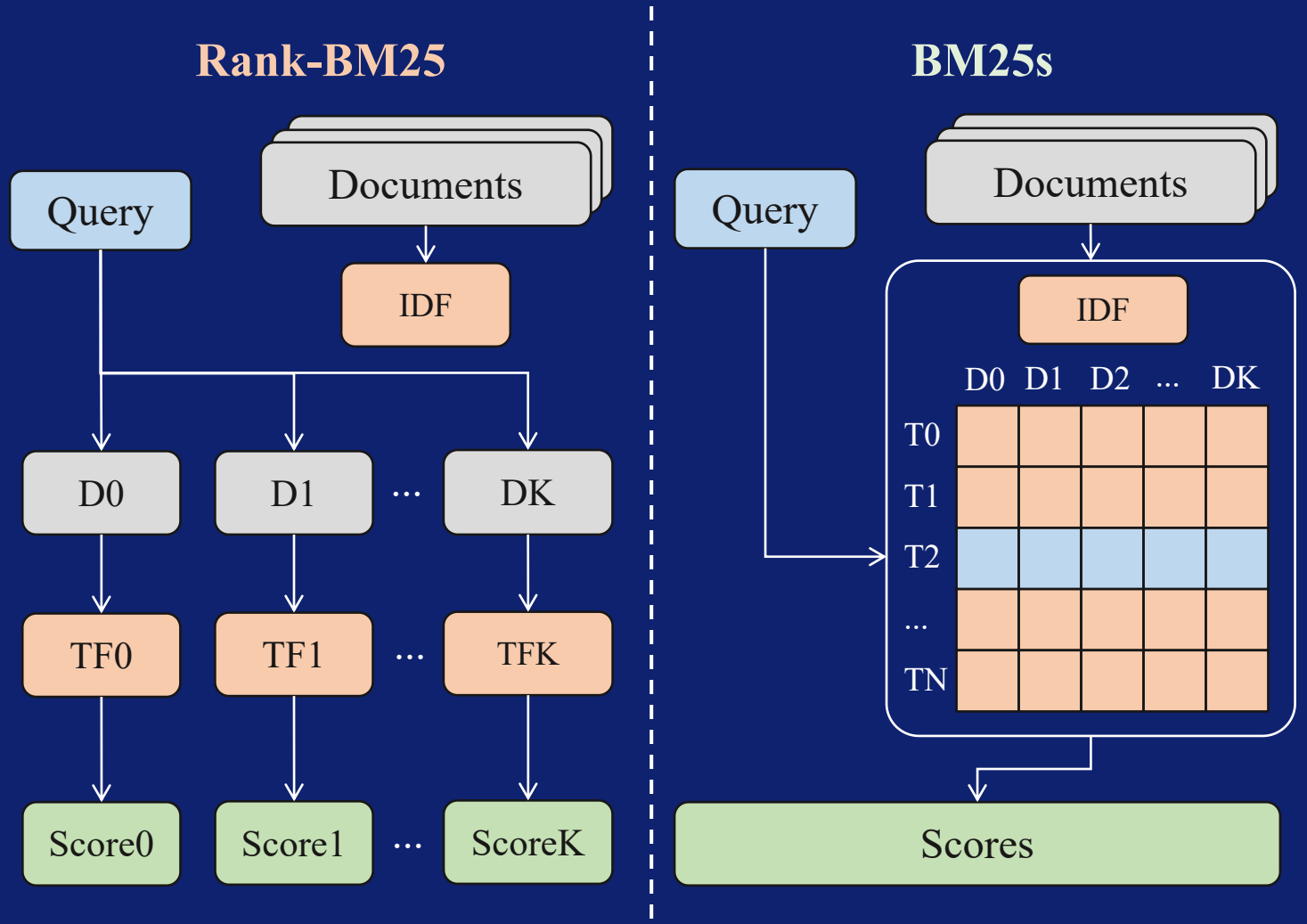
K_Path (Knowledge Path)：-emsplus-安装与调测-软件安装-安装准备-版本文件准备

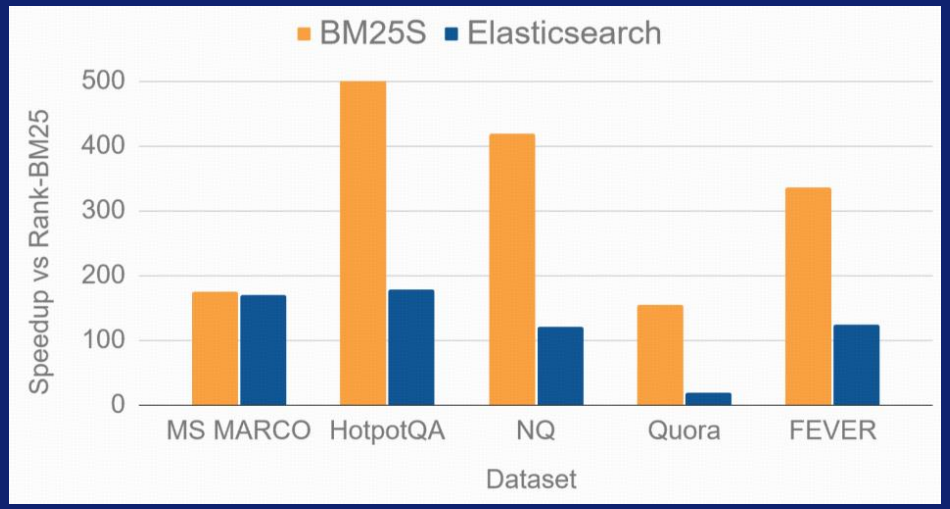
第三章节

高效性

速度优化1-高效稀疏检索

- 索引TF分数+Scipy高效稀疏矩阵运算
- 速度提升300倍，100次检索仅耗时50ms

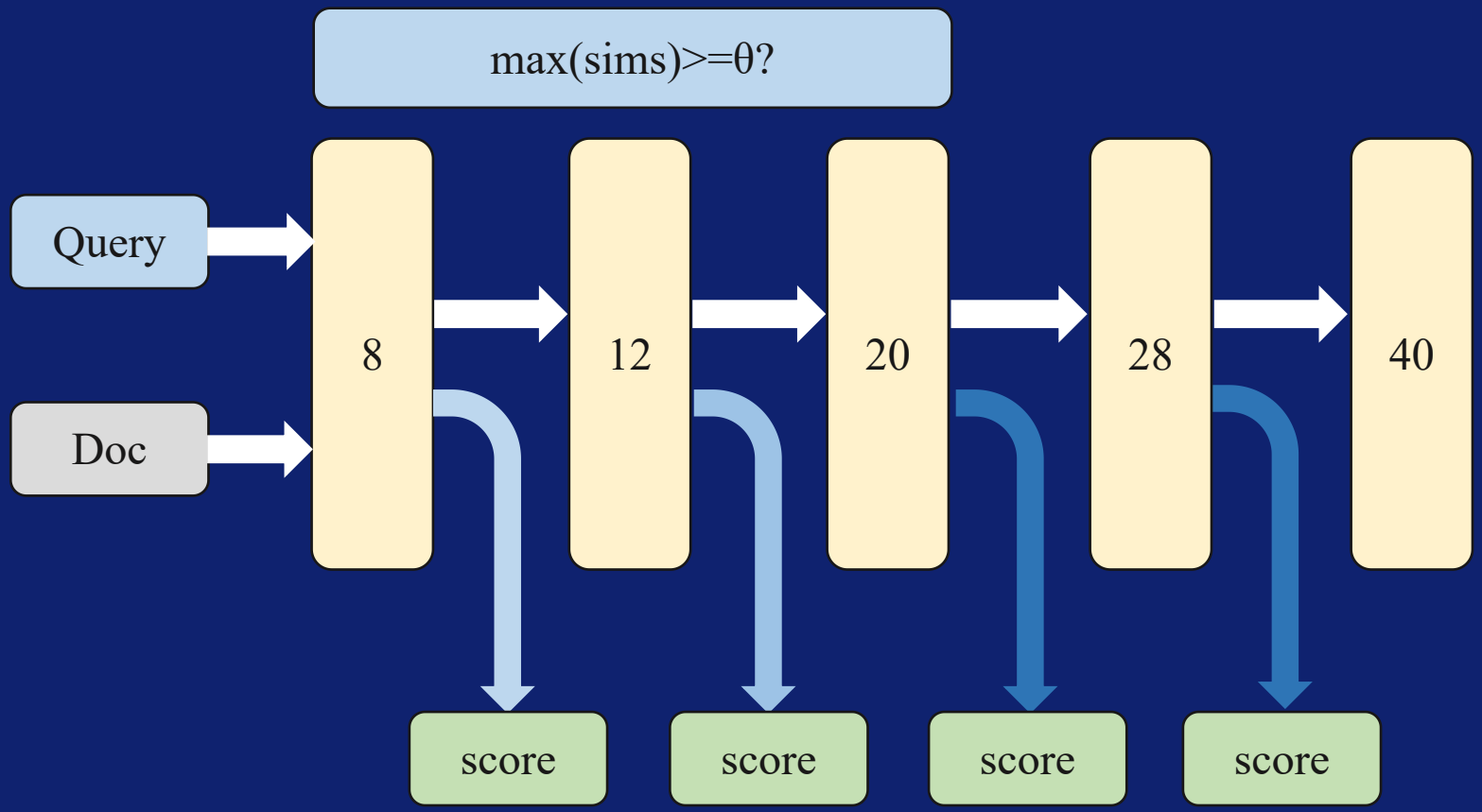




方法	103条查询时间(s)	准确度
Rank-BM25	17	94.49
BM25s	**0.05**	94.24

速度优化2-高效重排

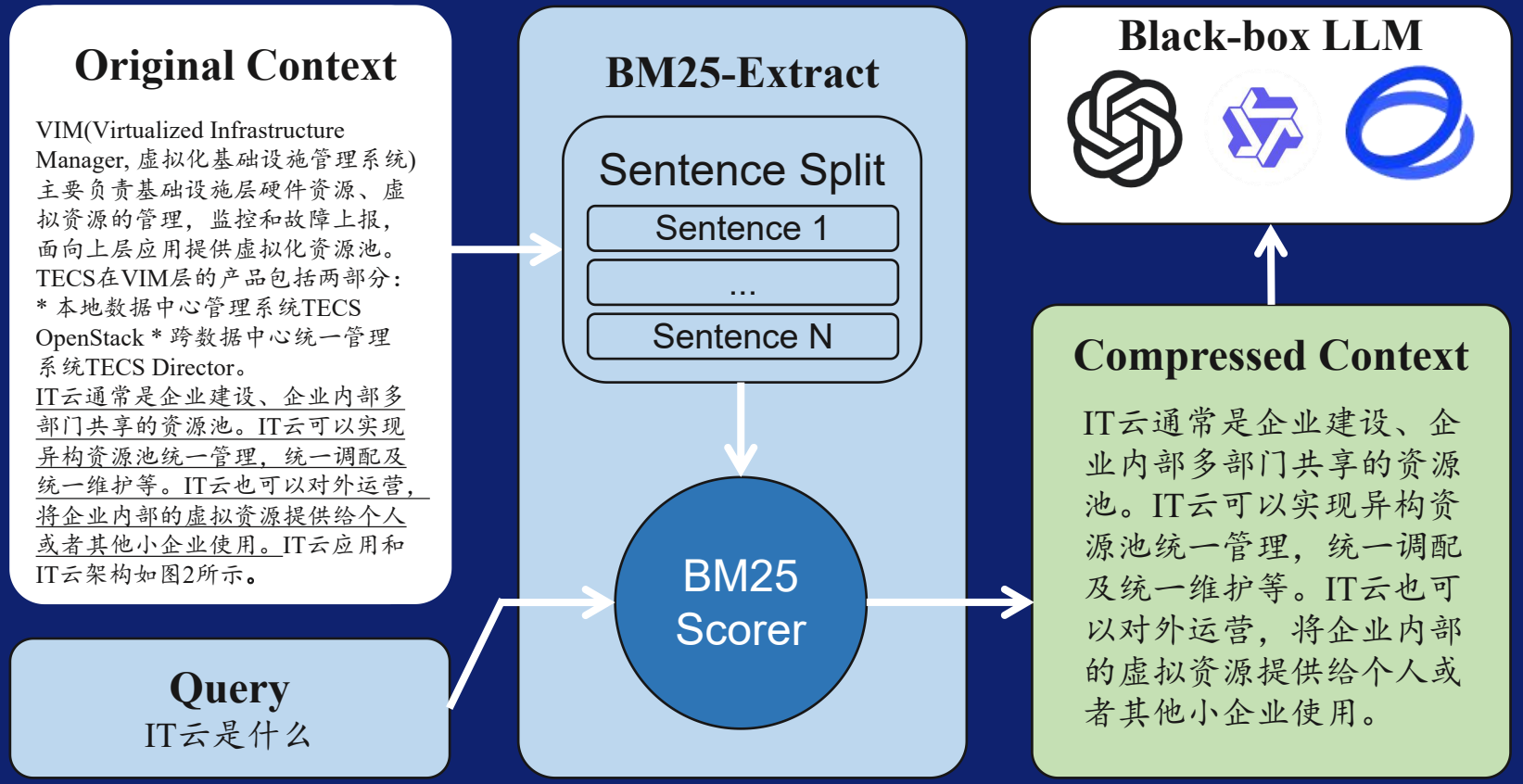
- 层最大相似度选择算法
- 速度提升**2倍**，平均每次查询**降低2s**



方法	查询时间(s)	精度 (%)
8层	1.67	73
12层	2.20	88
20层	3.58	86
28层	5.25	100
40层	7.71	100
最大值选择 (0.1)	**2.59**	90
最大值选择 (0.2)	3.55	96
最大值选择 (0.4)	4.57	**97**
熵选择 (0.2)	2.74	89
熵选择 (0.4)	3.37	91
熵选择 (0.6)	4.01	91

速度优化3-高效LLM推理

- BM25计算查询-句子相似度
- 基于抽取式摘要的上下文压缩
- 平均每个查询节省 1k token使用，降低 1.5s 推理时延



方法	压缩率 (%)	节省 token	准确度 (%)	时间 (s)
原始上下文	100	0	94.49	9.30
LLMLingua	62.80	143k	83.44	10.47
Long LLMLingua	62.80	143k	80.86	10.52
BM25-Extract (0.5)	55.92	160k	86.48	7.70
BM25-Extract (0.8)	83.84	59k	89.00	8.12

第四章节

实用性

实用性

简单部署

- 我们的框架支持Docker 或cli**命令**部署，使得构建部署过程更加简便
  - ◆ 用户可以通过 Docker 镜像实现快速**批量化**的问题回答，无需处理复杂的依赖配置；
  - ◆ 同时，也可以通过命令行灵活地定制化设置响应线程和快速启动以适应多种应用场景；

目前我们的代码已经完全开源，只需要简单的几行代码就可以**quick start**！

★ 批量处理

★ API接口部署

★ 用户友好交互启动

```
git clone https://github.com/BUAADreamer/EasyRAG.git && cd EasyRAG

# docker run
chmod +x ./run.sh && bash ./run.sh

# API deploy
cd src && uvicorn api:app --host 0.0.0.0 --port 8000 --workers 4

# Web-UI start
cd src && streamlit run webui.py
```

学术论文

EasyRAG: Efficient Retrieval-Augmented Generation Framework for Automated Network Operations

Zhangchi Feng[1], Dongdong Kuang[1], Zhongyuan Wang[1], Zhijie Nie[1], Yaowei Zheng[1], Richong Zhang[1]*
[1]CCSE, School of Computer Science and Engineering, Beihang University, Beijing, China
{zcmuller,kuangdd,wangzy23,hiyouga}@buaa.edu.cn, {niezj,zhangrc}@act.buaa.edu.cn

社区博客

AIOps RAG竞赛优秀方案EasyRAG解读：兼看SimRAG:自适应检索增强微调思路

所以我们来看看具体工作，写成了论文，《easyrag：efficient retrieval-augmented generation framework for automated network operations》，https://arxiv.org/abs/2410.103...

老刘说NLP 6天前 最近读过

【RAG】aiops第一名方案-EasyRAG：自动网络运营的高效检索增强生成框架

因此，easyrag框架探索了使用不同层数的模型来平衡推理时间和排名准确性。例如，可以选用28层或40层的模型，根据资源限制和性能需求进行选择。）

大模型自然语言... 15天前 最近读过

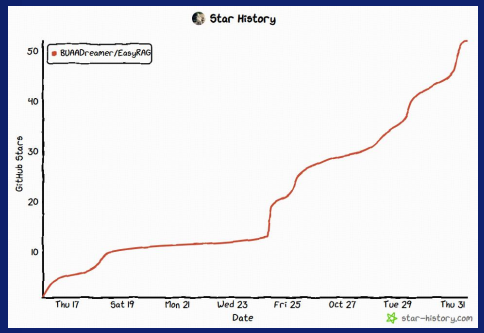
AIOps RAG 比赛获奖项目 EasyRAG 深度解读

易迟 分享有用有趣的内容 已关注

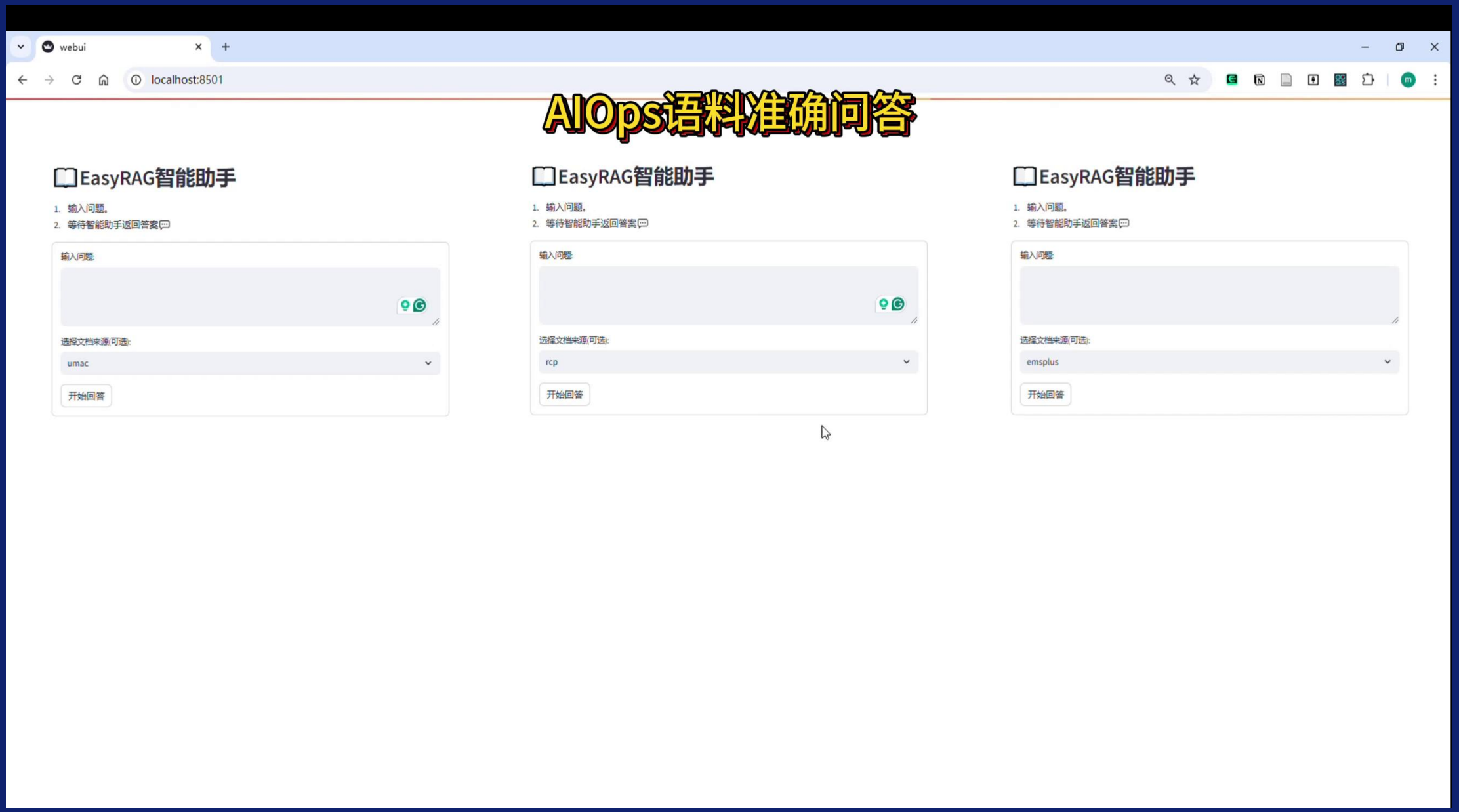
28 人赞同了该文章

Github仓库

Fork 7 Star 60



实用性


webui
localhost:8501
AIOps语料准确问答
EasyRAG智能助手
1. 输入问题。
2. 等待智能助手返回答案
输入问题:
选择文档来源(可选):
umac
开始回答
EasyRAG智能助手
1. 输入问题。
2. 等待智能助手返回答案
输入问题:
选择文档来源(可选):
rcp
开始回答
EasyRAG智能助手
1. 输入问题。
2. 等待智能助手返回答案
输入问题:
选择文档来源(可选):
emsplus
开始回答

OpenAIOps AIOPS | 2024 CCF国际AIOps挑战赛
2024 CCF International AIOps Challenge

THANKS

主办单位：中国计算机学会（CCF）

承办单位：中国计算机学会互联网专委会、清华大学、中国科学院计算机网络信息中心

协办单位：OpenAIOps社区、中兴通讯、北京智谱华章科技有限公司、清华大学计算机科学与技术系、中南大学、北京必示科技有限公司

特别感谢：ModelScope社区、南开大学软件学院