中国计算机学会 CHINA COMPUTER FEDERATION
CCF计算经济学专业组
北京大学前沿计算研究中心 Center on Frontiers of Computing Studies, Peking University
北大创新评论 PKU Innovation Review
博时基金 BOSERA FUNDS
阿里云
金证股份
未名数创 Undiscovered Innovation

2023 CCF 计算经济学比赛
博金挑战赛

果树技术灌溉组

团队介绍

北京快确信息科技有限公司

业内领先的金融科技与监管科技公司；
公司2018年推出的产品QTrade，已形成中国最活跃固收市场交易网络；
致力推进固收市场科技的发展和生态建设，助力固收市场迈向智能交易时代。

目录

- 赛制介绍
- 赛题分析
- 方案架构
  - 数据库查询-NL2SQL
  - 文本理解-RAG+LLM
- 项目总结

赛制介绍

赛题内容

本次比赛要求选手基于通义千问金融大模型或通义千问7B模型(不限制pretrain和chat)构建一个问答系统，问答内容涉及基金/股票/债券/招股书等不同数据来源

评估标准

$$score = 0.6 * score_{result} + 0.4 * score_{semantic}$$

$score_{result}$ （结果得分）：占比60%

$score_{semantic}$ （语义得分）：占比40%

赛题分析

数据库查询题

- 请问2019年三季度有多少家基金是净申购?它们的净申购份额加起来是多少?请四舍五入保留小数点两位。
- 2019年中期报告里，华夏基金管理有限公司管理的基金中，机构投资者持有份额比个人投资者多的基金有多少只?

文本理解题

- 2014年大连派思燃气系统股份有限公司经营业绩下降的主要原因是什么？
- 中国铁路通信信号股份有限公司的主要经营模式是怎样的？

赛题分析

赛题挑战

数据库查询

- SQL查询语句书写的准确性约束
- 多表联合查询列名之间复杂关系

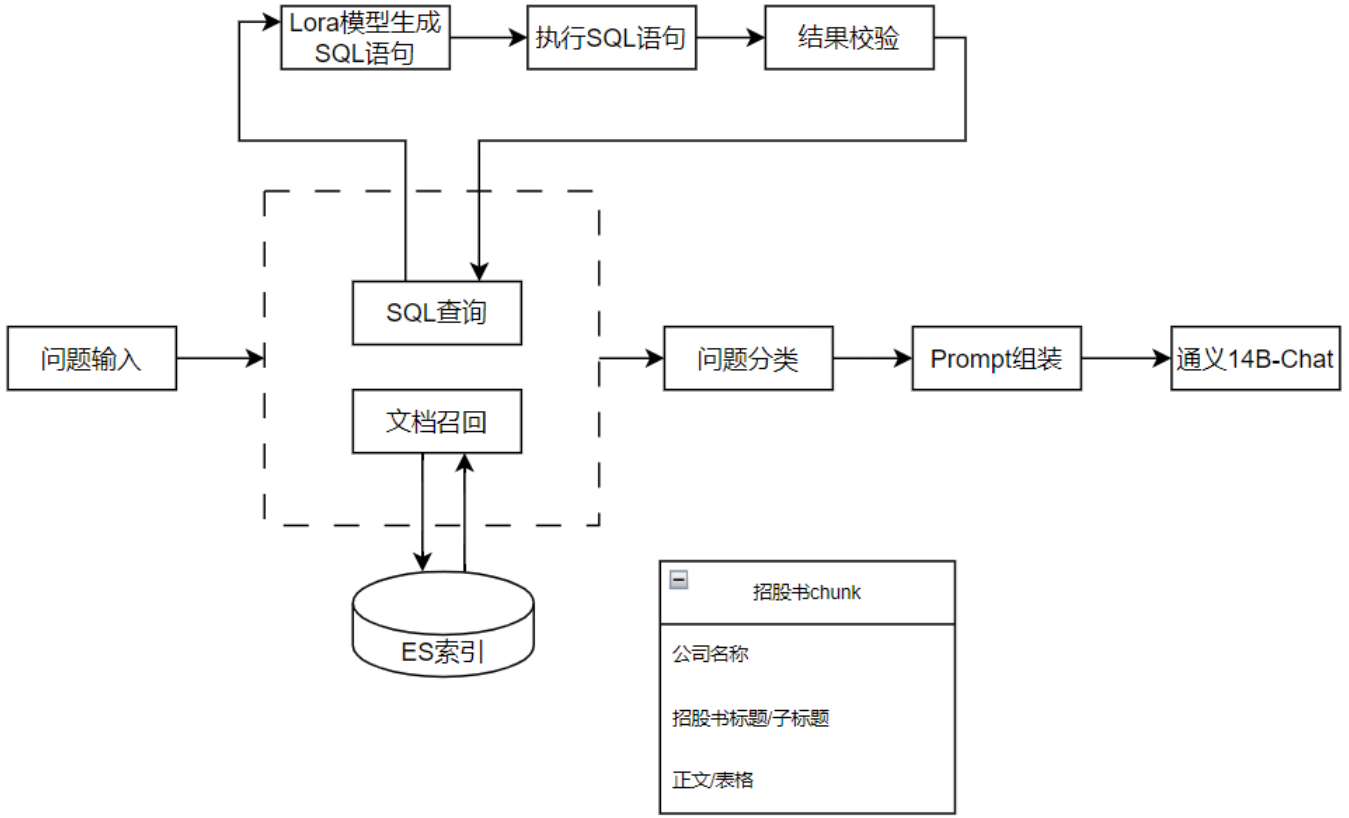
文本理解

- 招股书文本过长，大模型输入有限
- 文本噪音干扰，事实概念冲突

方案架构

整体方案

- 数据查询-NL2SQL
  - Lora模型生成SQL
  - 数据库Agent
- 文本理解-RAG+LLM
  - 结构化片段召回
  - Prompt工程优化



数据库查询

问题拆解

数据库查询，主要分为以下集中类型：

- 单表的选择查询
- 单表的复杂查询
- 。。。
- 多表联合的选择查询
- 多表联合的聚合排序查询
- 。。。

id	quesiton	问题分类	细分类别
0	请帮我计算，在20210105，中信行业分类划分的一级行业为综合金融行业中，涨跌幅最大股票的股票代码是？涨跌幅是多少？百分数保留两位小数。股票涨跌幅定义为：（收盘价 - 前一日收盘价 / 前一日收盘价）* 100%。	表格	多表联合的计算排序查询
2	请帮我查询出20210415日，建筑材料一级行业涨幅超过5%（不包含）的股票数量。	表格	多表联合的聚合查询
3	请查询在2021年度，688338股票涨停天数？ 解释：（收盘价/昨日收盘价-1）》=9.8% 视作涨停	表格	单表的聚合查询
4	20210304日，一级行业为非银金融的股票的成交量合计是多少？取整。	表格	多表联合的聚合查询
5	嘉实基金管理有限公司2019年成立了多少基金?	表格	单表的聚合查询
7	2021年三季度，有多少家基金发生了净赎回?总共赎回了多少份额?记得给我四舍五入到小数点后两位哦。	表格	多表联合的复杂查询
8	请帮我查询下，在20190605，中信行业分类里一级行业为石油石化行业的所有股票里, 成交金额(元)最多的股票的代码是什么？成交金额是多少?	表格	多表联合的计算排序查询
9	股票002244在20191220日期中的收盘价是多少？(小数点保留3位)	表格	单表的选择查询
11	请问2019年三季度有多少家基金是净申购?它们的净申购份额加起来是多少?请四舍五入保留小数点两位。	表格	单表的聚合查询
12	帮我查一下湘财长弘灵活配置混合A基金在20210419的资产净值和单位净值是多少?	表格	多表联合的选择查询

数据库查询

NL2SQL

尝试一：
使用模板填充方法，使用通义千问14B-Chat模型抽取问题中涉及的关键信息及对应的列名称，如债券名、基金简称、股票代码等，将这些信息填入给定模板。效果并不理想，会出现识别不全和列名称错误问题。

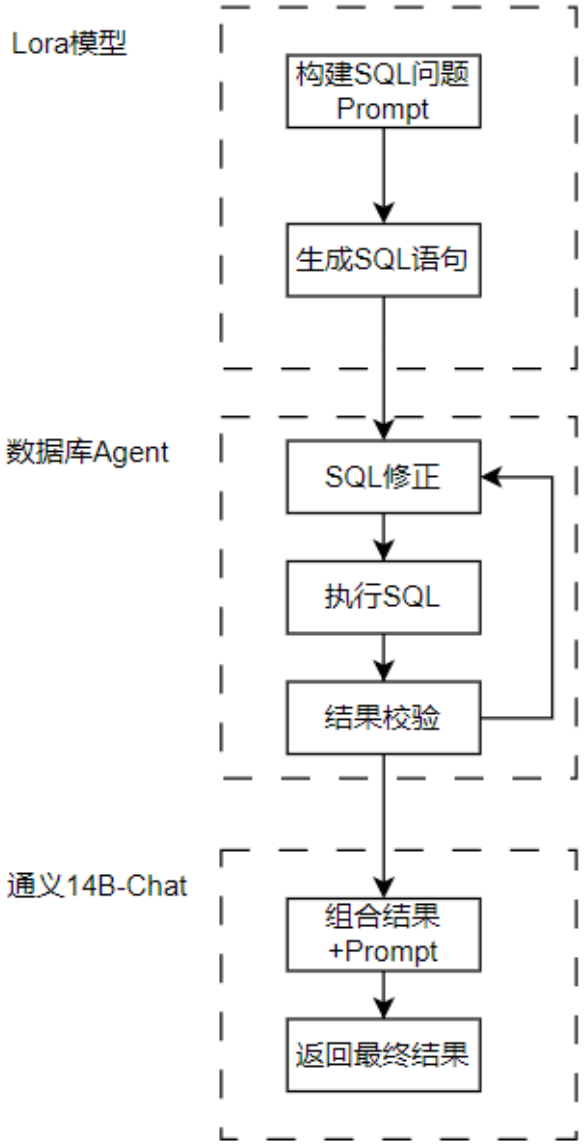
尝试二：
使用通义千问14B-Chat模型，设计多个prompt模板，直接输入问题，在多表及公式定义问题生成的sql存在很多问题，如表名、列名不匹配，sql语法错误。即使用few shot learning方法，在Prompt中增加样例，仍然存在表名列名不匹配问题，少量sql语法问题。

尝试三：
经过前面几个步骤的尝试，依然无法达到期待的效果，此时采用人工标注少量30-60个样本，微调通义千问14B-Chat模型，基本解决了表名列名不匹配问题，能较好理解单表多表问题

NL2SQL

核心思想

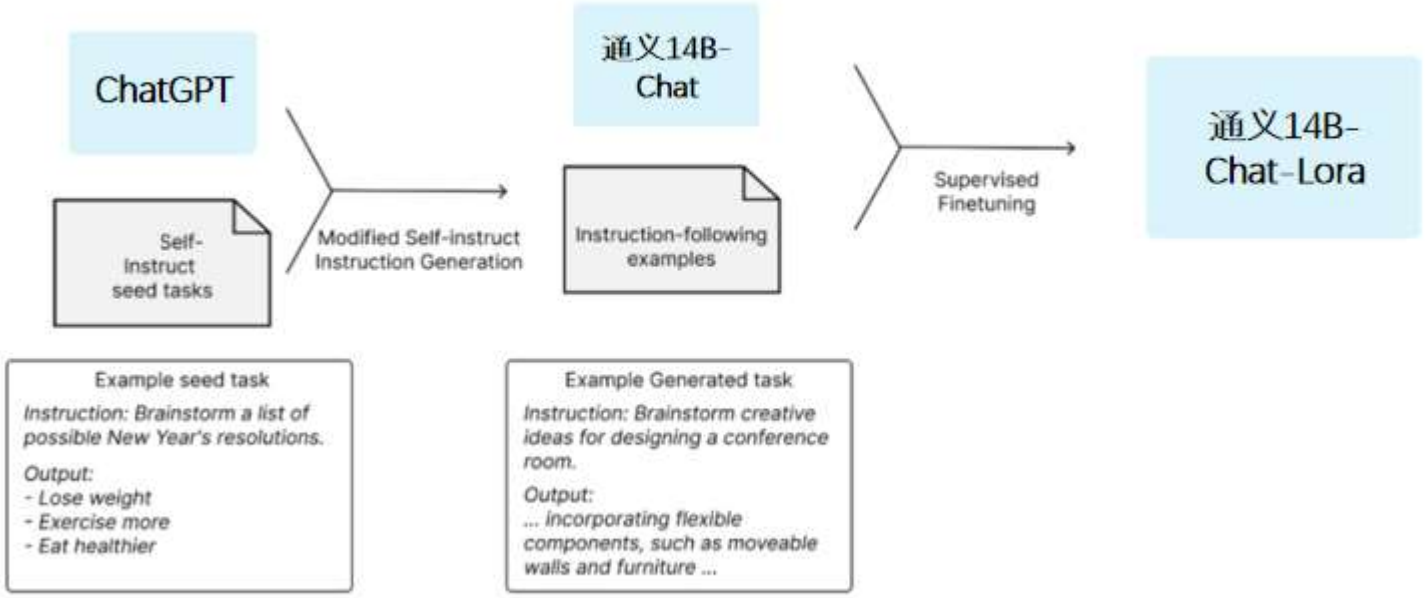
- Lora模型生成SQL
- 数据库Agent
- Chat模型生成答案



NL2SQL

训练数据构建

1. 问题抽象
   - 简单直接查询
   - 简单多表联合查询
   - 聚合函数
2. 人工编写种子
   - 拓展问题类型
   - 自定义公式计算
3. ChatGPT数据增强



Lora模型训练迭代过程

NL2SQL

Prompt模板

假设你是SQL专家，结合下面的问题描述、数据库表名、列名类型信息等，生成可以执行的SQL代码

数据库描述:{数据库基本信息}

问题描述:{Question}

数据库基本信息

数据库表名+列名

- 基金基本信息：包括基金代码、基金全称、基金简称、管理人...

特殊字段类型

- 枚举型：定期报告所属年度(2019、2020、2021)...
- 数值类型：数量、管理费率...
- 字符类型：基金代码、债券...

问题描述

请帮我查询出20210415日，建筑材料一级行业涨幅超过5%（不包含）的股票数量。

NL2SQL

数据构建-自定公式

问题-SQL对

请帮我计算在2019年，A股回撤最大的股票代码和回撤比例分别是多少，百分比保留2位小数。股票报告期内**最大回撤：（最高价(元)- 最低价(元)）/ 最高价(元) * 100%**

```
SELECT "股票代码",
       ROUND((MAX("最高价(元)") - MIN("最低价(元)")) / MAX("最高价(元)") * 100, 2) AS "最大回撤"
FROM "A股票日行情表"
WHERE "交易日" like '2019%'
GROUP BY "股票代码"
ORDER BY "最大回撤" DESC
LIMIT 1
```

Self-instrcut 增强

- 2019年，A股中哪只股票经历了最大的回撤？请提供该股票的代码和回撤比例，百分比保留两位小数。
- 在2019年，A股市场中，哪支股票的回撤幅度最大？请给出该股票的代码以及相应的回撤比例，百分比保留两位小数。
- 请计算2019年A股市场中回撤最大的股票的代码和回撤比例是多少，百分比精确到小数点后两位。
- 在2019年，A股市场上，哪只股票经受了最大的回撤？请提供该股票的代码和回撤比例，百分比保留两位小数。

NL2SQL

Lora模型效果对比

问题
2021年年度报告里，凯石基金管理有限公司管理的基金中，机构投资者持有份额比个人投资者少的基金有多少只?

原版结果

```
SELECT COUNT(DISTINCT 基金代码) AS 基金数量
FROM 基金份额持有人结构
WHERE 定期报告所属年度 = 2021
AND 管理人 = '凯石基金管理有限公司'
AND 机构投资者持有的基金份额占总份额比例 > 个人投资者持有的基金份额占总份额比例;
```

Lora结果

```
SELECT COUNT(DISTINCT 基金代码) AS 基金数量
FROM 基金份额持有人结构
WHERE 基金简称 IN (
SELECT 基金简称
FROM 基金基本信息
WHERE 管理人 = "凯石基金管理有限公司"
)
AND 定期报告所属年度 = "2021"
AND 报告类型 = "年度报告"
AND 机构投资者持有的基金份额 < 个人投资者持有的基金份额
```

NL2SQL

Lora数据规模以及效果

- 种子问题对，50+
- 自动化扩展效果+5%

Lora数据规模	模型效果
种子数据x10	0.85
种子数据x20	0.88
种子数据x30	0.89

数据库Agent

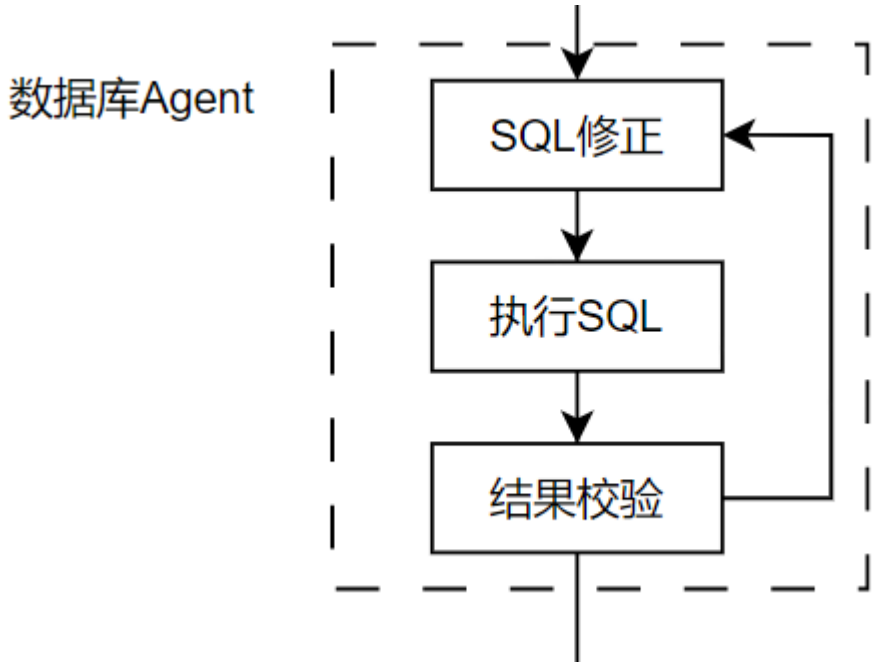
SQL语句校验

条件值更正

- 判断条件值是否在文本内，基金简称 = '海富通安益对冲策略灵活配置混合C'

数表名及列名对齐

- 基金股票持仓明细，如债券名称 = "20交投Y4"
- 基金债券持仓明细，如债券名称 = '20交投Y4'



答案生成-通义14B-Chat

Prompt模板

角色与目标

你是一个会组织语言的专家。下面将提供一些问题、sql语句及sql执行返回答案的样例，仿照样例输出答案

参考实例

样例1问题：我想知道华夏创业板动量成长ETF基金，在2021年年度报告中，前10大重仓股中，有多少只股票在报告期内取得正收益。
样例1的sql语句：SELECT COUNT(*) AS 股票数量
样例1的sql执行结果：2
样例1输出结果：华夏创业板动量成长ETF基金，在2021年年度报告中，前10大重仓股中，有2只股票在报告期内取得正收益。

查询结果

问题：{question}
sql语句：{title_desc}
sql执行结果：{sql_rst}
输出结果：

文本理解

问题拆解

问题一：
报告期内山东海看网络科技有限公司因进行宣传和活动策划活动共产生的业务宣传费为多少？
难点：招股书中只明确列举了宣传经费和活动策略经费，需要计算生成最终结果。

问题二：
烟台杰瑞石油服务集团股份有限公司的上下游行业是什么？。
难点：招股书中单独说明了上游行，以及下游行业，需要同时召回并总结。

问题三：
兰州海默科技股份有限公司的主要产品有哪些？
难点：招股书针对每个产品单独列举说明，需要上下文信息总结。

1、公司主要产品及用途

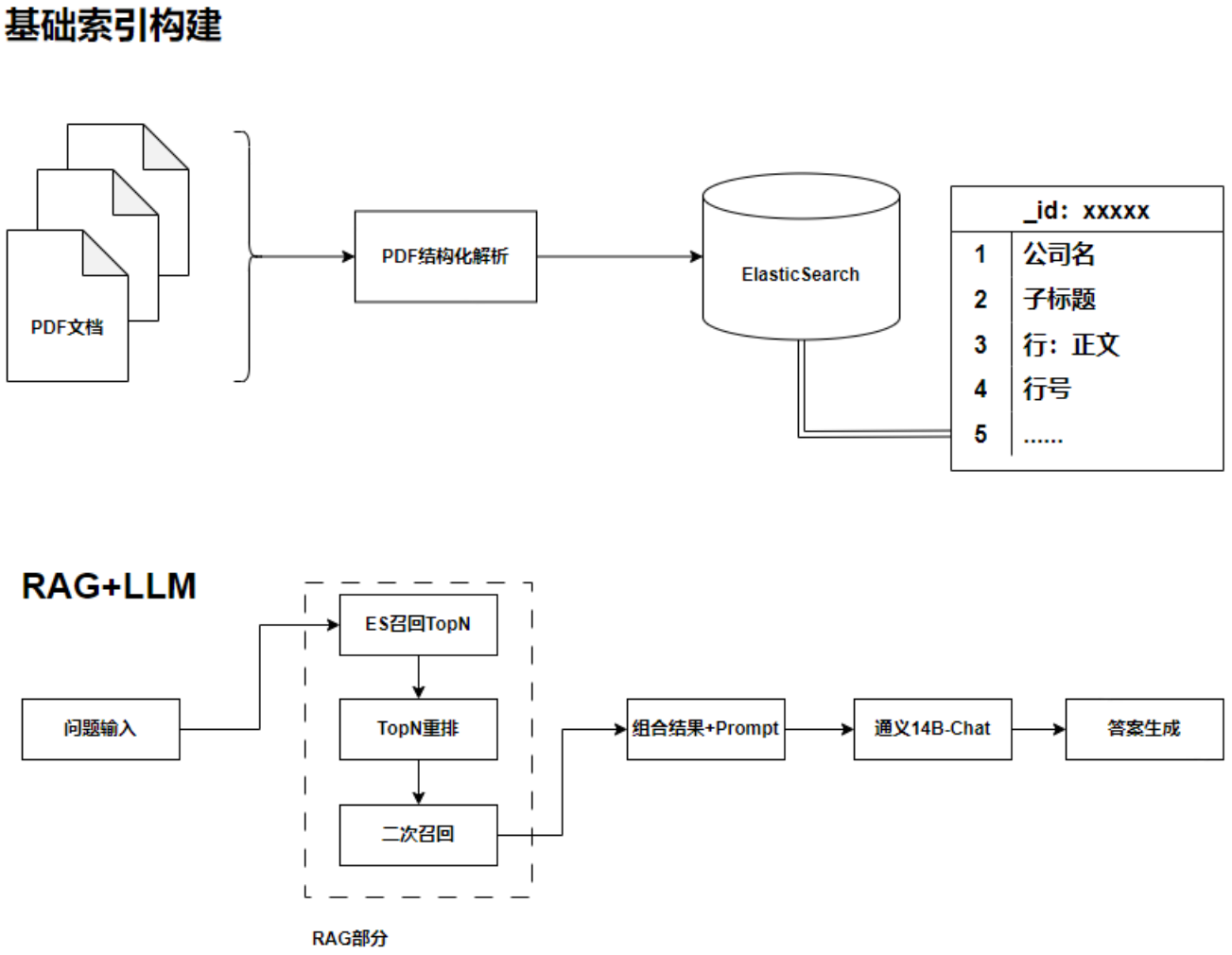
公司的主要产品为多相流量计系列产品，具体包括：

（1）油井多相流量计

RAG+LLM

核心思想

- ES粗召，LLM模型“精排
- 分块策略，以小召大
- Prompt工程优化



RAG增强

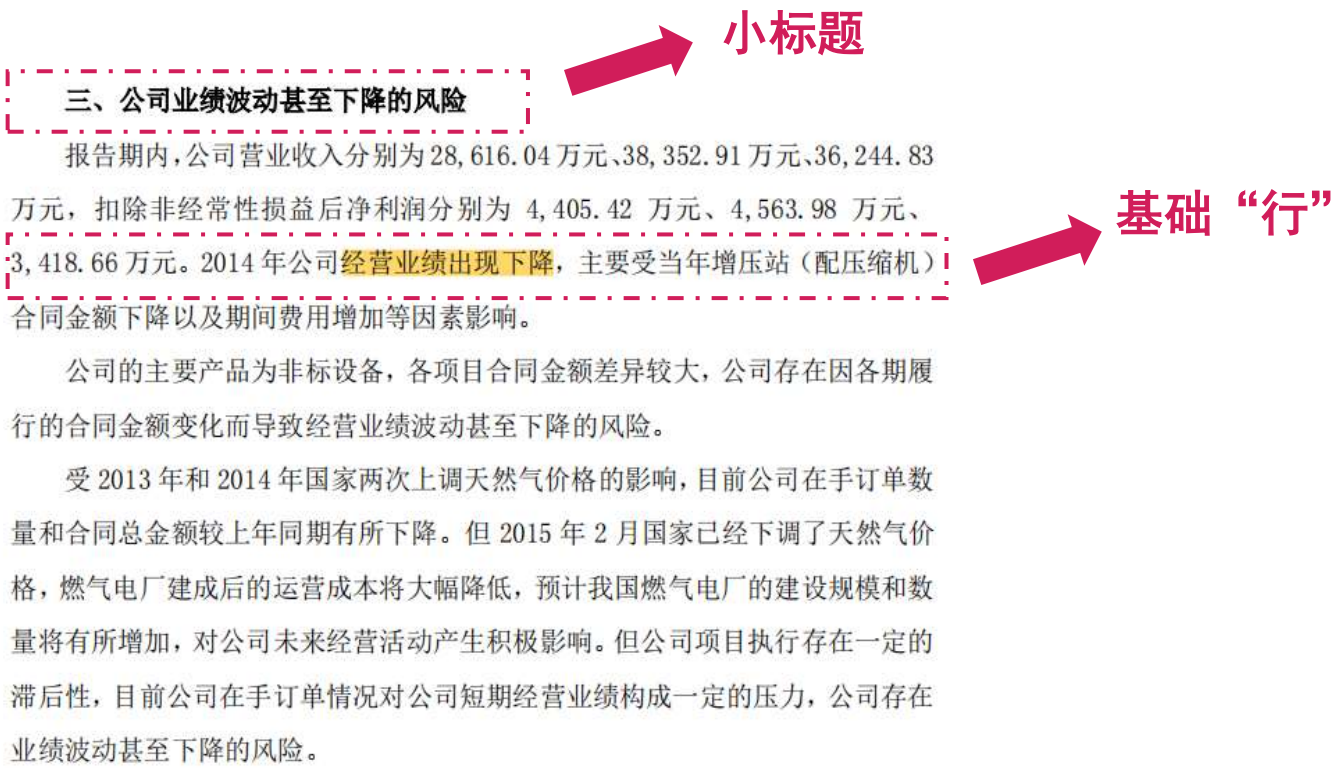
文本解析

文档分块

- 基于PyMuPDF解析PDF
- “行”作为最小的召回单元
- 招股书标题结构化存储

文档索引

- 基于Jieba分词搭建最小索引
- ElasticSearch搭建关键词召回



RAG增强

ES召回

分级索引

1. 匹配公司招股书
2. 检索TopN相似

TopN重排

- 多源相关性计算
- 最小分块融合

以小召大

问题：
2014年大连派思燃气系统股份有限公司经营业绩下降的主要原因是什么？

三、公司业绩波动甚至下降的风险

报告期内，公司营业收入分别为28,616.04万元、38,352.91万元、36,244.83万元，扣除非经常性损益后净利润分别为 4,405.42 万元、4,563.98 万元、3,418.66万元。2014年公司经营业绩出现下降，主要受当年增压站（配压缩机）合同金额下降以及期间费用增加等因素影响。

公司的主要产品为非标设备，各项目合同金额差异较大，公司存在因各期履行的合同金额变化而导致经营业绩波动甚至下降的风险。

受2013年和2014年国家两次上调天然气价格的影响，目前公司在手订单数量和合同总金额较上年同期有所下降。但2015年2月国家已经下调了天然气价格，燃气电厂建成后的运营成本将大幅降低，预计我国燃气电厂的建设规模和数量将有所增加，对公司未来经营活动产生积极影响。但公司项目执行存在一定的滞后性，目前公司在手订单情况对公司短期经营业绩构成一定的压力，公司存在业绩波动甚至下降的风险。

重排TopN

召回部分

答案生成-通义14B-Chat

Prompt模板

Prompt1

您是文档问答系统高级专家。
您可以通过在文档中查找相关内容并根据该内容来回答问题
文档内容：{content}
问题：{question}

Prompt2

您是文档问答系统高级专家。您可以通过在文档中查找相关内容并根据该内容来回答问题　　原则是
1、判断下面问题中是否有多个问题，若存在多个问题，将问题拆分出多个问题；
2、根据拆分出来的问题，精确定位文档中的内容，可能存在一个或多个位置，存在多个时需要都找出来
3、根据定位出来的位置前后内容，回答每个问题；
4、问题答案在文档中是分点/条、系列措施时列举时，需要全部列出不遗漏；
5、只需要给出问题的答案，不需要列出中间的思考过程，不需要开头部分的问候语，不需要拓展
6、只能根据提供的文档来回答每个问题

内容：{content}
问题：{question}

缺少示例，执行步骤与要求混合，指令跟踪不好

答案生成-通义14B-Chat

Prompt模板

角色与目标

任务要求

任务拆解

参考示例

```
    您是文档问答系统高级专家。根据以下原则、步骤及示例，您可以通过在文档中查找相关内容并仿造示例进行输出；
原则：1、输出一定在提供的文档内容中，且只能根据提供的文档内容来回答问题
      2、直接回答问题，不需要列出中间的思考过程，不需要拓展
      3、同一个问题，文本中有不同的答案，需要综合起来回答
      4、问题答案在文档中是分点/条、系列措施时列举时，需要全部列出不遗漏；
      5、文本中数值与两边的文字可能有空格或换行符，与没有这两类的符号的含义是相同的。例如 "持有8926万股" 与
        "持有 8926 万股" 、"持有 8926万股"、"持有 8926 \n万股"含义是相同的

步骤  1、判断下面问题中是否有多个问题，若存在多个问题，将问题拆分出多个问题；
      2、根据拆分出来的问题，精确定位文档中的内容，可能存在一个或多个位置，存在多个时需要都找出来；
      3、根据定位出来的位置前后内容，回答每个问题；
      4、优先根据文档原文来回答每个问题；若没有，再概况总结回答

示例
下面将提供示例文档内容、问题及输出，仿照样例输出
示例1文档内容：惠州光弘科技股份有限公司
报告期内2014年、2015年、2016年和2017年1-6月份 客户供料、非客户供料合作模式下营业成本构成及占比变动情况。客户供料模式下，
材料成本比重分别为\n13.15%、12.69%、13.38%和11.40%，\n人工成本比重分别为56.88%、58.17%、54.71%和51.52%，
制造费用比重分别为29.97%、29.14%、31.91%和37.08%,比例结构稳定;非客户供料模式下，材料成本比重分别为81.55%、85.69%、
91.33%和 \n90.17%。该业务模式下，其生产制造所需主要物料由发行人自行采购，按照产品整机的销售价格收取产品销售费，
因此其材料成本占比较高
示例1问题：报告期内，2016年年惠州光弘科技股份有限公司非客户供料模式下材料成本比重?
示例1输出：报告期内，2016年年惠州光弘科技股份有限公司非客户供料模式下材料成本比重91.33%

...
文档内容：{content}
问题：{question}
```

答案生成-通义14B-Chat

Prompt优化

召回内容重构，提升大模型的事实参考和问答效果

相关性高的内容前置

项目总结

最终比赛效果

复赛　　初赛

排名	参与者	组织	分数	data_query	text_comprehension	最优成绩提交日
1	饺子研究院	comm	83.27	90.46	72.48	2023-12-12
2	果树灌溉技术组	QTrade	82.02	89.72	70.48	2023-12-12
3	hxjj	华夏基金	81.05	91.02	66.09	2023-12-12

项目总结

Prompt对于LLM模型很重要

设计原则

- 系统角色定义指令
- 清晰明确的任务描述
- 详细的问题拆解与提示
- Few-shot提示样本的多样化
- 参考内容在模型输入中的位置

项目总结

未来工作优化

数据库查询

- 丰富长尾SQL指令生成，并融合后处理数据训练
- 迭代自动化样本增强流程，提升泛化效率和效果

文档内容问答

- 混合搜索方式，利用关键词+向量提升召回效果
- 探索长跨度信息的表示方式
- LLM对query改写的尝试

项目总结

金融大模型生态

通义千问金融大模型

- 金融知识更加丰富
- 金融指令表现优异，可微调性好

拓展生态

- API函数指令调用
- 更多类型文档问答

中國計算機學會 CHINA COMPUTER FEDERATION
CCF计算经济学专业组
北京大学前沿计算研究中心 Center on Frontiers of Computing Studies, Peking University
北大创新评论 PKU Innovation Review
博时基金 BOSERA FUNDS
阿里云
金证股份
未名数创 Undiscovered Innovation

谢谢大家