2023 CCF 计算经济学比赛
博金挑战赛

演讲人：吕亚龙

大模型说的队

目录

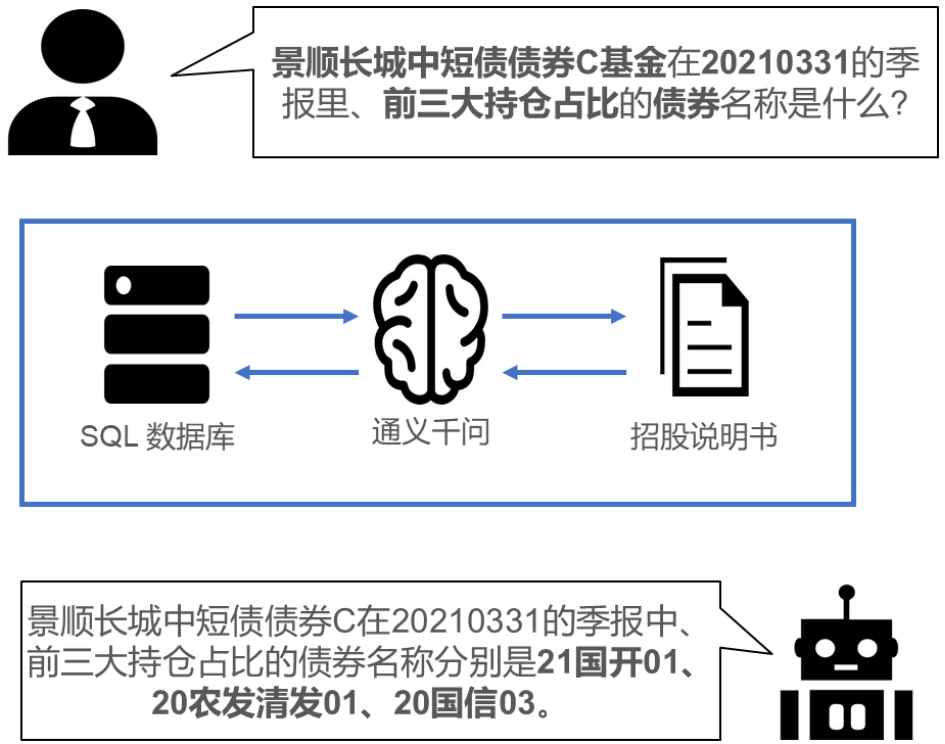
CONTENTS

第一部分

Part 1

赛题理解

赛题理解：四个精准



理解用户意图：分类精准

将用户问题分类为SQL查询和文本理解

转化机器语言：转码精准

将用户的自然语言转化为SQL查询语句

召回正确数据：搜索精准

将用户的问题与正确的招股书片段匹配

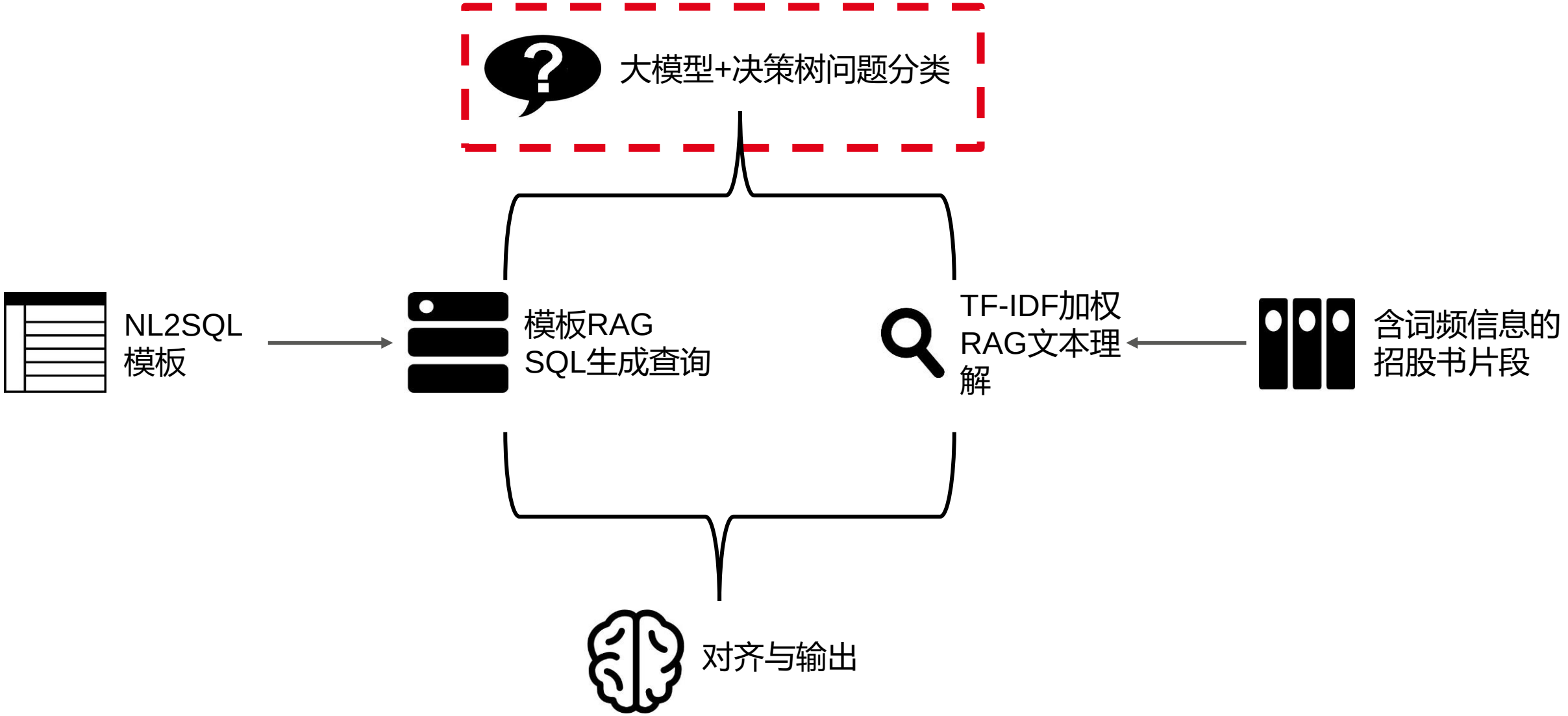
获得规范答案：输出精准

将回答与标准格式对齐，方便用户理解

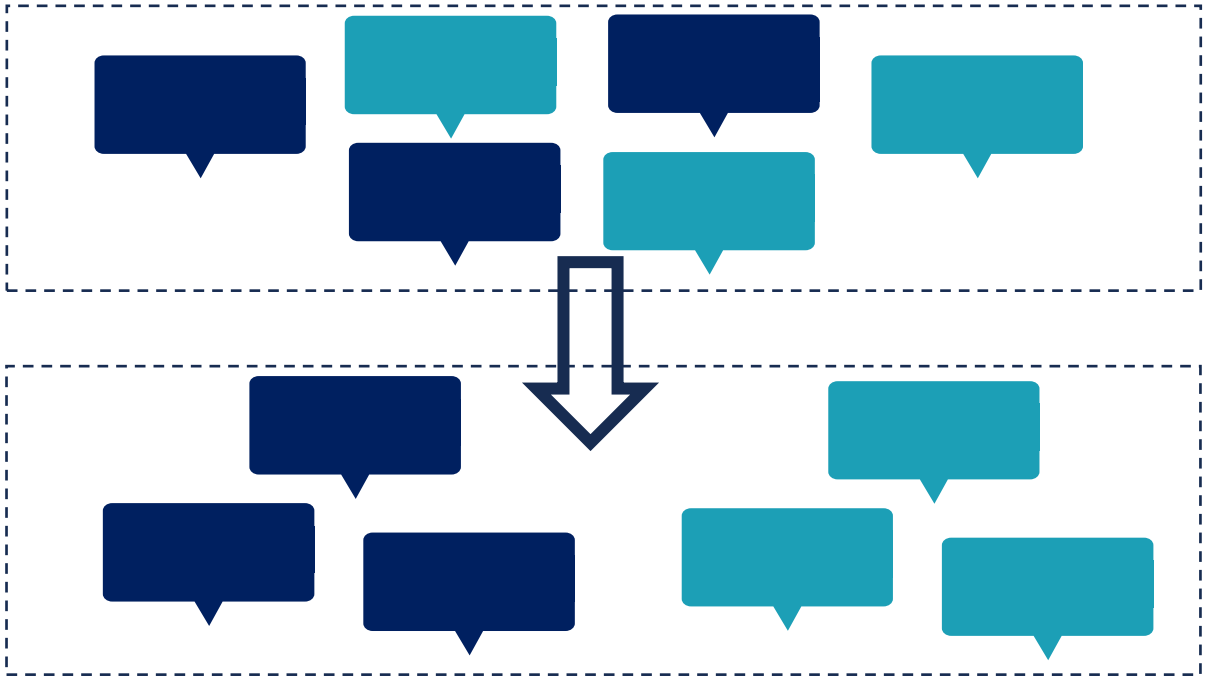
第二部分

Part 2

挑战与方案


大模型+决策树问题分类
NL2SQL
模板
模板RAG
SQL生成查询
TF-IDF加权
RAG文本理
解
含词频信息的
招股书片段
对齐与输出

深处种菱浅种稻：如何精准分类



例如：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%。

例如：湖南长远锂科股份有限公司变更设立时作为发起人的法人有哪些？

招股说明书

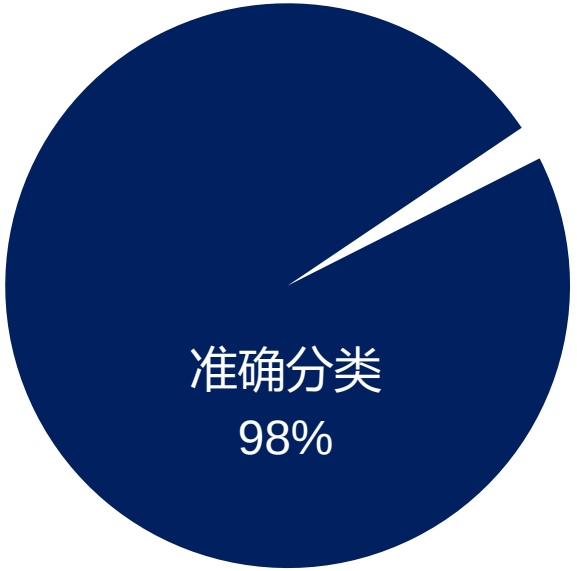
SQL数据库

深处种菱浅种稻：如何精准分类

大模型ICL

使用上下文学习（In Context Learning）方法构造prompt

根据大模型的回答是否包含关键字进行问题分类初赛问题集准确性**约为98%。**



深处种菱浅种稻：如何精准分类

你是一个问题分类器。对于每个提供给你的问题，你需要猜测答案是在该公司的**招股说明书**中还是在**基金股票数据库**里。以下是一些例子：

问题： “在2019年的中期报告里，金鹰基金管理有限公司管理的基金中，有多少比例的基金是个人投资者持有的份额超过机构投资者？希望得到一个精确到两位小数的百分比。”
回答： **“基金股票数据库”**

问题： “湖南长远锂科股份有限公司变更设立时作为发起人的法人有哪些？”
回答： **“该公司的招股说明书”**
…..（更多例子）
问题：**新问题**

403
问题: 请帮我查询下，在2019年09月的报告中，报告期基金总申购份额和报告期基金总赎回份额差额最大的一只基金的简称是什么？差额有多少？保留两位小数。
回答: **基金股票数据库**

404
问题: 2008年度、2009年度深圳市铁汉生态环境股份有限公司园林绿化工程收入的增长率分别为多少?
回答: 该问题的描述中没有提到任何具体公司的名称或代码，因此无法确定这是关于哪家公司的信息。同时，这个问题似乎也不是在询问基金的详细信息。因此，我无法将这个问题归类为关于**招股书或股票数据库**的问题。如果您能提供更多的上下文或相关信息，我将尽力回答您的问题。

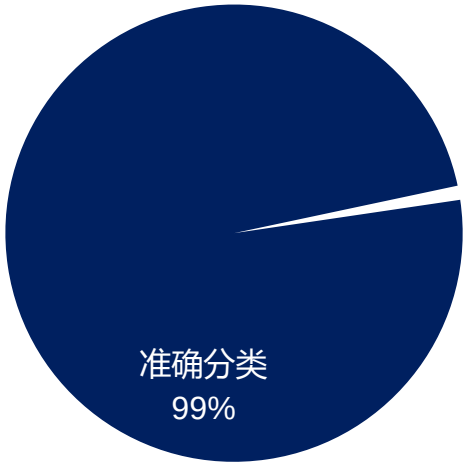
上下文学习输入 → **模型输出**

深处种菱浅种稻：如何精准分类

决策树

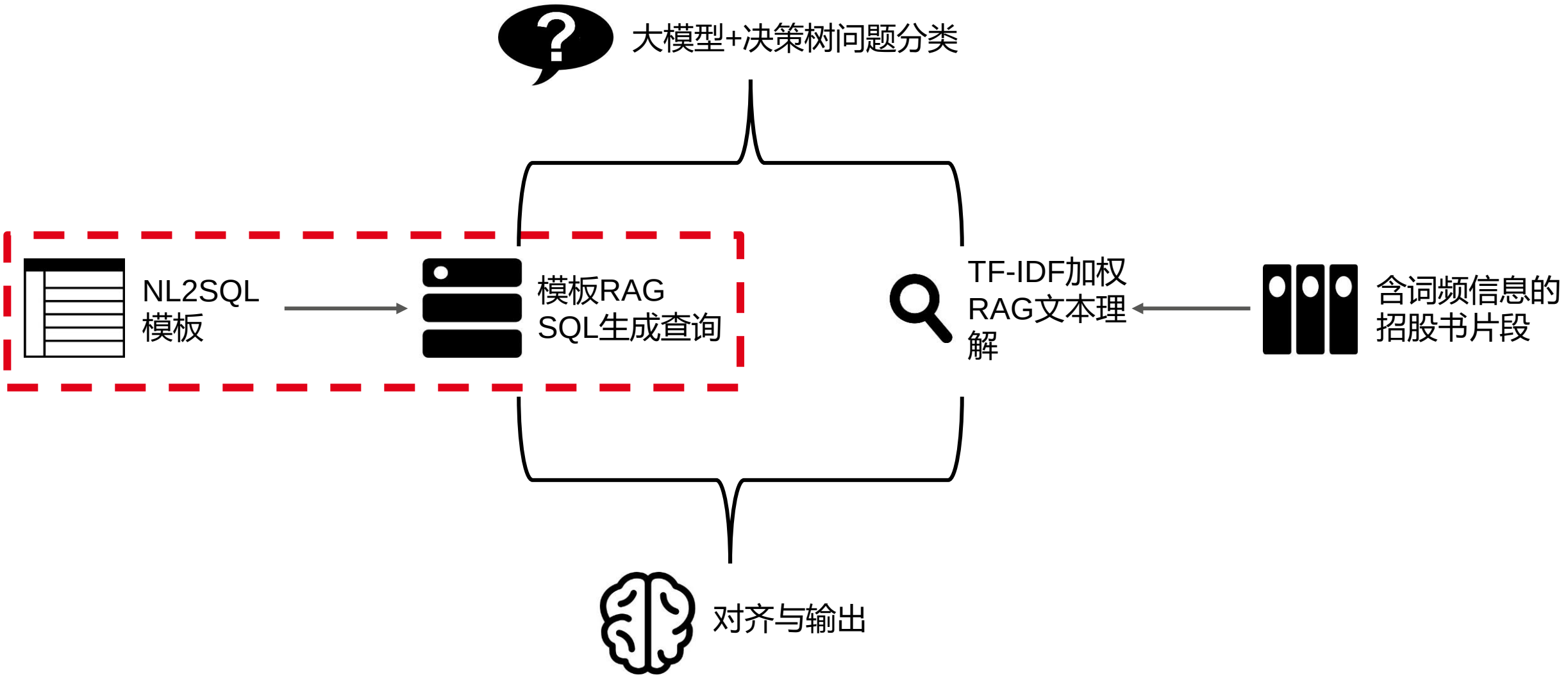
- 大模型输出含有“招股说明书”且不含“基金股票数据库”→**文本理解**
- 否则，问题含有公司全称→**文本理解**
- 其余→ **SQL查询**

初赛问题集准确性达到99%。



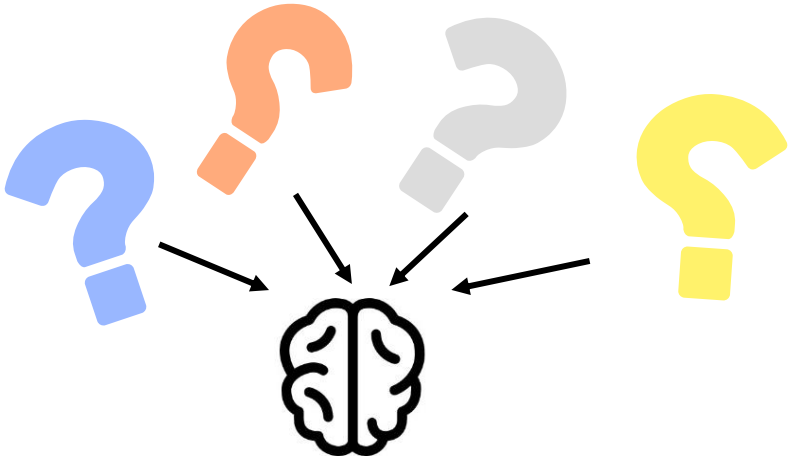
13
Q: 近三年，**广东银禧科技股份有限公司**总资产周转率分别为多少?
A: 该问题的描述中没有提到任何具体公司的名称或股票代码，因此无法确定这是在哪家公司的**招股说明书**中找到的答案，也无法确定这是在**基金股票数据库**中的哪个位置可以找到答案。
……
404
Q: 2008年度、2009年度**深圳市铁汉生态环境股份有限公司**园林绿化工程收入的增长率分别为多少?
A: 该问题的描述中没有提到任何具体公司的名称或代码，因此无法确定这是关于哪家公司的信息。同时，这个问题似乎也不是在询问基金的详细信息。因此，我无法将这个问题归类为关于
招股书或股票数据库的问题。如果您能提供更多的上下文或相关信息，我将尽力回答您的问题。

模型输出

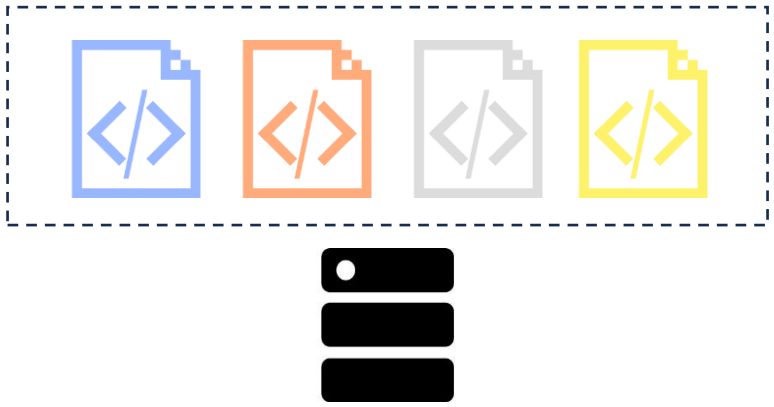

大模型+决策树问题分类
NL2SQL
模板
模板RAG
SQL生成查询
TF-IDF加权
RAG文本理
解
含词频信息的
招股书片段
对齐与输出

解一卷而众篇明：如何精准转码

自然语言：



SQL代码：



信码由缰不可取

官方提供了2019-2021年度的基金股票信息、共包含10张表，数量多、字段语义近，根据自然语言从零直接生成SQL语句的baseline方案分数仅为40+。

信码由缰精度低、难度大！

探索性数据分析

Jaccard 相似度： $Jaccard\ Similarity = |A \cap B| / |A \cup B|$

探索性数据分析

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%。

204: 请帮我计算，在**20210715**，**中信行业分类**划分的一级行业为**消费者服务**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%。

128: 请帮我计算，在**20200521**，**申万行业分类**划分的一级行业为**采掘**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%。

高 Jaccard 相似度→SQL语句结构相似

…… WHERE [交易日] = 20210105 AND [行业划分标准] = 中信行业分类 AND [一级行业名称] = 综合金融 ……

…… WHERE [交易日] = 20210715 AND [行业划分标准] = 中信行业分类 AND [一级行业名称] = 消费者服务 ……

…… WHERE [交易日] = 20200521 AND [行业划分标准] = 申万行业分类 AND [一级行业名称] = 采掘 ……

谱聚类

聚类个数=75，每类内问题基本同质。

聚 类 1

我想知道招商安润灵活配置混合A基金，在2021年半年度报告中，前10大重仓股中，有多少只股票在报告期内取得正收益。

我想知道圆信永丰高端制造混合基金，在2020年半年度报告中，前10大重仓股中，有多少只股票在报告期内取得正收益。

我想知道鹏扬景科混合A基金，在2020年半年度报告中，前10大重仓股中，有多少只股票在报告期内取得正收益。

我想知道国寿安保国证创业板中盘精选88ETF基金，在2021年半年度报告中，前10大重仓股中，有多少只股票在报告期内取得正收益。

我想知道东方红品质优选两年定期开放混合基金，在2021年半年度报告中，前10大重仓股中，有多少只股票在报告期内取得正收益。

我想知道国泰CES半导体芯片行业ETF联接A基金，在2020年半年度报告中，前10大重仓股中，有多少只股票在报告期内取得正收益。

我想知道长城品质成长混合C基金，在2021年半年度报告中，前10大重仓股中，有多少只股票在报告期内取得正收益。

聚 类 2

20200102日，一级行业为建筑的股票的成交金额合计是多少？取整。

20200706日，一级行业为机械的股票的成交金额合计是多少？取整。

20200722日，一级行业为化工的股票的成交金额合计是多少？取整。

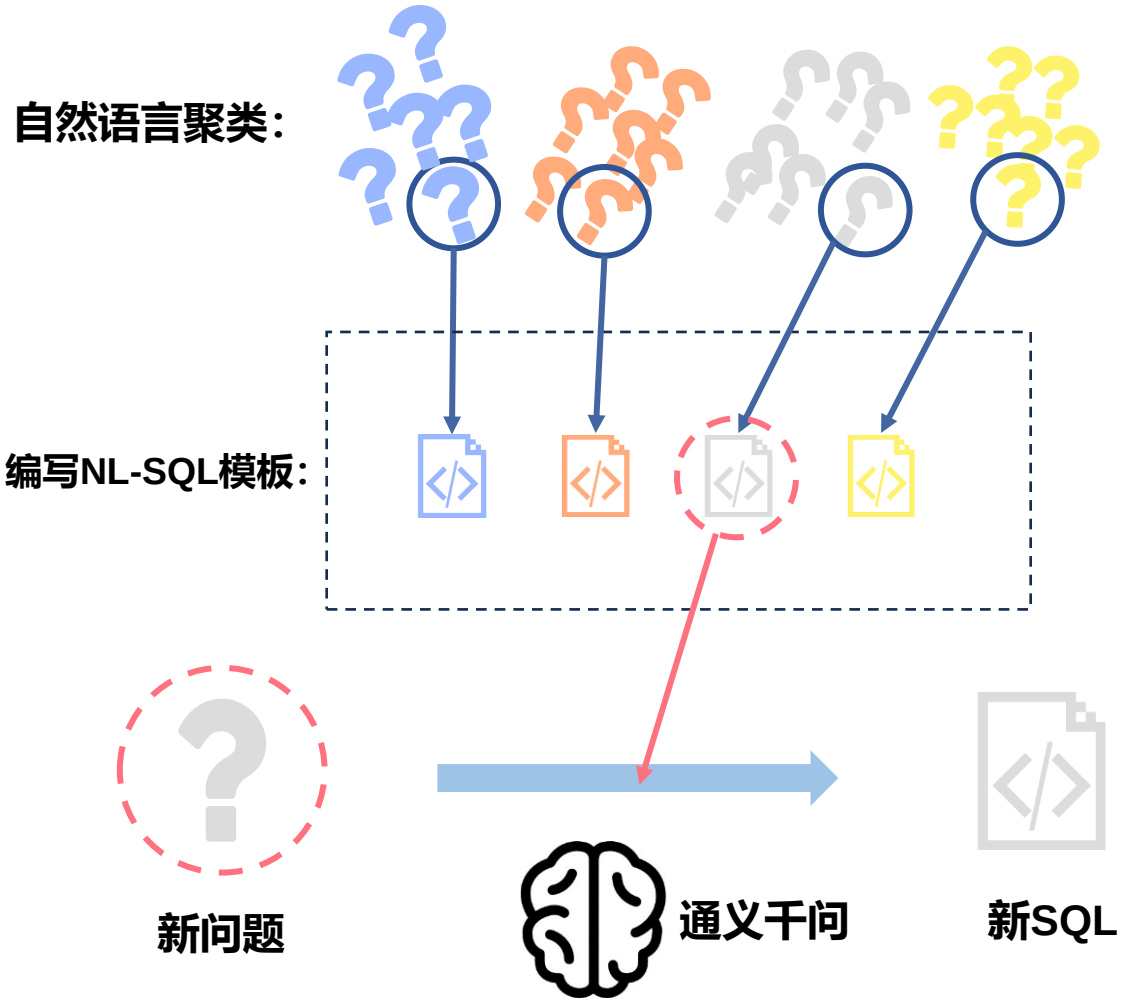
20201116日，一级行业为电气设备的股票的成交金额合计是多少？取整。

20200225日，一级行业为房地产的股票的成交金额合计是多少？取整。

20211126日，一级行业为休闲服务的股票的成交金额合计是多少？取整。

20200424日，一级行业为汽车的股票的成交金额合计是多少？取整。

解一卷而众篇明：如何精准转码



检索增强生成（RAG）

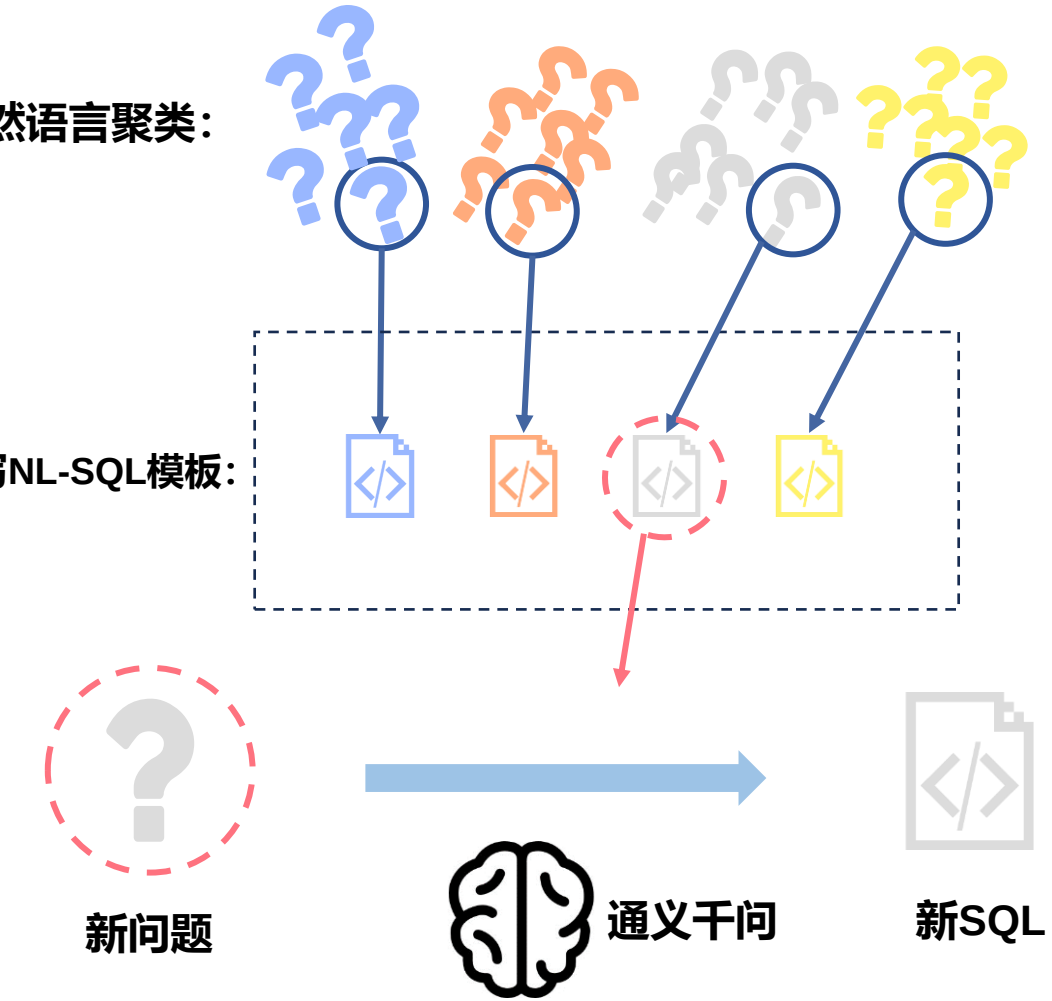
将新问题和与新问题最相似的模板输入大模型，大模型输出新问题对应的SQL语句.

- **高可靠性：** “填空题”比“作文题”简单，可以处理3-4层子查询和JOIN
- **高可解释性：**模板与生成结果强相关
- **高可拓展性**：面对新问题无需重新训练，只需补充少量模板

解一卷而众篇明：如何精准转码

自然语言聚类：

编写NL-SQL模板：



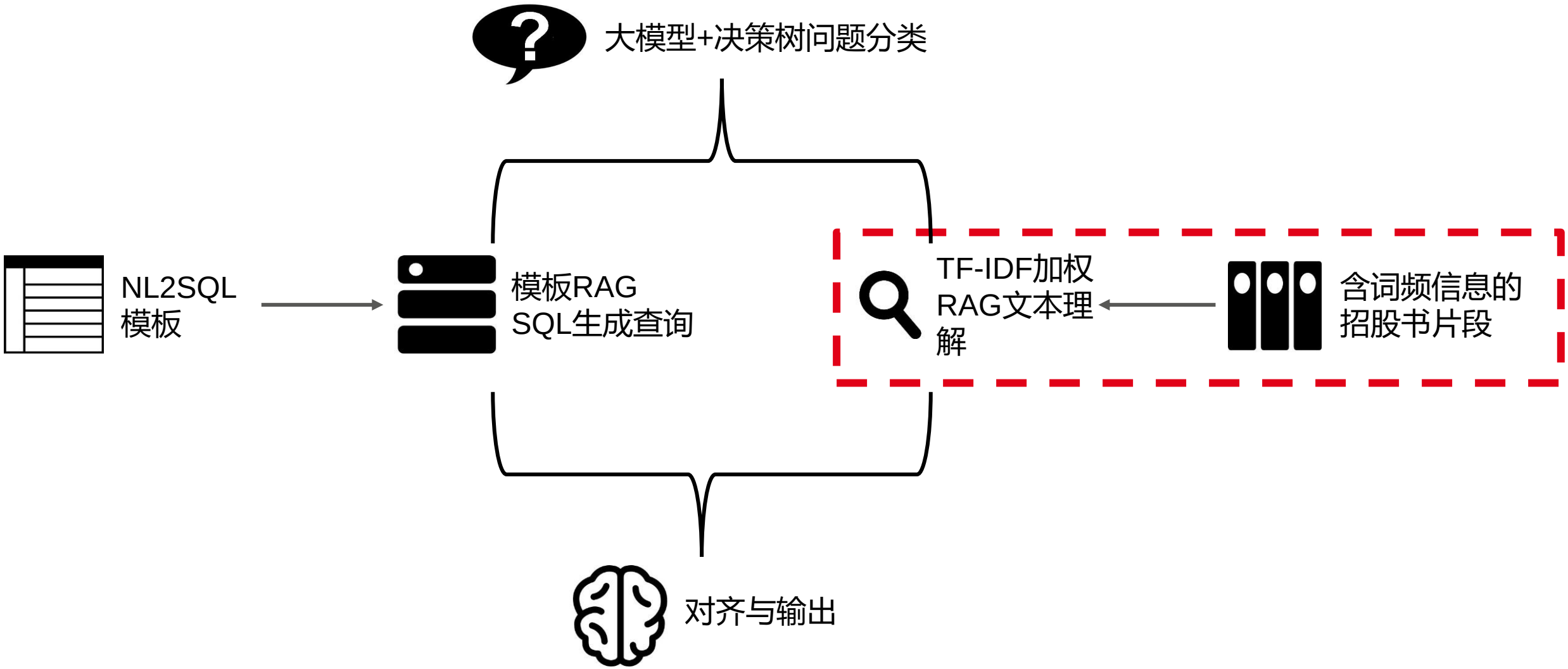
检索增强生成输入/输出

将下面的问题转化为SQL语句。下面是一些例子。

对于问题：请帮我计算，在**20210720**，中信行业分类划分的一级行业为**彩虹小马饲养**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%。
你应该回答：SELECT t1.[股票代码], ((t1.[收盘价(元)] - t1.[昨收盘(元)]) / t1.[昨收盘(元)]) * 100 AS 涨跌幅
FROM A股票日行情表 t1
JOIN A股公司行业划分表 t2 ON t1.[股票代码] = t2.[股票代码] AND t1.[交易日] = t2.[交易日期]
WHERE t1.[交易日] = '**20210720**'
AND t2.[行业划分标准] = '中信行业分类'
AND t2.[一级行业名称] = '**彩虹小马饲养**'
ORDER BY 涨跌幅 DESC
LIMIT 1;

对于问题：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%。
你应该回答：

```
SELECT t1.[股票代码], ((t1.[收盘价(元)] - t1.[昨收盘(元)]) / t1.[昨收盘(元)]) * 100 AS 涨跌幅
FROM A股票日行情表 t1
JOIN A股公司行业划分表 t2
ON t1.[股票代码] = t2.[股票代码]
AND t1.[交易日] = t2.[交易日期]
WHERE t1.[交易日] = '20210105'
AND t2.[行业划分标准] = '中信行业分类'
AND t2.[一级行业名称] = '综合金融'
ORDER BY 涨跌幅 DESC LIMIT 1;
```

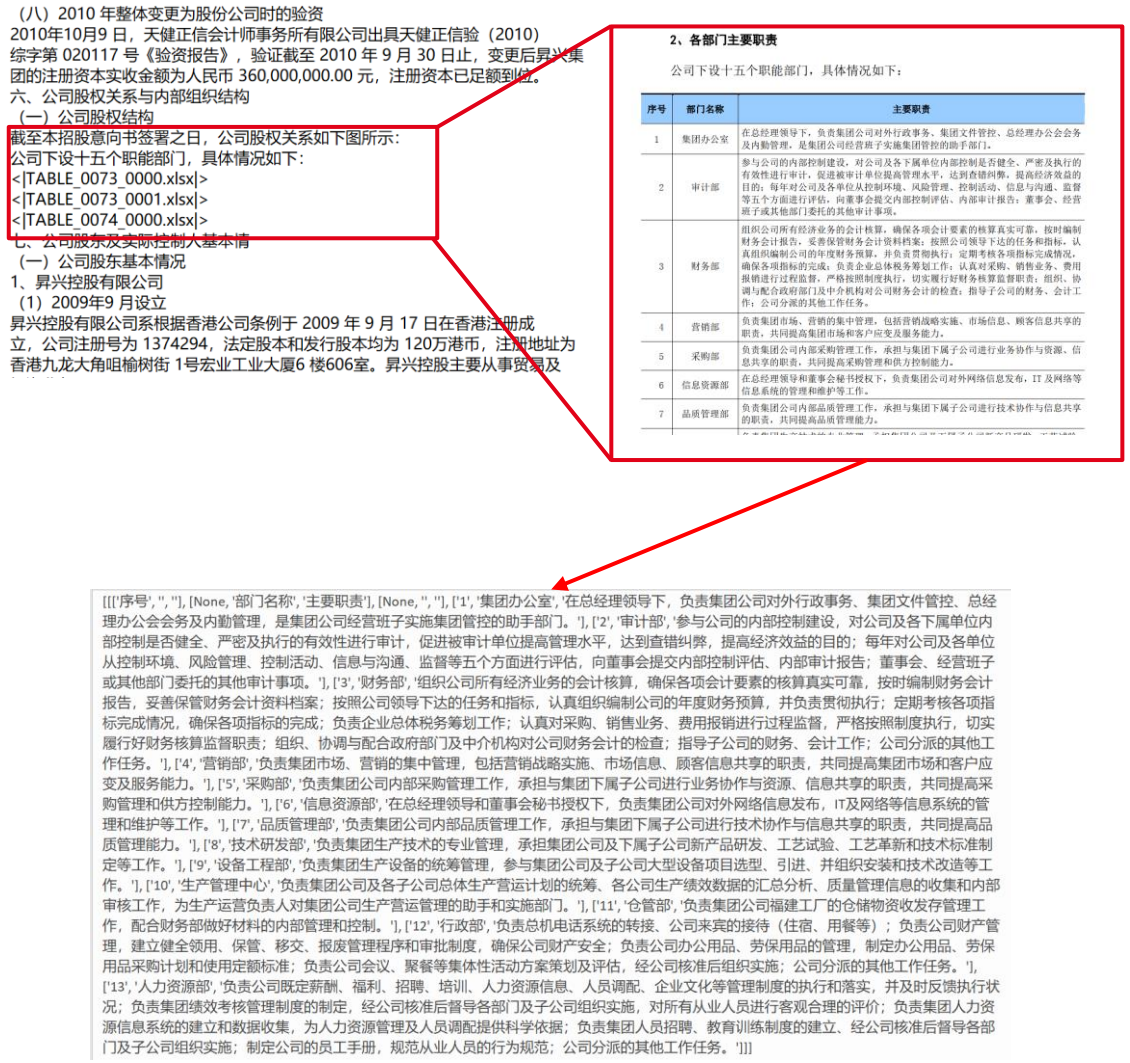

大模型+决策树问题分类
NL2SQL
模板
模板RAG
SQL生成查询
TF-IDF加权
RAG文本理
解
含词频信息的
招股书片段
对齐与输出

磨刀不误砍柴工：补充表格信息

基于pdfplumber的有限状态机

官方提供的TXT格式不含表格。

我们编写基于pdfplumber的有限状态机辅助抽取了无侧线表格，使信息更加全面精准。

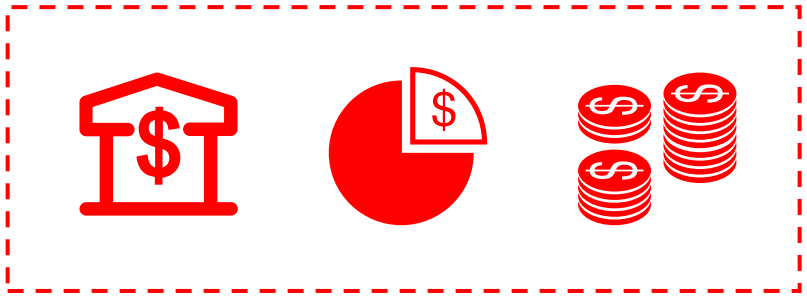


[[['序号', '', ''], [None, '部门名称', '主要职责'], [None, '', ''], ['1', '集团办公室', '在总经理领导下，负责集团公司对外行政事务、集团文件管控、总经理办公会会务及内勤管理，是集团公司经营班子实施集团管控的助手部门。'], ['2', '审计部', '参与公司的内部控制建设，对公司及各下属单位内部控制是否健全、严密及执行的有效性进行审计，促进被审计单位提高管理水平，达到查错纠弊，提高经济效益的目的；每年对公司及各单位从控制环境、风险管理、控制活动、信息与沟通、监督等五个方面进行评估，向董事会提交内部控制评估、内部审计报告；董事会、经营班子或其他部门委托的其他审计事项。'], ['3', '财务部', '组织公司所有经济业务的会计核算，确保各项会计要素的核算真实可靠，按时编制财务会计报告，妥善保管财务会计资料档案；按照公司领导下达的任务和指标，认真组织编制公司的年度财务预算，并负责贯彻执行；定期考核各项指标完成情况，确保各项指标的完成；负责企业总体税务筹划工作；认真对采购、销售业务、费用报销进行过程监督，严格按照制度执行，切实履行好财务核算监督职责；组织、协调与配合政府部门及中介机构对公司财务会计的检查；指导子公司的财务、会计工作；公司分派的其他工作任务。'], ['4', '营销部', '负责集团市场、营销的集中管理，包括营销战略实施、市场信息、顾客信息共享的职责，共同提高集团市场和客户应变及服务能力。'], ['5', '采购部', '负责集团公司内部采购管理工作，承担与集团下属子公司进行业务协作与资源、信息共享的职责，共同提高采购管理和供方控制能力。'], ['6', '信息资源部', '在总经理领导和董事会秘书授权下，负责集团公司对外网络信息发布，IT及网络等信息系统的管理和维护等工作。'], ['7', '品质管理部', '负责集团公司内部品质管理工作，承担与集团下属子公司进行技术协作与信息共享的职责，共同提高品质管理能力。'], ['8', '技术研发部', '负责集团生产技术的专业管理，承担集团公司及下属子公司新产品研发、工艺试验、工艺革新和技术标准制定等工作。'], ['9', '设备工程部', '负责集团生产设备的统筹管理，参与集团公司及子公司大型设备项目选型、引进、并组织安装和技术改造等工作。'], ['10', '生产管理中心', '负责集团公司及各子公司总体生产营运计划的统筹、各公司生产绩效数据的汇总分析、质量管理信息的收集和内部审核工作，为生产运营负责人对集团公司生产营运管理的助手和实施部门。'], ['11', '仓管部', '负责集团公司福建工厂的仓储物资收发存管理工作，配合财务部做好材料的内部管理和控制。'], ['12', '行政部', '负责总机电话系统的转接、公司来宾的接待（住宿、用餐等）；负责公司财产管理，建立健全领用、保管、移交、报废管理程序和审批制度，确保公司财产安全；负责公司办公用品、劳保用品的管理，制定办公用品、劳保用品采购计划和使用定额标准；负责公司会议、聚餐等集体性活动方案策划及评估，经公司核准后组织实施；公司分派的其他工作任务。'], ['13', '人力资源部', '负责公司既定薪酬、福利、招聘、培训、人力资源信息、人员调配、企业文化等管理制度的执行和落实，并及时反馈执行状况；负责集团绩效考核管理制度的制定，经公司核准后督导各部门及子公司组织实施，对所有从业人员进行客观合理的评价；负责集团人力资源信息系统的建立和数据收集，为人力资源管理及人员调配提供科学依据；负责集团人员招聘、教育训练制度的建立、经公司核准后督导各部门及子公司组织实施；制定公司的员工手册，规范从业人员的行为规范；公司分派的其他工作任务。']]

众里寻他千百度：如何精准搜索

预训练语境

金融语境

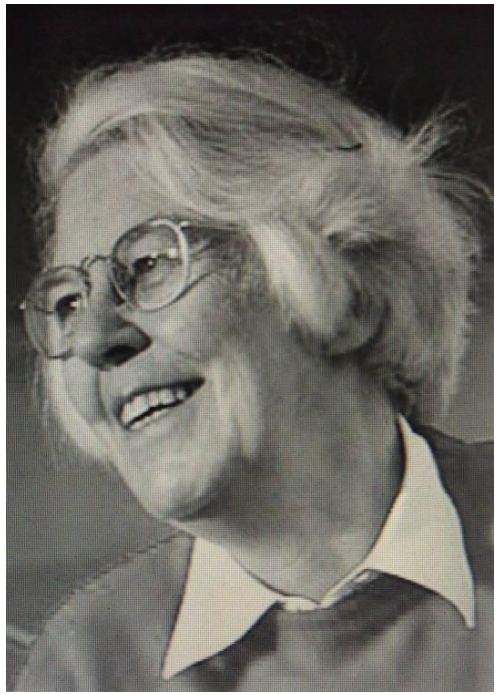


预训练语境与金融语境差别大

精准搜索是文本理解的核心。必须在数十万字的招股书中准确找到含有答案的片段，才能让大模型在合理的时间内给出正确答案。

预训练Embedding语义搜索效果不佳，预训练语境与金融语境差别大。

众里寻他千百度：如何精准搜索



克伦·施拜克·琼斯

词频-逆文件频率（TF-IDF）

token在单个片段中的权重随着它在单个片段中出现的次数成正比增加，但同时会随着它在所在整个语料库文件中出现的频率成反比下降。

仅词频

词频-逆文件频率

众里寻他千百度：如何精准搜索

TF-IDF余弦相似度

高可靠性和泛化能力：自适应权重，无需过滤停用词

$$s_{q,p_i} = \frac{\sum_{t \in \mathcal{T}} f_{q,t} f_{p_i,t} f_{a,t}^{-2}}{\sqrt{\sum_{t \in \mathcal{T}} f_{q,t}^2 f_{a,t}^{-2} \sum_{t \in \mathcal{T}} f_{p_i,t}^2 f_{a,t}^{-2}}}$$

- $f_{q,t}$ 为问题中 token t 的词频；
- $f_{p_i,t}$ 为第 i 段文本中 token t 的词频；
- $f_{a,t}$ 为招股书中 token t 的词频。

众里寻他千百度：如何精准搜索

Q1: 湖南长远锂科股份有限公司变更设立时作为发起人的法人有哪些?

token	招股书全文中token出现次数
有	100
作为	84
人	69
变更	36
设立	27
法人	14
发起	2

Vanilla余弦

相似度被“有” “作为”主导

TF-IDF余弦

“发起”的权重最大

众里寻他千百度：如何精准搜索

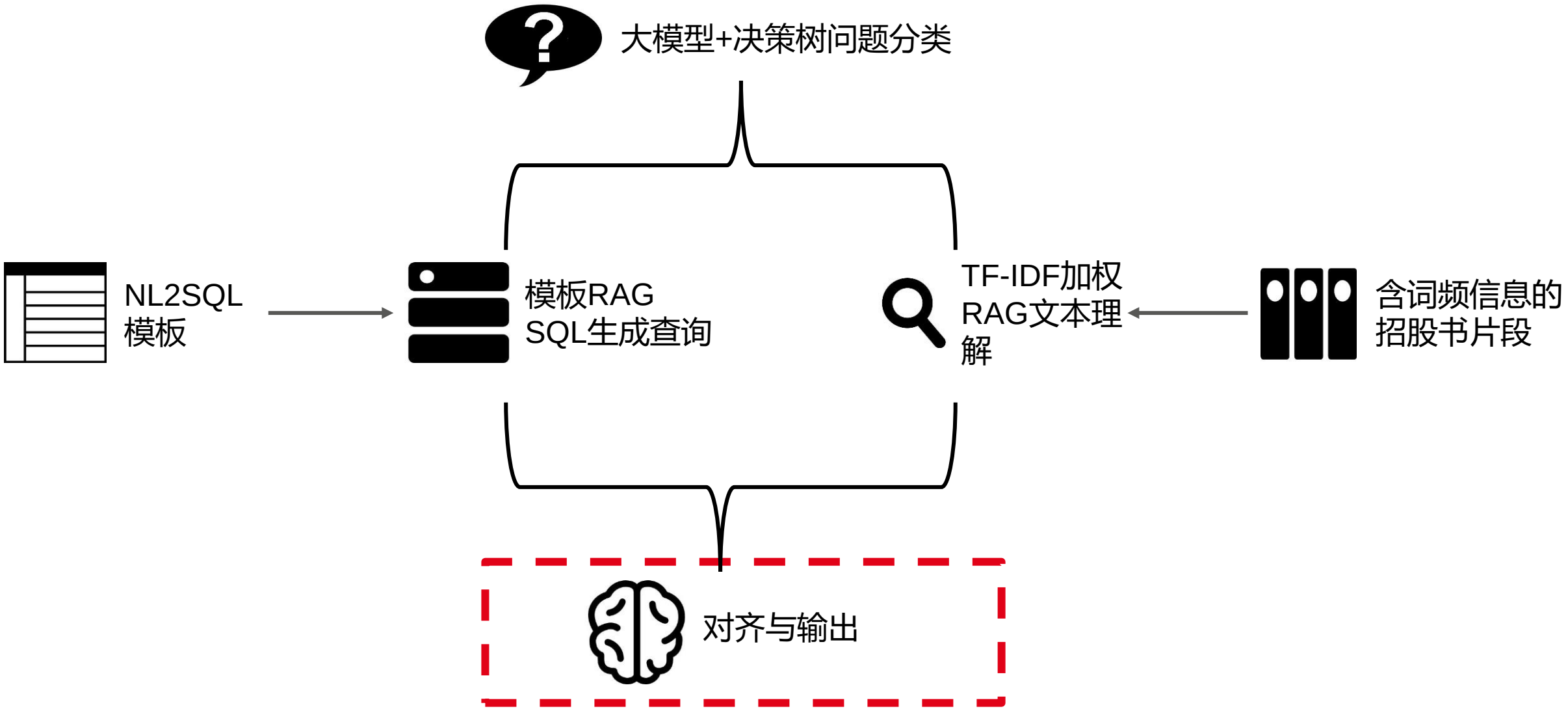
Q1: 湖南长远锂科股份有限公司变更设立时作为发起人的法人有哪些?

Vanilla-余弦相似度top1召回

……公司**法人**治理结构，在资产、人员、财务、机构、业务等方面与控股股东、实际控制人及其控制的其他企业之间相互独立，具有完整的业务体系及面向市场独立经营的能力，具有独立完整的供应、生产和销售系统。（一）资产独立公司系由锂科有限整体**变更设立**而来，……

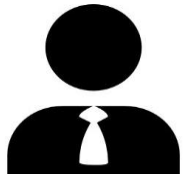
TF-IDF余弦相似度top1召回

……（二）股份公司设立情况公司由五矿股份、长沙矿冶院、宁波创元、深圳安晏、尚颀颀旻、安鹏智慧、国调基金、建信投资、信石信远、华能融科、中信投资、三峡金石、伊敦基金、中启洞鉴14 家**法人**作为**发起人**，由锂科有限于 2019年4 月整体**变更设立**股份有限公司。2019 年 4 月，中国五矿出具《关于湖南长远锂科有限公司实施股份制改革的意见》（中国五矿企管[2019]239 号），同意锂科有限以 2018 年 12 月 31 日作为整体改制基准日，以有限责任公司整体**变更**方式，**发起**设立股份有限公司。……

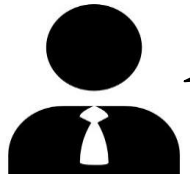

大模型+决策树问题分类
NL2SQL
模板
模板RAG
SQL生成查询
TF-IDF加权
RAG文本理
解
含词频信息的
招股书片段
对齐与输出

书同文字车同轨：如何精准输出

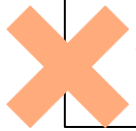
完整、准确、全面的答案输出



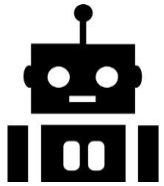
景顺长城中短债债券C基金在20210331的季报里、前三大持仓占比的债券名称是什么?

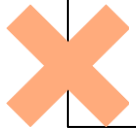


请帮我计算，在20210105，中信行业分类划分的一级行业为综合金融行业中，涨跌幅最大股票的股票代码是？涨跌幅是多少？百分数保留两位小数。

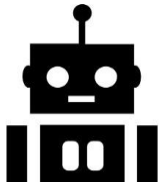


21国开01、20农发清发01、20国信03。

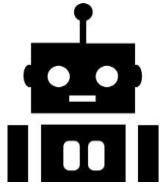




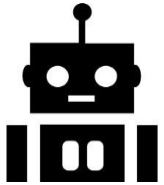
股票代码是600120，涨跌幅是0%。



景顺长城中短债债券C在20210331的季报中、前三大持仓占比的债券名称分别是21国开01、20农发清发01、20国信03。

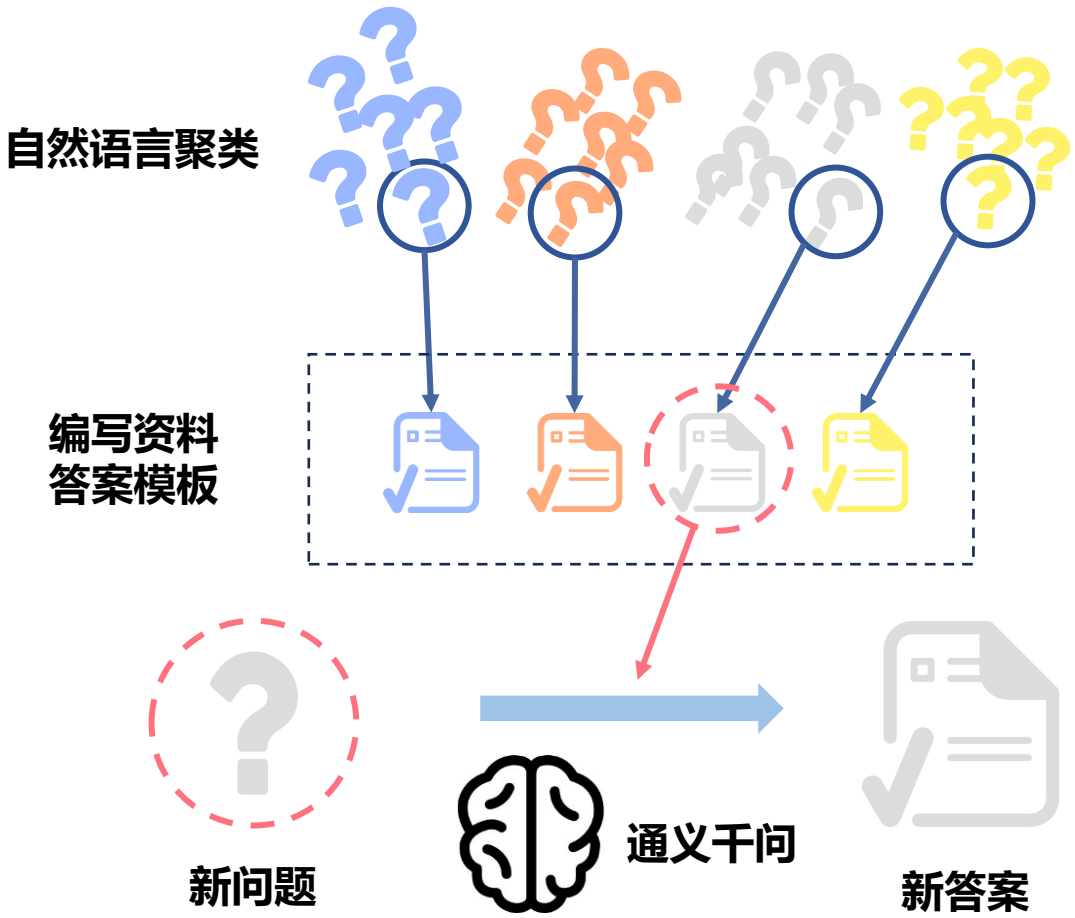


在20210105，中信行业分类划分的一级行业为综合金融行业中，涨跌幅最大股票的股票代码是600120，涨跌幅是0.00%。



书同文字车同轨：如何精准输出

使用回答模板库进行答案对齐生成，统一回答格式。



检索增强生成输入/输出

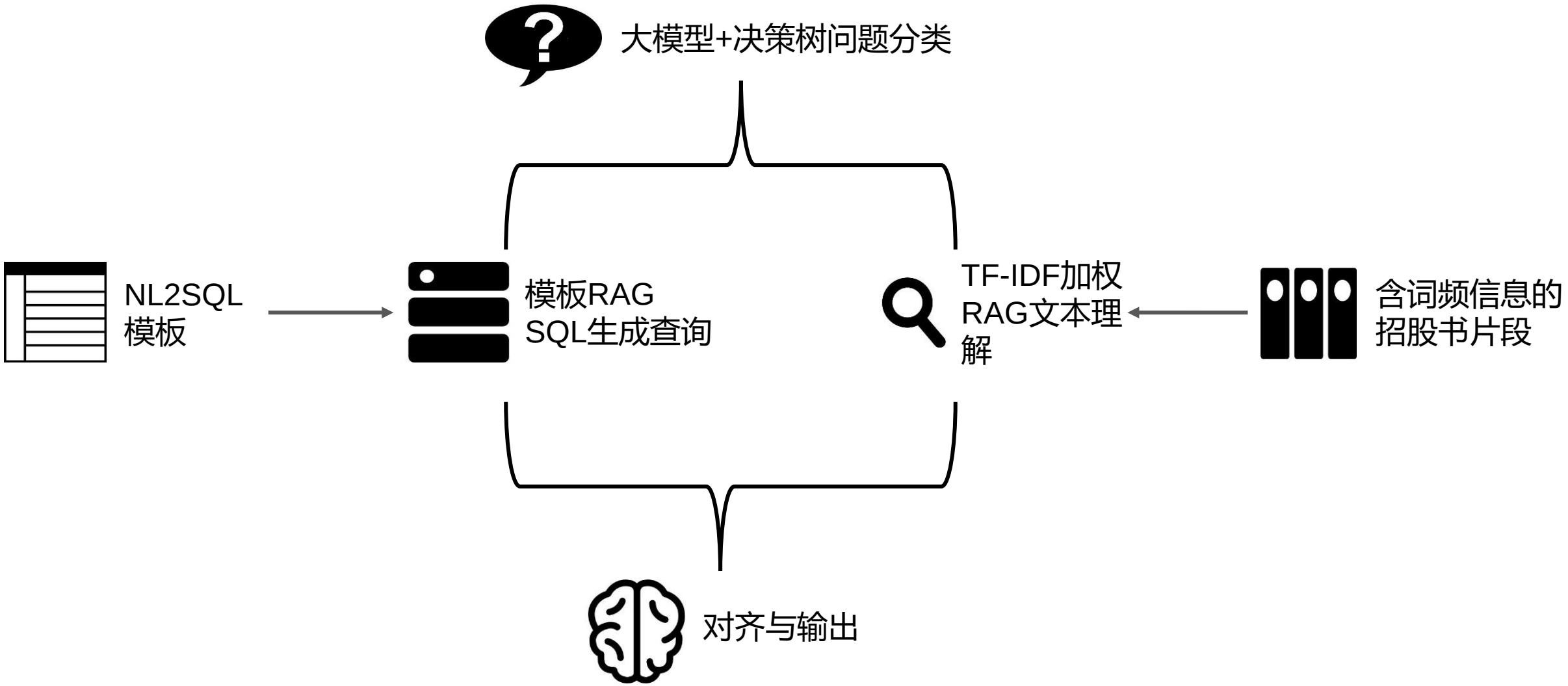
你要进行句子生成工作，根据提供的资料来回答对应的问题。下面是一些例子，注意问题与回答中的小数位数。
例1：
问题：“请帮我计算，在20210720，中信行业分类划分的一级行业为彩虹小马饲养行业中，涨跌幅最大股票的股票代码是？涨跌幅是多少？百分数保留两位小数。股票涨跌幅定义为：（收盘价 - 前一日收盘价 / 前一日收盘价）* 100%。”
资料：“[('114514', **4.81481481481481**)]”
你应该回答：在20210720，中信行业分类划分的一级行业为彩虹小马饲养行业中，涨跌幅最大股票的股票代码是114514，涨跌幅是**4.81**%。

例2：
问题：“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%”
资料：” [('600120', **0.0**)]”
你应该回答：

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600120，涨跌幅是**0.00**%。

第三部分
Part 3

架构总览


大模型+决策树问题分类
NL2SQL
模板
模板RAG
SQL生成查询
TF-IDF加权
RAG文本理
解
含词频信息的
招股书片段
对齐与输出

第四部分

Part 4

总结与展望

总结与展望：

步骤	优点	缺点	改进方向
问题分类	结构简单、识别准确	对无公司名和多公司名的问题支持较差	利用 p-tunning 改善大模型分类能力
SQL生成	高可靠性：支持复杂SQL 低标注：全程仅标注不到200段SQL 低代码：泛化无需重新训练 高可解释性与可控性	当问题类别数过多时，模板召回流程或无法命中	训练根据NL相似性预测SQL相似性的“小模型”，改进模板召回流程
文本理解	召回准确 高泛化能力：无需更新停用词	不支持基于语义的模糊搜索	引入金融语境下训练/微调过的Embedding模型
答案对齐	高可解释性与可控性	依赖于大模型的指令跟随能力	使用正则表达式识别小数位数要求并检查答案