

# SMP 2023 ChatGLM 金融大模型挑战赛方案-结婚买房代代韭菜队

比赛地址

## 一、简介

### 1. 代码仓库

- [推理代码](#)
- [训练代码](#)

### 2. 描述约定

- 问题类型

问题大类型	子类型	描述
type1	type11	可以直接完成的客观基础查询。 如： 2020 年一品红药业股份有限公司其他非流动金融资产是多少元？
	type11-2	可以直接完成的隐含类基础查询，在赛题内主要指法定代表人是否相同的问题。 如： 南京寒锐钴业股份有限公司 2021 年的法定代表人与 2019 年的法定代表人相比相同吗？
	type12	统计类客观查询。 如： 2019-2021 年哪些家上市公司货币总额均位列前十？
type2	type21	在直接完成的客观基础查询之上再进行数学运算的查询，数学运算的公式属于特定公式，不具有广泛推导的性质。 如： 2019 年贵研铂业股份有限公司速动比率为多少？保留两位小数。
	type22	在直接完成的客观基础查询之上再进行数学运算的查询，数学运算的公式固定。 如：

		在 2020 年的财务数据中，新华文轩管理费用增长率是多少？保留两位小数
type3	type31	需要结合财报知识进行总结回答的无精确答案问题。 如： 根据 2019 年的年报数据，哈空调研发投入的情况，请做简要分析。
	type32	金融知识理解的无精确答案问题。 如： 什么是其他债权投资？

• 模型

本方案中所有的微调模型均使用ptuningv2方法。

中文名称	英文名称	描述
路由模型	router	问题类型分类器
自然语言-数据库结构化查询语言转换模型	nl2sql	将自然语言转换为 SQL 语句
规范化回答模型	normalize	结合 SQL 查询结果与问题输出对应问题的规范化回答
无微调模型	no-tune chatglm2-6b	原始chatglm2-6b

3. 总体架构

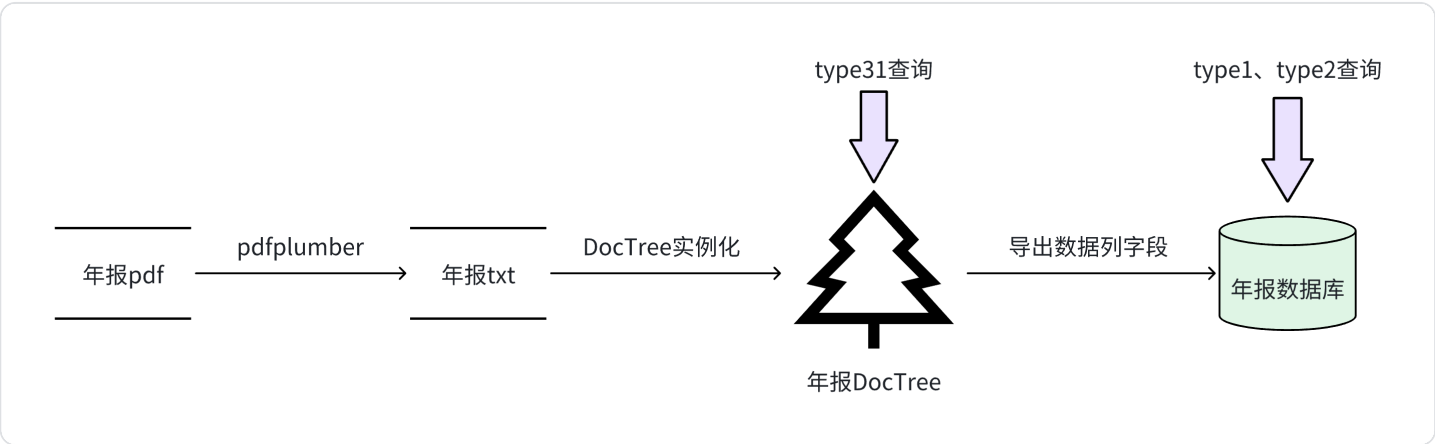


图 1 数据处理

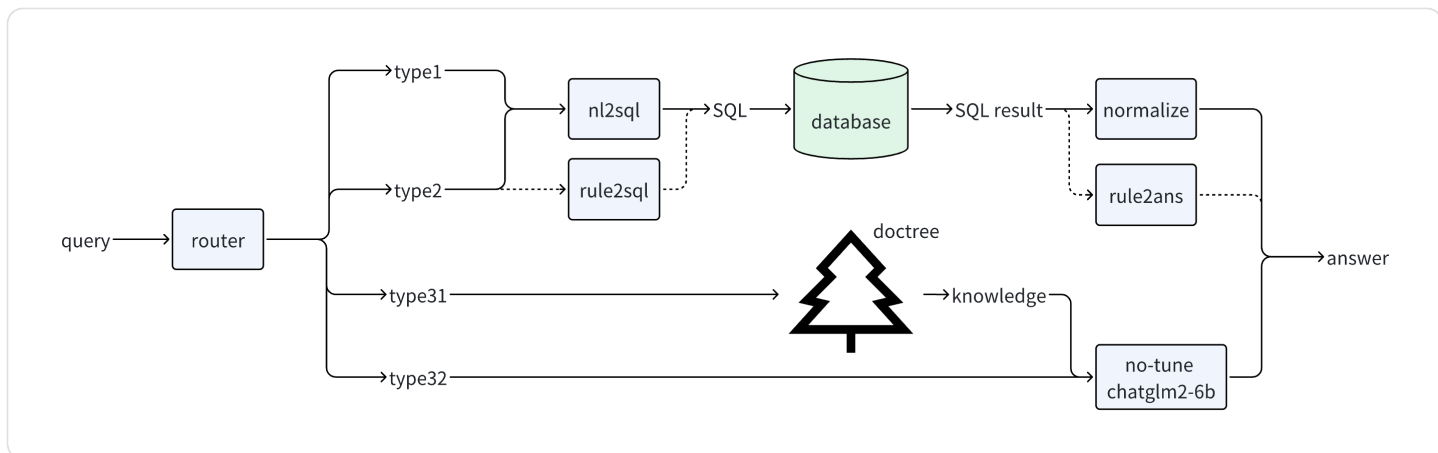


图 2 QA 具体流程

## 二、比赛思路

### 1. 得分思路

本次大赛赛题中，type1 需要精准的客观答案；type2 本质上是 type1 的进一步延伸组合；type31 需要从年报中找到相关的知识来进行总结回答；type32 依赖于模型的金融知识，不需要其他的输入。

本次大赛中，type32 类型基本无分差；type31 类型分差也较小；影响排名的主要因素来自于 type1，type2 得分。

### 2. 赛题难点

- 对于 type1、type2、type31，都需要额外的知识输入。**构建合理的数据结构来提供需要的知识是比赛的第一个难点。**
- 在细分问题类型中，type11-2 法定代表人是否相同、type12 统计类查询题目两类问题的问题形态较具有多样性，使用关键词来匹配提取问题的意图较为复杂且十分难以迭代。**完全、清晰的理解具有多样形态的问题以及泛化是比赛的第二个难点。**
- 根据得分规则，回答中除了包括直接的答案以外，还需要包括问题中的关键词。如问题为“2019 年中国工商银行财务费用是多少元？”，回答中也需要包括“2019 年”、“中国工商银行”、“财务费用”，这些关键词占比分的 25%，在比赛后期 type1、type2 都在 80 甚至 90 分以上竞争时，这一部分的得分非常关键。此外，得分规则中包括了相似度的要求，而标准答案保持着和问题高度一致的规范性，也占据了相当的得分比例。**保持问题中的关键词进行规范化的回答是比赛的第三个难点。**

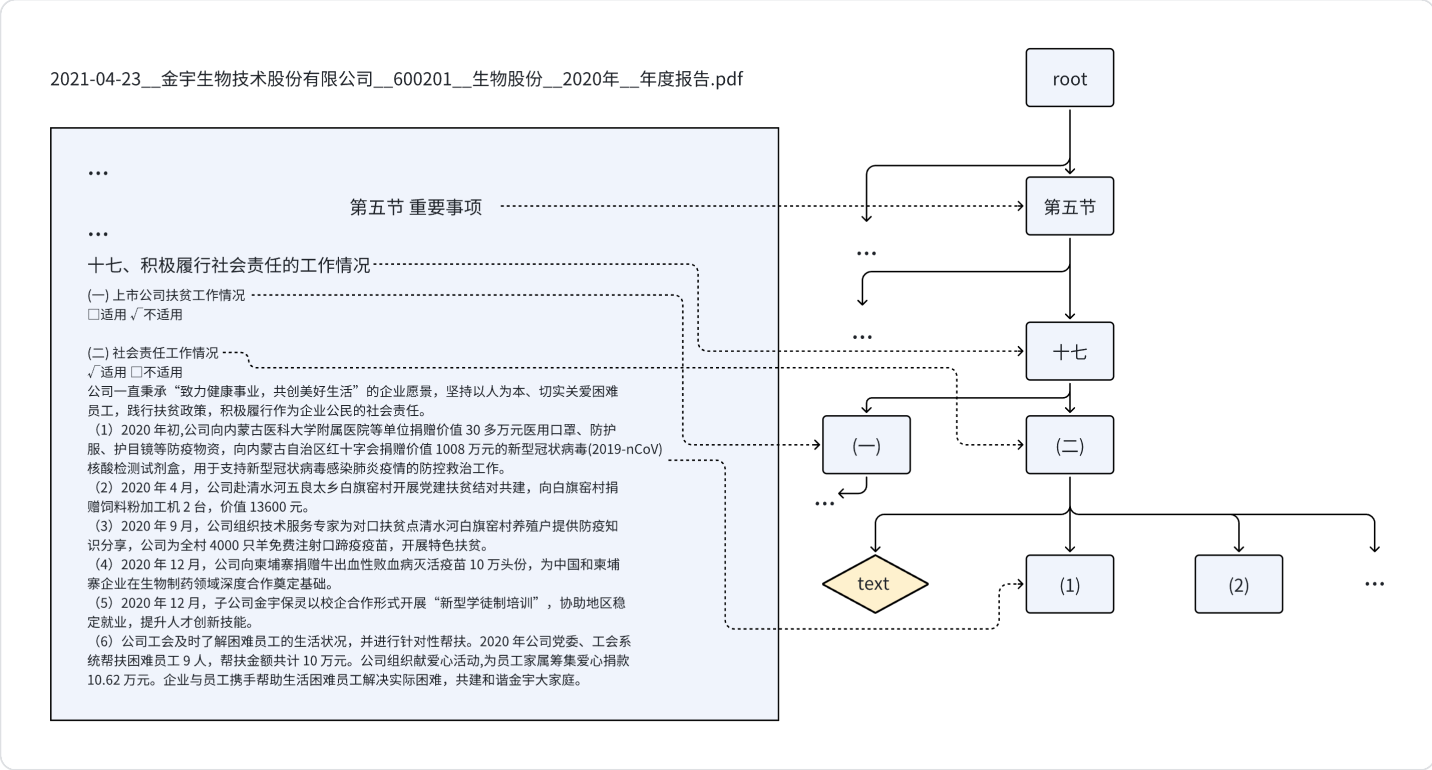
## 三、比赛方案

### 1. 数据结构

DocTree

采用 DocTree 来作为第一层数据结构主要有两个理由：

- 年报的格式具有较高的一致性，我们需要答案内容，如三表数据，基本信息等，都在特定章节的特定大标题之下，使用树的结构进行存储可以有效剪枝搜索范围，快速搜索到想要的答案。
- type31 类问题需要找到文档中的特定段落，通过建树可以对树的中间节点（即各级标题）进行高效针对的搜索，找到想要的段落。此外搜索标题相比于搜索正文更加适配于向量化搜索，避免无关内容挤占召回位置。




## Database

由于type1、type2类型问题对于答案的精确性需求，以及type12类型问题的统计需求，将考察范围（即基础信息，人员信息，财务三表信息）通过DocTree使用正则搜索的方式导入数据库。数据列主要来自于基础信息表，人员信息表，财务三表的相关字段。

## 2. 微调模型

### router



router的主要作用是分发问题。使用router主要的理由为：几大问题分类的解决方式不同，需要的输入也不一样。有效的分发问题能规范化系统的输入，得到理想的答案，也能分治问题进行迭代，并剪枝不必要的多余输入和输出。

## 训练数据

- 在比赛的早期，本方案使用关键词的形式来分析query的意图，可以捕捉到query中的年份、公司名称/公司简称以及年报关键词实体，可以通过对这三类实体的有无来对query的类型进行初步划分。

年份	公司名称	年报关键词	类型分析
	√	√	type11或者type2。根据年报关键词是否为公式可以区分type1和type2
√	√	×	type31，小概率为type1或者type2（关键词不在词表内）
√	×	√	type12，小概率数据库范围外
×	√	√	type12
√	×	×	type12或者type31
×	×	√	type32
×	√	×	type12，关键词不在词表内
×	×	×	type32

- 通过上述分析做出第一版的分类数据，然后人工审查容易弄混的部分得到第二版的分类数据训练。第二版的数据里同时使用初赛数据对替换了大量的关键词以及公式名称以增强模型的鲁棒性。

### 效果

- 在B榜问题集上应该已经过拟合了，B榜2k条相对于label有2条估错

### nl2sql



nl2sql模型的设计主要用来解决两个问题。

- 统计类型问题能够由SQL非常好的解决。
- 找出问题中的不规范字段并映射到数据库中包含的规范字段。

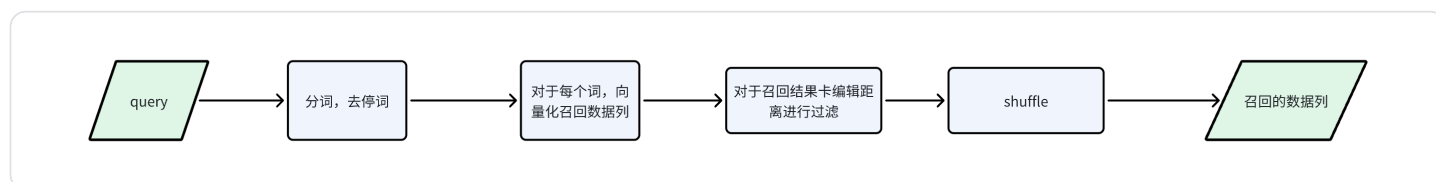
### 训练数据

- type11、type2类型的数据，通过解析问题的实体，可以机器自动生成对应的SQL

- type11-2, type12类型的数据, 通过chatgpt生成少量样本 (大概50~100左右) 生成第一版模型, 然后人工审查生成结果, 重训第二版。
- 为了更好的解决不规范字段映射到规范字段的问题, 训练数据中将部分数据的规范字段通过alias词表替换为了不规范字段。
- 训练数据加入了部分非原来的题目中包含但是存在于数据列中的规范字段替换原本的字段来作为增强, 提高模型的视野。
- 曾经尝试使用过开源的nl2sql数据集, 但是迁移效果不够好, 主要原因是因为赛题pattern过于严重, 因此可能过拟合反而是比较好的得分方式。

## 数据列的召回

nl2sql模型最大的难点在于数据列的召回。因为模型需要把问题中可能不存在于数据库的实体名称映射到存在的数据列上, 需要在prompt里包含该数据列信息。



## 其他tricks

nl2sql模型在早期训练的并不好的一个原因是捕捉query中的公司名称/简称的准确率只有并不是100%。nl2sql模型时常将query中的其他字符串识别为公司名称, 这是训练数据不足的一个表现。本方案中, 由人工从全集中检索出公司名称或者公司简称替换为“公司名称为XXX的公司”, 直接免去了模型的公司实体识别任务, 准确率大大增加。

## 效果

- type11基本cover, 从来没有扫到相关badcase
- type12, 整体效果也很不错, 基本上可以生成正确的SQL, 部分的错误可能来自于训练数据中的错误
- type11-2, 基本都是对的

## normalize



normalize的作用是使得answer能够遵循query的形式, 并保持其中的关键词, 这样能够在关键词以及相似度标准上更多拿分。在具体的用户体验里, 确认自己的所有意图都被理解并执行也是相当重要的事情, 这也是赛题标准的制定缘由之一。

## 训练数据

- 大部分的数据由解析得出的关键词信息以及查询答案进行规则构建。
- 部分的数据由chatgpt完成。
- 训练第一版后，由人工校验，纠正后再训一版。

## 效果

- 大部分（85%能表现得很好如预期）
- 少部分会丢关键词或者答案。检查训练数据后发现是由于训练数据质量导致，个人的精力所限训练数据的质量不足。这也是后面分流部分type1不走normalize而走rule2ans得原因。

## 四、模型效率

- 测试环境
  - 1张3090 python调用
- 测试数据
  - B榜中200条query
- 效率

问题类型	router	nl2sql	normalize	no-tune	total
type1	0.16s/it	2.85s/it	4.51s/it	\	7.52s/it
type2					
type31		\		14.90s/it	15.06s/it
type32					

## 五、数据流程示例

query 哪家上市公司，在北京注册，2020年总负债最高？金额为？

1. router
  - prompt

1 哪家上市公司，在北京注册，2020年总负债最高？金额为？

◦ response

1 type12

## 2. nl2sql

◦ prompt

1 你的任务是将问题转化为SQL。  
2 1. SQL语句查询的表名为：big  
3 2. 涉及到的列名有：注册地址,负债合计,预计负债,股票简称,公司名称,股票代码,年份  
4  
5 【问题】哪家上市公司，在北京注册，2020年总负债最高？金额为？  
6 【SQL】

◦ response

1 select 公司名称, 负债合计 from big  
2 where 注册地址 like '%北京%' and 年份 = '2020年'  
3 order by 负债合计 desc limit 1

## 3. normalize

◦ prompt

1 根据查询结果回答问题。  
2 【查询结果】  
3 {'公司名称': ['华能国际电力股份有限公司'], '负债合计': ['296730398479']}  
4 【问题】哪家上市公司，在北京注册，2020年总负债最高？金额为？  
5 【回答】

◦ response

1 华能国际电力股份有限公司在2020年总负债最高，金额为296730398479元。



