

# 文档问答投研 Agent 系统方案报告

## 一、整体方案介绍

本章将从模型选型和方案的整体流程两个部分介绍系统方案的设计。

### 1.1 模型选型

模型选型决定了系统各个模块的基础能力，而需要选择和考虑哪些种类的模型，决定了模块设计的流程。比赛建议部署本地开源大模型作为问答基础模型，故需要在有限的计算资源范围内尽可能选择能力足够强的模型，以实现更好的组合效果、降低误差的累计。通常，在 RAG 系统的设计上，除了 LLM 模型外还会涉及到计算稠密特征的 Embedding 双塔模型，用以对文段进行特征向量的表示。而囿于本方案涉及到多路召回，故需要有一个高效的召回排序融合的策略实现对多路结果的融合，以经验来看，单塔的 Reranker 模型具有良好的效果。所以，模型选型主要涉及到 LLM、Embedding 模型和 Reranker 模型。

#### 1.1.1 大语言模型

LLM 的选择对于建议使用开源模型本地部署的要求来说并不算多，考虑到比赛会涉及到 Agent 的构建，故要优先选择具备 Function Call 或 ReAct 能力的模型。我的实验室资源为两张 A6000，共 48 G \* 2 显存，所以目光首先放在了 13、14B 的模型之中。个人测试了前不久 Meta 开源的 Llama3-8B-Instruct 模型，虽然相比上一代词表扩充到很大的数量，但是对于中文 Byte 的 token 和语料都训练不足。不过我在尝试使用长度较长的（2K 左右）RAG prompt 进行问答时，Llama3 的输出中文较为稳定，对召回的文段的总结 and 发现也效果不错。但是考虑到比赛会涉及到一些较短 prompt 和效率要求较高的模块，在短输入的测试下，Llama3 的中文输出指令并不稳定，故暂时放弃了选择 Llama3 的想法。最后综合考虑的是今年年前开源的 Qwen1.5-14B-Chat 版本，因为它具备同参数量下较强的通用能力之外，还支持 ReAct 调用，除此之外 1.5 版本的滑动窗口机制对推理的加速效果有很大的提升。最终，囿于服务器没有 root 权限、以及英伟达驱动版本过低的问题，舍弃了 vLLM 和 TGI 推理加速框架，选择了 ModelScope 的 SWIFT 框架进行本地的部署，并开放兼容 OpenAI 的 API 接口方便随时调用。不过虽然显存资源勉强够用，但是后期的测试集推理还是遇到了显存溢出的问题，最大长度 prompt 和模型本地占用了接近 48G 的显存，故后期对 RAG 的 prompt 进行了长度上的优化。

#### 1.1.2 嵌入模型

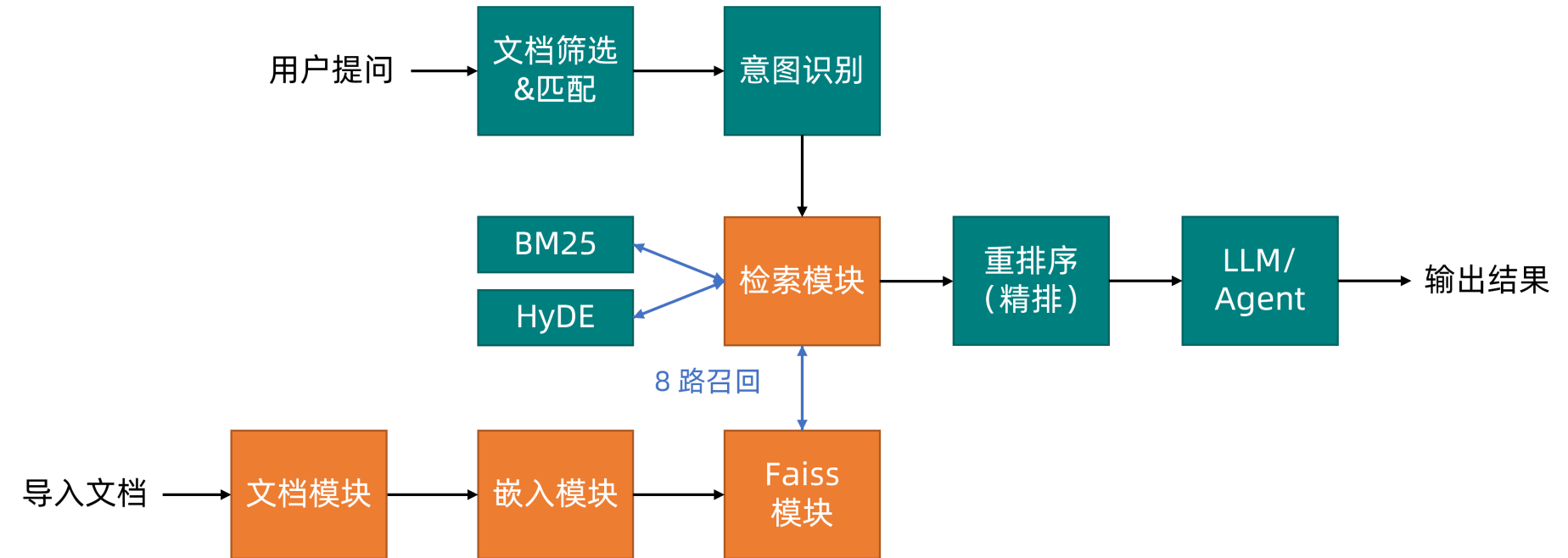
双塔的嵌入模型的选择对于 RAG 系统来说至关重要，双塔模型是牺牲了精度换取速度的典型方案，所以在我们要求高召回率的条件下，嵌入模型的精度至关重要。个人测试了之前常用的 BGE-large-zh、BGE-large-zh-v1.5 和 BGE-m3 模型，发现 BGE-large-zh 和 v1.5 版本能力相当，但是后者在 [0.6, 1] 的相似度分布区间上进行了优化，分布更加合理；而 BGE-m3 模型虽然支持 8K 的文段输入，且支持三种文段特征（稠密特征、稀疏特征、交互稠密特征），但是特点在于多语言上，中文的表现不如前两者。所以，虽然很可惜 BGE-large-zh-v1.5 只支持 512 的长度，但是目前的最优选择还是它了。embedding 模型的部署、负采样和微调均使用官方的 FlagEmbedding 框架，官方的实现非常便捷，推理速度也基本在 1400 itmes/s 左右，能够满足比赛的 96 个文档的需求。模型参数大概 0.7 B，占用显存 1.5 G。

#### 1.1.3 重排序模型

Reranker 重排序模型对于多路召回的融合至关重要，单塔的排序模型是以牺牲性能的代价提高精度，所以同样需要能力较强的模型。而归功于 Reranker 模型优秀的语义建模，其可以在很多模块发挥小模型的分类作用，且泛化性能很强。目前主流的 Reranker 模型基本就是智源的 BGE-reranker-large，所以按照经验也直接选择了该模型。同样地，Reranker 模型的部署、负采样和微调也使用了官方的 FlagEmbedding 框架，推理速度相比嵌入模型稍慢一些，基本在 120 itmes/s 左右。模型参数大概也是 0.7 B，占用显存 1.5 G。

### 1.2 整体流程

系统的整体流程图如下所示：



如图所示，蓝绿色流程为普通的可插拔模块，方便进行二次开发；而橙色流程为系统核心模块，提供基础的 RAG 功能。整个 RAG 系统的构建基于个人之前自己搭建的简易 RAG 框架，实现了多路召回、PDF 和 Excel 文档解析、NL2Pandas 查询等特点。基于比赛的数据特点和应用场景，重点优化了文档模块、嵌入模块、文档筛选和匹配、意图识别和 LLM/Agent 模块。下面小结分别从文档解析流程（流程图下面分支），和用户提问流程（流程图上面分支），以及公共流程（流程图后半部分）三个角度介绍模块的流程。

### 1.2.1 文档解析流程

文档解析流程如图所示，需要依次经过文档模块、嵌入模块、Faiss 模块，最后到达检索模块与用户提问汇合。

具体来说，文档解析流程在比赛中是对官方提供的 96 个 PDF 文档进行解析、处理、存储和使用的过程。我们需要将每一个 PDF 文档分别读取，方便建立每个文档的处理过程存储，以灵活地组合或者筛选其中一个或几个文档。

对于处理后的文档需要进行分段做嵌入，以便于文档的信息能分块被索引到。这里使用双塔模型做嵌入，故文段和用户提问的 embedding 分布一致，方便自行构建指标对其进行对齐和召回。而文档的文段需要为用户的问题所复用，故我们将文档处理后的数据放入数据库进行存储不仅方便召回、更便于快速调用和召回文段。为了增加召回的稳健性，我们还使用了 TF-IDF 的优化算法 BM25 作为文段特征之一参与检索。我们需要把文段和用户提问进行 BM25 得分计算，取得分最高的 topk 个文段进行召回。

最后，文档解析流程进入检索模块，赋予检索模块 embedding 生成和向量数据库检索的服务。至此，文档解析的部分就结束了，理论上本系统只需要开启后首次上传 96 个文档后，就无须进行重复解析，后续只需要用户提问即可。

文档解析流程在代码 chat\_page.py 的第 258-366 行。如下图所示为文档解析流程各个模块的代码：

```
1 # 若上传的文件存在
2 if uploaded_files is not None:
3     file_names = [".".join(uploaded_file.name.split(".")[::-1]) + "." + uploaded_file.name.split(".")[::-1].lower()
4         for uploaded_file in uploaded_files]
5
6     # 拟保存的文件的完整路径
7     file_paths = [os.path.join(DOCUMENT_TMP_PATH, file_name) for file_name in file_names]
8     st.session_state.file_list += file_paths
9     file_types = [postfix_type.get(file_name.split(".")[::-1]) for file_name in file_names]
10
11     # 保存文件
12     ...
13     # 分句
14     ...
15     # 嵌入
16     ...
17     # 存入 Faiss
18     ...
19     # 计算 BM25
20     with st.spinner('正在计算 BM25 ...'):
21         for sentences_long, sentences_short in zip(st.session_state.sentences_long,
22 st.session_state.sentences_short):
23             if "bm25_long" not in st.session_state:
24                 st.session_state.bm25_long = [BM25(sentences_long)]
25             else:
26                 st.session_state.bm25_long.append(BM25(sentences_long))
27             if "bm25_short" not in st.session_state:
28                 st.session_state.bm25_short = [BM25(sentences_short)]
29             else:
30                 st.session_state.bm25_short.append(BM25(sentences_short))
31         placeholder.success('BM25 计算完成!', icon="✅")
32         my_bar.progress(1.00, "保存文档✅ --> 文本分段✅ --> 段落 Embeddings✅ --> 存入 Faiss✅ --> 计算 BM25✅")
33
34     placeholder.success(f"{len(st.session_state.document[file_types[0]]['long'])} 个文档处理完成!",
35 icon="✅")
```

### 1.2.2 用户提问流程

用户提问流程依赖于文档解析流程提供的 embedding 服务和检索服务。如图所示，用户输入问题后，会首先针对提问筛选和匹配所涉及到的文档数量和具体是哪些文档。

文档筛选就是简单的使用 prompt 工程，让大模型判断用户提问所涉及到的文档数量。例如只提及“第一季度”则判定为只有一篇文档；若提及到“中期和年度报告”，则判定为 2 篇文档；若提及到“四个季度”，则认为是从第一季度到第四季度的 4 篇文档。用以进一步降低检索范围，确定召回通路数量。而文档匹配则是将具体所需的 n 个文档，通过 96 分类的形式进行匹配，取其前 n 名。这里我们使用 reranker 模型作为分类模型，测试发现无需微调，reranker 模型也具备较好的泛化性能，能够准确的匹配到用户所提及的文档。

意图识别环节主要用于判断用户询问的问题类别，以提供不同的 prompt 模型响应和后处理服务。意图识别为三分类：金融问答计算、json 信息抽取和统计图表绘制，分别对应比赛题目的要求一二、要求三和要求四。信息抽取意图和图表绘制意图就要求模型输出的一些特殊格式化和 Agent 构建，以完成复杂的任务。

BM25 则是用户基于所选择的文档的计算得分，取其得分前 topk 的文段进行多路召回。

HyDE（Hypothetical Document Embeddings）（论文）的改写召回思路非常直观，假设为“长度相当的文段做双塔 embedding 相似度后效果更好”这一前提，作者基于这一前提将用户 query 输入给大模型，生成一段对 query 的虚拟回答 dummy\_query，再将这个 dummy\_query 代替原 query 进行 embedding 相似召回，达到提升召回效果的目的。同样地，也可以对文档解析后做 chunk → query(dummy chunk) 改写等等。

我们采用长/短文本、原 query/HyDE、dense/sparse 召回策略，进行排列组合，实现了针对同一个文档的 8 路召回方案。但是在实测中发现，HyDE 的生成效率很低，但对召回排序的效果提升不显著，故后续为了增加召回效率，关闭 HyDE 只使用 4 路召回。

Route 的实现代码在 Retriever.py，如下：

```
1 class Route:
2     def __init__(self,
3         topk: int = 3,
4         embedding: Embedding = None,
```

```

5         db: VectorDatabase = None,
6         bm25: BM25 = None,
7         hyde: HyDE = None
8     ):
9         self.topk = topk
10        self.embedding = embedding
11        self.bm25 = bm25
12        self.db = db
13        self.hyde = hyde
14        assert bool(self.embedding) ^ bool(self.bm25), "You must select one of dense or bm25."
15        assert (self.db and self.embedding) or self.embedding == None, "You must provide a database
and a embedding."
16
17        def __call__(self, query: str) -> List[int]:
18            logger.info(f"Route running ... HyDE: {bool(self.hyde)}, Embedding: {'dense' if
bool(self.embedding) else 'bm25'}, DB: {self.db.db_name if self.db else None}({self.db.amount if
self.db else None})")
19            # 1. HyDE or not
20            ...
21            # 2. embedding or bm25
22            ...
23            # 3. bm25: long or short
24            ...

```

用户提问流程的代码实现主要在 `chat_page.py` 的第 370-416 行。如下所示：

```

1  # 若用户提问
2  if user_query := st.chat_input("您好，请问有什么可以帮到您? "):
3      # 文档选择
4      ...
5      # 匹配文档
6      ...
7      # 意图识别 0-金融问答计算，1-json信息抽取，2-统计图表绘制
8      ...
9      # 配置多路召回
10     routes = []
11     for i in file_id:
12         routes += [
13             Route(8, st.session_state.embedding, st.session_state.db_long[i], None, None),
14             Route(8, None, None, st.session_state.bm25_long[i], None),
15             Route(8, st.session_state.embedding, st.session_state.db_short[i], None, None),
16             Route(8, None, None, st.session_state.bm25_short[i], None),
17             # Route(3, st.session_state.embedding, st.session_state.db_long, None,
st.session_state.hyde["concise"]),
18             # Route(3, None, None, st.session_state.bm25_long, st.session_state.hyde["glm"]),
19             # Route(5, st.session_state.embedding, st.session_state.db_short, None,
st.session_state.hyde["step"]),
20             # Route(5, None, None, st.session_state.bm25_short, st.session_state.hyde["qwen"])
21         ]
22     retriever = Retriever(routes, st.session_state.rerank)

```

### 1.2.3 公共流程

公共流程包括从检索模块到最后的处理，是从定义好后的多路召回开始执行后的操作。

多路召回以多线程的方式并行执行，每个线程对应一个 **Route**，最后收集 **Route** 排序后的多个结果，并对结果进行融合。这里我个人设计的融合方法有两种，①基于 **Reranker** 的单塔融合：将多路召回的结果全部作为候选项，再重新排序作为最终的结果；②基于自定义权重的倒排序：用户根据设定的 **Route** 的权重，对每个排序排名的倒数做加权，最终对所有排序做最终的权重排序。方法一适合通用情形下的高精度细排，方法二适合自定义业务的针对性控制排序。我们选择方法一来融合多路召回的结果。

除此之外，有论文提出 *Lost in the middle*（[论文](#)）的注意力分布现象，即关键信息出现在开头和结尾更容易检索出来，因此我们最后的 **chunk** 排序将相关度高的放在 **prompt** 的开头和结尾位置，相关度低的放在中间。

检索重排序部分的代码在 `Retriever.py`，**Retriever** 的基类代码如下：

```

1  class Retriever:
2      def __init__(self,
3                  route: Union[Route, List[Route]],
4                  reranker: RerankEmbedding = None,
5                  weights: List[float] = None):
6          self.route = route if isinstance(route, list) else [route]
7          self.n_routes = len(self.route)
8          self.reranker = reranker
9          self.weights = weights
10         assert self.n_routes == 1 ^ bool(self.weights) or self.reranker, \
11             "If routes > 1, you must provide weights. Or if routes = 1, you don't need to provide
weights."

```



```
12         assert bool(self.reranker) ^ bool(self.weights), "You must select one of reranker or weights."
13         assert self.weights is None or (len(self.weights) == self.n_routes), "The length of weights
must be equal to the number of routes."
14
15     def __call__(self,
16                 query: str,
17                 sent_list_list: List[List[str]],
18                 topk: int = 3,
19                 lost_in_the_middle: bool = True,
20                 return_lists: bool = False
21                 ) -> Union[List[str], List[int]]:
22         assert len(sent_list_list) == self.n_routes, "The length of `sent_list_list` must be equal to
the number of routes."
23
24         # 1. run multi-routes
25         ...
26         # 2. merge results
27         ...
28         # 3. Lost in the middle
29         ...
```

重排序后，我们将召回的文段（这里设置的是 **top4**）列表和用户的问题拼接好 **prompt**，输入给模型，得到模型预测的答案。针对信息抽取任务，我们还需要对输出做 **json** 格式的格式化处理。而对于统计图表绘制任务，则需要构建一个 **Agent** 实现画图功能。下面章节会介绍该 **Agent** 的实现。

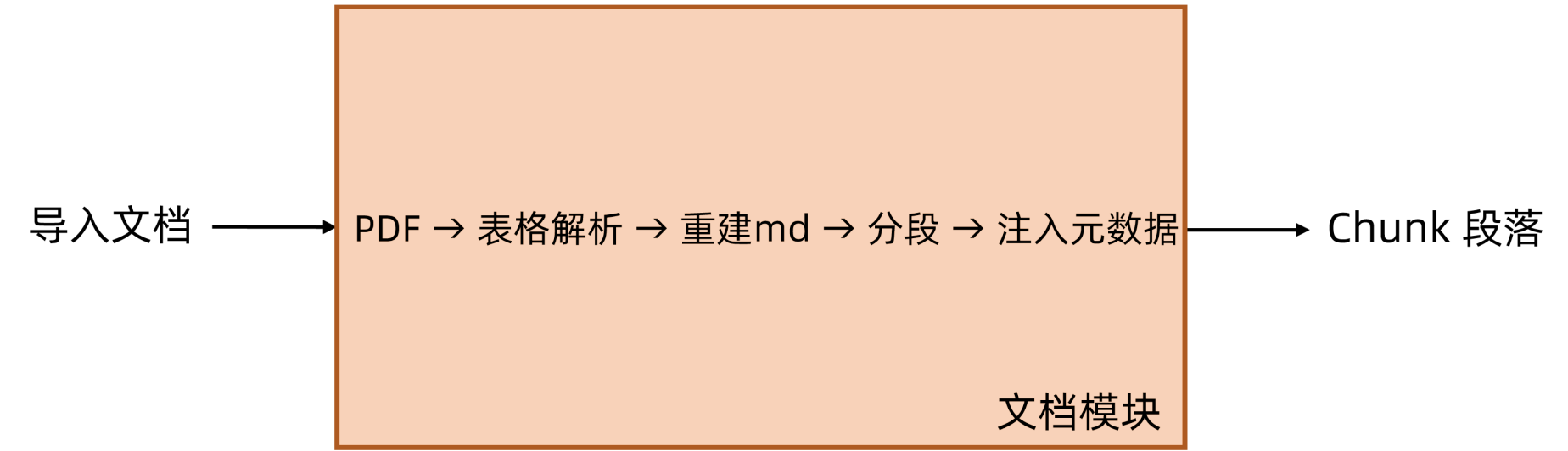
## 二、关键技术介绍

本章节介绍系统在 **RAG** 模块上的优化策略，以及在模型优化上的策略。

### 2.1 RAG 模块优化

**RAG** 的核心模块包括文档模块、嵌入模块、向量数据库模块和检索模块。这几个模块的内部集成了较多的步骤，相比于其他可插拔的小模块，核心模块的质量需要有所保障。而针对不同模块的优化手段非常之多，下面介绍一些我自己这个框架下实现的一些优化策略。

#### 2.1.1 文档模块



文档模块我个人实现了三种文档的解析：文本文件（**txt**, **md** 等）、**PDF** 文件和 **Excel** 文件。框架的 **PDF** 解析在将文档解析成 **plaintext** 后，直接继承文本文件的类，做相似的处理方法。但是针对我们金融文档的表格来说，这种文本表示大模型的识别较差。故最后采用其他比赛中开源的一个 **pdf2txt.py**（[项目](#)）的处理脚本，来按行将 **PDF** 的表格解析出来。而我针对解析出来的 **txt** 进行标题识别、表格提取和去噪的脚本编写，实现了 **txt2md.py**。具体效果如下：

原 PDF

转为 txt

转为 md

3.2 基金净值表现

3.2.1 基金份额净值增长率及其与同期业绩比较基准收益率的比较

阶段	净值增长率①	净值增长率标准差②	业绩比较基准收益率③	业绩比较基准收益率标准差④	①-③	②-④
过去三个月	0.68%	0.01%	1.08%	0.01%	-0.40%	0.00%
过去六个月	1.36%	0.01%	2.16%	0.01%	-0.80%	0.00%
过去一年	2.83%	0.01%	4.40%	0.01%	-1.57%	0.00%
过去三年	8.30%	0.01%	13.80%	0.01%	-5.50%	0.00%
自基金合同生效起至今	9.31%	0.01%	15.52%	0.01%	-6.21%	0.00%

注：(1)上述基金业绩指标不包括持有人认购或交易基金的各项费用，计入费用后实际收益水平要低于所列数字；

(2)本基金业绩比较基准为：三年定期存款利率（税后）+1.5%。

3.2.2 自基金合同生效以来基金累计净值增长率变动及其与同期业绩比较基准收益率变动的比较

[“page”:4,“alIrow”:51,“type”:“text”,“inside”:“3.2基金净值表现”]

[“page”:4,“alIrow”:52,“type”:“text”,“inside”:“3.2.1基金份额净值增长率及其与同期业绩比较基准收益率的比较”]

[“page”:4,“alIrow”:53,“type”:“excel”,“inside”:“阶段：净值增长率①，净值增长率标准差②，业绩比较基准收益率③，比较基准收益率标准差④，①-③，②-④”]

[“page”:4,“alIrow”:54,“type”:“excel”,“inside”:“过去三个月，0.68%，0.01%，1.08%，0.01%，-0.40%，0.00%”]

[“page”:4,“alIrow”:55,“type”:“excel”,“inside”:“过去六个月，1.36%，0.01%，2.16%，0.01%，-0.80%，0.00%”]

[“page”:4,“alIrow”:56,“type”:“excel”,“inside”:“过去一年，2.83%，0.01%，4.40%，0.01%，-1.57%，0.00%”]

[“page”:4,“alIrow”:57,“type”:“excel”,“inside”:“过去三年，8.30%，0.01%，13.80%，0.01%，-5.50%，0.00%”]

[“page”:4,“alIrow”:58,“type”:“excel”,“inside”:“自基金合同生效起至今，9.31%，0.01%，15.52%，0.01%，-6.21%，0.00%”]

[“page”:4,“alIrow”:59,“type”:“text”,“inside”:“”]

[“page”:4,“alIrow”:60,“type”:“text”,“inside”:“注：(1)上述基金业绩指标不包括持有人认购或交易基金的各项费用，计入后实际收益水平要低于所列数字；”]

[“page”:4,“alIrow”:61,“type”:“text”,“inside”:“(2)本基金业绩比较基准为：三年定期存款利率（税后）+1.5%。”]

[“page”:4,“alIrow”:62,“type”:“text”,“inside”:“3.2.2自基金合同生效以来基金累计净值增长率变动及其与同期业绩比较收益率变动的比较”]

[“page”:4,“alIrow”:63,“type”:“text”,“inside”:“”]

[“page”:5,“alIrow”:64,“type”:“text”,“inside”:“第4页，共12页”]

[“page”:5,“alIrow”:65,“type”:“text”,“inside”:“”]

[“page”:5,“alIrow”:66,“type”:“text”,“inside”:“”]

[“page”:5,“alIrow”:67,“type”:“text”,“inside”:“”]

3.2 基金净值表现

3.2.1 基金份额净值增长率及其与同期业绩比较基准收益率的比较

阶段	净值增长率①	净值增长率标准差②	业绩比较基准收益率③	业绩比较基准收益率标准差④	①-③	②-④
过去三个月	0.68%	0.01%	1.08%	0.01%	-0.40%	0.00%
过去六个月	1.36%	0.01%	2.16%	0.01%	-0.80%	0.00%
过去一年	2.83%	0.01%	4.40%	0.01%	-1.57%	0.00%
过去三年	8.30%	0.01%	13.80%	0.01%	-5.50%	0.00%
自基金合同生效起至今	9.31%	0.01%	15.52%	0.01%	-6.21%	0.00%

注：(1)上述基金业绩指标不包括持有人认购或交易基金的各项费用，计入费用后实际收益水平要低于所列数字；

(2)本基金业绩比较基准为：三年定期存款利率（税后）+1.5%。

3.2.2 自基金合同生效以来基金累计净值增长率变动及其与同期业绩比较基准收益率变动的比较

注：(1)本基金合同生效日为2020年2月25日；

(2)本基金的建仓期为合同生效日起6个月；截至建仓期末和本报告期末，本基金的资产配置符合基金合同的相关要求。

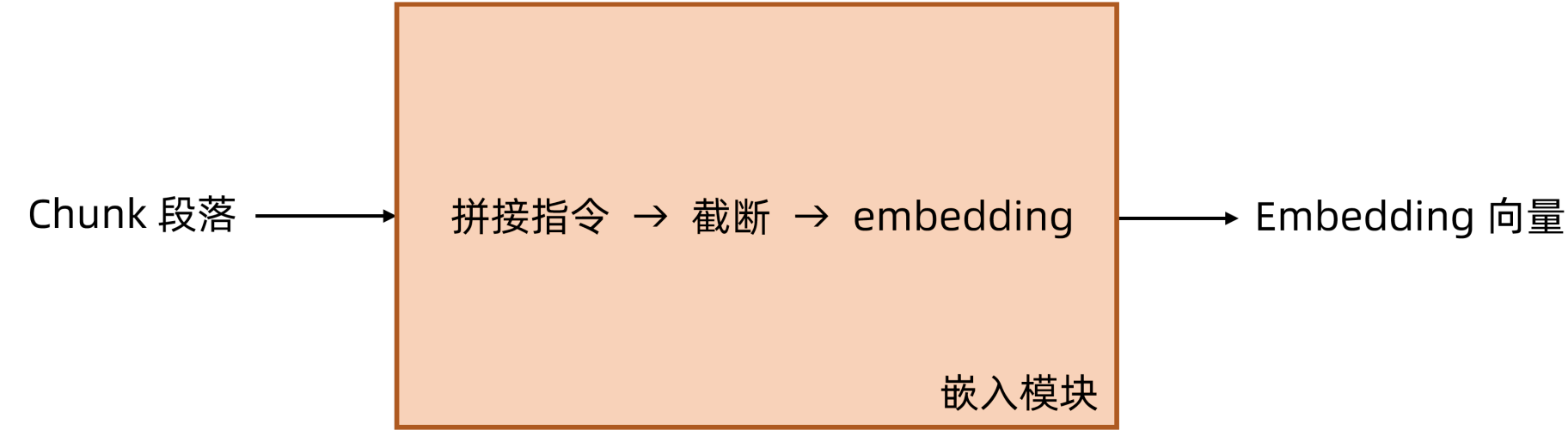
\$4 管理人报告 |

基金管理人 基金管理人 基金管理人

大模型对 **markdown** 格式的表格按行输入的识别比较准确，故文本分段时我们也将整个表格（包括前后两行）作为一个 **chunk** 不截断。保障了阅读表格的连续性和完整性。

具体的文档实现代码见 **document.py**。

2.1.2 嵌入模块



embedding 模型的优化一般是针对垂直领域的数据集进行进行微调。BGE 的模型微调使用的是对比学习的方式，官方也实现了数据集负样本的困难样本挖掘脚本（FlagEmbedding 实现有 BUG，需要拉取最新的 GitHub）。

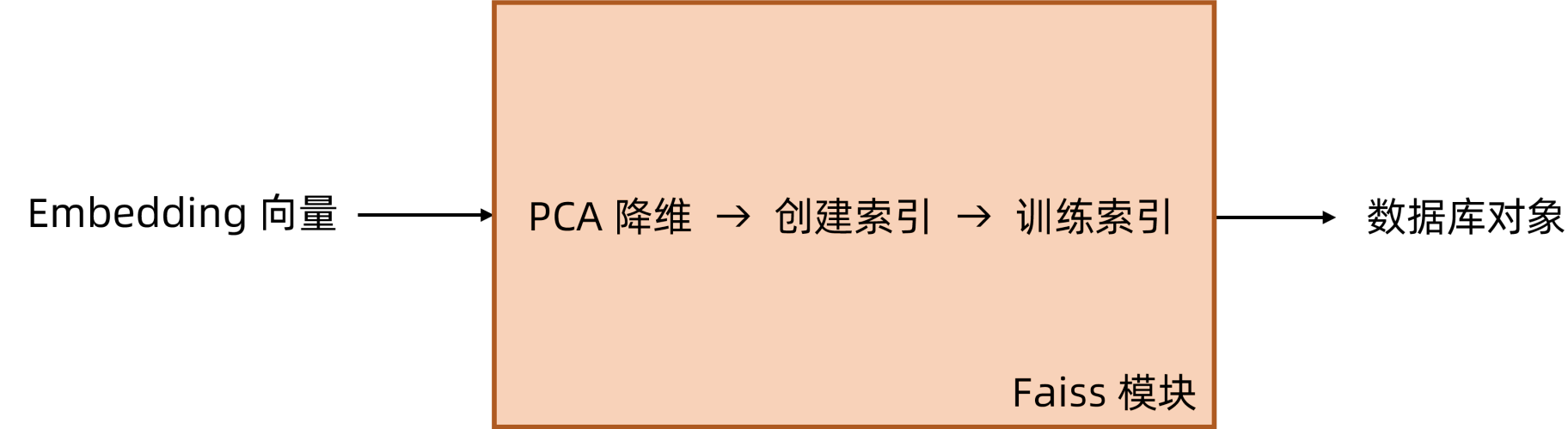
从经验上看，embedding 模型的微调性价比非常高，可以使用少量的训练集提升较为显著的效果。所以我们需要根据 chunk 构造能够微调 embedding 模型的数据集。

代码 `create_dataset.py` 实现了调用智谱 GLM4 和阶跃星辰 Step 模型的 API 实现对 chunk 生成 query 的数据集构建脚本。一共选取了每个文档中的 6 个 chunk，每个 chunk 生成 5 个 query，一共生成了近 4000 条的微调数据集。最终通过 9:1 的方式划分训练集和测试集，进行微调。

微调脚本在 `finetune` 文件夹下，半个小时微调 3 个 epoch 即可，测试的 top1 准确率从原模型的 0.79，上升到微调后模型的 0.91，效果可观。

为了保障 reranker 模型的通用能力能够在我设计的分类模块中继续保持，个人没有对 reranker 模型经行微调。当然针对像文档名匹配、意图识别等任务也可以单独微调实现更稳定的分类效果。

2.1.3 向量数据库模块

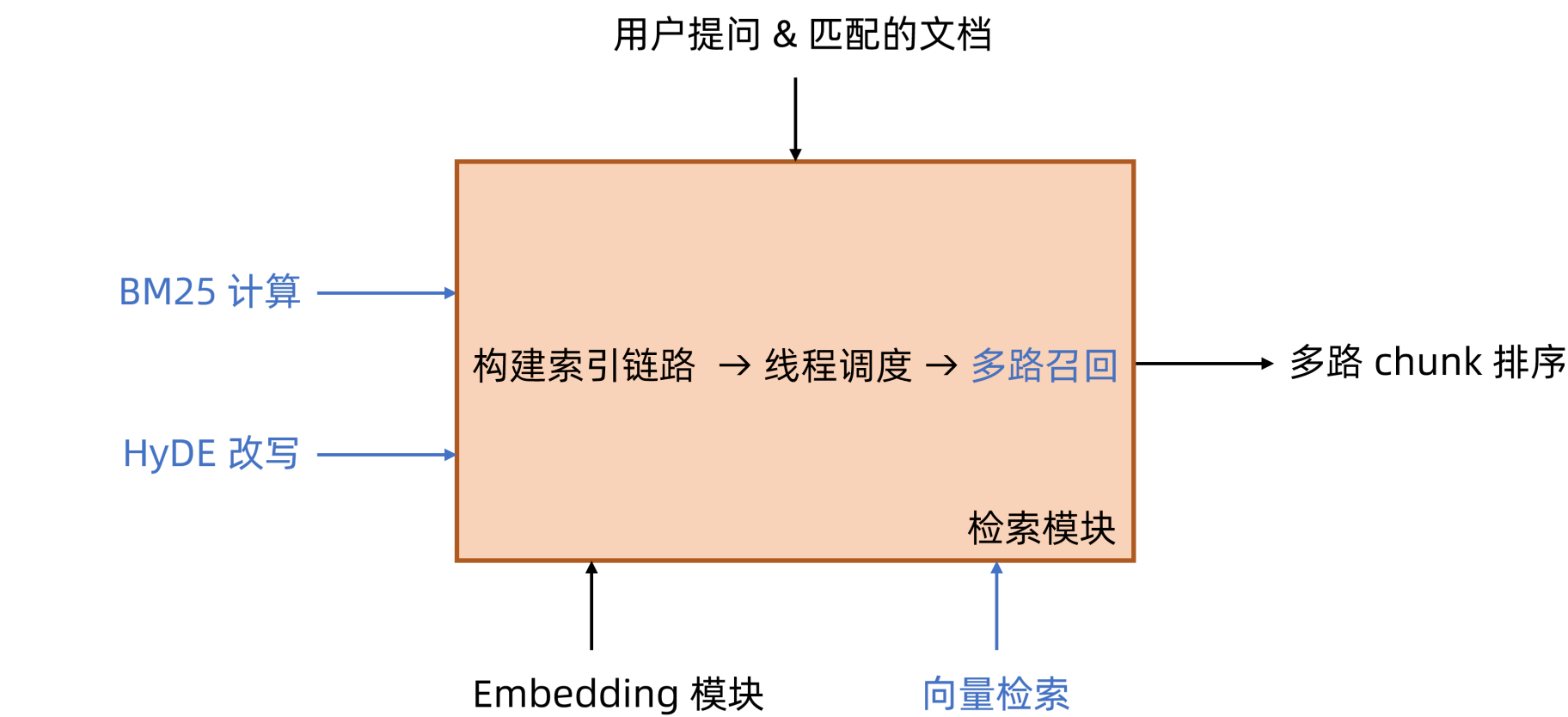


向量数据库的选择也比较有限，主流的 Faiss、ElasticSearch 和 Milvus 的部署很多，但是考虑到比赛时间较短，我们选择较为轻量化的 Faiss 进行向量数据库的构建。不过考虑到我设置的多路召回方案，使用 ES 数据库可能更为合适，可以在下一次优化中进行尝试。

Faiss 数据库我是运行在 GPU 上的，加速创建索引的速度。具体采用的索引构建算法是 HNSW64，指标为内积。优点是能够保证召回精度损失较小的同时召回速度比较快，缺点是训练索引的时候速度较慢。而测试的其它索引算法，如 IVF256 需要提前创建索引，对于灵活的向量 IO 操作来说，比较麻烦。

在多文档任务中，我将每个文档的长文本（长度 768）和短文本（长度 384）都分别构建单独的 Faiss 对象。所以对于 96 个文档的初始化步骤中，一共创建了  $96 * 2 = 192$  个 Faiss 对象。具体的实现代码详见 `VectorDatabase.py`。

2.1.4 检索模块



检索部分的设计在 1.2.2 用户提问流程和 1.2.3 公共流程 中讲的较为详细，这部分就不再赘述了。详细代码可以查阅 `Retriever.py`。

## 2.2 模型优化

模型部分的优化主要是针对模型自身能力的优化，即微调；以及对模型工作流的优化，即工具调用和 Agent 的构建。

### 2.2.1 模型微调

大模型的微调能够更好地提升特定任务对格式输出的指令跟随能力，但是不适用于知识的注入，否则会以大模型通用能力较大损失的代价呈现。在测试大模型对召回文段的总结和分析过程中，我发现针对 **去除千分位分隔符** 和 **保留小数点后两位** 这两个指令需求，大模型遵守的并不好。所以打算针对这两个任务进行微调，使得模型具备遵守这两个指令的能力。

数据集的构建很简单，使用 step 模型的 api 生成带有指令的 **有/无千分位分隔符的文本对** 和 **有/无小数点后保留两位的文本对**，生成数量为 200:200。为了避免简单任务微调对模型能力的较大损失，我使用了 100 余条强逻辑的 弱智吧数据集（**huggingface**）进行混合训练。最终使用 500 余条数据集的随机打乱进行微调，使用 Llama-Factory 框架进行 AdaLoRA 微调，占用 40 G 显存，训练 3 个 epoch，最终实现了格式上的指令跟随能力。

### 2.2.2 模型 Agent

模型的 Agent 构建环节主要在统计图表绘制的场景下，根据用户问题生成相应的统计图。

具体来说，画图 Agent 分为两步：①常规的 RAG 文段和提问的信息总结；②将上一步总结的信息与用户提问拼接，要求其根据上下文生成 python 的 **matplotlib** 的图表代码。最后我们解析和格式化代码，将保存图片的语句注入代码，然后执行画图部分代码，并将图片路径引用到模型回复中，用于 **streamlit** 的展示。若画图代码执行出现问题，则会尝试一次重试：将报错信息与原 **matplotlib** 生成的 prompt 进行拼接，要求其重新修改代码。通过构建这样一个简单的 Agent 则可以实现统计图表的绘制任务。

这部分 Agent 代码在 **Retriever.py** 的 **ChartRetriever** 类中，代码如下：

```
1 class ChartRetriever(APIModel):
2     def __init__(self, api_name: str = "qwen1half-14b-chat", **kwargs):
3         super().__init__(IMAGE_GEN, api_name, **kwargs)
4
5     def __call__(self, query: str, desc: str) -> str:
6         codes = super().__call__(query=query, desc=desc)
7         codes = re.findall(r'```.*?```', codes, re.DOTALL)
8         codes = "" if codes == None else codes[0]
9         codes = codes.replace("plt.show()", "" + "plt.rcParams['font.sans-serif'] = ['SimHei']
10 plt.rcParams['axes.unicode_minus'] = False
11 plt.savefig('./tmp/figure.png')""").replace("```python\n", "").replace("\n```", "")
12         logger.info(f"the chart code: {codes}")
13         try:
14             exec(codes)
15             desc += " \n![img](./tmp/figure.png)"
16         except Exception as e:
17             logger.error(f"生成图片失败! {str(e)}")
18             try:
19                 codes = super().__call__(query=query, desc=desc + f"请根据报错信息{str(e)}, 修改以下代码:
\n\n{codes}")
20                 codes = re.findall(r'```.*?```', codes, re.DOTALL)
21                 codes = "" if codes == None else codes[0]
22                 codes = codes.replace("plt.show()", "" + "plt.rcParams['font.sans-serif'] = ['SimHei']
23 plt.rcParams['axes.unicode_minus'] = False
24 plt.savefig('./tmp/figure.png')""").replace("```python\n", "").replace("\n```", "")
25                 logger.info(f"the retry chart code: {codes}")
26                 exec(codes)
27                 desc += " \n![img](./tmp/figure.png)"
28             except Exception as e:
29                 desc += f" \n*已尝试两次生成图片均失败! {str(e)}*"
30         return desc
```

## 三、Prompt 工程介绍

系统的大部分功能均依赖 Prompt 完成，所以在 Prompt 的构建上需要注重。

项目的所有 Prompt 模版均保存在 **template.py** 文件中。

### 3.1 文段的元信息模版

针对需要分段的 chunk，我设计了能够具备元信息（来源文档、该段所属的段落标题）的模版，以辅助多文档问答时对 chunk 的定位和理解。模版如下表所示：

原文	模版	Prompt
报告期期末持有的本基金份额占基金总份额比例（%） <b>02 2</b> 基金管理人运用固有资金投资本基金交易明细 本报告期基金管理人未运用固有资金投资本基金。 <b>§8</b> 影响投资者决策的其他重要信息 <b>1</b> 报告期内单一投资者持有基金份额比例达到或超过 <b>20%</b> 的情况 本基金本报告期内不存在单一投资者持有基金份额比例达到或超过 <b>20%</b> 的情况。 <b>§9</b> 备查文件目录 <b>1</b> 备查文件目录 一、《东方新能源汽车主题混合型证券投资基金基金合同》 二、《东方新能源汽车主题混合型证券投资基金托管协议》 三、东方基金管理股份有限公司批准成立批件、营业执照、公司章程 四、本报告期内公开披露的基金资产净值、基金份额净值及其临时公告 <b>2</b> 存放地点 上述备查文本存放在本基金管理人办公场所。 <b>3</b> 查阅方式 投资者可免费查阅，在支付工本费后，可在合理时间内取得上述文件的复制件或复印件。	<b>{document_name}</b> 文档里，目录层级为： <b>{contents_title}</b> ， 以下为正文内容： <b>{context}</b>	《东方新能源汽车主题混合型证券投资基金 <b>2023</b> 年第 <b>3</b> 季度报告.md》文档里，目录层级为： <b>7.2</b> 基金管理人运用固有资金投资本基金交易明细 <b>§8</b> 影响投资者决策的其他重要信息 <b>8.1</b> 报告期内单一投资者持有基金份额比例达到或超过 <b>20%</b> 的情况 <b>§9</b> 备查文件目录 <b>9.1</b> 备查文件目录 <b>9.2</b> 存放地点 <b>9.3</b> 查阅方式， 以下为正文内容： 报告期期末持有的本基金份额占基金总份额比例（%） <b>02 2</b> 基金管理人运用固有资金投资本基金交易明细 本报告期基金管理人未运用固有资金投资本基金。 <b>§8</b> 影响投资者决策的其他重要信息 <b>1</b> 报告期内单一投资者持有基金份额比例达到或超过 <b>20%</b> 的情况 本基金本报告期内不存在单一投资者持有基金份额比例达到或超过 <b>20%</b> 的情况。 <b>§9</b> 备查文件目录 <b>1</b> 备查文件目录 一、《东方新能源汽车主题混合型证券投资基金基金合同》 二、《东方新能源汽车主题混合型证券投资基金托管协议》 三、东方基金管理股份有限公司批准成立批件、营业执照、公司章程 四、本报告期内公开披露的基金资产净值、基金份额净值及其临时公告 <b>2</b> 存放地点 上述备查文本存放在本基金管理人办公场所。 <b>3</b> 查阅方式 投资者可免费查阅，在支付工本费后，可在合理时间内取得上述文件的复制件或复印件。

### 3.2 HyDE 模版

通过模版对 **query** 进行虚拟文档的生成。

query	模版	Prompt
东方新能源汽车主题混合在 <b>2023</b> 年第三季度报告期末的基金净值增长率是多少？ （以百分数表示，保留小数点后 <b>2</b> 位）	请编写一小段严谨、简洁的中文段落来回答以下问题，不得回答不知道、不得要求用户求助他人或拒绝回答，不要超过 <b>500</b> 字：\n问题：{query}\n段落：	请编写一小段严谨、简洁的中文段落来回答以下问题，不得回答不知道、不得要求用户求助他人或拒绝回答，不要超过 <b>500</b> 字：\n问题：东方新能源汽车主题混合在 <b>2023</b> 年第三季度报告期末的基金净值增长率是多少？（以百分数表示，保留小数点后 <b>2</b> 位）\n段落：

### 3.3 知识问答计算 **RAG** 模版

针对任务一和任务二的 **RAG** 模版，如下：



问题id	问题	Prompt	答案
0	财通裕惠63个月定期债券在2023年第三季度报告期末的债券投资占基金总资产的比例是多少？（以百分数表示，保留小数点后2位）	<p>你是一个知识丰富的学者，擅长根据搜索结果，回复用户的问题。你的性格也很温和，回复都很礼貌。</p> <p>对于用户的每一个问题，基于这些搜索结果，尽量引用搜索结果中原文回复，信息要全面，避免任何主观臆断或虚构，且遵循以下回答要求：</p> <p>1. 回复要全面，不漏信息，语言流畅，逻辑清晰。</p> <p>2. 当问题比较模糊，而检索结果包含多种可能的答案时，反向提问用户想问的具体内容；</p> <p>3. 可以对答案中的核心部分进行markdown加粗（加粗内容）；</p> <p>4. 当提问需要进一步计算或者调用外部工具进行辅助时，可以调用外部工具进行辅助回答；</p> <p>5. 使用引用标记来标注回答内容参考的搜索结果序号，例如<sup>[1]</sup>（引用单个搜索结果），<sup>[1][2]</sup>（引用多个搜索结果），其中方括号中的数字是搜索结果序号。引用标记只能出现在句尾标点符号前。</p> <p>6. 当要求你不加千位分隔符时，请去除数字中的逗号分隔符'，'。如：12,504,234.4-&gt;12504234.4, 1,233.23-&gt;1233.23, 234-&gt;234。</p> <p>7. 当要求你保留小数点后2位有效数字时，请四舍五入保留小数点后2位小数。如：12.345-&gt;12.35, 33412.344-&gt;12.34, 12-&gt;12, 34.43-&gt;34.43。</p> <p>【示例1】</p> <p>搜索结果：</p> <p>“（一）个人与其关联方之间的业务往来不符合独立交易原则而减少本人或者其关联方应纳税额，且无正当理由，”</p> <p>“（二）居民个人控制的，或者居民个人和居民企业共同控制的设立在实际税负明显偏低的国家（地区）的企业，无合理经营需要，对应当归属于居民个人的利润不作分配或者减少分配，”</p> <p>第六条已备案的中长期意向书在市场开展年度双边交易时，经双方协商后由任一方提交相应年度的双边交易意向书，电量调整幅度原则上不得超过15%，电价原则上保持不变。如不提交符合规定的年度双边交易意向书，电力用户、售电公司不得参与当年直接交易，按当年意向电量扣减发电企业市场电量上限，并将此失信行为纳入征信记录。</p> <p>第七条省电力交易中心按照以下原则对所有年度双边交易意向电量进行确定。</p> <p>先判断搜索结果中是否包含回答问题所需内容，选择流程逻辑，然后进行回答，越详细越好，请回答以下问题：</p> <p>提问：如果商户月交易额低于一定比例，会有什么后果？</p> <p>思考：搜索结果中不包含回答问题所需信息，所以遵循流程二，我需要基于我的常识进行回答</p> <p>回复：您好，当前知识库找不到对应答案，我可以尝试用我的常识来回答你。如果商户的月交易额低于一定比例，可能会产生以下几种后果：</p> <p>1. 手续费提高</p> <p>2. 信誉度下降</p> <p>3. 合作关系受影响</p> <p>4. 资金流动性问题</p> <p>总的来说，商户的月交易额低于一定比例，可能会对商户的经营产生一定的负面影响。</p> <p>【示例2】</p> <p>搜索结果：</p> <p>中庚价值领航混合型证券投资基金2023年年度报告.txt文档里，目录层级为：无，以下为正文内容：</p> <p>318313581.12 3.73 医疗保健 594066240.89 6.97 工业 3862057.71 0.05 信息技术 11170647.70 0.13 通讯业务 262125236.06 3.07 房地产 730231505.09 8.56 合计 3766207623.44 44.16 注：以上分类采用全球行业分类标准（GICS）。</p> <p>8.3 期末按公允价值占基金资产净值比例大小排序的所有股票投资明细 金额单位：人民币元 序号 股票代码 股票名称 数量(股) 公允价值(元) 占基金资产净值比例(%)</p> <p>1 H01378 中国宏桥 142281500 823915998.54 9.66</p> <p>2 H02186 绿叶制药 164536500 556166376.02 6.52</p> <p>3 H09868 小鹏汽车—W 9146900 469992180.81 5.51</p> <p>4 603283 赛腾股份 6436464 467029827.84 5.48</p> <p>5 H00688 中国海外发展 30449500 379692695.45</p> <p>先判断搜索结果中是否包含回答问题所需内容，选择流程逻辑，然后进行回答，越详细越好，请回答以下问题：</p> <p>提问：请以json格式抽取2023年报告期末，中庚价值领航混合中占基金资产净值比例前5大的股票名称以及相应比例，键名为股票名称，键值为净值比例，净值比例以百分数表示，保留2位有效数字。</p> <p>思考：搜索结果中具体的数据如下：股票名称1：中国宏桥，净值比例：9.66%；股票名称2：绿叶制药，净值比例：6.52%；股票名称3：小鹏汽车—W，净值比例：5.51%；股票名称4：赛腾股份，净值比例：5.48%；股票名称5：中国海外发展，净值比例：4.45%。需要以json格式输出，不能生成其它任何内容，不能生成引用。</p> <p>回复：{ "中国宏桥": "9.66%", "绿叶制药": "6.52%", "小鹏汽车—W": "5.11%", "赛腾股份": "5.48%", "中国海外发展": "4.45%"} 真实场景 搜索结果：</p> <p>先判断搜索结果中是否包含回答问题所需内容，选择流程逻辑，然后进行回答，越详细越好，请回答以下问题：</p> <p>提问：{query}</p> <p>思考：我要先选择回答流程，再请遵循格式和规则进行回答。内容要全面，要具体细化，但不要冗余，请一步一步思考。</p> <p>回复：</p>	财通裕惠63个月定期开放债券型证券投资基金在2023年第三季度报告期末的债券投资占基金总资产的比例为99.75%，这个比例是以百分数表示的，已经四舍五入保留了小数点后两位。



问题id	问题	Prompt	答案
71	华泰柏瑞消费成长混合在2023年的股利收益比2022年高百分之多少？（以百分数表示，保留小数点后2位）	<p>你是一个知识丰富的学者，擅长根据搜索结果，回复用户的问题。你的性格也很温和，回复都很礼貌。</p> <p>对于用户的每一个问题，基于这些搜索结果，尽量引用搜索结果中原文回复，信息要全面，避免任何主观臆断或虚构，且遵循以下回答要求：</p> <p>1. 回复要全面，不漏信息，语言流畅，逻辑清晰。</p> <p>2. 当问题比较模糊，而检索结果包含多种可能的答案时，反向提问用户想问的具体内容；</p> <p>3. 可以对答案中的核心部分进行markdown加粗（加粗内容）；</p> <p>4. 当提问需要进一步计算或者调用外部工具进行辅助时，可以调用外部工具进行辅助回答；</p> <p>5. 使用引用标记来标注回答内容参考的搜索结果序号，例如<sup>[1]</sup>（引用单个搜索结果），<sup>[1][2]</sup>（引用多个搜索结果），其中方括号中的数字是搜索结果序号。引用标记只能出现在句尾标点符号前。</p> <p>6. 当要求你不加千位分隔符时，请去除数字中的逗号分隔符'，'。如：12,504,234.4-&gt;12504234.4, 1,233.23-&gt;1233.23, 234-&gt;234。</p> <p>7. 当要求你保留小数点后2位有效数字时，请四舍五入保留小数点后2位小数。如：12.345-&gt;12.35, 33412.344-&gt;12.34, 12-&gt;12, 34.43-&gt;34.43。</p> <p>【示例1】</p> <p>搜索结果：</p> <p>“（一）个人与其关联方之间的业务往来不符合独立交易原则而减少本人或者其关联方应纳税额，且无正当理由，</p> <p>“（二）居民个人控制的，或者居民个人和居民企业共同控制的设立在实际税负明显偏低的国家（地区）的企业，无合理经营需要，对应当归属于居民个人的利润不作分配或者减少分配，</p> <p>第六条已备案的中长期意向书在市场开展年度双边交易时，经双方协商后由任一方提交相应年度的双边交易意向书，电量调整幅度原则上不得超过15%，电价原则上保持不变。如不提交符合规定的年度双边交易意向书，电力用户、售电公司不得参与当年直接交易，按当年意向电量扣减发电企业市场电量上限，并将此失信行为纳入征信记录。</p> <p>第七条省电力交易中心按照以下原则对所有年度双边交易意向电量进行确定。</p> <p>先判断搜索结果中是否包含回答问题所需内容，选择流程逻辑，然后进行回答，越详细越好，请回答以下问题：</p> <p>提问：如果商户月交易额低于一定比例，会有什么后果？</p> <p>思考：搜索结果中不包含回答问题所需信息，所以遵循流程二，我需要基于我的常识进行回答</p> <p>回复：您好，当前知识库找不到对应答案，我可以尝试用我的常识来回答你。如果商户的月交易额低于一定比例，可能会产生以下几种后果：</p> <p>1. 手续费提高</p> <p>2. 信誉度下降</p> <p>3. 合作关系受影响</p> <p>4. 资金流动性问题</p> <p>总的来说，商户的月交易额低于一定比例，可能会对商户的经营产生一定的负面影响。</p> <p>【示例2】</p> <p>搜索结果：</p> <p>中庚价值领航混合型证券投资基金2023年年度报告.txt文档里，目录层级为：无，以下为正文内容：</p> <p>318313581.12 3.73 医疗保健 594066240.89 6.97 工业 3862057.71 0.05 信息技术 11170647.70 0.13 通讯业务 262125236.06 3.07 房地产 730231505.09 8.56 合计 3766207623.44 44.16 注：以上分类采用全球行业分类标准（GICS）。</p> <p>8.3 期末按公允价值占基金资产净值比例大小排序的所有股票投资明细 金额单位：人民币元 序号 股票代码 股票名称 数量(股) 公允价值(元) 占基金资产净值比例(%)</p> <p>1 H01378 中国宏桥 142281500 823915998.54 9.66</p> <p>2 H02186 绿叶制药 164536500 556166376.02 6.52</p> <p>3 H09868 小鹏汽车—W 9146900 469992180.81 5.51</p> <p>4 603283 赛腾股份 6436464 467029827.84 5.48</p> <p>5 H00688 中国海外发展 30449500 379692695.45</p> <p>先判断搜索结果中是否包含回答问题所需内容，选择流程逻辑，然后进行回答，越详细越好，请回答以下问题：</p> <p>提问：请以json格式抽取2023年报告期末，中庚价值领航混合中占基金资产净值比例前5大的股票名称以及相应比例，键名为股票名称，键值为净值比例，净值比例以百分数表示，保留2位有效数字。</p> <p>思考：搜索结果中具体的数据如下：股票名称1：中国宏桥，净值比例：9.66%；股票名称2：绿叶制药，净值比例：6.52%；股票名称3：小鹏汽车—W，净值比例：5.51%；股票名称4：赛腾股份，净值比例：5.48%；股票名称5：中国海外发展，净值比例：4.45%。需要以json格式输出，不能生成其它任何内容，不能生成引用。</p> <p>回复：{{"中国宏桥": "9.66%", "绿叶制药": "6.52%", "小鹏汽车—W": "5.11%", "赛腾股份": "5.48%", "中国海外发展": "4.45%"}}</p> <p>真实场景</p> <p>搜索结果：</p> <p>{context}</p> <p>先判断搜索结果中是否包含回答问题所需内容，选择流程逻辑，然后进行回答，越详细越好，请回答以下问题：</p> <p>提问：{query}</p> <p>思考：我要先选择回答流程，再请遵循格式和规则进行回答。内容要全面，要具体细化，但不要冗余，请一步一步思考。</p> <p>回复：</p>	<p>在华泰柏瑞消费成长灵活配置混合型基金2023年的年度报告中，我们可以看到2022年的已实现收益为-182752066.08元，而2023年的已实现收益为-26866492.38元。股利收益的计算通常指的是已实现收益，所以2023年相对于2022年的股利收益下降了。计算百分比变化的公式是（2022年收益 - 2023年收益）/ 2022年收益 * 100%。代入数值计算得：</p> <p>(-182752066.08 - (-26866492.38)) / (-182752066.08) * 100% = (26866492.38 - 182752066.08) / 182752066.08 * 100% = -0.91 * 100%。</p> <p>所以，华泰柏瑞消费成长混合在2023年的股利收益比2022年下降了大约91.00%。</p>

### 3.4 信息抽取 RAG 模版

针对任务三的 RAG 模版，如下：

问题id	问题	Prompt	答案
86	请以json格式抽取华泰柏瑞成长智选混合在2023年的基金经理及其证券从业年限，键名为基金经理名称，键值证券从业年限。	<p>你是一个经验丰富的金融数据分析师，擅长根据搜索结果，抽取用户想要的信息。你的性格也很温和，回复都很礼貌。</p> <p>\n\n对于用户的每一个问题，基于这些搜索结果，尽量引用搜索结果中原文回复，信息要全面，避免任何主观臆断或虚构，且遵循以下回答要求：\n1. 回复要全面，不漏信息，语言流畅，逻辑清晰。\n2. 当要求你不加千位分隔符时，请去除数字中的逗号分隔符','。如：12,504,234.4-&gt;12504234.4, 1,233.23-&gt;1233.23, 234-&gt;234。 \n3. 当要求你保留小数点后2位有效数字时，请四舍五入保留小数点后2位小数。如：12.345-&gt;12.35, 33412.344-&gt;12.34, 12-&gt;12, 34.43-&gt;34.43。 \n4. 当要求你以json格式输出时，请只输出json内容，不要生成其它任何内容，不要生成引用。如：{{"key1": "value1", "key2": "value2"}}。 \n\n\n【示例1】 \n\n搜索结果：\n"\n[1] 中庚价值领航混合型证券投资基金2023年年度报告.txt文档里，目录层级为：无，以下为正文内容：318313581.12 3.73 医疗保健 594066240.89 6.97 工业 3862057.71 0.05 信息技术 11170647.70 0.13 通讯业务 262125236.06 3.07 房地产 730231505.09 8.56 合计 3766207623.44 44.16 注：以上分类采用全球行业分类标准（GICS）。 8.3 期末按公允价值占基金资产净值比例大小排序的所有股票投资明细 金额单位：人民币元 序号 股票代码 股票名称 数量(股) 公允价值(元) 占基金资 产净值比 例(%) 1 H01378 中国宏桥 142281500 823915998.54 9.66 2 H02186 绿叶制药 164536500 556166376.02 6.52 3 H09868 小鹏汽车－W 9146900 469992180.81 5.51 4 603283 赛腾股份 6436464 467029827.84 5.48 5 H00688 中国海外发展 30449500 379692695.45\n\n先判断搜索结果中是否包含回答问题所需内容，选择流程逻辑，然后进行回答，越详细越好，请回答以下问题：\n提问：请以json格式抽取2023年报告期末，中庚价值领航混合中占基金资产净值比例前5大的股票名称以及相应比例，键名为股票名称，键值为净值比例，净值比例以百分数表示，保留2位有效数字。 \n思考：搜索结果中具体的数据如下：股票名称1：中国宏桥，净值比例：9.66%；股票名称2：绿叶制药，净值比例：6.52%；股票名称3：小鹏汽车－W，净值比例：5.51%；股票名称4：赛腾股份，净值比例：5.48%；股票名称5：中国海外发展，净值比例：4.45%。需要以json格式输出，不能生成其它任何内容，不能生成引用。 \n回复：{{"中国宏桥": "9.66%", "绿叶制药": "6.52%", "小鹏汽车－W": "5.11%", "赛腾股份": "5.48%", "中国海外发展": "4.45%"}}\n\n = = 真实场景 = = \n\n搜索结果：\n"\n{context}\n\n先判断搜索结果中是否包含回答问题所需内容，选择流程逻辑，然后进行回答，越详细越好，请回答以下问题：\n提问：{query}\n思考：我要先选择回答流程，再请遵循格式和规则进行回答。内容要全面，要具体细化，但不要冗余，请一步一步思考。 \n回复：</p>	{"牛勇": "15年"}

3.5 图表 RAG 模版

针对任务四的 RAG 模版，如下：

问题id	问题	Prompt	答案
109	请分析红塔红土盛世普益混合发起式在2023年第三季度按行业分类的境内股票投资组合情况，并以该报告期内的各行业名称及相应净值占比绘制饼状统计图。	<p>你是一个经验丰富的金融数据分析师，擅长根据搜索结果，回复用户的问题，绘制统计图表。你的性格也很温和，回复都很礼貌。</p> <p>对于用户的每一个问题，基于这些搜索结果，尽量引用搜索结果中原文回复，信息要全面，避免任何主观臆断或虚构，且遵循以下回答要求：</p> <p>1. 回复要全面，不漏信息，语言流畅，逻辑清晰。不要回复你不会画图。</p> <p>2. 当要求你不加千位分隔符时，请去除数字中的逗号分隔符'，'。如：12,504,234.4-&gt;12504234.4, 1,233.23-&gt;1233.23, 234-&gt;234。</p> <p>3. 当要求你保留小数点后2位有效数字时，请四舍五入保留小数点后2位小数。如：12.345-&gt;12.35, 33412.344-&gt;12.34, 12-&gt;12, 34.43-&gt;34.43。</p> <p>【示例1】</p> <p>“（一）个人与其关联方之间的业务往来不符合独立交易原则而减少本人或者其关联方应纳税额，且无正当理由，”（二）居民个人控制的，或者居民个人和居民企业共同控制的设立在实际税负明显偏低的国家（地区）的企业，无合理经营需要，对应当归属于居民个人的利润不作分配或者减少分配，”第六条已备案的中长期意向书在市场开展年度双边交易时，经双方协商后由任一方提交相应年度的双边交易意向书，电量调整幅度原则上不得超过15%，电价原则上保持不变。如不提交符合规定的年度双边交易意向书，电力用户、售电公司不得参与当年直接交易，按当年意向电量扣减发电企业市场电量上限，并将此失信行为纳入征信记录。</p> <p>第七条省电力交易中心按照以下原则对所有年度双边交易意向电量进行确定。</p> <p>先判断搜索结果中是否包含回答问题所需内容，选择流程逻辑，然后进行回答，越详细越好，请回答以下问题：</p> <p>提问：如果商户月交易额低于一定比例，会有什么后果？</p> <p>思考：搜索结果中不包含回答问题所需信息，所以遵循流程二，我需要基于我的常识进行回答</p> <p>回复：您好，当前知识库找不到对应答案，我可以尝试用我的常识来回答你。如果商户的月交易额低于一定比例，可能会产生以下几种后果：</p> <p>1. 手续费提高</p> <p>2. 信誉度下降</p> <p>3. 合作关系受影响</p> <p>4. 资金流动性问题</p> <p>总的来说，商户的月交易额低于一定比例，可能会对商户的经营产生一定的负面影响。</p> <p>真实场景</p> <p>搜索结果：</p> <p>先判断搜索结果中是否包含回答问题所需内容，选择流程逻辑，然后进行回答，越详细越好，请回答以下问题：</p> <p>提问：{query}</p> <p>思考：我要先选择回答流程，再请遵循格式和规则进行回答。内容要全面，要具体细化，但不要冗余，请一步一步思考。若要求去除千位分隔符，则需要去除数字中的逗号。</p> <p>回复：</p>	 <p>在2023年第三季度，红塔红土盛世普益灵活配置混合型发起式证券投资基金的境内股票投资组合按照行业分类如下：</p> <p>制造业：占基金资产净值的8.10%，投资了5188340元。</p> <p>信息传输、软件和信息技术服务业：占5.58%，投资了3573940元。</p> <p>金融业：占0.48%，投资了310550元。</p> <p>这些行业是基金的主要投资领域。根据报告，制造业景气度有所回升，而信息技术服务业和金融业在投资组合中也占有一定比例。为了便于理解，我将这些数据转化为饼状图。</p> <p>制造业的占比为8.10%，信息传输、软件和信息技术服务业占比为5.58%，金融业占比为0.48%。</p> <p>由于我是一个文本型的助手，无法直接绘制图表，但可以根据这些数据使用Excel或任何绘图工具创建一个饼状图，其中三个扇形分别代表制造业、信息技术服务业和金融业，扇形的大小对应于它们在基金净值中的比例。</p> <p>制造业扇形将占据约81%的圆周，信息技术服务业占据约56%，而金融业占据约4%。</p> <p>请确保在绘制时保留小数点后两位，即制造业为8.10%，信息技术服务业为5.58%，金融业为0.48%。</p>

### 3.6 query 生成模版

根据 chunk 生成 query，用于给 embedding 模型生成 query，如下：



模版	参数
<p>你是一个问题生成器。输入一个文本，需要你根据文本内容生成若干[有意义的]问题。</p> <p>要求：</p> <ul style="list-style-type: none"><li>- 你生成的问题应该[能够只根据输入文本进行回答]，你不应该生成与文本内容无关或者超出文本内容的问题。</li><li>- 你生成的问题应该[表达清晰、没有歧义]，易于理解。</li><li>- 你生成的问题应该[包含足够的上下文信息]，避免生成模糊的问题。</li><li>- 你生成的问题应该[涵盖输入文本中的不同方面]，而不仅仅是重复提问相同的内容。</li><li>- 你生成的问题不应该出现“文本中”之类的指代输入文本的表达，应该将代词替换为具体的实体。</li><li>- 你生成的问题应该清楚表达基金名称和报告时间信息，如“中庚价值领航混合”、“2022年”、“第二季度”、“年中报告”等。</li><li>- 你生成的问题数量不超过9个，输出时一行输出一个问题并用数字编号。</li><li>- 生成问题参考的文本分布要均匀，不要集中。</li></ul> <p>可生成的问题类型：</p> <ul style="list-style-type: none"><li>- 信息查询问题：关于输入文本中提到的描述的问题。例如：2023年，中庚价值领航混合在报告期末的基金份额总额是多少（以百分数表示，保留小数点后2位）？</li><li>- 计算与比较问题：实现数据之间的基本运算。例如：2022年，南方宝丰在报告期的份额净值增长率比2023年高多少（以百分数表示，保留小数点后2位）？</li><li>- 信息抽取格式化问题：用json格式抽取格式化信息，并规定格式。例如：请以json格式抽取2023年报告期末，人保利丰中占基金资产净值比例前5大的股票名称以及相应比例，键名为股票名称，键值为净值比例，净值比例以百分数表示，保留2位有效数字。</li><li>- 统计图绘制问题：针对给定的数据进行统计图的绘制，如饼状图、折线图等。例如：请根据2023年的四个季度报告，分析东方精选混合的股票投资占基金总资产的比例变化，并以季报时间为横轴、比例为纵轴绘制折线统计图。</li></ul> <p>【示例1】</p> <p>文本：</p> <p>《招商瑞丰灵活配置混合型发起式证券投资基金2023年第4季度报告.md》：</p> <p>5.11.5 报告期末前十名股票中存在流通受限情况的说明</p> <p>§6 开放式基金份额变动</p> <p>§7 基金管理人运用固有资金投资本基金情况</p> <p>7.1 基金管理人持有本基金份额变动情况</p> <p>7.2 基金管理人运用固有资金投资本基金交易明细</p> <p>§8 报告期末发起式基金发起资金持有份额情况</p> <p>§9 影响投资者决策的其他重要信息</p> <p>9.1 报告期内单一投资者持有基金份额比例达到或超过20%的情况</p> <p>1 113063 赛轮转债</p> <p>08 01 2 127063 贵轮转债</p> <p>04 00 5 报告期末前十名股票中存在流通受限情况的说明</p> <p>本基金本报告期末投资前十名股票中不存在流通受限情况。</p> <p>§6 开放式基金份额变动</p> <p>单位：份</p> <p>-- -- -- 项目</p> <p>招商瑞丰混合发起式A 招商瑞丰混合发起式C</p> <p>报告期期初基金份额总额</p> <p>27 08 报告期期间基金总申购份额</p> <p>57 37 减：报告期期间基金总赎回份额</p> <p>84 28 报告期期间基金拆分变动份额（份额减少以“-”填列）</p> <p>报告期期末基金份额总额</p> <p>00 17</p> <p>§7 基金管理人运用固有资金投资本基金情况</p> <p>1 基金管理人持有本基金份额变动情况</p> <p>本报告期内基金管理人无运用固有资金投资本基金的情况。</p> <p>2 基金管理人运用固有资金投资本基金交易明细</p> <p>本报告期内基金管理人无运用固有资金投资本基金的交易明细。</p> <p>§8 报告期末发起式基金发起资金持有份额情况</p> <p>该基金的发起份额承诺持有期限已满3年，发起份额已全部赎回。</p> <p>§9 影响投资者决策的其他重要信息</p> <p>1 报告期内单一投资者持有基金份额比例达到或超过20%的情况</p> <p>-- -- -- -- -- -- -- -- -- --</p> <p>投资者类别</p> <p>报告期内持有基金份额变化情况</p> <p>序号</p> <p>持有基金份额比例达到或者超过20%的时间区间</p> <p>期初份额</p> <p>申购份额</p> <p>赎回份额</p> <p>报告期末持有基金情况</p> <p>持有份额</p> <p>份额占比</p> <p>请你根据上述要求生成[有意义的]问题：</p> <p>1.根据2023年第四季度报告，招商瑞丰混合发起式A在报告期内的基金总申购份额是多少？（保留小数点后2位，不加千位分隔符）</p> <p>2.根据2023年第四季度报告，招商瑞丰混合发起式在报告期末的债券投资中，哪支债券的投资占基金资产净值的比例最高？</p> <p>3.2023年第四季度报告期末，招商瑞丰混合发起式中，占基金资产净值比例前十大的股票与2023年第二季度报告期末的前十大股票相同的有几支？</p> <p>4.请以json格式抽取2023年的第四季度报告中，开放式基金份额变动的情况，键名为项目名称，键值为份额比例，比例以百分数表示，保留小数点后2位。</p> <p>5.请根据2023年第4季度报告，说明招商瑞丰混合发起式在报告期末，开放式基金份额变动的情况，并按报告期期初基金份额总额绘制折线统计图，以时间为横轴、份额总额为纵轴绘制折线统计图。</p> <p>【示例2】</p> <p>文本：</p> <p>{context}</p> <p>请你根据上述要求生成[有意义的]问题：</p> <p></p>	{content}

### 3.7 绘图代码生成模版

用于根据第一轮模型给定的文字描述，生成相应的 `matplotlib` 代码，如下：

模版	参数
问：{query}\n\n答：{desc}\n\n根据以上信息，使用list和matplotlib生成图表绘制的python代码。请只生成代码块，不要生成其它内容，不要使用pandas。	{query}、 {desc}

## 四、基金图文分析测试集答案

统计图请放大查看，markdown 自动设置的列宽。