# 基于知识库和LLM的问答技术方案一

## 1 技术背景

LLM可应用于招联的会议纪要转录结果问答、公开课视频转录结果问答以及庭审视频的摘要提取，可以提取会议、公开课、庭审转录结果的摘要、总结核心观点等。LLM结合招联现有的音视频转录技术，可以极大程度提高会议、公开课以及庭审记录的效率和体验。但直接将LLM进行应用存在如下问题：

1）答非所问。通常公司的会议、公开课以及视频庭审过程会涉及部分专有知识点，LLM对于相关内容问题的回答很有可能出现超出具体会议、公开课以及具体某场庭审内容范围的现象，即答非所问、不知所云。

2）LLM输入限制。将包含专有知识点的文本直接作为LLM的输入内容一定程度上可以解决第1个问题，但LLM的输入有长度限制，超出限制长度无法进行正常问答。

本方案主要通过基于知识库和LLM的问答技术解决上述2个核心问题，同时计划构建LLM+知识库问答的pipline，实现问答技术的可配置、易拓展、易应用。

## 2 目标

### 2.1 技术目标

1）实现将LLM的回答限制在给定内容范围内。（prompt+知识库）

2）在1的基础上解决LLM输入长度限制问题。（知识抽取）

3）构建知识库+LLM问答的pipline，提高知识库+LLM问答的可拓展性。

### 2.2 应用目标

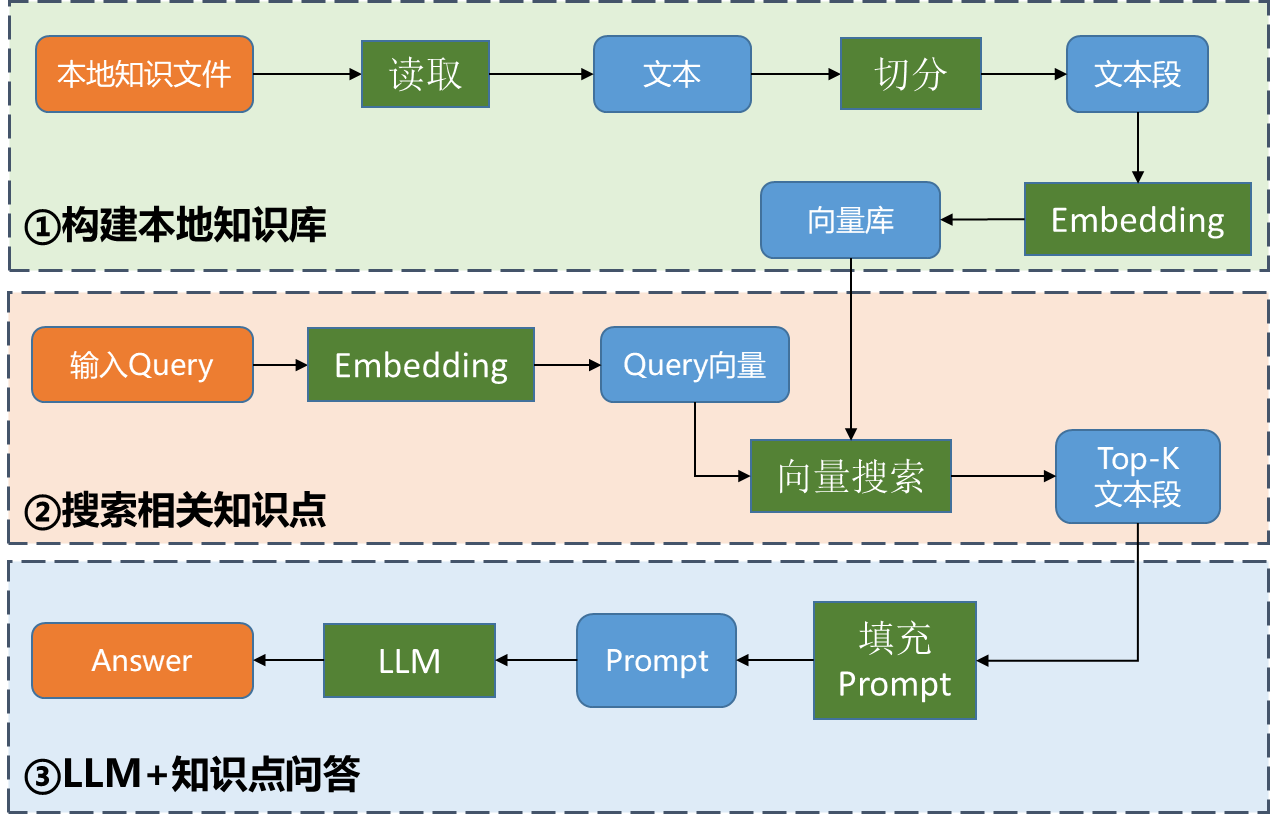
1）应用于会议纪要二期建设。基于本方案实现：①主动提取摘要、关键词、主题等内容；②结合转录内容，为用户提供半开放式的问答窗口。

2）应用于AMC视频庭审，实现提取庭审转录的摘要。

3）应用于招联公开课视频，应用方式与1类似。

## 3 技术流程

本方案的技术流程如下图所示，主要包含3个部分：构建本地知识库、搜索相关知识点和LLM+知识点问答。



1. 构建本地知识库

①读取本地知识文件得到文本；

②对文本按照规则进行切分，得到文本段；

③对所有文本段进行Embdedding（OpenAI、BERT等）并存储在向量库中。

1. 搜索相关知识点

①对输入Query进行Embdding得到Query向量；

②使用Query向量，结合向量库，通过向量搜索技术（Flat、IVFx Flat、HNSWx、PQx等）得到Top-K个最相关的文本段作为知识点；

1. LLM+知识点问答

①基于2得到的Top-K个知识点，设计填充Prompt模板，为LLM提供回答的参考；

②将Prompt输入至LLM得到回答结果。LLM可采用api（chatgpt）、本地部署（chatglm）。

## 4 关键研究点

* 本地知识库构建部分：代码实现及文本切分Embedding策略设计
* 知识点搜索部分：代码实现及向量搜索算法、搜索策略设计
* LLM+知识点问答部分：chatglm本地部署与Prompt模板设计。

## 5 研究成本

### 5.1 技术难度

基于知识库和LLM的问答技术难度主要在于如何设计合适的知识库构建和知识抽取策略，为LLM提供最相关的知识点，实现“答即所问”的问答体验。

### 5.2 风险点

（1）本方案主要通过给定的知识文件构建知识库并进行后续应用，因此，给定知识库的范围和质量会对最终问答结果产生较大影响。

（2）本方案通过知识库限制LLM的问答范围，LLM的回答结果在限制范围内仍有一定的随机性。

### 5.3 计划

（1）时间：3个月

（2）人力：2人

（3）GPU资源：A100\*1

（4）进度安排：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 任务 | 时间 | 目标 | 进度 | 结论 |
| LLM+知识库问答初步验证 | 2023.06.05  -  2023.06.15 | 验证LLM+直接输入文本填充prompt的问答效果。 | 已完成 | 效果见：<https://docs.mucfc.com/weboffice/p/988535>  结论为：在测试案例中，通过prompt+输入填充prompt能够有效地将LLM的回答限制在给定知识库范围内。同时存在输入不能超过LLM字符限制的问题。 |
| 本地知识库构建策略设计与代码实现 |  |  | -- |  |
| 知识点搜索策略设计与代码实现 |  |  | -- |  |
| chatglm本地部署与Prompt模板设计 |  |  | -- |  |