

参赛学生姓名：陈我风

中学：深圳国际交流书院

省份：广东省

国家/地区：中国

指导老师姓名：陈凯、吴邵汐

指导老师单位：上海交通大学、深圳国际交流书院

论文题目：基于大语言模型的古诗词智能搜索系统设计
与实现

基于大语言模型的古诗词智能搜索系统设计与实现

Design and Implementation of an Intelligent Ancient Poetry Search System Based on Large Language Models

陈我风 深圳国际交流书院

摘要

古诗词智能搜索系统在文化遗产、提高学术研究效率等众多方面具有重要的研究意义。目前一般采用基于传统搜索引擎、专业古诗词网站，以及直接向大语言模型提问的方法，这些方法在搜索意图理解、权威性及完整性方面还存在比较明显的缺陷。

针对上述问题，本论文探讨利用大语言模型技术构建古诗词智能搜索系统的方法，具体包括如下工作：（1）论文整理了并发布了 22 万首包括人物及地名结构化信息的古诗词数据集；（2）设计并实现了一种基于大语言模型的两阶段古诗词智能搜索系统的方法，构建数据库阶段利用大模型抽取结构化语义信息，搜索阶段利用大语言模型把搜索意图转换数据库检索指令。（3）设计并实现一种基于思维链和外部知识辅助的古诗词结构化信息提取提示词方法，以及将古诗词自然语言查询转换为数据库查询的提示词方法，从而有效提取古诗词的结构化信息，有效理解用户意图。

论文实现了基于大语言模型的古诗词智能搜索系统的原型，并设计了典型测试样本进行了测试，准确达到 75%，说明了我们的方法通过结合大语言模型能力，不需要进行定制训练，可以有效获取古诗词的多维结构化语义信息，有效获取用户搜索意图，初步达到预期效果，为古诗词智能搜索系统的实现积累了一定的经验。本论文所有代码和数据集都在 GitHub 发布[23]。

关键词：大语言模型、古诗词搜索、提示词、思维链

目录

1. 绪论	4
1.1 研究意义	4
1.2 相关研究工作	4
1.2.1 利用搜索引擎的方法	4
1.2.2 利用大语言模型提问的方法	5
1.2.3 利用专业的古诗词文网站搜索的方法	6
1.3 本文工作	7
2. 我们的方法	8
2.1 总体思路和流程	8
2.2 古诗词结构化抽取方法	9
2.2.1 相关研究工作	9
2.2.2 我们的古诗词结构化提取方法	9
2.3 古诗词搜索意图理解	12
2.3.1 相关研究工作	12
2.3.2 我们的古诗词搜索意图理解方法	13
3. 系统实现及测试	16
3.1 数据处理	16
3.1.1 原始数据	16
3.1.2 处理后的数据	16
3.2 系统实现及测试	16
3.2.1 系统架构	16
3.2.2 Demo 系统	16
3.2.3 测试讨论	18
4. 结论及展望	19
4.1 结论	19
4.2 展望	19
参考文献	20
致谢	21
附录 1	22

1. 绪论

1.1 研究意义

古诗词智能搜索系统通过计算机技术，特别是自然语言处理（NLP）与人工智能技术，对大量古代诗歌数据进行高效处理与检索。它不仅仅局限于传统意义上的关键词匹配搜索，更进一步地，能够理解用户的查询意图，并根据上下文环境提供相关性更高的搜索结果。

古诗词智能搜索系统具有重要的研究意义：

- **文化传承与普及：**古诗词是中华民族文化宝库中的瑰宝，对于弘扬传统文化具有不可替代的作用。智能搜索系统有助于让这些文化遗产更好地融入现代生活，让更多人尤其是年轻人能够方便地接触、学习并欣赏到古典文学的魅力。
- **提升学术研究效率：**传统的文献检索方式耗时费力，而智能化的搜索工具可以极大提高学者们查找资料的速度与准确性，从而促进学术界对古诗词领域更深入细致的研究。例如，学者可以在几秒钟内找到发生在“河西走廊”相关的所有诗句，极大提升研究效率。
- **促进古诗词的普及和教育：**通过提供便捷的检索工具，古诗词能更易于被大众和学生所了解。比如，教师可以在课堂上即时找到适合教学的诗词内容，增加教学的互动性和趣味性。
- **促进跨学科交叉融合：**随着技术的发展，古诗词智能搜索系统可以与其他领域相结合，如音乐、绘画等艺术形式，产生更多创新应用，丰富文化传播手段。比如，音乐制作人利通过智能搜索系统匹配歌词灵感，创作融合古风旋律与现代编曲技法的新歌。

1.2 相关研究工作

目前古诗词智能检索系统一般采用如下方法，（1）利用搜索引擎，例如：百度/搜狗/必应/谷歌等等；（2）利用大语言模型提问，例如：通义千问、文心一言、ChatGPT；（3）利用专业的古诗词文网站，例如：古诗文网；中国哲学书电子化计划[1,2,3,4]等等。下面我们对各种方法进行描述。

1.2.1 利用搜索引擎的方法

利用搜索引擎进行古诗词检索的技术原理[5,6]如下，（1）首先进行索引构建，搜索引擎使用网络爬虫抓取网页内容，并将其分解成单独的词汇或短语。每个词汇或短语的出现位置及其相关信息被记录在索引中。这种索引允许搜索引擎快速找到包含特定词汇的网页。（2）查询处理，用户输入查询词后，搜索引擎会对这些词进行处理，包括词干提取、去停用词、同义词扩展等。处理后的查询词与索引中的词汇进行匹配，找到相关网页。（3）排名算法，搜索引擎使用复杂的排名算法（如 PageRank、TF-IDF、BM25）对匹配的网页进行排序。排名因素包括关键词出现频率、网页权重、用户行为数据等，以确保返回的结果最相关。

该方法优点：（1）高效快速：搜索引擎能够在短时间内处理大量数据并返回相关结果。这使得用户能够迅速获取古诗词的相关信息和文本。（2）广泛覆盖：搜索引擎能够访问互联网上的大量资源，包括古诗词的原文、解释、背景信息和评论，提供全面的信息获取途径。

但该方法也存在明显的不足，（1）上下文理解不足：传统搜索引擎对古诗词的深层含义、

历史背景和文学风格理解有限，可能无法准确处理复杂的查询，导致返回的结果不够精准或相关。(2) **关键词依赖**：搜索结果高度依赖于用户输入的关键词。如果用户的查询不够准确或具体，搜索引擎可能返回大量不相关的信息，增加了信息筛选的难度。(3) **完整性和权威性不够**：搜索引擎结果中会出现很多重复的结果，并且会漏掉很多重要结果。

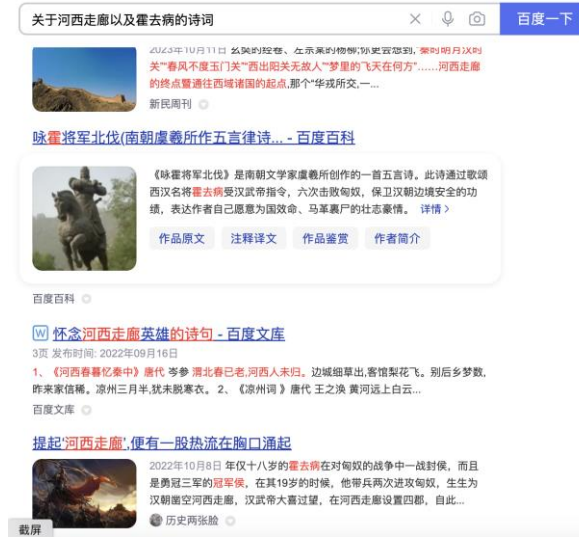


图 1 (a) 百度搜索结果

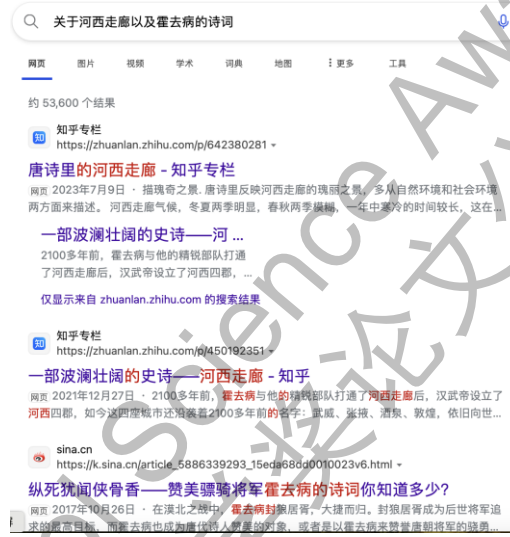


图 1 (b) 必应搜索结果

1.2.2 利用大语言模型提问的方法

利用大语言模型（如通义千问、文心一言、ChatGPT）进行古诗词检索的方法涉及以下几个技术步骤[7]：(1) **模型训练**：大语言模型通常通过大量的文本数据进行训练，包括古诗词及其相关的文学评论和背景信息。这些模型通过学习语言的模式和结构，能够生成和理解自然语言中的复杂含义。(2) **查询解析**：用户输入的查询会被转化为语言模型可以理解的格式。大语言模型可以解析用户的意图，包括对古诗词的具体要求，如作者、地点、人物、朝代、主题或特定的情感表达。(3) **生成答案**：基于训练数据和用户查询，大语言模型生成相关的古诗词或其解释。模型可以结合上下文，提供更精确的回答，同时考虑用户的查询意图和历史背景。(4) **上下文理解**：大语言模型在生成答案时，能够考虑查询的上下文，包括历史背景、诗词的文学风格和文化内涵。这使得生成的回答更符合用户的实际需求。

该方法优点：(1) **上下文理解**：大语言模型能够深入理解古诗词的上下文和文化背景，从而提供更具深度和准确性的答案。它们可以处理复杂的查询，提供更为精确的古诗词及其解释。(2) **灵活性高**：与传统搜索引擎相比，大语言模型能够更自然地处理各种类型的查询，包括那些非常具体或模糊的问题。用户可以用自然语言提问，模型能够理解并生成相关的回应。

但该方法也有明细不足：(1) **生成内容的准确性**：尽管大语言模型在生成内容时能够考虑上下文，但模型的回答有时可能不完全准确或包含误导即幻觉信息。由于模型的训练数据可能存在偏差，生成的回答也可能受到影响。(2) **背诵原文能力不足**：大语言模型可能在生成古诗词的原文时存在不准确的情况，尤其是在处理长篇文本或复杂查询时，可能无法准确背诵或引用相关的诗词原文。(3) **完整性和权威性不够**：和搜索引擎方法一样，该方法的结果中会漏掉很多重要结果。

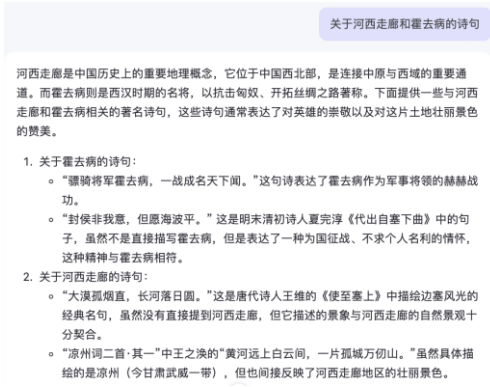


图 2 (a) 通义千问结果

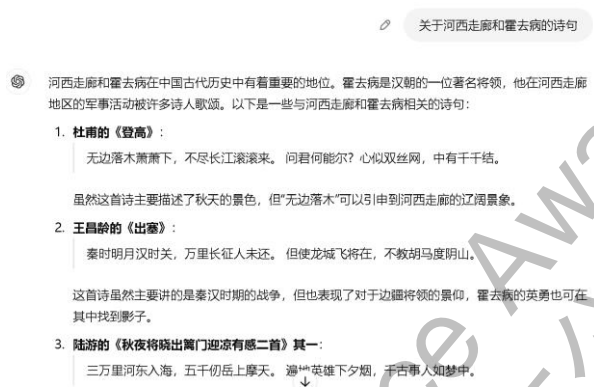


图 2 (b) GPT4o 搜索结果

1.2.3 利用专业的古诗词文网站搜索的方法

利用专业的古诗词文网站(如古诗文网、中华诗词网、中国哲学书电子化计划等[1,2,3,4])进行古诗词检索的方法技术原理如下:(1) 数据整合与管理:专业古诗词网站通常拥有专门的数据库,收录了大量的古诗词及相关文献。这些网站会系统地整理和分类古诗词,根据作者、朝代、体裁等信息进行标记,以方便检索。例如,古诗文网和中国哲学书电子化计划都会对古诗词进行系统化的整理,提供详细的注释和解释。(2) 检索功能实现:用户可以通过这些网站的搜索功能输入关键词、作者名字、朝代或其他相关信息进行检索。网站利用预先构建的索引和数据库中的结构化数据,快速匹配用户的查询。检索功能通常包括模糊匹配和精确匹配,以适应不同类型的查询需求。(3) 结果呈现与辅助功能:检索结果通常包括古诗词的原文、作者信息、创作背景、相关注释等。专业网站还可能提供额外的功能,如诗词的朗读、注释、历史背景介绍等,以帮助用户更好地理解 and 欣赏古诗词。

方法的优点:(1) 高准确性与权威性:专业古诗词网站的内容经过严格审核和整理,具有较高的准确性和权威性。这些网站通常由文学专家或学术机构维护,确保古诗词及其解释的真实性和准确性。(2) 深度解析与附加信息:除了提供原文外,专业网站通常还会提供详尽的注释、背景信息和文学分析。这些附加信息有助于用户更全面地理解古诗词的内涵和背景。(3) 易用性和专业性:这些网站设计通常注重用户体验,提供直观的检索界面和丰富的功能,使得用户能够方便地找到所需的古诗词及其相关信息。

该方法的不足之处在于:(1) 智能搜索能力差:尽管专业古诗词网站提供了大量的资源,但其检索功能主要依赖于关键词匹配,智能搜索能力相对较弱。这意味着用户需要准确输入关键词,否则可能无法找到相关的古诗词。参考文献[3]的工具提供图谱功能,可以通过图的方式展示链接,但并没有提供一个有效的智能搜索方式;(2) 对搜索意图理解不好:专业古诗词网站在解析用户搜索意图方面存在局限。它们通常无法深入理解用户的复杂查询或意图,只能基于输入的关键词提供匹配结果,这可能导致返回的结果与用户的实际需求不完全匹配。(3) 搜索维度受限:搜索的维度一般包括关键词、作者、朝代,但更深度的维度存在欠缺,例如:任务、地点、关联关系等等。



图 3 古诗文网结果

方法	明细优势	权威性	完整性	搜索维度	意图理解
搜索引擎方法	快速便捷	好	较差	一般	一般
大语言模型方法	理解搜索意图很好	差	差	丰富	很好
专业古诗文网站搜	权威性/完整性很好	好	好	较差	较差

表 1: 各方法优势总结及对比

1.3 本文工作

如上所述，古诗词智能搜索系统具有重要的研究意义，但目前的三种方法都还存在明显的缺点（如表 1 所示）。虽然大语言模型的结果还有很多待提高的地方，甚至很多地方出现了“胡说”幻觉，但大语言模型方法带来的对搜索意图的理解力，对内容的理解和认知能力，也给我们展现了大语言模型的巨大潜力，基于此，我们计划开展本项研究工作：

本文的研究内容是：探讨利用大语言模型技术构建古诗词智能搜索系统，利用大语言模型的能力辅助进行古诗词库的结构化信息抽取工作，以及辅助用户搜索意图理解工作，最终从从古诗词数据库中完整和权威的获取数据，如果上述能验证成功，可以有效的利用大语言模型的能力，并规避其准确性和权威性的问题，实现更好的古诗词智能搜索体验。

本文的主要工作如下：

- 1) 设计并实现了一种基于大语言模型的古诗词智能搜索系统，据我们最大知识能力所知，公开的论文/专利等文献没有检索到采用基于大语言模型的古诗词智能搜索系统的工作，我们的方法具有较好的创新性；
- 2) 设计并实现了一种基于思维链的古诗词结构化信息提取提示词方法，该方法可以有效的利用大语言模型自身能力，并结合辅助信息，提取古诗词的结构化信息；设计并实现了一种把古诗词自然语言查询转换数据库查询的提示词方法，并把查询结果根据搜索意图总结回复，该方法可以有效的理解用户意图，提取相应准确的后续数据并答复；
- 3) 我们实现基于大语言模型的古诗词智能搜索系统原型，采集并处理了公开数据集进行处理，结构化提取了包括篇名、作者、作者介绍、古诗词中地址（古代名/现代名）、古诗词中人物等较多维度的结构化数据信息。我们把原型上线了测试，达到了较好的效果。

2. 我们的方法

2.1 总体思路和流程

我们的方法流程图如图 4 所示，图 5 是一个用户搜索示例。方法包括 2 个步骤：(1) 古诗词数据库构建过程：利用大模型能力提取结构化语义信息，例如：人物、地址、关系、事件等，这些结构化语义信息可以大大扩充搜索的维度；(2) 用户搜索过程：利用大语言模型能力对用户的自然语言查询进行更精确的分析，精准理解搜索意图并转换为数据库查询指令 SQL，直接从古诗词数据库中完整和权威的获取数据。最后汇总得到搜索结果返回。

我们的方法的特点是：

- 采用标准的数据库方式存储数据，从而保证数据的权威性和完整性；
- 利用大模型对原始数据库进行语义分析，提取更多结构化信息，扩充搜索维度；
- 利用大模型对原始搜索意图进行理解，提升意图理解能力。

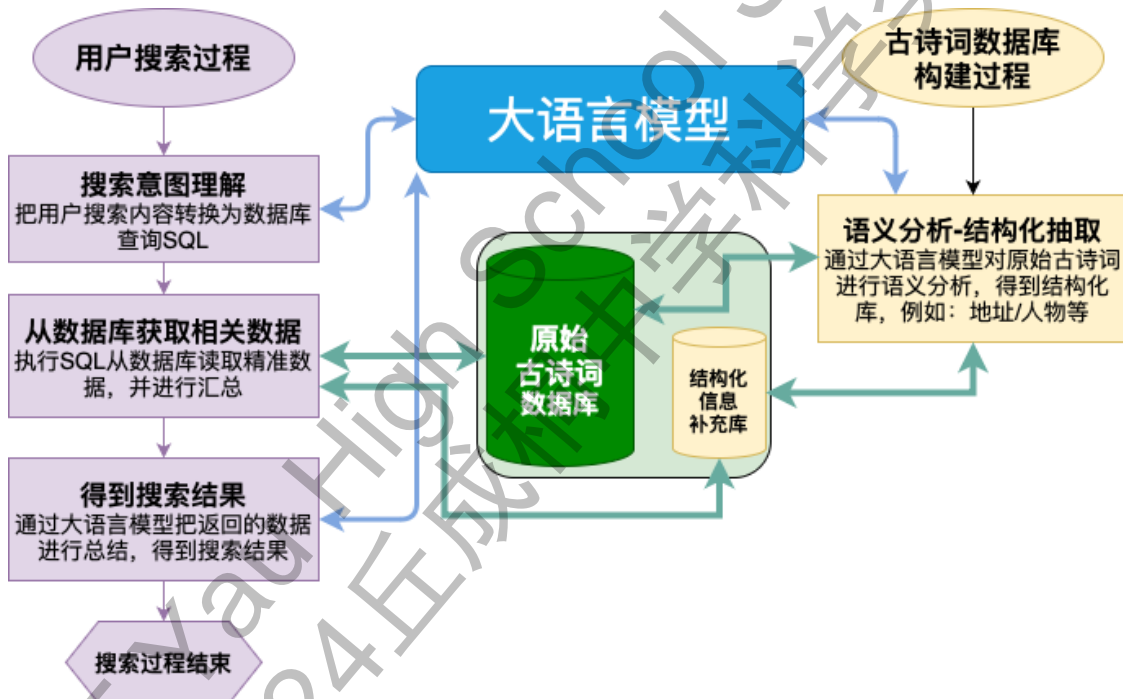


图 4：我们方法的总体流程图

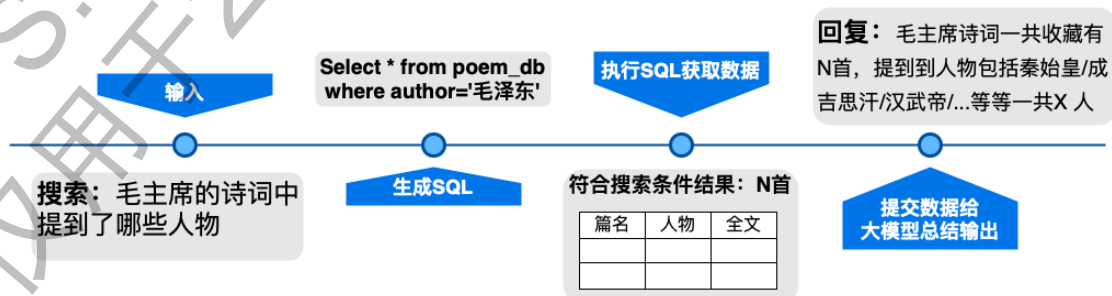


图 5 用户搜索示例

2.2 古诗词结构化抽取方法

2.2.1 相关研究工作

文献[9]在新的预训练语言模型 CP-ChineseBERT 基础上,采用图注意力网络获取古诗知识图谱的语义表示。文献[10]致力于解决中小学生在古诗词学习过程中遇到的信息获取效率低下问题,构建了一个综合性的古诗词知识图谱来提升学习体验。古诗词结构化抽取的工作是基于中小学语文教材及相关网络资源,通过实体识别、关系抽取等自然语言处理技术,整合了大量的古诗词及相关人物信息,形成了结构化的古诗词知识图谱,并存储于 HugeGraph 图数据库中。

文献[11]提出了一种基于大语言模型的学科知识图谱自动构建与更新方法。通过训练用户模型、实体关系识别和知识三元组抽取,实现知识图谱的自动化构建,减少人工和时间成本。加入用户交互机制,实现知识图谱的动态扩展和个性化更新,同时减少模型幻觉现象,提升图谱的准确性和可解释性。

专利[12]提出了一种基于大语言模型的非结构化文本数据知识抽取方法:对获取的原始文档进行命名实体识别,并将命名候选项进行分类;基于对命名实体对候选项的关系推理,确定对应的实体关系三元组候选项。本发明提出的基于大语言模型的方法来实现对非结构化文本数据的知识抽取,有效解决大批量、多领域非结构化文本数据自动知识抽取问题,提高知识抽取的效率和通用性。

OneKE[13]是由蚂蚁集团与浙江大学联合发布的中英双语大模型知识抽取框架,旨在解决当前大语言模型存在的推理能力不足、事实知识匮乏、生成结果不稳定等问题,从而推动大模型在各行业的实际应用。该框架不仅优化了基于 Schema 的轮询指令构造技术,增强了模型在结构化信息抽取方面的泛化能力,还通过开源形式贡献给了 OpenKG 开放知识图谱社区,提供了完整的工具链支持。此外,OneKE 已经在多个内部业务领域得到应用,并支撑了蚂蚁集团的知识图谱构建能力。

思维链 (Chain-of-Thought, CoT) [14]是一种提示工程技术,旨在引导大型语言模型进行更深入、更有逻辑的推理。其核心原理是将复杂问题分解为一系列中间步骤,鼓励模型逐步思考,显式地展现推理过程。CoT 通过在提示中加入“让我们一步步思考”等引导性短语,或提供具体的示例,可以激发模型生成中间推理步骤。这些步骤明确展示了模型如何从已知信息推导出答案,从而提高了模型在复杂任务上的表现,特别是在涉及计算、逻辑推理等领域。思维链提示工程的优势在于:(1)提高模型可解释性:显式推理过程有助于理解模型决策,增强用户信任。(2)增强模型解决复杂问题能力:分解问题,降低推理难度,提升模型准确性。

2.2.2 我们的古诗词结构化提取方法

根据上述调研,古诗词结构化提取方法主要还是基于传统 NLP、NER、知识图谱方法抽取,用大语言模型辅助的方法暂时还没看到,但受到上述[12,13]等调研文献启发,我们设计了基于思维链及外部知识参考的结构化提示词方法采用大语言模型进行提取古诗词结构化提取。



图 6-a 原始提示词及结构化抽取结果

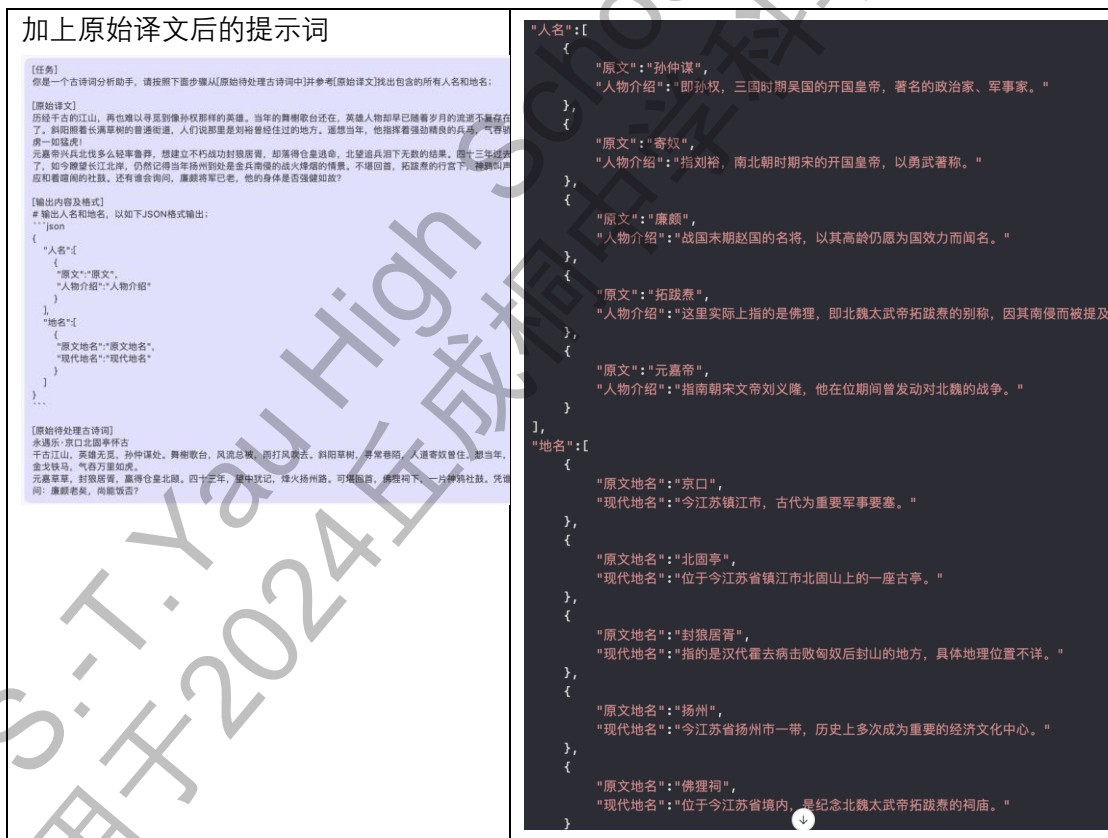


图 6-b 加上参考译文后的提示词及结构抽取结果

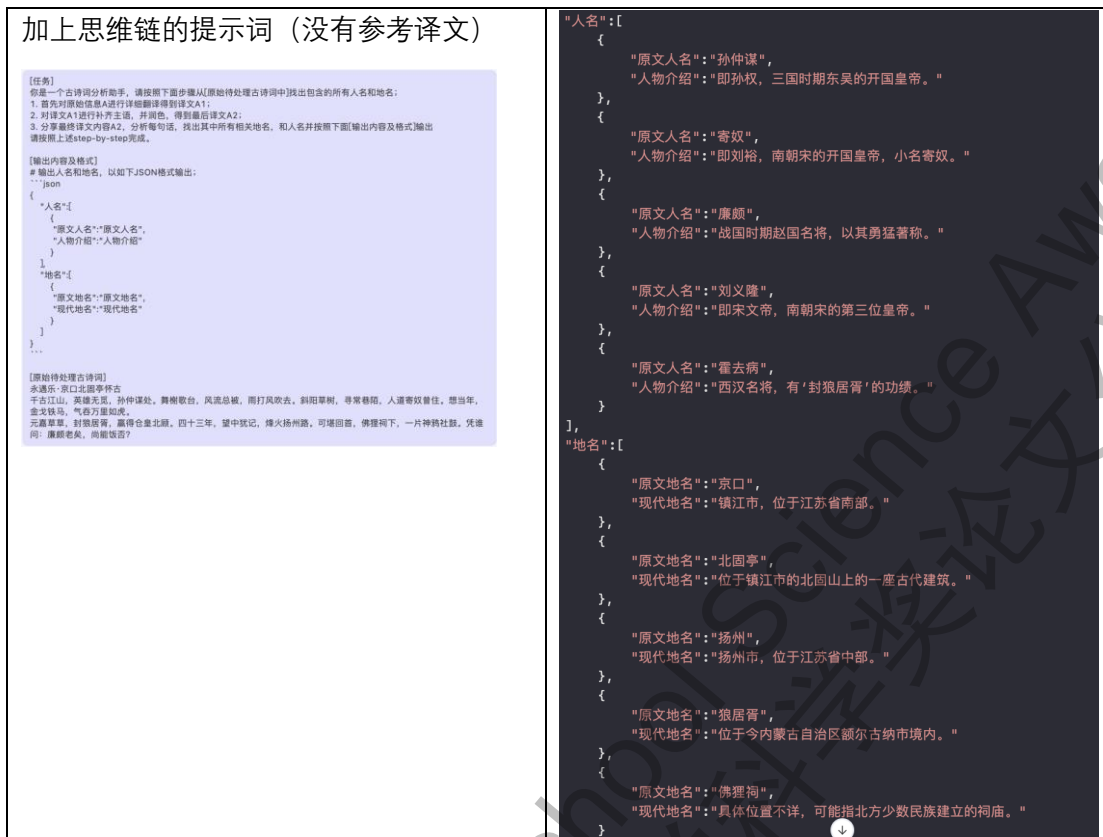


图 6-c 基于思维链方法的提示词及结构抽取结果

首先，我们的方法把提示词分为几个重要的结构化模块，包括：角色指定、任务步骤拆解、参考内容、输出格式等，每个模块通过明确的括号及分隔符分开；这样对大语言模型的指令跟随具有重要大的作用；

其次，我们在提示词中加入外部参考译文，拓展大语言模型的外部知识。通过图 6-a、及图 6-b 的对比结果，提示词在加上参考译文后，查全率得到明显提升，准确补上了在原始提示词中没有找到的“拓跋焘”和“元嘉帝”。

此外，我们参考思维链[14]的思路，把提取工作进行分步骤细化说明，从而有效激发大语言模型的能力。通过图 6-b 和图 6-c 的对比，加上思维链方法后，即使没有用外部参考译文，也准确提取了“拓跋焘”和“元嘉帝（刘义隆）”。

我们最后设计的提示词如下：

[任务]
你是一个古诗词分析助手，请按照下面步骤从[原始待处理古诗词中]找出包含的所有人名和地名：
1. 首先对原始信息 A 进行详细翻译得到译文 A1；
2. 对译文 A1 进行补齐主语，并润色，得到最后译文 A2；
3. 分析最终译文内容 A2，并结合[外部参考译文]，找出其中所有相关地名，和人名并按照下面[输出内容及格式]输出
请按照上述 step-by-step 完成。

[外部参考译文]
.....

[输出内容及格式]

输出人名和地名，以如下 JSON 格式输出：

```
```json{
 "人名": [{
 "原文人名": "原文人名",
 "人物介绍": "人物介绍"
 }],
 "地名": [{
 "原文地名": "原文地名",
 "现代地名": "现代地名"
 }]
}```
```

### [原始待处理古诗词]

。。。。

## 2.3 古诗词搜索意图理解

### 2.3.1 相关研究工作

在大语言模型之前，传统的方法一般是采用 NLP 的 NER 实体抽取、知识图谱的方式辅助进行意图识别，类似文献[9,10,15]是开发专门针对中文古诗的智能问答技术，都是采用这种方式。

文献[16]在医疗领域采用下面方法进行意图理解，预处理：包括文本清洗、分词、词性标注等步骤，准备数据以供后续模型使用。特征提取：利用 ALBERT 提取问句的语义特征。卷积与池化：使用 TextCNN 对 ALBERT 提取到的特征进行进一步处理，通过卷积层捕捉重要信息并通过池化层减少冗余信息。分类：最终将经过处理后的特征送入分类器进行意图分类。

百度在其专利[17]中公开提出了一种意图识别方法，该方法通过将用户的搜索条件信息与预设的品类词库中的下位标签进行匹配，筛选出匹配标签数量超过特定阈值的目标品类。在此基础上，进一步依据匹配标签的字符长度、权重以及词向量相似度等多维指标综合评估，以最终判定该目标品类是否真正符合用户的搜索意图。

和其它方法不同，文献[18]采用 NL2SQL 技术将用户的自然语言查询转化为计算机可执行的结构化查询语言，从而实现了从自然语言问题到数据库查询结果的直接转换。通过在某市人口数据上的实验验证，结果显示该系统能够显著简化数据应用流程，支持多维度复杂查询，降低了政府部门工作人员操作难度，提升了整体工作效率。专利文献[19]实现了一种基于 NL2SQL 的问答系统，具体如下步骤：获取用户输入的待回复语句，获取目标意图和目标实体列表，若目标实体列表中的时间实体不是预设时间格式，将目标实体列表的时间实体转化为预设时间格式，基于目标意图和目标实体列表，确定检索数据库，基于检索数据库，获取目标表格或目标段落，将待回复语句和目标表格或将待回复语句和目标段落列表输入大语言模型，获取待回复语句的答复语句；使得大语言模型能够更加准确的输出待回复语句的答案。

## 2.3.2 我们的古诗词搜索意图理解方法

根据上述调研，古诗词搜索意图理解主要还是基于传统 NLP、NER 等方法，采用大语言模型辅助的方法暂时还没看到，但受到上述[18, 19]等调研文献启发，我们设计了结构化提示词方法采用大语言模型，首先把古诗词搜索意图转换为数据库查询语言，然后根据查询结果内容进行总结答复。

首先，我们用户搜索的搜索条件进行分类汇总如下：

- 搜索条件原文：例如：锄禾日当午的下一句是什么？廉颇老矣出自哪首诗？
- 搜索条件更多维度：包括作者、人物、地点等等。例如：关于河西走廊的诗歌有哪些？哪些诗是关于珠海的？霍去病在哪些诗歌中出现过？
- 上述搜索条件组合：例如：毛主席诗词中有哪些人物？李白中的诗出现过什么地名？苏轼和李白的古诗中共同出现过的地名是什么？等等

针对上述搜索条件，我们设计了基于思维链和少样本的结构化提示词方法，把古诗词搜索意图转换为 SQL。下面的[转换示例]部分即少样本，通过少样本，可以把大语言模型易混淆的转换更好的纠正过来，只需要加上很少的示例即可，部分复杂的情况，还可以在示例中加上思维链的方式，把整个过程描述完整，从而提高准确率。

### **[任务]**

你是一个古诗词搜索意图分析助手，把[原始搜索请求]转换为 SQL，具体步骤如下：

- 1、首先对原始搜索请求进行语义分析，提取其中的篇名、作者、任务、地点等信息
- 2、根据第 1 步语义分析结果，参考[数据表的定义]，生成 SQL 查询指令；
- 3、生成的查询内容带上 id，便于后续进一步查询；
- 4、请参考[转换示例]。

### **[转换示例]**

#例 1

## 搜索请求：在哪些白居易的诗中包括黄山？

## SQL : SELECT id, title FROM poem\_db WHERE author\_and\_description LIKE '%白居易%' AND locations\_and\_description LIKE '%黄山%';

### **[数据表的定义]**

```sql

```
CREATE TABLE poem_db (  
  id INT PRIMARY KEY AUTO_INCREMENT COMMENT '主键，自增',  
  title VARCHAR(255) NOT NULL COMMENT '古诗文篇名',  
  full_text TEXT NOT NULL COMMENT '古诗文全文',  
  author_and_description TEXT COMMENT '作者及简介',  
  persons_and_description TEXT COMMENT '人物描述',  
  locations_and_description TEXT COMMENT '地点描述',  
  other_description TEXT COMMENT '其它描述'  
) COMMENT '古诗文数据表';
```

```

### **[输出内容及格式]**



```
输出 SQL 预计, 以如下 JSON 格式输出:
```json{
  "SQL": "生成的 SQL",
}
```

[原始搜索请求]
....
```

表 2 是部分搜索意图转换的示例, 可以看出简单的提示词即可达到较好的效果。

| 搜索内容          | 生成的 SQL                                                                                                                                      |
|---------------|----------------------------------------------------------------------------------------------------------------------------------------------|
| 霍去病在哪些诗歌中出现过? | SELECT id, title FROM poem_db WHERE persons_and_description LIKE '%霍去病%'                                                                     |
| 廉颇老矣出自哪首诗?    | SELECT id, title FROM poem_db WHERE full_text LIKE '%廉颇老矣%';                                                                                 |
| 封狼居胥和什么人物相关?  | SELECT id, title, persons_and_description FROM poem_db WHERE locations_and_description LIKE '%封狼居胥%'                                         |
| 李白中的诗出现过什么地名? | SELECT id, title, locations_and_description FROM poem_db WHERE author_and_description LIKE '%李白%' AND locations_and_description IS NOT NULL; |
| 李白的诗中哪些包括庐山?  | SELECT id, title FROM poem_db WHERE author_and_description LIKE '%李白%' AND locations_and_description LIKE '%庐山%';                            |

表 2: 搜索意图转 SQL 结果

我们也在 mongodb 数据库中也做了测试 (需要用更新的提示词, 由于篇幅问题, 就不放在论文中), 也得到了很好的结果

| 搜索内容          | 生成的 mongodb 数据库查询指令                                                                                                                                       |
|---------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| 霍去病在哪些诗歌中出现过? | "people_list": {       \$elemMatch: {         \$or: [           {"原文人名": {\$regex: "霍去病"}},           {"人物介绍": {\$regex: "霍去病"}}         ]       }     }  |
| 廉颇老矣出自哪首诗?    | "content": {\$regex: "廉颇老矣"}                                                                                                                              |
| 封狼居胥和什么人物相关?  | "people_list": {       \$elemMatch: {         \$or:[           {"人物介绍": {\$regex: "封狼居胥"}},           {"原文人名": {\$regex: "封狼居胥"}}         ]       }     } |

|               |                                                                                                                                                               |
|---------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 李白中的诗出现过什么地名? | <pre> "author": "李白", "places_list": {   \$elemMatch: {     \$or: [       {"原文地名": {"\$regex": ".*" }},       {"现代地名": {"\$regex": ".*" }}     ]   } } </pre> |
| 李白的诗中哪些包括庐山?  | <pre> "author": "李白", "places_list": {   \$elemMatch: {     \$or: [       {"原文地名": {"\$regex": "庐山" }},       {"现代地名": {"\$regex": "庐山" }}     ]   } } </pre> |

表 3: 搜索意图转 MongoDB 查询指令结果

2024 S.-T. Yau High School Science Award  
 仅用于2024丘成桐中学科学论文公示

## 3. 系统实现及测试

### 3.1 数据处理

#### 3.1.1 原始数据

我们主要从两个公开的数据中获取原始古诗文数据。

第 1 个是从开源 GitHub[20]，数据库收录 10000 首古文(诗、词、歌、赋以及其它形式的文言文)，近 4000 名作者，10000 名句。其中古文不仅包含原文，还提供注释、翻译、赏析以及朗诵音频地址链接；作者包含了详细生平介绍；名句包含原句以及出处。

第 2 个是从开源 GitHub[21]，数据来源于：古诗词网，包括 21 万首古诗词，以及注释、赏析等信息，包含近 8000 多名诗人以及诗人的介绍、生平等等，同时包含 1600 多个词牌介绍，中国 70 多个朝代解析，和古诗文的近 200 个分类标签。

#### 3.1.2 处理后的数据

我们首先对上述 2 个原始数据进行了合并操作，处理后的数据总共是 22 篇 (220175)，作者信息 8763 条。

然后根据本文 2.2.2 部分的提示词，调用通义千问的 72B 模型补充人物、地点等信息，由于大语言模型调用的费用比较贵，时间比较长，我们设计了一个 50 个人的作者名单的诗篇进行处理。后续我们会持续更新，完成所有 22 万篇的结构化信息抽取工作。

## 3.2 系统实现及测试

### 3.2.1 系统架构

整个系统架构如下：

- 大语言模型采用实验室部署的阿里 Qwen72B
- 数据库采用实验室共享的 MongoDB 数据库
- 系统放在阿里云的一台服务器，操作系统 ubuntu24
- 开发用的代码是 python3
- 前端工具采用 Streamlit，可以直接用 python 代码构建前端

### 3.2.2 Demo 系统

Demo 系统根据架构包括 3 个阶段，第 1 阶段是搜索意图转换，把用户的输入转换为 MongoDB 数据库查询指令；第 2 阶段是执行数据库查询指令，得到结果；第 3 阶段是利用大模型总结回复。具体界面如图 7 所示。

# 古诗词智能搜索

请输入

毛泽东的哪些诗中提到了长沙

15/256

## (1) 搜索意图转换

```
{
 "author": "毛泽东",
 "places_list": {
 "$elemMatch": {
 "$or": [
 {"原文地名": {"$regex": "长沙"}},
 {"现代地名": {"$regex": "长沙"}}
]
 }
 }
}
```

图 7-a Demo 系统的第一阶段部分

## (2) 执行数据库查询

|   | title  | author | content                                                                                                                                     | places_list                                                                                                                                                                                                                                                              |
|---|--------|--------|---------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 0 | 沁园春长沙  | 毛泽东    | 独立寒秋，湘江北去，橘子洲头。看万山红遍，层林尽染；漫江碧透，百舸争流。鹰击长空，鱼翔浅底，万类霜天竞自由。怅寥廓，问苍茫大地，谁主沉浮？携来百侣曾游，忆往昔峥嵘岁月稠。恰同学少年，风华正茂；书生意气，挥斥方遒。指点江山，激扬文字，粪土当年万户侯。曾记否，到中流击水，浪遏飞舟？ | [{"原文地名": "长沙", "现代地名": "湖南省长沙市"}, {"原文地名": "橘子洲", "现代地名": "湖南省长沙市岳麓区橘子洲"}, {"原文地名": "湘江", "现代地名": "湖南省境内的湘江"}]                                                                                                                                                          |
| 1 | 水调歌头游泳 | 毛泽东    | 才饮长沙水，又食武昌鱼。方里长江横渡，极目楚天舒。不管风吹浪打，胜似闲庭信步，今日得宽馥。子在川上曰：逝者如斯夫！风樯动，龟蛇静，起宏图。一桥飞架南北，天堑变通途。更立西江石壁，截断巫山云雨，高峡出平湖。神女应无恙，当惊世界殊。                          | [{"原文地名": "长沙", "现代地名": "湖南省长沙市"}, {"原文地名": "武昌", "现代地名": "湖北省武汉市武昌区"}, {"原文地名": "长江", "现代地名": "长江，中国第一大河流，世界第三长河"}, {"原文地名": "楚天", "现代地名": "泛指长江中下游地区，今湖北、湖南一带"}, {"原文地名": "西江", "现代地名": "西江，长江上游支流，流经四川、重庆、湖北等地"}, {"原文地名": "巫山", "现代地名": "巫山，位于重庆市与湖北省交界处，巫山山脉的一部分"}] |
| 2 | 贺新郎别友  | 毛泽东    | 挥手从兹去。更那堪凄然相向，苦情重诉。眼角眉梢都似恨，热泪欲零还住。知误会前番书语。过眼滔滔云共雾，算人间知己吾和汝。人有病，天知否？今朝霜重东门路，照横塘半天残月，凄清如许。汽笛一声肠已断，从此天涯孤旅。凭割断愁丝恨缕。要似昆仑崩绝壁，又恰象台风扫寰宇。重比翼，和云翥。    | [{"原文地名": "东门路", "现代地名": "不确定，可能指长沙东门，毛泽东与友人分别之地"}, {"原文地名": "横塘", "现代地名": "横塘，位于江苏苏州，古代为水乡，今为苏州高新区的一部分"}]                                                                                                                                                               |

图 7-b Demo 系统的第二阶段部分

### (3) 总结回复

- 毛泽东的《沁园春·长沙》中提到了长沙，描述了他在寒秋时节独立于橘子洲头，眺望湘江的壮丽景色，回忆往昔与同学的峥嵘岁月。
- 《水调歌头·游泳》中也提到了长沙，诗人在饮长沙水后，又食武昌鱼，表达了对长江的豪情壮志，以及对建设宏图的展望。
- 《贺新郎·别友》虽然主要描述的是离别之情，但也间接提到了长沙的东门路，这是毛泽东与友人分别的地方，透露出离别的凄清与不舍。

图 7-c Demo 系统的第三阶段部分

## 3.2.3 测试讨论

我们构建了 30 个测试案例（见附录 1，由于篇幅原因，没有放入所有结果）进行测试。完全正确的结果有 75%。错误大致分为下面几种情况：

- 在搜索意图转换数据库指令时，还是以关键词为主，导致生成的指令语法正确，但语义偏差，例如：搜索“毛泽东”诗词成功，但搜索“毛主席”时失败；搜索“张掖、武威”等成功，但搜索“河西走廊”时，会出现丢失的情况；
- 查询的结果很长，超过了大模型的 token 长度，导致错误。
- 构建数据库阶段时信息没有提取，或者提取错误，例如：食物名称并没有提取。

总结测试结果，总体达到我们的预期，但还存在一些问题，（1）虽然搜索意图理解基于大语言模型，但搜索数据库的时候还是有限的，基于关键词的，因此查全率存在问题。针对这个问题，后续可以用向量库搜索/RAG[22]的方法，采用语义匹配的方法来改进；（2）大语言模型本身存在错位，包括提取的结构化信息不全，以及上下文长度受限等等，后续也可以通过外部数据增强。



## 4. 结论及展望

### 4.1 结论

为了解决古诗词搜索中搜索意图理解、权威性及准确性，我们设计了一种利用大语言模型的两阶段古诗词智能搜索系统，在数据库构建简单通过大语言模型抽取丰富的语义信息；在搜索阶段通过大语言模型辅助理解搜索意图。

我们构建了 22 万条包含了大语言模型辅助抽取的语义信息的古诗文数据库。设计并实现一种基于思维链的古诗词结构化信息提取提示词方法，该方法能够有效利用大语言模型自身能力，结合辅助信息提取古诗词的结构化信息；设计并实现一种将古诗词自然语言查询转换为数据库查询的提示词方法，并根据用户的搜索意图总结回复，从而有效理解用户意图，提取相应的准确数据并作出恰当回应；最终实现了基于大语言模型的古诗词智能搜索系统的原型，并进行了测试，取得了良好的效果。

这些成果表明，利用大语言模型技术可以显著提升古诗词智能搜索系统的性能与用户体验。据我们最大知识能力所知，公开的论文/专利等文献没有检索到用大语言模型辅助智能搜索系统的案例，我们的工作具有较好的创新性。我们正在为我们的工作申请发明专利。

我们再 GitHub 上[23] 公开了本次研究工作的所有数据和代码（包括数据处理代码，Demo 系统代码等等）。

### 4.2 展望

我们的最终目标是实现了一个实用的古诗文智能搜索系统，要达到这个目标，本文还存在不足及很多后续需要完成的工作：

本次项目的时间特别紧，还存在很多不足，目前大语言模型在做结构化抽取时，还存在一些误判，我们发现如果加上已有的注释、翻译等会有明显的提升，我们后续可以不断优化。此外，大模型目前的输入的文本量有限制，导致了测试中的一些错误，后续我们可以采用更大的上下文尺寸的模型进行测试优化。由于对 MongoDB 不熟悉，一些复杂的查询条件目前不支持，下一阶段需要完善。

本次论文没有对大语言模型进行任何微调，只是采用的提示词的方法，后续可以收集样本，采用微调的方法提升整体性能。

到真正应用还有很多工作要做，例如：数据权威性校验，提取的数据的准确率验证，操作界面设计等等。我们计划会一直把这个项目做下去，让更多人用到本系统，感受中国优秀传统文化的传承。

# 参考文献

- [1] 中国哲学书电子化计划, <https://ctext.org/zhs>
- [2] 书格 <https://www.shuge.org>
- [3] 古诗文知识图谱 <https://cnkgraph.com>
- [4] 古诗文网 <https://www.gushiwen.cn>
- [5] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. <https://nlp.stanford.edu/IR-book/>
- [6] Büttcher, S., Clarke, C. L. A., & Cormack, G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. Springer. doi:10.1007/978-0-387-85566-9
- [7] Smith, A. B. (2012). The limitations of search engines in understanding natural language. *Journal of Information Retrieval*, 15(4), 320-332. <https://www.jstor.org/stable/30100543>
- [8] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). A Comprehensive Overview of Large Language Models. arXiv preprint arXiv:2307.06435. <https://arxiv.org/abs/2307.06435>
- [9] 洪婕. 面向中文古诗的智能问答技术研究[D]. 湖北: 华中师范大学, 2023.
- [10] 陈思宇. 基于知识图谱的古诗词学习系统的设计与应用[D]. 重庆: 重庆师范大学, 2021.
- [11] 陕西师范大学. 基于大语言模型的学科知识图谱自动构建与更新方法:CN202410154643. 5[P]. 2024-04-12.
- [12] 中国电子科技集团有限公司电子科学研究院. 一种基于大语言模型的非结构化文本数据知识抽取方法:CN202311073935. 8[P]. 2024-05-14.
- [13] <http://oneke.openkg.cn/>
- [14] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. *Advances in neural information processing systems*, 2022, 35: 24824-24837.
- [15] 谢项. 基于古诗知识图谱的智能问答研究[D]. 湖北: 华中师范大学, 2020.
- [16] 黄星宇. 基于 ALBERT 的医疗问答系统设计与实现[D]. 四川: 电子科技大学, 2022.
- [17] 百度在线网络技术(北京)有限公司. 意图识别方法、装置、设备以及存储介质:CN202211532937. 4[P]. 2023-03-07.
- [18] 姚俊华, 汤代佳. 基于自然语言处理技术的政务智能搜索引擎应用探索 [J]. *软件工程*, 2023, 26(2): 59-62, 58. DOI:10.19644/j.cnki.issn2096-1472.2023.002.012.
- [19] 每日互动股份有限公司. 一种基于 NL2SQL 的问答系统:CN202410098147. 2[P]. 2024-04-26.
- [20] <https://github.com/caoxingyu/chinese-gushiwen>
- [21] <https://github.com/yxcs/poems-db>
- [22] Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey[J]. arXiv preprint arXiv:2312.10997, 2023.
- [23] <https://github.com/eiklnv/Intelligent-Ancient-Poetry-Search-System>

# 致谢

我喜欢旅游，也很喜欢中国古诗文。去年暑假我们全家在河西走廊旅游时候，美丽祁连山景色、嘉峪关的雄伟，霍去病的豪气给我印象深刻，我很想找到相关的诗句，但出人意料的诗，找到了一些，但更多的话很难找到，在百度上总是那几句，点进去链接后收获也不大，古诗文网也只能用关键词搜索，而基于厚望的 ChatGPT 的回答大部分都是在胡说。从那时开始，我就想能否自己做一个古诗的搜索系统，我虽然有一些编程的基础，但具体如何实现，还没有好的想法，一直到今年春节时候和上海交通大学陈凯老师见面后，才有了比较相对可行的思路。陈凯老师当时和珠海一家企业有个合作，我爸爸知道后让我找时间和陈老师交流，陈凯老师听我后，很支持我的想法，并马上建议用大语言模型来辅助完成，并邀请我暑期到上海交通大学实验室实习。

今年 6 月份到上海后，陈凯老师让我首先进行了大模型文献的调研，让我知道了大语言模型的优势及落实，也特别是比较系统的了解了提示词的设计方法。6 月底开始，通过和陈凯老师的多次交流，明确了大模型辅助的两阶段方案来实现古诗文搜索系统。从 7 月份开始，从数据处理，到提示词设计，到后续 Demo 系统开发等等都是我一个人完成。因为之前有比较好的 python 基础，数据处理比较轻松一下，但提示词设计时候花了不少功夫细调，经过陈老师指导和查资料发现结构化的提示词方法后，并逐步理解思维链、少样本的基本原理后，后续工作就比较顺利了。系统实现期间遇到的另外一个问题是 MongoDB 数据的指令问题，到现在我对其还没有完全熟悉，我只是能查询基本的命令，后续这个要加强。

感谢陈凯老师从选题、系统方案设计、提示词设计、最后系统实现等过程的无私帮助，感谢陈凯老师实验室黄韵文同学在调用大模型，在访问 MongoDB 给的帮助。感谢我们高中的吴邵汐老师在选题、论文撰写给予的指导。

也要感谢开源社区，我们用到的基础的古诗文数据都是从开源社区获取的。感谢通义千问，让我能快速阅读文献，即使是英文文献，大大提高了效率。感谢 GitHub copilot，可以自动补全代码，我的第一个 Streamlit 代码就是由它辅助完成的。

# 附录 1

| 序号 | 搜索内容         | 是否正确                 | 生成的内容                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|----|--------------|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1  | 廉颇老矣出自哪首诗?   | 正确                   | 廉颇老矣出自辛弃疾的《永遇乐·京口北固亭怀古》。在诗中，辛弃疾借用了廉颇的故事，表达了自己虽年老但仍渴望为国效力的心情。原文：“凭谁问，廉颇老矣，尚能饭否？”                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| 2  | 李白的诗中哪些包括庐山? | 正确                   | <p>以下是李白的诗中提及庐山的内容总结：</p> <p>望庐山瀑布<br/>原文引用：“日照香炉生紫烟，遥看瀑布挂前川。飞流直下三千尺，疑是银河落九天。”</p> <p>经乱离后天恩流夜郎忆旧游书怀赠江夏韦太守良宰<br/>内容摘要：诗中提到了香炉顶，即庐山香炉峰，如“仆卧香炉顶，餐霞漱瑶泉。”<br/>原文引用：“仆卧香炉顶，餐霞漱瑶泉。”</p> <p>庐山谣寄卢侍御虚舟<br/>内容摘要：诗中详细描绘了庐山的美景，如“庐山秀出南斗傍，屏风九叠云锦张。影落明湖青黛光，金阙前开二峰长，银河倒挂三石梁。”<br/>原文引用：“庐山秀出南斗傍，屏风九叠云锦张。影落明湖青黛光，金阙前开二峰长，银河倒挂三石梁。”</p> <p>望庐山瀑布水二首<br/>内容摘要：诗中再次描绘了庐山瀑布的壮丽景色，如“日照香炉生紫烟，遥看瀑布挂前川。飞流直下三千尺，疑是银河落九天。”<br/>原文引用：“日照香炉生紫烟，遥看瀑布挂前川。飞流直下三千尺，疑是银河落九天。”</p> <p>下寻阳城泛彭蠡寄黄判官<br/>内容摘要：诗中提到了石镜和香炉，即庐山的石镜山和香炉峰，如“石镜挂遥月，香炉灭彩虹。”<br/>原文引用：“石镜挂遥月，香炉灭彩虹。”</p> |
| 3  | 霍去病出现过在哪些诗中  | 漏掉了许多，主要原因是结构化提取时的问题 | <p>杜甫的《自京赴奉先县咏怀五百字 / 自京赴奉先咏怀五百字》中提到了霍去病，用以对比当朝的将领，表达了对理想中英勇善战、为国效力的将领的向往。原文：“况闻内金盘，尽在卫霍室。中堂舞神仙，烟雾散玉质。”</p> <p>李白的《塞下曲六首》中，以霍去病为典，赞扬其英勇善战的精神。原文：“功成画麟阁，独有霍嫖姚。”</p> <p>李白的《胡无人》中，再次提及霍去病，将其作为英勇善战、为国效力的象征。原文：“汉家战士三十万，将军兼领霍嫖姚。”</p>                                                                                                                                                                                                                                                                                                             |

|   |                 |                      |                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|---|-----------------|----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|   |                 |                      | 杜甫的《广州段功曹到得杨五长史谭书功曹却归聊寄此诗》中，虽然未直接提及霍去病，但提到了与霍去病同时代的名将卫青，以及与霍去病相关的典故，反映了对汉代英勇将领的怀念。原文：“卫青开幕府，杨仆将楼船。”                                                                                                                                                                                                                                                                                                                                  |
| 4 | 哪些诗歌的地点和河西走廊相关  | 漏掉了很多，主要原因是结构化提取时的问题 | <p>以下是与河西走廊相关的诗歌地点总结：</p> <p>杜甫的《宿府》中提到了“江城”，虽然泛指长江沿岸城市，但具体可能指成都，杜甫晚年曾居住于此，但与河西走廊无直接关联。</p> <p>杜甫的《牵牛织女》中明确提到了“河西”，这里指的是河西走廊，位于中国甘肃省西北部，是古代丝绸之路的重要通道。直接与河西走廊相关的诗歌为杜甫的《牵牛织女》。</p>                                                                                                                                                                                                                                                     |
| 5 | 汉武帝在哪些诗词中人名出现过  | 错误                   | 数据库返回内容太多，超过大语言模型的输入上限。                                                                                                                                                                                                                                                                                                                                                                                                              |
| 6 | 苏轼的诗词中出现过什么食物名称 | 错误                   | 由于食物名称并没有在结构化提取信息中，导致没有找到任何结果                                                                                                                                                                                                                                                                                                                                                                                                        |
| 7 | 描述张掖的诗有哪些       | 正确                   | <p>根据您的搜索问题“描述张掖的诗有哪些”，以下是与张掖相关的古诗词总结：</p> <p>李白的《秋思》中提到了“燕支”，燕支即今甘肃省张掖市高台县附近，古代为匈奴地，后为汉朝边疆。诗中通过燕支黄叶落的景象，描绘了边疆秋色的苍凉与壮阔。</p> <p>燕支黄叶落，妾望自登台。</p> <p>陆游的《睡起》和《睡起二首 其一》中均提到了“山丹”，山丹县位于甘肃省张掖市。诗中通过山丹石竹的零落，以及孤蝶飞来的情景，表达了诗人内心的寂寥与孤独。</p> <p>山丹石竹俱零落，孤蝶飞来伴寂寥。</p>                                                                                                                                                                           |
| 8 | 和珠江相关的诗有哪些      | 正确                   | <p>和珠江地点相关的诗有以下几首：</p> <p>毛泽东的《七律·长征》中提到了“五岭”、“乌蒙”、“金沙”、“大渡桥”和“岷山”，其中“五岭”、“乌蒙”、“金沙”和“岷山”都与珠江流域的地理环境有关。五岭山脉是长江水系与珠江水系的分水岭；乌蒙山脉位于珠江上游；金沙江是珠江的上游之一；岷山虽非直接关联，但体现了长江与珠江流域的壮丽景色。</p> <p>原文：“红军不怕远征难，万水千山只等闲。五岭逶迤腾细浪，乌蒙磅礴走泥丸。金沙水拍云崖暖，大渡桥横铁索寒。更喜岷山千里雪，三军过后尽开颜。”</p> <p>辛弃疾的《西江月·人道偏宜歌舞》、《西江月 其二 春晚》、《西江月 渔父词》、《沁园春（送赵江陵东归，再用前韵）》、《西江月·渔父词》、《西江月·贪数明朝重九》和《西河·西江水》中多次提及“西江”，这里的“西江”指的是珠江的主干流，流经中国广东省。</p> <p>例如：“西江水。道是西风人泪。无情却解送行人，月明千里。”</p> |



|    |                    |    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|----|--------------------|----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    |                    |    | <p>文天祥的《过零丁洋》中提到了“零丁洋”，这是珠江口外的海域，与珠江地理位置紧密相关。</p> <p>原文：“辛苦遭逢起一经，干戈寥落四周星。山河破碎风飘絮，身世浮沉雨打萍。惶恐滩头说惶恐，零丁洋里叹零丁。人生自古谁无死？留取丹心照汗青。”</p> <p>白居易的《盐商妇》中提到了“西江”，这里的“西江”同样指的是珠江的主干流。</p> <p>原文：“盐商妇，多金帛，不事田农与蚕绩。南北东西不失家，风水为乡船作宅。本是扬州小家女，嫁得西江大商客。”</p> <p>李商隐的《赠司勋杜十三员外》和《时二公从事商隐座主府》中分别提到了“西江”和“二江”，其中“西江”指的是珠江的主干流，“二江”可能指长江与珠江，但具体需结合上下文或历史背景确定。</p> <p>例如：“汉江远吊西江水，羊祜韦丹尽有碑。”</p> <p>陆游的《东斋杂书》中提到了“江”，虽然未具体指明哪条江，但古代中国多江，如长江、珠江等，可能泛指江河。</p> <p>原文：“区芋常愿雨，秧菜常愿晴，吾儿行渡江，晨起愁风生。人生各徇私，夫岂造物情。孰能均此意，万里皆春耕？”</p> |
| 9  | 白居易的诗有和长安相关的吗      | 正确 | 结果太多，超出大模型字数输入限制                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| 10 | 哪些诗中和现代地名海南有关      | 正确 | 略                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| 11 | 哪些诗歌同时出现了秦始皇和汉武帝   | 错误 | 生成数据库查询指令错误                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| 12 | 又食武昌鱼的下一句是什么       | 正确 | 略                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| 13 | 关于黄鹤楼的诗            | 正确 | 略                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| 14 | 和长城相关的诗            | 正确 | 略                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| 15 | 原文中出现了荔枝的诗         | 正确 | 略                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| 16 | 原文中出现了荔枝，并且地点和福建相关 | 正确 | 略                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| 17 | 原文中出现了岭南           | 正确 | 略                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| 18 | 京口是哪个地方            | 正确 | 略                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| 19 | 和镇江、孙权相关的诗歌        | 正确 | 略                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| 20 | 和曹植相关，并且地点包括终南山    | 正确 | 略                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |