ScholarLens：基于RAG的科研文献阅读助手

含义为“学者之镜”，寓意通过它看清文献的重点与本质

开题报告　　自选命题

组长：张敬涵

组员：李晓欧　王美真

前期开题的小组分工：

共同任务：

小组内三人共同进行讨论，确定选题。

任务分配：

张敬涵：整理材料、上台汇报

李晓欧：相关技术调研

王美真：PPT制作

已经完成工作：

1.确定选题

2.了解相关技术，制定初步技术路线

目录-CONTENTS

01

研究背景及意义

Research Background and Significance

01. 研究背景及意义

Research Background And Significance

3000万

据统计，全球每年发表的科技文献数量已超3000万篇，这一数字仍在持续增长。

近些年来，随着科技的快速发展和研究手段的不断提升，科学研究的产出能力呈指数级增长。各学科领域的新成果、新理论、新技术被迅速发表并广泛传播，科研知识的更新速度远远超过以往。这种趋势一方面体现了科研活跃度的提高，但另一方面也带来了前所未有的挑战：**科研人员要在海量文献中精准、高效地获取所需信息变得愈发困难。**

传统的信息检索

核心思路：基于关键词的匹配与筛选

传统信息检索的基本逻辑是将用户的查询转化为关键词，然后在文档库中寻找包含这些关键词的内容。例如，当用户搜索 “人工智能医疗应用” 时，系统会扫描所有包含 “人工智能” 和 “医疗” 这两个词的文档，并返回匹配结果。这种方式依赖于预先建立的索引（如记录每个词出现在哪些文档中的 “倒排索引”），就像图书馆管理员根据书名或目录中的关键词快速定位书籍一样。

传统的信息检索的缺点

“关键词陷阱”

无法区分一词多义（如 “苹果” 可能指水果或科技公司），导致无关结果泛滥。

缺乏深度理解

只能匹配文字表面，无法分析文档内容的逻辑或隐含关系。

难以处理复杂查询

对于需要跨文档整合或逻辑推理的问题，传统技术无法自动分析多篇文献中的数据并提炼结论。

无法适应个性化需求

对不同用户（如科研人员、普通读者）采用统一检索策略，无法根据专业背景调整结果。

01. 研究背景及意义

Research Background And Significance

大模型能力较强，能回答很多问题；
但可解释性差，难以验证回答的真实性。

回答错比没有回答更可怕！！！

02

预期目标

Expected Targets

02. 预期目标
Expected Targets

ScholarLens：基于RAG（检索增强生成）的科研文献阅读助手

针对传统信息检索系统和大语言模型的局限性，我们提出构建一个基于RAG（检索增强生成）的科研文献阅读助手。**提供一个简洁、美观的交互界面**，融合检索系统与大语言模型的优势，从外部知识库的科研论文中快速筛选出相关文献片段，重排顺序后与原有提示词合理结合，并通过语言模型生成准确、连贯的答案，从而帮助科研人员高效地进行论文检索，提升文献检索的精准度和效率。此系统在提升回答质量的同时，也有效缓解了大语言模型可能出现的**“幻觉”问题**，增强了生成内容的可靠性和可验证性。

本系统的研究与开发，不仅有助于解决科研人员在文献检索过程中遇到的实际问题，还能够推动自然语言处理技术在科学文献领域的应用与发展，具有重要的理论意义和实践价值。

03

技术方案

Technical Solution

03. 技术方案
Technical Solution

RAG（检索增强生成）的大致工作流程：

首先用户提出问题时，系统会先将问题转化为向量表示，随后在向量数据库中进行相似性搜索，向量数据库中储存的时外部知识库信息，这些信息往往时大模型原生状态下无法知晓的（例如公司内部产品信息、特定项目的专属资料等）。需要注意的是：纯向量数据库存储的并非大量外部知识库的原始内容，而是经过一系列处理，将外部知识库中的知识转化后所得到的向量数据，当系统检索出相关信息后，将作为问题的上下文相关信息(context)来使用，这些上下文相关信息(context)将被整合进提示词模板中，用户的问题也会被嵌入提示词模板内与上下文相关信息(context)相结合，行出一个全新的提示词。接下来新提示词被发送到大语言模型中，利用其强大的推理和文本生成能力，生成一个答案。

03. 技术方案
Technical Solution

知识库文档的处理

对知识库文档进行分割是一个至关重要的步骤，将文本分割成有意义的片段或“块”的过程叫做文本分块，其质量直接决定落地检索的准确性和生成模型的效果。

片段提取后重排

文本块将由嵌入模型（将高维数据转化为低维向量）转化为向量存入向量数据库，与用户输入信息计算距离后挑出top k，再进行重排使其更贴合用户需求。

智能问答

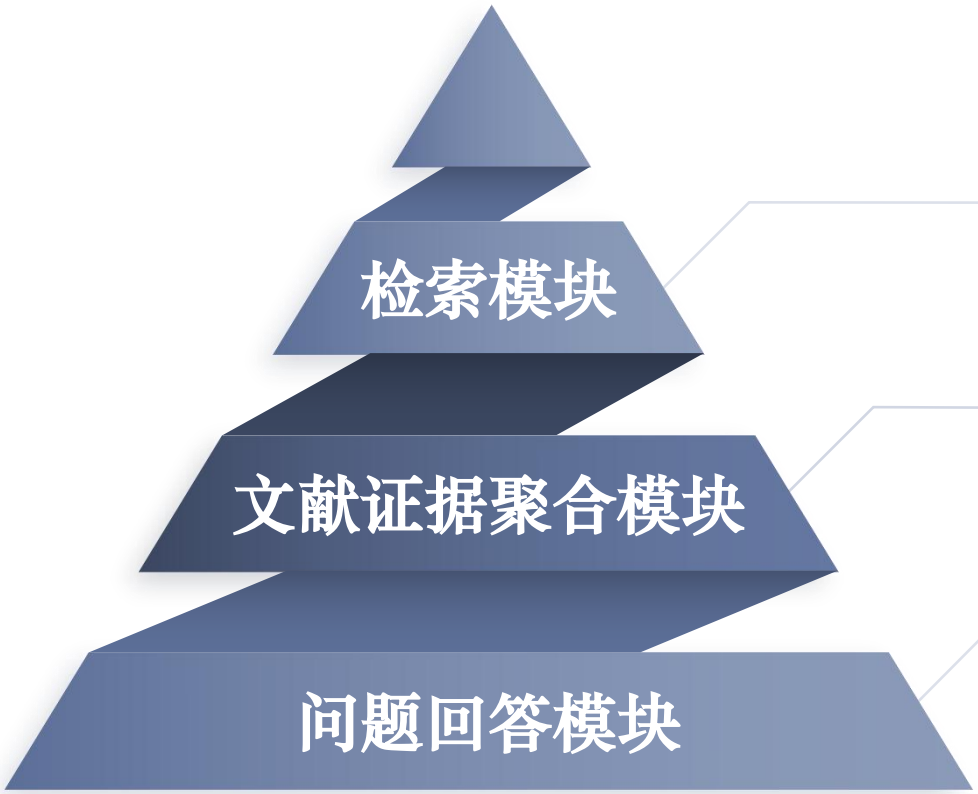
用户可以通过自然语言提问，进过一系列流程后，大语言模型对经过处理后的新问题进行理解和分析，即能结合知识库中的信息，生成准确、连贯的答案。

整合新prompt

对于重排后的最相关的内容，与用户的输入内容相结合，类似于形成新的prompt，再提供给大模型，这样合理地限制其回答内容，减少“幻觉”问题。

03. 技术方案
Technical Solution

使用Python进行功能开发，并会提供问答界面便于交互。



负责接收用户输入的检索内容，据此查询本地文献库。此模块的目标是快速定位与用户查询相关的论文内容，并将这些内容进行向量转换存储，供后续步骤使用。

负责将与查询最相关的论文片段整合到上下文库中。这一步骤是系统生成准确答案的关键，因为它确保了系统能够基于最相关的文献信息来回答问题。

系统的核心输出部分，负责根据累积的上下文提出答案。这一步骤将结合预训练的大型语言模型和上下文库中的内容，生成准确、连贯且带有引用来源的答案。

03. 技术方案
Technical Solution

01 检索模块

文档检索：系统将优先从本地文献库中检索信息，以确保获取更真实相关的文献资源。

高效索引：为了实现高效的相似性搜索，系统将创建重叠的文本片段，并使用文本嵌入模型将这些片段转换为向量表示。这些向量将被插入到向量数据库中，以便快速检索。

错误处理：系统将实施完善的错误处理机制，并记录详细的日志信息，以便后续分析和优化。

02 文献证据聚合模块

向量搜索：系统将使用向量搜索技术从向量数据库中返回与查询最相关的文本片段。

最大边缘相关性搜索：为了增加返回文本的多样性，系统将采用最大边缘相关性搜索算法，确保搜索到的内容既相关又具有代表性。

相关性评分：每个检索到的文本片段都将被输入到一个摘要大型语言模型（LLM）中，该模型将总结文本内容，并提供与问题的相关性评分。

03 问题回答模块

先验判断：在生成最终答案之前，系统将使用一个提问LLM来提供来自预训练LLM的任何有用信息（先验判断）。这些先验判断将作为生成答案的基础。

上下文结合：系统将把先验判断与上下文库中的文本相结合，形成一个全面的答案草稿。

答案生成：最终系统将通过一个回答LLM来生成答案。

03. 技术方案
Technical Solution

模型

可以集成多个模型
包括gpt、claude、gemini等

涉及的知识点

检索增强生成(RAG)
大语言模型的微调
Maximal Marginal Relevance
LLM重排
“先思考-后验证”双阶段推理范式

数据集：

PubMedQA　　2019　闭卷QA测试

它是首个专门针对生物医学研究性问题的 YesNo/Mavbe 问答数据集，旨在考察模型对 PubMed 摘要中定量结论的推理能力。

BioASQ　　2015　领域迁移评估

它是自 2013 年以来每年举办的生物医学语义检索与问答竞赛，其 QA 数据集由多领域医学专家持续扩充，反映“真实临床/科研信息需求”。截至2023 年，已累计4721个带金标的 QA 实例。

03. 技术方案——涉及知识点

Technical Solution

Maximal Marginal Relevance(MMR)

在 通常的向量检索模块中，初始候选往往来自同一篇论文或相似语段：它们与查询的相关性（Relevance）都很高，却缺乏多样性（Diversity）。

- 如果直接把这批“高度同质”的段落送入 LLM，总上下文利用率会被冗余信息挤占，导致证据覆盖面不足、答案片面或遗漏。
- 我们为此在向量 Top-N 结果上应用 MMR 重排：每次从剩余候选里选择既相关、又与已选证据差异最大的片段，从而让送入 LLM 的 k 段文本在语义空间上形成“扇形覆盖”。这种“相关-新颖”并重的策略显著提升了系统检索一次就能找到互补证据的概率。

03. 技术方案——涉及知识点

Technical Solution

LLM重排

执行步骤

1. **向量初筛** → 取 Top-N (默认 20) 并用 MMR 去冗余。
2. **并行调用 summary LLM** 对每个块输出
   - 行内 **摘要**
   - 最后一行独占 **1-10 分数字**
3. **按分值降序排序**，选取前 *k* (默认 8) 写入 *context library*；最高分的块立即返回给 Agent 作为“当轮最佳证据”。
4. 若全部返回 “Not applicable” 或高分块不足 5 条，则 Agent 视为**证据不足**，回到 *search* 或重新 *gather*。

03. 技术方案——涉及知识点

Technical Solution

LLM重排所需提示词：

Summarize the text below to help answer a question.
Do not directly answer the question, instead summarize to give evidence.
Reply 'Not applicable' if text is irrelevant.
Use summary_length.
At the end of your response, provide a score from 1-10 on a newline indicating relevance to question.
Do not explain your score.

<chunk>
Excerpt from citation
Question: <original question>
Relevant Information Summary:

03. 技术方案——涉及知识点

Technical Solution

“先思考–后验证”双阶段推理范式

Ren Ruiyang 等人在 Investigating the Factual Knowledge Boundary of Large Language Models with Retrieval Augmentation（2023）中提出了一种 “先思考-后验证” （Think-First & Verify-Later） 的双阶段推理范式，旨在让大型语言模型（LLM）更好地感知自己的知识边界，并在必要时借助检索补充证据。

维度	传统单阶段 RAG	两阶段推理 (Ren et al.)
判断环节	无	**Think**：能否答？ **Verify**：答得对吗？
触发检索	固定一次	**条件式、多轮**
误答处理	直接输出（可能幻觉）	不确定→检索或放弃
评价指标	仅 QA 准确率	QA + Give-up + Eval-Acc 等边界感知指标

04

工作计划

Work Plan

04. 工作计划
Work Plan

系统框架搭建阶段

明确各模块输入输出与依赖关系
搭建系统基础架构（代理框架、模块接口）
构建本地文献库并完成基本索引
完成搜索模块，包括向量化和检索逻辑实现

5月16日 - 5月25日

证据整合与问答开发阶段

完成文献证据聚合模块的向量搜索、摘要与打分机制
实现问答模块（包括提问LLM与回答LLM调用）
初步测试从问题到答案的完整流程
设计简洁且美观的交互界面

5月26日 - 6月7日

测试优化与总结阶段

系统整体联调，测试不同类型问题的表现
优化模块调用顺序、错误处理、运行稳定性
准备结题展示材料
撰写最终报告

6月8日 - 6月20日

04. 工作计划
Work Plan

姓名	检查点1：5月16日	检查点2：5月25日	检查点3：6月7日	检查点4：6月13日
张敬涵	1.共同讨论确定选题 2.整理资料 3.上台汇报	1.明确各模块输入输出与依赖关系 2.搭建系统基础架构（代理框架、模块接口）	1.初步测试从问题到答案的完整流程 2.设计简洁且美观的交互界面	1.配合测试工作 2.整理最终提交材料 3.上台汇报
李晓欧	1.共同讨论确定选题 2.相关技术调研	1.构建本地文献库并完成基本索引 2.搭建系统基础架构	1.完成文献证据聚合模块的向量搜索、摘要与打分机制	1.系统整体联调，测试不同类型问题的表现 2.撰写最终报告
王美真	1.共同讨论确定选题 2.PPT制作	1.完成搜索模块，包括向量化和检索逻辑实现	1.实现问答模块（包括提问LLM与回答LLM调用） 2.设计简洁且美观的交互界面	1.优化模块调用顺序、错误处理、运行稳定性 2.配合测试工作 3.准备结题展示PPT。

感谢倾听

Thanks For The Listening

组长：张敬涵

组员：李晓欧　王美真