國立臺灣大學社會科學院經濟學系

學士論文

Institute of Department of Economics

College of Social Science

National Taiwan University

Bachelor Thesis

基於檢索增強生成的經濟學術期刊論文問答系統開發

EconLitQA: A Retrieval-Augmented Question-Answering System for Papers in Economics Journals

馬松鐸

Song-Duo Ma

指導教授: 陳由常 博士, 林明仁 博士

Advisor: Yu-Chang Chen Ph.D., Ming-Jen Lin Ph.D.

中華民國 113 年 12 月

December, 2024

摘要

本研究提出一套專為經濟學術研究設計的檢索增強生成問答系統——EconLitQA，旨在解決現有文獻回顧工具在特定領域與語言上的不足。系統整合大語言模型與語義檢索技術，分為三個階段運作：關鍵字提取、語義檢索和答案生成。研究中選取臺灣九個具代表性的經濟學術期刊，針對中文文獻進行處理，並建構了向量資料庫用以檢索。實驗結果說明，EconLitQA 在提取用戶查詢核心概念、檢索相關文獻以及生成準確且語境相關的回答方面表現出色，特別是基於 GPT-4o 的模型在答案的忠實性與關鍵字提取的精確性上表現最佳。未來改進方向包括強化語義檢索的準確性、優化答案生成的簡潔性與可讀性，以及探索多領域擴展與互動功能，進一步提升系統在學術研究中的應用價值。

關鍵字：檢索增強生成、經濟學術期刊、文獻回顧、語義檢索、大語言模型、問答系統

Abstract

This study presents EconLitQA, a Retrieval-Augmented Generation (RAG) system tailored for economics research to address limitations of existing literature review tools in specific domains and languages. The system integrates large language models (LLMs) with semantic search techniques, operating in three stages: keyword extraction, semantic search, and answer generation. The research utilizes data from nine representative Taiwanese economics journals, focusing on Chinese-language literature and constructing a vector database to facilitate retrieval. Experimental results demonstrate that EconLitQA excels in extracting core concepts from user queries, retrieving relevant academic papers, and generating accurate, contextually grounded answers. Among the evaluated models, GPT-4o achieves the highest fidelity in answer generation and the highest accuracy in keywords extractions. Future enhancements include improving the precision of semantic retrieval, refining the conciseness and readability of answer generation, and exploring cross-domain extensions and interactive features, further enhancing the system’s utility

in academic research.

Keywords: Retrieval-Augmented Generation, Economics Journals, Literature Review, Semantic Search, Large Language Models, Question-Answering System

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background

With the rapid advancement of Large Language Models (LLMs), a growing number of applications leveraging these technologies have emerged, significantly enhancing efficiency in addressing routine and repetitive tasks in everyday life. Inspired by this transformative potential, we aim to extend the utility of LLMs to the domain of academic research. One critical phase in the research process is the literature review, which serves as a foundational step in understanding and contextualizing existing scholarly work on a specific topic. A literature review is not only a demonstration of familiarity with the academic discourse but also a means to identify knowledge gaps and opportunities for further investigation. By engaging in comprehensive literature reviews, researchers can deepen their understanding of their chosen subject and produce more robust and impactful contributions to their field.

LLMs can play a pivotal role in optimizing the literature review process by efficiently summarizing large volumes of academic content, extracting key themes, and identifying trends across a broad range of publications. These models can assist in keyword extraction, query expansion, and even the synthesis of findings, enabling researchers to quickly grasp the state of the art in their field. Moreover, by automating repetitive tasks such as citation analysis or clustering related articles, LLMs empower researchers to focus on critical thinking and analysis, ultimately enhancing both the quality and productivity of the research process.

1.2 Research Motivation and Objective

Numerous tools leveraging Large Language Models (LLMs) have been developed to facilitate the literature review process. For instance, Scite[1] provides citation context to enhance literature discovery, while Consensus[2] summarizes insights from scholarly articles to deliver evidence-based answers. These tools, among others, provide significant support for researchers during the literature review process.

Despite their utility, many of these tools primarily focus on English-language content and often rely on extensive databases that encompass literature from diverse fields and languages. This broad scope can pose challenges for researchers seeking focused analyses within specific domains or linguistic boundaries. For example, when employing LLMs to study a particular research area, they may retrieve information from irrelevant fields or fail to prioritize content in the specified language or domain, leading to responses based on unsuitable or misaligned data sources.

To address these limitations, we develop EconLitQA, a localized question-answering system specifically tailored to the field of economics in Taiwan. EconLitQA focuses on Chinese-language articles from nine prominent local economics journals, offering researchers a more targeted and streamlined tool for engaging with domain-specific literature. By narrowing the scope to a defined corpus, EconLitQA seeks to enhance the precision and relevance of literature review efforts within this specialized context.

Chapter 2 Literature Review

2.1 Challenges in Literature Retrieval and the Role of LLMs

A well-conducted literature review is a cornerstone of academic research, enabling scholars to build upon existing knowledge, develop innovative techniques, and propose new theories without redundantly duplicating prior work [3]. To support this critical process, various academic search engines have been developed, including Google Scholar, PubMed, and Web of Science. Each of these systems offers unique strengths and weaknesses.

For instance, some studies suggest that Google Scholar is inappropriate as a primary search tool [4], despite other studies recognizing its comprehensiveness as one of the leading academic search engines [5]. This highlights the challenges of relying exclusively on a single search engine, as limitations such as incomplete coverage, varying levels of specificity, or algorithmic biases can hinder the retrieval of relevant literature. Therefore, it is crucial to explore alternative methods that improve the efficiency and accuracy of literature retrieval, reducing the dependence on potentially flawed search systems.

The rapid advancements in Large Language Models (LLMs) have catalyzed a wide range of applications across diverse fields. For instance, Gao et al. [6] introduced a framework that leverages LLMs to enhance both the interactivity and explainability of recommendation systems. Building on such capabilities, the use of LLMs has been extended to academic contexts, such as recommending and analyzing scholarly literature. Łaniewski and Ślepaczuk [7] further demonstrated that LLMs, in combination with natural language processing (NLP) techniques, can significantly improve the depth and efficiency of literature reviews. Additionally, studies like those by Antu et al. [8] and Scherbakov et al. [9] have explored approaches to optimize the literature review process, emphasizing efficiency and accuracy. These advancements collectively highlight the transformative potential of LLMs in academic research, particularly in conducting comprehensive and efficient literature reviews.

2.2 Applications of LLMs in Economics and the Integration of RAG

The potential of Large Language Models (LLMs) to transform economic research, particularly in the realm of literature review, is widely acknowledged in the literature. Korinek [10] highlights that LLMs can automate cognitive tasks such as idea generation, writing, and background research, significantly enhancing the efficiency of literature reviews. In his 2024 work[11], Korinek further emphasizes that LLMs, with their advanced reasoning and interactive capabilities, can help researchers quickly identify key papers and theories, improving the scope and accuracy of literature reviews. These advancements demonstrate the transformative role of LLMs in supporting and refining literature review practices in economics.

Large Language Models (LLMs) have achieved significant breakthroughs across various domains. However, one persistent challenge is the issue of hallucinations, particularly when the domain-specific knowledge is underrepresented in their training corpora. To address this, Lewis et al. [12] introduced Retrieval-Augmented Generation (RAG), a novel approach that combines a parametric memory, represented by a pre-trained seq2seq model, with a non-parametric memory in the form of a dense vector index (e.g., Wikipedia), accessed via a pre-trained neural retriever. This architecture effectively mitigates hallucinations when the model encounters unfamiliar or sparse knowledge domains.

Moreover, recent studies have explored the application of RAG in the context of academic literature. Thüs et al. [13] investigated how RAG can enhance student engagement with scientific literature, while Mezhlumyan [14] leveraged RAG to improve search capabilities for academic papers by integrating retrieval strategies with LLM-based text generation. These studies highlight the promising potential of RAG in advancing literature review processes and suggest future avenues for its integration in academic research.

Chapter 3 Methodology

3.1 Overview of Methodology

This study proposes a three-stage approach to develop EconLitQA, a Retrieval-Augmented Question-Answering (RAG) system tailored for papers in economics journals. The methodology integrates the capabilities of large language models (LLMs) with semantic search techniques to answer domain-specific questions effectively. The three stages of the methodology are as follows:

1. **Keyword Extraction:** Utilizing a large language model (LLM), keywords are extracted from user queries to encapsulate the core concepts and themes of the questions.

2. **Semantic Search:** The extracted keywords are used to retrieve semantically similar papers from a pre-processed vector database containing paper embeddings, enabling precise selection of relevant literature.

3. **Answer Generation:** Based on the retrieved papers, the LLM generates a concise and accurate response that directly addresses the user's query.

This approach leverages the strengths of LLMs in understanding and contextualizing natural language queries while relying on a vector-based retrieval mechanism to ensure accurate and efficient information retrieval. The flow chart in Figure 3.1 illustrates the overall process.

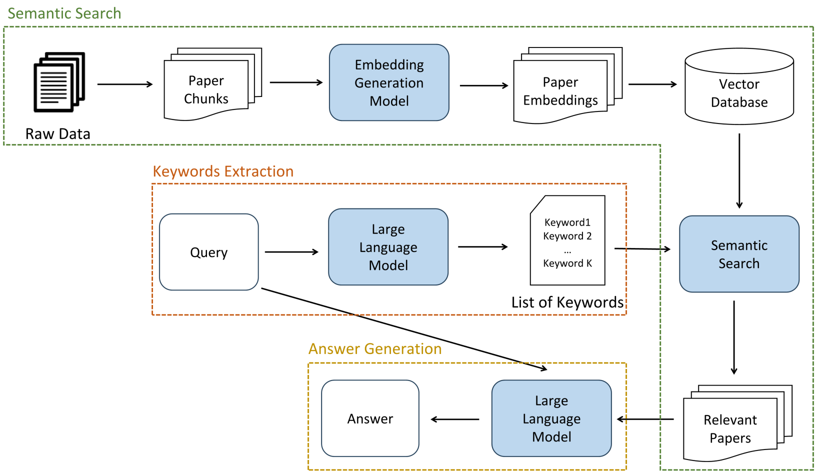

Figure 3.1: Flow chart illustrating the methodology of EconLitQA.

3.2 Keywords Extraction

The first stage of the methodology involves extracting keywords from user queries using a large language model (LLM). This process distills the query into its essential concepts and themes, facilitating an efficient and targeted search in the subsequent semantic retrieval stage. To enhance the relevance and domain specificity of the extracted keywords, a custom-designed prompt is utilized (see Appendix A.1).

The prompt incorporates techniques inspired by few-shot learning, as demonstrated by Brown et al. [15]. By including carefully curated examples within the prompt, the LLM's ability to generate accurate and contextually appropriate keywords is significantly improved. Additionally, the output format is constrained to a bullet-point list of keywords, focusing exclusively on Chinese terms and limiting the number of keywords generated.

These constraints ensure that the extracted keywords are concise and aligned with the thematic focus of the user's query, thereby improving the quality of the retrieval process.

3.3 Semantic Search

The second stage of the methodology involves conducting a semantic search to identify papers relevant to the user query. Initially, the raw data, comprising a collection of academic papers, is preprocessed into manageable chunks to ensure efficient handling and analysis.

Unlike the commonly used approach of recursively splitting text by character, we directly segment each paper into independent chunks. This decision is informed by the unique nature of academic paper content, where information across paragraphs is typically self-contained and does not rely on contextual relationships between sections. This approach preserves the semantic integrity of each chunk, ensuring that the embeddings accurately reflect the distinct information in each segment.

To facilitate semantic search, we employ FAISS (Facebook AI Similarity Search), a highly efficient library for similarity search and clustering of dense vectors, developed by Meta's Fundamental AI Research group [16]. The paper chunks are embedded using the Multilingual E5 Text Embeddings model, as introduced by Wang et al. [17]. These embeddings, representing the semantic content of each chunk, are stored in a vector database created using FAISS.

In the semantic search process, the keywords extracted in the previous stage are matched against the embeddings in the vector database. A similarity threshold is applied to retrieve the top 20 most relevant papers based on their semantic closeness to the user query. This ensures that the search results are both accurate and contextually aligned with the query, forming a robust foundation for the subsequent answer generation stage.

3.4 Answer Generation

The final stage of the methodology focuses on generating an accurate and contextually relevant answer to the user query. This is achieved by integrating the extracted keywords, the retrieved relevant paper information, and the initial user query into a custom-designed prompt (see Appendix A.2).

To mitigate the inherent risk of hallucination in large language models (LLMs), the prompt explicitly instructs the model to focus solely on papers directly relevant to the query during the answer generation process. Additionally, the prompt specifies that if the retrieved papers lack sufficient relevance, the model should respond with "insufficient information." These measures not only reduce the likelihood of generating unsupported or fabricated answers but also enhance the reliability and accuracy of the model's outputs.

By applying these safeguards at the final stage, the system ensures that the generated answers are grounded in the retrieved papers, maintaining accuracy and reliability while minimizing the risk of introducing erroneous or irrelevant information.

Chapter 4 Experiment

4.1 Data Description

The dataset utilized in this study is sourced from Airiti Library, which hosts over 2,000 academic journals and conference proceedings from Taiwan and abroad, alongside master's and doctoral theses from more than 60 universities and colleges. As the most comprehensive collection of academic documents in Taiwan, it provides a valuable resource for this research. To gather the necessary information, we employed web scraping techniques such as Selenium to extract data from nine of the most prominent local journals.

To ensure precision and relevance, we focused on extracting metadata, including the title, authors, affiliations, abstract, and keywords, instead of the full text of each paper. This approach was adopted to maintain consistency and simplify processing. Additionally, to align with the scope of our research on local academic publications, only papers written in Chinese were included. The distribution of papers across the selected journals is summarized in Table 4.1.

In addition, a supplementary dataset was constructed for the comprehensive evaluation of the RAG performance of each LLM. This dataset comprises 100 user queries and their corresponding keywords, designed to assess not only the effectiveness of the

Journal	Number of Papers
中國大陸研究	504
人文與社會科學集刊	406
經濟論文	405
經濟論文叢刊	392
農業經濟叢刊	205
應用經濟論叢	199
經濟研究	187
台灣經濟預測與政策	184
文化創意產業研究學報	154
農業與經濟	149

Table 4.1: Distribution of Papers Across Selected Journals

keyword extraction process but also the quality of answer generation. It includes 10 distinct question templates, each representing a common structure of user queries, populated with 126 diverse topics, effects, aspects, or phenomena related to the field of economics. The dataset ensures a random and balanced distribution across the templates, serving as a benchmark for evaluating the entire RAG process, from capturing core concepts in user queries to generating accurate and relevant answers.

4.2 Results

In this section, we evaluate the performance of EconLitQA by comparing three different Large Language Models (LLMs) on our dataset: Llama-3.1-8B-Instruct [18], developed by Meta; Mistral-7B-Instruct-v0.2 [19], developed by Mistral AI; and GPT-4o [20], developed by OpenAI. The comparison is conducted using five distinct types of user queries to assess the system's capability in addressing various information needs: (1) Analysis of Journals and Publishers, (2) Keyword Analysis and Research Trends, (3) Trend and Research Hotspot Analysis, (4) Author Contribution Analysis, and (5) Literature Review Writing.

For each LLM, the results include the user query, the extracted keywords, and the final answer generated by EconLitQA. Among the three models, GPT-4o consistently demonstrates the most accurate and coherent results. To highlight its performance, the results for each type of query are primarily illustrated using GPT-4o in Tables 4.2 to 4.6, showcasing its superiority in both keyword extraction and answer generation.

Aspect	**Llama-3.1-8B-Instruct**
User Query	假設你是一位經濟學領域的學者，請根據提供的的資料，請幫我找到入學管道這個主題相關的論文中找出最常出現的學術期刊或出版單位，並分析這些來源的影響力。** 請用繁體中文回答 **
Extracted Keywords	入學管道、學術期刊
Generated Answer	根據提供的資料，涉及「入學管道」這個主題的多篇論文主要刊登於《經濟論文叢刊》和《人文及社會科學集刊》。這兩個期刊在台灣的學術界中具有一定的影響力，尤其是《經濟論文叢刊》，其經常發表與經濟學相關的研究，並且涵蓋面廣泛，學術影響力較為顯著。 1. **《經濟論文叢刊》**： - 該期刊多次刊載與入學管道相關的研究，例如探討多元入學政策對學業表現的影響、入學管道與學業退學的關聯等，顯示其在此領域的研究具有一定的深度和廣泛性。 -《經濟論文叢刊》在台灣經濟學術界中具有重要地位，經常被引用和討論，對政策制定和學術研究都有一定的影響力。 2. **《人文及社會科學集刊》**： - 該期刊也刊載了關於入學管道的研究，尤其是從家庭經濟背景和入學制度改革的角度進行分析。 - 雖然其影響力可能不如《經濟論文叢刊》在經濟學界那麼明顯，但在社會科學領域中仍然具有重要地位，常成為政策研究和社會問題分析的資料來源。 總體來說，這兩個期刊在討論入學管道和教育制度方面提供了豐富的學術資源，對相關政策的評估和制定具有參考價值。

Table 4.2: GPT-4o: Analysis of Journals and Publishers

4.3 Evaluation Procedure

To comprehensively evaluate the performance of the LLMs in EconLitQA, we divide the procedure into two main components: keyword extraction and answer generation.

For keyword extraction, the dataset described in Section 4.1 is used. For each of the 100 user queries and their corresponding keywords, a custom-designed prompt (de-

Aspect	GPT-4o
User Query	假設你是一位經濟學領域的學者，請根據提供的的資料，請幫我找到入學管道這個主題最常使用的關鍵字，並推測這些關鍵字可能的研究熱點。** 請用繁體中文回答 **
Extracted Keywords	入學管道、關鍵字研究
Generated Answer	根據提供的資料，有關「入學管道」這個主題最常使用的關鍵字包括： 1. 多元入學 2. 個人申請 3. 繁星推薦 4. 入學考試 5. 學業成績 6. 家庭經濟背景 7. 教育公平 8. 高中畢業學校 這些關鍵字可能的研究熱點包括： 1. ** 多元入學政策的成效 **：研究不同入學管道對學生學業表現的影響，並評估多元入學政策在提高教育公平性和效率方面的成效。 2. ** 家庭背景與入學機會 **：分析學生的家庭經濟背景如何影響他們選擇的入學管道，以及這對社會階級流動的影響。 3. ** 教育公平與效率 **：探討多元入學對教育公平的影響，特別是如何為弱勢群體提供更多進入高等教育的機會。 4. ** 入學管道與學業表現 **：比較不同入學管道（如個人申請、繁星推薦、考試分發）的學生在大學的學業和非學業表現差異。 5. ** 區域和學校特定效果 **：研究繁星推薦等計畫對於不同地區或學校（例如明星高中）學生進入頂尖大學的影響。 這些研究熱點反映了學者們對於入學管道在教育機會、社會流動性以及教育政策效果等方面的廣泛關注。

Table 4.3: GPT-4o: Keyword Analysis and Research Trends

tails in Appendix A.1) is employed to extract keywords from each LLM. To account for minor discrepancies where extracted keywords differ slightly from the true keywords, a similarity-based mechanism is applied. Keywords are considered equivalent if the cosine similarity between the extracted and true keywords exceeds a threshold of 0.9, calculated using the embedding model described in Section 3.3. Based on this comparison, precision, recall, and F1-scores are computed.

For answer generation, we adopt RAGAS, an automated framework for evaluating Retrieval-Augmented Generation (RAG) proposed by Es [21]. This framework includes three evaluation metrics: faithfulness, answer relevance, and context relevance. In our

Aspect	GPT-4o
User Query	假設你是一位經濟學領域的學者，請根據提供的的資料，請幫我以入學管道這個主題的相關文章綜合關鍵字、摘要和來源，是否能觀察到這個主題的研究熱點或趨勢。** 請用繁體中文回答 **
Extracted Keywords	入學管道、研究熱點、趨勢分析
Generated Answer	根據提供的資料，可以觀察到入學管道這個主題的研究熱點和趨勢如下： 1. ** 多元入學制度的影響 **： - 多篇文章探討了多元入學制度對學生學業表現的影響。例如，《經濟論文叢刊》中李維倫等人的研究分析了台大學生在不同入學管道下的學業表現，發現繁星生的表現優於其他管道的學生。而蔡炎涓的研究則指出個人申請與繁星推薦在成績表現上優於考試分發。 2. ** 家庭背景與入學管道的關係 **： - 施琇涵的研究顯示學生的家庭經濟背景與其選擇的入學管道存在顯著關聯，特別是弱勢家庭的學生在繁星推薦入學機率較高。李浩仲等人的研究也指出多元入學對中、低社經地位家庭學生的入學具正面影響。 3. ** 學校特定效果與學生表現 **： - 莊奕琦的研究探討了明星高中對進入頂尖大學的影響，發現明星高中對進入頂尖大學仍具顯著學校特定效果。這顯示出明星高中在入學競爭中的獨特地位。 4. ** 學生滿意度與媒合效率 **： - 銀慶貞等人的研究發現，多元入學制度提升了學生與校系的媒合效率，申請入學者的校系滿意度高於其他管道。 綜合來看，入學管道的研究熱點集中在多元入學制度對學業表現、家庭背景影響、學校特定效果，以及對學生滿意度與媒合效率的影響等方面。這些研究提供了多元入學制度在公平性與效率上的不同視角和見解。

Table 4.4: GPT-4o: Trend and Research Hotspot Analysis

evaluation, we focus on faithfulness and answer relevance, as EconLitQA prioritizes retrieving comprehensive paper information rather than optimizing conciseness.

Faithfulness is assessed by verifying whether the generated answers are grounded in the retrieved paper context. The evaluation involves the following steps:

1. Decomposing answers into shorter, focused statements using an LLM with a custom-designed prompt (Appendix A.3).

2. Verifying each statement against the retrieved paper information using another LLM with a verification prompt (Appendix A.4).

Aspect	GPT-4o
User Query	假設你是一位經濟學領域的學者，請根據提供的資料，幫我觀察哪些作者在入學管道這個主題發表的文章最多 ** 請用繁體中文回答 **
Extracted Keywords	入學管道、作者發表數量
Generated Answer	根據提供的資料，陶宏麟（Hung-Lin Tao）在入學管道這個主題上發表的文章最多。他參與的相關文章包括： 1. 大學入學管道與學業表現－以北部某私立大學為例 2. 由大學多元入學者的個人背景與滿意度評估多元入學的成效 3. 高中職多元入學管道選擇與家庭背景及學習表現之關係－兼論多元入學的公平與效率 4. 從效率與公平評估臺灣的大學入學制度改革 這些文章都涉及入學管道及其相關議題。

Table 4.5: GPT-4o: Author Contribution Analysis

3. Computing the faithfulness score F as:

$$F = \frac{|V|}{|S|}$$

where $|V|$ represents the number of verified statements, and $|S|$ denotes the total number of statements.

Answer relevance evaluates whether the generated answers directly address the posed questions. The procedure involves the following steps:

1. Generating potential questions for each answer using an LLM with a custom-designed prompt (Appendix A.5).

2. Calculating the cosine similarity between the original question q and each generated question q_i using their embeddings.

3. Computing the answer relevance score AR as:

$$AR = \frac{1}{n}\sum_{i=1}^{n} \text{sim}(q, q_i)$$

where n is the number of generated questions, and $\text{sim}(q, q_i)$ represents the cosine simi-

Aspect	GPT-4o
User Query	假設你是一位經濟學領域的學者，請根據提供的的資料，幫我寫一篇有關入學管道的文獻回顧。** 請用繁體中文回答 **
Extracted Keywords	入學管道、文獻回顧
Generated Answer	以下是有關入學管道的文獻回顧： 在台灣，高等教育的入學管道多樣化引起廣泛的學術研究和討論。李維倫等人（2020）使用台灣大學學生的資料，探討個人申請和指考入學學生的學業表現差異，發現學測成績是主要影響因素，申請生在學業表現上優於指考生，但控制學測成績後兩者無顯著差異。此外，繁星計畫的學生學業表現優於其他管道，尤其是來自非明星高中的繁星生（李維倫等，2020）。 陶宏麟等人（2020）研究北部某私立大學的學生，指出繁星生在學業表現上優於其他管道的學生，且繁星生的優勢在數理相關科系不如其他科系明顯，但整體仍優於指考生。這些發現顯示多元入學方案在一定程度上能夠提供優質的學生配置。 施琇涵（2024）則探討家庭經濟背景與入學管道的關係，發現弱勢家庭的學生較不傾向透過個人申請入學，但透過繁星推薦的機率較高，這表明入學管道的選擇與家庭背景密切相關。 郭祐誠和陶宏麟（2013）利用「台灣教育長期追蹤資料庫」，分析高中職多元入學管道的公平性，結果顯示高社經背景的學生更傾向選擇聯考入學，推薦甄試則提供了較聯考更有效率的入學方式。 整體而言，文獻顯示多元入學管道在一定程度上改善了入學的公平性和效率，尤其對於來自非明星高中及弱勢家庭的學生提供了更多的機會。然而，家庭背景仍在某些程度上影響著入學管道的選擇和學業表現，這顯示出多元入學制度在公平性和效率的平衡上仍有改進的空間。

Table 4.6: GPT-4o: Literature Review Writing

larity between q and q_i.

4.4 Evaluation Result

Table 4.7 summarizes the precision, recall, and F1-scores for the keyword extraction task. GPT-4o achieves the highest performance across all three metrics, demonstrating superior accuracy in identifying relevant keywords compared to the other models.

Model	Precision	Recall	F1-Score
Llama-3.1-8B-Instruct	0.771	0.940	0.847
GPT-4o	**0.880**	**0.986**	**0.930**
Mistral-7B-Instruct-v0.2	0.816	0.967	0.885

Table 4.7: Performance Metrics for Keyword Extraction Across Models

Table 4.8 presents the performance metrics for answer generation evaluation. GPT-4o achieves the highest faithfulness score of 0.879, outperforming the other models. In terms of answer relevance, all models show similar performance, with Llama-3.1-8B-Instruct slightly outperforming the others with a score of 0.928, followed closely by GPT-4o at 0.927 and Mistral-7B-Instruct-v0.2 at 0.926. These results suggest that while GPT-4o excels in faithfulness, the answer relevance across the models is nearly identical.

Model	**Faithfulness**	**Answer Relevance**
Llama-3.1-8B-Instruct	0.790	**0.928**
GPT-4o	**0.879**	0.927
Mistral-7B-Instruct-v0.2	0.743	0.926

Table 4.8: Performance Metrics for Answer Generation Across Models

Chapter 5 Conclusion

5.1 Conclusion

In conclusion, EconLitQA has been successfully developed as a Retrieval-Augmented Generation (RAG) system tailored to answer questions based on economics journal articles. By employing a three-step process, the system demonstrates its ability to extract essential topics from user queries, retrieve relevant papers aligned with these topics, and generate answers grounded in the retrieved data. This functionality provides researchers with an effective and precise tool for conducting literature reviews, streamlining the process of extracting meaningful insights from academic sources.

Evaluation results reveal that GPT-4o outperforms other large language models (LLMs) in keyword extraction and answer faithfulness, highlighting its robust capability to ground generated answers in the retrieved context. However, the answer relevance metric shows near-identical performance across all evaluated models, suggesting that this aspect may not serve as a distinguishing factor under the current evaluation framework. These findings establish GPT-4o as the most suitable choice for tasks requiring high factual accuracy and contextual grounding in EconLitQA.

5.2 Limitations and Future Research Directions

EconLitQA demonstrates strong potential as a Retrieval-Augmented Generation (RAG) system for economics journals, but several limitations remain. Its reliance on general-purpose pre-trained language models, constrains its ability to fully capture domain-specific nuances. Additionally, the evaluation framework primarily focuses on keyword extraction, faithfulness, and relevance, overlooking aspects like conciseness, logical coherence, and user experience, which are critical for practical applications. The quality of the retrieved documents further impacts answer generation, as irrelevant or partially accurate contexts can diminish the system's effectiveness.

Future improvements include fine-tuning LLMs on domain-specific datasets to enhance understanding of economics-related queries and expanding evaluation metrics to provide a more holistic assessment of performance. Strengthening the retrieval component with advanced ranking algorithms could improve answer quality, while extending EconLitQA to other academic fields would broaden its applicability. Integrating interactive features, such as user feedback mechanisms, could also make the system more adaptive to diverse research needs.

By addressing these limitations and pursuing these enhancements, EconLitQA can evolve into a more robust and versatile tool, offering researchers a reliable means to explore and synthesize academic knowledge.

References

[1] scite. Ai for research. https://scite.ai/. Accessed: 2024-12-24.

[2] Consensus. Consensus: Ai-powered academic search engine. https://consensus.app/home/. Accessed: 2024-12-24.

[3] Julia H. Littell. Conceptual and practical classification of research reviews and other evidence synthesis products. Campbell Systematic Reviews, 14(1):1–21, July 2018.

[4] Michael Gusenbauer and Neal R. Haddaway. Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed, and 26 other resources. Research Synthesis Methods, 11(2):181–217, March 2020.

[5] Michael Gusenbauer. Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics, 118:177–214, November 2018.

[6] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. Chat-rec: Towards interactive and explainable llms-augmented recommender system. arXiv preprint arXiv:2303.14524, v2, April 2023.

[7] Stanisław Łaniewski and Robert Ślepaczuk. Enhancing literature review with llm and nlp methods. algorithmic trading case. arXiv preprint arXiv:2411.05013, 2024.

[8] Shouvik Ahmed Antu, Haiyan Chen, and Cindy K. Richards. Using llm (large language model) to improve efficiency in literature review for undergraduate research. In Proceedings of the Workshop on Large Language Models at the International Conference on Artificial Intelligence in Education (LLM@AIED 2023), 2023.

[9] Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A. Lenert. The emergence of large language models (llm) as a tool in literature reviews: an llm automated systematic review. arXiv preprint arXiv:2409.04600, 2024.

[10] Anton Korinek. Language models and cognitive automation for economic research. Technical Report 30957, National Bureau of Economic Research, Inc., 2023.

[11] Anton Korinek. Generative ai for economic research: Llms learn to collaborate and reason. Technical Report w33198, National Bureau of Economic Research (NBER), 2024.

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. arXiv preprint arXiv:2005.11401, 2020.

[13] Dominik Thüs, Sarah Malone, and Roland Brünken. Exploring generative ai in higher education: a rag system to enhance student engagement with scientific literature. Frontiers in Psychology, 15:1474892, October 2024.

[14] Artashes Mezhlumyan. Building a retrieval-augmented generation (rag) system for academic papers. Bs thesis, Zaven and Sonia Akian College of Science and Engineering, 2024.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[16] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. arXiv preprint arXiv:2401.08281, 2024.

[17] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. arXiv preprint arXiv:2402.05672, 2023.

[18] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. The llama 3 herd of models, 2024.

[19] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[20] OpenAI et al. Gpt-4 technical report, 2024.

[21] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation, 2023.

Appendix A — Prompt Template

A.1 Keywords Extraction Prompt Template

假設你是一位專業助手，請仔細閱讀以下問題。如果有人要根據這些問題來檢索經濟學論文，請給出三個以下可以幫助找到相關論文的核心關鍵字，並按照編號列出，請不要提供題目沒有提到的關鍵字。

舉例來說：

問題：

假設你是一位經濟學領域的研究者，請根據提供的的資料，分析行為經濟學中決策理論的應用，並根據提供的資料提出可能的研究方向。

** 請用繁體中文回答 **

關鍵字：

1. 行為經濟學

2. 決策理論

問題：

假設你是一位經濟學領域的研究者，請根據提供的的資料，觀察有關最低工資政策對就業影響的研究，並找出相關的研究熱點。** 請用繁體中文回答 **

關鍵字：

1. 最低工資

2. 就業影響

問題：

假設你是一位經濟學領域的研究者，請根據提供的的資料，分析環境經濟學中碳排放交易市場的研究現況，並找出相關文章的研究熱點。** 請用繁體中文回答 **

關鍵字：

1. 環境經濟學

2. 碳排放交易

問題：

假設你是一位經濟學領域的研究者，請根據提供的的資料，分析貿易保護政策對全球供應鏈的影響，並根據資料找出核心研究方向。** 請用繁體中文回答 **

關鍵字：

1. 貿易保護

2. 全球供應鏈

現在換你回答：

問題：

假設你是一位經濟學領域的學者，請根據提供的的資料，(Query)。** 請用繁體中文回答 **

關鍵字：

A.2 Answer Generation Prompt Template

假設你是一位經濟學領域的學者。以下是根據檢索關鍵字「(keywords)」從資料庫中獲取的相關內容，若有跟問題無關之資料，請忽略它：

(retrieved_content)

請根據以上內容回答以下問題：

(query)

若文件中無法找到相關資訊，則請務必回覆「資訊不足」。另外，也不要生成不存在的文獻，僅提供實際存在的學術研究。

** 請務必使用繁體中文回答。**

A.3 Decomposition Prompt Template

根據以下的問題與回答，請從回答中創建一個或多個具體的陳述（每句話拆分為具體簡短的陳述句），並按照編號列出。

問題：(question)

回答：(answer)

A.4 Verification Prompt Template

考慮以下的上下文與陳述句，判斷每個陳述句是否被上下文中的資訊支持。

請為每個陳述句判斷正確與否（請只要回答是或否，不要其他解釋），

並按照編號列出。

(statement)

上下文：(context)

A.5 Question Generation Prompt Template

請根據以下答案產生一個對應的問題

答案：(answer)