Enhancing Career Guidance for Students through Retrieval Augmented Generation in Assistive Software

Akhil Swarop S*, Ganeshkaran M*, Hanish K R*, Hidesh Balaji C U*, Bindu K R*
*Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Coimbatore, India

***Abstract*—Navigating career choices can be overwhelming for students, who often face a flood of information without clear guidance on what's truly relevant to them. Traditional career advice tends to be one-size-fits-all, which misses the mark for students needing insights tailored to their unique goals and the changing job market. This paper introduces an AI-driven career guidance system based on Retrieval-Augmented Generation (RAG). By pulling together real-time data from academic sources and the job market, this system delivers personalized career recommendations that require minimal input from students. Our approach aims to simplify decision-making by providing students with relevant, data-informed options that align with their interests and the latest industry trends, helping them make confident, informed career choices.**

***Index Terms*—RAG, LLM, Career Guidance, Chatbot, Resume Parsing**

I. Introduction

Identifying specialized and developing subjects is critical in a competitive study field for academics who want to make important and novel contributions. Nonetheless, there is a significant obstacle due to the abundance of previously published works and the steady stream of fresh data. In order to locate certain topics that complement their specialized knowledge and interests, researchers frequently find it difficult to sort through this massive amount of data. When academics are trying to focus on newly emerging research possibilities and have specific areas of knowledge, this difficulty becomes even more apparent. We provide a software approach based on Retrieval Augmented Generation (RAG) to tackle this problem. This tool helps academics find unique study topics based on their individual queries by utilizing cutting-edge natural language processing and retrieval techniques. The system offers tailored recommendations by letting users add or remove specific hobbies and talents. By streamlining the research process, lowering information overload, and assisting researchers in concentrating on novel and uncharted territory, this strategy seeks to match their unique interests and strengths with their task. This instrument not only increases the effectiveness of research but also encourages more focused and significant scholarly contributions.

II. Related Work

In recent years, retrieval-based speculative decoding techniques have gained significant attention due to their potential to enhance efficiency in large language model (LLM) applications. For instance, REST [1] introduces a retrieval-based approach to accelerate text generation, using datastore-driven token retrieval, demonstrating speed improvements of up to 2.36x on models like CodeLlama and Vicuna. Similarly, SEED [2] optimizes reasoning tree construction in LLMs for complex tasks, achieving up to a 1.5x speedup by employing scheduled speculative execution strategies across datasets like GSM8K and Blocksworld.

Speculative decoding has also been advanced through adaptive candidate length optimization in SpecDec++ [3], which employs a Markov Decision Process to determine candidate token lengths, yielding up to a 2.26x speedup. Another notable contribution, RAMO [4], addresses the "cold start" problem in MOOCs by integrating Retrieval-Augmented Generation (RAG) techniques, enhancing personalized course recommendations through precision and recall metrics.

Frameworks for evaluating RAG systems, such as RAGCHECKER [5], provide insights into modular retrieval and generation interactions by using claim-based entailment checking and metric designs, enabling fine-grained diagnostics of RAG efficacy. To improve retrieval quality in RAG systems, eRAG [6] leverages document-based relevance scoring, yielding correlation improvements and significantly reducing GPU memory consumption on tasks involving datasets like TriviaQA and HotpotQA.

In question-answering tasks, retrieval augmentation methods like Self-Knowledge Guided Retrieval (SKR) [7] aim to enhance LLMs by adaptively integrating internal model knowledge with external resources, resulting in improved accuracy across commonsense and temporal reasoning tasks. The Uni-Parser model [8] further contributes by offering a unified semantic parsing framework for structured question-answering tasks on knowledge bases and databases, achieving notable Exact Match accuracy on datasets like GrailQA and Spider.

Multi-hop question answering has been advanced through hierarchical frameworks like HiRAG [9], which employs a multi-component retrieval and summarization pipeline to handle complex reasoning tasks. In educational contexts, customized retrieval-augmented chatbots like Professor Leodar [10] aim to reduce misinformation ("Botpoop") while improving student engagement and comprehension, achieving a

97.1% positive user feedback rate.

For documents with multimodal content, a novel approach in [11] introduces a multimodal RAG pipeline that integrates text and image data for more accurate information retrieval and generation. Semantic search advancements, as seen in the work on similarity graphs [12], enable robust retrieval of semantically similar documents, showing notable improvements over traditional keyword-based models.

In the recruitment domain, Resspar [13] presents an AI-driven resume parsing system utilizing NLP and Generative AI to streamline candidate selection, enhancing data extraction accuracy and processing speed.

III. Research Gaps

On conducting a thorough literature survey, we have identified the following research gaps. Current Retrieval-Augmented Generation (RAG) systems have not been widely explored for career path recommendations.

A. Limitations in Prior Studies

Focus primarily on educational applications. Depend on static datasets that lack essential contextual details, such as user skills and educational background.

B. Challenge

Lack of dynamic, personalized data in existing RAG systems. This limitation reduces the effectiveness of RAG in providing accurate career guidance.

C. Future Research Directions

Investigate how RAG systems can use real-time industry trends, psychometric analysis, and individual user profiles. Aim to create personalized, data-driven career recommendations.

D. Optimization Considerations

Efforts are underway to minimize datastore size without compromising the performance of RAG systems, following methods that evolved from traditional decoding to RESTful approaches.

E. Model Architecture

The modified VGG-19 model architecture includes three dense layers and three dropout layers, with the final dense layer using sigmoid activation for binary classification. The model effectively captures features from retinal images relevant to glaucoma diagnosis.

IV. Problem statement

To develop and evaluate a Retrieval-Augmented Generation (RAG) system that provides personalized career guidance by integrating user-specific data, including academic history and psychometric profiles, to enhance the accuracy and relevance of career recommendations.

V. Module Description

A. User Profile and Data Collection Module

This module collects and manages comprehensive user data, including academic history, research interests, skills, hobbies, and psychometric profiles. This module serves as the foundation for generating personalized recommendations. The functions of the module are as follows:

- **Academic Data Collection**:Import transcripts, course records, publications, and certifications.
- **Psychometric Profiling**: Incorporate results from personality tests and cognitive assessments.
- **Interest and Skill Input**: Allow users to specify or update their interests, hobbies, and skills.
- **Data Privacy Management**: Ensure all collected data is stored securely and user consent is obtained.

The module takes user-provided information (academic records, psychometric assessments, interests, skills) as Input and returns a comprehensive and secure user profile for use in recommendation generation as an output

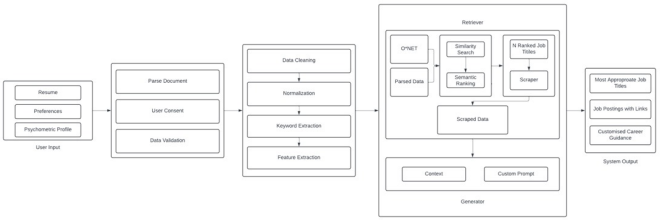

Fig. 1. Proposed Architecture

B. Data Preprocessing and Normalization Module

This module processes raw user data and external datasets to prepare them for effective retrieval and analysis. It normalizes and enriches data to align with standardized formats and taxonomies. The functions of the module are as follows:

- **Data Cleaning**: Remove inconsistencies and errors in user and external data.
- **Normalization**: Standardize data formats (e.g., unify grading scales, course codes).
- **Enrichment**: Map courses, skills, and interests to standardized classifications and ontologies.
- **Feature Extraction**: Identify key attributes from textual data for better matching.

The module takes raw user data, external academic and job market data as Input and returns cleaned and enriched data ready for retrieval and recommendation processes as output

C. Knowledge Base Construction and Management Module

This module builds and maintains a dynamic knowledge base that aggregates real-time academic publications, emerging research topics, and job market trends relevant to various fields. The functions of the module are as follows:

- **Data Aggregation**: Collect data from academic journals, conference proceedings, preprint servers, job postings, and industry reports.

- **Indexing and Storage**: Organize data for efficient retrieval using indexing and semantic embeddings.
- **Continuous Updating**: Regularly update the knowledge base to include the latest information and trends.

The module takes external data sources (academic databases, job market APIs, industry reports) as input and returns an up-to-date, searchable knowledge base of academic and market information as output.

D. Retrieval-Augmented Generation (RAG) Module

This module integrates retrieved information with generative AI models to create personalized and context-rich recommendations for research topics or career paths. The functions of the module are as follows:

- **Contextual Generation**: Collect data from academic journals, conference proceedings, preprint servers, job postings, and industry reports.
- **Personalization**: Use retrieved data as context for generating tailored suggestions.
- **Novelty Emphasis**: Highlight emerging and underexplored areas that offer opportunities for significant contributions.
- **Consistency Checks**:Ensure generated recommendations are coherent and relevant.

The module takes retrieved information and user profile data as input and returns a personalized recommendations for career paths as output.

REFERENCES

[1] Z. He, Z. Zhong, T. Cai, J. D. Lee, and D. He, "Rest: Retrieval-based speculative decoding," *arXiv preprint arXiv:2311.08252*, 2023.

[2] Z. Wang, J. Wu, Y. Lai, C. Zhang, and D. Zhou, "SEED: Accelerating Reasoning Tree Construction via Scheduled Speculative Decoding," *arXiv preprint arXiv:2406.18200*, 2024.

[3] K. Huang, X. Guo, and M. Wang, "SpecDec++: Boosting Speculative Decoding via Adaptive Candidate Lengths," *arXiv preprint arXiv:2405.19715*, 2024.

[4] J. Rao and J. Lin, "RAMO: Retrieval-Augmented Generation for Enhancing MOOCs Recommendations," *arXiv preprint arXiv:2407.04925*, 2024.

[5] D. Ru, L. Qiu, X. Hu, T. Zhang, P. Shi, S. Chang, J. Cheng, C. Wang, S. Sun, H. Li, and Z. Zhang, "RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation," *arXiv preprint arXiv:2408.08067*, 2024.

[6] A. Salemi and H. Zamani, "Evaluating retrieval quality in retrieval-augmented generation," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2395-2400, 2024.

[7] Y. Wang, P. Li, M. Sun, and Y. Liu, "Self-knowledge guided retrieval augmentation for large language models," *arXiv preprint arXiv:2310.05002*, 2023.

[8] Y. Liu, S. Yavuz, R. Meng, D. Radev, C. Xiong, and Y. Zhou, "Uniparser: Unified semantic parser for question answering on knowledge base and database," *arXiv preprint arXiv:2211.05165*, 2022.

[9] X. Zhang, M. Wang, X. Yang, D. Wang, S. Feng, and Y. Zhang, "Hierarchical Retrieval-Augmented Generation Model with Rethink for Multi-hop Question Answering," *arXiv preprint arXiv:2408.11875*, 2024.

[10] M. Thway, J. Recatala-Gomez, F. S. Lim, K. Hippalgaonkar, and L. W. Ng, "Harnessing GenAI for Higher Education: A Study of a Retrieval Augmented Generation Chatbot's Impact on Human Learning," *arXiv preprint arXiv:2406.07796*, 2024.

[11] P. Joshi, A. Gupta, P. Kumar, and M. Sisodia, "Robust Multi Model RAG Pipeline for Documents Containing Text, Table & Images," in *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pp. 993-999, IEEE, 2024.

[12] L. Stanchev, "Semantic search using a similarity graph," in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, Anaheim, CA, USA, 2015.

[13] A. D, K. S, N. E. R, K. K, J. M. S, and R. R, "Resspar: AI-Driven Resume Parsing and Recruitment System using NLP and Generative AI," in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, Coimbatore, India, 2024.