Integrating Fine-Tune Large Language Models With Retrieval-Augmented Generation For Enhanced Career Guidance

Ms. Veda N[1]
Assistant professor
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru, India

Mr. Dhanunjaiah R[2]
(7th Sem, 4th year)
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru, India

Mr. Akash G B[3]
(7th Sem, 4th year)
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru, India

Mr. Hemanth G N[4]
(7th Sem, 4th year)
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru, India

ABSTRACT

Individualized career guidance becomes a very difficult task with the rapid changes going on in various industries as well as in the field of career opportunities. Traditional career counselling methods lack depth and flexibility to cater to students and professionals based on different types of personalization through aspirations, abilities, experience, and market trends. Thus, we offer the development of an AI-Enhanced Career Guidance System based on fine-tuned Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) to make advice extremely personal but very actionable on the basis of career paths.

User inputs such as domain preference, career expectations, location, salary range, experience level, technological skills, education background, aptitude scores, and awareness of current trends are captured in the system. The information is then matched to channel enhanced fine-tuning of LLMs for career guidance tasks; it can give tailored recommendations for three suggested job roles (from which the specific end-user can select) along with detailed guidelines for career progression. To enhance accuracy and the ability to dominate context, a RAG framework features in it, thus allowing the model to retrieve and append latest industry information, related trends, and thereby offer "live," or near real-time recommendations.

This enables the evaluation of potential gaps in skills and what targeted learning can be used to bridge these gaps in future pathways. It accommodates everyone from fresh students who want to jump into their careers to those who are bringing new directions or advancements with their already professional lives. The proposed system aligns individual strengths, aspirations, and market demands to enable better satisfaction with the professional world, improved outcomes of progress in that world, and scalable, adaptable applicability over diverse educational and professional contexts.

The present work demonstrates how powerful AI-based systems can be in transforming the face of career counselling by providing customized, data-driven insights that keep pace with industry and individual profiles over time.

KEYWORDS: *AI-powered career guidance, personalized career pathways, fine-tuned LLMs, Retrieval-Augmented Generation (RAG), skill gap analysis, career progression recommendations*

INTRODUCTION

Besides, provide a personalized career guidance customized to the dreams, capabilities, and experiences of that particular individuals in a highly developed framework where it is a necessity but also a challenge. Designing a system that would have the potential to make very accurate career recommendations would entail analysing user inputs at a very divisive level and generating tons of information. However, the fine-tuning of large language models (LLMs) has built huge

challenges owing to the computational resource requirement and cost. The next critical area would be setting a solid knowledge base through the Retrieval-Augmented Generation (RAG) system for context-aware, updated, and relevant outputs.

Fine-tuning LLMs has widely been accepted as the best means to obtain precise and cutting outputs and would typically be more personalized and pertinent understanding than from an entirely new base model. Fine-tuning requires highly computational resources, which really hikes the cost, thus posing quite a hurdle. The process of bringing RAG into the fold of fine-tuning LLMs would ultimately be successful in alleviating these challenges through efficient retrieval of relevant knowledge according to user queries.

The paper's main contribution lies in the development of the AI-Enhanced Career Guidance System, a system using fine-tuning and efficiency gained through other computationally efficient techniques. The model is built with quantization techniques and knowledge installation strategies via QLoRA for high-grade performance at substantially less computational cost. Also, it integrates with existing models in providing real-time guidance, up to date with current trends in the market, and specific to individual inputs to the system; thus, it has a dynamic knowledge base.

The rest of the paper is organized as follows: "Literature Survey," which examines the various existing methodologies in AI-based career counselling and fine-tuned models; "Proposed Solution," which describes the methods and systems for fine-tuning and integration of the RAG; and "Experiments Setup" along with "Conclusions," which are immersed by implementation and evaluation with findings and directions for the future.

It shows the effect along with work being possible through coupling scalable personalized career advice meeting individual's demands and needing something from the market. It appropriates the usage of efficient methods for fine-tuning LLM using the RAG systems.

LITERATURE REVIEW

Real meagre advancements have occurred concerning AI-supported personalization recommendations in the domain of career decision-making, travel planning, and recommender systems. Prior research on AI techniques, such as natural language processing (NLP), collaborative filtering, and machine learning, to enhance personalization and user experience includes the works of Srinivasan et al. (2023), who study applications of NLP in chatbots regarding how AI interfaces with humans in knowledge and opinions, and eventually what it might mean concerning career guidance applications where personal dialogue may be important to comprehend personalized preferences and needs; as well as Gokul Krishna et al. (2021), and Dasari et al. (2023) who investigate personalized systems in travel planning through collaborative filtering and hybrid methods that emphasize the contribution of user data in context-aware recommendations-the same goes for the need for user-centric career ladder suggestions in the proposed system. However, not all such solutions satisfy the above qualities. Many AI models are built on personalized recommendation systems but do not have enough scalability and computational efficiency for their work. This is particularly true for LLMs fine-tuning-they become expensive and not scalable for even particular tasks when used for developing personalized recommendations. There isn't much literature discussing the balance between the cost computation and accuracy of recommendations, which can be bridged by our research. While LLMs tend to be good at generating customized outputs, it costs in high terms in terms of computation for fine-tuning on-the-fly jobs, such as dynamic career guidance.

New advances in model optimization or quantization techniques like QLoRA (Quantized Low-Rank Adapter) have enabled resource saving without compromising model fidelity. The research proves by the studies under the knowledge-based question answering domain (Yu & Lu, 2022) that knowledge bases and advanced AI models can lead to higher contextual relevance as well as accuracy in personalized career recommendations, plus such improvements for performance. Meanwhile, distillation and retrieval-augmented generation (RAG) advances can make a huge difference in increasing richness and quality in recommendations by supplementing their LLMs with real-time data retrieval, as already proposed by Salhi et al. (2023).

These limitations of current solutions like extensive computations and static to non-dynamic adaptability to real-time trends mark the need for such an improved system suitable for personal career guidance. It is an LLM fine-tuned solution that uses quantization methods on QLoRA and RAG framework to save computational resources while personalizing and contextualizing career recommendations. This will take care of the scalability and cost but present a highly flexible system in tune with current industry trends.

With these innovations, our work will open up a new and practical solution to the problem of personalized career guidance that is computationally efficient and extremely relevant to the user's evolving needs.

METHODOLOGY

This part states the technical approach for developing AI-based career guidance system in terms of specifying how the system will take personalized input from users and process it to yield custom career recommendations. The methodology comprises several major stages such as data collection, model fine tuning, RAG systems integration, and API development.

System Architecture

The system architecture consists of multiple components that collaborate to provide personalized career guidance.The main components are as follows:

Data Collection: The first phase involves gathering data from the user through a comprehensive quiz. The quiz is divided into four sections, and the user provides responses related to the following areas:

- Aptitude Score: Measures the user's analytical and problem-solving skills.
- Domain-Specific Score: Assesses the user's knowledge in their selected career domain.
- Current Trends Awareness: Gauges the user's awareness of current industry trends andinnovations.
- Education and Knowledge Gaps: Evaluates areas where the user may require additional learning or certifications.

These responses are scored for each section, providing the basis for personalized career recommendations.

Dataset Preparation: The data collected from the quiz is processed to ensure that it is ready for fine-tuning. Thesteps involved include:

- Tokenization (using NLP tools such as NLTK)
- Data Cleaning: Removing irrelevant or incorrect information.
- Normalization: Standardizing input values to ensure consistency across different responses.

Fine Tuning the Model: Now you will carry out fine- tuning to the previously trained large language model (transformers) to personalize it for the task of career guidance. Such a step will make the model capable of handling user queries through interpretation and response. Fine-tuning is usually a task that requires a lot of resources; therefore, quantization methods are applied to lessen the computational expenses while still maintaining high performance.

System Integration RAG: Application of Retrieval-Augmented Generation (RAG) integration is the main feature of such a system. This system essentially improves the performance of the model in terms of specific contextualization and correctness of guidance. In this system, external knowledge can access information retrieved by the model using inputs such as job-role specificities, qualifications, trends, etc. for supplementing the final output with the model's own learned knowledge for generating most accurate and livecareer guidance.

Prediction and Personalization: The fine-tuned model, enhanced by the RAG system, processes the user's input data to generate personalized career recommendations. The system produces:

- First Job Role & Guidelines
- Second Job Role & Guidelines
- Third Job Role & Guidelines

The recommendations all steer towards specific practical actions that one can instantly take towards improving the profile - in terms of certifications, upskilling in a few modern technologies, and the respective job roles that one can take, depending on the skills and aspirations. It then provides some estimated salary ranges based on the Location Preference of the user so that it becomes a really relevant and realistic guidance.

API Development: After the career guidelines have been created, the outcomes are available through API. This includes prediction results obtained from the API so that users may receive personal career suggestions in real-time. It is also responsible for the creation of the endpoints from which users access employable career insights based on their input data and preferences.

Performance Evaluation: The model's performance is assessed by evaluating its predictions against real-worlddata. The evaluation includes:

- Model Metrics Evaluation: Analysing the effectiveness of the model in generating accurate recommendations.

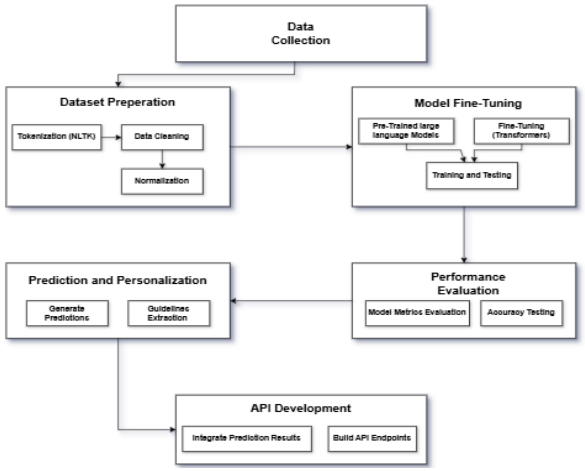

Fig. 1: Methodology

- Accuracy Testing: Comparing the model's predictions with actual career outcomes to measure the quality of the advice.

IMPLEMENTATION

This section outlines the tools, technologies, and process for deploying AI-boosted career advisory systems. The implementation includes fine-tuning of a large language model, state of the art techniques to reduce computational cost, and a fast API to deploy the system for interaction with users. The steps, in summary, are asfollows:
Tools and Technologies

1. **Hugging Face Transformers:** Hugging Face has pre-trained language models used for fine- tuning on the specific task of career guidance. The Hugging Face library can support a plethora of transformer models, which are necessary for processing and generation of human-like text. For fine-tuning, we have used these models on a customized dataset intended to understand the requirements of the career guidance system and to generate personalized recommendations.
2. **Knowledge Distillation and Quantization:** We have designed the fine-tuning process to be less cost-effective in terms of computation for model efficiency improvement regarding knowledge distillation and quantization. Knowledge distillation is transferring knowledge from a better, very expensive model to a smaller model which runs faster while inferences. Quantization is to lower the precision of model weights, which highly shrinks the resource requirements of the model while retaining the performance.
3. **Retrieval-Augmented Generation (RAG) System:** RAG implementation was incorporated into the system for generating more relevant and accurate career guidelines. Therefore, the RAG system can simply apply an external knowledge base to retrieve relevant context and generate more accurate career advice. This makes the system more alive and in-tune with industry standards on certifications and skill requirements.
4. **FastAPI:** FastAPI was the tool that created the backend service for interacting with the fine- tuned model. FastAPI is an excellent choice for this kind of activity because it is very performance-oriented and is one of the easiest for creating RESTful APIs. This application programming interface (API) takes the user input, runs it through the model, and brings back personalized guidance.
5. **React.js & UI Libraries:** The frontend was created with React, which brings a lot of interactivity and friendliness for the users. The UI libraries picked (such as next-UI and shadcn UI) had had their contribution in styling and making more user-friendly. Their functionality covers a lot of prebuilt modules so that the design could be less strenuous in making well-designed and responsive user interfaces through which users can enter their information, view recommendations, and interact with the system.

Implementation Process

1. **Data Collection and Dataset Preparation**: The first step in the implementation was the collection of user input through a comprehensive quiz. The dataset consists of responses to questions about:
   - Career Expectations
   - Technological Skills
   - Experience Level
   - Location Preferences
   - Education Background
   - Aptitude and Domain-Specific Scores

We have done various cleaning, tokenization, and normalization of the data to make it ready for the fine-tuning process. Generally, tokenization is carried out through NLTK libraries, while cleaning and normalization of data are done using personalized scripts.

2. **Model Fine-Tuning:** We fine-tuned both the transformer models developed by Hugging Face, which were initially generic models trained for NLP, on our specific career guidance dataset. Hence, the model was fine-tuned so that the parameters were modified to work on generating career guidance using user input data. As a result, since the whole process of fine-tuning was supposed to be cost-effective, we applied techniques such as knowledge distillation and quantization. The major fact was here that it was possible to transfer some knowledge from a larger, complex model into a smaller, quicker model and that this was possible without losing any progress. Thus, knowledge distortion and quantization were used here.
3. **Development of the RAG System:** To enhance the capabilities of providing precise careers to recommend for, it was decided to incorporate a Retrieval-Augmented Generation (RAG) system into the established system. This RAG

system retrieves pertinent information from a carefully structured external knowledge base about job roles, skills necessary for such careers, qualifications in the form of certifications, as well as real-time industry trends. Thus, the recommendations will not only be personalized but will also be closely tied to real-time developments within the industry.

4. **API Development:** When the model was ready after fine-tuning, FastAPI was used to deploy it as an API. This API helps interact with the model for user input processing, model prediction, and presenting results back to the frontend. This API is designed for high concurrency to ensure users do not have to wait in line even at times of heavy system usage.
5. **Frontend Development:** Generally, to give the user an experience of query and submitting data for quiz results to access personalized career guidelines as predicted through the model, the challenging frontend was also developed using React.js because it avails a faster and dynamic user interface. The frontend also permits a smooth, interactive experience, where users can view recommendations, job areas, and career paths according to their profiles. In terms of libraries, UI design libraries such as Material-UI were used for developing the design through responsiveness, across devices.
6. **Performance Evaluation and Testing:** The authorities undertook a thorough test of the system to ensure that the captured performance parameters reached the desired accuracy level. Subsequent evaluation of model metrics and accuracy testing ensured that recommendations were valid and beneficial in line with real-world outcomes. In addition, performance measured the degree of system speed and scalability. Hence, the API and frontend should be able to accommodate large volumes of users without latency.

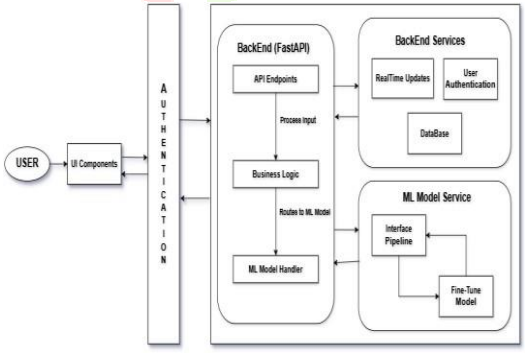

Fig. 2: System Architecture

CONCLUSION

There has been this new way of having personalized career guidance between fine-tuned large language models and Retrieval-Augmented Generation technologies. The major contribution of this research is that AI integrates personalized recommendations for the special needs of the users and profiles, gathered using their input parameters like aptitude, experience, and tech skills. Using techniques like knowledge distillation and model quantization actually decreased costs, making the solution efficient even when scaled up.

An important consideration of the work is that it revolutionizes career guidance systems-as one under these conditions offers personalized recommendations based on the user's aspirations or the reality behind industry trends. It serves to improve job matching in totality, not only helping users identify the right career paths but also directing them on how to upskill and prepare for the job market-in an extremely focused manner.

There is a possibility that future studies may consider improving the knowledge base that will be used by the RAG system for getting even more precise and up-to-the-minute career advice. Also, it could integrate this system to include more dynamic datasets, thus improving the real-time update capability and presenting career opportunities to the users more in line with what industries are doing at the moment. In addition, improved user interfaces and user experiences might also consider more interactive and immersive information to increase accessibility and involvement.

In summary, our work lays a solid foundation for a career guidance system that is extremely efficient and individualized yet scalable. The integration of state-of-the-art machine learning techniques and dynamic knowledge integration reveals the potential of AI in career development and opens new avenues for innovative action in the future.

REFERENCES

[1] R. Srinivasan, M. Kavitha, and Uma S., "Natural Language Processing: Concepts and Applications using Chatbot," *Proc. of the 7th Int. Conf. on I-SMAC (I-SMAC 2023)*, 2023.

[2] G. Krishna M, M. Haseeb, M. Siyad B, P. A. Mohamed Zameel, and S. Vyshnav Raj, "Budget and Experience Based Travel Planner Using Collaborative Filtering," *2021 1st Odisha Int. Conf. on Electrical Power Engineering,*

Communication and Computing Technology(ODICON), 2021.

[3] S. Babu Dasari, V. Vandana, A. Bhharathee, and P., "Smart Travel Planner using Hybrid Model," *Int. Conf. on Intelligent Data Communication Technologies and Internet of Things (IDCIoT 2023)*, IEEE Xplore, 2023.

[4] X. Zhang, Q. Ke, and X. Zhao, "Travel Demand Forecasting: A Fair AI Approach," *IEEE Trans. on Intelligent Transportation Systems*, vol. 25, no. 10,Oct. 2024.

[5] A. Salhi, A. C. Henslee, J. Ross, J. Jabour, and I. Dettwiller, "Data Preprocessing Using AutoML: A Survey," *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, 2023.

[6] Y. Yu and X. Lu, "Research on knowledge base question and answer methods based on the joint subgraph structure of interrogative features," *2022 Int. Conf. on Machine Learning and Knowledge Engineering (MLKE)*, 2022.

[7] U. Ependi A, M. Muzakir, M. Bunyamin, A., D. Irawan, and Fatoni, "Model for Mobile Application Development on Traveling Guide: A General Proposal," *Int. Conf. on Electrical Engineering and Computer Science (ICECOS)*, 2019.

[8] X. Chen, L. Luo, Z. Hu, X. Pei, and Q. Peng, "A Travel Recommendation Method Based on User Personalized Characteristics with Collaborative Fusion Matrix," *2020 5th Int. Conf. on Mechanical, Control and Computer Engineering (ICMCCE)*, 2020.

[9] M. Xu, "Research on Smart Tourism System Based on Artificial Intelligence," *2023 IEEE 3rd Int. Conf. on Information Technology, Big Data and Artificial Intelligence (ICIBA 2023)*, 2023.

[10] L. Benaddia, C. Ouaddia, A. Jakimia, and B. Ouchaoa, "Towards A Software Factory for Developing the Chatbots in Smart Tourism Mobile Applications," *The 4th Int. Workshop on the Advancements in Model Driven Engineering & Software Engineering (AMDE 2024)*, 2024.

[11] Y. Zuo, "Intelligent Tourism Route Planning System Based on Data Mining Algorithm," *2022 Int. Conf. on Artificial Intelligence of Things and Crowdsensing (AIoTCs)*, 2022.

[12] Z. Lin et al., "Quantization Techniques for Reducing Computational Cost in Large Language Models," *Journal of Machine Learning Research*, vol. 23, no. 105, 2024.

[13] Y. Li, J. Xu, and P. Xie, "Knowledge Distillation and Fine-Tuning for Efficient Language Model Deployment," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 7, July 2024