**AI-Powered Career Guidance Chatbot Using Retrieval-Augmented Generation (RAG)**

**Team 02**

**Simran Sattar**  
**Sandeep Raj Katipagala**  
**Nertila Cahani**

**Abstract**

In today’s evolving education and job landscape, students require reliable, personalized, and accurate career advice to make informed decisions. This paper introduces an AI-powered **Career Guidance Chatbot** built on a **Retrieval-Augmented Generation (RAG)** framework that leverages open-source large language models (LLMs) such as **Phi-2**, **GPT-2**, and **Falcon-RW**. The chatbot uses **semantic retrieval** through **FAISS indexing** and constructs context-aware responses using **Chain-of-Thought (CoT) prompting**. A custom dataset of career-related articles serves as the knowledge base, enabling the system to deliver grounded, factual, and coherent answers. We evaluate our approach using BLEU and ROUGE metrics, revealing that retrieval-enhanced generation significantly improves personalization, accuracy, and trustworthiness in AI-driven advisory systems.

**1. Introduction**

Large Language Models (LLMs) like GPT and LLaMA have shown great promise in general-purpose language generation. However, their direct application in sensitive domains like career counseling presents challenges due to **hallucination**, **lack of context-awareness**, and **difficulty in personalizing output**. We address this limitation by proposing a **Retrieval-Augmented Generation (RAG)** framework tailored for student career advice. Our system retrieves relevant documents from a curated knowledge base using FAISS and constructs prompts that guide the generation model to deliver responses grounded in real-world data. By integrating **modular retrievers**, **flexible prompt templates**, and **evaluated generative models**, we ensure that students receive **contextualized, safe, and domain-relevant answers**.

**2. Related Work**

RAG was introduced by [Lewis et al., 2020] as a way to fuse document retrieval with generation, offering a hybrid architecture that grounds outputs in verifiable facts. It has since been applied in tasks like open-domain QA, document summarization, and task-oriented dialogue systems. **Chain-of-Thought (CoT) prompting** [Wei et al., 2022] improves the reasoning abilities of LLMs, especially in step-by-step decision-making tasks. Researchers like Menick et al. (2022) emphasized the importance of **factual grounding** and retrieval for reducing hallucinations. The combination of retrieval, reasoning, and generation, as explored in [Izacard & Grave, 2021], inspires our architecture for trustworthy student guidance.

**3. System Overview**

Our chatbot system follows a **modular and extensible pipeline architecture**:

* **Knowledge Base Construction**: A handpicked collection of ~500 career advice articles and FAQs are preprocessed.
* **Semantic Indexing**: These articles are embedded using **Sentence-BERT (all-MiniLM-L6-v2)** and indexed using **FAISS**.
* **Retriever Module**: For each query, the top-k semantically similar documents are retrieved based on cosine similarity.
* **Prompt Builder**: The retrieved passages are merged into a structured CoT prompt that introduces step-wise reasoning.
* **LLM Inference**: The prompt is passed into one of the LLMs (Phi-2, GPT-2, Falcon-RW) to generate personalized responses.
* **Response Evaluation**: BLEU and ROUGE scores compare generated output with curated references to assess relevance and structure.

The chatbot supports **single-model deployment** or **side-by-side model comparisons**, offering flexibility for user testing and academic benchmarking.

**4. Technical Implementation**

The system is implemented in **Python** and designed for **scalable deployment**. The key technical layers include:

* **Data Processing**: Text cleaning, tokenization, and embedding using sentence-transformers.
* **Indexing**: FAISS is used for vector indexing of document embeddings, enabling high-speed retrieval.
* **Retrieval Strategy**: A hybrid scoring mechanism combining **semantic similarity** and **keyword overlap** improves document relevance.
* **Prompt Engineering**: Prompts follow the Chain-of-Thought structure, guiding the LLM to reason through retrieved information.
* **Model Invocation**: Responses are generated using pre-trained versions of Phi-2, GPT-2, and Falcon-RW, loaded via HuggingFace Transformers.
* **Evaluation Layer**: Uses nltk.translate and rouge-score libraries to compute BLEU, ROUGE-1, ROUGE-L metrics for benchmarking.

**5. Domain-Specific Questions**

Our system is evaluated against 10 commonly asked career queries:

1. How do I start a career in data science?
2. What skills are important for digital marketing?
3. How to transition from engineering to product management?
4. What certifications help in cybersecurity?
5. How can I build a career in entrepreneurship?
6. What is the future of AI in healthcare careers?
7. How to get internships in finance as a student?
8. What tools are used in business analytics?
9. Is UX/UI design a good career in 2025?
10. How important is networking for career growth?

**6. Evaluation and Comparative Results**

We evaluated the outputs from all three LLMs for the same queries using standardized metrics:

* **BLEU Scores**: Assessed the overlap between generated and reference responses. Phi-2 scored highest (avg. 0.68).
* **ROUGE-L Scores**: Measured sequence overlap and linguistic fluency. Falcon-RW and GPT-2 showed moderate scores (0.6–0.75).
* **Response Length & Structure**: GPT-2 generated the longest and most fluent responses but showed a tendency to hallucinate.

Each model performed well in different areas. **Phi-2** excelled in **factual grounding**, **GPT-2** in **elaboration**, and **Falcon-RW** in **balanced coherence**.

**7. Analysis of Model Behavior**

* **Phi-2**: Compact and informative answers, highly precise, best grounded to the retrieved context, but slightly minimal.
* **GPT-2**: Produced verbose, naturally flowing responses with good structure but more prone to introducing speculative or hallucinated content.
* **Falcon-RW**: Balanced accuracy and readability but occasionally lacked deep detail.

**Chain-of-Thought prompting** played a pivotal role in improving logical flow and factual accuracy. The impact of token limitations, particularly on GPT-2, influenced the depth and compression of responses.

**8. Conclusion and Future Work**

Our RAG-based chatbot shows that **retrieval grounding**, when combined with **open-source LLMs**, significantly enhances response relevance in the career guidance domain. The system is modular, reproducible, and educationally impactful.

Future enhancements include:

* **Multi-turn memory** for conversational context
* **Cross-lingual support** for global student access
* **Improved reranking algorithms** for retrieved passages
* **UI Enhancements**: voice commands, confidence scores, and visual summaries

These additions will make the chatbot a more interactive, inclusive, and intelligent advisory tool.

**9. Model Output Comparisons**

**Query:** How do I start a career in data science?

* **Phi-2:** Learn Python, statistics, and machine learning. Start with online courses and practice with datasets on Kaggle. Internships and certifications help gain real-world exposure.
* **GPT-2:** To begin a career in data science, you should build a strong foundation in mathematics, programming, and data analysis. Start by learning Python, explore libraries like pandas, NumPy, and scikit-learn. Engage with real datasets, participate in projects, and consider earning certifications in machine learning and data analysis.
* **Falcon-RW:** Start by learning core data skills including Python, SQL, and data visualization. Move on to machine learning and deep learning. Practice using real datasets. Having a portfolio on GitHub will significantly help with hiring.

**Query:** What skills are important for digital marketing?

* **Phi-2:** Learn SEO, content writing, email marketing, and Google Ads. Understanding analytics tools is important. Certifications from HubSpot or Google help.
* **GPT-2:** Key skills in digital marketing include search engine optimization (SEO), pay-per-click (PPC) advertising, social media strategy, email marketing, and web analytics. Familiarity with tools like Google Analytics, SEMrush, and Canva is beneficial. Strong communication and copywriting are essential.
* **Falcon-RW:** Digital marketing requires knowledge of SEO, social media platforms, paid ads, and content creation. Google Ads and Meta Business Suite are commonly used. Analytical thinking and audience targeting are crucial.

**Observations:**

* **Phi-2** offers brief, high-precision answers.
* **GPT-2** excels in language fluency and depth.
* **Falcon-RW** strikes a middle ground between clarity and completeness.

**References**

[1] Lewis et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, NeurIPS 2020  
[2] Wei et al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, arXiv:2201.11903  
[3] Izacard & Grave, Leveraging Passage Retrieval with Generative Models for Open-Domain QA, arXiv:2007.01282  
[4] Guu et al., REALM: Retrieval-Augmented Language Model Pretraining, ICML 2020  
[5] Menick et al., Teaching Language Models to Support Answers with Verified Facts, NeurIPS 2022  
[6] Kenton et al., Sentence-Transformers: Sentence Embeddings using BERT & RoBERTa, EMNLP 2020  
[7] NLP Progress by Sebastian Ruder, [https://nlpprogress.com](https://nlpprogress.com/)