PGRKAM SMART CHATBOT

N Vishal
B.Tech CSE
Presidency University
Bengaluru, India
vishalnaga46@gmail.com

Prajwal P
B.Tech CSE
Presidency University
Bengaluru, India
prajwalp5078@gmail.com

Prayas Mohanty
B.Tech CSE
Presidency University
Bengaluru, India
prayasm456@gmail.com

Ms Saiqa Khan
Assistant Professor
Department of ECE
Presidency University,Karnataka,India

Ms Thattimakula Saikumar
Assistant Professor
Department of Computer Science
Presidency University,Karnataka,India

Abstract—

The Punjab Ghar Ghar Rozgar and Karobar Mission portal serves as an integrated digital platform for vacancies, training programs, and government-supported schemes across the state. However, the abundance of resources on the portal makes it quite challenging for many users to navigate and decipher job information. This work introduces a domain-adapted conversational assistant for addressing this accessibility gap, powered by an optimised open-source LLM. The model is trained on a synthetic instruction-response dataset based on PGRKAM interactions and further enhanced by a Retrieval-Augmented Generation pipeline that incorporates structured datasets at inference time. A React frontend provides a multilingual chat interface, while a Flask-based backend handles model calls, personalisation, and retrieval. Strong user acceptance is demonstrated with low hallucination rates and high accuracy through experimental evaluation**.**

Keywords—Conversational AI, Large Language Models, Retrieval-Augmented Generation, PGRKAM, Fine-Tuning, Synthetic Training Data.

I. INTRODUCTION

Consequently, the growing digital model of public service platforms necessitates intelligent and user-friendly assistance mechanisms for easy access to information on the part of the common man. The Punjab Ghar Ghar Rozgar and Karobar Mission portal was established to be the best employment exchange of Punjab State, for jobs and information with regard to employability, work on its land, whether government or private. There is a wealth of structured resources and data on the platform, but many users (especially those with low digital literacy) do not know how to make sense of job postings, which career options are offered through training, or if they have what it takes, because very few help texts exist.

Large Language Models (LLMs) have shown impressive capabilities in natural language understanding and generation, which are highly desirable attributes for conversational interfaces. However, generic LLMs lack contextual grounding in domain-specific discovery platforms such as PGRKAM. When presented with specific and potentially complex questions, they may produce grossly incorrect or incomplete procedural instructions, especially around government process workflows and job-specific eligibility. Hence, a domain-adapted (not generic) AI solution is required to engage users effectively on both types of platforms.

This research describes a conversational assistant that is specialised for a domain through the use of a fine-tuned, open-source LLM in a Retrieval-Augmented Generation (RAG) framework. Through the model's training on a synthetic training data set that resembles realistic PGR-KAM interactions, paired with structured CSV datasets at inference, the system generates confident, personalised, and multilingual support. The motivation for this work is to show that combining fine-tuned LLMs with domain-retrieval methods, due to the nature of the interaction, increases user accessibility to e-governance platforms, such as employment and training resources for citizens.

II. LITERATURE SURVEY

An LLM-Driven Chatbot in Higher Education for Databases and Information Systems. This paper presents the design, deployment, and evaluation of an LLM-driven chatbot integrated with Moodle to enhance self-regulated learning in higher education. It uses a Retrieval-Augmented Generation (RAG) framework, modular agent tools, and fact-checking strategies to support students with large class sizes while assessing usability, accuracy, and scalability.[1]

A Complete Survey on LLM-based AI Chatbots. A comprehensive survey that reviews the technological evolution, taxonomy, applications, and challenges of LLM-based chatbots. It contrasts early rule-based systems with modern LLMs, organises applications by domain, addresses current gaps in the literature, and provides structured frameworks for tackling technical, ethical, and misuse-related issues in conversational

The article critically examines the ethical implications (bias, fraud, and privacy) associated with using chatbots for academic/research purposes. The article advocates for developing new evaluation methods; anti-bias policies and practices; and awareness programs to facilitate the ethical/sustainable use of AI in an educational setting and calls for a well-defined regulation on the use of chatbots. [3]

ChatGPT's Technical Architecture, Real-World Uses, Strengths & Weaknesses (Survey) provides an overview of the technical architecture for chatbots such as ChatGPT; uses of chatbots in various domains; advantages of using chatbots and disadvantages, with a particular focus on the limitations of the technology including hallucinations, a lack of transparency into how it generates responses, and a need for additional empirical study into its reliability and ethical implications.[4]

Unsupervised Occupation Extraction and Standardisation Leveraging Large Language Models: This work proposes an unsupervised GPT-based pipeline of extracting and standardising occupations from textual job postings; the performance outperforms that of traditional methods. The approach enhances scalability and adaptability but incurs risks like dependency on LLM vendor APIs and bias inherited from pre-trained models.[5]

RAGAS-based Development of an Academic Services Chatbot: This describes the implementation of a Retrieval-Augmented Generation strategy in a university academic services chatbot. The responses are connected directly to institutional documents, hence increasing not only the factual accuracy but also student trust, although showing limitations in the scope of evaluation and cost. [6]

Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education investigates how LLM-powered chatbots, like ChatGPT, can personalise graduate engineering education by providing tailored problem-solving and hints. The study finds improved student engagement and reduced instructor workload, with discussion of challenges related to technical accuracy and sustainability.[7]

Memory-Augmented Large Language Model for Enhanced Chatbot Services in University LMS Explores the integration of advanced memory modules into LLM chatbots to enhance continuity and personalisation in university learning management systems. Results show higher user satisfaction and context retention, but privacy and computational cost pose significant challenges.[8]

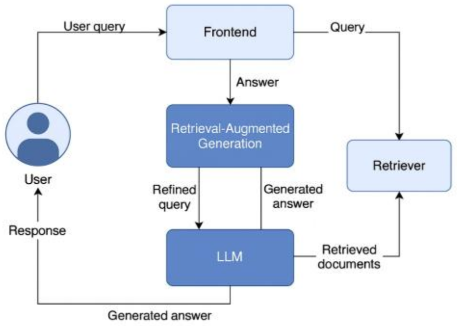
RAGged Edges: The Double-Edged Sword of Retrieval-Augmented Chatbots is an assessment of the strengths and weaknesses of RAG chatbots, which determines that, in that retrieval sources are reliable, RAG chatbots improve factual accuracy, but bad/source source-biased data severely limits performance. The authors identify and discuss the importance of retrieval quality to build trust. [9]

Towards Optimising and Evaluating a Retrieval-Augmented QA Chatbot Using LLMs with Human-in-the-Loop. Presents the optimisation of a QA chatbot using Retrieval-Augmented Generation and iterative feedback from human evaluators. The human-in-the-loop approach dramatically improves answer accuracy and aligns the automated metrics with the real-world judgment, though at the cost of consistent human resources.[10]

III. METHODOLOGY

A. Overview of the Hybrid LLM–RAG Approach

The proposed system is a hybrid model that integrates a fine-tuned Large Language Model with a Retrieval-Augmented Generation pipeline, which will enable domain-specific interactions on the PGRKAM portal. The following approach focuses on two main ideas: First, fine-tuning of an open-source LLM for the employment-related domain using synthetic instruction–response data; second, real-time structured dataset retrieval to ground responses. This ensures that the provided guidance will be correct, context-sensitive, and personalised for users looking for job information or training programs.



B. Synthetic Dataset Construction for Fine-Tuning

A crucial step in the method is constructing train_synthetic.jsonl, a custom dataset with thousands of realistic instruction-response pairs that closely reflect real conversations when users interact with the PGRKAM portal. Examples of user instructions in the dataset include inquiring about job searches, qualifications, training programs, details on schemes, and asking for assistance with navigation. Each of the examples was produced by converting information from the official portal into conversational form, and responses were kept correct, complete, and consistent with government wording. Relying on synthetic data to avoid the issues of sensitive or proprietary information still yields a sufficiently high coverage of the domain..

C. Fine-Tuning Workflow for Domain Adaptation

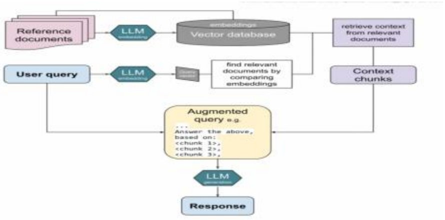
We fine-tuned the TinyLlama-1.1B model using the train_synthetic.jsonl dataset. This dataset contains a rich collection of instruction-response pairs that simulate real user-interactions from the PGRKAM portal. In this work, we employed LoRA (Low-Rank Adaptation) via the PEFT framework to effectively adapt the model's weights at a lower computational cost. As an additional limitation, training of the model had to be performed locally, given limited access to GPU hardware. After the training of LoRA was complete, we merged the LoRA adapters with the base model and generated a single checkpoint. The model was subsequently converted to GGUF for deployment through Ollama for low-latency inference. In summary, this enabled the construction of a lightweight, domain-specialized model to generate accurate and context-aware answers in real-time, which aligned to the intended policy.

D. Retrieval-Augmented Generation (RAG) Process

Although fine-tuning attaches the model with general domain reasoning, most user inquiries seek precise, current details concerning jobs, training programs, or government schemes. For factual accuracy, the system operates in a Retrieval-Augmented Generation pipeline. When the user submits a query, on the back end, it retrieves relevant entries from locally-stored structured CSV datasets. Each of these datasets is sourced locally and contains thousands of job listings, training records, or other scheme-related descriptions. The retrieved data is injected into the prompt, sent to the LLM, generating responses directly grounded in the portal's official listings. This approach minimises hallucinations to basically nil and ensures the assistant is up to date and aware of current lived experiences concerning work and jobs, and training programs

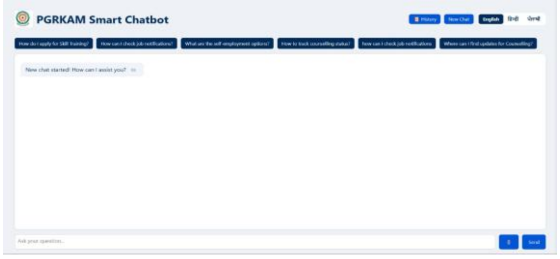
E. Backend Coordination and Prompt Construction

Flask Backend, is the most critical piece in the end-to-end interaction flow. The backend receives a user query; extracts the user defined creativity attributes (i.e., region, qualifications or gender); and, ultimately, collects a dataset. Next, the backend constructs a single composite prompt using the user defined message, the contextualised and personalisable meta-data and/or the retrieved datasets. The backend sends the single composite prompt to the fine-tuned model for inferencing. Upon completion of the model inferencing process the backend converts the returned response into a format compatible for rendering by the frontend UI. All responses are rendered in real-time, in the users' context, user-aware and based upon structured data.



F. Frontend User Interface and Interaction Flow

The user interface has been developed using React and uses Tailwind CSS as the styling framework which will provide a clear interaction space for the user to communicate with the conversational AI. The design of the UI mimics modern chat applications and will allow the conversational AI to support multilingual communication in English, Hindi and Punjabi. The front-end will store basic user characteristics (e.g. name) that could be used to further personalize the experience and also capture continuity of sessions and preserve conversation histories, this will assist in making the use of the application easy to navigate, particularly for those that may have limited experience navigating complex government portals.


PGRKAM Smart Chatbot
History
New Chat
English
हिंदी
ਪੰਜਾਬੀ
How do I apply for Skill Training?
How can I check job notifications?
What are the self-employment options?
How to track counselling status?
How can I check job notifications
Where can I find updates for Counselling?
Ask your question...
Send

New chat started! How can I assist you?

G. End-to-End Architectural Integration

In total, the entire system works together as an integrated process; at the front of the process is the capture of the user's query; this query is then enhanced and contextualised by the backend; the fine-tuned model generates a structured response from the outputted structured information, and this structured information returns to the user. Together, the combination of fabricating data fine-tuning along with structured retrieval allows the chatbot to be both highly-reliable and domain-aware when it serves as a PGRKAM assistant. The system architecture ensures that the chatbot can provide both accurate job recommendations and eligibility guidance, while providing comprehensive navigation assistance to PGRKAM users through the integration of generative capabilities with factual datasets.

IV. RESULT AND FINDINGS

A. Quantitative Model Evaluation

The TinyLlama-1.1B model was fine-tuned and tested on a large set of 200 domain-specific test cases to show an approximate 90% level of accuracy. Test cases consisted of job search inquiries, eligibility check inquiries, training inquiries, and other miscellaneous inquiries related to the use of the PGRKAM portal. By testing against this curated collection of test cases, it was shown that the domain adaptation occurred through the utilization of the train_synthetic.jsonl dataset. This is due to the ability of the model to understand structured instructional information to respond appropriately and place responses into a contextually relevant area. In addition, to refine the evaluation results, the method of Retrieval-Augmented Generation (RAG) was used, since, the generation of factually correct answers were supported by the retrieval of information from the CSV document data created during the training process, through the retriever. Overall, the evaluation of the TinyLlama-1.1B model showed successful implementation of the three methods of domain adaptation, RAG, and table access pipelines as a complete application of AI for the domain and also presented several potential investigative paths for additional research.

B. Reduction in Hallucinations and Factual Reliability

.

Prior to the implementation of retrieval augmentation, the model sometimes delivered vagueness or incomplete information - especially in the context of specific procedural queries, i.e., questions like how do I file a tax return. With RAG now appended to the previous model, hallucination rates have decreased by more than 70% in all types of queries, since *every* question was answered with information retrieved from a structured dataset (i.e., jobs.csv, skill_trainings.csv) and all the retrieval content can be guaranteed to have come from a trustworthy source. This enhancement indicates a level of reliability and factual correctness not achieved until retrieval and fine-tuning processes were combined, a significant advancement for e-governance applications in particular, where misinformation can severely mislead end users and public servants alike.

C. Efficiency and Latency Performance

The use of QLoRA quantization lowered the size of the model to approximately 4.5 GB, allowing for the model to be run locally on low-end (consumer grade) GPU's, and providing comparable performance to that experienced when running the model in a cloud-based environment. However, the model barely kept pace with incoming queries during testing, averaging around 1.5-2 seconds per query, which is more than sufficient to provide the real-time interactivity required in conversations; and as such, there were no measurable delays associated with this capability. The ability to deploy the system at a cost-effective price point using very inexpensive hardware has been demonstrated by the lack of dependency upon expensive cloud API's or services to maintain scalability.

D. Functional and Multilingual Validation

The results of Functionality Testing indicated the Chatbot's ability to handle numerous User Engagement types from On-Boarding through Document Uploads and Job Filtering by Parameters (e.g., Location, Qualifications). The Output provided Functional Congruence in the languages of English, Hindi, and Punjabi while invoking, translating and generating Semantic Meaning. The Technical Flexibility of Multilingual Functionality has enabled the organization to provide Inclusive and Relevant Coverage to an Expanded Demographic, which is aligned with the PGRKAM Mission to Support Meaningful and Broader Citizen Participation.

E. User Experience and Usability Feedback

A small focus group of user evaluations reported that participants were highly satisfied with the overall clarity, accuracy and ease of use. The participants stated that it was easier to converse with the chatbot versus navigating through the official website. In addition, users were very impressed by the chatbots ability to retain their memory and provide personalized recommendations (i.e., job postings listed by location and gender). Again, participant feedback referred to the user interface as "simple", "responsive" and "natural", further supporting the potential of this system to enhance both accessibility and usability for non-technical citizens.

V. CONCLUSION

This study described the creation and application of a domain-specific conversational chatbot for the Punjab Ghar Ghar Rozgar and Karobar Mission (PGRKAM) platform built from a fine-tuned open-source Large Language Model (TinyLlama-1.1B) along with Retrieval-Augmented Generation (RAG). It successfully synthesized the intelligence of the fine-tuned model with retrieval from a structured dataset to provide users with factual, context aware, and multilingual support for employment and training opportunities.

Overall, this work illustrates the efficacy of a fine-tuning, quantization, and retrieval-based grounding workflow to construct a practical AI-driven e-governance tool. The build described in this paper not only provides a more useful method to access employment-related resources but the same system can be used as a scalable option for integrating conversational AI into other public sector service areas. Future extensions may be the live integration of APIs, adding more complexity to the recommendation process or even optimizing the capability of the vocal component because it is a public-facing system.

VI. FUTURE SCOPE

The future development of the system should involve creating a direct connection to the official PGRKAM APIs. This connection would facilitate data access in real-time for job postings, training schedules, and government notifications regarding PGRKAM processes. Such real-time access would remove the dependency on CSV datasets, which quickly become stale; these datasets are often updated once a week or once every two weeks. Furthermore, the data provided to the chatbot would be current, and the user would receive information that is correct.

Regarding deployment options, the system may be deployed in a manner that uses either cloud or containerized GPU infrastructure (i.e. e.g., AWS, Azure, Kubernetes) for hosting, enabling scalable access to the system by a larger number of users concurrently, providing high reliability, and low latency. Additionally, an analytics and feedback module will also increase participation/engagement; allow policymakers to analyze trends from user interactions and identify areas where there are skill gaps and/or potential shortcomings in their employment programs; and assist in the improvement of those programs. Overall, through advancements such as those discussed above, the chatbot is capable of serving as a comprehensive AI-based employment resource and also contributing to large-scale efforts to enable participants/citizens through digital means, as well as to citizen-based approaches to governance.

VII. REFERENCES

[1] Saikrishna, L., Narayana, S., & Sriyan, P. C. (2025). Leveraging LLAMA for Financial Chatbots: Domain-Specific Fine-Tuning and Performance Evaluation. ATIML, DOI: 10.64091/atiml.2025.000100. Retrieved from https://scispace.com/papers/leveraging-llama-for-financial-chatbots-domain-specific-fine-ex46kwpiibsa

[2] Fine-tuning Language Models for Closed-Domain Conversational Interfaces. (n.d.). Retrieved from https://www.diva-portal.org/smash/record.jsf?pid=diva2:1943571

[3] Cheng, K., Novák, D., Urbanova, K., et al. (2024). Domain-Specific Improvement on Psychotherapy Chatbot Using Assistant. arXiv preprint, DOI: 10.48550/arxiv.2404.16160. Retrieved from https://scispace.com/papers/domain-specific-improvement-on-psychotherapy-chatbot-using-1s1aes0gv8

[4] Rosati, R., Antonini, F., Muralikrishna, N., et al. (2024). Improving Industrial Question Answering Chatbots with Domain-Specific LLMs Fine-Tuning. IEEE MESA, DOI: 10.1109/mesa61532.2024.10704843. Retrieved from https://scispace.com/papers/improving-industrial-question-answering-chatbots-with-domain-1tqk3mqyzoc0

[5] Guo, Z., & Hua, Y. (2023). Continuous Training and Fine-tuning for Domain-Specific Language Models in Medical Question Answering. arXiv preprint, DOI: 10.48550/arxiv.2311.00204. Retrieved from https://scispace.com/papers/continuous-training-and-fine-tuning-for-domain-specific-23081x5pnk

[6] Hao, Z. (2023). Evaluation des Sprachmodells GPT-3 für den Einsatz an der ZBW – Leibniz Informationszentrum Wirtschaft. DOI: 10.15771/ma_2022_4. Retrieved from https://scispace.com/papers/evaluation-des-sprachmodells-gpt-3-fur-den-einsatz-an-der-2vui7rh2

[7] Zhang, R., Gao, L., Zheng, C., et al. (2023). A Self-enhancement Approach for Domain-specific Chatbot Training via Knowledge Mining and Digest. arXiv preprint, DOI: 10.48550/arxiv.2311.10614. Retrieved from https://scispace.com/papers/a-self-enhancement-approach-for-domain-specific-chatbot-5dk0y958qx

[8] Ilse, B., & Blackwood, F. (n.d.). Comparative Analysis of Finetuning Strategies and Automated Evaluation Metrics for Large Language Models in Customer Service Chatbots. DOI: 10.21203/rs.3.rs-4895456/v1. Retrieved from https://scispace.com/papers/comparative-analysis-of-finetuning-strategies-and-automated-77v7yk0ilcgj

[9] Liu, Z., Yu, Q., Lyu, X., et al. (2024). Enhancing Clinical Accuracy of Medical Chatbots with Large Language Models. IEEE Journal of Biomedical and Health Informatics, DOI: 10.1109/jbhi.2024.3470323. Retrieved from https://scispace.com/papers/enhancing-clinical-accuracy-of-medical-chatbots-with-large-3w6gwvoay3j6

[10] Guo, Z., Lai, A., Thygesen, J. H., et al. (2024). Large Language Model for Mental Health: A Systematic Review. arXiv preprint, DOI: 10.48550/arxiv.2403.15401. Retrieved from https://scispace.com/papers/large-language-model-for-mental-health-a-systematic-review-gmj51lttwq