iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70999

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

LLM in Public Service Applications: CPGRAMS

Sudarshan Goyal[1], Rimjhim Singh[2]
Student at Department of Computer Science & Engg, MIT School of Computing, MIT Art, Design and Technology University, Loni Kalbhor, Pune, India

Abstract: ***CPGRAM Portal is a government platform introduced for the citizens to file their grievances related to specific ministries. With the increasing population, it's difficult to handle and manage a large number of complaints. Also manual addressing to basic user queries is also time taking and requires a lot of labour. This project proposes an AI/ML driven chatbot designed to handle users will ease the grievances filing process. This chatbot is powered by Large Language Models(LLMs) which can address user queries efficiently. This chatbot will be made for specific ministries, guiding users step by step in filing grievances and also address frequently asked questions. Use of machine learning enables the chatbot to continuously learn and adapt to user queries which will help in increasing the response time and accuracy. So this will help to reduce the manual support, enhance user satisfaction and optimize the overall experience of the user. This solution focuses on addressing common queries faced by user, timely and efficient support for grievance resolution in ministry-specific context.***
Keywords: Large Language Models (LLMs),Cloud Computing, AI for Public Administration, Digital Government Solutions, RAG, CPGRAMS Portal.

I. INTRODUCTION

CPGRAMS stands for Centralized Public Grievance Redress and Monitoring System. This forum helps citizens in filing grievance complaints related to specific ministries. With the increasing population, the number of complaints are increasing leading to difficulty in managing and handling these complaints. Resolving each user query quickly is also becoming a problem.

These problems can be addressed by introducing AI/ML. This will help to ease up the process and improve the query handling time. AI/ML models can continuously learn from ongoing user handling processes and improve their accuracy with training, this will help them to enhance their performance and enable them to address the queries and look for the patterns in order to improve efficiency, user experience and ultimately give citizens satisfaction. LLMs introduction to the chatbot is like a whole library with a special section of related information stored which can be selected when and where needed. Such a chatbot with a LLM, hosted on a cloud platform, will greatly improve the user experience. This chatbot would be developed to help users in particular ministry areas by providing real-time support, assisting with the grievance submission process, and responding to often requested inquiries.

By decreasing manual labor will help in increasing response time and also guarantees a more effective grievance filing procedure. Cloud resources are scalable as well as dependable and AI/ML usage allows the chatbot to learn and grow over time, this will lead to the ever-changing landscape of government services. The suggested AI-driven chatbot seeks to optimize the CPGRAMS platform. This chatbot will revolutionize the citizen-government relations by utilizing these technologies, making them more effective, accessible and responsive.

II. LITERATURE SURVEY

The research study clearly argues a key transformative role for AI in government services. Importantly, it shows how AI can improve efficiency in operations, decision making, data analysis, and data processing to develop evidence-based policy. AI also has the capacity to reduce red tape, improve enforcement, and improve strategic planning within government. Despite the benefits, there are challenges - data privacy, ethics, and public trust - that must be addressed. In this way, AI needs to have solid policies, transparent in their frameworks for accountable and responsible use to fully leverage AI's value.[1] The research study delves into the application of large language models (LLMs) within a national security context and how they can revolutionize workflows related to making decisions, analysing data and operational readiness. Examples of current use cases include wargaming, summarisation, and training services. However, challenges are quite clear, including hallucinations and the risk of adversaries using the data for ill. This is important to highlight in the study - whatever other benefits LLMs can bring about, safeguards must also be in place for the responsible use of these technologies. The study highlights the potential of LLMs to supercharge processes, but that caution must always be a constant checkpoint to meet objectives of accuracy and reliability and responsible use in defence contexts.[2]

The paper reviewed considers the impact of AI in public administration and its potential to improve the provision of public services, as well as decision making in public administration. If Ai needs to be used in public administration then careful measures need to be taken in terms of bias, accuracy, metrics and all other important things. The paper notes that ethical and legal issues will be paramount when considering this emerging technology, and governance responses need to be developed. The paper identified explainable AI as a key component to deliver accountability and trust in public service applications.[3]The paper reviewed e-government data management with a focus on issues of citizen-administration dialogue, privacy, and standardization when providing case management of the e-government application. By utilizing natural language understanding and case-based reasoning for some elements of the application, which includes an experimental system that allows support for decision making in case assessment, while filtering out routine cases for the administration or experts to examine. The authors conclude that while monitoring data actively was possible, human communication was preferred for unresolved cases. The authors argue any unresolved legal and technical issues must be addressed with the aim of improving citizen satisfaction and improving the efficiency of the systems.[4]The paper examines e-government data management, focusing on challenges in citizen-administration dialogue, privacy, and standardization. Using natural language understanding and case-based reasoning, an experimental system supports decision-making and filters routine cases for experts. The results show active data management is feasible, with human communication prioritized for unresolved cases. Legal and technical issues are addressed to improve citizen satisfaction and system efficiency.[4]The paper explores using Cognitive Computing to develop a legal AI chatbot, focusing on improving user interaction, usability, and security. The chatbot, LegalBot, handles appointments and FAQs, connecting to information corpora and APIs. IBM Watson services were used for functionality, with user testing emphasizing trust and interaction effectiveness. Results demonstrate enhanced customer experience in legal services, highlighting future potential for broader applications.[5]The paper introduces a Personalized Complaint Assistant (PCA) for managing and validating complaints without domain-specific knowledge. The paper analyzes the management of e-government data relative to citizen civic engagement/administration engagement dialogues and touches on legalities surrounding privacy, and standardizing practices for those dialogues. The experimental system automates selected decision-making by employing natural language understanding and case-based reasoning, to assist human decision-makers while filtering out cited decision-making and routine efforts for social design experts to address. The findings demonstrate that employing active data management in social design, or legal design, is feasible and supports citizen engagement and administration engagement dialogues, with human communication reserved for situations where cases are unresolved. A series of legalities, social design factors, and technical issues were examined with the intent to enhance citizen satisfaction and improve efficiency of operations, communication, and engagement with the system by citizens and administrators. [4]The paper discusses a case study using Cognitive Computing to develop a legal AI chatbot, which is intended to enhance user interaction, usability, and security during user engagement. LegalBot is designed to schedule appointments and answer frequently asked questions while linking to information corpora and APIs. It used IBM Watson services to support scalability and trust in its functionality, while user testing indicated trust, and design effectiveness during interaction were the two most important factors for the users trust. The findings illustrate marked improvement in the customer experience for legal services, and the application broadens the potential for future explorations of the field. [5]The paper presents a description of a Personalized Complaint Assistant (PCA) to manage and validate complaints from people without knowledge of the particulars of the domain involved in their complaint. PCA utilizes hybrid systems of reasoning with interactive forms to improve the structure and consistency of the complaint, and the speed of resolution. Trust development, and established advanced interfaces are required sustains its effectiveness to consider a range of complaints environment. Deductive reasoning and inductive reasoning allow for expediting a range of conflict that require analysis in uncertainty.[6]The paper reviews the use of Large Language Models (LLMs), such as GPT-3.5 and GPT-4, in governance in terms of a legal assistant - especially in relation to GDPR and the law within the EU. Using Retrieval Augmented Generation (RAG), the system proposed provides promising accuracy in regards to legal related questions, however, witnessed variability within response quality as complexity increased. Beginner questions and intermediate structures derived successful responses, however responses to some expert questions varied between strong and poor. Future assessment will be to refine system performance, as well as assess response reliability.[7] The paper also introduces Mistral 7B, a proposed 7-billion parameter language model that aims to find a sweet spot between performance and efficient use of computational resources. It shows improvements over Llama 2 and Llama 1 models on multiple benchmarks and also implements grouped-query attention and sliding window attention to maintain quicker inference and accommodate long-range sequences. The model showed improvements in code generation, mathematics, and reasoning. Mistral 7B -Instruct, a fine-tuned version of Mistral 7B showed even greater performance on MT-Bench. Some consideration is ongoing on how to optimize performance of smaller models while remaining high performant.[8]

The paper also summarizes the space of Retrieval-Augmented Large Language Models (RA-LLMs), which seek to enhance LLMs by integrating externally reliable, document-based knowledge to proactively reduce overgeneration of hallucinated and out-dated information. It offered surveys of RAG-based architectures, training division, RAG applications, issues of retrievability and honesty, privacy concerns and limited domain-based knowledge. The report identifies early phase RA-LLM research as well as highlighting future directions. The methods defined are training-free, independent, sequential, and joint training. RA-LLMs depict improved generative AI performance as in the allows for QA systems, chatbots, and factual verification. [9]

III. ALGORITHMIC OVERVIEW

The system utilizes cutting edge algorithms to drive efficient data processing, data retrieval, and lastly, generation response. In the successful deployment of a conversational, AI-driven chatbot, these algorithms are key to facilitating the dialogue between the user and the system:

1) Sentence Embedding with SentenceTransformer: The model selected for this step is called SentenceTransformer (all-MiniLM-L6-v2). The model looks at the words in a sentence and generates a high-dimensional vector embedding for textual data. These embeddings represent semantic (the meaning) of sentences and provides similarity-based searches along with contextual understanding. Importantly, when looking at unstructured text, the process of transforming to a machine readable representation is essential in order to facilitate retrieval or analysis.

2) Similarity Search with FAISS: For the similarity search, we employed Facebook AI Similarity Search (FAISS) to provide fast retrieval of pertinent information. First, FAISS organizes the vector embeddings from the previous example into an index structure (IndexFlatL2) and then FAISS applies nearest-neighbor search algorithms, relative to Euclidean distance, to show the most relevant sentences or documents in relation to a query. This structure allows for an efficient and accurate search process relative to potentially large collections of data.

3) Transformer-based Generative Model: For text generation, we used the Mistral-7B model, a generative, transformer-based language model. Mistral-7B uses sophisticated and complex attention mechanisms to "understand" the context provided based on the sentences retrieved from this step and the capability to generate natural language response. This algorithm represents the architectures that generate response in the system.

4) Database Access Algorithms: The operations in MongoDB perform structured access algorithms like validating user credentials, retrieving registration details, and retrieving grievance statuses. These operations ensure that you are efficiently and accurately accessing the appropriate records in the database related to a user.

IV. PROPOSED METHODOLOGY

The proposed methodology for the AI-driven chatbot focuses on building and deploying these tools to support ministry-specific support for the grievance filing process in CPGRAMS. The methodology starts with the data collection and pre-processing of a set of relevant textual data in CPGRAMS, including user guides, frequently asked questions, and document web pages. This raw data is cleaned, tokenized, and embedded in a vectorizable representation using SentenceTransformer with FAST Sentence Transformers Framework, allowing for better semantic interpretation and understanding of meaning in regards to grievance knowledge. The chatbot also builds a searchable FAISS index with the sentence representations of the relevant text data, which allows a user search to be quickly matched with text to retrieve the most contextually relevant text data. The Mistral-7B transformer-based model is used for language response, using the retrieved context and user question as context to produce coherent, context-rich natural language outputs as a response. The MongoDB system is used to manage user credentials, registration numbers, and grievance statuses, while also allowing users to recall their work in a specific and accurate way. The chatbot is deployed on a vague cloud system for manageable scalability while anticipating user demand and resource allocation in support of real time responsiveness.In function, the chatbot is in fact an application program interface, with endpoints for login, signup, query division, and have query status capability, Systematically guiding users through the grievance and complaint filing process, responding to frequently asked questions, and giving status based upon their registration and credential files. The system will undergo operational testing to examine valid functionality, valid interaction protocols, as well as winding in feedback on an ongoing basis to improve performance and ensure accuracy. This approach fused data sourcing and retrieval, advanced natural language processing, and cloud-based software to provide a lean and effective solution to these users, to minimize manual intervention, improve critical user satisfaction, and centralize the operations from the CPGRAMS portal.

V. TECHNOLOGY USED

We are utilizing the programming language Python(3.11) for the project, as it is versatile and has a vast number of libraries available.
For API testing we used postman.

A. Backend

The backend is built using Flask, a python-based web application framework that is lightweight and flexible in nature. Flask provides a solid foundation for building APIs based on the REST architecture, allowing for user requests, data processing, and component interaction.

B. PDF Processing

We have used PyPDF2 to extract text from PDF files. We can efficiently parse both the structured data and unstructured text using this library in order to build the knowledge base for the chatbot.

C. Natural Language Processing (NLP)

- SentenceTransformers: We have implemented this library to extract high-dimensional embeddings for sentences and queries. We apply the embeddings to the queries enabling the semantic meaning to be captured and similarity-based searching.
- FAISS: Stands for Facebook AI Similarity Search. We use FAISS to efficiently retrieve the relevant information from very large datasets with nearest- neighbor search.
- Mistral-7B: Hugging Face is a company that has provided this transformer that is a language model. Mistral-7B produces context-based and coherent user query responses. Mistral-7B uses pre-trained model characteristics.

D. Inference and Computation

- Torch: This machine learning framework powers model execution and inference, offering high performance for the Mistral-7B language model.
- Vertex AI and Compute Engine: With cloud-based solutions, Vertex AI and Compute Engine offer the necessary underpinning infrastructure to support the scaling and running of the chatbot. Vertex AI supports model deployment easier through its environment, while Compute Engine supports our orchestration, resource needs and operational needs.

VI. SYSTEM ARCHITECTURE

A. Input Data Set

At this stage, we are collecting URL links to various web pages, as well as general FAQ resources that are usually accessed by users, and will compile everything collected into one PDF for improved processing.

B. Preprocessing

The preprocessing stage is where raw data is transformed into a structured and machine-readable format for effective down-stream processing. The process begins by parsing text from the PDF documents using a parsing software (PyPDF2), moving from unstructured data to plain text. Once the text has been extracted, it is broken down into their respective sentences, to facilitate a more meaningful representation of the data. Next, with the extracted text represented as sentences, we generate sentence embeddings using a pre-trained SentenceTransformer model (all-MiniLM-L6-v2), which takes textual data and encodes it into hi-dimensional vectors that represent semantic meanings. To prepare for retrieval, the embeddings, as well as the added metadata (e.g. URL links) are indexed in FAISS (Facebook AI Similarity Search) using an IndexFlatL2 structure. When users perform queries, the sentence embedding of the query is similarly pre-processed to allow for nearest neighbor searches in the index. This process creates a structured pipeline that will prepare raw data for retrieval-augmented generation (RAG) and improve the system's ability to return accurate answers and context-aware responses.

C. Analyzing Data

The data analysis employs advanced embedding methods and machine learning models to strip away the subjectivity and provide solid insights and accurate answers. Sentence embeddings are produced from the extracted text and related URLs through a pre-trained SentenceTransformer model (all-MiniLM-L6-v2), capturing the semantic relationships contained in the data.

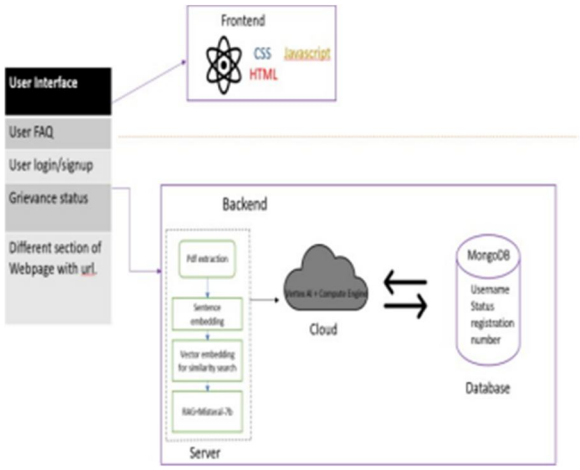


Fig 6.1: Architecture

D. No Training

The system does not rely on training a model, but RAG, and this enabled it to effectively use existing models, because RAG enables the system to find a query and accurately respond using the conversation history. The approach of reducing computation and development time by developing an RAG system is less problematic that creating a model and buidling more development time.

E. Implicit/Operational Testing

In this research, as we used operational testing to assess the system functionality during its operational phase, the functionally of particular system components, and their integration into the larger application, could not be sufficiently examined. While there was not any formalized testing dimension, we did confirm the following:

1) API Endpoint Testing

We tested each of the API endpoints (/signup, /login, /get-status, and /query) for their ability to accept user input and ultimately return an appropriate response. We also verified the mechanism for managing error handling in ensuring that invalid inputs would return an appropriate user-defined error message and re.strip().strip()lative HTTP status code

2) Pre-Processor Testing

4-6 of the RAG functions such as extract_text_from_pdf and create_vector_database could be tested operationally during execution. The content corresponding to these functions, for example the extracted text from the pdf files, and the embeddings for vectors for sentences, were verified against expected responses. This confirms that data is successfully prepared in the pre-processing phase of the retrieval phase and the generation phase of the RAG.

3) RAG Workflow Verification

The Retrieval-Augmented Generation (RAG) was implicitly tested through the /query endpoint. The system's ability to conduct information retrieval from the FAISS vector index and generate a contextually relevant response from the Mistral-7B language model proved that embedding generation, similarity search, and response generation were all functioning properly

4) Database Query Execution

MongoDB queries to verify our operations, such as user registration, user log-in validation, user status: were performed in runtime queries. We checked the response from the database to ensure we were returning the correct information to the user and validating the user's credentials.

So this is flow of architecture incorporated into Flask app for the public to have user friendly interface through a mistral- 7b model and pdf file to load the models in the app.

VII. ALGORITHMS USED

The system uses advanced algorithms to process, retrieve, and generate responses with the data it collects. An embedding is produced for each sentence using the SentenceTransformer model (all-MiniLM-L6-v2) which converts text data into an embedding, a high-dimensional vector of data that represents the semantic meaning of it. For effective and fast retrieval of relevant sections of text, Facebook AI Similarity Search (FAISS) organizes the embeddings into an index and with a nearest-neighbor search achieves an approximate nearest neighbor around an index of embeddings leveraging the Euclidean distance.

To create a coherent response that is contextually aware, the Mistral-7B transformer-based language model is used to generate a response while taking into account the input as an attention mechanism that combines the retrieved context embeddings and the user's query as input. The MongoDB query for querying user's credentials, registration numbers, and grievance statuses utilize algorithms that use efficient operations while following the CRUD process (Create-Read-Update-Delete). This coming together of components creates a system that utilizes a collection of algorithms which, as a whole, create a Retrieval-Augmented Generation (RAG) framework for the chatbot that creates the capability to provide contextually relevant and correct assistance to the user.

VIII. OUTPUT SCREENSHOTS

Fig 1.1 Prompt 1: how to register

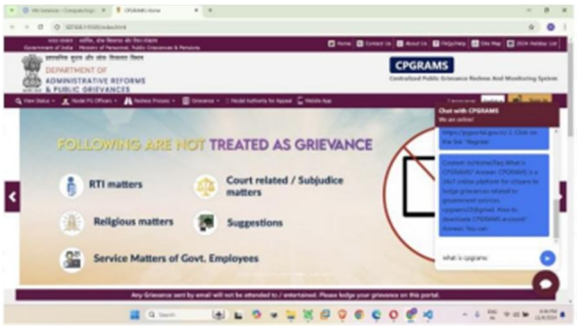


Fig 1.2 Prompt 2: what is cpgrams

IX. CONCLUSION

This project showcases the creation and implementation of an AI enabled chatbot that improves the functionality and user experience of CPGRAMS grievance filing process. Using recent advancements, such as Large Language Models (LLMs), FAISS-based vector retrieval, and cloud computing, the chatbot effectively helped users navigate the grievance process and answer frequently asked questions. Through the use of natural language processing for training and semantics, combined with a cloud-based, scalable backend platform for generating responses and other functionality, the project demonstrated the chatbot's ability to adapt and respond quickly in a public service situation.

By providing a remote and scalable intelligent user interface to users while minimizing the need to interact with manual support, the proposed solution has dramatically improved users' access to government services.

Moreover, the chatbot has the capability of continuing to improve over time as users interact it and provide feedback, ensuring its relevance in a dynamic and complex operational environment. Overall, the project demonstrates how the use of AI and ML technologies can be used to change the engagement of citizens and modernize public sector engagement to make it more efficient and agile, and is an important example for future projects leveraging AI-based governance solutions.

REFERENCES

[1] The Role of Artificial Intelligence in Government Services: A Systematic Literature Review Hamirul1*, Darmawanto1, Nova Elsyra1, Syahwami1 1Setih Setio Institute of Administration and Health, Muarabungo, Indonesia

[2] On Large Language Models in National Security Applications William N. Caballeroa, Phillip R. Jenkinsa aDepartment of Operational Sciences, Air Force Institute of Technology, WPAFB, OH 4543

[3] Paul Henman (2020): Improving public services using artificial intelligence: possibilities, pitfalls, governance, Asia Pacific Journal of Public Administration, DOI: 10.1080/23276665.2020.1816188 To link to this article: https://doi.org/10.1080/23276665.2020.1816188

[4] Data Management and AI in E-government□ Tibor Vámos and István Soós Computer and Automation Research Institute, Hungarian Academy of Sciences H-1111 Budapest, Lágymányosi u. 11., Hungary vamos@sztaki.hu

[5] APPLYING COGNITIVE COMPUTING TO LEGAL SERVICES Deborah Whittle Faculty of Technology University of Sunderland, Sunderland. bg19tz@student.sunderland.ac.uk Lynne Hall Faculty of Technology University of Sunderland, Sunderland Lynne.hall@sunderland.ac.uk http://dx.doi.org/10.14236/ewic/HCI2022.35

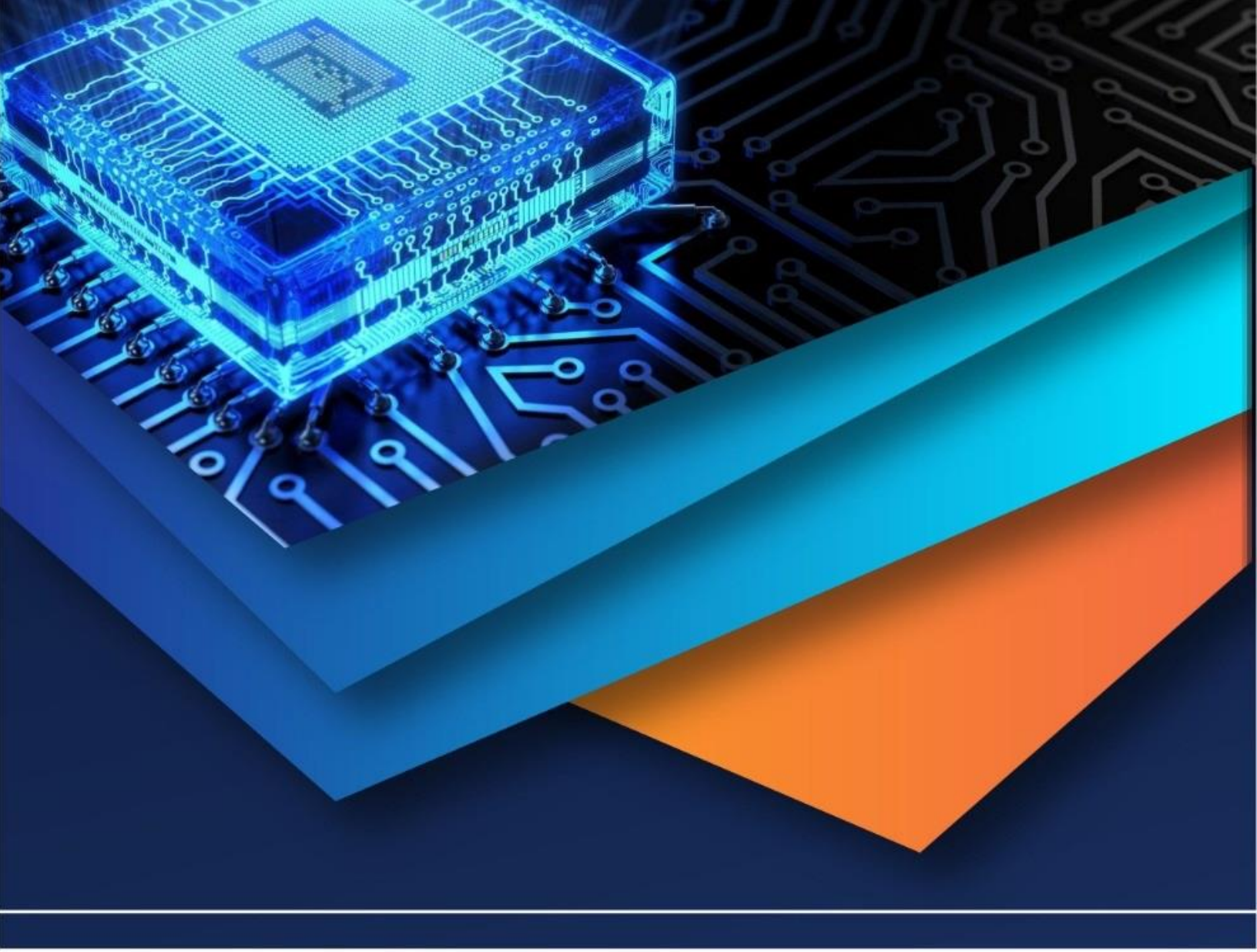
[6] A Personalized Assistant for Customer Complaints Management Systems Boris Galitsky School of Computer Science and Information Systems Birkbeck College, University of London Malet Street, London WC1E 7HX, UK galitsky@dcs.bbk.ac.uk http://www.dcs.bbk.ac.uk/~galitsky/ComplaintE ngineIntro.html

[7] A Large Language Model based legal assistant for governance applications Mamalis, Marios Evangelos□ marios.mamalis@uom.edu.gr Fitsilis, Fotios† fitsilisf@parliament.gr Kalampokis, Evangelos□ ekal@uom.edu.gr Theodorakopoulos, Georgios‡ g.theodorakopoulos@nsk.gr Tarabanis, Konstantinos□ kat@uom.edu.gr

[8] Mistral 7B 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, William El Sayed

[9] ASurvey on RAGMeetingLLMs: Towards Retrieval-Augmented Large Language Models Wenqi Fan wenqifan03@gmail.com The Hong Kong Polytechnic University, HK SAR Shijie Wang arXiv:2405.06211v3 [cs.CL] 17 Jun 2024 shijie.wang@connect.polyu.hk The Hong Kong Polytechnic University, HK SAR Tat-Seng Chua Yujuan Ding□ dingyujuan385@gmail.com The Hong Kong Polytechnic University, HK SAR Hengyun Li neilhengyun.li@polyu.edu.hk The Hong Kong Polytechnic University, HK SAR dcscts@nus.edu.sg National University of Singapore, Singapore Liangbo Ning BigLemon1123@gmail.com The Hong Kong Polytechnic University, HK SAR Dawei Yin yindawei@acm.org Baidu Inc, China Qing Li csqli@comp.polyu.edu.hk The Hong Kong Polytechnic University, HK SAR

-



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY