ISSN No:-2456-2165

Enhancing Drug Information Access: AI-Powered System with Large Language Models and Chatbot Integration

Authors: Surendra Digumarthi[1], Sarita Padhi[2], Sai Raghava[3], Sreekanth Putsala[4], Shirish Kumar Gonala[5] Bharani Kumar Depuru[6]
[1,2,3]Research Associate, Innodatatics, Hyderabad, India.
[4]Team Leader, Research and Development, Innodatatics, Hyderabad, India.
[5,6]Director, Innodatatics, Hyderabad, India.
***Corresponding Author:** Bharani Kumar Depuru
OCR ID: 0009-0003-4338-8914

Abstract:- Chatbot, the AI Powered Pharmaceutical Classification Systems using LLMs, can assist users in efficiently navigating and understanding complex drug classification. This research introduces the chatbot framework empowered by LLM, to provide users with a conversational user interface (UI) for Pharmaceutical Classification inquiries.

The training of chatbot happened on a diverse dataset, enabling it to grasp the intricate relationship between drugs, dosage form, product type, pack size etc. Through continuous interactions, the chatbot leverages its contextual understanding to deliver real-time and accurate information to users, ranging from healthcare professionals seeking specific drug classifications to consumers inquiring about medication details.

The research involves fine-tuning the pre-trained language model such as Palm2, Llama2 and Meditron, T5, Mistral 7B, TAPEX, BERT on a curated dataset of drug related texts to enhance its understanding of pharmaceutical concepts, molecular structure etc. The fine-tuned model is then employed to classify drugs based on multiple criteria including mechanism of action, therapeutic class etc. The model's ability to comprehend complex relationships and contextual information enables it to make accurate predictions and handle ambiguous cases.

The Practical implication of this research extends to pharmaceutical education, healthcare decision support, and public health awareness. By offering a user - friendly and conversational interface, the chatbot provides an accessible and efficient means for individuals across diverse backgrounds to obtain reliable drug classification information. The study underscores the transformative potential of LLMs in developing intelligent chatbot tailored for pharmaceutical knowledge dissemination, thereby contributing to the evolving landscape of healthcare informatics.

Keywords:- *Artificial Intelligence, Large Language Models, Chatbot Framework, Drug Classification, ,Conversational Interface, Real-Time Information.*

I. INTRODUCTION

The Indian government is responsible for overseeing the prices of essential drugs which include bulk drugs and formulations, dividing them into scheduled and non-scheduled categories. Government regulation applies to scheduled drugs, which are deemed critical for public health, and their prices are determined by governmental authorities. In contrast, non-scheduled drugs are not subject to government regulation, giving manufacturers the autonomy to establish their pricing strategies.

'Bulk Drugs' denotes the raw materials employed in drug manufacturing, encompassing salts and stereoisomers utilized as ingredients in various formulations.

Categories of Active Pharmaceutical Ingredients (APIs) / Bulk Drugs:

- Chemical Synthetic Drugs
- Inorganic – Examples: Aluminium hydroxide and Magnesium trisilicate
- Organic – Examples: Aspirin, Chloramphenicol, Caffeine, and so forth.

'Formulations' refer to medications derived from bulk drugs or drugs, available in diverse formats like tablets, capsules, syrups, creams, and ointments.

'Drugs' encompass all formulations or medicines designed for internal and external use in humans or animals for health purposes. Drugs can manifest in different forms, including chemical, biological, or pharmaceutical-based products.

ISSN No:-2456-2165

II. METHODOLOGY

This article proposes a Pharmaceutical Classification System architecture based on the CRISP-ML(Q) methodology available in the 360DigiTMG website (ak.1) [Fig.1]. The architecture is designed for creating a medical chatbot for classifying medicines and medical devices [5]. Categories encompass scheduled/non-scheduled, new drug (patented/non-patented), and Para 19.

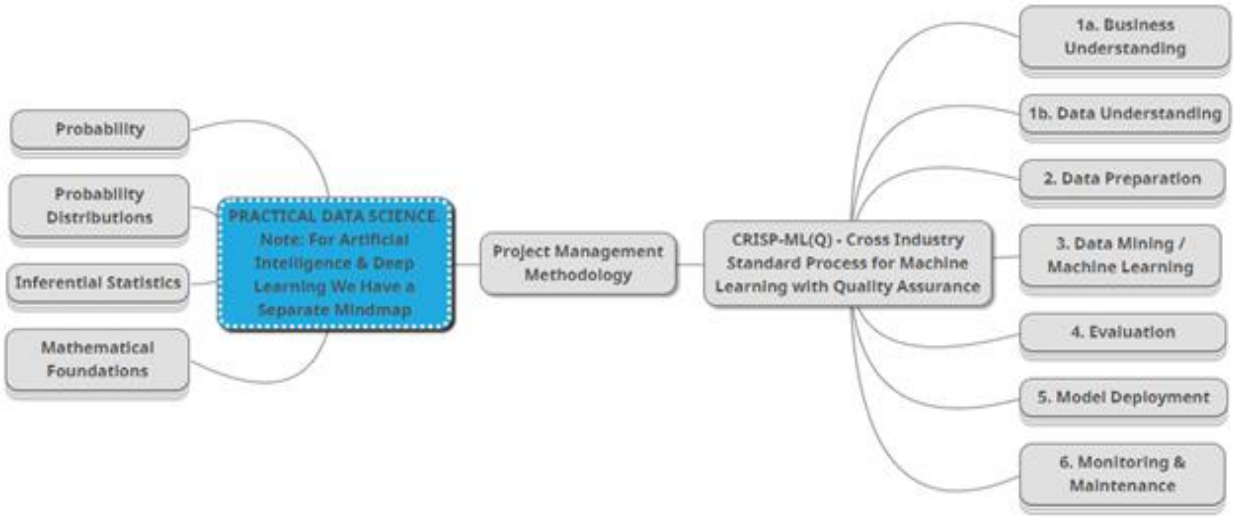

Fig.1: *The above Figure Displays the CRISP-ML(Q) Architecture that we used for this Research Project. (Source: Mind Map - 360DigiTMG)*

In the Business Understanding phase, we identified the following business problem: Manually classifying pharmaceuticals requires a considerable amount of time and effort, involving the examination of diverse sources [10]. The implementation of a pharmaceutical classification chatbot, leveraging large language models, expedites the pharmaceutical classification process, ensuring swift and accurate results [8].

In the Data Understanding phase, we explored and understood the data that we would be using to train the Large Language Model [11]. The dataset comprises information on drug names, categories, dosages, compositions, prices, and manufacturers. Additionally, during the analysis, we identified certain data cleaning and preprocessing tasks essential for refining the dataset, including the removal of extraneous information and the consolidation of data into a unified file [9].

In the Data Preparation phase, we conducted cleaning, preprocessing, and transformation procedures on the data, resulting in the creation of a CSV file optimized for model training [9].

In the Modelling phase of the CRISP-ML(Q) methodology, we will alter and add the Large Language Models mentioned above to improve the performance of our pharmaceutical classification system [11,12]. We will do this by:

- Choosing the right large language model for our task.
- Fine Tuning the model to improve its performance.
- Adding new data to the training set to improve the model's performance.

Once the model is trained, we will evaluate its performance on a held-out test set to ensure that it meets our requirements [13]. If the model does not meet our requirements, we will go back to the Modelling phase and alter the model further [11,12].

After confirming the satisfactory performance of the model, deployment to production will follow, enabling its utilization for drug classification [13]. The chatbot will seamlessly integrate into a client's website, providing users with the ability to engage in conversations and obtain answers.

The proposed architecture for a Pharmaceutical Classification System, based on the CRISP-ML(Q) methodology and emphasizing modifications and incorporations of Large Language Models, holds the potential to enhance the efficiency and precision of addressing client inquiries [11]. This approach aims to diminish the dependency on manual responses, ultimately resulting in substantial time savings.

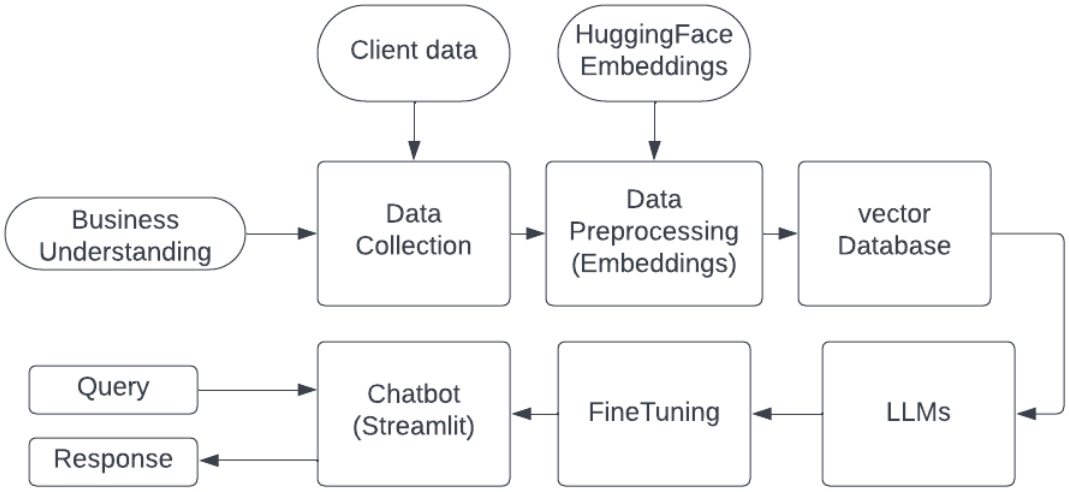

Fig.2 *ML Workflow Architecture used for the Research - A Detailed overview of Large Language Model based chatbot for answering client inquiries. (Source: ML Workflow - 360DigiTMG)*

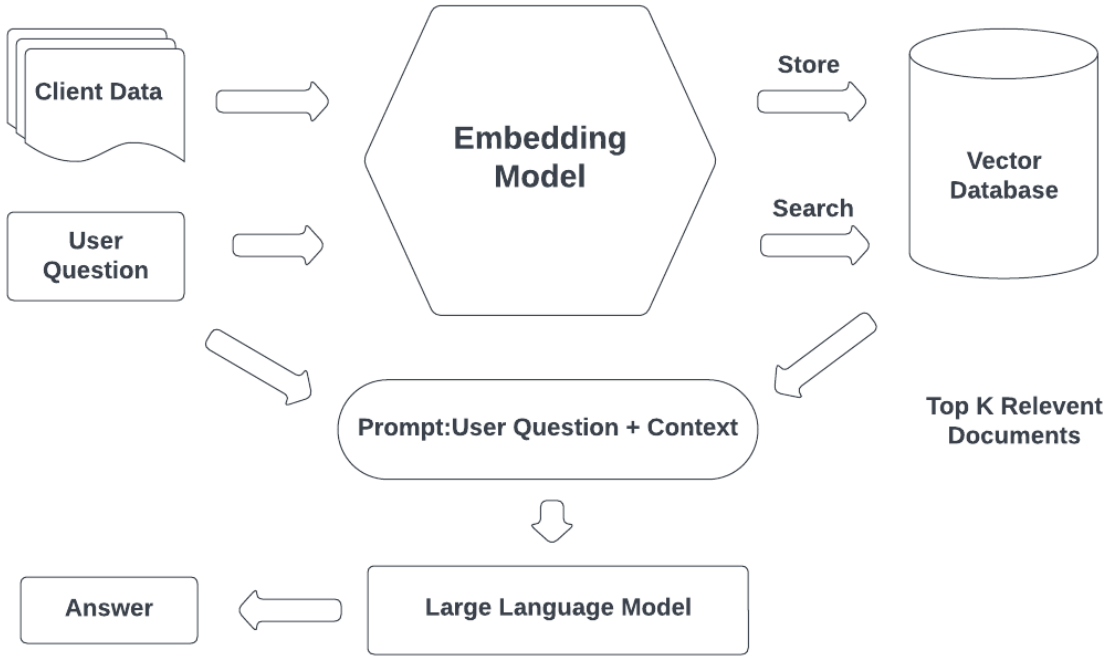

Fig.3: *Retrieval Augmented Generation Architecture*

A. Data Collection

In this section, we describe the methodology employed for the acquisition of pharmaceutical data [Fig.2, 3], which forms the foundation of our research dataset. We have utilized the interactive ML workflow diagram hosted as open source by 360DigiTMG (ak.2) [Fig.2]. The collected dataset encompasses the category, composition, dosage, and manufacturer.

➢ *Data Source*

We obtained pharmaceutical-related data from two primary sources:

- *Internet Sources:*

We obtained information on scheduled drugs and their prices [Table.1] from the National List of Essential Medicines (NLEM) on the Ministry of Health & Family Welfare website.

- *Client Data:*

The remaining portion of our dataset [Table.1] was procured from the client database .

➢ Types of drugs classification: the Dataset represents a wide range of drugs classification, Including:

ISSN No:-2456-2165

Category: The drug is scheduled or non-scheduled.
Composition: The composition of the drug.
Dosage: The dosage and strength of the drug.
Price: the price fixed by the NPPA.
Manufacture: The manufacture of the drug.
Drug type: whether the drug is injection, tablet, or syrup.

Table 1: *Data Description*

Data Description	Records	Format
Products 1 from Client	487	Excel
Products 2 from Client	346	Excel
Products 3 from Client	92,390	Excel
NPPA Price list from Internet	3551	PDF
NLEM from Internet	418	PDF

B. Data Preprocessing

In our data preprocessing approach, we partitioned the dataset into smaller chunks [Fig.3], each containing 1000 characters and overlapping by 100 characters [4,6,7]. This segmentation strategy aligns with the input token limit imposed by the large language model employed, facilitated through the Lang chain library in Python [7]. This ensures that the data can be efficiently processed without exceeding the model's constraints.

Following the segmentation, we proceeded to create embeddings [Fig.3] for each individual chunk [4,7]. Leveraging the capabilities of the Hugging Face embeddings and the FAISS vector store, this step enhances the representation of the data, contributing to the overall effectiveness of our language model in capturing nuanced patterns and information within the segmented chunks [1,2,3].

C. Large Language Models:

PaLM2 by google ai is a transformer-based technique excels in classification Q&A problems and NLG tasks. Llama 2 showcases fluency in query and solution dialogue coherence and text summarization boasting 137b parameters. Meditron, ranging from 70 billion to 540 billion parameters specializes in medical question answering including differential diagnosis support. T5 a versatile language model demonstrates excellence in various domains particularly machine translation leveraging extensive training across diverse datasets. Mistral 7B a resource-efficient language model demonstrates powerful interpretation and precise handling of vast data. Tapex empowers models by pre-training them to comprehend and execute sql commands on synthetic tabular data. BERT, “bert-large-uncased” variant achieves impressive accuracies in multiple nlp tasks such as SQUAD 1.1, Multi NLI and MNLI [Table.2].

Table.2*: LLMs Comparison*

Model	Release Time	Parameter Size	Open-Source	Provider
PaLM 2	May 2023	340 B	Yes	Google
Llama 2	July 2023	70 B	Yes	Meta AI
Mistral 7B	September 2023	7.3 B	Yes	Mistral AI
Meditron	November 2023	70B	Yes	Meditron
T5	October 2019	220M	Yes	Google
BERT	October 2018	336M	Yes	Google
TAPEX	October 2023	137M	Yes	Microsoft

The architecture of above-mentioned models contains hyperparameters that influences training across tasks such as classification generation comprehension optimizing efficiency in diverse NLP endeavors. [Table.3].

Table.3: *LLMs and its Hyperparameters*

Model	Hyperparameters
Llama 2	Max_new_tokens=512, temperature=0.9
PaLM 2	Max_new_tokens=512, Temperature=0.5
Meditron	Max_new_tokens=256, Temperature=0.1
TAPEX	Max_length=50, Temperature=0.7
Mistral 7B	Max_length=512
T5	Max_length=50, Temperature=0.7
BERT Large Uncased	-

D. Deployment Strategy

Model deployment signifies a pivotal stage in transitioning that has undergone training from its developmental stage to an operational state inside the confines of an operational environment; this crucial step enables the model to actively engage in making predictions or outputs synthesis.

Streamlit is a python-centric open-source framework utilized for constructing interactive and data-driven applications it empowers developers to rapidly deploy applications utilizing python by furnishing a straightforward and instinctive api for constructing web interfaces.

- Streamlit Application

The Streamlit interface [Fig.4] facilitates users in posing queries. The chatbot, powered by Large Language Models (LLM), responds to these queries utilizing the vector database.

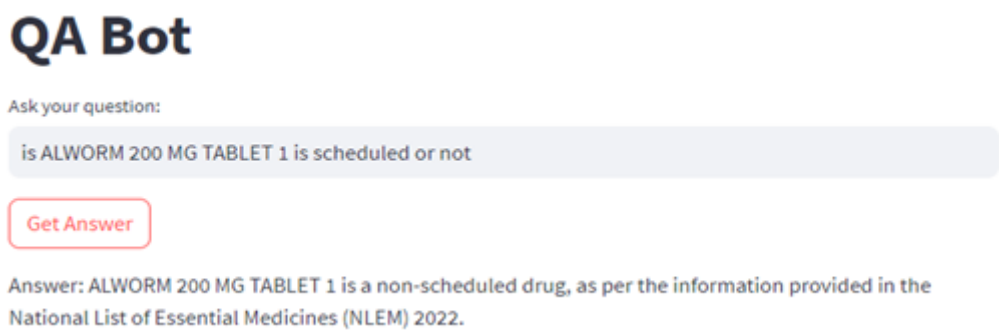

***Fig.4**: The figure above illustrates the deployment using Streamlit and the drug category in question.*

***Table.4:** Table shows the accuracies and response time of different LLMs.*

LLM	Accuracy	Response Time
Llama2	85%	2 Min
BERT Large Uncase	98%	3sec
Palm2	85%	6sec
Meditron	90%	10 sec
T5	80%	36 sec
TAPEX	90%	5 sec
Mistral 7B	70%	13 -20 sec

III. RESULTS AND DISCUSSION

In the model training phase, we meticulously assessed the performance of seven distinct models [Table.4,3] using a dataset comprising over 80,000 records and leveraged RAG implementation and vector databases [Fig.3]. Notably, the bert-large-uncased-hole-word-masking-fine tuned-squad model emerged as the top performer, showcasing exceptional accuracy. However, recent observations have indicated that Llama 2 is experiencing extended response times, raising concerns about its efficiency. Additionally, Palm2 has exhibited occasional hallucinations, posing potential challenges to its reliability.

Despite these concerns, all models demonstrated commendable performance in pharmaceutical classification tasks. The bert-large-uncased-hole-word-masking-fine tuned-squad model, in particular, continued to stand out. To ensure the overall robustness of our model selection, it becomes imperative to conduct further evaluations on a larger dataset. Moreover, making comparisons with other state-of-the-art models is crucial before considering deployment in a production environment. Addressing the issues with Llama 2's response time and Palm2's hallucinations will be essential in ensuring the overall success and dependability of the selected model.

IV. CONCLUSION

In summary, our study utilized state-of-the-art language models to improve the categorization of pharmaceutical information. Employing cutting-edge methods, we effectively created an advanced chatbot capable of addressing queries about medications. By tapping into the capabilities of the latest language models, our research not only advances drug information systems but also underscores the prospect of enhancing accessibility and user engagement in the pharmaceutical field.

REFERENCES

[1]. Feifei Pan, Mustafa Canim, Michael Glass, Alfio Gliozzo, James Hendler. End-to-End Table Question Answering via
[2]. Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z. Pan, Kam-Fai Wong. UniMS-RAG: A Unified Multi-Source Retrieval-Augmented Generation for Personalized Dialogue Systems. https://doi.org/10.48550/arXiv.2401.13256
[3]. Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. https://doi.org/10.48550/arXiv.2312.10997
[4]. Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, Lei Li. On the Sentence Embeddings from Pre-trained Language Models. https://doi.org/10.48550/arXiv.2011.05864
[5]. Keivalya Pandya, Mehfuza Holia. Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organization. https://doi.org/10.48550/arXiv.2310.05421
[6]. Qiao Jin, Bhuwan Dhingra, William W. Cohen, Xinghua Lu. Probing Biomedical Embeddings from Language Models. https://doi.org/10.48550/arXiv.1904.02181
[7]. Jimmy Lin, Ronak Pradeep, Tommaso Teofili, Jasper Xian. Vector Search with OpenAI Embeddings: Lucene Is All You Need. https://doi.org/10.48550/arXiv.2308.14963
[8]. Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan & Daniel Shu Wei Ting. Large language models in medicine. https://doi.org/10.1038/s41591-023-02448-8
[9]. Hamza Ed-douibi, Javier Luis Cánovas Izquierdo, Gwendal Daniel & Jordi Cabot. A Model-Based Chatbot Generation Approach to Converse with Open Data Sources. https://doi.org/10.1007/978-3-030-74296-6_33
[10]. Ran Elgedawy, Sudarshan Srinivasan, Ioana Danciu. Dynamic Q&A of Clinical Documents with Large Language Models. https://doi.org/10.48550/arXiv.2401.10733

[11]. Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, You Zhang . ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. https://doi.org/10.48550/arXiv.2303.14070
[12]. Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, Jimmy Lin. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. https://doi.org/10.48550/arXiv.2310.08319
[13]. Samaneh Shafee, Alysson Bessani, Pedro M. Ferreira. Evaluation of LLM Chatbots for OSINT-based Cyberthreat Awareness. https://doi.org/10.48550/arXiv.2401.15127
[14]. PaLM 2 Technical Report. https://doi.org/10.48550/arXiv.2305.10403
[15]. Llama 2: Open Foundation and Fine-Tuned Chat Models. https://doi.org/10.48550/arXiv.2307.09288
[16]. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. https://doi.org/10.48550/arXiv.2311.16079
[17]. Clinical-T5: Large Language Models Built Using MIMIC Clinical Text. https://doi.org/10.13026/rj8x-v335.
[18]. Improving Text Embeddings with Large Language Models. https://doi.org/10.48550/arXiv.2401.00368
[19]. TAPEX: Table Pre-training via Learning a Neural SQL Executor. https://doi.org/10.48550/arXiv.2107.07653
[20]. SciBERT: A Pretrained Language Model for Scientific Text. https://doi.org/10.48550/arXiv.1903.10676