***An AI-Powered Chatbot for FDA Drug Labeling Information Retrieval: Leveraging OpenAI's GPT for Enhanced Document Understanding and Query Response***

Manasa Koppula [manasakoppula@utexas.edu](mailto:manasakoppula@utexas.edu)

Navya

Madhulika

University of Texas at Austin

Austin, Texas, USA

**ABSTRACT**

This work presents the development of a Chatbot designed to assist users in retrieving information from FDA drug labeling documents. Using OpenAI’s GPT model, the Chatbot provides users with answers derived from specific sections of drug labeling PDFs. The tool features a streamlined interface built with Streamlit, allowing users to easily upload PDF files for analysis. Key features include semantic similarity scoring to assess answer relevance and the imposition of constraints to ensure that responses are limited to information contained within the uploaded document. The performance of the Chatbot is evaluated using a ground-truth comparison, and semantic similarity scores are calculated to assess the effectiveness of the model in delivering authentic and relevant information.

1. **Introduction**

**Problem Statement:**  
The FDA drug labeling documents are dense and complex, often requiring time-consuming manual reading to extract key information. This paper proposes a solution to automate the process of querying these documents for specific information through the development of a Chatbot.

**Objective:**   
The goal of this work is to build an AI-based tool that can query FDA drug labeling PDFs, extract information from key sections in Highlights. (e.g., Indications and Usage, Dosage and Administration, Warnings and Precautions, Adverse Reactions, and Drug Interactions), and provide users with authentic and relevant answers to their questions, while ensuring the answers are strictly grounded in the document content.

* **Scope:**   
  The scope includes the development of the user interface, integration with OpenAI’s GPT model, and the implementation of semantic similarity scoring to validate the chatbot's responses.

**2. Background and Related Work:**

* **FDA Drug Labeling Documents:**   
  Every FDA-approved drug has a document called product labels or package inserts are legally binding resources that define exactly how a drug must be used. They protect safety, guide clinician decision-making, and ensure manufacturers stay compliant with strict federal regulations. A standard drug labels include sections like “Indications and Usage”, “Dosage and Administration”, “Warnings and Precautions”, “Adverse Reactions”, “Drug Interactions”, “Use in Specific Populations”, “Patient Counseling Information”, and “Reporting and Manufacturer Information”[1]. FDA labels are dense, technical, and often spanning dozens of pages with detailed medical language, study data, and regulatory references[2]. Though they are structured in standard sections, due to their highly detailed nature and complexity, they require careful navigation and exact extraction of information.
* **Existing Solutions:**   
  Recent years have seen strong progress in document retrieval and question answering (QA) for medical and regulatory texts. Classic examples include search engines for FDA databases, Biomedical NLP models like (BioBERT, PubMedBERT for biomedical QA) , and recent large language model (LLM) frameworks that perform open-domain QA. However, these solutions either generate answers freely using the entire pretrained model’s knowledge or general web sources or broad medical corpora, which can introduce hallucinations.

This work specifically addresses this gap by combining ChatGPT’s advanced natural language capabilities with controlled, document-grounded QA. No external or pretrained knowledge leaks into the response, maintaining traceable authenticity and legal alignment. This approach ensures that responses are factually constrained, avoiding the risk of the model generating information not present in label.

* **AI and NLP for Medical Data:**   
  p8),(p9),(p10),(p11),(p12),(p13)

Large Language Models like GPT and Natural Language Processing (NLP) excel at generating fluent, human-like text, which makes them promising for tasks like summarization, and conversational QA. They have been tested on tasks such as medical QA benchmarks (eg., MedQA, PubMedQA) and specialized applications like clinical note drafting or explanation of medical literature.

But for high-stake domains like regulatory labeling, additional controls are needed to ensure accuracy, context fidelity, and legal compliance.

**3. Methodology:**

* **System Architecture:**   
  Our solution integrates a modern document-grounded question-answering (QA) pipeline for FDA Drug labeling documents, combining an easy-to-use interface, robust document processing, and OpenAI’s GPT-35-turbo model with carefully designed constraints to ensure authentic, reliable answers.
* **User Interface (Streamlit):**   
  The system utilizes Streamlit as a lightweight, interactive web framework. Users can:
  + Upload the FDA labeling document in PDF format.
  + Enter natural language questions related to the uploaded document.
  + Receive answers generated by the AI model that are constrained strictly to the uploaded document’s content.
* **Document Processing:**   
  Once a PDF is uploaded, it is processed to extract plain text. We leverage PymuPDF to parse the PDF. After extracting, the raw text is pre-processed (encoded properly) and then segmented into relevant sections commonly found in FDA labels, such as:
* Indications and Usage
* Dosage and Administration
* Warnings and Precautions
* Adverse Reactions
* Drug Interactions
* **GPT Integration:**   
  To perform QA, we integrate Azure Open AI’s chatGPT-35-tubo model.

The extracted sections are stored as context for the GPT prompt.

* When the user asks a question, a prompt template explicitly instructs the model to answer only using content found in the uploaded text.
* While no additional fine-tuning is performed in this setup, prompt engineering plays a crucial role - prompts are crafted with clear system instructions and dynamic insertion of the extracted sections.
* **Text Extraction and Constraints:**
* **Section Extraction:**   
  A regex-based approach identifies section headings (e.g., “INDICATIONS AND USAGE”, “DOSAGE AND ADMINISTRATION”) and slices the document into structured segments. This ensures the model can locate context-specific answers rather than searching the entire raw text.
* **Answer Constraints:**

To enforce factual consistency, the system includes only the relevant extracted text in the GPT input prompt.

* **Semantic Similarity Scoring:**

Semantic similarity is a concept in NLP that helps us see how close the meaning of the generated answer instead to the exact word match. A score of 1.0 means identical meaning, and 0.0 means unrelated.

**4. Results:**

* **Test Dataset:**   
  30 Oncology drugs related to breast cancer were used to test the Chatbot. 8 questions were given as input to the Chatbot.
* What are the indications?
* What is the usage?
* What is the dosage?
* What are the warnings?
* What are the precautions?
* What are adverse reactions?
* What are side effects?
* What are the drug interactions?
* **Performance Evaluation:**

All the answers generated by Chatbot for each drug were recorded manually to a CSV file along with the Ground truth (true text from the FDA label). Both columns were converted into vector embeddings and then cosine similarity is calculated to understand the semantic similarity.

**5. Discussion:**

* **Interpretation of Results:**   
  Most of the scores cluster between 0.7 and 0.9, showing that most of the answers align well with the original FDA text. A few scores fall below 0.7, indicating cases where the generated answers paraphrased or partially summarized more complex sections.

Higher Scores (0.8-0.9) typically appear for sections like Warnings and Precautions.

Moderate scores (0.6 – 0.75) can occur when the answer summarizes multiple lines or rephrases legal language into simpler sentences, reflecting desirable paraphrasing without loss of meaning.

Add example here

This demonstrates that the model is serving the purpose by condensing instead of repeating all specifics. This shows a balance between natural language output and regulatory accuracy, which is crucial for real-world usability.

* **Challenges and Limitations:**   
  Address the limitations of the current system, such as issues with document quality, handling complex queries, or incomplete extractions.
* **Potential Enhancements:**

Going forward, this solution can be extended with Automatic flagging of low-similarity responses for human review. 

**6. Conclusion:**

This project demonstrates a practical, document-grounded question-answering (QA) system for navigating complex FDA drug labeling documents. The system addresses key gaps found in general–purpose LLM’s namely the risk of hallucination and the lack of traceability. This approach illustrates that LLMs can be responsibly deployed in sensitive, high-stakes  domains like medical regulation.

**7. References:**

1. Fang H et al. FDA drug labeling: rich resources to

facilitate precision medicine, drug safety, and

regulatory science. Drug Discov Today.

2016;21:1566–1570.

1. Refered

**Appendices:**

Include any additional information such as:

* Example inputs and outputs from the chatbot.
* Code snippets or detailed descriptions of the algorithm used for text extraction and answer generation.
* Sample semantic similarity results for ground truth vs. generated answers.