AI-Powered Chatbot for FDA Drug Labeling Information Retrieval: OpenAI's GPT for Grounded Question Answering

Manasa Koppula (mk47542)
manasakoppula@utexas.edu
Master of Science in Artificial Engineering
University of Texas at Austin
Austin, Texas, USA

Fnu Madhulika (fm22739)
madhulika27@utexas.edu
Master of Science in Artificial Engineering
University of Texas at Austin
Austin, Texas, USA

Navya Sreeramoju (ns38687)
navyasreeramoju@utexas.edu
Master of Science in Artificial Engineering
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 FDA drug labeling PDFs. A streamlined interface was built using Streamlit, allowing users to upload PDF files for analysis easily. Major features include semantic similarity scoring for determining answer relevance and the use of constraints to ensure that responses are limited to information contained within the uploaded document, overcoming hallucinations. The performance of the Chatbot is evaluated using a ground-truth comparison and semantic similarity scores, which are calculated to assess the effectiveness of the model in delivering authentic and relevant information.

1. INTRODUCTION

Every FDA-approved drug has a document called a product label, also known as package insert, 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.[1] 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". FDA labels are dense, technical, and often span dozens of pages with detailed medical language, study data, and regulatory references. Though they are structured in standard sections, due to their highly detailed nature and complexity, they require careful navigation and exact extraction of information.[2]

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 [3] and PubMedBERT for biomedical QA[4]), 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.[5]

Large Language Models, such as GPT, and Natural Language Processing (NLP) excel at generating fluent, human-like text, which makes them promising for tasks like summarization and conversational QA[6] [7]. They have been tested on functions such as medical QA benchmarks (eg., MedQA, PubMedQA) and specialized applications like clinical note drafting or explanation of medical literature.[8] [9] [10] [11]

AI Language models are trained using vast amounts of data from many sources, such as books, research papers and the internet [12]. But for high-stakes domains like regulatory labeling, additional controls are needed to ensure accuracy, context fidelity, and legal compliance of the response generated.[8] [13]

This study specifically addresses this gap by combining ChatGPT's advanced natural language capabilities with controlled, document-grounded QA. No external or pretrained knowledge contaminates 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 the label.

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. 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 Chabot's responses.

2. METHODOLOGY

Our solution integrates a modern document-grounded question-answering (QA) pipeline for FDA Drug labeling documents. It combines a user-friendly interface, robust document processing, and OpenAI's GPT-35-turbo model with carefully designed constraints to ensure authentic, reliable responses.

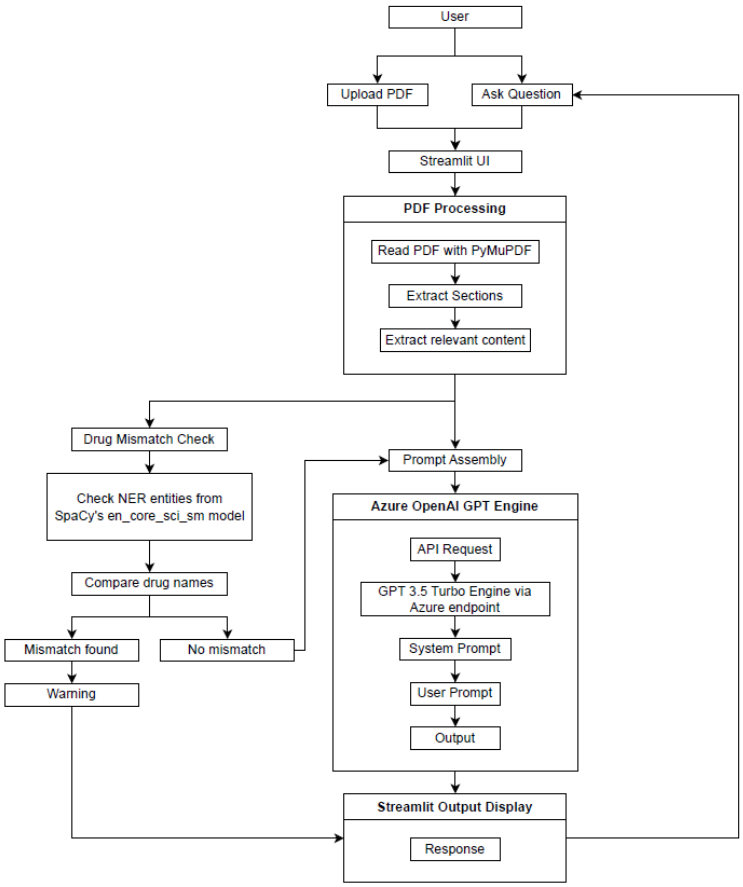

Figure 1: Architecture

2.1 User Interface (Streamlit)

The system utilizes Streamlit, which is a lightweight, interactive web framework that simplifies the process of creating interactive applications.[14]

Users can:

- Upload the FDA labeling document of a drug in PDF format. Figure 2
- Ask natural language questions about the drug.
- Receive answers generated by the AI model that are limited strictly to the uploaded document's content.

2.2 Document Processing:

After the user uploads a PDF of a labeling document of FDA FDA-approved drug, file is parsed to extract the plain text. We leverage PyMuPDF to parse the PDF. After extracting, the raw text, a regex-based approach is used to identify section headings (e.g., "INDICATIONS AND USAGE", "DOSAGE AND ADMINISTRATION", etc.) and slices the document into structured segments. This ensures the model can locate context-specific answers rather than searching the entire raw text.
The five sections mentioned below are the focus of this study.

- Indications and Usage
- Dosage and Administration
- Warnings and Precautions
- Adverse Reactions
- Drug Interactions

2.3 GPT Integration:

To perform Question-Answering, we integrate Azure OpenAI's chatGPT-35-turbo model. The extracted sections are stored as context for the GPT prompt.

- A prompt template specifically tells the model to respond to user questions by utilizing only the information provided in the document uploaded.

- While no additional fine-tuning is required in this setup, prompt engineering plays a crucial role, as prompts are crafted with clear system instructions and dynamic insertion of the extracted sections.

2.4 Semantic Similarity Scoring:

Semantic similarity is a concept in NLP that helps us see how close two texts (words, sentences, documents) are in conveying the same meaning despite using different words.[15] For example, the words "Bus" and "Vehicle" show high similarity, while "Tiger" and "Lion" exhibit some similarity as they both are animals. But "Bus" and "Tiger" show Little to no similarity.

3. DATASET:

30 Oncology FDA-approved drugs related to breast cancer were used to test the Chatbot. Eight questions were given as input to the Chatbot, and responses were recorded.

- 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 the side effects?
- What are the drug interactions?

Performance Evaluation

The Sentence Transformers model, which is based on top of a pre-trained transformer model MiniLM(all-MiniLM-L6-v2), is used for calculating Semantic similarity. These models convert sentences, paragraphs, or documents into dense vectors and measure semantic similarity between text pairs using cosine similarity. Later, these embeddings are used for the task of question-answering.[16]

All the answers generated by the 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 the cosine similarity was calculated to understand the semantic similarity.

4. RESULTS

As shown in Figure 2, Most of the scores fall between 0.7 and 0.9, showing that the majority of the responses align well with the original FDA text. A few scores fall below 0.7, indicating cases where the generated responses are paraphrased or partially summarized from longer passages.

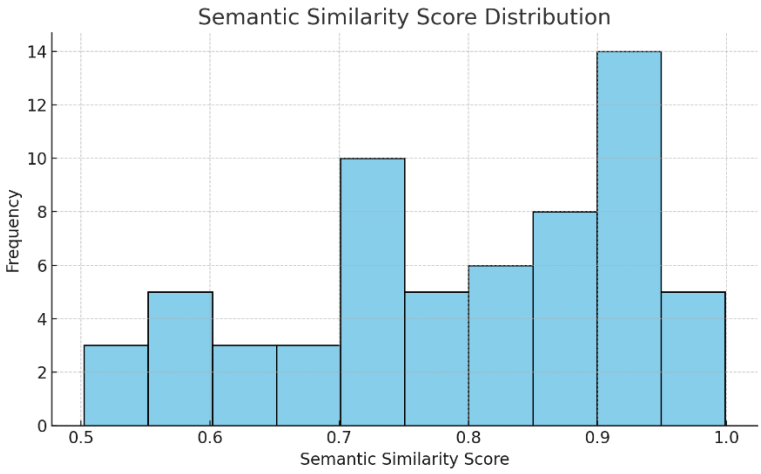

Figure 2: Semantic Similarity Score Distribution

As per Table 1, Sections having shorter paragraphs where there isn't much scope to summarize, or sections that are one-liners have higher scores (0.8 – 0.9).

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. For example, the DATROWAY drug, as shown in Table 2, has length passage, yet the model simplified it effectively without losing meaning and gave appropriate information. Another example is the drug AROMASIN, where all the numerical values were cut by the model and provided relevant information.

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.

Table 1: Comparing ChatGPT-generated responses with the Ground Truth text in the FDA label document with Semantic Similarity Score

Ground truth	Generated answer	Semantic score
PIQRAY is a kinase inhibitor indicated in combination with fulvestrant for the treatment of adults with hormone receptor (HR)-positive, human epidermal growth factor receptor 2 (HER2)-negative, PIK3CA-mutated, advanced or metastatic breast cancer as detected by an FDA-approved test following progression on or after an endocrine-based regimen. (1)	PIQRAY is a kinase inhibitor indicated in combination with fulvestrant for the treatment of adults with hormone receptor (HR)-positive, human epidermal growth factor receptor 2 (HER2)-negative, PIK3CA-mutated, advanced or metastatic breast cancer as detected by an FDA-approved test following progression on or after an endocrine-based regimen.	0.99
Tamoxifen: Do not use in combination with ARIMIDEX. No additional benefit seen over tamoxifen monotherapy. (7.1, 14.1) â€¢ Estrogen-containing products: Combination use may diminish activity of ARIMIDEX. (7.2)	Tamoxifen should not be used in combination with ARIMIDEX as no additional benefit is seen over tamoxifen monotherapy. Additionally, combination use of ARIMIDEX with estrogen-containing products may diminish the activity of ARIMIDEX.	0.97
Recommended Dose: 300 mg (two 150 mg tablets) taken orally once daily with food. (2.2) â€¢ For adverse reactions, consider dose interruption, dose reduction, or discontinuation. (2.3)	Recommended Dose: 300 mg (two 150 mg tablets) taken orally once daily with food.	0.96
Strong CYP 3A4 inducers: Concomitant use of strong CYP 3A4 inducers decreases exemestane exposure. Increase the AROMASIN dose to 50 mg (2.2,7).	The document states that strong CYP 3A4 inducers can decrease exemestane exposure. It recommends increasing the AROMASIN dose to 50 mg when used concomitantly with strong CYP 3A4 inducers.	0.95
One 1 mg tablet taken once daily (2.1)	One 1 mg tablet taken once daily.	0.95

Table 2a: Comparing ChatGPT-generated responses with the Ground Truth text in FDA label document with Semantic Similarity score between (0.6-0.75)

Drug name	Ground truth	Generated answer	Semantic score
DATROWAY (What are the warnings?)	Interstitial Lung Disease (ILD) and Pneumonitis: DATROWAY can cause severe and fatal cases of ILD/pneumonitis. Monitor for new or worsening signs and symptoms of ILD/pneumonitis. If ILD/pneumonitis is suspected, withhold DATROWAY and initiate corticosteroids. Permanently discontinue DATROWAY in patients with confirmed Grade 2 or higher ILD/pneumonitis. (5.1)? Ocular Adverse Reactions: DATROWAY can cause ocular adverse reactions including dry eye, keratitis, blepharitis and meibomian gland dysfunction, increased lacrimation, conjunctivitis, and blurred vision. Monitor patients for ocular adverse reactions during treatment with DATROWAY. Advise patients to use preservative-free lubricating eye drops and to avoid using contact lenses during treatment with DATROWAY. Dose delay, dose reduce, or permanently discontinue DATROWAY based on the severity of ocular adverse reactions. Refer patients to an eye care professional for any new or worsening ocular signs and symptoms. (2.2, 2.3, 5.2)? Stomatitis/Oral Mucositis: DATROWAY can cause stomatitis, including mouth ulcers and oral mucositis. Advise patients to use a steroid-containing mouthwash when starting treatment and to hold ice chips or ice water in mouth during the infusion of DATROWAY. Based on the severity of the adverse reaction, withhold, dose reduce, or permanently discontinue DATROWAY. (2.2, 2.3, 5.3)? Embryo-Fetal Toxicity: DATROWAY can cause fetal harm. Advise patients of potential risk to a fetus and to use effective contraception. (5.4, 8.1, 8.3)	The warnings for DATROWAY include Interstitial Lung Disease (ILD) and Pneumonitis, Ocular Adverse Reactions, Stomatitis/Oral Mucositis, and Embryo-Fetal Toxicity.	0.67

Drug name	Ground truth	Generated answer	Semantic score
DATROWAY (What are the precautions?)	Interstitial Lung Disease (ILD) and Pneumonitis: DATROWAY can cause severe and fatal cases of ILD/pneumonitis. Monitor for new or worsening signs and symptoms of ILD/pneumonitis. If ILD/pneumonitis is suspected, withhold DATROWAY and initiate	The precautions for DATROWAY include monitoring for interstitial lung disease (ILD) and pneumonitis, monitoring for ocular adverse reactions such as dry eye and blurred vision, advising patients to use a steroid-containing mouthwash for stomatitis, and cautioning about embryo-fetal toxicity and the need for effective contraception.	0.64
AROMASIN (What are the side effects?)	Early breast cancer: Adverse reactions occurring in ?10% of patients in any treatment group (AROMASIN vs. tamoxifen) were hot flushes (21% vs. 20%), fatigue (16% vs. 15%), arthralgia (15%vs. 9%), headache (13% vs. 11%), insomnia (12% vs. 9%), and increased sweating (12% vs.10%). Discontinuation rates due to AEs were similar between AROMASIN and tamoxifen (6% vs. 5%). Incidences of cardiac ischemic events (myocardial infarction, angina, and myocardial ischemia) were AROMASIN 1.6%, tamoxifen 0.6%. Incidence of cardiac failure: AROMASIN 0.4%, tamoxifen 0.3% (6, 6.1). Advanced breast cancer: Most common adverse reactions were mild to moderate and included hot flushes (13% vs. 5%), nausea (9% vs. 5%), fatigue (8% vs. 10%), increased sweating (4% vs. 8%), and increased appetite (3% vs. 6%) for AROMASIN and megestrol acetate, respectively (6, 6.1). To report SUSPECTED ADVERSE REACTIONS, contact Pfizer Inc at 1-800-438-1985 or FDA at 1-800-FDA-1088 or www.fda.gov/medwatch.	The document lists the following common side effects of AROMASIN: For early breast cancer: hot flushes, fatigue, arthralgia, headache, insomnia, increased sweating For advanced breast cancer: hot flushes, nausea, fatigue, increased sweating, increased appetite	0.69`

5. POTENTIAL ENHANCEMENTS

- In the future, this solution can be extended with Automatic flagging of low-similarity responses for human review in critical cases and implement follow-up question handling so the Chatbot can ask clarifying questions if the query is vague.
- Link the system with the FDA's official database to automatically fetch the latest drug label.
- Additionally, this model can also be customized for specific departments within an organization for domain-specific operational needs.

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] H. Fang *et al.*, "FDALabel for drug repurposing studies and beyond," *Nat Biotechnol*, vol. 38, no. 12, pp. 1378–1379, Dec. 2020, doi: 10.1038/s41587-020-00751-0.

[2] L. Ying *et al.*, "Text summarization with ChatGPT for drug labeling documents," *Drug Discovery Today*, vol. 29, no. 6, p. 104018, Jun. 2024, doi: 10.1016/j.drudis.2024.104018.

[3] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.

[4] S. Pudasaini and S. Shakya, "Question Answering on Biomedical Research Papers using Transfer Learning on BERT-Base Models," in *2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Oct. 2023, pp. 496–501. doi: 10.1109/I-SMAC58438.2023.10290240.

[5] Y. Kim *et al.*, "Medical Hallucinations in Foundation Models and Their Impact on Healthcare," Feb. 26, 2025, *arXiv*: arXiv:2503.05777. doi: 10.48550/arXiv.2503.05777.

[6] Y. Liu *et al.*, "Toward a Large Language Model-Driven Medical Knowledge Retrieval and QA System: Framework Design and Evaluation," *Engineering*, vol. 50, pp. 270–282, Jul. 2025, doi: 10.1016/j.eng.2025.02.010.

[7] H. Yang *et al.*, "Large Language Model Synergy for Ensemble Learning in Medical Question Answering: Design and Evaluation Study," *Journal of Medical Internet Research*, vol. 27, no. 1, p. e70080, Jul. 2025, doi: 10.2196/70080.

[8] H. Yang *et al.*, "LLM-MedQA: Enhancing Medical Question Answering through Case Studies in Large Language Models," Jan. 18, 2025, *arXiv*: arXiv:2501.05464. doi: 10.48550/arXiv.2501.05464.

[9] V. Srinivasan, V. Jatav, A. Chandrababu, and G. Sharma, "On the Performance of an Explainable Language Model on PubMedQA," Apr. 07, 2025, *arXiv*: arXiv:2504.05074. doi: 10.48550/arXiv.2504.05074.

[10] J. Han, J. Park, J. Huh, U. Oh, J. Do, and D. Kim, "AscleAI: A LLM-based Clinical Note Management System for Enhancing Clinician Productivity," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, in CHI EA '24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–7. doi: 10.1145/3613905.3650784.

[11] I. Salehin, M. T. A. Sajib, N. H. Badhon, M. S. H. Rifat, N. Amin, and N. N. Moon, "Systematic Literature Review of LLM-Large Language Model in Medical: Digital Health, Technology and Applications", Accessed: Aug. 06, 2025. [Online]. Available: https://www.authorea.com/doi/full/10.22541/au.174587258.81848862?commit=3b3611e288615e28cf0823750da3e6ca177cd775

[12] M. U. Hadi *et al.*, “Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects,” Sep. 05, 2024. doi: 10.36227/techrxiv.23589741.v7.
[13] D. K. Pham and B. Q. Vo, “Towards Reliable Medical Question Answering: Techniques and Challenges in Mitigating Hallucinations in Language Models,” Aug. 25, 2024, *arXiv*: arXiv:2408.13808. doi: 10.48550/arXiv.2408.13808.
[14] S. Patil and V. Lokesha, “Live Twitter Sentiment Analysis Using Streamlit Framework,” May 25, 2022, *Social Science Research Network, Rochester, NY*: 4119949. doi: 10.2139/ssrn.4119949.
[15] D. Chandrasekaran and V. Mago, “Evolution of Semantic Similarity—A Survey,” *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–37, Mar. 2022, doi: 10.1145/3440755.
[16] M. T. Colangelo, M. Meleti, S. Guizzardi, E. Calciolari, and C. Galli, “A Comparative Analysis of Sentence Transformer Models for Automated Journal Recommendation Using PubMed Metadata,” *Big Data and Cognitive Computing*, vol. 9, no. 3, Art. no. 3, Mar. 2025, doi: 10.3390/bdcc9030067.