

PP23 A framework enabling LLMs into regulatory environment for transparency and trustworthiness and its application to drug labeling document

Leihong Wu², Joshua Xu², Shraddha Thakkar¹, Magnus Gray², Yanyan Qu², Dongying Li², Weida Tong²

¹Office of Computational Sciences (OCS), Office of Translational Sciences (OTS), Center for Drug Evaluation and Research (CDER), US FDA

² Division of Bioinformatics and Biostatistics (DBB), National Center for Toxicological Research (NCTR), US FDA,

Introduction

Regulatory agencies consistently deal with extensive document reviews, ranging from product submissions to both internal and external communications. Large Language Models (LLMs) like ChatGPT can be invaluable tools for these tasks, however present several challenges, most importantly the potential use of proprietary and private information. This combined with the need for customized functions for specific review needs and transparency and trustworthiness of the model's output, make the need for a local and customized solution imperative.

To tackle these challenges, we formulated a framework named AskFDALabel on FDA drug labeling documents that is a crucial resource in the FDA drug review process. AskFDALabel operates within a secure IT environment and comprises two key modules: a semantic search (Module S) and a Q&A/text-generation module (Module T). The Module S is built on word embeddings to enable comprehensive semantic queries within labeling documents. The Module T utilizes a tuned LLM to generate responses based on references from Module S. As the result, our framework enabled small LLMs to perform comparably to ChatGPT leveraging a computationally inexpensive solution for regulatory application.

To conclude, through AskFDALabel, we have showcased a pathway that harnesses LLMs to support agency operations within a secure environment, offering tailored functions for the needs of regulatory research.

Objectives and Challenges

- Objective: Develop a framework to Integrate LLMs in regulatory application
- Challenges:
- Review-centric functionalities
 - Reviewers often need functionalities to support internal ID search
 - Customized function with specific review requirements must be developed on top of the foundational models
- Transparency and explainability
 - Reviewers need a more transparent chain-of-thought and traceable evidence from the reliable data sources
 - Concern may arise regarding trustworthiness of the model, especially when the model is not transparent
- Data security
 - Regulatory data is sensitive and cannot be shared with public LLMs
 - Some text data used in regulatory research are not publicly accessible due to confidentiality or privacy considerations

Proposed Solutions

- A proof-of-concept study aimed at developing a framework to address these needs
- Developed a generic framework and demonstrated on FDA Drug Labeling documents.
 - It is a localized system (internal IT environment), customized to support review-centric queries, and search results linked back to the original documents.
- It has been tested on two LLMs (i.e., Llama-13b and Falcon-40b)
- Compare performance of these two models judged by GPT-4

Framework to facilitate LLMs application in Regulatory Setting

To address these challenges, we developed a Retrieval-Augmented Generation (RAG)-based framework to enhance the customization and transparency of LLMs. The proposed framework was also implemented in a local, secured IT environment, to meet the requested security needs of many regulatory documents.

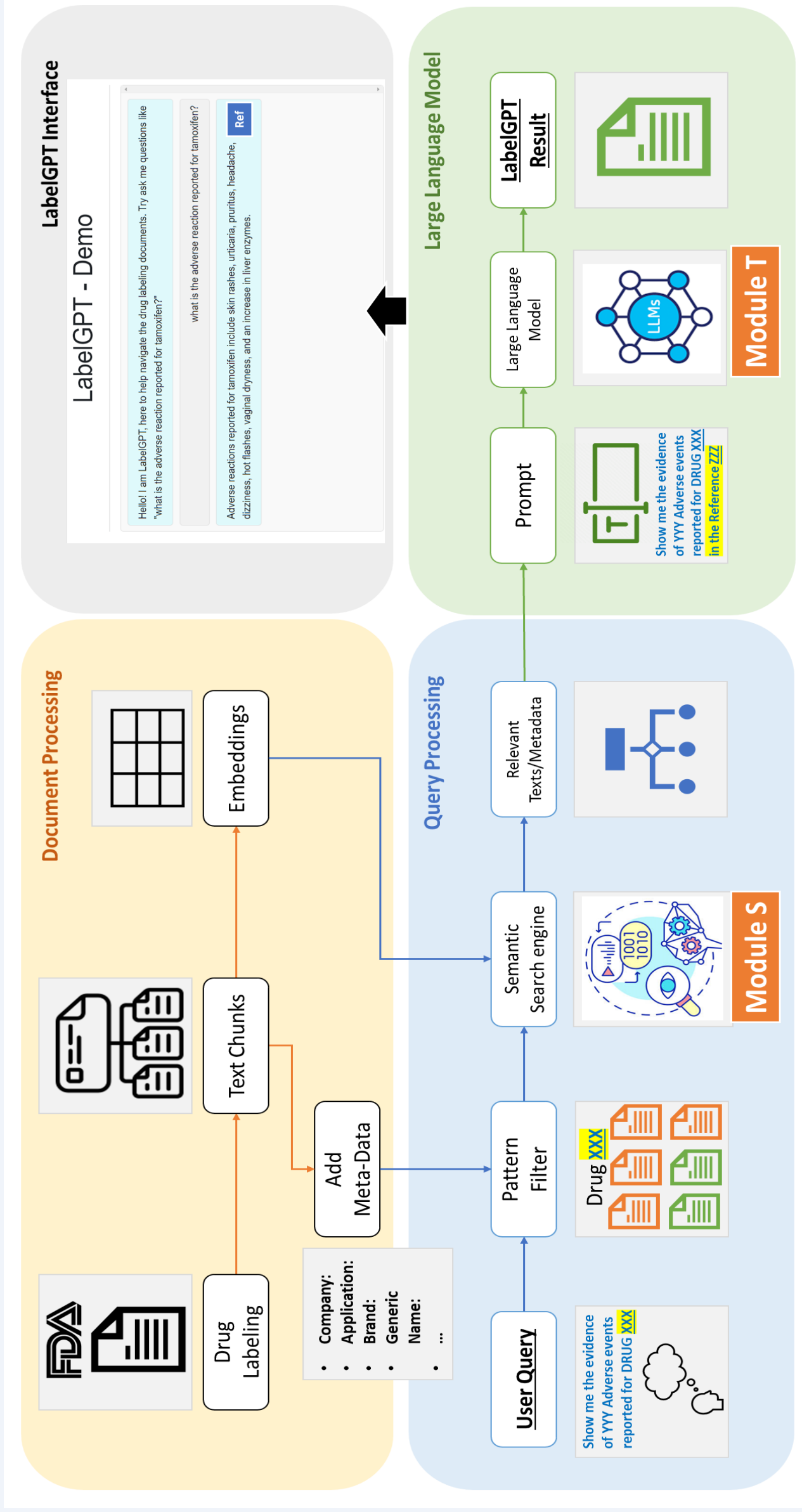


Figure. 1. AskFDALabel Framework. AskFDALabel contains four components: (a) Document processing, (b) Query processing, (c) LLM inference, and (d) User interface. The main purpose of Document and Query processing is to establish semantic search a query which subsequently forms the prompt for LLM to enhance the transparency and trustworthiness of the generative output.

Improving search relevance by module optimization

Entity	Search Type	Relevance	Improvement (over Regular)
Generic name	Regular	24.9% (462/1859)	
Generic name	Optimized	85.1% (1582/1859)	+ 60.2%
Product name	Regular	48.8% (1073/2201)	
Product name	Optimized	95.0% (2091/2201)	+ 46.2%
Application Number	Regular	0.05% (1/2153)	
Application Number	Optimized	99.16% (2135/2153)	+ 99.11%

Table 1. One important measure to assess search relevance was based on whether the query result originated from the correct labeling documents. We developed three experiments to mimic the real review usage scenario, involving search on three different types of entities – generic name, product name, and application number. For example, the query could be in the form of “*Tell me the adverse reactions of Lipofen?*” or “*Tell me the adverse reactions of NDA022006?*”. The output of module S was then compared to the ground truth – which was obtained by FDALabel database query of the designated entities. The result was considered a match if the output chunks of module S was originally from the true label. The overall relevance was measured by the ratio of matched cases to total tested cases. For each type of entity, around 2000 cases were tested for LLM to enhance the transparency and trustworthiness of the generative output.

Disclaimer: The information in this poster does not necessarily represent FDA's position or policy.



Case Study

DICTRANK	MODEL ANSWER	STATISTICAL TYPE	CHATGPT (V3.5)	ASKFDALABEL
MOST	Yes	TP	268	175
MOST	No	FN	27	113
NO	No	TN	165	280
NO	Yes	FP	146	18
		Accuracy	71.45%	77.65%
		Sensitivity	90.85%	60.76%
		Specificity	53.05%	93.96%

Table 2: Here we asked AskFDALabel the following question: “Does [Drug_Name] contain any cardio-related adverse events or risks?“, where the [Drug_Name] is from the list of DICTrank. AskFDALabel returned answers from the top 3 references, and the drug is considered as DICT concern if at least one answer said “Yes”. The averaged performance is summarized in Table 2. We also compared the AskFDALabel inference to the ChatGPT result. As shown in Table 2, AskFDALabel outperformed ChatGPT by overall accuracy as 77.65%–71.45%, where ChatGPT was better at Sensitivity and AskFDALabel was better at Specificity.

Summarizing multiple answers into one response using an additional summary LLM

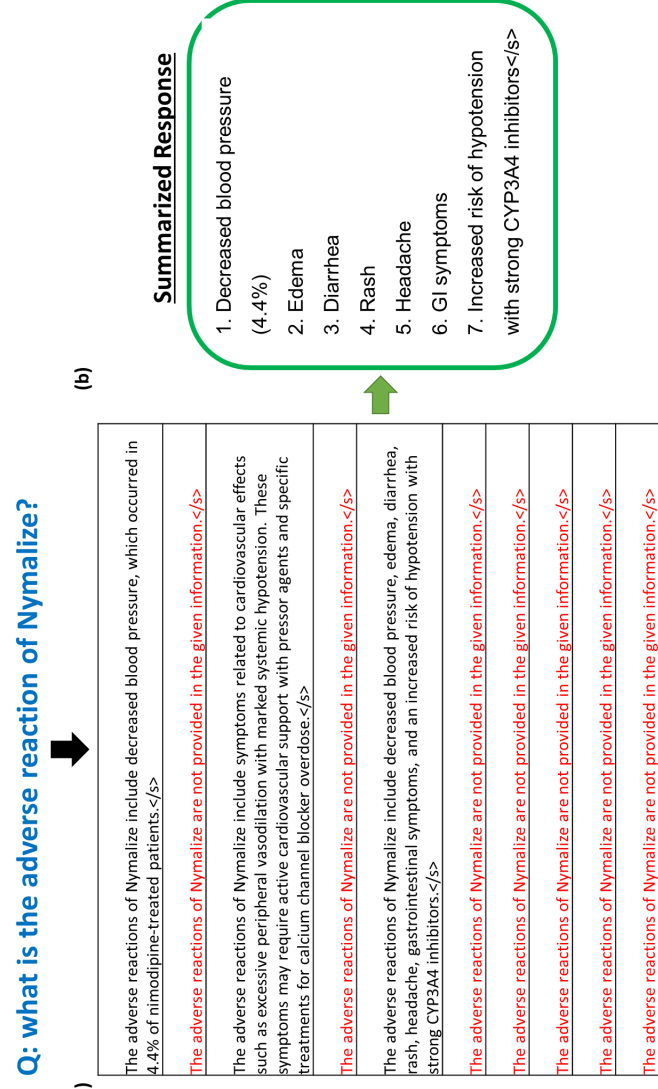


Figure 2. Based on the current Module S and T, the most relevant text chunks from Module S are passed to Module T separately, and each text chunk would then be used to generate one answer.

For example, AskFDALabel generated 10 answers if the top 10 text chunks were processed. It is worth noting that Module T may respond like “The adverse reactions of Nymalizer are not provided in the given information” if the provided text chunk is not considered to be a valid reference to the query.

Conclusion

In conclusion, while AskFDALabel framework presents substantial advancements in the analysis of drug labeling documents, it simultaneously highlights key areas for further research and development. The importance of respecting the confidentiality of sensitive data, optimizing system performance, and accommodating the unique needs of reviewers underscores the complexities of integrating AI within the pharmaceutical domain. Future work should continue to focus on these areas, with the ultimate goal of optimizing the AskFDALabel system to best support the vital work of FDA scientists and reviewers.

Acknowledgements

We would like to thank the CDER Innovation Board for feedback for use case within the FDA. This research project was funded and supported by NCTR, FDA (project ID: E0777301 and E0782201)