AskFDALabel: Enhancing AE Detection, Profiling, Classification, and Monitoring using FDA Labeling Document and Emerging Large Language Models

This presentation reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration. Any mention of commercial products is for clarification only and is not intended as approval, endorsement, or recommendation.

[1]Leihong Wu, [1]Joshua Xu, [2]Hong Fang, [1]Weida Tong,
[1]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. FDA, 3900 NCTR Rd, Jefferson AR, 72079
[2]Office of Scientific Coordination, National Center for Toxicological Research, U.S. FDA, 3900 NCTR Rd, Jefferson AR, 72079

Abstract

Adverse drug events (AEs) are a leading cause of mortality in the United States, with over 70,000 cases related to death reported annually since 2020.

FDA drug labeling documents are a critical resource for AE research, providing comprehensive and reliable drug safety information. However, manually extracting and classifying AE data from these documents is labor-intensive and time-consuming. Recent advancements in natural language processing (NLP) and large language models (LLMs) offer a promising, modernized solutions.

This presentation introduces AskFDALabel, an LLM-powered framework, to enhance the efficiency and effectiveness of AE research with drug labeling document.

We applied AskFDALabel on three experiments related to AE research: (1) DILI Classification, (2) DICT Classification, and (3) Drug AE Profiling. As a result, we observed promising results over all experiments.

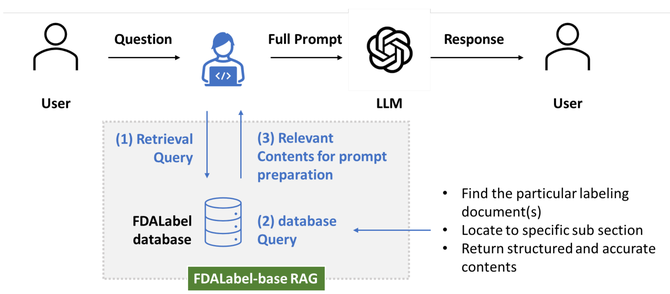

Figure 1. The AskFDALabel Framework with FDALabel-based RAG. Three key steps in the FDALabel-based RAG are (1) retrieval query process, (2) FDALabel database query, and (3) relevant contents for prompt preparation.

Experiment 1: Toxicity Classification

[Drug-Induced Liver Injury (DILI)] In total, 226 drug names were process in AskFDALabel with a query like "What is the DILI Class of KETAMINE?" Among them, 18 drugs cannot be retrieved due to their discontinuation status and labeling information is no longer available through FDALabel database. For the other 208 available drugs, we further combined the Most and Less DILI concern as DILI Positive, where No DILI concern is considered as DILI Negative, for statistical analysis.

[Drug-Induced Cardiotoxicity (DICT)] Similarly, we processed 1,184 drugs labeled with DICT concern levels—Most, Less, or No DICT concern. We successfully found and processed the labeling for 1,153 of these drugs. As with the DILI classification analysis, we combined the Most and Less DICT concern groups into a DICT-positive category.

Prompt used in DILI Classification, strictly followed the criteria defined by DILI scientists [1]

```
Drug Induced Liver Injury (DILI) is the topic of interest for your study on drug labeling document.
Check the given paragraph to see if any of the following DILI-relevant terms (or with similar meaning) are mentioned in the paragraph.
if one term occurs multiple times in the paragraph, only list once.
In addition, determine whether the keyword(s) reflect severe events of Drug Induced Liver Injury
### Severity determination
  The severity of Drug Induced Liver Injury should be determined by the following criteria and scores.
  [Score: 8] Fatal. [Description]hepatotoxicity: Death; fatal liver failure; or needed liver transplantation.
  [Score: 7] Acute liver failure.        [Description]Liver/hepatic failure; fulminant hepatic necrosis
  [Score: 6] Liver necrosis. [Description]Histologically confirmed liver necrosis caused by drug
  [Score: 5] Jaundice.        [Description]Jaundice (clinically apparent), if caused by drug-induced hepatocellular injury
  [Score: 4] Hyperbilirubinemia.      [Description]Hyperbilirubinemia without visible jaundice, if not due to other causes like Gilbert syndrome or cholestasis
  [Score: 3] Liver aminotransferases increase.            [Description]Liver aminotransferases increase (e.g. ALT, AST, transaminase, aminotransferase); abnormal liver/hepatic function test; liver/hepatic injury
  [Score: 2] Cholestasis; steatohepatitis.        [Description]Steatohepatitis, cholestasis, cholestatic hepatitis; liver/hepatic damage/disorder/impairment/toxicity/reaction; hepatitis; hepatopathy
  [Score: 1] Steatosis.        [Description]Steatosis; fatty liver;
### Output format:
  (Found.) keyword1; keyword2; [Severity]: [Score: 5] Jaundice
  (Not Found) No Keyword was found in the given paragraph
```

Table 1. Statistical Summary of DILI and DICT Classification Result.

	Approach	Accuracy	Recall	Precision	Specificity	F1-Score
DILI Classification	**Current Study**	**0.976**	0.993	**0.974**	**0.930**	**0.984**
	BERT Model [29]	0.927	**1.000**	0.784	0.900	0.879
	Keyword [29]	0.822	**1.000**	0.581	0.764	0.735
	XGBoost [28]	0.839	0.651	0.856	0.941	0.740
DICT Classification	**Current Study**	**0.895**	**0.894**	**0.967**	**0.898**	**0.929**
	Prev. LLM [20]	0.776	0.608	0.907	0.940	0.728
	ChatGPT (3.5) [20]	0.715	0.908	0.647	0.531	0.756

Experiment 2: Drug AE Profiling

To evaluate the performance of creating AE profile from labeling documents, we processed 200 drugs for which their AE profiling have been manually annotated by human in the TAC 2017 challenge.

First, we used the original 476 samples from the TAC dataset to for a comparative study between using LLM and other approaches. Both exact and semantic based matching were measured.

Then, we used AskFDALabel template to perform a real-time AE profiling generation, which instead of using the original contents from TAC 2017, will retrieve the most up-to-date labeling documents based on the user query which included the drug name.

34066-1 - The Adverse Events reported in this section are listed as below:
[AE1]: Hypersensitivity reactions, including anaphylaxis, have been reported during or after the administration of raxibacumab by intravenous infusion. $$infusion reaction$$
[AE2]: Anaphylaxis has been reported during or after the administration of raxibacumab by intravenous infusion. $$infusion-related reaction$$
… …
43685-7 - The Adverse Events reported in this section are listed as below:
[AE1]: Rash ##rash## $$rash papular$$; $$rash erythematous$$; $$infusion-related rash$$
[AE2]: Urticaria ##urticaria##
[AE3]: Pruritus ##pruritus##
[AE4]: Chills $$infusion-related reaction$$
[AE5]: Chest tightness
… …
34084-4 - The Adverse Events reported in this section are listed as below:
[AE1]: Injection site reaction $$infusion-related reaction$$
[AE2]: Erythema $$rash erythematous$$
[AE3]: Pain
[AE4]: Headache
[AE5]: Rash ##rash## $$infusion-related rash$$; $$rash papular$$; $$rash erythematous$$
[AE6]: Pain in extremity ##pain in extremity##
[AE7]: Pruritus ##pruritus##
[AE8]: Somnolence ##somnolence##
… …

Figure 2. An example of Drug AE profiling contents generated by AskFDALabel (Specific Mode1) for RAXIBACUMAB. Hash tags indicated human-annotated AE either exactly (##) or semantically ($$) matched to this finding. Yellow highlights are similar AE terms that reported by human annotator, but not exactly found in the generated AE profile.

Table 2. Performance on TAC 2017 dataset and Drug AE Profiling

	Approaches	Precision	Recall	F1-Score
TAC Original Text Contents	Current Study - Exact	0.858	0.825	0.827
	Current Study - Semantic	**0.906**	**0.939**	**0.911**
	TAC2017 – SOTA	0.851	0.853	0.852
	RxBERT	0.893	0.885	0.889
Drug Name Only	Exact – up-to 1024 Tokens output	0.758	0.769	0.753
	Semantic - up-to 1024 Tokens output	0.840	0.895	0.859
	Semantic - up-to 4096 Tokens output	**0.848**	**0.904**	**0.870**

Conclusion

- For DILI and DICT Classification, the overall consistency between AskFDALabel and previous human annotation were **0.984** and **0.929** by F1-score, respectively, significantly better than other toxicity classification approaches.
- For Drug AE detection, the overall F1-score was **0.911**, outperformed than previous NLP approaches. Moreover, even using the drug name instead of actual text contents, the F1-score still reached **0.870**.
- These high consistency performance results indicated AskFDALabel can automate the process of manual review process offering unprecedented speed and accuracy in AE detection and classification.
- Additionally, AskFDALabel provided the cited reference as well as the detailed explanations for the answer, which could further assist the reviewers to manual check and verify the LLM-generated answers.