Leveraging Large Language Models (LLMs) to Support Regulatory Assessments

AAPS 2024 PHARMSCI 360:
Track: Discovery and Basic Research

Theme 1- Leveraging AI in the Processes of Drug Discovery, Lead Validation, Preclinical Development, and Regulatory Submissions

Jing (Jenny) Wang, Ph.D.

Staff Fellow, Division of Quantitative Methods and Modeling
Office of Research and Standards, Office of Generic Drugs
Center for Drug Evaluation and Research | U.S. Food and Drug Administration

October 21, 2024

Disclaimer

- This presentation reflects the views of the author and should not be construed to represent FDA's views or policies.

Learning Objectives

- To acknowledge the potential of large language models (LLMs) in supporting generic drug product development and regulatory assessment.
- To present a case study – leveraging LLMs to generate food effect summaries from new drug review documents.

The contents of this presentation are from our recent published paper "*Leveraging GPT-4 for Food Effect Summarization to Enhance Product-Specific Guidance Development via Iterative Prompting*." Journal of Biomedical Informatics (2023): 104533.

Background

- Text summary of key elements from lengthy documents in the regulatory assessments can be labor-intensive and time-consuming.
- LLMs have raised interest in their potential for automatic text summarization.
- Evaluation for the quality of LLM-generated summaries for regulatory review-related tasks may provide foundational evidence for future LLM-based applications in the regulatory field.

Background

- Food effect summary was used as an example for the evaluation of LLM-generated summaries
  - Both source documents (new drug reviews from Drugs@FDA) and golden reference documents (drug labeling) are publicly available
  - Food effect summary provides critical information for product-specific guidance (PSG) development, especially for oral drug products, as it is an essential aspect of drug absorption and it can be influenced by drug formulations through various mechanisms.
  - FDA publishes PSGs to describe FDA's current thinking and expectations on how to develop generic drug products therapeutically equivalent to specific reference listed drugs (RLDs) to facilitate generic drug development and accelerate Abbreviated New Drug Application (ANDA) submissions.

Research Questions

1. To what extent are LLMs capable to generate a summary factually consistent with the golden reference summary?
2. Which LLM model (ChatGPT vs. GPT-4) performs better for the summary-generation tasks with the food effect-related text?
3. Can the multi-turn iterative conversation approach with a chatbot enhance the quality of LLM-generated summary?

Data

- The food effect study documents from NDA review files available to the public via Drugs@FDA website, which serve as the article to be summarized – LLM input
- The food effect summary from the FDA-approved drug labeling – golden reference summary
- One hundred selected New Drug Application (NDA) drugs from the past 5 years (from 2019 to present), which provide a comprehensive coverage of the NDA drugs in terms of clinical category, drug substance, and dosage form

Methods

Models

- ChatGPT (GPT-3.5)
- GPT-4

Developed Prompt Method - Iterative Prompting

Evaluation Tasks

Task 1: Comparison between Summaries of Different Prompt Iterations

Task 2: Comparison between ChatGPT and GPT-4

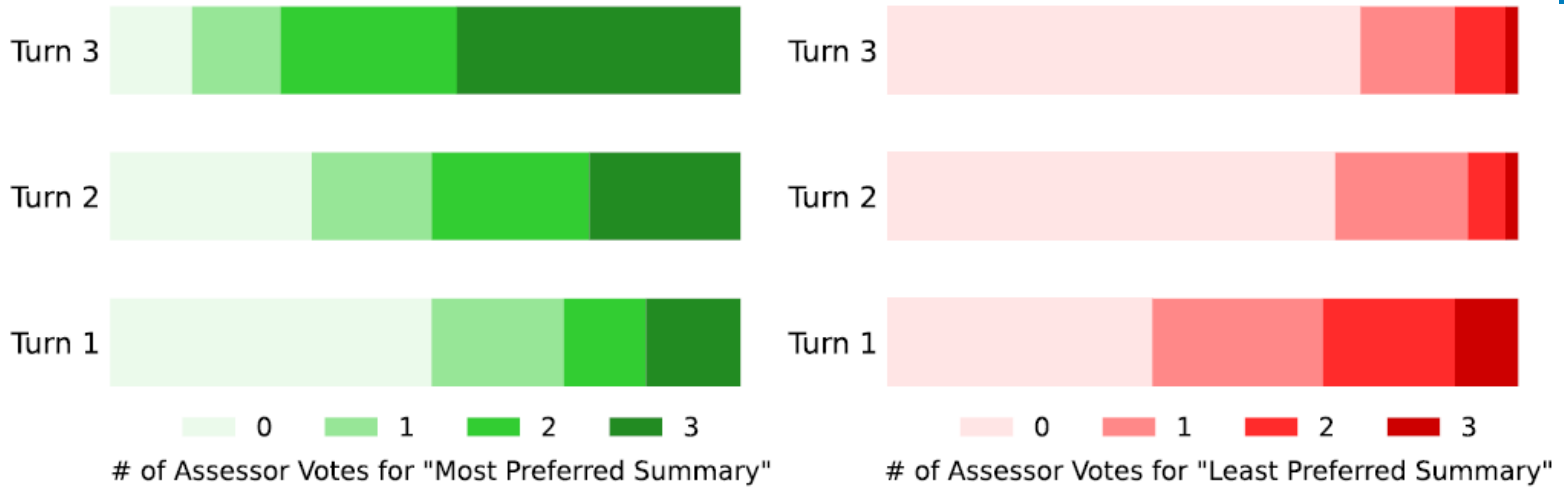
Task 3: GPT-4 Consistency Evaluation

Methods

Evaluation metrics

- Automated Metrics: Recall-Oriented Understudy for Gisting Evaluation (ROUGE)
- Human Evaluation: three independent FDA professionals
- GPT-4 Evaluation: a text evaluator using large language models based on relevance, coherence, consistency, and fluency

Results

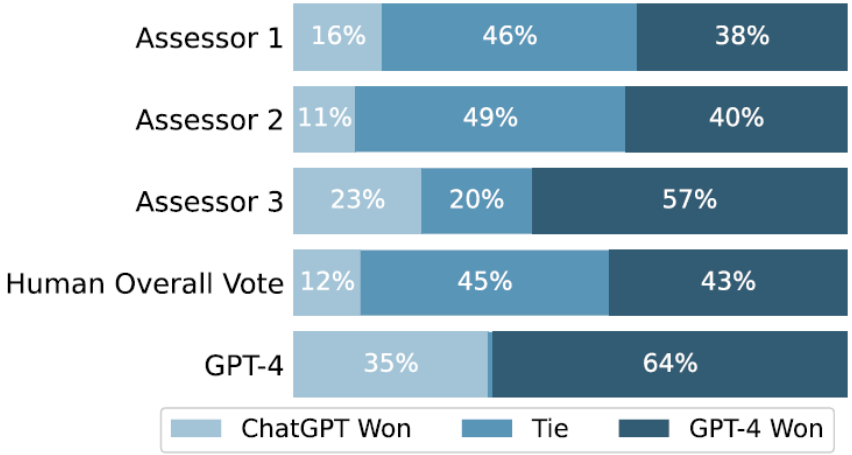


- Distribution of assessor votes for the most and least preferred summaries at each turn during the iterative prompting process. Summaries from Turn 3 (the last turn) received the highest number of votes in the “Most Preferred Summary”.

Results

Table 1
ROUGE scores (1/2/L) of summaries at each turn. Bold indicates the best score among the turns for each metric.

	ROUGE-1/2/L	
	ChatGPT	GPT-4
Turn1	29.40/11.24/19.48	31.44/12.72/20.45
Turn2	32.67/**14.06**/21.75	31.37/**12.88**/20.98
Turn3	**34.04**/13.36/**22.60**	**33.64**/11.44/**22.12**



- Results indicate that the summary becomes increasingly similar to the golden reference summary as the iterative process progresses.
- Human assessors and GPT-4 preferred the summaries generated by GPT-4 over ChatGPT.

Results

- The majority vote of the human annotations indicates that 85% of the GPT-4 generated summaries are factually consistent with the golden reference summary.
- GPT-4 evaluation shows a consistency rate of 72% of the GPT-4 generated summaries are factually consistent with the golden reference summary.
- A strong consistency of rating between human and GPT-4 as both share 69% overlap.

Summary

- Extensive evaluations were conducted on the quality of LLM-generated summaries with the food effect text from one hundred NDA drugs selected over the past 5 years.
- The LLM-generated summary quality is progressively improved throughout the iterative prompting process through keyword-focused and length-controlled prompts in consecutive turns.
- All the three FDA professionals unanimously rated that 85% of the food effect summaries generated by GPT-4 are factually consistent with the golden reference summary.
- A great potential of LLM was indicated to provide draft text summaries for regulatory review-related tasks that could be reviewed by FDA professionals, thereby improving the efficiency of the PSG development and promoting generic drug product development.

Acknowledgement

Office of Research and Standards, Office of Generic Drugs, Center for Drug Evaluation and Research, U.S. Food and Drug Administration

- Meng Hu, PhD
- Ping Ren, PhD
- Yi Zhang, PhD
- Biao Han, PhD
- Liang Zhao, PhD (currently work in the School of Pharmacy, University of California San Francisco)
- Andrew Babiskin, PhD
- Lanyan Fang, PhD
- Lei K. Zhang, PhD
- Robert Lionberger, PhD

Drexel University, School of Biomedical Engineering, Science and Health Systems

- Yiwen Shi, PhD
- Taha ValizadehAslani, PhD
- Felix Agbavor, MS
- Hualou Liang, PhD

U.S. FOOD & DRUG ADMINISTRATION

Questions?

Jing (Jenny) Wang, Ph.D.

Staff Fellow, Division of Quantitative Methods and Modeling
Office of Research and Standards, Office of Generic Drugs
CDER | U.S. FDA

jing.wang1@fda.hhs.gov


FDA
U.S. FOOD & DRUG
ADMINISTRATION