Summarizing FAERS Narratives with Generative AI: Methods, Resource Requirements, and Quality Assessment

Kate Dowdy[1], Anna Hoffman[1], Tyree Giles[1], David Kugele[1], Jacqueline Guill[1], Isaac Chang[2], Gregory Jackson[2]

[1]Booz Allen Hamilton

[2]Center for Drug Evaluation and Research, U.S. Food and Drug Administration

Abstract

Problem Statement: Can we trust artificial intelligence to summarize adverse event reports?

In the last decade, over 20 million cases have been submitted to the FDA Adverse Event Reporting System (FAERS). Understanding the signals and trends in adverse event reports is critical to the FDA's mission to keep U.S. consumers safe from emerging and evolving threats to public health. Due to the volume and complexity of these cases, **reviewing and synthesizing the information in FAERS currently requires significant manual effort** from experts within the Office of Surveillance and Epidemiology (OSE).

Given the advances in Large Language Models (LLMs) and generative Artificial Intelligence (AI), the Real-time Application Platform for Innovation and Development (RAPID) team set out to explore the possibilities and limitations of using generative AI to summarize FAERS narratives. We tested two methods for creating machine-written summaries:

1) finetuning an LLM with FAERS data
2) prompting out-of-the-box LLMs with example summaries.

We evaluated model outputs by comparing them to human-written summaries; analyzing token probabilities to detect hallucinations; and comparing LLM-detected entities in summaries vs. those in the narratives. Overall, we found that our **models produced fluent, human-readable results that contained relevant information, but that AI-written summaries often fell short of encompassing the key facts or accurately representing case details.**

Introduction

The launch of Open AI's ChatGPT in 2022 ushered in a new level of excitement and interest in AI for the workplace. Thanks to the availability of LLMs trained on enormous amounts of data, generative AI is sophisticated enough to passably perform a variety of administrative tasks – from retrieving and synthesizing information across documents, to drafting new content from prompts. However, there are good reasons to be skeptical of generative AI:

- **Models are biased by their training data.** Furthermore, many state-of-the-art foundation models are trained on closed datasets, which makes it harder to anticipate the ways in which the model will be biased.
- **Models do not ascribe meaning to language the same way humans do.** This does not prevent LLMs from generating convincing, confident responses to prompts that may be factually inaccurate, misleading, or entirely fabricated.
- **There is no one-size-fits-all method for ensuring models generate truthful and trustworthy content.** As a result, human reviewers must assume the responsibility of auditing and correcting AI-generated content.

Generative AI promises to revolutionize knowledge curation, but can we trust it? To explore this question, we attempted to generate the best possible AI-written summaries of adverse event reports using resources on the RAPID workbench and two different methods of tailoring model outputs for the FAERS domain.

Materials and methods

Method 1: Finetuning a custom LLM

Model: We chose the **Long T5 Base** from Google's Text-to-Text Transfer Transformer (T5) suite of models. We chose T5 because of its versatility, open-source status, and solid performance on summarization tasks. T5 was trained on the Colossal Cleaned Corpus Dataset - a massive repository of text from the internet that is cleaner than previous versions (filtered out non-narrative text and abusive language). The *longformer* is a version of the standard transformer specifically designed to handle longer input sequences (ex. summarizing long documents).

Training: Training was completed in two steps.

a) **First, we *domain-trained* our model for one epoch on FAERS narratives** (~1M cases from 2021), in addition to a sample from our finetuning data. In this step, the model learns to guess parts of text sequences that are hidden (the "denoising objective"), which adjusts the pretrained model weights to better conform to the specific syntax, vocabulary, and style of a given domain.

b) **Then, we *finetuned* our model for the task of summarizing medical documents**. Because we did not have many examples of cases and their summaries from FAERS itself, we used open-source datasets for this finetuning step: the **Pubmed Summarization** dataset (260k scientific articles and their abstracts), and the **Multidocument Summarization for Literature Review** (MSLR 2022) dataset (22k sets of collections of medical article abstracts and corresponding literature reviews). These datasets are in the scientific/medical domain (FAERS-adjacent) and popular for summarization training and benchmarking.

Training a model to summarize long (or multiple) documents presents unique challenges. The default behavior for handling long sequences of text is to truncate them to fit within the maximum context window of the model (for most transformers, 512 tokens/~400 words). Truncation eliminates any information provided later in the document (or entire documents in a series of documents.) Even for models designed to take longer inputs (like the longformer), the hardware for training constrains how long the context window can be. For our finetuning run, we discarded all very short/ long examples and chunked any inputs longer than ~4000 tokens. We finetuned on this data limited to a max input length of 4096 tokens for 10 epochs.

Hardware: Training was completed on a p3dn.24xlarge AWS GPU instance (8.5 hours).

Method 2: Prompting an out-of-the-box (OOB) LLM

Models: We chose to compare three small out-of-the-box models: **T5 Base**, **Gemma 2B**, and **Falconsai for Medical Summarization**. These open-source models had been tested by RAPID previously and performed decently on summarization benchmarks.

Prompting: To improve results from a model that had never seen FAERS data before, we included an example case and corresponding human-written summary from OSE Pharmacovigilance Reviews in our prompts. We found a single, high-quality example tended to outperform zero-shot prompting (vs. using more examples of varying quality.)

Hardware: Inferencing was completed on the RAPID CPU Workbench.

Case Study: AI-Generated Summary vs. Human-Written Summary

At first glance, these summaries of a 160-word FAERS narrative appear comparable. On closer examination, there are some key points where they diverge...

An endocrinologist reported a 60-year-old male patient of Hispanic/Latino origin who experienced an adverse event after receiving tirzepatide (Mounjaro). The patient developed a rash and throat closed 30 minutes after receiving the first dose of tirzepatide. He was admitted to the hospital and received steroids and diphenhydramine hydrochloride. The event was considered as serious by the company due to its medical significance. The patient was recovered from the event.

This summary (69 words) was generated by Gemma 2B (Method 2).

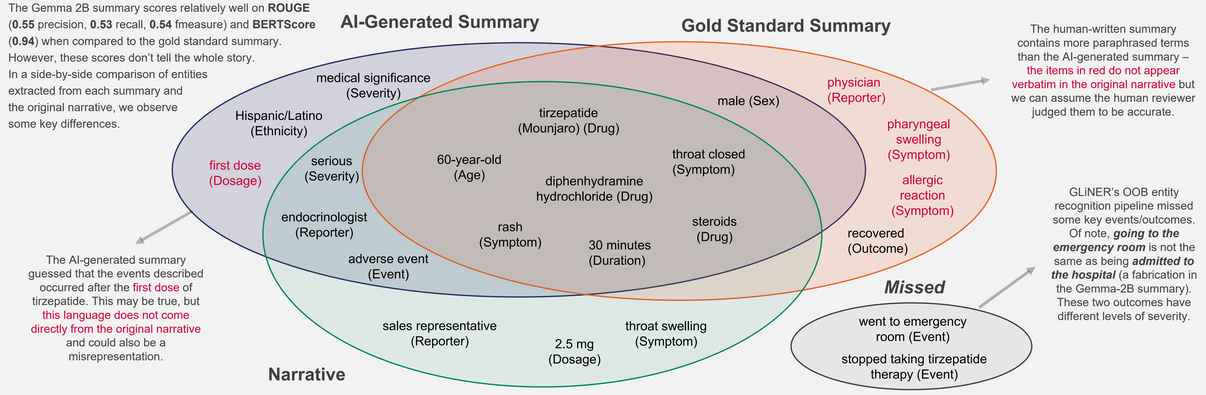
A physician reported a 60-year-old male patient who experienced an allergic reaction 30 minutes after tirzepatide injection. Past medical history and concomitant medications were not reported. It was reported that the patient developed "a rash and throat closed" (PTs Rash and Pharyngeal swelling). The patient went to the emergency department and received "steroids (unspecified) and diphenhydramine hydrochloride." It was reported that the patient recovered and stopped tirzepatide.

This summary (67 words) was written by a human reviewer.

How Similar Are These Summaries?

The Gemma 2B summary scores relatively well on **ROUGE** (**0.55** precision, **0.53** recall, **0.54** fmeasure) and **BERTScore** (**0.94**) when compared to the gold standard summary. However, these scores don't tell the whole story. In a side-by-side comparison of entities extracted from each summary and the original narrative, we observe some key differences.

Entities Extracted from Text (GLiNER)



The AI-generated summary guessed that the events described occurred after the first dose of tirzepatide. This may be true, but this language does not come directly from the original narrative and could also be a misrepresentation.

The human-written summary contains more paraphrased terms than the AI-generated summary – the items in red do not appear verbatim in the original narrative but we can assume the human reviewer judged them to be accurate.

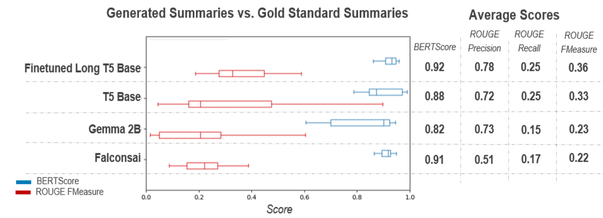
GLiNER's OOB entity recognition pipeline missed some key events/outcomes. Of note, ***going to the emergency room*** is not the same as being ***admitted to the hospital*** (a fabrication in the Gemma-2B summary). These two outcomes have different levels of severity.

Results and discussion

Evaluating machine-written text

OSE stakeholders shared a set of 11 pharmacovigilance reports focused on hypersensitivity, and from those reports, the RAPID team extracted 26 human-written summaries of FAERS cases that comprise our gold standard dataset.

Comparison to the Gold Standard: We compared AI-written summaries to the gold standard summaries using two metrics that compare text similarity: **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)**, which calculates the n-gram overlap between two texts and rewards exact phrasing, and **BERTScore**, which measures semantic similarity and is more forgiving of paraphrasing. We found the **Finetuned Long T5 Base** (Method 1) on average outperformed other models on these metrics.



Average Scores

	BERTScore	ROUGE Precision	ROUGE Recall	ROUGE FMeasure
Finetuned Long T5 Base	0.92	0.78	0.25	0.36
T5 Base	0.88	0.72	0.25	0.33
Gemma 2B	0.82	0.73	0.15	0.23
Falconsai	0.91	0.51	0.17	0.22

Measuring Veracity: To test whether our models were producing *hallucinations* (fabricated information), we analyzed the generated token probability scores for our model outputs - a method to detect how much the model is guessing in its answer. This method failed to detect meaningful patterns, even when applied to a full-blown hallucinated answer produced by Gemma 2B in response to an incomplete prompt. **We observed no total hallucinations for complete prompts.**

Extracting the Facts: Stakeholders indicated that it was most important for a summary to cover the key facts of a FAERS case accurately and completely. To evaluate this aspect of AI summaries, we used an LLM-based Named Entity Recognition (NER) and Relation Extraction (RE) pipeline called **GLiNER** to extract key pieces of information from FAERS narratives and their summaries.

Hand Evaluation: Hand evaluation was necessary to assess the overall quality of our model outputs, which poses a significant challenge for automating assessment of generative AI. **We observed the best fluency (readability) from the T5 Base and Gemma 2B.** The finetuned models repeated information more often. All models were susceptible to skipping or misrepresenting important narrative sequences.

Conclusion

In use cases such as ours, where limited gold standard data is available and the task for the model is relatively straightforward, **prompting an OOB LLM may produce more coherent results with a smaller investment in time and resources than finetuning a model on a domain-adjacent dataset**. Automated evaluation of AI summaries remains challenging, but a combination of text comparison metrics and entity comparison provided insight into the strengths and weaknesses of different models we tested for summarizing FAERS cases. Leveraging LLMs for knowledge extraction from FAERS is a promising avenue for future exploration and may have benefits for any generative AI evaluation or multi-case pattern analysis.

We would like to acknowledge the support and assistance in interpreting the context of FAERS reports provided by Monica Munoz, Lisa Wolf, and Oanh Dang from the Office of Surveillance and Epidemiology (OSE). This project was approved by the CDER Innovation Board.