Mining Data Outside the Box: Internet as a New Source for Common Data Model

< Min-Gyu Kim MD>[1,2], < Min Ho An, MD >[1,2], <GyuBeom Hwang MD>[1,2], <Rae Woong Park, MD, Ph.D.>[1]

[1]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea

[2]Department of Medical Sciences, Graduate School of Ajou University, Suwon, Republic of Korea

Background

While the Observational Medical Outcomes Partnership(OMOP) Common Data Model (CDM) standardizes data acquired in healthcare settings, EHR data is not the only source of healthcare data. The internet such as social media, patient forums, and other online sources can also be a valuable source of real-world health data.[1] It can potentially be used to study a variety of healthcare topics, such as identifying risk factors, detecting adverse drug reactions, or patient experience. However, internet data is not as easy to handle as CDM. It is often unstructured and can be difficult to extract meaningful information from.

In this paper, we present our first step in extracting and formatting medical data mined from the internet into OMOP-CDM. A certain degree of deduction is necessary to use texts from internet as a source to feed OMOP-CDM. To tackle this problem, we used a generative large language model (LLM) to generate text about the logical flow of extraction.

The objectives of this study are:

- Proof-of-concept in building CDM out of a data source outside the healthcare system
- An extraction flow that uses generative language models to perform logical deductions necessary for extracting data

Methods

We focused on extracting the date of diagnosis from posts submitted by diabetes patients on the internet community “Reddit”. We designed a method consisting of two steps: first, we used text generation models to create text explaining why the date of diagnosis is estimated as such; second, we evaluated the output in three aspects: factual, logical, mathematical and formatting correctness. If the output did not contain false information, it was considered factually correct. When the output was not logically coherent, it was scored as illogical. For cases where the output contained any sort of mathematical computation, its accuracy was scored as right or wrong. Finally, if the output was formatted as required by the instruction, it was deemed correct, otherwise wrong.

First, we gathered data from Reddit through pushshift API. All submissions in January 2021 to the community for diabetes patients, the subreddit “r/diabetes”, were collected. The user ID of the author, upload date, title and content of the post were then filtered, removing any posts that does not contain any of the four columns. From this, the first 200 submissions of January were collected and used.

We used LLaMA-30b supercot, a variation of the large language model (LLM) “Large Language Model Meta AI” (LLaMA) from Meta. This version of LLaMA was fine-tuned to follow instructions, using a specific prompt that passes an instruction and an input. The LLM extracts information related to the date of diagnosis for diabetes mentioned in the posts, and answers with the estimated date of diagnosis. It was

explicitly asked to include the reasoning about how it came up with that date. Prompting can be found on Figure 1.

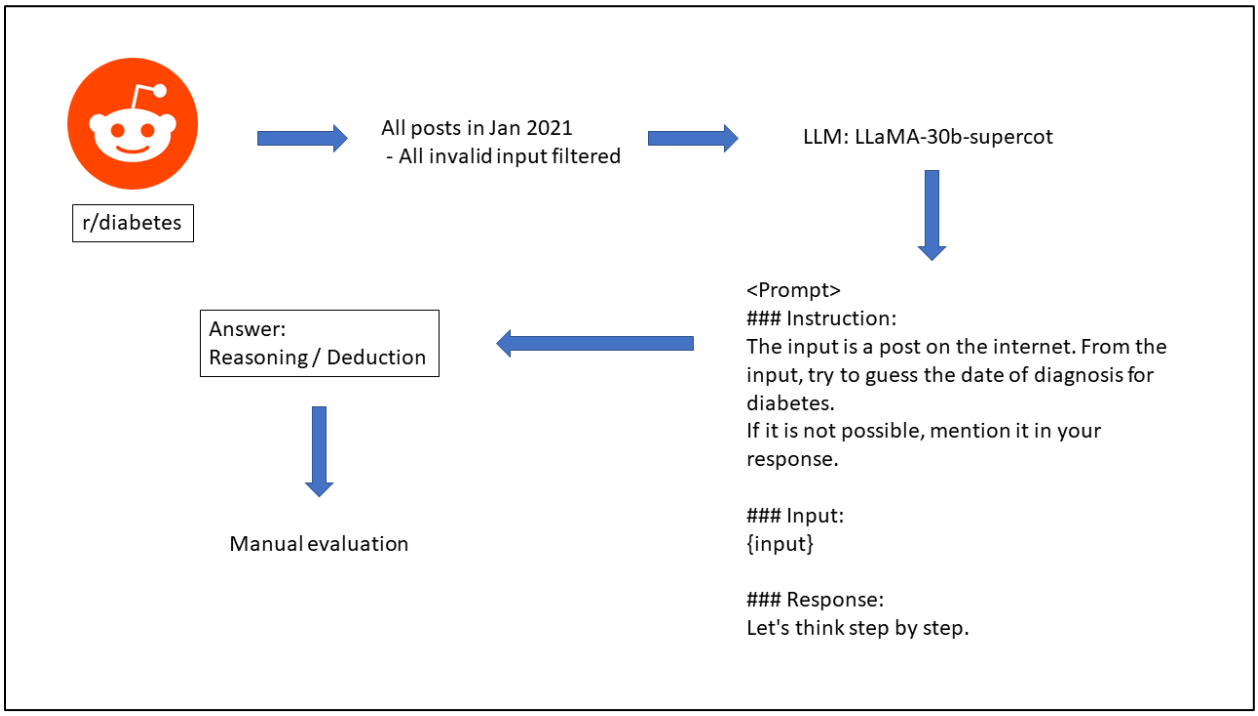

Figure 1 Study design and flow

Final results were assessed by manual review. First, the posts were divided into three groups: has explicit mentioning of date (group 1), has information about the date (group 2), has no information about the date at all (group 3). The output by the model was scored in 4 aspects: factual, logical, mathematical and formatting correctness.

Results

Among the 200 outputs generated, only 4 of them included factual inaccuracies. Furthermore, when focusing specifically on the 23 post submissions that provided context regarding the date of diagnosis, none of the outputs were found to be factually incorrect. However, in terms of logical deductions, out of the same 23 post submissions, 18 outputs were logically correct while 5 were deemed incorrect.

34 posts led the LLM to conduct mathematical calculations related to the date. Among these outputs, 27 were approximately correct, while 7 were identified as incorrect by a large margin. This finding aligns with the well-known fact that LLMs are bad at simple calculations.

Out of the 200 posts passed to the LLM, 3 explicitly included the date of diagnosis in its content, and 23 had implications. In total, 26 posts were eligible for producing date information to be formatted to OMOP-

CDM. Out of those 26 posts, 21 dates of diagnosis were accurately extracted.

The model had faults in its logic with 26 posts that did not include any information about the date of diagnosis for diabetes. Out of the 26 outputs, 14 outputs falsely generated wrong dates.

In total, 40 dates were generated. 21 were correct, 5 were inaccurate. 14 were falsely generated from illogical deductions. (Figure 2)

	CORRECT	WRONG	%
FACT	196	4	98.0
Logic	35	14	71.4
Math	27	7	79.4

Table 1 Scorings of each aspect

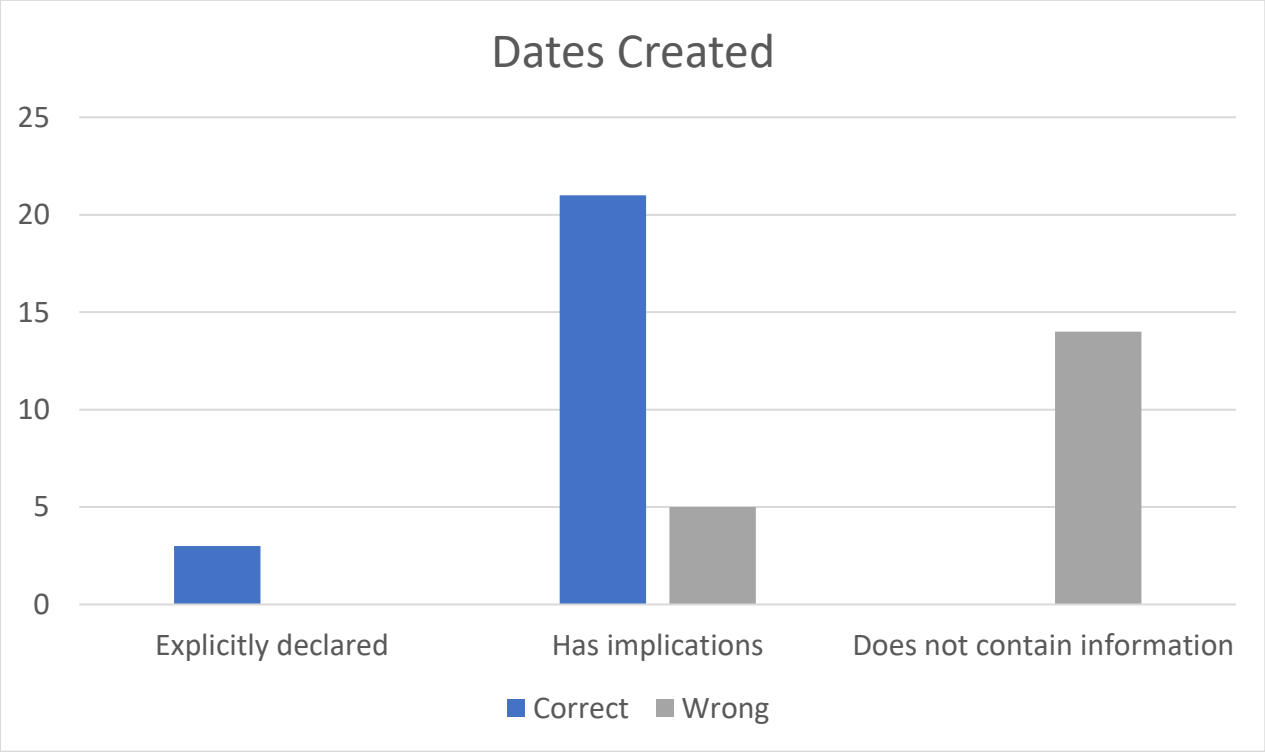

Figure 2. Correct and wrong proportions of all dates created

Conclusion

This paper suggests the potential of generative language models being utilized in mining medical data from the internet, and formatting them for convenient usage. At the moment, its accuracy is not optimal yet. Nonetheless, our work shows the feasibility of building CDM out of a data source that is not a part of the healthcare system. We believe similar approaches could be used on a variety of internet data sources and conventional EHR alike. With the development of additional modules to assist LLMs, the internet may become a new source of medical data to feed OMOP-CDM.

References

1. Somani S, van Buchem MM, Sarraju A, Hernandez-Boussard T, Rodriguez F. Artificial Intelligence–Enabled Analysis of Statin-Related Topics and Sentiments on Social Media. JAMA Network Open. 2023;6(4):e239747-e.