A comparison of large language model versus manual chart review for extraction of data elements from the electronic health record

Jin Ge, MD, MBA[1]; Michael Li, MD, MPH[1]; Molly B. Delk, MD[2]; Jennifer C. Lai, MD, MBA[1]

[1] Division of Gastroenterology and Hepatology, Department of Medicine, University of California – San Francisco, San Francisco, CA

[2] Section of Gastroenterology and Hepatology, Department of Medicine, Tulane University School of Medicine, New Orleans, LA

Financial/Grant Support:

The authors of this study were supported in part by the KL2TR001870 (National Center for Advancing Translational Sciences, Ge), P30DK026743 (UCSF Liver Center Grant, Ge, Li, and Lai), ACG Junior Faculty Development Award (American College of Gastroenterology Institute, Li), and R01AG059183/K24AG080021 (National Institute on Aging, Lai). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or any other funding agencies. The funding agencies played no role in the analysis of the data or the preparation of this manuscript.

Abbreviations:

CI, confidence interval; FrAILT, Functional Assessment in Liver Transplantation; GAI, generative artificial intelligence; HCC, hepatocellular carcinoma; LI-RADS, Liver Imaging Reporting and Data System; LLM, large language model; NLP, natural language processing; PHI, protected health information; UCSF, University of California, San Francisco

Correspondence:

Jin Ge, MD, MBA

513 Parnassus Avenue, S-357

San Francisco, CA 94143

E-mail: jin.ge@ucsf.edu

Fax: 415-476-0659

Disclosures:

The authors of this manuscript have the following potential conflicts of interest to disclose:

- Dr. Jin Ge receives research support from Merck and Co; and consults for Astellas Pharmaceuticals/Iota Biosciences.
- Dr. Jennifer C. Lai receives research support from Lipocene and Vir Biotechnologies; receives an education grant from Nestle Nutrition Sciences; serves on an advisory board for Novo Nordisk; and consults for Genfit, Third Rock Ventures, and Boehringer Ingelheim.

Writing Assistance:

None.

Author Contributions:

Authorship was determined using ICMJE recommendations.

Ge: Study concept and design; data extraction; analysis and interpretation of data; drafting of manuscript; critical revision of the manuscript for important intellectual content; statistical analysis; study supervision

Li: Analysis and interpretation of data; data extraction; critical revision of the manuscript for important intellectual content

Delk: Analysis and interpretation of data; data extraction; critical revision of the manuscript for important intellectual content

Lai: Study concept and design; analysis and interpretation of data; critical revision of the manuscript for important intellectual content; study supervision

Data Acknowledgement:

- The authors thank the UCSF AI Tiger Team, Academic Research Services, Research Information Technology, and the Chancellor's Task Force for Generative AI for their software development, analytical, and technical support related to the use of Versa API gateway (the UCSF secure implementation of large language models and generative AI via API gateway), Versa chat (the chat user interface), and related data assets.

Word Count:

1,190 (without references)

Structured Abstract:

Importance: Large language models (LLMs) have proven useful for extracting data from publicly available sources, but their uses in clinical settings and with clinical data are unknown.

Objective: To determine the accuracy of data extraction using "Versa Chat," a chat implementation of the general-purpose OpenAI gpt-35-turbo LLM model, versus manual chart review for hepatocellular carcinoma (HCC) imaging reports.

Design: We engineered a prompt for the data extraction task of six distinct data elements and input 192 abdominal imaging reports that were also manually tagged. We evaluated performance by calculating accuracy, precision, recall, and F1 scores.

Setting/Participants: Cross-sectional abdominal imaging reports of patients diagnosed with hepatocellular carcinoma enrolled in the Functional Assessment in Liver Transplantation (FrAILT) study.

Background:

Large language models (LLMs) hold tremendous potential for accelerating clinical research and augmenting clinical care.[1] One of the most promising LLM use cases is natural language processing (NLP) and extraction of structured elements from unstructured clinical text, such as imaging reports.[2] LI-RADS (Liver Imaging Reporting and Data System) was created by the American College of Radiology and provides standardized and reproducible reporting of hepatocellular carcinoma (HCC) imaging for clinical care and research.[3] Due to the LI-RADS reporting system, HCC imaging provides an ideal test case for LLM-enabled NLP extraction of structured data from unstructured clinical text. We sought to assess the performance of a commercially available general-purpose LLM, deployed in an isolated protected environment and permitted to be used with protected health information (PHI), versus human manual chart review in extracting six distinct data elements from abdominal imaging reports.

Methods:

We used "Versa Chat," the chat user interface of the general purpose Microsoft Azure OpenAI gpt-35-turbo LLM model ("Versa") that is implemented in a protected environment at the University of California, San Francisco (UCSF) to accommodate the use of PHI and intellectual property, for this study.[4] "Versa," like other gpt-35-turbo implementations, has a token limit of 4,096 tokens, defined as the unit that OpenAI generative artificial intelligence (GAI) models use to compute text length. One token approximates to about four characters or one word. This 4,096 token limit includes the count from both the user prompt and completion of the task for each session.[5] We manually reviewed 182 CT or MRI abdomen imaging reports without evidence of locoregional treatments from 169 patients diagnosed with HCC enrolled in the Functional Assessment in Liver Transplantation (FrAILT) study at UCSF.[6] The imaging reports, therefore, may or may not contain evidence of HCC as a diagnosis could have occurred subsequent to the date of imaging. We manually tagged the imaging reports for six distinct data elements: 1. Maximum LI-RADS score for any HCC lesions (defined as 4 or 5), 2. Number of HCC lesions, 3. Diameter (cm) of the largest lesion, 4. Sum of diameters (cm) of all HCC lesions, 5. Presence or absence of macrovascular invasion, or 6. Presence or absence of extrahepatic metastases.

All 182 imaging reports were trimmed to only include the findings and impressions sections. Due to the limitation of 4,096 tokens per session in "Versa Chat," we iteratively developed a "zero-shot" prompt (defined as a prompt that does not contain training data) with testing on the first five records (Figure 1).[7] As snowballing of data passed per prompt often led to execution failure from exceeding the token limit,

we ran 26 sessions of the final "zero-shot" extraction prompt in "Versa Chat" with approximately seven records per session for data extraction (see Figure 2 for an example exchange using mock data with "Versa Chat"). If "Versa Chat" produced an output that required minor additional formatting, we made those changes within the chat interface prior to collecting and aggregating the data. The total amount of time required to process all 182 records was 45 minutes.

We evaluated the accuracy of "Versa Chat" data extractions versus manual chart review with each imaging report as a separate record. We calculated performance metrics, notably accuracy, precision, recall, and F1 score (harmonic mean of precision and recall commonly used to evaluate classification in machine learning) for each of the six data elements whenever possible. For multilevel classifications (maximum LI-RADS score, number of HCC lesions, diameter of the largest lesion, and sum of tumor diameters), we calculated weighted-average precision, recall, and F1 score. For binary classifications (macrovascular invasion and extrahepatic metastases), we defined the presence of these features as a positive case for precision, recall, and F1 score.[8] We estimated 95% confidence intervals (CI) for performance metrics whenever possible through bootstrapping with 2,000 iterations. All statistical analyses were conducted in R, version 4.3.1 "Beagle Scouts" (R Core Team, Vienna, Austria),[9] and R packages *boot*, version 1.3-28.1,[10] and *caret*, version 6.0-94.[11] This study was approved by the UCSF Institutional Review Board in Study #11-07513.

Results:

The performance metrics for the six data elements extracted by the gpt-35-turbo "Versa Chat" model versus manual chart review are featured in Table 1. The overall accuracy of "Versa Chat" was 0.889 (95% CI 0.869-0.907) versus manual review. The accuracy rate varied between 0.725 (95% CI 0.643-0.780) for sum of tumor diameters to 0.989 (95% CI 0.956-0.995) for macrovascular invasion. In general, accuracy was higher for simple classification tasks (maximum LI-RADS score, macrovascular invasion, and extrahepatic metastases) compared to those that required comparison (maximum tumor diameter) or summation (number of tumors and sum of tumor diameters). As macrovascular invasion and extrahepatic metastases did not have any true positive cases, the precision for these two data elements were both zero. Similarly, as there were no false negative cases, the recall and F1 score for macrovascular invasion could not be calculated. As the precision, recall, and F1 score statistics for maximum LI-RADS score, number of tumors, maximum tumor diameter, and sum of tumor diameters were calculated as weighted-average values due to multilevel classifications, these values may be biased

as accurate predictions of absence of an imaging feature (e.g. “Versa Chat” noted zero tumors when there were no tumors by manual chart review) were included in the statistics.

Discussion:

This is one of the first studies that has demonstrated and compared the performance of the chat interface of a general-purpose LLM versus manual chart review for extraction of clinical data. We demonstrated high accuracy for simple extraction tasks, which degraded with more complex use cases. Of note, iterative development (“prompt engineering”) of a “zero-shot” prompt to specify the operations to be executed by the LLM was necessary to achieve this level of accuracy.[7] Our use of a “zero-shot” prompt and limiting the amount of data processed per session, however, prevented the gpt-35-turbo model from maintaining a persistent memory to allow in-context “learning” based on previous data.[12] These are known limitations of the gpt-35-turbo model, which have been improved upon in gpt-35-turbo-16k (which supports up to 16,384 tokens), gpt-4 (up to 8,192 tokens), and gpt-4-32k (up to 32,768 tokens).[5] Despite these limitations, our study demonstrated two important concepts: 1. Feasibility of using general purpose LLMs to extract structured information from clinical data with *minimal* technical expertise, and 2. Use of a LLM deployed in isolated protected environment that accommodates PHI (as opposed to ChatGPT, which is often not permitted for use with PHI) for clinical use cases.

Table 1 – Performance evaluation statistics of "Versa Chat" versus manual chart review

	Accuracy	Precision	Recall	F1 Score
Max LIRADS	0.945 (0.901-0.967)	0.955 (0.921-0.977)*	0.945 (0.907-0.973)*	0.946 (0.906-0.973)*
Number of Tumors	0.830 (0.764-0.879)	0.826 (0.760-0.878)*	0.830 (0.769-0.879)*	0.822 (0.756-0.872)*
Max Tumor Diameter	0.868 (0.808-0.907)	0.903 (0.853-0.940)*	0.868 (0.808-0.912)*	0.870 (0.816-0.918)*
Sum of Tumor Diameters	0.725 (0.643-0.780)	0.751 (0.681-0.825)*	0.725 (0.644-0.786)*	0.714 (0.642-0.781)*
Macrovascular Invasion	0.989 (0.956-0.995)	0.000 (0.000-0.000)[#]	N/A[#]	N/A[#]
Extrahepatic Metastases	0.978 (0.941-0.989)	0.000 (0.000-0.000)[$]	0.000 (0.000-0.000)[$]	N/A[$]
Overall Accuracy	0.889 (0.869-0.907)			

*Due to multilevel classification, these statistics were calculated as weighted-average values.
[#] There were no true positives or false negatives for the evaluation of this data element.
[$] There were no true positives for the evaluation of this data element.

Figure 1 – Final prompt used for data extraction from "Versa Chat" (gpt-35-turbo)

I am a clinical researcher reviewing CT and MRI abdomen imaging reports for evidence of hepatocellular carcinoma (HCC). The imaging reports include only those patient who have not undergone treatment. I will provide you with a dataset with the patient identifier (study_id), imaging note identifier (note_id) and the full imaging report (findings), I would like your help in extracting specific data from these imaging reports for the following fields:

1. The maximum LIRADS score for any lesions. Entries should be from 1 through 5.
2. The number of HCC lesions. This is defined as total number of LIRADS 4 and LIRADS 5 lesions only. Do not include lesions that are LIRADS score 3 or less.
3. The size of the largest lesion in centimeters. This is defined as the size of the largest LIRADS 4 or LIRADS 5 lesions only. Do not include lesions that are LIRADS score 3 or less.
4. Sum of sizes of all lesions in centimeters. This is defined as the sum all LIRADS 4 and LIRADS 5 lesions only. Do not include lesions that are LIRADS score 3 or less.
5. Presence or absence of macrovascular invasion
6. Presence or absence extrahepatic metastases

If there are multiple records for each study_id, the note_id records are sequential for that patient. If there are previous imaging reports for the same patient (study_id), please evaluate the previous report for any progression of disease (defined as the lesions becoming bigger and/or if there is upgrading of the LIRADS rating), and would extract the data based on progression from previous imaging.
Please give the output in csv file with the following column headers: study_id, note_id, max_lirads_score, number_of_hcc_lesions, size_of_largest_lesion, sum_of_sizes, macrovascular_invasion, extrahepatic_metastases

Figure 2 – Example of an exchange with "Versa Chat" (gpt-35-turbo) using mock data

Greetings [user]! I'm Versa, how may I assist you today?

I am a clinical researcher reviewing CT and MRI abdomen imaging reports for evidence of hepatocellular carcinoma (HCC). The imaging reports include only those patient who have not undergone treatment. I will provide you with a dataset with the patient identifier (study_id), imaging note identifier (note_id) and the full imaging report (findings), I would like your help in extracting specific data from these imaging reports for the following fields: 1. The maximum LIRADS score for any lesions. Entries should be from 1 through 5. 2. The number of HCC lesions. This is defined as total number of LIRADS 4 and LIRADS 5 lesions only. Do not include lesions that are LIRADS score 3 or less. 3. The size of the largest lesion in centimeters. This is defined as the size of the largest LIRADS 4 or LIRADS 5 lesions only. Do not include lesions that are LIRADS score 3 or less. 4. Sum of sizes of all lesions in centimeters. This is defined as the sum all LIRADS 4 and LIRADS 5 lesions only. Do not include lesions that are LIRADS score 3 or less. 5. Presence or absence of macrovascular invasion 6. Presence or absence extrahepatic metastases If there are multiple records for each study_id, the note_id records are sequential for that patient. If there are previous imaging reports for the same patient (study_id), please evaluate the previous report for any progression of disease (defined as the lesions becoming bigger and/or if there is upgrading of the LIRADS rating), and would extract the data based on progression from previous imaging. Please give the output in csv file with the following column headers: study_id, note_id, max_lirads_score, number_of_hcc_lesions, size_of_largest_lesion, sum_of_sizes, macrovascular_invasion, extrahepatic_metastases

Sure, I can help you with extracting specific data from the imaging reports. Please provide me with the dataset containing the patient identifier (study_id), imaging note identifier (note_id), and the full imaging report (findings).

study_id note_id findings Patient1 Note1 Findings: A 3.2 cm arterial enhancing lesion in segment VII of the liver. This lesion also demonstrates washout on portal venous phase and pseudocapsule. LI-RADS 5. No evidence of tumoral thrombus in the hepatic vasculature. Background liver parenchyma reveals features consistent with cirrhosis. Impression: 3.2 cm LI-RADS 5 lesion in segment VII of the liver consistent with HCC.

To extract the specific data from the imaging report, I will analyze the provided findings and extract the required information. Here is the extracted data:
study_id, note_id, max_lirads_score, number_of_hcc_lesions, size_of_largest_lesion, sum_of_sizes, macrovascular_invasion, extrahepatic_metastases
Patient1, Note1, 5, 1, 3.2 cm, 3.2 cm, Absent, Absent
Please note that the information extracted is based on the provided imaging report. If there are any previous reports for the same patient, please provide them so that I can evaluate the progression of the disease.

Please remove the cm and change "Absent" to 0. Please give the output as a csv file.

Certainly! I will remove the "cm" units and change "Absent" to "0". Here is the updated output in CSV format:

```
study_id,note_id,max_lirads_score,number_of_hcc_lesions,size_of_largest_lesion,sum_of_sizes,macrovascular_invasion,extrahepatic_metastases
Patient1,Note1,5,1,3.2,3.2,0,0
```

Please let me know if there are any additional imaging reports for the same patient or if you have any further requirements.

References:

1. Ge J, Lai JC. Artificial intelligence-based text generators in hepatology: ChatGPT is just the beginning. *Hepatol Commun*. 2023;7(4). doi:10.1097/HC9.0000000000000097
2. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2
3. Chernyak V, Fowler KJ, Kamaya A, et al. Liver Imaging Reporting and Data System (LI-RADS) Version 2018: Imaging of Hepatocellular Carcinoma in At-Risk Patients. *Radiology*. 2018;289(3):816-830. doi:10.1148/radiol.2018181494
4. Azure OpenAI Service – Large Language Models for Generative AI. https://azure.microsoft.com/en-us/products/ai-services/openai-service-b. Accessed August 25, 2023.
5. Azure OpenAI Service models - Azure OpenAI | Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models. Accessed August 26, 2023.
6. Lai JC, Covinsky KE, Dodge JL, et al. Development of a novel frailty index to predict mortality in patients with end-stage liver disease. *Hepatology*. 2017;66(2):564-574. doi:10.1002/hep.29219
7. Wang J, Shi E, Yu S, et al. Prompt Engineering for Healthcare: Methodologies and Applications. *arXiv*. 2023. doi:10.48550/arxiv.2304.14670
8. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*. 2009;45(4):427-437. doi:10.1016/j.ipm.2009.03.002
9. Team RC. R: A language and environment for statistical computing. 2013.
10. Bootstrap Functions (Originally by Angelo Canty for S) [R package boot version 1.3-28.1]. https://cran.r-project.org/web/packages/boot/index.html. Published November 22, 2022. Accessed January 6, 2023.
11. Kuhn M. Classification and Regression Training [R package caret version 6.0-94]. March 2023.
12. Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. [2101.06804] What Makes Good In-Context Examples for GPT-3? *arXiv*. January 2021.