

Use of large language models to optimize poison center charting

Nikolaus Matsler, Lesley Pepin, Shireen Banerji, Christopher Hoyte & Kennon Heard

To cite this article: Nikolaus Matsler, Lesley Pepin, Shireen Banerji, Christopher Hoyte & Kennon Heard (12 Jun 2024): Use of large language models to optimize poison center charting, Clinical Toxicology, DOI: [10.1080/15563650.2024.2348107](https://doi.org/10.1080/15563650.2024.2348107)

To link to this article: <https://doi.org/10.1080/15563650.2024.2348107>




View supplementary material 



Published online: 12 Jun 2024.



Submit your article to this journal 



Article views: 50



View related articles 



View Crossmark data 

Use of large language models to optimize poison center charting

Nikolaus Matsler^{a,d,e,*}, Lesley Pepin^{a,c†}, Shireen Banerji^a, Christopher Hoyte^{a,b} and Kennon Heard^{a,b}

^aRocky Mountain Poison and Drug Safety, Denver Health and Hospital Authority, Denver, CO, USA; ^bDepartment of Emergency Medicine, University of Colorado, Aurora, CO, USA; ^cDepartment of Emergency Medicine, Hennepin Healthcare, Minneapolis, MN; ^dOregon Poison Center, Oregon Health and Science University, Portland, OR; ^eSalem Health Emergency Physician Services at Salem Health, Salem, OR

ABSTRACT

Introduction: Efficient and complete medical charting is essential for patient care and research purposes. In this study, we sought to determine if Chat Generative Pre-Trained Transformer could generate cogent, suitable charts from recorded, real-world poison center calls and abstract and tabulate data.

Methods: De-identified transcripts of real-world hospital-initiated poison center consults were summarized by Chat Generative Pre-Trained Transformer 4.0. Additionally, Chat Generative Pre-Trained Transformer organized tables for data points, including vital signs, test results, therapies, and recommendations. Seven trained reviewers, including certified specialists in poison information and board-certified medical toxicologists, graded summaries using a 1 to 5 scale to determine appropriateness for entry into the medical record. Intra-rater reliability was calculated. Tabulated data was quantitatively evaluated for accuracy. Finally, reviewers selected preferred documentation: original or Chat Generative Pre-Trained Transformer organized.

Results: Eighty percent of summaries had a median score high enough to be deemed appropriate for entry into the medical record. In three duplicate cases, reviewers did change scores, leading to moderate intra-rater reliability ($\kappa = 0.6$). Among all cases, 91 percent of data points were correctly abstracted into table format.

Discussion: By utilizing a large language model with a unified prompt, charts can be generated directly from conversations in seconds without the need for additional training. Charts generated by Chat Generative Pre-Trained Transformer were preferred over extant charts, even when they were deemed unacceptable for entry into the medical record prior to the correction of errors. However, there were several limitations to our study, including poor intra-rater-reliability and a limited number of cases examined.

Conclusions: In this study, we demonstrate that large language models can generate coherent summaries of real-world poison center calls that are often acceptable for entry to the medical record as is. When errors were present, these were often fixed with the addition or deletion of a word or phrase, presenting an enormous opportunity for efficiency gains. Our future work will focus on implementing this process in a prospective fashion.

ARTICLE HISTORY

Received 2 January 2024

Revised 11 April 2024

Accepted 22 April 2024

KEYWORDS

ChatGPT; Artificial intelligence (AI); poison center charting; Artificial intelligence summary

Introduction

Medical charting is often derided as a necessary evil, a cumbersome, time-consuming process that medical professionals seek to make more efficient through various pre-written texts or pre-filled click boxes. While time-saving, these strategies may lead to documentation errors, less nuance, or limitations in scope [1]. Ultimately, this demonstrates the trade-off of upfront work for backend convenience; detailed, nuanced documentation may take more time to create but may also contain information that later becomes relevant for patient care and/or research. This is never more prescient than in the world of medical toxicology in poison centers.

United States poison centers receive roughly 3 million calls annually; all require various levels of documentation, which includes coding the case for collection into the National


Poison Data System [2,3]. Standards do exist for coding cases to capture national level data. However, documentation requirements for prose style entry into the electronic medical record are at the discretion of individual poison centers [4]. A preliminary survey conducted to inform this study demonstrated that specialists in poison information at a large regional poison center spend 40% of their shift on documentation. Similar time commitments for documentation have been described for outpatient physicians who provide direct bedside care [5]. Aside from the time burden, incomplete data collection and prose format often limit the ability to abstract data in an efficient manner. Therefore, a new technique of documentation may provide profound benefits in both efficiency and capture of subtle and minable data points.

In 2017, a landmark paper was published that put forth the “transformer”, the functional unit of what has become

CONTACT Nikolaus Matsler  nik.matsler@denverrem.org  Rocky Mountain Poison and Drug Safety, Denver Health and Hospital Authority, Denver, CO, USA.

*Present address: Oregon Poison Center, Oregon Health and Science University, Portland, OR, USA; Salem Health Emergency Physician Services at Salem Health, Salem, OR, USA.

†Present address: Department of Emergency Medicine, Hennepin Healthcare, Minneapolis, MN, USA.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/15563650.2024.2348107>.

© 2024 Informa UK Limited, trading as Taylor & Francis Group

contemporary large language models [6]. Part of the breakthrough of this technology was the ability for large language models to learn the relative importance of words in a sentence, paragraph, or text as they relate to one another in a time-efficient manner. Doing so allows large language models to excel at tasks such as summarization and abstraction of data into minable formats.

In this study, we sought to determine if large language models, such as Chat Generative Pre-Trained Transformer (ChatGPT), could generate documentation from poison center calls that then were judged as acceptable for entry into the medical record in terms of accuracy, readability, and completeness, while improving data mine-ability.

Methods

This study was performed at a large regional poison center. Ten real-world calls originating from hospitals were selected and transcribed over six months (01/01/2023 to 06/30/2023). They were hand-selected to maximize variation in length, number of speakers, and toxin. Only cases from hospitals were considered to ensure complexity and increase the chance of abstractable data. Transcriptions of case calls were performed by two of the authors (NM and LCP). Patient identifiers were removed in a systematic fashion. Otherwise, calls were documented verbatim, including vocal disfluencies natural to speech.

Transcripts were entered into the large language model ChatGPT version 4.0. ChatGPT summarized each transcript in a unique instance and abstracted specific clinical details into tables based on the following prompt:

The following is a call to a poison center for a toxicology consult. Please summarize the call and include all the clinically relevant information. In addition to this, please also create separate tables for each of the following and abstract any information from the call into these tables: Laboratory values, vital signs, electrocardiogram (ECG) data, interventions, and recommendations. Here is the transcript of the call:

A panel of seven reviewers were asked to review the combination of phone transcripts and summaries. Reviewers graded summaries after being trained on a pre-formulated rubric. The rubric was a one to five scale in which scores less than three meant summaries were unacceptable for entry into the medical record (Table 1). The reviewers included three board certified toxicologists with experience in poison center management and four certified specialists in poison

information. All reviewers completed a 30-minute training with example phone transcripts and summaries prior to study participation. Reviewers were compensated for their time with a \$50 Amazon gift card.

Reviewers independently evaluated transcripts and ChatGPT summaries using the survey platform Qualtrics. Week one, seven case transcripts and summaries were reviewed and scored. Week two, three new transcript and summary combinations and three repeat combinations were reviewed and scored. Repeat combinations were evaluated for intra-rater reliability. Week three, reviewers were asked to select the preferred documentation for the ten cases, ChatGPT summary or original poison center charting. Preferred documentation was specifically held until week three so that reviewers would not be biased with initial scoring based on the actual charts generated by specialists in poison information.

During weeks one and two, if reviewers scored a summary less than three, they were prompted to provide their reasoning. Errors were collected in free text format. Following completion of data collection, error categories were then defined, and errors were sorted as follows: alias/interpretation – data ascribed to the wrong person or interpretation of data is incorrect; firmness – inaccurate degree of certainty about information; confabulation – detail entirely made up; missing detail – important information excluded from summary. Further, all tabulated data were independently reviewed for accuracy and completeness of data points.

Categorical data was described using frequencies and proportions. Ordinal data was described using medians, and continuous data using averages. Intra-rater reliability was described using the kappa statistic based on rubric scores. All analyses were performed using Microsoft Excel.

The study was approved and deemed exempt by the Institutional Review Board of Colorado. The authors confirm that the data supporting the findings of this study are available within the article and/or its [supplementary materials](#).

Results

Ten calls were selected to be transcribed with word counts ranging from 553 to 2,551 words per transcript (Table 2). Half of the calls involved a medical provider calling the poison center and a specialist in poison information providing guidance, while the other half involved the addition of a medical toxicologist in the conversation. Most calls (nine/ten)

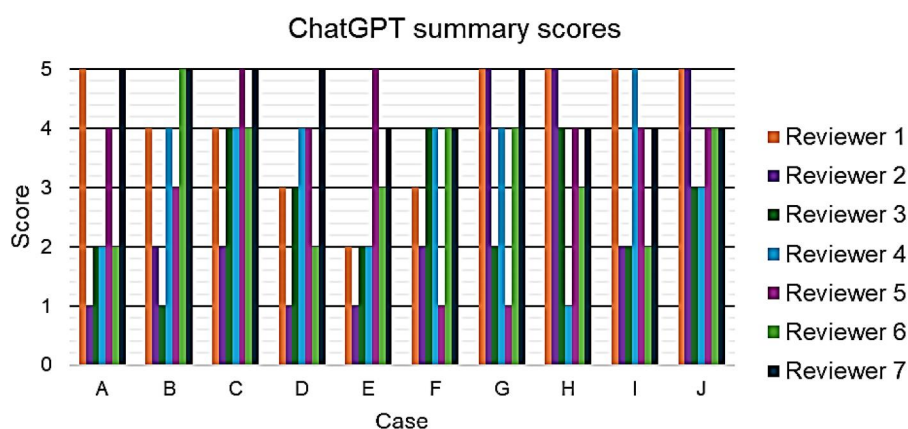
Table 1. Rubric for grading ChatGPT summaries.

| Score | Description |
|-------|--|
| 1 | Unacceptable with gross inaccuracies or confabulations: Summary contains confabulated information AND/OR inaccuracies/missing pertinent details distort interpretation of the case AND/OR summary is unreadable. This would be unacceptable to enter into the medical record. |
| 2 | Unacceptable without gross inaccuracies or confabulations: Summary does NOT contain confabulated information. Inaccuracies/ missing pertinent details are present, but do NOT distort interpretation of the case. This would be unacceptable to enter into the medical record. |
| 3 | Equivalent to transcript: Summary is accurate and complete. It is readable however there is a robotic quality AND/OR it contains considerable extraneous information. This would be acceptable for entering into the medical record. |
| 4 | Superior with some extraneous info: Summary is accurate, complete, and readable. Some extraneous details are present, but do NOT detract from the summary. This would be acceptable for entering into the medical record. |
| 5 | Superior without extraneous info: Summary is accurate, complete and readable, but concise. It includes little to no extraneous details. This would be acceptable for entering into the medical record. |

Scores of 3 or higher would be acceptable charting for the medical record.

Table 2. Characteristics of selected cases for transcription.

| Case content | Number of patients | Speakers involved in call | Words |
|---------------------------------|--------------------|--|-------|
| Salicylate overdose | 1 | Physician caller, specialist in poisons information | 2,551 |
| Psilocybin ingestion | 1 | Physician caller, specialist in poisons information | 1,311 |
| Scorpion sting | 1 | Physician caller, specialist in poisons information, medical toxicologist | 1,543 |
| Viscous lidocaine ingestion | 1 | Physician caller, specialist in poisons information, medical toxicologist | 980 |
| Cocaine and opioid insufflation | 3 | Two physician callers, specialist in poisons information, medical toxicologist | 3,442 |
| Unexplained metabolic acidosis | 1 | Physician caller, specialist in poisons information | 553 |
| Flecainide overdose | 1 | Physician caller, specialist in poisons information | 1,089 |
| Fentanyl body stuffer | 1 | Physician caller, specialist in poisons information, medical toxicologist | 1,242 |
| Pregabalin overdose | 1 | Physician caller, specialist in poisons information | 1,056 |
| Lamotrigine overdose | 1 | Physician caller, specialist in poisons information, medical toxicologist | 1,598 |

**Figure 1.** Distribution of scores across raters for first review of individual cases.**Table 3.** Scores and reasoning for low individual scores for cases after initial review.

| Case | Median score | Number of individual scores ≥ 3 (n, %) | Most preferred documentation (n, %) | Reasons for score < 3 | Tabulated data abstracted |
|----------|--------------|---|-------------------------------------|---|---------------------------|
| A | 2 | 3 (43 %) | Original (4, 57 %) | Alias/interpretation – 1; firmness – 1; confabulation – 1; missing detail – 4 | 24/27 (89 %) |
| B | 4 | 5 (71 %) | ChatGPT (5, 71 %) | Firmness – 2 | 13/14 (93 %) |
| C | 4 | 6 (86 %) | Original (4, 57 %) | Firmness – 1; missing detail – 1 | 15/15 (100 %) |
| D | 3 | 5 (71 %) | ChatGPT (7, 100 %) | Confabulation – 1; alias/interpretation – 1 | 19/19 (100 %) |
| E | 2 | 3 (43 %) | ChatGPT (4, 57 %) | Missing detail – 3; alias/interpretation – 2, firmness – 1 | 13/14 (93 %) |
| F | 4 | 5 (71 %) | ChatGPT (5, 71 %) | Alias/interpretation – 2 | 15/16 (94 %) |
| G | 4 | 5 (71 %) | ChatGPT (6, 86 %) | Alias/interpretation – 2 | 11/12 (92 %) |
| H | 4 | 6 (86 %) | ChatGPT (4, 57 %) | Not available | 24/27 (89 %) |
| I | 4 | 4 (58 %) | ChatGPT (4, 57 %) | Alias/interpretation – 1; confabulation – 1; missing detail- 1 | 20/27 (74 %) |
| J | 4 | 7 (100 %) | ChatGPT (6, 81 %) | Not available | 12/12 (100 %) |
| Repeat A | 3 | 4 (58 %) | Not applicable | Missing detail – 3 | Not applicable |
| Repeat E | 3 | 5 (71 %) | Not applicable | Alias/interpretation – 1; missing detail- 1 | Not applicable |
| Repeat F | 3 | 5 (71 %) | Not applicable | Alias/interpretation – 2 | Not applicable |

Quantitative results of tabulated data. Reviewers preferred documentation for each case. Original refers to the poison center chart, and Chat Generative Pre-Trained Transformer (ChatGPT) refers to the ChatGPT generated summary and table.

pertained to a single patient. Each exposure discussed was unique, with most involving ingestions (eight/ten).

The summaries and abstracted tables were independently graded by reviewers across a two-week period using the rubric. Reviewers graded 13 total case summaries, which included three duplicate case presentations to evaluate intra-rater reliability. There was variability across reviewers in terms of scoring. The proportion of cases receiving a passing score on first review by an individual reviewer ranged from 30 % to 100 % (Figure 1). On first review, most of the summaries were deemed appropriate for the medical record, with 70 % receiving a passing score. Eighty percent of cases

on first review had a median score of three or higher. Across all 13 summary reviews, 85 % had a median score of three or higher (Table 3). Medical toxicologists scored cases lower than the certified specialists in poison information. Across all 13 summary reviews, medical toxicologist reviewers gave passing scores in 63 %, while the certified specialists in poison information reviewers gave passing scores in 75 %. On re-review of the three duplicate cases, reviewers changed scores, leading to moderate intra-rater reliability (kappa = 0.6).

Individual reviewers were prompted to provide reasoning for any score less than three. Only one case received passing

scores from all reviewers. Reasoning for lower scores was prompted in 28 (31%) of the 91 summary reviews. Reasoning was further categorized into error types as described in the methods. Missing detail or alias/interpretation errors were most cited (Table 3). Reasoning for scores less than three was often the same across reviewers. More than one reviewer cited the same error type in 63.6% of lower scored cases. In case H, reviewer 4 scored the summary less than three due to concern that ChatGPT had inserted a confabulated detail (denoted by asterisk in Table 3). However, this was a reviewer error as the detail in question was present in the original transcript. On average, 1.61 unique errors were reported per case.

When evaluating the abstracted tabulated information, a quantitative review of all data points was performed. Number of unique data points ranged from 12 to 27 per case. Among all cases, 91% of data points were correctly abstracted overall, and there was no confabulated data in the tables (Table 3). Most of the inaccuracies were due to misplacement of “recommendations” in the “interventions” table or vice versa. When analyzing numerical data alone (e.g., laboratory values, ECG, vital signs), 95 out of 103 (92.2%) were correctly abstracted. Eight out of ten of the cases had no errors in the numerical data abstraction. One case contained the preponderance of numerical errors, with seven in total, due to only abstracting final repeat laboratory values instead of the full laboratory trends. The remaining error was a failure to abstract “pregnancy was negative” into the laboratory data. The ChatGPT abstraction of numeric data points was also compared to the data captured in the original poison center documentation created by specialists in poison information. Seventy-seven percent (81/103) of numeric data was abstracted in prose format in the original documentation. For an example of transcript and abstracted data, please see Appendix A.

The majority of reviewers preferred the ChatGPT summary over the original poison center documentation in 80% of cases (Table 3). The medical toxicologist reviewers preferred the ChatGPT summary in 77% (23/30) and the certified specialists in poison information reviewers in 60% (24/40). The ChatGPT documentation was preferred on 12 (57%) individual reviews in which the reviewer initially graded the ChatGPT summary less than three. Original documentation was preferred in 11 (22%) cases where the initial ChatGPT summary scored greater than three.

Discussion

Medical charting, especially at poison centers, has never been more important. Charts need to be accurate, efficient, and readable to allow multiple providers to take care of the same patient across different states and over variable time periods. Contemporaneously capturing thorough, nuanced data can aid in future research and patient care. Large language models have dramatically improved over recent years and may be leveraged to achieve these goals.

In this study, we demonstrated that large language models can summarize and abstract key information into table format from conversational poison center encounters. Seventy

percent of the large language model summaries were deemed appropriate for direct entry into the medical record. Further, when compared to the real-world documentation generated for these calls, the ChatGPT summaries were preferred in most cases. Interestingly, even in lower scoring cases, 57% still preferred the ChatGPT summary over original documentation. Similarly, capture of numerical data increased by 14% in the ChatGPT tables compared to original documentation.

There is a distinct advantage to data collection in tables over the prose style format often used in poison center charts. Free text collection of data creates natural differences across charts. Future research can be limited by the need for individual chart review to abstract this information. Additionally, there can be missing data points that were deemed irrelevant or mistakenly missed on initial documentation. Tabulated data is minable; it can be exported into any desired format and collated easily. The National Poison Data System requires cases to be coded using discrete variables for this exact reason: to capture specific case details in a minable fashion. However, data must be accurate for this advantage to be realized.

The prompt used in our study specifically instructed the large language model to generate several different tables: laboratory values, ECG data, vital signs, poison center recommendations, and interventions done so far. Looking at all tables combined, the accuracy of abstraction was good at 91%. Errors primarily arose when the large language model placed a poison center “recommendation” in the “interventions” table or vice versa. This likely reflects a need for further training to classify these requests. However, when examining the numerical data alone (e.g., laboratory values, ECG, vital signs), the large language model successfully captured more data points than the original poison center documentation (81 versus 95 data points). In eight cases, the large language model correctly abstracted all numerical data. One case contained seven of the missing data points, which was due to the large language model only abstracting the most recent set of labs, as opposed to the full trends listed in the transcript (which the poison center generated chart correctly captured). This overall improvement presents opportunities for increased efficiency both at the time of the call and in future projects.

Our study did not directly evaluate efficiency gains in real world use cases. However, given that ChatGPT generates responses in seconds, there is likely a large opportunity for improvement in documentation. Currently, specialists in poison information answer calls, take histories and provide appropriate triage and recommendations over a recorded line. Documentation often occurs in parallel, requiring a specialist in poison information to take brief pauses to complete data entry or ask for repeat values. In fact, more complex calls may require the specialist in poison information to listen back through the recording to confirm certain details. By altering the workflow from a specialist in poison information completing all documentation by hand, to simply reviewing and editing large language model generated charts/tables, this potentially can lead to a significantly reduced time burden of charting. Further, this may improve the conversational flow of a call, as it could potentially eliminate the need for

pausing for data entry or clarification of lengthy lists of data. This also has an advantage over dictation software, such as Dragon[®], as specialists in poison information would not have to spend time synthesizing the call themselves, which would still suffer from dictation errors that would require proofreading. In our pre-study survey, 14/17 specialists in poison information also noted difficulty interpreting other specialists' notes at least once per shift. Given the readability of the generated summaries, this could further improve specialist to specialist communication and efficiency. These potential efficiency gains, however, directly depend on how much time is needed to proofread and correct errors.

The preponderance of errors in the ChatGPT summaries identified by reviewers could be fixed with the addition/deletion of a single word or phrase. As examples, an error described by a reviewer noted that in several places a caller described the ingestion happening "exactly" 30 min prior, but the generated summary documented "approximately" 30 min prior. Another noted a patient had "intentionally" overdosed, but was listed as an "accidental" ingestion in the summary. More interpretation errors were noted, such as a summary listing a patient as "hypertensive" when the accompanying numerical blood pressure was within normal range later in the call. Please see [Appendix A](#) for an example of a de-identified transcript and summary/tables that were utilized. The current study did not evaluate the time required to identify and edit these errors. However, it would be reasonable to expect this process to take less time than the 5 min most specialists in poison information reported spending to chart after a call.

Large language models have had a meteoric rise in popularity, with myriad publications evaluating various use cases. Many of these studies have utilized vanilla ChatGPT (as in our case), which has not been specially trained in medicine or toxicology [7]. Recent publications have pointed out the flaws in attempting to use large language models to directly answer complex medical questions or replace critical thinking [8–10]. These research studies offer an important juxtaposition as they start to define ideal use cases for generic models. In these generic models, prediction of the next best token (e.g., part of or entire words) is done based on internet data or learned by positive/negative reinforcement from humans, eventually training an optimized policy. As it implies, this policy has not had specialized training in medicine and is therefore expected to fail when applied in critical thinking or complex decision making. But that does not mean that it has not trained on the medical or toxicological literature.

Training, in this sense, should not be thought of as purely the material confronted. In other words, while we do not know the exact training data set presented to ChatGPT, we can infer it has "read" significant literature in the field of toxicology. This is apparent when asking it to describe certain toxicological entities or treatments. Because the model does not use internet search (though newer versions allow this), it means that the data are abstracted away in the "hidden" layers of the network. So, in a sense, it has been trained in the world of toxicology, but how to best make timely medical decisions based on disparate patient presentations with incomplete information it has not. This idea extends to charting as well; it

has simply not had the opportunity to learn the exact style that suits each poison center best. However, it has learned a great deal about summarization as a task through its training, making it particularly useful for this case. This is why the vanilla (standard) model was chosen for this study, as the success or failure of this application should be testable without extra training. However, utilizing further training specifically crafted to teach medical charting may improve performance, which can be achievable locally.

Locality of data has become increasingly important in medicine as each online transfer of data represents a potential point of failure of protected health information. The architecture of ChatGPT 4 boasts 1.76 trillion parameters, which cannot be feasibly run on most local hardware. However, newer models, capable of running locally on a desktop computer, have since emerged. These models are smaller and do not require data transfer via the internet. Securing data locally is universally done by poison centers currently and could similarly be done for these smaller large language models. These models are still capable of being trained allowing individual poison center preferences for documentation to be achieved. In order to maintain patient confidentiality these records had to be manually de-identified before submitting them to ChatGPT. As such, we caution against attempts to replicate our findings using actual recordings of sensitive patient data in an open environment.

This study has several limitations to consider. In addition to the errors discussed above, we also found a high variance in intra-rater-reliability, that threatened the internal validity of the study. When looking at specific cases and across raters, this mostly came down to some raters considering a phrase or word as an error, while some did not. This was despite a training session given at the beginning of the study, where sample cases were discussed. Even when raters were given summaries they had seen before, there was still variation if they considered something an error the second time or not, with many of them changing scores. This could represent an opportunity for an improved training session but also suggests that individuals may place different emphasis on what constitutes an acceptable chart. Towards this end, future studies should consider including additional training and more systematic tools that define acceptability. Further, even with cases that raters identified as unacceptable for entry into the medical record as is, they still often preferred those documents to the actual charts that were created based on the call, which may speak to variation in external validity, as different poison centers may have different explicit or implicit standards. This would also be a concern if attempting to generalize this work to non-poison center documentation, as changing the format of input can significantly affect the output, as demonstrated in the lone case that averaged a score less than three in our study.

Conclusions

In this study, we sought to determine if a large language model could generate charts that were acceptable for entry into the medical record and successfully abstract data into a

more minable format. Our results suggest that the summaries are accurate, complete, and acceptable for entry to the medical record “as is” the majority of the time while also being subjectively preferred over extant charts generated by the same call. Likewise, data abstraction was successful 91 % of the time, and eight of ten cases contained no errors in numerical abstraction. Even when the model made errors, these were easily fixable with either simple insertion/deletion operations or by tweaking the generic prompt that was initially crafted. Future steps will be to in-line this model into real-world poison center medical charting, to determine if this will lead to improvements in efficiency, patient care, and future research.

Disclosure statement

The authors have nothing to disclose.

Funding

The authors have no financial support to declare.

References

- [1] Weis JM, Levy PC. Copy, paste, and cloned notes in electronic health records. *Chest*. 2014;145(3):632–638. doi: [10.1378/chest.13-0886](https://doi.org/10.1378/chest.13-0886).
- [2] Gummin DD, Mowry JB, Beuhler MC, et al. 2021 Annual report of the national poison data system© (NPDS) from america’s poison centers: 39th annual report. *Clin Toxicol (Phila)*. 2022;60(12):1381–1643. doi: [10.1080/15563650.2022.2132768](https://doi.org/10.1080/15563650.2022.2132768).
- [3] Gummin DD, Mowry JB, Beuhler MC, et al. 2020 Annual report of the American association of poison control centers’ national poison data system (NPDS): 38th annual report. *Clin Toxicol (Phila)*. 2021;59(12):1282–1501. doi: [10.1080/15563650.2021.1989785](https://doi.org/10.1080/15563650.2021.1989785).
- [4] National Poison Data System (NPDS) Coding Users’ Manual. America’s poison centers. 2022.
- [5] Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med*. 2016;165(11):753–760. doi: [10.7326/M16-0961](https://doi.org/10.7326/M16-0961).
- [6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30. Available from: <https://proceedings.neurips.cc/paper/7181-attention-is-all>
- [7] Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in clinical toxicology. *JMIR Med Educ*. 2023;9:e46876. doi: [10.2196/46876](https://doi.org/10.2196/46876).
- [8] Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11(6):887. doi: [10.3390/healthcare11060887](https://doi.org/10.3390/healthcare11060887).
- [9] Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*. 2023;15(2):e35179. doi: [10.7759/cureus.35179](https://doi.org/10.7759/cureus.35179).
- [10] Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312. doi: [10.2196/45312](https://doi.org/10.2196/45312).