





## BRIEF REPORT

## Cardiology

# Automated HEART score determination via ChatGPT: Honing a framework for iterative prompt development

Conrad W. Safranek BS<sup>1</sup>   | Thomas Huang BS<sup>1</sup> | Donald S. Wright MD, MHS<sup>2</sup> | Catherine X. Wright MD<sup>3</sup> | Vimig Socrates MS<sup>1</sup> | Rohit B. Sangal MD, MBA<sup>2</sup> | Mark Iscoe MD<sup>1,2</sup> | David Chartash PhD<sup>1,4</sup> | R. Andrew Taylor MD, MHS<sup>1,2</sup>  

<sup>1</sup>Section for Biomedical Informatics and Data Science, Yale University School of Medicine, New Haven, Connecticut, USA

<sup>2</sup>Department of Emergency Medicine, Yale University School of Medicine, New Haven, Connecticut, USA

<sup>3</sup>Department of Cardiovascular Medicine, Yale University School of Medicine, New Haven, Connecticut, USA

<sup>4</sup>School of Medicine, University College Dublin, National University of Ireland, Dublin, Republic of Ireland

**Correspondence**

Richard Andrew Taylor, MD, MHS, 464 Congress Ave., Suite 260, New Haven, CT 06519, USA.  
Email: [richard.taylor@yale.edu](mailto:richard.taylor@yale.edu)

Supervising Editor: Lara Goldstein, MD, PhD

**Prior Presentation:** CWS presented some of the findings of this research at the NLP Working Group at the AMIA 2023 Annual Symposium. No abstract or figures will be published online.

**Funding information**

National Heart, Lung, and Blood Institute of the National Institutes of Health, Grant/Award Number: T35HL007649; National Heart, Lung, and Blood Institute, Grant/Award Number: T35HL007649

**Abstract**

**Objectives:** This study presents a design framework to enhance the accuracy by which large language models (LLMs), like ChatGPT can extract insights from clinical notes. We highlight this framework via prompt refinement for the automated determination of HEART (History, ECG, Age, Risk factors, Troponin risk algorithm) scores in chest pain evaluation.

**Methods:** We developed a pipeline for LLM prompt testing, employing stochastic repeat testing and quantifying response errors relative to physician assessment. We evaluated the pipeline for automated HEART score determination across a limited set of 24 synthetic clinical notes representing four simulated patients. To assess whether iterative prompt design could improve the LLMs' ability to extract complex clinical concepts and apply rule-based logic to translate them to HEART subscores, we monitored diagnostic performance during prompt iteration.

**Results:** Validation included three iterative rounds of prompt improvement for three HEART subscores with 25 repeat trials totaling 1200 queries each for GPT-3.5 and GPT-4. For both LLM models, from initial to final prompt design, there was a decrease in the rate of responses with erroneous, non-numerical subscore answers. Accuracy of numerical responses for HEART subscores (discrete 0–2 point scale) improved for GPT-4 from the initial to final prompt iteration, decreasing from a mean error of 0.16–0.10 (95% confidence interval: 0.07–0.14) points.

**Conclusion:** We established a framework for iterative prompt design in the clinical space. Although the results indicate potential for integrating LLMs in structured clinical note analysis, translation to real, large-scale clinical data with appropriate data privacy safeguards is needed.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *Journal of the American College of Emergency Physicians Open* published by Wiley Periodicals LLC on behalf of American College of Emergency Physicians.

**KEYWORDS**

artificial intelligence in medicine, ChatGPT, clinical decision support systems, clinical note analysis, emergency department risk algorithms, HEART score, large language models, natural language processing, prompt engineering

## 1 | INTRODUCTION

### 1.1 | Background

Large language models (LLMs), such as ChatGPT, have drawn attention in healthcare due to their ability to rapidly analyze clinical text and produce coherent, generally accurate responses to detailed instructional prompts.<sup>1,2</sup> By extracting critical, timely information from clinical encounter notes, LLMs can access a key component of the electronic health record that until now has largely been locked in unstructured, text-based formats. Recent clinical developments, including Epic's beta testing of GPT-4 integration, highlight the growing excitement in this domain.<sup>3,4</sup>

### 1.2 | Importance

Despite growing interest in the potential of LLMs to extract meaningful insights from clinical notes, a critical gap persists: the absence of a design framework to improve and evaluate LLM prompts for clinical encounter note analysis. While computer science research has begun to explore iterative prompt development, or "prompt engineering," limited work has been applied in the clinical space.<sup>1,5-7</sup>

### 1.3 | Goals of this investigation

We developed a design framework to improve LLM prompts for clinical note analysis, aiming to produce LLM response choices more accurately aligned with physician assessments. To demonstrate this framework, we used a limited set of synthetic patient notes to refine prompts for automated LLM determination of the HEART (History, ECG, Age, Risk factors, Troponin risk algorithm) score, an established risk-stratification tool for emergency department chest pain evaluation.<sup>8,9</sup> In addition to quantitative evaluation of patient age and troponin, the HEART score integrates detailed assessment of a patient's history, electrocardiogram (ECG) interpretation, and risk factors, each requiring the LLM to analyze unstructured textual data, extract complex clinical concepts, and apply rule-based translation to distinct, verifiable subscores. Through this proof-of-concept HEART score LLM assessment, we sought not only to underscore the feasibility of iterative clinical prompt improvement, but also to highlight emerging considerations for clinical LLM prompt development. The objective of this research was to present and assess a framework for iterative LLM prompt refinement via a pilot, proof-of-concept with a limited set of synthetic notes.

## 2 | METHODS

### 2.1 | Study design

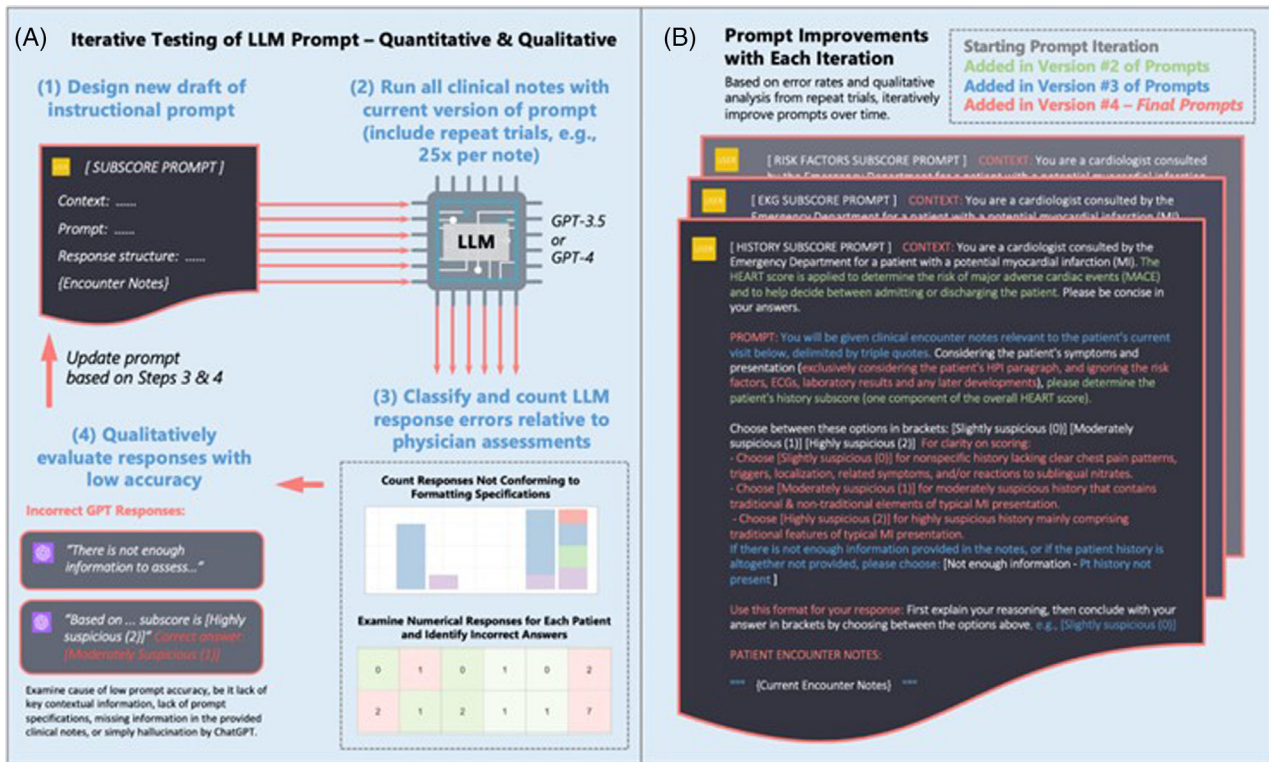
In this proof-of-concept cohort simulation study, we assessed the ability of a novel design framework to iteratively improve LLM prompts. We evaluated this design framework in the context of refining prompts to automate determination of the HEART score across a limited set of synthetic notes representative of the clinical encounter notes typically made available during chest pain work-up in the emergency department.

### 2.2 | Development of synthetic patient notes

We generated 24 total synthetic clinical notes representing four simulated patients. These notes encompassed a range of electronic health record note formats and were furthermore diverse with regard to patient characteristics (eg, varied age, sex, and ethnicity) and presentation (eg, varied symptom type and severity, comorbidities, and past medical history). All emergency department notes and ECG interpretations were written by an emergency medicine physician and all primary care notes by an internal medicine physician.

For each of the four patients, synthetic clinical encounter notes were composed of the following elements:

1. An emergency medicine physician note including history of present illness, past medical history, physical exam, and medical decision making. No HEART scores were provided in any of the physician notes. The emergency department notes frequently excluded some past historical elements mentioned in other notes to ensure the models could not rely entirely on the contents of this note to calculate the HEART score reliably.
2. An ECG interpretation.
3. An emergency department nursing note for the chest pain visit, frequently with additional information not captured in the physician note.
4. A prior emergency physician note for a noncardiac complaint.
5. A prior emergency department nursing note for the noncardiac visit.
6. A prior internal medicine physician note for a routine annual health maintenance visit.
7. A prior internal medicine physician note for a nonroutine visit for a noncardiac complaint.



**FIGURE 1** Iterative framework to improve large language model (LLM) prompts for clinical applications. (A) Flowchart of prompt development framework to iteratively improve prompt design for analysis of clinical encounter notes, with the goal of increasing accuracy of LLM response choices relative to physician gold-standard assessments. (B) Iterative prompt design results from stepwise application of prompt development framework in the context of the HEART (History, ECG, Age, Risk factors, Troponin risk algorithm) score for evaluating acute coronary syndrome risk across four synthetic patient encounter note sets.

Gold-standard HEART scores were separately calculated a priori from the synthetic notes before LLM prompt generation by two blinded emergency physician raters, with adjudication by a third when necessary.

### 2.3 | Iterative prompt evaluation framework

Notes were compiled to mirror a typical electronic health record data export and parsed into an R-based data pipeline to sort and combine the respective notes necessary for each subscore prompt, with logical delimiters and labels separating notes. This study did not include the Age and Troponin subscores due to their structured, formulaic nature, which obviates the need for LLM interpretation.

An initial draft of three prompts for the history, ECG, and risk factor HEART subscores were designed, integrating features of prompt design from previous research.<sup>5,6,8-11</sup> The prompts were deployed via the pipeline, automatically querying and retrieving responses from OpenAI's GPT-3.5 (ChatGPT's underlying LLM) and GPT-4. The pipeline was optimized for parallelization and rate limiting and handled key model settings such as response stochasticity.

This pipeline facilitated Steps 1–3 of the overall framework for iterative prompt design (Figure 1A). For each round of prompt testing, 25 repeat trials are applied to each prompt for each unique note set. Subscore responses were extracted from the LLM responses and com-

pared to gold-standard physician interpretations. For each subscore prompt, non-numerical responses (failure to follow specified response structure or "insufficient information") and numerical errors (incorrect 0, 1 or 2 subscore choice) were counted.

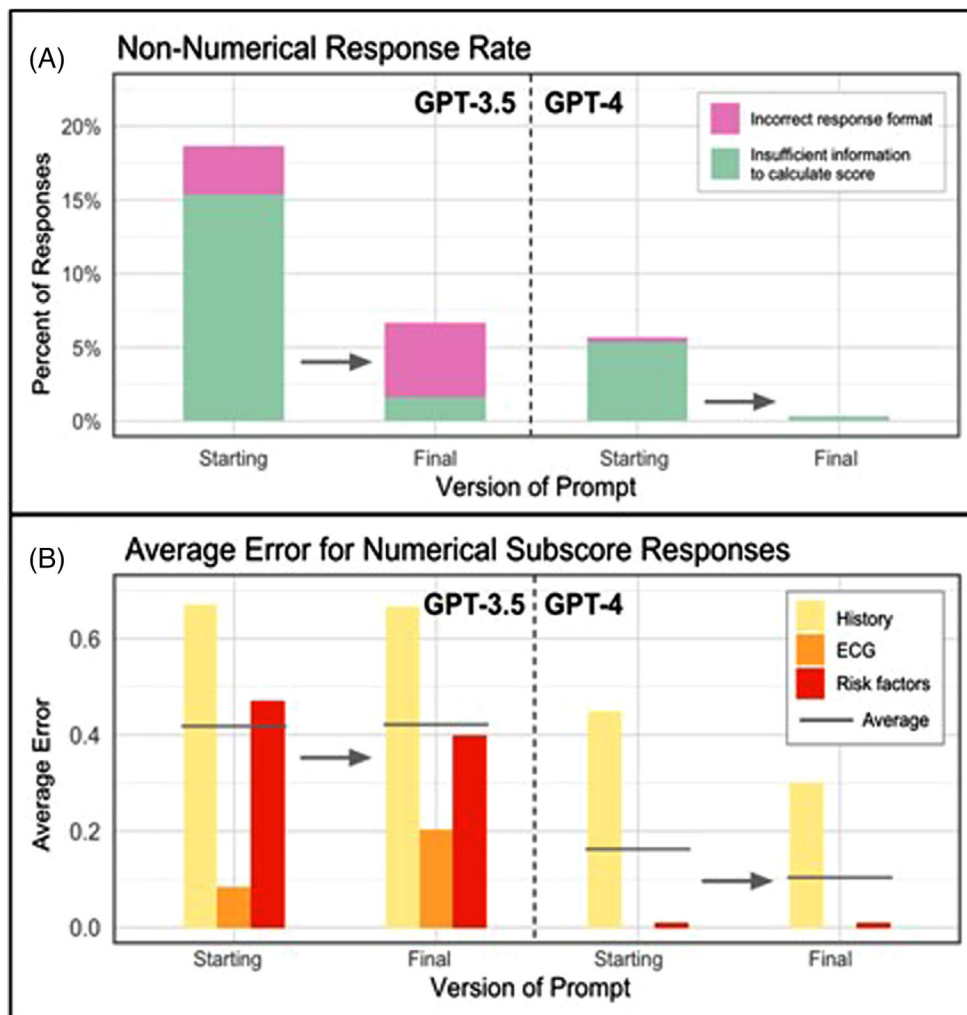
By quantifying errors and identifying which prompts and corresponding note sets had low answer validity or accuracy, it was possible to focus qualitative analysis of response errors. For prompts performing poorly on a specific patient's clinical note set, the interdisciplinary research team—including physicians and computer scientists—could examine the LLM's reasoning explanation in its response to potentially identify the source of discrepancy and guide subsequent prompt refinement for the next iteration (as shown in Figure 1B).

Wilson score intervals were used to calculate 95% confidence intervals for percentages. The full prompt versions, synthetic patient notes, LLM responses, and R data pipeline are open source and available via repository (Supporting Information).

## 3 | RESULTS

### 3.1 | Gold-standard physician HEART scores

For the a priori physician HEART score assessments across the four simulated patients (considering only the unstructured elements of the HEART score: history, ECG, and risk factors), there was one



**FIGURE 2** Errors in large language model (LLM) responses across the three starting versus final subscore prompt versions, tested on four synthetic patient encounter note sets with 25 repeat trials per patient. (A) Rates of LLM responses with non-numerical subscore answers, either due to failure to adhere to the specified response format or an erroneous response of insufficient information, as complete data was provided. (B) Average degree of error in the HEART (History, ECG, Age, Risk factors, Troponin risk algorithm) subscore (on a discrete 0–2 point scale) responses for numerical responses, as compared to gold-standard physician scores. Starting versus final prompt versions from iterative prompt improvement, tested with two LLM models, GPT-3.5 and GPT-4.

discrepancy for a history subscore and one for a risk factors subscore, leading to an overall Cohen's kappa of 0.733. For these specific disagreements, a third blinded emergency physician reviewed the encounters and independently provided scores as adjudication to determine gold standards.

### 3.2 | HEART score prompt results

The three initial subscore prompts underwent three rounds of iteration within the overall framework, with 25 repeat trials for each of the four sets of synthetic patient notes, resulting in 1200 LLM queries each for GPT-3.5 and GPT-4.

From initial to final prompt designs, the overall non-numerical response rate decreased for both LLM models, from 18.7% (95% confidence interval [CI]: 14.7%–23.5%) to 6.7% (4.4%–10.1%) for GPT-3.5

and 5.7% (3.6%–8.9%) to 0.3% (0.1%–1.9%) for GPT-4 (Figure 2A). Among numerical subscore responses (on a discrete 0–2 point scale), average error remained relatively constant for the initial versus final prompt versions for GPT-3.5, with a mean error of 0.42 (0.33–0.50) versus 0.42 (0.36–0.48) points across the subscore prompt results (Figure 2B). For GPT-4, this same average error decreased from 0.16 (0.11–0.22) to 0.10 (0.07–0.14) points. GPT-3.5 had higher variability in its responses, with the mean numerical subscore for the final prompts having a standard deviation of 0.52 for GPT-3.5 compared to 0.33 for GPT-4. Given the final overall HEART score risk stratification buckets (0–3, 4–6, or 7–10 points when summing all 5 HEART subscores), 81.5% (71.7%–88.4%) of GPT-3.5 and 100% (96.3%–100%) of GPT-4 numerical final HEART score calculations predicted the correct risk group.

During the process of iterative qualitative error analysis, certain decision points regarding clinical prompt design emerged as

**TABLE 1** Key considerations and decision points when developing large language model (LLM) prompts for clinical notes, with examples in the context of prompts to automate the HEART (History, ECG, Age, Risk factors, Troponin risk algorithm) score: We identified several key considerations for clinical prompt development based on computer science and industry guidelines, as well as from deployment of our iterative prompt design framework with a limited set of synthetic patient notes. These design considerations require future research to determine best practices for LLM prompt optimization in the clinical setting.

Prompt design consideration	Description of decision point ("Option 2" corresponds to the design choices in the final iteration of our LLM prompts for automated HEART scores)
One-pass prompt vs. 3 subscore questions?	Option 1: Ask for the History, ECG, and Risk factors subscores in one prompt. Option 2: Ask for each of the three subscores in three separate prompts, and then subsequently sum numerical responses to determine overall HEART score.
Explain reasoning?	Option 1: Request LLM only provides final answer in brackets; "No prose." Option 2: Request LLM explains step-wise reasoning before final answer.
One-shot learning?	Option 1: Provide example of step-wise answer structure for each subscore. Option 2: No example; "zero-shot learning."
Stochasticity of model?	Option 1: Model "temperature" = 0.7 → more randomness in responses. Option 2: Model "temperature" = 0.3 → responses more determinant.
Order of notes?	Option 1: For prompts with multiple notes, list notes from least to most recent. Option 2: For prompts with multiple notes, list notes from most to least recent.
Instructional phrase?	Option 1: Include contextual phrase as a "system" wide instructional phrase. Option 2: Include "CONTEXT:" phrase at beginning of each prompt.
Delimiting notes?	Option 1: Do not delimit notes. Option 2: Delimit encounter notes (e.g., triple quotations ""(PatientNotes)"" and "#####" between encounters).

Abbreviation: ECG, electrocardiogram.

particularly influential on response validity and accuracy, with key considerations outlined in **Table 1**.

Each round of prompt testing (300 queries per model) cost \$1.89 and \$21.62 for GPT-3.5 and GPT-4, respectively.

## 4 | LIMITATIONS

Due to the small sample size and synthetic nature of our encounter notes, several limitations emerged.

First, while our synthetic notes encompassed a variety of structures, they did not fully represent the breadth of real-world clinical notes. Our notes were largely interpretable when transposed to our sample extract without specific textual formatting, a condition that might not hold for all unstructured clinical notes. Furthermore, our longest prompt with synthetic notes was approximately 2100 words, which may be significantly shorter than some real-world patient's compiled clinical notes. Iterative prompt improvement with our current design framework may not be able to overcome potential inherent performance loss due to long, convoluted patient medical records. Future updates to our framework may thus need to incorporate strategies for effectively selecting and processing longer patient notes with diverse note formatting.

Second, this study was limited by its sample size with regard to model performance and generalizability. The small sample size likely affected type II error, suggestive of possible falsely measured change

in LLM error performance. A larger sample size of real-world clinical notes would address core variability of the data, decreasing standard error as well as allowing for more extensive experimentation with the stochastic processes of the LLM. This experimentation would reduce both the type II error and the additional source of error of the inherent stochastic nature of the LLM. Future experiments could also reinforce this assessment by splitting data into "test" and "training" sets to assess tuned accuracy of the LLM.

At present, data privacy concerns necessitated the use of synthetic notes for this pilot study given the serious risks of transmitting real-world private health information through nonsecure programming interfaces or even unintentionally embedding patient data into self-training LLMs, as is the case with some versions of ChatGPT such as the web browser deployment.<sup>12</sup> The recent announcement of nontraining, HIPAA compliant solutions is a promising development for future research.<sup>13</sup>

## 5 | DISCUSSION

We developed a design framework to iteratively improve LLM prompts for clinical note analysis, aiming to increase accuracy of LLM response choices relative to physician gold-standard assessments. As a pilot proof-of-concept study with a limited set of synthetic patient notes, we demonstrated the framework's ability to improve LLM prompt design for automated HEART scores—a task requiring extraction

of clinical concepts from unstructured encounter notes followed by application of rule-based logic to determine distinct, verifiable subscores.

Our framework integrated quantitative error assessment to guide subsequent qualitative identification of the mechanism of prompt failure, be it due to misunderstanding of instructions, insufficient contextual information, or a more nuanced discrepancy in prompt interpretation. By prompting the LLM to explain its reasoning in each query before providing its answer, it facilitated error troubleshooting. Moreover, employing repeat trials with nonzero model stochasticity leveraged the inherent variability of LLM responses, offering insights into potential prompt issues or errors that could surface during subsequent tests with novel patient notes. This process of iterative prompt design was more impactful for GPT-4 relative to GPT-3.5, with a greater proportional improvement in valid and correct response rates across our limited data set; this result suggests that as LLM models continue to advance and can understand more complex, contextually nuanced clinical prompts, the process of iterative prompt design will become increasingly important to achieving optimal accuracy.

In summary, this study presented a framework for systematic prompt design to optimize structured analysis of clinical encounter notes, serving as a translation of the emerging best practices from industry and computer science literature.<sup>5-7,10,11</sup> We demonstrated this framework via LLM prompt improvement for automated HEART score determination across a limited set of synthetic patient notes. Our results have shown how, as LLMs continue to advance and gain clinical popularity, physician input will be needed to shape LLM prompt optimization to ensure reliability and validity of LLM outputs. While the results are promising, challenges related to note diversity, overfitting, and patient privacy persist. This study has emphasized the potential of iterative, systematic prompt engineering in optimizing LLMs for healthcare applications.

#### AUTHOR CONTRIBUTIONS

Donald S. Wright, Vimig Socrates, and R. Andrew Taylor conceived of the research idea. Donald S. Wright and Catherine X. Wright created all synthetic patient notes. Mark Iscoe and R. Andrew Taylor reviewed these synthetic notes to ascertain true HEART scores. The data pipeline was developed by Conrad W. Safranek and Thomas Huang. Conrad W. Safranek was responsible for data collection. Data analysis was undertaken by Conrad W. Safranek, Thomas Huang, Donald S. Wright, Vimig Socrates, Rohit B. Sangal, David Chartash, and R. Andrew Taylor. Conrad W. Safranek, Thomas Huang, and Donald S. Wright drafted the initial manuscript. All authors engaged in the review and editing of the manuscript, and approved the final draft.

#### ACKNOWLEDGMENTS

The authors thank David Yang, MD for reviewing and adjudicating any disagreements in emergency medicine physician HEART score assessments. OpenAI's GPT-4 (2023 version) was used to develop some of the

R code used in the data pipeline; all GPT-4 code outputs were reviewed and edited for accuracy prior to deployment. Research reported in this publication was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award T35HL007649 (author CWS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

#### CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

#### DATA AVAILABILITY STATEMENT

The full prompt versions, synthetic patient notes, LLM responses, and R data pipeline are open source and available via repository ([Supporting Information](#)).

#### ORCID

Conrad W. Safranek BS  <https://orcid.org/0000-0003-1985-9432>

R. Andrew Taylor MD, MHS  <https://orcid.org/0000-0002-9082-6644>

#### TWITTER

Conrad W. Safranek BS  <https://twitter.com/ConradSafranek>

R. Andrew Taylor MD, MHS  [https://twitter.com/Yale\\_EM](https://twitter.com/Yale_EM)

#### REFERENCES

- Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA*. 330:866-869. doi:10.1001/jama.2023.14217
- Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMSr2214184
- Epic. Epic and microsoft bring GPT-4 to EHRs. Accessed August 22, 2023. <https://www.epic.com/epic/post/epic-and-microsoft-bring-gpt-4-to-ehrs>
- Diaz N. 6 hospitals, health systems testing out ChatGPT. 2023. Accessed August 22, 2023. <https://www.beckershospitalreview.com/innovation/4-hospitals-health-systems-testing-out-chatgpt.html>
- Giray L. Prompt engineering with ChatGPT: a guide for academic writers. *Ann Biomed Eng*. 2023;51:2629-2633. doi:10.1007/s10439-023-03272-4
- White J, Fu Q, Hays S, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv*. 2023. doi:10.48550/arXiv.2302.11382
- Strobel H, Webson A, Sanh V, et al. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE Trans Vis Comput Graph*. 2023;29(1):1146-1156. doi:10.1109/TVCG.2022.3209479
- Six AJ, Backus BE, Kelder JC. Chest pain in the emergency room: value of the HEART score. *Neth Heart J*. 2008;16:191-196. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2442661/>
- Backus BE, Six AJ, Kelder JC, et al. Chest pain in the emergency room: a multicenter validation of the HEART score. *Crit Pathw Cardiol*. 2010;9(3):164. doi:10.1097/HPC.0b013e3181ec36d8
- OpenAI. GPT best practices – OpenAI API documentation. OpenAI Platform. Accessed January 23, 2024. <https://platform.openai.com/docs/guides/prompt-engineering>
- Ng A, Fulford I. ChatGPT prompt engineering for developers – learning platform beta. DeepLearning AI. Accessed August 23,

2023. <https://learn.deeplearning.ai/chatgpt-prompt-eng/lesson/1/introduction>
12. Marks M, Haupt CE. AI chatbots, health privacy, and challenges to HIPAA compliance. *JAMA*. 2023;330(4):309-310. doi:[10.1001/jama.2023.9458](https://doi.org/10.1001/jama.2023.9458)
13. Microsoft. Microsoft Azure compliance offerings. Accessed January 23, 2024. [https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-database-security/?ef\\_id=\\_k\\_CjwKCAiA5L2tBhBTEiwAdSxJX23x8aKG3EUQyBI8BfhxzeXLIayK8tsUfrAGGJZtWt8jEDmYjHxrNBoCulsQAvD\\_BwE\\_k\\_&OCID=AIDcmme9zx2qiz\\_SEM\\_k\\_CjwKCAiA5L2tBhBTEiwAdSxJX23x8aKG3EUQyBI8BfhxzeXLIayK8tsUfrAGGJZtWt8jEDmYjHxrNBoCulsQAvD\\_BwE\\_k\\_&glid=CjwKCAiA5L2tBhBTEiwAdSxJX23x8aKG3EUQyBI8BfhxzeXLIayK8tsUfrAGGJZtWt8jEDmYjHxrNBoCulsQAvD\\_BwE](https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-database-security/?ef_id=_k_CjwKCAiA5L2tBhBTEiwAdSxJX23x8aKG3EUQyBI8BfhxzeXLIayK8tsUfrAGGJZtWt8jEDmYjHxrNBoCulsQAvD_BwE_k_&OCID=AIDcmme9zx2qiz_SEM_k_CjwKCAiA5L2tBhBTEiwAdSxJX23x8aKG3EUQyBI8BfhxzeXLIayK8tsUfrAGGJZtWt8jEDmYjHxrNBoCulsQAvD_BwE_k_&glid=CjwKCAiA5L2tBhBTEiwAdSxJX23x8aKG3EUQyBI8BfhxzeXLIayK8tsUfrAGGJZtWt8jEDmYjHxrNBoCulsQAvD_BwE)

### SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Safranek CW, Huang T, Wright DS, et al. Automated HEART score determination via ChatGPT: Honing a framework for iterative prompt development. *JACEP Open*. 2024;5:e13133. <https://doi.org/10.1002/emp2.13133>

### AUTHOR BIOGRAPHY



**Conrad W. Safranek** is a medical student in the Yale School of Medicine in New Haven, Connecticut.