Prompt Engineering GPT-4 to Answer Patient Inquiries: A Real-Time Implementation in the Electronic Health Record across Provider Clinics

Majid Afshar, MD, MS\(^1,2\); Yanjun Gao, PhD\(^1\); Graham Wills, PhD\(^2\); Jason Wang, BS\(^1\); Matthew M Churpek, MD, MPH, PhD\(^1\); Christa J Westenberger\(^2\), BSN, RN; David T Kunstman\(^{2,3}\), MD; Joel E Gordon\(^{2,3}\), MD; Frank J Liao\(^{2,4}\), PhD; Brian Patterson, MD, MPH\(^{2,4}\)

\(^1\)Department of Medicine, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA

\(^2\)Information Systems, University of Wisconsin Health System, Madison, WI, USA

\(^3\)Department of Family Medicine and Community Health, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA

\(^4\)Department of Emergency Medicine, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA

Corresponding author:

Majid Afshar, MD, MS

600 Highland Avenue

Madison, WI 53792

Majid.afshar@wisc.edu

608-263-6400

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
ABSTRACT

Background: The integration of Large Language Models like GPT-4 in healthcare has opened new avenues for improving patient-provider communication. However, the effectiveness of prompt engineering strategies, which have been shown to affect the quality and accuracy of model output, remains unexplored in this context. This study aims to evaluate the impact of manual versus semi-automated prompt engineering to improve the usability of AI-generated responses to patient inquiries with real-time integration into an electronic health record.

Methods: A pre-post study over eight months was conducted at University of Wisconsin Health, involving 27 providers across multiple specialties. The study compared GPT-4 in a pre-period use of manual prompts with a post-period semi-automated engineered prompt that incorporated iterative design and prompt evaluation scoring. Testing by informaticists was completed before deployment of the new prompt. The primary outcome was the number of AI-generated draft messages used in a mixed effects model accounting for multiple messages by same provider. Secondary outcomes included message editing metrics and sentiment analysis.

Results: Of the 7,605 draft messages generated by GPT-4 and seen by providers during the study period, 17.5% (n=1,327) were used by the providers and 2.6% (n=202) were left identical or nearly identical in editing by the providers. The number of messages used decreased in the post-period with the new prompt (beta coefficient -0.10; 95% CI: -0.11 - -0.09, p<0.01); however, there was a reduction in negative sentiment with an odds ratio of 0.43 (95% CI: 0.36 - 0.52, p<0.01) with the new prompt design.

Discussion: The decrease in negative sentiment with the new prompt demonstrated improvement in the AI-generated content quality but usage remained low by providers. This study highlights a need for better alignment between prompt engineering and human factors engineering.

Trial Registration number: Not Applicable
INTRODUCTION:

The emergence of Large Language Models (LLMs), especially OpenAI's ChatGPT, has marked a pivotal turn in generative Artificial Intelligence (AI), opening novel avenues in healthcare delivery.\(^1\), \(^2\) The introduction of GPT-3 in 2020 demonstrated the first capabilities of a pre-trained language model trained on a large corpus of text with 175 billion parameters to perform well on benchmark tasks in the general domain without any additional training or fine-tuning.\(^3\) This was evidenced by the model's capacity for a new method of prompt engineering through natural language prompts with examples, called in-context learning, with demonstrated success also in “zero-shot” prompts where no examples are provided. The subsequent evolution of the field of prompt engineering has been extraordinary with many advancements in more rigorous methods for designing prompts. In addition, the advent of GPT-4, heralded additional enhancements that now establish it as the state-of-the-art LLM in many tasks, including question answering.\(^4\) Nonetheless, as LLMs continue to evolve, prompt engineering is increasingly more important, as small changes in the natural language input can impact the quality and relevance of the generated output, necessitating rigorous optimization for the realization of diverse use cases.\(^5\)

Beginning in 2023, a leading electronic health records (EHR) vendor, Epic Systems Corp (Epic), expanded its collaboration with Microsoft (Microsoft Corp®) to incorporate GPT-4, marking one of the seminal deployments of generative AI in healthcare.\(^6\) Despite the ability to bring generative AI into real-time applications, there remained a knowledge gap in the design of prompts that can cater to patients and providers, which are necessary to generate responses that are detailed, accurate, empathetic, and anchored in the most recent medical evidence. The University of Wisconsin (UW Health) is one of the inaugural sites to operationalize a use case of GPT-4\(^7\), with the responsibility for prompt engineering being delegated to the health system. Initial strategies encompassed the development of customized prompts, sourced through a collaborative ecosystem, established between Epic and the participating health systems, including UW Health.

To examine the effectiveness of different prompt engineering strategies, this case study examines a pre-post design comparing crowd-sourced manual prompts versus a novel semi-automated approach.
We hypothesized that using a semi-automated approach to prompt engineering will increase the usability of the generated output in providers responding to patient inquiries. We also shared user-friendly prompt engineering and library software available at https://cliniprompt.medicine.wisc.edu/ for other health systems to use.

METHODS:

Hospital Setting and Study Period

The GPT-4 use case was implemented at UW Health with 27 physician providers across family medicine, internal medicine, dermatology, oncology, psychiatry, and pediatric clinics. UW Health uses Epic (Epic© 2023 vNovember22) as its electronic health record (EHR) and MyChart is the patient engagement platform of the Epic EHR. MyChart facilitates secure online access for patients to portions of their medical records, allowing them to manage their healthcare information, schedule appointments, and communicate with their healthcare providers. The study was a quality improvement initiative by the health system to facilitate provider responses to patient medical questions in MyChart about their health and healthcare. The use case was evaluated in the UW Health infrastructure using a pre-post quasi-experimental study design with April 30, 2023, to August 29, 2023, serving as the pre-period with the manual, crowd-sourced prompt design and August 30, 2023, to December 23, 2023, serving as the post-period with the semi-automated engineered prompt design. The study was deemed exempt by the UW-Madison Institutional Review Board.

Real-time implementation of GPT-4 large language model (LLM)

In one of the first real-time use cases of an LLM in healthcare operations, UW Health activated the Epic interface of GPT-4 to generate draft responses to MyChart messages for providers. The prompts were built as Epic SmartText and combined with other Epic SmartText data from the EHR as the input to the model along with the patient query. Prompts were developed for the following categories of queries
that were pre-determined by Epic and Microsoft: (1) General; (2) Medication; (3) Results; and (4) Paperwork. For instance, for the General category, UW Health used a prompt that began with the following prompt: “Your job is to create a draft message for a provider to use in response to patient messages…” followed by the patient question and included the SmartText to bring in the patient’s medication list from the EHR’s Medication Administration Record. No access to the GPT-4 Application Program Interface (API) was provided to modify any model hyperparameters. The hyperparameters were set by Epic with a temperature of 1 and an output token limitation of 500.

**Prompt Engineering Pre-Period with Manual Design**

The prompts were manually designed by informaticians at UW Health with testing of the prompts in the Epic testing environment. The manual prompt deployed at UW Health is shown in Appendix 1. During the rollout of the LLM, Epic provided examples of prompt in their Epic Galaxy User Web that were crowd-sourced from eight health systems nationwide, including UW Health. Recommended prompts were provided for GPT 3.5 DaVinci and GPT-4 and were binned into the four pre-determined Epic categories mentioned above.

**Prompt Engineering Post-Period with Novel Semi-Automation Design**

Due to low initial usage rates, the informatics team engineered a new prompt using in-context learning with few-shot examples, comprised of the following components: (1) semi-automated prompt design using ChatGPT; (2) few-shot examples generated by experts; and (3) synthetic data generated by ChatGPT for a reference dataset to examine prompt performance metrics. For the first component, manual task-specific prompts were provided to ChatGPT with a prompt to provide 20 paraphrases. These prompts were ranked by their perplexity score, selecting the ones with the lowest perplexity. Gonen and colleagues introduced BETTERPROMPT for perplexity scoring, utilizing the Selecting Prompts by Estimating Language model Likelihood (SPELL) method.8 The top five paraphrased prompts with the lowest perplexity scores were advanced to the second component.
In the second component, the top five prompts were combined with few-shot examples, using expert input from end-users and informatics specialists to provide synthetic examples for each category. The final five prompts with the synthetic few-shot examples were a representation of in-context learning. In the third component, self-consistency was measured by generating 25 outputs for each prompt and then measuring its cosine similarity through pre-trained text representation (using Google’s BERT pretrained language model to encode embeddings) against a reference dataset. The reference dataset was a synthetic dataset generated by ChatGPT using the prompts with examples shown in Appendix 2. The top 5 lowest perplexity prompts with few-shot examples (in-context learning) yielded 500 self-consistency evaluations across the four categories. The generated output with the highest mean cosine similarity score in each category (General, Medicine, Results, Paperwork) was selected as the final prompt and the related SmartText of relevant tabular EHR data was added to provide a final prompt template. The full design framework is depicted in Figure 1 and the final semiautomated engineered prompt is shown in Appendix 3. We also provided an open-source website resource for engineering prompts with a prompt library for healthcare use cases, including our own (https://cliniprompt.medicine.wisc.edu/).

Analysis Plan:

Baseline characteristics of providers were compared using nonparametric tests with Wilcoxon rank-sum tests for continuous variables and chi-square tests for proportions. The primary outcome was usage defined by the number of AI-generated draft messages used by providers between the pre- and post-period. Messages used represented messages seen by the provider and sent to the patient after final editing. A mixed effects linear regression model was applied to examine the number of draft messages used by providers between the pre- and post-period. A random intercept for each provider addressed the within-group correlation to account for multiple messages seen and used per provider.

Providers edited the generated AI response before sending it back to patients. In a secondary analysis, edit metrics were also measured between the AI-generated response and the final accepted response with the Damerau-Levenshtein distance as a string metric to measure the edit distance between
the AI-generated message and the final message used by the provider. The results were normalized on a scale between 0 and 1 and values >0.9999 were considered “identical”, ≥0.9750 and <0.9999 as “nearly identical”, and <0.9750 and ≥0.6666 as “similar”, with values under 0.6666 considered “different.” Sentiment analysis was also conducted with a “thumbs down” symbol entered by the provider in Epic to indicate negative sentiment. To analyze the proportion of messages rated as negative sentiment by providers between the pre- and post-period, a Generalized Linear Mixed Model (GLMM) was employed with a random intercept for each provider to account for individual variations among providers. All analyses were performed in Python (Python Software Foundation, v3.10) and R (R Foundation, v4.3)

RESULTS:

The top five prompts from the group of 50 that had the lowest perplexity measures are shown in Table 1. A synthetic randomly generated data set of n=20 was used for the four categories to perform self-consistency measures. The final optimized prompt template with the lowest perplexity prompt with few-shot examples had the highest mean cosine similarity score of 99.1 (95% CI: 98.9-99.3). An example of the semi-automated prompt template is shown in Figure 2 in comparison to a manual prompt.

A total of 7,605 draft messages were generated by GPT-4 and 1,327 (17.5%) were used as a reply to a patient’s inquiry. The median age of providers was 40.1 years (IQR 36.3 - 48.0) and 55.6% (n=15) were female. The top three specialties represented were family medicine (39.3%, n=11), internal medicine (25.0%, n=7), and dermatology (21.4%, n=6).

There was no significant increase in the number of messages used between the pre- and post-period in unadjusted analysis (778 vs. 549, p=0.89). In mixed linear regression accounting for the provider and number of messages per provider, the engineered prompt in the post-period was associated with a decrease in message usage (beta coefficient -0.10; 95% CI: -0.11 - -0.09, p<0.01). The amount of editing performed for all messages seen by providers is shown in Table 2. Across all messages, those that were left identical or nearly identical in editing by providers was 2.6% (n=202). The proportion of
messages rated as negative sentiment with a ‘thumbs down’ in the Epic interface decreased from 12.0% (n=468) in the pre-period to 8.2% (n=307) in the post-period (p<0.01). In mixed effects logistic regression, the odds ratio (OR) of having a negative sentiment was lower with the semi-automated engineered prompt versus the manually designed prompt with an OR of 0.43 (95% CI: 0.36 - 0.52, p<0.01).

During the post-period implementation, a new wave of providers was included in the GPT-4 system rollout using the semi-automated prompt, and they had higher usage rates than the providers in the previous wave used for the primary analyses (p<0.01) as depicted in Figure 3.

**DISCUSSION:**

In the digital era, several studies have showcased the potential of LLMs to aid healthcare providers in responding to patient inquiries.\(^{10-12}\) As of this publication's date, no studies have been published on real-time LLM implementations integrated in an EHR for clinical operation. Our study contributes to the healthcare field by offering user-friendly prompting software and sharing one of the pioneering experiences with Epic’s GPT-4 interface for managing patient communications. The open-source CliniPrompt software serves as a library for knowledge sharing of prompt designs and use cases.

Our results indicate that while our prompting strategies enhanced the quality of the generated responses in the design phase and decreased negative sentiments in production, usage rates with the engineered prompt did not increase. The disconnect between improved system performance based on sentiment analysis and lack of a corresponding increase in usage is not necessarily surprising. In general, the uptake of new clinical technologies is challenging\(^ {13}\), with multiple barriers to physicians incorporating AI input into workflows.\(^ {14}\) Among the messages reviewed by healthcare providers, a significant number were either modified substantially or not utilized. However, when new groups of providers were introduced to the updated prompt system, there was an observable increase in usage (although this was not part of our primary analysis). It is possible that subjects formed an opinion (and usage pattern) of draft usefulness which was less amenable to change even when draft quality increased. Our study illustrates the
delicate balance between enhancing output quality through prompt engineering and the effective design and implementation of the systems which deliver output to clinicians. As LLMs become more ingrained in healthcare, understanding and managing this balance will be crucial for optimizing the integration and effectiveness of these advanced technologies in clinical settings.

Alternatively, the disconnect between improvement in sentiment and lack of increased uptake may suggest that sentiment analysis did not capture the alignment with provider preferences or communication styles necessary to improve uptake. Prior work has shown how persona is important in prompting strategies. Our provider mixture included a diverse group of providers from several specialties who may have had different preferences in writing style and approach to patient questions necessitating personalized prompts. At this point, we can only provide one system level prompt; in the future, we may find it necessary to split prompts by specialty or other grouping to improve quality and uptake.

Prompt engineering can be used to access the model’s internal knowledge via natural language, but the prompts can perform worse when compared to direct prompts when they are too verbose with metalinguistic text. Optimizing prompts to lower perplexity is one promising direction achieved by our CliniPrompt software. The manual prompt designed by end-users was reduced in its metalinguistic properties, allowing for more token space for few-shot examples to provide in-context learning. Other prompt frameworks in the general domain (i.e., PromptSource, Comet and LangChain) and healthcare domain (i.e., MediPrompt) are available with similar approaches but they are designed for data engineers and programmers. Our software interface allows clinical users to refine and improve their prompts without the need for programming. While our software was utilized for a specific use case in GPT-4, it is also flexible to enter new use cases, build a library of prompts and examples, and incorporate other LLMs for testing prompts.

CONCLUSION:
While prompt engineering showed some improvement in the sentiment of generated responses, the findings reflect ongoing challenges in AI integration, including aligning AI output with provider expectations and communication styles. The decrease in usage with the engineered prompt suggests a need for further refinement in understanding provider needs and preferences. This study sheds light on the complex dynamics of AI implementation in healthcare, highlighting both the potential benefits and the challenges that need to be addressed to optimize the use of LLMs in clinical practice.

**ACKNOWLEDGEMENTS:** UW Health Epic team, including Samantha Garwood, Katie J Edwardson, and Karen Campbell, for experimenting with different prompt designs in the Epic test environment. We also thank the University of California-Berkely Masters students in Artificial Intelligence for their help in building the CliniPrompt software.
### Table 1. Perplexity Measures of top 5 paraphrased prompts from ChatGPT

<table>
<thead>
<tr>
<th>Prompts</th>
<th>Perplexity Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your role as a primary care provider is marked by helpfulness, respectfulness, and honesty when answering patient questions submitted electronically. It's crucial to provide useful answers while ensuring safety. Avoid mentioning other physicians and stay away from harmful, inappropriate content. Maintain a socially unbiased and positive tone in your responses. Address &lt;User Message&gt; with a formal greeting and integrate relevant patient details from the electronic health record enclosed in [ ] brackets. The</td>
<td></td>
</tr>
<tr>
<td>Embrace the role of a supportive, respectful, and truthful primary care provider who addresses electronically submitted patient questions. Your focus lies in providing helpful guidance while ensuring safety. Avoid references to other physicians and avoid content that might be harmful or inappropriate. Maintain a socially unbiased and positive tone in your responses. Craft replies to &lt;User Message&gt; with a formal greeting and integrate relevant patient details from the electronic health record enclosed in [ ] brackets. The</td>
<td></td>
</tr>
<tr>
<td>Your role involves being a helpful, respectful, and honest primary care provider, answering patients' electronically submitted questions. Always aim to provide assistance while prioritizing safety. Do not make referrals to other physicians or include any content that could be harmful or inappropriate. Keep your responses unbiased and positive. Respond to &lt;User Message&gt; with a formal greeting and incorporate relevant patient details from the electronic health record enclosed in [ ] brackets. The</td>
<td></td>
</tr>
<tr>
<td>You embody a supportive, courteous, and truthful primary care provider who addresses electronically submitted patient inquiries. Your goal is to offer helpful responses while maintaining safety standards. Avoid references to other doctors and steer clear of harmful or inappropriate content. Keep your responses socially neutral and optimistic. Respond to &lt;User Message&gt; with a professional salutation and include pertinent patient information from the electronic health record enclosed in [ ] brackets. The</td>
<td></td>
</tr>
<tr>
<td>Embrace the role of a supportive, respectful, and honest primary care provider who responds to electronically submitted patient questions. Your focus lies in providing helpful guidance while ensuring safety. Avoid references to other doctors and steer clear of harmful, inappropriate content. Maintain a socially unbiased and positive tone in your responses. Respond to &lt;User Message&gt; with a formal greeting and integrate pertinent patient information from the electronic health record enclosed in [ ] brackets. The</td>
<td></td>
</tr>
</tbody>
</table>
**Table 2.** Editing by Providers of all GPT-4 generated messages seen by the provider (n=7,605)

<table>
<thead>
<tr>
<th>Editing Distance, n (%)</th>
<th>Pre-Period (Manual Prompt) n=3,882</th>
<th>Post-Period (Semi-automated Prompt) n=3,723</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical</td>
<td>4 (0.1)</td>
<td>14 (0.4)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Nearly Identical</td>
<td>110 (2.8)</td>
<td>74 (2.0)</td>
<td></td>
</tr>
<tr>
<td>Similar</td>
<td>278 (7.2)</td>
<td>224 (6.0)</td>
<td></td>
</tr>
<tr>
<td>Different/Not Used</td>
<td>3490 (89.9)</td>
<td>3411 (91.6)</td>
<td></td>
</tr>
</tbody>
</table>

* Damereau-Levenshtein distance was used as the string metric to measure the edit distance between the GPT-4 generated message and the final message used by the provider. The results were normalized on a scale between 0 and 1 and values >0.9999 were considered “Identical”, >0.9750 and <0.9999 as “Nearly Identical”, and <0.9750 and >0.6666 as “Similar”, with values under 0.6666 considered “Different”. GPT-4 messages not used by the provider and fully edited are included in the “Different” category.
REFERENCES


Figure 1. Prompt Auto-selection Framework and Synthesized Exemplars for In-Context Learning

LLM = large language model; API = application program interface

Back-end development involves: (1) automatically extend a small seed set of manually written prompts by paraphrasing using GPT3.5; (2) select prompts with the lowest perplexity score; and (3) choose the prompt with the highest self-consistency. For each paraphrased prompt, we iterate over the input 5-10 times with temp = 0.5. The prompt with the lowest perplexity is chosen. Examples are then added to the prompt to finalize in-context learning. The generated output is then tokenized along with the ground truth from a synthetic dataset. Using cosine similarity, we measure the similarity of each response to the ground truth and average these results to provide top 5 results for user testing. To make our database as easy as possible to manage, we chose to utilize Firebase, a popular BaaS (Backend-as-a-Service). This service provides an easy way to implement authentication while allowing us to parse the database quickly using precalculated indexes. This is especially important to allow our users to find prompts created by themselves as well as other users.
**Figure 2. Example pre- and post-prompt engineering**

<table>
<thead>
<tr>
<th>Manual Pre-Period: Test patient question and generated output</th>
<th>Semi-automated Post-period: Test patient question and generated output</th>
</tr>
</thead>
<tbody>
<tr>
<td>I have some lower back pain that is going on 3 days now. I lifted a heavy tree limb and now have this pain and it is radiating down my left leg. Thanks</td>
<td>I have some lower back pain that is going on 3 days now. I lifted a heavy tree limb and now have this pain and it is radiating down my left leg. What should I do?</td>
</tr>
</tbody>
</table>

© 2023 Epic Systems Corporation.
Figure 3. Provider waves of GPT-4 System Rollout in the Electronic Health Record

The figure shows usage rates as messages used by the provider per week. The vertical line is the implementation of the newly engineered prompt and delineates the pre- and post-period for primary analysis. A number of clinicians were added after the new prompt had been implemented (blue and red waves were from beginning of the study). The new wave of providers has higher usage rates than the providers in prior waves.