

GPT-4 Prompt Engineering For Automated Patient Response Generation

Alex Budko (COL 2026)¹, Katrina Liu (COL 2025)¹, Kevin B. Johnson, MD, MS^{1,2}

Funded by: Penn Undergraduate Research Mentoring Program

¹ School of Engineering and Applied Science, University of Pennsylvania, ² Perelman School of Medicine, University of Pennsylvania

Abstract

In recent years, the growing demands of electronic health records (EHR), particularly in-basket patient messages, have intensified physician burnout.^{1,2} As GPT-4 is often considered as the golden standard for generating conversational content when evaluating other large language models (LLMs), this poster focuses on the potential of using GPT-4 to automate doctor responses as a solution to the issue. Our automation pipeline incorporates prompt engineering techniques, such as zero-shot, few-shots, and chain-of-thought. We evaluated our pipeline using a survey based on 5 metrics: Empathy, Relevance, Medical Accuracy, Readability, and if the LLM-generated messages can be differentiated from human-written messages. Our survey is currently pending IRB approval.

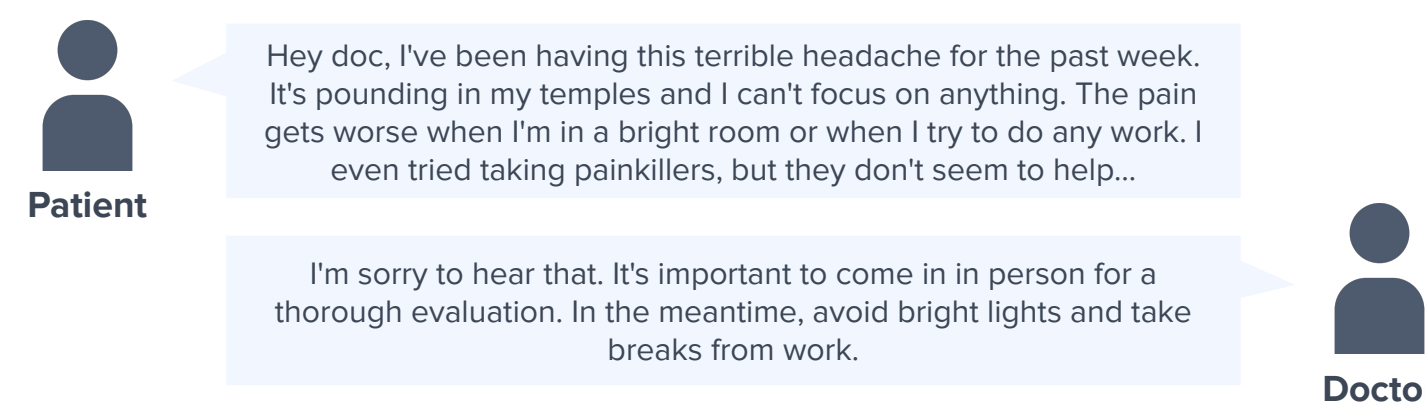


Fig 1. Diagram representation of a patient-doctor dialogue using the EHR system's in-basket messaging feature (not based on actual data).

Introduction

The rising prominence of electronic health records (EHR), especially in-basket patient messages, is a current focal point in the healthcare sector, given its purported link to increasing physician burnout.^{1,2} Patient portals act as crucial communication bridges between healthcare providers and patients, underscoring their significant role in today's medical ecosystem. Accurate automation of these portals holds the potential to substantially mitigate administrative challenges and enrich patient understanding.

Parallel to this, the realm of artificial intelligence continues to progress at a brisk pace, with the emulation of nuanced human conversation remaining a central pursuit. Recently, there's been a notable interest in prompt engineering techniques, such as few-shot learning and chain-of-thought prompting.³

With these prevailing challenges and advancements in mind, we present our exploratory work on Katralex, an automated response generation framework. Leveraging the capabilities of the conversational language model GPT-4, Katralex aims to embed the developments of prompt engineering to optimize physician responses to in-basket messages. This is a step towards our aspiration of not only improving digital patient interfaces but also fostering a deeper, more intuitive connection between patients and AI systems. As part of our ongoing research, we have undertaken an evaluation rooted in parameters such as Empathy, Relevance, and Medical Accuracy, hoping to glean insights on the potential and limitations of our system in healthcare settings.

Methodology

Prompt Engineering

We utilized prompt engineering techniques such as zero-shot and few-shot learning (with synthetic patient message and real doctor responses), as well as a pipeline to generate responses to patient messages.

Zero-Shot Learning Prompt:

Imagine you are a doctor's healthcare provider. When responding to a patient's message, it's crucial to offer information grounded in the latest medical research, ensuring it aligns with established medical standards. Your response should be both clear and tailored to the patient's specific concerns, avoiding any unnecessary jargon. Always approach each message with empathy, understanding the emotions and concerns the patient may be experiencing. Lastly, maintain a comforting and reassuring tone, ensuring the patient feels heard and cared for.

Few-Shot Learning Prompt:

Your mission, as an advanced language learning model (LLM), is to emulate the intricate balance of clinical expertise, heartfelt empathy, and rigid confidentiality exercised by proficient medical professionals. This task involves creating a distinct yet equally comprehensive response to each simulated patient message, encapsulating the nuanced elements of a real-world medical consultation. Your responses should not only deliver medically sound advice but also provide emotional support, reflecting the human touch in healthcare services. All responses should adhere to confidentiality guidelines, ensuring that no personally identifiable information is used within the simulation.

Key Elements for Consideration:

"Clinical Accuracy": Every response crafted by you should align with the most recent, evidence-based medical guidelines and standards of practice. The information should be pertinent to the patient's inquiry, symptoms, or concerns. Furthermore, the content should be transparent, comprehensive, and easily understandable to ensure effective communication.

"Empathy and Compassion": The clinical component of healthcare is undeniably important, but so is the emotional and psychological well-being of the patient. As such, your responses should reflect the empathy, understanding, and compassion characteristic of a seasoned physician. Acknowledge the patient's feelings, validate their concerns, and offer reassurances as appropriate to help alleviate their anxieties or fears.

"Strict Confidentiality": Given the simulated nature of this task, you must absolutely abstain from incorporating any personally identifiable patient information in the generated responses. The same applies to hypothetical situations that could compromise the privacy or anonymity of any potential real-world individuals. Here are examples of doctor responses: <REDACTED>

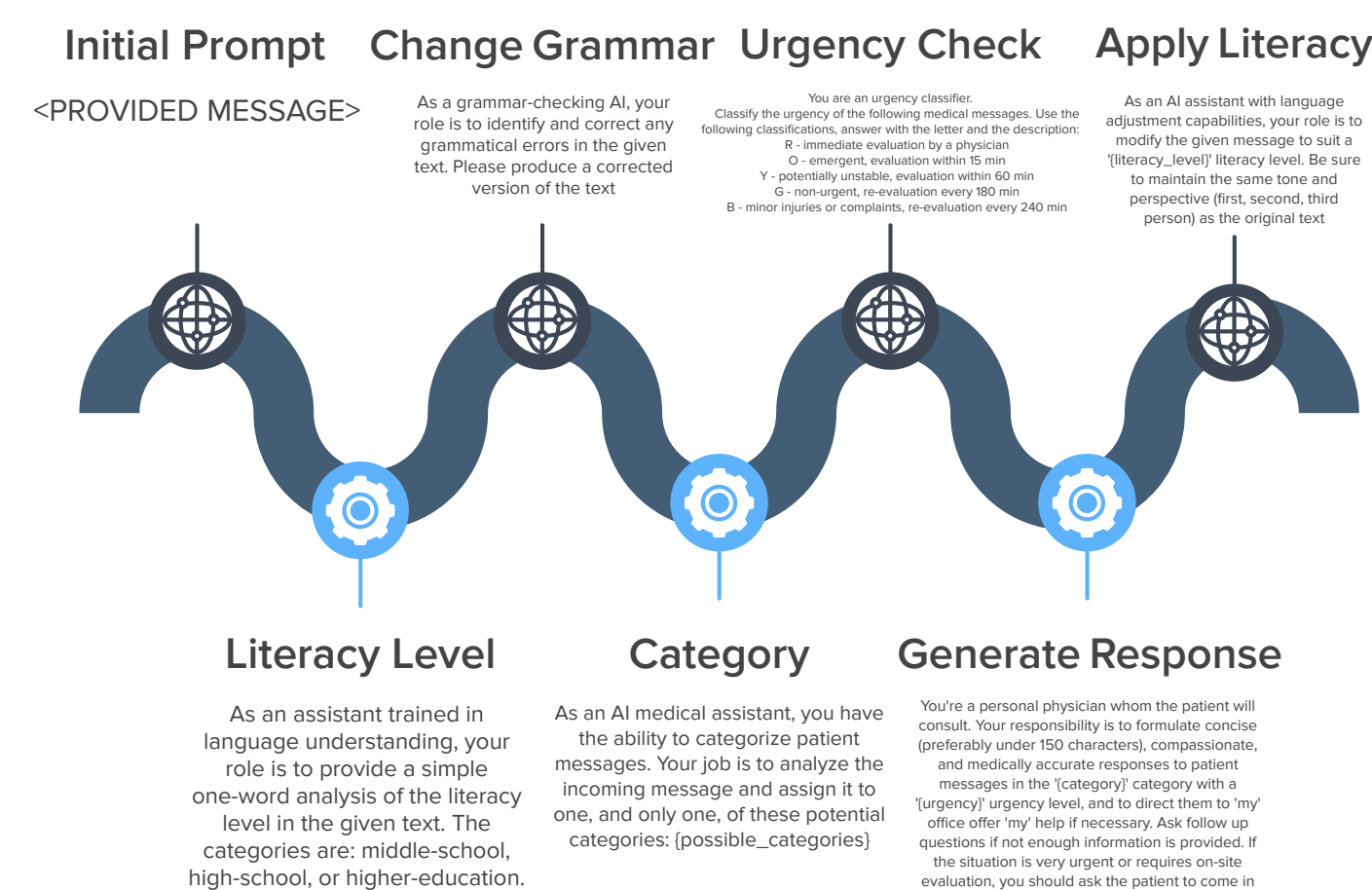


Fig 2. Diagram representation of the Katralex pipeline

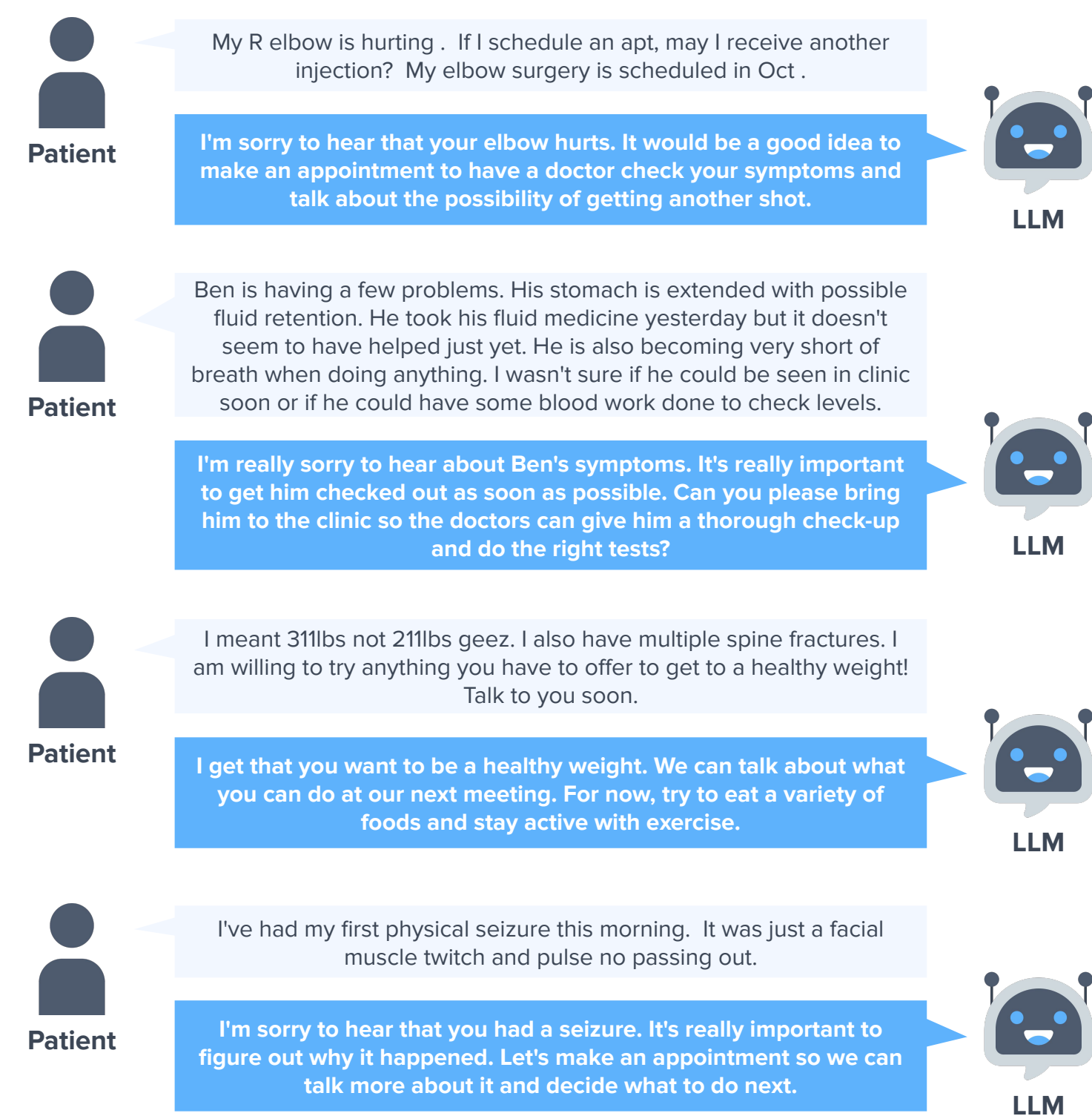
Evaluation

The evaluation of GPT-4 generated patient responses was conducted through a survey sent to 60 medical workers.

The survey aimed to gauge the quality and authenticity of the messages across several dimensions: **1) Empathy:** Medical workers rated the level of empathy perceived in the messages, reflecting the compassionate communication often required in medical settings; **2) Relevance:** Participants evaluated the alignment of the responses with the context and inquiries of the patient scenario, focusing on how closely the content matched patient needs; **3) Medical Accuracy:** The accuracy of the information in the generated messages was assessed, considering its alignment with established medical practices and guidelines; **4) Readability:** The medical workers rated the clarity, coherence, and simplicity of the language used; **5) If LLM-generated messages can be differentiated from human-written messages:** This provided insights into the authenticity and human-like quality of the generated responses, helping to assess how effectively GPT-4 could mimic human interaction in a medical context.

Results

As our survey is still under the process of obtaining IRB approval, we were unable to provide the results of the evaluation. However, below are some examples of messages generated by our system using the GPT-3 model since GPT-4 was not available at the time of response generation.



Conclusion

The development and implementation of the Katralex system serves as an exploration into the potential of conversational AI within medical settings, emphasizing the importance of mimicking human-like conversational patterns to enhance user experience. By integrating mechanisms such as the chain of thought process, category-specific processing, and iterative linguistic refinements, we endeavor to produce responses that are not only relevant but also contextually rich and human-like. The methodology presented underscores the intricate balancing act between retaining the system's vast knowledge and ensuring that its interactions remain relatable and comprehensible to diverse users. This balance, pivotal in enhancing user trust and satisfaction, echoes the broader challenge in the field of AI – harnessing immense computational power without alienating the very humans it's designed to serve.

The reported surge in physician burnout, attributed to administrative loads and patient communications, accentuates the need for systems like ours to offer tangible solutions. In the ever-evolving landscape of technology and, more specifically, conversational AI, our intent with Katralex is not just to mimic but to genuinely support and alleviate some pressing human-centric challenges.

While our system showcases numerous innovative features, we are cognizant of its inherent limitations such as the lack of offline support and privacy concerns. More than just recognizing these shortcomings, we see them as an opportunity to invite further refinement and exploration.

Acknowledgements

We extend our heartfelt gratitude to Dr. Kevin B. Johnson and everyone at the Kevin Johnson Lab for their invaluable insights and unwavering guidance throughout the course of this research. Special thanks to Dr. Bryan Steitz for providing the synthetic messages from Vanderbilt. Lastly, we acknowledge and appreciate the financial support from the Penn Undergraduate Research Mentorship Program (PURM), without which this research would not have been possible.

Illustrations used in this poster are from Flowicon on Flaticon.com.

References

- Tai-Seale, M., et al. (2023). Association of physician burnout with perceived EHR work stress and potentially actionable factors. *Journal of the American Medical Informatics Association*, ocad136. <https://doi.org/10.1093/jamia/ocad136>
- Johnson, K.B., Neuss, M.J., & Detmer, D.E. (2021). Electronic health records and clinician burnout: A story of three eras. *Journal of the American Medical Informatics Association*, 28(5), 967-973. <https://doi.org/10.1093/jamia/ocaa274>. PMID: 33367815; PMCID: PMC8068425.
- Brown, T., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33, 1877-1901.