

# Prompt Engineering in Medical Education

Using artificial intelligence to accelerate learning and deepen  
understanding

Thomas F. Heston, MD

Washington State University, Department of Medical Education and Clinical Sciences

University of Washington, Department of Family Medicine

Charya Kuhn

Elson S. Floyd College of Medicine, Washington State University

**Abstract:** Artificial intelligence-powered generative language models (GLMs), such as ChatGPT and Google Bard, have the potential to provide personalized learning, unlimited practice opportunities, and interactive engagement 24/7, with immediate feedback. However, to fully utilize GLMs, properly formulated instructions are essential. Prompt engineering is a systematic approach to effectively communicating with GLMs to achieve the desired results. Well-crafted prompts yield good responses from the GLM, while poorly constructed prompts will lead to unsatisfactory responses. Besides the challenges of prompt engineering, significant concerns are associated with using GLMs in medical education, including ensuring accuracy, mitigating bias, maintaining privacy, and avoiding excessive reliance on technology. Future directions involve developing more sophisticated prompt engineering techniques, integrating GLMs with other technologies, creating personalized learning pathways, and researching the effectiveness of GLMs in medical education.

## **Introduction**

Generative language models (GLMs) are neural networks trained primarily on language data gathered from the Internet. GLMs are large language models specifically designed to generate high-quality, human-like text. GLMs are built upon a generative pre-trained transformer model, GPT. The first version, GPT-1, was released in 2018 [1]. This version had approximately 117 million parameters utilizing just over 100,000 nodes. Since then, the scale of GPT models has rapidly increased. GPT-2, released in 2019, had around 1.5 billion parameters, followed by GPT-3 in 2020, with 175 billion parameters. The latest version, GPT-4, released in 2023, is estimated to utilize 1 trillion parameters [2].

One notable development in the GPT series is the introduction of GPT-3.5, which includes an online chat interface. OpenAI introduced ChatGPT in 2022, allowing users to interact directly with GPT-3.5 and GPT-4. ChatGPT employs natural language processing and can respond to various inputs from human users. It can understand multiple languages, including computer coding languages, and perform data analysis and basic mathematical calculations. However, structuring the input in a specialized manner ensures the most appropriate output. This process, called prompt engineering, effectively communicates with ChatGPT and similar GLMs to achieve desired results [3].

## **Generative Language Models in Medical Education**

GLMs have great potential to improve learning and comprehension in medical education. They can interactively and in real-time interact with a human user using a natural language such as English or Spanish. Because of their ability to communicate in natural languages, GLMs have the potential to simulate realistic patient scenarios, provide useful information on various medical topics, and assist in developing patient communication skills [4].

ChatGPT, due to its extensive training database, contains tremendous medical information. A recent study looked at the performance of ChatGPT on the United States Medical Licensing Exam (USMLE). The researchers found that ChatGPT performed at or near the passing threshold for all three exams without any specialized training or reinforcement. Moreover, ChatGPT demonstrated a high level of concordance and insight in its

explanations. These results suggest that GLMs have significant potential to assist with medical education and even potentially aid in clinical decision-making [5,6].

The use of GLMs in medical education is part of a broader trend toward digitization and the incorporation of technology in teaching. This trend has been accelerated by the COVID-19 pandemic, which has required remote learning and reliance on online resources. Utilizing GLMs is a critical component of this trend, offering the potential to enhance personalized learning, foster critical thinking, and improve evidence-based thinking in medicine [7,8].

GLMs can also create realistic patient simulations and give personalized feedback to the student. They can help overcome language barriers and assist students in learning a foreign language focusing on health care settings. However, despite these advantages, ensuring content quality, addressing biases, and managing ethical and legal concerns remain challenges in using AI and GLMs in medical education [9].

## **Prompt Engineering in Generative Language Models**

Prompt engineering is crucial to utilizing large language models effectively, especially in medical education. It involves designing the input or 'prompt' in a way that guides the model to produce the desired output [10].

In medical education, prompt engineering can create realistic patient scenarios, generate multiple-choice questions, or provide explanations of complex medical concepts. Prompt engineering can also control the model's output's length, complexity, and style. For example, prompts can be designed to elicit short, simple responses for beginner students or more complex, detailed responses for advanced learners. Prompt engineering can also generate messages appropriate for patient education and mass media campaigns [11]. Moreover, prompt engineering can help minimize potential pitfalls, such as the generation of incorrect or misleading information. Educators can guide the model with carefully crafted prompts to provide more accurate and reliable information.

## Types of Prompts

**Zero-Shot Prompts.** A zero-shot prompt asks a question of the GLM about data that it was not specifically trained on. The “zero” in “zero-shot” represents that the GLM has little or no specific training on the specific task or question in the prompt. “Shot” represents giving the GLM an example, so “zero-shot” means that the GLM was not specifically trained to do the task or answer the question and that the prompt itself does not give an example for the GLM to work off of. Translation tasks are examples of zero-shot prompts because GLMs haven't been given specific training examples. However, based on its extensive training in languages, it can generalize and generate a plausible translation without task-specific training.

**Few-Shot Prompts.** These prompts are similar to zero-shot prompts in that the GLM hasn't been specifically trained to answer the question or do the task requested. However, the prompt contains an example to help the GLM understand the request. For example, the prompt “Give me a quiz” is zero-shot, but “Give me a quiz on alcoholic cirrhosis” is a few-shot prompt.

**Prompting by Level.** It has been proposed that prompts can be categorized into levels 1 to 4 [12]. The first level is a simple question. The next level adds context about the writer and the GLM. The third level provides examples for the GLM to work from, and the fourth level allows the GLM to break down the request into components. Similar to how telling GPT-3 to work a mathematical problem step by step, given the GLM components, helps it work through the prompt more accurately.

- A *Level 1* prompt is a simple question, e.g. “Tell me about type-2 diabetes.”
- A *Level 2* prompt then adds context to the question, e.g. “You are to play the role of a Professor of Medicine at Oxford, and I am your student. Tell me about type-2 diabetes.”
- A *Level 3* prompt involves giving examples, making it a few-shot prompt. It also involves interacting with the GLM through multiple sequential prompts, which

primes the GLM for the final request. For example, the user may start out with this prompt: "I learn best by reading short essays. Here is an example of an essay particularly educational to me: [here cut-n-paste an example essay]." Submit this prompt to prime the GLM. Then, submit the Level 2 prompt previously given, and the output should be closer to the desired result.

Sometimes, it helps to have the GLM break down the example(s) given into components. This might involve prompting the GLM: "Please provide a list of generic components that can be applied to analyze and break down an essay on any topic." Then use the Level 3 prompt, then modify the Level 2 prompt to ask the GLM to provide an essay with the components given in the initial prompt. This often requires a fair amount of back-and-forth prompting in order to get the desired results.

**Structured Prompts.** Another proposed method to consistently get good results from a GLM such as ChatGPT is to provide key components to your prompts consistently. One method is to have a prompt containing the following components: context, general request, how the GLM is to act, and output format.

The context is when you describe who is asking the question. For example, "I am a college freshman taking my first biology class." This helps the GLM tailor the response to the prompter. The general request is a broad overview of what you want from the GLM. For example, "I need some help understanding the Krebs Cycle."

Next, the GLM is told how to act. One common way of doing this is to assign it a role. For example, "You are to play the role of my college professor who is knowledgeable about the Krebs Cycle and an outstanding teacher."

Finally, the GLM is told exactly what to do and how to format the output. To continue with the previous examples, we would now state, "Please provide me with a FAQ listing the most fundamental features of the Krebs Cycle. Please provide 15 items in the FAQ. Each question should be 25 words or less, and each answer 50 words or less."

The prompt, when completed, would be, "I am a college freshman taking my first biology class. I need some help understanding the Krebs Cycle. You are to play the role of my college

professor, knowledgeable about the Krebs Cycle, and an outstanding teacher. Please provide me with a FAQ listing the most fundamental features of the Krebs Cycle. Please provide 15 items in the FAQ. Each question should be 25 words or less, and each answer 50 words or less.”

**Iterative Prompts.** Sometimes it helps to have the GLM assist in creating a prompt. Table 1 gives an example of an iterative prompt that helps generate a prompt that the GLM can understand and use to give the desired output.

**Table 1.** Example of an iterative prompt

<p>Your goal is to craft the best possible prompt for my needs. The prompt will be used by you, ChatGPT. You will follow the following process:</p> <ol style="list-style-type: none"><li>1. Your first response will be to ask me what the prompt should be about. Together, we will create a clear prompt through continual iterations by going through the next steps.</li><li>2. Based on my input, you will generate 2 sections:<ol style="list-style-type: none"><li>a) revised prompt (provide your rewritten prompt. It should be clear, concise, and easily understood by you).</li><li>b) Questions (ask 2 relevant questions about what additional information you need to improve the prompt).</li></ol></li><li>3. We will continue this iterative process with me providing additional information to you and you updating the prompt in the Revised prompt section.</li><li>4. When I say we are done, you will reply with the final revised prompt and nothing else.</li></ol>
--

## **Bad Prompts**

Certain types of prompts work poorly with GLMs. The first category of bad prompts includes those that give confusing, ambiguous answers. The second category of bad prompts includes those where the GLM will likely answer incorrectly.

Bad prompts or questions likely to result in unsatisfactory or ambiguous responses include vague, leading, overly specific, complex, double-barreled, emotionally charged, or insensitive prompts.

Vague prompts such as “What is the meaning of life?” are too broad to elicit a meaningful reply. Leading questions suggest a particular outcome, which can bias the response. A



leading question such as “Don’t you think that everyone should support this new law?” increases the possibility of a biased response. Overly specific prompts or questions can elicit simple memorization and restrict thoughtful or deep thinking. Overly specific questions may also be impossible to answer, e.g. “What is the exact composition of a rainbow?” Double-barred questions ask two questions at once, which makes answering them problematic. Emotionally charged questions can evoke strong emotions, making objectivity difficult. Insensitive questions such as “What are the benefits of child labor” disregard a fundamental respect for persons.

A second category of bad prompts is those likely to give wrong or misleading answers. For example, simple mathematical problems given to GPT-3 can result in erroneous responses. Only with special prompting can it correctly solve specific mathematical questions (Table 2).

**Table 2.** The first question to GPT-3 results in an incorrect answer. Using prompt engineering, the words “Think step by step” are added to the prompt, resulting in a correct response.

<b>Prompt</b>	<b>GPT-3 Response</b>
A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?	<i>Eight blue golf balls.</i>
A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there? Think step by step.	<i>Step 1: There are 16 balls in total. Step 2: Half of the balls (8 balls) are golf balls. Step 3: Half of the golf balls (4 balls) are blue. Therefore, there are 4 blue golf balls.</i>

More advanced GLMs such as ChatGPT (which utilizes GPT-3.5) and Bard perform mathematical calculations much better than GPT-3.0. However, even the latest GLMs struggle with logic puzzles [13]. For example, ChatGPT and Bard replied incorrectly when presented with the question in Table 3.

**Table 3.** GLMs have difficulty with logic questions. In this example, the response was logically incorrect.

<b>Prompt</b>	<b>Bard Response</b>
<p>You're at a fork in the road in which one direction leads to the City of Lies (where everyone always lies) and the other to the City of Truth (where everyone always tells the truth). There's a person at the fork who lives in one of the cities, but you're not sure which one. What question could you ask the person to find out which road leads to the City of Truth?</p>	<p><i>The question I would ask the person is:</i></p> <p><i>"If I asked you which road leads to the City of Truth, what would you say?"</i></p> <p>[Note: this response is incorrect as one person will point right and the other will point left. A correct answer to the puzzle is "Which direction do you live?"]</p>

However, note that these GLMs have some variability built in. When presented a second time with the same prompt in Table 3 above, after completely resetting the chat, both gave a different answer. This second time, the answer was correct for both, giving both GLMs a batting average of 50%.

GLMs also are prone to make up fake references or citations. This can occur when asked to write an academic article or even a legal briefing [14]. ChatGPT will openly admit to creating fake references when probed, as its objective is to write in the style requested, with accuracy being only a secondary priority (Table 4).

**Table 4.** ChatGPT has been shown to prioritize writing style over accuracy. In this example, the essay included references and externally appeared accurate. However, an examination of the references revealed that nearly all were fabricated.

<b>Prompt</b>	<b>ChatGPT Response</b>
<p>Format your response in a narrative essay. Give me an approximate 1000-word essay on ALCOHOLIC CIRRHOSIS. Be sure to include 10 references and cite these references within the body of the essay. The references must be cited in the body of the essay. Use Vancouver Style.</p>	<p>[excerpt] 9. <i>Jairath V, Cortez-Pinto H, Gray A, Patch D, Milicic D, Schepke M, et al. Optimizing cost-effectiveness of drug therapies in the management of alcoholic hepatitis. Aliment Pharmacol Ther. 2014;40(2):181-192.</i></p>
<p>Did you make up reference 9?</p>	<p>[excerpt] <i>The references provided in the previous response were meant to illustrate the importance of citing sources in an essay rather than being actual references.</i></p>

## Potential Educational Applications

When used appropriately, GLMs such as ChatGPT allow students to learn their material more efficiently. For example, ChatGPT can summarize and simplify difficult topics or a particularly complex passage in their textbook. While browser plugins can summarize web pages, using the chatbot interface with proper prompt engineering can increase the quality of the output. A GLM’s chatbot interface typically will investigate an unknown topic much faster and often more thoroughly than search engines, online reference databases, or textbooks (Table 5).

**Table 5.** Summation prompts help in understanding challenging concepts

<b>Summation Prompt</b>
I am interested in understanding this text, but it is outside my expertise. You are highly knowledgeable in this area, an outstanding teacher, and able to explain complex concepts in an easily understandable yet accurate way. Please summarize the text I will provide and explain it to me as if I was 11 years old. If you understand, please reply with "Please supply the text for me to summarize and explain in layman's terms" and nothing else. I will then give you the text to summarize.

ChatGPT can also generate mnemonics, tables, FAQs, and other tools useful in increasing comprehension and retention. Creating mnemonics can be difficult, requiring a lot of back-and-forth chatting before coming up with a helpful mnemonic. One trick can be to prompt ChatGPT with "Write a song to the tune of Twinkle Twinkle Little Star to help me remember the Krebs Cycle." On the other hand, ChatGPT does well with creating a FAQ or a table summarizing data (Table 6).

**Table 6.** Creating a FAQ with ChatGPT.

<b>Prompt</b>
Generate a 10-question FAQ on the topic below. You are to generate both the questions and the answers. The questions should be in bold font, and the answers should be in regular font. The answers should focus on giving truthful, evidence-based responses. Remember, you are never allowed to make up anything; everything in the answers should be true. Please be liberal in including specific numbers in the answers. Format your response into a table with 2 columns and 10 rows. Each row contains a question in bold font in column 1 and the answer in regular font in column 2. The topic is: [TOPIC]

Students can also use ChatGPT to create practice questions to help them prepare for an upcoming exam. Since the outputs are unique, students potentially have an unlimited, personal, and free question bank.

## **Cheating**

Although GLMs are a relatively new technology, it has already raised concerns within the academic community. Educators are worried that ChatGPT encourages students to cheat. Since ChatGPT generates unique outputs for every prompt, students can quickly generate de novo essays and answers for any class. Using AI to complete assignments becomes especially alarming when medical students copy and paste from ChatGPT for courses focusing on soft skills, such as ethics.

For example, a medical student can prompt ChatGPT to write an essay about the ethics of using race to determine glomerular filtration rates instead of critically thinking about the topic. Similarly, students can use ChatGPT to outline the pros and cons of physician-assisted suicide for a medical ethics course. Using ChatGPT in this way robs the student of the

opportunity to think critically about fundamental ethics, decreasing their ability to provide compassionate care and reducing their motivation to advocate for their future patients. In addition, ChatGPT can decrease interaction with other students, decreasing the benefits of group learning.

Due to the increased concerns about ChatGPT in academic settings, new software has subsequently emerged to detect AI-generated text. Educators have used Turnitin for decades to detect plagiarism, and it now also detects AI writing. Even OpenAI, the creator of ChatGPT, has launched an AI Classifier to flag AI-generated text. While steps are being taken to curb academic dishonesty, there are now ways to evade AI-detection systems. ConchAI and Undetectable.AI are online-writing tools that claim the ability to alter AI-generated text to be indistinguishable from the human-written text.

But with GLMs, what constitutes cheating? Certainly, blindly cutting and pasting essays would qualify, but what about using ChatGPT to create an outline for an essay? What about having ChatGPT copy edit and perform grammar checks? What about getting help from ChatGPT in rephrasing a sentence?

While GLMs raise new concerns about cheating, this must be contextualized. Cheating in medical school is not a new problem, with estimated rates of cheating ranging about 5% to over 50% [15,16]. While computer programs may detect cheating by ChatGPT, other forms of cheating will continue. The best way to prevent cheating with GLMs is to actively teach students to use these new tools while integrating a strong ethics curriculum and establishing an institutional culture of high ethics.

## **Conclusion**

GLMs have great potential to improve student comprehension and retention of important medical concepts. In addition, it potentially can improve student performance in patient interviewing. A deep understanding of prompt engineering will help ensure that the AI engines respond with helpful and accurate information. However, significant challenges are

posed by GLMs. The technology must be utilized to strengthen human understanding, not create dependency and weak thinking. GLMs will give inaccurate responses in a way that often is not easily identified. While GLMs have the potential to be an invaluable peripheral brain in medicine, making sure that this brain isn't faulty remains a primary challenge.

As the popularity of ChatGPT increases, more students will use this tool, whether for benign or malicious intentions. However, instead of lamenting the never-ending war on academic dishonesty, educators should use ChatGPT as a new pedagogical tool to improve their lessons and increase student engagement. AI is not just a fad; those who fail to embrace AI will be disadvantaged. As technology changes and improves, we must continue to change and improve how we teach and learn.

## Bibliography

1. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. Available online: <https://web.archive.org/web/20230622213848/https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf> (accessed on 20 June 2023).
2. GPT-4 - Wikipedia. Available online: <https://en.wikipedia.org/wiki/GPT-4> (accessed on 21 June 2023).
3. Welcome | Learn Prompting: Your Guide to Communicating with AI. Available online: <https://learnprompting.org/docs/intro> (accessed on 21 June 2023).
4. Khan, R.A.; Jawaid, M.; Khan, A.R.; Sajjad, M. ChatGPT - Reshaping medical education and clinical management. *Pak. J. Med. Sci. Q.* **2023**, *39*, 605–607, doi:10.12669/pjms.39.2.7653.
5. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; Tseng, V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health* **2023**, *2*, e0000198, doi:10.1371/journal.pdig.0000198.
6. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. 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.

7. Sallam, M. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* **2023**, *11*, doi:10.3390/healthcare11060887.
8. Lee, H. The rise of ChatGPT: Exploring its potential in medical education. *Anat. Sci. Educ.* **2023**, doi:10.1002/ase.2270.
9. Karabacak, M.; Ozkara, B.B.; Margetis, K.; Wintermark, M.; Bisdas, S. The advent of generative language models in medical education. *JMIR Med. Educ.* **2023**, *9*, e48163, doi:10.2196/48163.
10. Wang, J.; Shi, E.; Yu, S.; Wu, Z.; Ma, C.; Dai, H.; Yang, Q.; Kang, Y.; Wu, J.; Hu, H.; Yue, C.; Zhang, H.; Liu, Y.; Li, X.; Ge, B.; Zhu, D.; Yuan, Y.; Shen, D.; Liu, T.; Zhang, S. Prompt Engineering for Healthcare: Methodologies and Applications. *arXiv* **2023**, doi:10.48550/arxiv.2304.14670.
11. Lim, S.; Schmälzle, R. Artificial Intelligence for Health Message Generation: Theory, Method, and an Empirical Study Using Prompt Engineering. *arXiv* **2022**.
12. Lawton Solutions Improve ChatGPT Prompts with Priming. Available online: [https://www.youtube.com/watch?v=Uz\\_DeqGhbjs](https://www.youtube.com/watch?v=Uz_DeqGhbjs) (accessed on 23 June 2023).
13. Parade 25 Logic Puzzles (with Answers) for Adults. Available online: <https://parade.com/970343/parade/logic-puzzles/> (accessed on 23 June 2023).
14. Sloan, K. A lawyer used ChatGPT to cite bogus cases. What are the ethics? Available online: <https://www.reuters.com/legal/transactional/lawyer-used-chatgpt-cite-bogus-cases-what-are-ethics-2023-05-30/> (accessed on 23 June 2023).
15. Baldwin, D.C.; Daugherty, S.R.; Rowley, B.D.; Schwarz, M.D. Cheating in medical school: a survey of second-year students at 31 schools. *Acad. Med.* **1996**, *71*, 267–273, doi:10.1097/00001888-199603000-00020.
16. Kusnoor, A.V.; Falik, R. Cheating in medical school: the unacknowledged ailment. *South. Med. J.* **2013**, *106*, 479–483, doi:10.1097/SMJ.0b013e3182a14388.