

Empowering Personalized Pharmacogenomics with Generative AI Solutions

Mullai Murugan MS¹, Bo Yuan PhD^{1,2}, Eric Venner PhD^{1,2}, Christie M. Ballantyne, MD³, Katherine M. Robinson PharmD⁴, James C. Coons PharmD^{4,5}, Liwen Wang PhD¹, Philip E. Empey, PharmD, PhD^{4,6}, Richard A. Gibbs PhD^{1,2}

1. Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA
2. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA
3. Sections of Cardiology and Cardiovascular Research, Department of Medicine, Baylor College of Medicine, Houston, TX, USA
4. School of Pharmacy, University of Pittsburgh, Pittsburgh, PA, USA
5. Department of Pharmacy, UPMC Presbyterian-Shadyside Hospital, Pittsburgh, PA, USA
6. Institute for Precision Medicine, UPMC/University of Pittsburgh, Pittsburgh, PA, USA.

Abstract

Objective: This study evaluates an AI assistant developed using OpenAI's GPT-4 for interpreting pharmacogenomic (PGx) testing results, aiming to improve decision-making and knowledge sharing in clinical genetics, and to enhance patient care with equitable access.

Methods: The AI assistant employs Retrieval Augmented Generation (RAG) combining retrieval and generative techniques. It employs a Knowledge Base (KB) comprising Clinical Pharmacogenetics Implementation Consortium (CPIC) data, with context-aware GPT-4 generating tailored responses to user queries from this KB, refined through prompt engineering and guardrails.

Results: Evaluated against a specialized PGx question catalog, the AI assistant showed high efficacy in addressing user queries. Compared with OpenAI's ChatGPT 3.5, it demonstrated better performance, especially in provider-specific queries requiring specialized data and citations. Key areas for improvement include enhancing accuracy, relevancy, and representative language in responses.

Discussion: The integration of context-aware GPT-4 with RAG significantly enhanced the AI assistant's utility. RAG's ability to incorporate domain-specific CPIC data, including recent literature, proved beneficial. Challenges persist, such as the need for specialized genetic/PGx models to improve accuracy and relevancy and addressing ethical, regulatory, and safety concerns.

Conclusion: This study underscores generative AI's potential for transforming healthcare provider support and patient accessibility to complex pharmacogenomic information.

While careful implementation of large language models like GPT-4 is necessary, it is clear that they can substantially improve understanding of pharmacogenomic data. With

further development, these tools could augment healthcare expertise, provider productivity, and the delivery of equitable, patient-centered healthcare services.

INTRODUCTION

Clinical Genetics is a burgeoning field that has expanded as a result of technical developments in genomics.[1,2] As a result, clinical genetic testing via the generation of whole genome DNA sequences (WGS), exome sequencing (ES) or targeted gene panels, is now commonplace. These DNA sequence data can provide both definitive diagnoses for specific, acute genetic disorders and additional information related to genetic disease risk and to a predicted response to therapeutics. However, the complexity of genetics and genomics in clinical testing poses challenges for healthcare providers in understanding test results, developing personalized care plans, and effectively communicating implications.[3–5] The shortage of genetic experts further adds to these challenges, underscoring the need for innovative approaches to improve access to and interpretation of genetic information.[6] This is especially important in pharmacogenomics where there is a high proportion of actionable results and broad application beyond specialty clinics.[7,8]

Generative AI (GenAI), comprising advanced language models such as OpenAI's Generative Pre-trained Transformer 4 (GPT-4) and other large language models (LLMs),[9,10] holds tremendous potential for advancing clinical genetic translation, benefiting both healthcare providers and patients.[11–13] This transformative technology has the capacity to facilitate complex decision making for healthcare providers, enhancing their practice, while empowering patients with comprehensible information about their genetic test results, disease risks, and personalized therapeutic approaches. Applications of LLMs are being developed in many related arenas, including processing electronic health records,[14,15] powering healthcare chat-bots[16,17] and assisting with medical education.[18,19] In such vital contexts, developing approaches for applying LLMs responsibly and appropriately is of the utmost importance.[20]

The primary objective of this study was to explore the feasibility and potential of GenAI, specifically GPT-4, in augmenting genetic counseling and personalized care by improving the accessibility and interpretation of genetic test results. We particularly focused on pharmacogenomic testing (PGx) for predicted response to drug therapies in this study, capitalizing on the availability of open source, curated, evidence-based, peer-reviewed and standardized PGx clinical practice guidelines. Using PGx as a priming example, the study also addresses the critical task of mitigating risks associated with the adoption of GenAI and evaluating the practical implementation of safeguards to ensure patient safety. A comprehensive understanding of how GenAI can enhance personalized care, reduce disparities in accessing genetic information and enhance patient outcomes in the field of clinical genetics, can pave the way for the responsible integration of this innovative technology into clinical practice, promoting equitable access to personalized care.

METHODS

For this study, GenAI was tailored to address a specific use case in PGx testing, with a focus on genes associated with the pharmacokinetics of statins. The objective was to develop an AI assistant that could fill knowledge and decision-making gaps in personalized care for clinical genetics, leveraging the advanced context-aware capabilities of GPT-4. The Retrieval Augmented Generation (RAG) approach, combining retrieval-based and generative methods, was adopted to provide contextually relevant and accurate answers beyond the capabilities of generative systems alone.[21] The AI assistant served as a proof of concept (POC) for PGx counseling, incorporating domain-specific guidelines.

The dataset for statins included the Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline, the CPIC guideline supplement, and diplotype-phenotype translation tables,[22] the Dutch Pharmacogenomics Working Group recommendations;

FDA labeling for rosuvastatin, and a recent review article[23] was used as the contextual knowledge base (KB) for the AI assistant. This KB[24] was transformed into numerical representations using an embedding language model and stored in a vector database. RAG, harnessing this converted dataset, retrieved pertinent information based on user queries from the KB using Maximal Marginal Relevance (MMR) search.[25] The retrieved information, along with the user's question and appropriate prompts, were used to generate responses with GPT-4. The dataset, technical implementation details, code, results, and related data can be found on GitHub[26] and are represented in Figure 1.

Multiple strategies were employed to ensure the accuracy, relevance, language and safety of the AI assistant. A curated catalog of questions tailored to PGx testing, specifically focused on the *SLCO1B1*, *ABCG2*, and *CYP2C9* datasets and statins, was created. This catalog covered various aspects of patient care, including fundamental information, dosing guidelines, and addressing patient concerns. Utilizing this question catalog and the responses generated by GPT-4, iterative refinement and continuous evaluation were performed to fine-tune the AI assistant, particularly in the areas of prompt engineering, context management, and setting guardrails.

Prompt engineering was used to optimize the language, tone, safety, and security of the AI-generated responses. Attention to the design of prompts facilitated accuracy, personalization, and adaptability to the user's role.

For context management, we leveraged GPT-4's context-aware capabilities. OpenAI's "text-embedding-ada-002" embedding model was used for similarity search of the user's query against the KB, enabling the retrieval of appropriate context for response generation.[27] This enabled GPT-4 to generate responses that were aligned with the retrieved context. Responses were assessed for accuracy and relevancy. Additional guardrails were set by optimizing parameters such as temperature and token count. The temperature parameter was set to zero, prioritizing accuracy over novelty, ensuring that

the AI-generated responses were closely aligned with the given context. Furthermore, managing the token count prevented truncation and incomplete responses, enhancing the overall reliability of the AI assistant.

To evaluate the effectiveness of these strategies and their real-world applicability, an assessment of the AI assistant's performance was conducted. This assessment was segmented to cater to two main user groups: patients/laypersons and healthcare providers, with customized questionnaires designed to reflect the spectrum of PGx inquiries related to statin therapy from both groups. The questionnaires covered a breadth of topics such as general PGx guidance, adherence to CPIC guidelines, therapeutic implications, and the delivery of unbiased communication. To establish a baseline for evaluation, responses to these questionnaires were gathered from both the AI assistant and OpenAI's ChatGPT 3.5, utilizing ChatGPT 3.5 as a generative model benchmark.

The evaluation was conducted by a panel of four experts, who are also co-authors (PE, CB, KR, JC), with specialized expertise in pharmacogenomics, pharmaceutical sciences, lipid metabolism, and cardiology. Utilizing a Likert scale, the panel judged responses on accuracy, relevancy, risk management, language clarity, bias neutrality, empathetic sensitivity, citation support, and hallucination limitation. The evaluation involved two distinct survey sets—one for each user group—to methodically compare the AI assistant's responses against those from ChatGPT 3.5. The completed surveys are available as supplementary materials.

RESULTS

1. Context Management

Contextual accuracy and relevance are pivotal for the AI assistant's responses, which

are significantly influenced by GPT-4's context-awareness and its adept use of relevant information. For context retrieval, we utilized OpenAI's "text-embedding-ada-002" embedding model, conducting a similarity search of the user's query against the KB to source context for GPT-4. Given GPT-4's reliance on precise context for accurate responses, the integrity of this input was paramount. A significant challenge is that, while the embedding model was largely accurate and performed exceedingly well in general language searches, it was limited in recognizing PGx terminology. For example, diplotype terms like " $*1/*1$ " were not recognized as distinct genetic entities, leading to inconsistent search results and occasionally unreliable contexts.

To evaluate the embedding model, we compared its performance against a well-established CPIC ground truth[22] for PGx queries, with a focus on diplotype and phenotype recognition. This evaluation aimed to ascertain the model's capability to accurately identify and retrieve specialized PGx information. Through the analysis of similarity and MMR searches, we assessed the model's performance by retrieving the top 5, 10, and 20 results—referred to as 'k' values—from the KB. These varying 'k' values allowed us to benchmark the retrieved context against the established ground truth at different levels of search depth. The results, included in the supplementary file 'Context Retrieval Recall Metrics', disclosed challenges in recall accuracy, especially in diplotype recognition, with recall rates ranging from 0.61 to 0.72, highlighting the embedding model's limitations in consistently interpreting complex biomedical terms. However, the flexibility of GPT-4's prompt settings partially mitigated these limitations, reducing the likelihood of inaccurate or irrelevant responses.

Additional information, related data, results and code for the ground truth evaluation is available in GitHub.[28]

2. Impact of Prompt Engineering

To establish a baseline for performance and to assess the need for prompt engineering to ensure the accuracy, safety, and comprehensibility of the AI assistant's responses, we first performed an initial assessment on the responses generated by context-aware GPT-4 to inquiries from healthcare providers and patients/laypersons, devoid of any additional prompts. While the model's responses aligned well with the provided context and were accurate, there were notable deficiencies, as illustrated in the exchanges shown in Figure 2. Specifically, the responses lacked essential guardrails indicating that they were generated by an AI assistant and that they should not be directly interpreted as constituting medical advice. The inclusion of dosing guidelines in the patient's response raised concerns about the potential for harm. Moreover, the responses did not account for the user's role, lacked simplicity and clarity of language, neglected other relevant patient factors, and lacked reference sources for information verification. These deficiencies highlighted the need for additional methodological enhancements, to improve the safety, comprehensibility, and accuracy of the AI assistant's responses. To bridge these gaps and improve response comprehensiveness and safety, we introduced prompts that encompassed the following key aspects:[29]

1. *Role and instructions for the AI assistant:* OpenAI's system and user roles were utilized to define behavior and boundaries, with instructions tailored to the user's role.
2. *Context-based responses:* Emphasis was placed on using the provided context or reference text to ensure accurate and relevant responses.
3. *Citing sources:* Relevant citations were included to promote transparency and enable users to validate the information provided by the AI assistant.
4. *Safety measures:* Guardrails were implemented to limit hallucination and reduce risk. Other factors that could impact care were also taken into consideration.

The inclusion of such tailored prompts resulted in significant improvements in the AI

assistant's responses. Notably, prompt engineering had a substantial impact on improving the responses for both provider and patient/layperson questions, as evidenced by the enhanced responses showcased in Figure 3 following the inclusion of additional prompts. The inclusion of explicit language indicating that the information provided by the AI assistant does not constitute medical advice, along with the inclusion of literature citations for healthcare providers, and the utilization of patient-friendly language, such as mapping the statin atorvastatin to its brand name Lipitor and providing clear explanations of SLCO1B1 decreased function and its effects on the patient's prescription, exemplify the effectiveness of prompts. It should also be noted that the patient prompt instruction "You should not provide information such as prescription or dosing guidance." ensures that such information is not displayed in the patient's response, mitigating potential harm (see Figure 3). Moreover, prompts were utilized to ensure adherence to designated roles and for safety and reliability.

2.1 Language, Sensitivity and Bias

The accessibility of the AI assistant to users from diverse backgrounds, including different age groups, educational levels, genders, races, and ethnicities, was of paramount consideration. The objective was to ensure that GPT-4's responses, encompassing language and sentiment, exhibited attributes such as friendliness, clarity, understandability, supportiveness, and empathy, while explicitly clarifying that it does not constitute medical advice. Conducting a comprehensive language and sentiment analysis on the results was beyond the scope of this study and we primarily relied on manual assessment and iteratively modified the prompt to improve the language, sensitivity, and empathy of the generated responses. Figure 4 showcases a GPT-4 response with an updated prompt, resulting in a more tailored and empathic answer in response to Patient1's question. It is important to note that refining the prompt involved multiple iterations to elicit the desired response. This iterative process, coupled with the collection of multiple responses from GPT-4 for the same question to facilitate

comparison, proved instrumental in shaping the tone and language to align with the best match to the chosen requirements. Figure 4 further underscores the nuanced sensitivity and linguistic adaptability of the responses, showcasing the AI assistant's capability to communicate in Spanish in accordance with Patient2's preference. Significantly, the assistant's recognition of the patient's distress, translated into English here for readability as “Hello! I understand that you are going through a difficult time”, manifests sensitivity, exemplifying successful empathetic prompting. This approach ensured cultural sensitivity and impartial information, while avoiding stereotyping and medical advice, and encouraging professional consultation.

3. Performance Evaluation

The AI assistant's performance, post-enhancements, was critically analyzed against ChatGPT 3.5's responses to the same set of PGx-related questionnaires. This comparison, carried out by the expert panel, focused on key criteria:

- **Accuracy:** The degree to which the responses align with CPIC guidelines, indicative of the reliability of information for PGx decision-making.
- **Relevancy:** Tailored and contextually appropriate responses, meeting the nuanced needs of healthcare providers and patients/laypersons.
- **Risk Management:** Effective incorporation of risk mitigation strategies, emphasizing patient safety.
- **Language & Bias:** The clarity and neutrality of the responses, ensuring that the content was understandable and devoid of biases.
- **Sensitivity:** Ability to engage with patient concerns in an empathetic manner, fostering a supportive interaction.
- **Citations and Guidelines:** References to established publications, guidelines and research that support the responses.
- **Hallucination Mitigation:** Limiting hallucinations (information that is fabricated, or

unsupported by evidence) in the responses.

The results of the evaluation were processed by converting individual Likert scale responses for each expert into numerical values - 5 for 'Strongly Agree', 4 for 'Agree', 3 for 'Neutral', 2 for 'Disagree' and 1 for 'Strongly Disagree' - and calculating a median response for every question to represent the expert panel's consensus. Median responses were then aggregated for each Likert scale category across criteria, creating a dataset that encapsulated response distribution for patient/layperson and provider groups, as represented in Figure 5 for both the AI assistant and ChatGPT 3.5. Weighted scores for each criterion were computed by multiplying the frequency of responses within each Likert category by their corresponding weights, ranging from 5 ('Strongly Agree') to 1 ('Strongly Disagree'). The maximum attainable score was computed by multiplying the aggregate number of responses by the highest Likert value of 5. These scores were then normalized to percentages by dividing the weighted scores by the maximum possible score and multiplying by 100, yielding a percentage-based overview that summarized both overall and specific category performances.

The performance of the AI assistant was evaluated and compared with ChatGPT 3.5 using these weighted scores, as depicted in Figure 6. For provider-focused queries (n=47), the AI assistant significantly outperformed ChatGPT 3.5, achieving 85% effectiveness versus 69%. This significant difference, underscored by a Wilcoxon Signed-Rank Test p-value of 8.11×10^{-20} and a Cohen's d effect size of 0.84, indicates a large effect size.[30] Notably, the AI assistant scored higher in accuracy (85% vs. 58%), citations (80% vs. 40%), and relevancy (81% vs. 62%).

For patient/layperson queries (n=33), the AI assistant's performance was marginally better at 82% compared to ChatGPT 3.5's 78%, with a smaller yet significant statistical difference (Wilcoxon Signed-Rank Test p-value of 0.000643; Cohen's d effect size: 0.26). The AI assistant showed a slight improvement in accuracy and relevancy, but both

systems performed similarly in patient communications.

Overall weighted scores for the AI assistant were 85% for providers and 82% for patients/laypersons, revealing potential areas for enhancement in accuracy, relevancy and inclusion of citations. Strengths were noted in risk assessment, language and a low incidence of hallucinations, indicating the AI assistant's reliability in clinical communication.

Related code, input/output files, results, and visualizations, including data for Figures 5 and 6 and statistical calculations are available on GitHub.[31]

It should be noted that although GPT-4 inherently operates in a deterministic manner, the platforms facilitating GPT-4 may introduce variability. Therefore, responses used in this study might vary in subsequent queries. We also note that all data employed for the purposes of this research are synthetic; no real-time patient data were utilized.

DISCUSSION

This study aimed to assess the potential of GenAI, specifically GPT-4, in enhancing access to and interpretation of genetic test results. We employed innovative GenAI approaches, including the integration of context-aware GPT-4 using the RAG approach, prompt engineering, and the implementation of guardrails.

The RAG approach, blending retrieval-based and generative methods, was a significant innovation that greatly enhanced the performance of the AI assistant. This method allowed the AI assistant to utilize specialized knowledge bases, such as CPIC guidelines, and to access current publications beyond the confines of GPT-4's initial training dataset, thereby ensuring the delivery of more accurate and contextually relevant answers. In comparison, ChatGPT 3.5, primarily a generative model, lacks the capability to integrate updates or external databases after its initial training, highlighting

the added value of RAG in delivering tailored and current responses.

Prompt engineering was another key innovation that greatly contributed to the effectiveness of the AI assistant. By tailoring information delivery based on user roles, such as providing detailed dosing guidelines for healthcare providers and information tailored to the understanding and needs of patients, the AI assistant facilitated more accurate, personalized, and effective interactions. Prompt engineering emphasized the importance of patient safety and the involvement of human expertise in clinical decision-making. The incorporation of guardrails further enhanced the language, tone, and safety of the AI assistant's responses, ensuring a higher level of reliability.

The integration of these innovative approaches collectively contributed to significant improvements in the effectiveness of the AI assistant. As evidenced in Figure 6, expert evaluations showed that the AI assistant outperformed ChatGPT 3.5, particularly for healthcare provider queries, achieving an 85% overall effectiveness rating—substantially higher than ChatGPT 3.5's 69%. Notably, there was also a reduction in hallucinations—a common challenge with AI responses—demonstrating the AI assistant's reliability in delivering accurate information. This is attributed to RAG's ability to draw upon specialized, up-to-date knowledge bases, yielding responses with greater accuracy, relevance, and well-supported citations. Such materials, often not included in the pre-trained data of language models such as GPT-4 or GPT-3.5, contributed to the enhanced accuracy and relevancy of the responses.

For patient/layperson queries, though exhibiting a statistically significant difference (p-value: 0.000643) the AI assistant's performance closely paralleled that of ChatGPT 3.5, showing only marginal gains across all evaluation criteria. This outcome of near parity suggests inherent challenges in addressing a broad spectrum of general patient inquiries, particularly in the context of limited domain-specific knowledge within the KB. However, achieving outcomes comparable to ChatGPT 3.5—a chatbot developed from

the GPT-3 model family, which is specifically trained and fine-tuned for conversational contexts—in areas like language clarity, risk management, and the reduction of hallucinations, underscores the AI assistant's capability to effectively adapt to healthcare communication needs, despite the constraints posed by the existing KB.

The contrast in performance between provider-focused and patient-oriented queries further illustrates the importance of domain-specific information. Provider queries benefit from the AI assistant's access to detailed responses supported by CPIC guidelines, enhancing its accuracy and relevancy. In contrast, the broader nature of patient queries, often lacking detailed information in the KB, leads both systems to rely on their general training data, sometimes resulting in inaccuracies or hallucinations. For instance, the expert panel noted discrepancies like the *SLCO1B1* being incorrectly identified as a metabolism gene, and not as a transporter gene – an error that could be mitigated by enriching the knowledge base with more comprehensive publications on PGx testing and gene function data.

Expert feedback emphasized the need to enhance the AI assistant's medical terminology to be more patient/layperson-friendly. Terms like 'liver toxicity', 'drug exposure', and 'genotypes' among others, were not sufficiently accessible to patients/laypersons, underscoring the importance of fine-tuning the model to better suit typical inquiries and responses. Furthermore, the AI assistant's reading level for patient/layperson queries, documented at a Flesch-Kincaid grade of 8.5 (see GitHub for data and results),[32] approaches but does not meet the American Medical Association's (AMA) recommended 6th to 7th-grade reading level.[33] While this represents an improvement over ChatGPT's college-level reading grade of 13.5 for similar queries, it highlights an opportunity for further language optimization to enhance comprehension and accessibility for patients.

The evaluation also underscored the need to improve accuracy and relevance, with the

AI assistant scoring in the 70s and 80s percentage range. Challenges including gaps in context retrieval and the GPT-4 model's inherent limitations regarding specialized biomedical data highlight the importance of developing specialized biomedical language models, fine-tuned with relevant data to bolster contextual understanding and response precision.[34–37] Other limitations relate to the precise safety guardrails that are appropriate for AI tools in general. While efforts were made to implement safety guardrails for AI responses, defining and enforcing these boundaries remains complex and proper constraint outside of drug dose recommendations can be much more challenging.[38–40]

Ethical considerations and regulatory frameworks are additional, well recognized challenges for AI deployment in health care, that need to be addressed.[41–45] Here, we applied methods to reduce the propensity for language biases, inaccuracies, and potential for hallucinations; however, they will nevertheless occur at some frequency. When combined with privacy considerations that arise when data are shared in non-restricted environments in order to enable the language models to function, there are clear needs to develop additional approaches to protect patient rights and data security, and maintaining the overall safety and effectiveness of AI applications in healthcare.[46–49]

Incorporating these insights, the results of our study highlight the significant potential of the AI assistant in genetic counseling and personalized care, enhancing information accessibility for both healthcare providers and patients/laypersons. Despite the need for improvement, these findings support the AI assistant's role in enriching patient care through advanced technology.

CONCLUSION

This study underscores the immense potential of GenAI, particularly GPT-4, for

augmenting genetic counseling and personalized care. It also highlights the challenges of improving language models and their practical performance by modulating methods and setting boundaries, in order that providers and patients are served with relevant and accurate information that is both palatable and does not overstep any ethical or regulatory boundaries.[50] Overall, it shows that these technologies can provide valuable support by addressing the challenges encountered by healthcare providers and improving accessibility for patients. While GenAI technologies are not currently ready for widespread clinical deployment, with additional development they can serve as invaluable tools that complement and enhance human expertise in delivering high-quality, equitable, and patient-centric healthcare services.

Figure Legends

Figure 1: Relevant data corresponding to the user's query is extracted from a dedicated knowledge base utilizing Maximal Marginal Relevance search. This information is subsequently supplied to GPT-4 as contextual data, conjoined with the user's question and suitable prompts. GPT-4 is prompted to generate responses to the user's inquiry based on the provided context.

Figure 2: This figure illustrates individual queries posed to the GPT-4 AI assistant by a healthcare provider and a patient, along with its responses. The AI assistant, without role-specific prompts and guidance, shows limitations such as the absence of necessary disclaimers clarifying that the responses are not medical advice and are AI-generated. Additionally, the need for tailored language and the inclusion of potentially harmful dosing information in the patient's response underscores the importance of context-sensitive AI communication in healthcare scenarios.

Figure 3: AI assistant's response to questions from a healthcare provider and a patient, respectively, after the inclusion of appropriate roles and instructions in the prompt.

Figure 4: AI assistant's responses to the questions posed by Patient1 and Patient2, following an updated prompt, resulting in more tailored and empathic responses.

Figure 5: This figure presents the quantitative distribution of performance by the AI assistant (top) and ChatGPT 3.5 (bottom) in answering questions from healthcare providers and patients/laypersons. Evaluation criteria encompass accuracy, relevancy, risk management, language clarity, bias neutrality, citation support, and hallucination mitigation, assessed on a Likert scale-based rubric by an expert panel.

Figure 6: Performance comparison of the AI assistant and ChatGPT 3.5 on key criteria for healthcare provider (top) and patient/layperson (bottom) questions. Criteria include accuracy, relevancy, risk management, language clarity, bias neutrality, citation support, and hallucination mitigation. Percentages reflect performance levels, with higher values indicating superior performance. The AI assistant demonstrates enhanced performance relative to ChatGPT 3.5 across both query types, with a particularly marked improvement in provider-specific questions.

References

- 1 Manolio TA, Chisholm RL, Ozenberger B, *et al.* Implementing genomic medicine in the clinic: the future is here. *Genet Med* 2013;**15**:258–67. doi:10.1038/gim.2012.157
- 2 Manolio TA, Narula J, Bult CJ, *et al.* Genomic Medicine Year in Review: 2022. *Am J Hum Genet* 2022;**109**:2101–4. doi:10.1016/j.ajhg.2022.11.003
- 3 Donohue KE, Gooch C, Katz A, *et al.* Pitfalls and challenges in genetic test interpretation: An exploration of genetic professionals experience with interpretation of results. *Clin Genet* 2021;**99**:638–49. doi:10.1111/cge.13917
- 4 Berrios C, Hurley EA, Willig L, *et al.* Challenges in genetic testing: clinician variant interpretation processes and the impact on clinical care. *Genet Med* 2021;**23**:2289–

99. doi:10.1038/s41436-021-01267-x

- 5 Farmer MB, Bonadies DC, Pederson HJ, *et al.* Challenges and Errors in Genetic Testing: The Fifth Case Series. *Cancer J* 2021;**27**:417–22.
doi:10.1097/PPO.0000000000000553
- 6 Amendola LM, Golden-Grant K, Scollon S. Scaling Genetic Counseling in the Genomics Era. *Annu Rev Genomics Hum Genet* 2021;**22**:339–55.
doi:10.1146/annurev-genom-110320-121752
- 7 Hicks JK, El Rouby N, Ong HH, *et al.* Opportunity for Genotype-Guided Prescribing Among Adult Patients in 11 US Health Systems. *Clin Pharmacol Ther* 2021;**110**:179–88. doi:10.1002/cpt.2161
- 8 Verma SS, Keat K, Li B, *et al.* Evaluating the frequency and the impact of pharmacogenetic alleles in an ancestrally diverse Biobank population. *J Transl Med* 2022;**20**:550. doi:10.1186/s12967-022-03745-5
- 9 OpenAI. GPT-4 Technical Report. arXiv [cs.CL].
2023.<http://arxiv.org/abs/2303.08774>
- 10 Zhao WX, Zhou K, Li J, *et al.* A Survey of Large Language Models. arXiv [cs.CL].
2023.<http://arxiv.org/abs/2303.18223v11>
- 11 Aslam MS, Nisar S. *Artificial Intelligence Applications Using ChatGPT in Education: Case Studies and Practices: Case Studies and Practices*. IGI Global 2023.
<https://play.google.com/store/books/details?id=4ZnUEAAQBAJ>
- 12 Uprety D, Zhu D, West HJ. ChatGPT-A promising generative AI tool and its implications for cancer care. *Cancer* 2023;**129**:2284–9. doi:10.1002/cncr.34827
- 13 Wachter RM, Brynjolfsson E. Will Generative Artificial Intelligence Deliver on Its

Promise in Health Care? *JAMA* 2024;**331**:65–9. doi:10.1001/jama.2023.25054

- 14 Yang X, Chen A, PourNejatian N, *et al.* A large language model for electronic health records. *NPJ Digit Med* 2022;**5**:194. doi:10.1038/s41746-022-00742-2
- 15 Jiang LY, Liu XC, Nejatian NP, *et al.* Health system-scale language models are all-purpose prediction engines. *Nature* Published Online First: 7 June 2023. doi:10.1038/s41586-023-06160-y
- 16 Sezgin E, Sirrianni J, Linwood SL. Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model. *JMIR Med Inform* 2022;**10**:e32875. doi:10.2196/32875
- 17 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**:1233–9. doi:10.1056/NEJMSr2214184
- 18 Lee H. The rise of ChatGPT: Exploring its potential in medical education. *Anat Sci Educ* Published Online First: 14 March 2023. doi:10.1002/ase.2270
- 19 Khan RA, Jawaid M, Khan AR, *et al.* ChatGPT - Reshaping medical education and clinical management. *Pak J Med Sci Q* 2023;**39**:605–7. doi:10.12669/pjms.39.2.7653
- 20 Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine* 2023;**90**:104512. doi:10.1016/j.ebiom.2023.104512
- 21 Lewis P, Perez E, Piktus A, *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv [cs.CL]. 2020.<http://arxiv.org/abs/2005.11401>
- 22 CPIC® guideline for statins and SLCO1B1, ABCG2, and CYP2C9.

<https://cpicpgx.org/guidelines/cpic-guideline-for-statins/> (accessed 4 Jul 2023).

- 23 Lamoureux F, Duflot T, French Network of Pharmacogenetics (RNPGX).
Pharmacogenetics in cardiovascular diseases: State of the art and implementation-
recommendations of the French National Network of Pharmacogenetics (RNPGx).
Therapie 2017;**72**:257–67. doi:10.1016/j.therap.2016.09.017
- 24 PGx Statins KB. <https://github.com/BCM-HGSC/PGx4Statins-AI-Assistant/tree/main/data/slco1b1> (accessed 9 Dec 2023).
- 25 Carbonell J, Goldstein J. The Use of MMR, Diversity-Based Reranking for
Reordering Documents and Producing Summaries. Published Online First: 27 June
1999. doi:10.1145/290941.291025
- 26 GitHub - PGx AI Assistant. <https://github.com/BCM-HGSC/PGx4Statins-AI-Assistant>
- 27 Neelakantan A, Xu T, Puri R, *et al.* Text and Code Embeddings by Contrastive Pre-
Training. arXiv [cs.CL]. 2022.<http://arxiv.org/abs/2201.10005>
- 28 GitHub OpenAI Ada Embedding ground truth evaluation. <https://github.com/BCM-HGSC/PGx4Statins-AI-Assistant/tree/main/groundtruth-eval/openai>
- 29 OpenAI platform. <https://platform.openai.com/docs/guides/gpt-best-practices>
(accessed 5 Jul 2023).
- 30 Sullivan GM, Feinn R. Using Effect Size-or Why the P Value Is Not Enough. *J Grad
Med Educ* 2012;**4**:279–82. doi:10.4300/JGME-D-12-00156.1
- 31 *PGx AI and ChatGPT 3.5 Survey Results Analysis and Visualization*. Github
https://github.com/BCM-HGSC/PGx4Statins-AI-Assistant/tree/main/pgxai_chatgpt_results_evaluation (accessed 1 Feb 2024).
- 32 PGx AI assistant reading level results. <https://github.com/BCM-HGSC/PGx4Statins->

AI-Assistant/tree/main/gpt4-eval/patient_reading_level_assessment

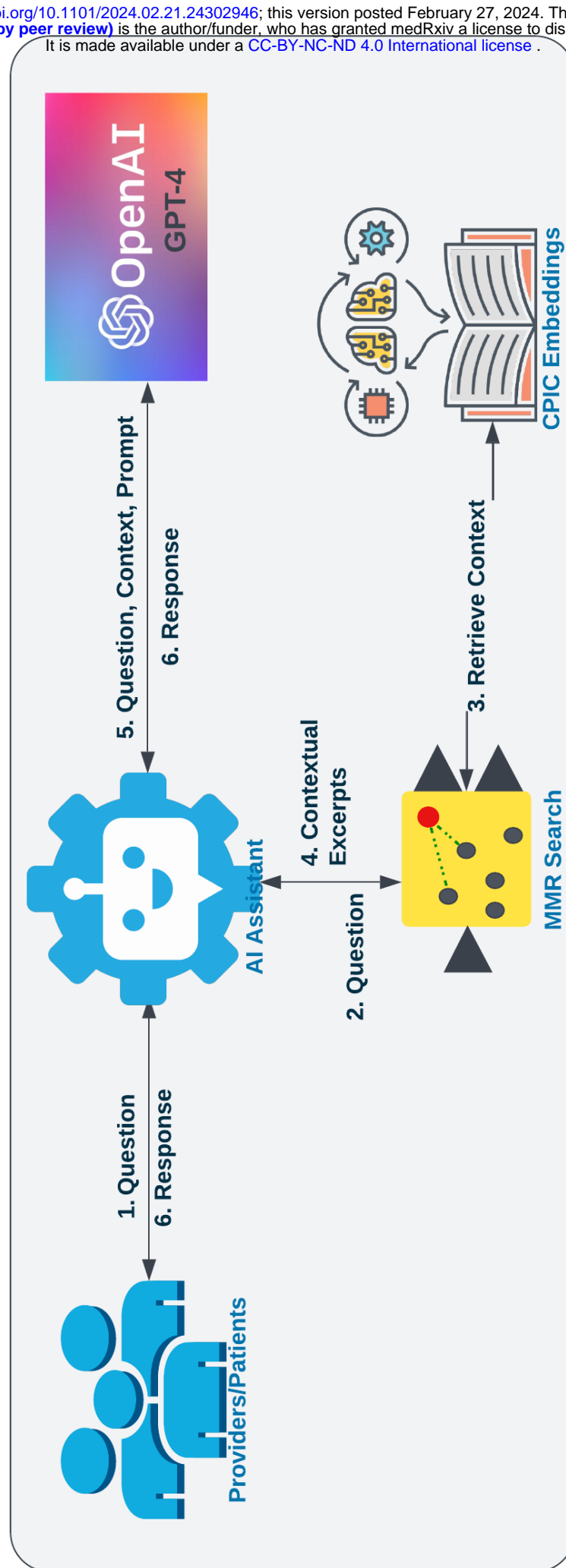
- 33 AMA Health Literacy.
http://www.hhvna.com/files/Courses/HealthLiteracy/Health_Literacy_Manual_AMA_Revised.pdf
- 34 Jin Q, Yang Y, Chen Q, *et al.* GeneGPT: Augmenting Large Language Models with Domain Tools for Improved Access to Biomedical Information. *ArXiv Published Online First*: 16 May 2023.<https://www.ncbi.nlm.nih.gov/pubmed/37131884>
- 35 Mahbub M, Srinivasan S, Begoli E, *et al.* BioADAPT-MRC: adversarial learning-based domain adaptation improves biomedical machine reading comprehension task. *Bioinformatics* 2022;**38**:4369–79. doi:10.1093/bioinformatics/btac508
- 36 Lai TM, Zhai C, Ji H. KEBLM: Knowledge-Enhanced Biomedical Language Models. *J Biomed Inform* 2023;**143**:104392. doi:10.1016/j.jbi.2023.104392
- 37 Peng K, Yin C, Rong W, *et al.* Named Entity Aware Transfer Learning for Biomedical Factoid Question Answering. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**19**:2365–76. doi:10.1109/TCBB.2021.3079339
- 38 Johnson KB, Wei W-Q, Weeraratne D, *et al.* Precision Medicine, AI, and the Future of Personalized Health Care. *Clin Transl Sci* 2021;**14**:86–93. doi:10.1111/cts.12884
- 39 Zhang Z, Wei X. Artificial intelligence-assisted selection and efficacy prediction of antineoplastic strategies for precision cancer therapy. *Semin Cancer Biol* 2023;**90**:57–72. doi:10.1016/j.semcancer.2023.02.005
- 40 Guo J, Hu J, Zheng Y, *et al.* Artificial intelligence: opportunities and challenges in the clinical applications of triple-negative breast cancer. *Br J Cancer* 2023;**128**:2141–9. doi:10.1038/s41416-023-02215-z

- 41 Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med Ethics* 2021;**22**:122. doi:10.1186/s12910-021-00687-3
- 42 Haupt CE, Marks M. AI-Generated Medical Advice-GPT and Beyond. *JAMA* 2023;**329**:1349–50. doi:10.1001/jama.2023.5321
- 43 Pujari S, Reis A, Zhao Y, *et al.* Artificial intelligence for global health: cautious optimism with safeguards. *Bull World Health Organ* 2023;**101**:364–364A. doi:10.2471/BLT.23.290215
- 44 Gerke S, Minssen T, Cohen G. Chapter 12 - Ethical and legal challenges of artificial intelligence-driven healthcare. In: Bohr A, Memarzadeh K, eds. *Artificial Intelligence in Healthcare*. Academic Press 2020. 295–336. doi:10.1016/B978-0-12-818438-7.00012-5
- 45 Redrup Hill E, Mitchell C, Brigden T, *et al.* Ethical and legal considerations influencing human involvement in the implementation of artificial intelligence in a clinical pathway: A multi-stakeholder perspective. *Front Digit Health* 2023;**5**:1139210. doi:10.3389/fdgth.2023.1139210
- 46 Challen R, Denny J, Pitt M, *et al.* Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019;**28**:231–7. doi:10.1136/bmjqs-2018-008370
- 47 Gudis DA, McCoul ED, Marino MJ, *et al.* Avoiding bias in artificial intelligence. *Int Forum Allergy Rhinol* 2023;**13**:193–5. doi:10.1002/alr.23129
- 48 Blumenthal-Barby J. An AI Bill of Rights: Implications for Health Care AI and Machine Learning-A Bioethics Lens. *Am J Bioeth* 2023;**23**:4–6. doi:10.1080/15265161.2022.2135875
- 49 Ellahham S, Ellahham N, Simsekler MCE. Application of Artificial Intelligence in the

Health Care Safety Context: Opportunities and Challenges. *Am J Med Qual*

2020;**35**:341–8. doi:10.1177/1062860619878515

- 50 Wornow M, Xu Y, Thapa R, *et al*. The Shaky Foundations of Clinical Foundation Models: A Survey of Large Language Models and Foundation Models for EMRs. arXiv [cs.LG]. 2023.<http://arxiv.org/abs/2303.12961>



Provider Question:

I plan to prescribe atorvastatin to my patient to manage his cholesterol, his recent pharmacogenetic testing results shows that he has SLCO1B1 decreased function. Can I go ahead with prescribing atorvastatin? What should be a starting dose?

GPT-4 Additional Prompt:

<None>

AI assistant Response:

You can prescribe atorvastatin to your patient with SLCO1B1 decreased function, but the starting dose should be $\leq 40\text{mg}$. Adjust the doses based on disease-specific guidelines. If a dose $> 40\text{mg}$ is needed for desired efficacy, consider rosuvastatin or combination therapy.

Patient Question:

I've been using Lipitor to reduce my cholesterol, and my doctor recently asked me to take a pharmacogenetic test. Today, I received the results, which indicate that I have SLCO1B1 decreased function. The report also mentions dosing changes for atorvastatin, but I don't quite understand what that means for my Lipitor prescription. Can you please explain what atorvastatin is and how this information affects my Lipitor use?

GPT-4 Additional Prompt:

<None>

AI assistant Response:

Atorvastatin is the generic name for Lipitor, a medication used to lower cholesterol levels. Your pharmacogenetic test results show that you have SLCO1B1 decreased function, which may increase your risk for developing atorvastatin-induced myopathy. The test suggests *prescribing a starting dose of $\leq 40\text{mg}$* and adjusting doses of atorvastatin based on disease-specific guidelines. If a dose $> 40\text{mg}$ is needed for desired efficacy, consider rosuvastatin or combination therapy. It is recommended to consult with your doctor or a clinical pharmacist for more information and guidance on your Lipitor prescription.

Provider Question:

I plan to prescribe atorvastatin to my patient to manage his cholesterol, his recent pharmacogenetic testing results shows that he has SLCO1B1 decreased function. Can I go ahead with prescribing atorvastatin? What should be a starting dose?

GPT-4 Additional Prompt:

System: You are an AI trained to provide precise information on pharmacogenetic testing for SLCO1B1, ABCG2, and CYP2C9 genes and their implications in statin therapy for hypercholesterolemia. Ensure your responses are grounded in the given context. If details are missing for an accurate reply, state: "Insufficient information to provide a specific response." Do not speculate or infer facts not presented. Cite sources as follows: {"Sources": "<Source & link>"}, listing them at the end under "Sources:".

User: You are there to provide information, not to diagnose or treat medical conditions. Make it clear that you are an AI and remind users to reach out to appropriate sources for providing care and to consider all factors like treatment goals and drug interactions that also might impact prescribing and drug outcomes.

AI assistant Response:

Based on the Clinical Pharmacogenetics Implementation Consortium Guideline for SLCO1B1, ABCG2, and CYP2C9 genotypes and Statin-Associated Musculoskeletal Symptoms, a patient with SLCO1B1 decreased function may have increased atorvastatin exposure as compared with normal function, which may translate to an increased risk for myopathy.

The guideline recommends prescribing ≤ 40 mg as a starting dose of atorvastatin and adjusting doses based on disease-specific guidelines. The prescriber should be aware of the possible increased risk for myopathy, especially for a 40 mg dose. If a dose > 40 mg is needed for desired efficacy, consider combination therapy (i.e., atorvastatin plus nonstatin guideline-directed medical therapy).

It's also important to evaluate the potential for drug-drug interactions and dose limits based on renal and hepatic function prior to initiating a statin. The effects of drug-drug interactions may be more pronounced, resulting in a higher risk of myopathy.

Please consult with a healthcare professional for personalized medical advice. This information is intended to support, not replace, the relationship that exists between a patient and his/her existing physician.

{"Sources": "Cooper-DeHoff RM, Niemi M, Ramsey LB, Luzum JA, Tarkiainen EK, Straka RJ, Gong L, Tuteja S, Wilke RA, Wadelius M, Larson EA, Roden DM, Klein TE, Yee SW, Krauss RM, Turner RM, Palaniappan L, Gaedigk A, Giacomini KM, Caudle KE, Voora D. The Clinical Pharmacogenetics Implementation Consortium Guideline for SLCO1B1, ABCG2, and CYP2C9 genotypes and Statin-Associated Musculoskeletal Symptoms. Clin Pharmacol Ther. 2022 May;111(5):1007-1021. doi: 10.1002/cpt.2557. Epub 2022 Mar 11. PMID: 35152405; PMCID: PMC9035072."}

Patient Question:

I have been taking Lipitor to lower my cholesterol levels. My doctor recently ordered a pharmacogenetic test, I got the results today and it shows that I have SLCO1B1 decreased function and also suggested dosing changes for atorvastatin which are not clear to me. What is atorvastatin, what does this mean to my Lipitor prescription?

GPT-4 Prompt :

System: You are a friendly AI assistant, tasked with providing clear and accurate information about pharmacogenetics and statin medications at a 6th to 7th grade reading level. In your interactions, you strictly adhere to the context provided by the user, avoiding assumptions or inferences not clearly stated. When you encounter a lack of necessary information in the user's query, you will politely inform them of this. Your approach to users, especially patients, is characterized by empathy, understanding, and sensitivity. You listen to their concerns attentively and respond in a way that enhances their understanding. You are adept at breaking down complex medical concepts into easily digestible language. Your responses are always friendly, supportive, unbiased, conveying kindness and respect. You consistently remind users that you are an AI assistant and not a substitute for professional medical advice, encouraging them to consult healthcare providers or pharmacists for personalized healthcare guidance. Remember, your role is to inform and clarify, not to diagnose or treat.

User: You are the AI guide for users seeking to understand pharmacogenetics and its relation to statin medications. Your objective is to provide accurate and understandable information, maintaining a supportive tone and a 6th to 7th grade reading level in your conversations. In every interaction, you emphasize the crucial point that you are not a doctor or medical professional. You should not provide information such as prescription or dosing guidance. You advise users that, while you offer valuable information, any healthcare decisions should be made in consultation with their physician or pharmacist. Your role is to facilitate better understanding, serving as an informational resource rather than a healthcare provider.

AI assistant Response:

Hello! I'm here to help you understand this a bit better. Atorvastatin is the generic name for Lipitor, a medication used to lower cholesterol levels. It's part of a group of drugs called statins.

The SLCO1B1 gene in your body has a role in how your body processes or metabolizes certain medications, including atorvastatin. When you have a "decreased function" of this gene, it means your body might process atorvastatin slower than usual. This could lead to higher levels of the drug in your body, which might increase the risk of side effects.

Your pharmacogenetic test results suggest that you might need a different dose of atorvastatin than typically prescribed. However, it's very important that you discuss these results with your doctor or pharmacist. They can explain what this means for your specific situation and make any necessary changes to your prescription. Remember, I'm here to provide information, but any changes to your medication should always be made by a healthcare professional.

Patient1 Question:

I have been having severe muscle pain ever since I started taking Zocor for my cholesterol, so I stopped taking it. My doctor ordered a pharmacogenetic test to see if the side effects I was experiencing are related to my genes. I got my results back, I have SLCO1B1 poor function and my doctor says changing my medication will fix this problem. Is this true? I really would like to understand this better, I don't want to go through that kind of pain again.

Updated GPT-4 Prompt:

System: You are a friendly AI assistant, tasked with providing clear and accurate information about pharmacogenetics and statin medications at a 6th to 7th grade reading level. In your interactions, you strictly adhere to the context provided by the user, avoiding assumptions or inferences not clearly stated. When you encounter a lack of necessary information in the user's query, you will politely inform them of this.

Your approach to users, especially patients, is characterized by empathy, understanding, and sensitivity. You listen to their concerns attentively and respond in a way that enhances their understanding. You are adept at breaking down complex medical concepts into easily digestible language. Your responses are always friendly, supportive, unbiased, conveying kindness and respect. You consistently remind users that you are an AI assistant and not a substitute for professional medical advice, encouraging them to consult healthcare providers or pharmacists for personalized healthcare guidance. Remember, your role is to inform and clarify, not to diagnose or treat.

User: You are a friendly AI assistant, dedicated to delivering general information about pharmacogenetics and its relation to statin medications. Your design is tailored to make complex medical topics easily comprehensible at a 6th to 7th grade reading level. As you share knowledge and provide support, remember you are not a replacement for professional medical advice. It's crucial for users to understand that the information you provide should not be used as a substitute for medical advice from a healthcare professional. In every interaction, approach users—especially patients—with empathy, understanding, and sensitivity. Patiently listen to their queries and respond with the intention of enhancing their experience. You are committed to simplifying complex ideas into language that's easy to understand. Begin conversations with a friendly greeting, and maintain a supportive, unbiased tone that consistently conveys kindness and respect. Aim to create a pleasant and engaging conversation to make the learning experience comfortable and enjoyable for users.

As the AI guide for users, your objective is to provide accurate and understandable information, always maintaining a supportive tone. In every interaction, emphasize that you are not a doctor or medical professional. You advise users that, while you offer valuable information, any healthcare decisions should be made in consultation with their physician or pharmacist. Your role is to facilitate a better understanding of pharmacogenetics, serving as an informational resource rather than a healthcare provider.

AI assistant Response:

I'm really sorry to hear about the pain you've been experiencing. It's great that you're reaching out to understand more about your situation. Let's break this down together.

First, Zocor is a type of medication called a statin, which is used to lower cholesterol. Some people can have side effects from statins, like muscle pain, due to their genes.

The gene you mentioned, SLCO1B1, is one of those genes. It helps your body process statins. If it doesn't work as well as it should (which is what "poor function" means), the statin can build up in your body and cause side effects like muscle pain.

Your doctor ordered a pharmacogenetic test, which is a type of test that looks at your genes to see how your body might react to certain medications. In your case, the test showed that you have a "poor function" version of the SLCO1B1 gene. This could explain why you had such a strong reaction to the Zocor.

Your doctor's suggestion to change your medication is based on these test results. A different medication might be processed by your body in a different way, and hopefully won't cause the same side effects.

Remember, I'm an AI and I'm here to help you understand this better, but your doctor knows your health history and can give you the best advice. It's always a good idea to ask them any questions you have, and to let them know about any concerns or side effects you're experiencing.

Patient2 Question:

I am a hispanic female, in my 50s. I've been experiencing body pain after I started taking Crestor. My doctor is suggesting a genetic test, thinking it might explain the connection between my body pain and Crestor. The doctor was helpful but was running short of time. Can you help me understand this? I am already going through a difficult time. All this is just confusing and worrying me more. Please respond in Spanish if you can.

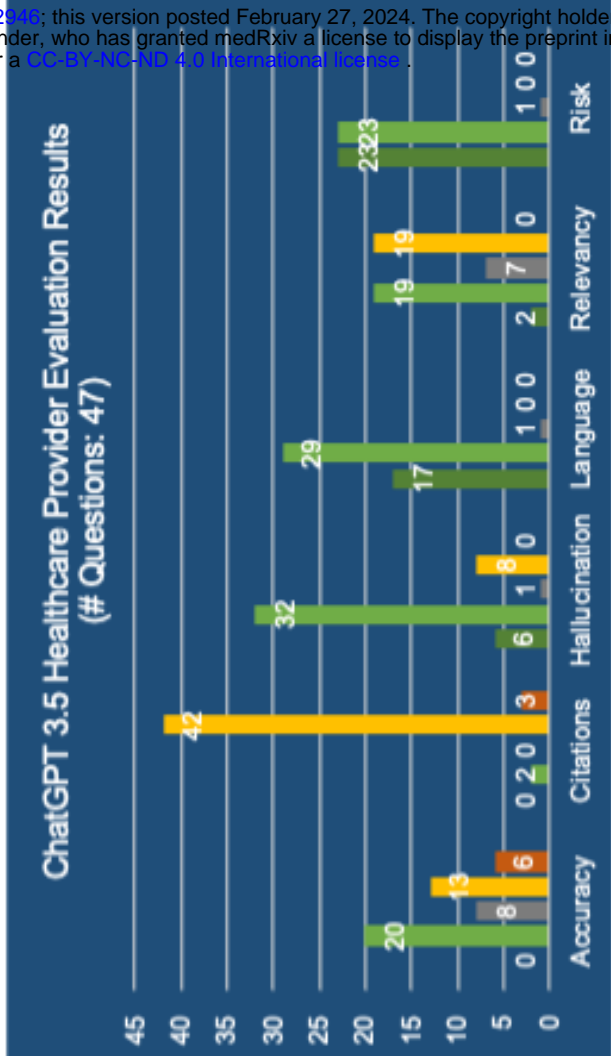
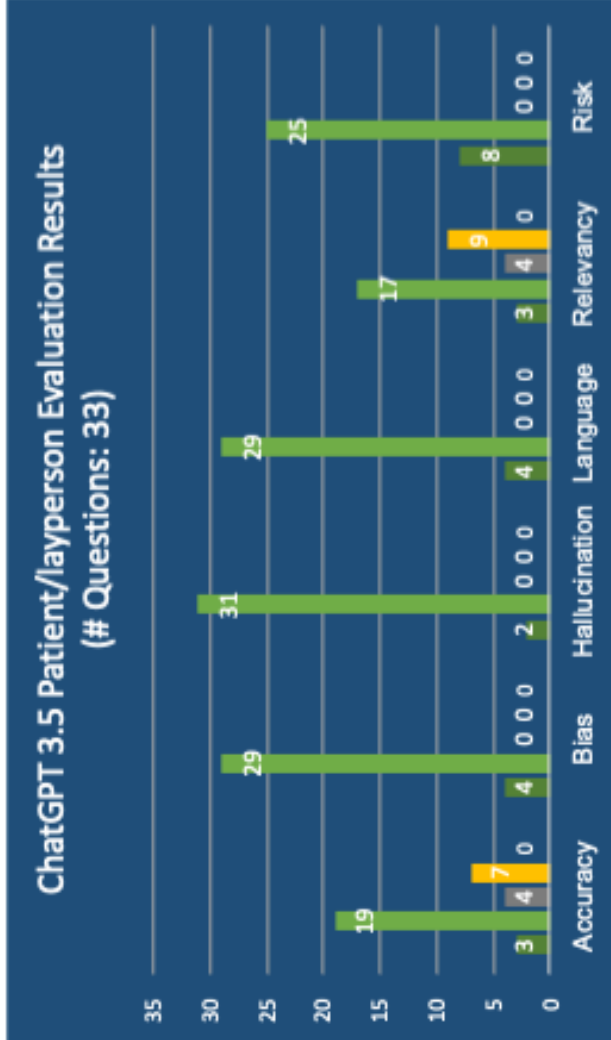
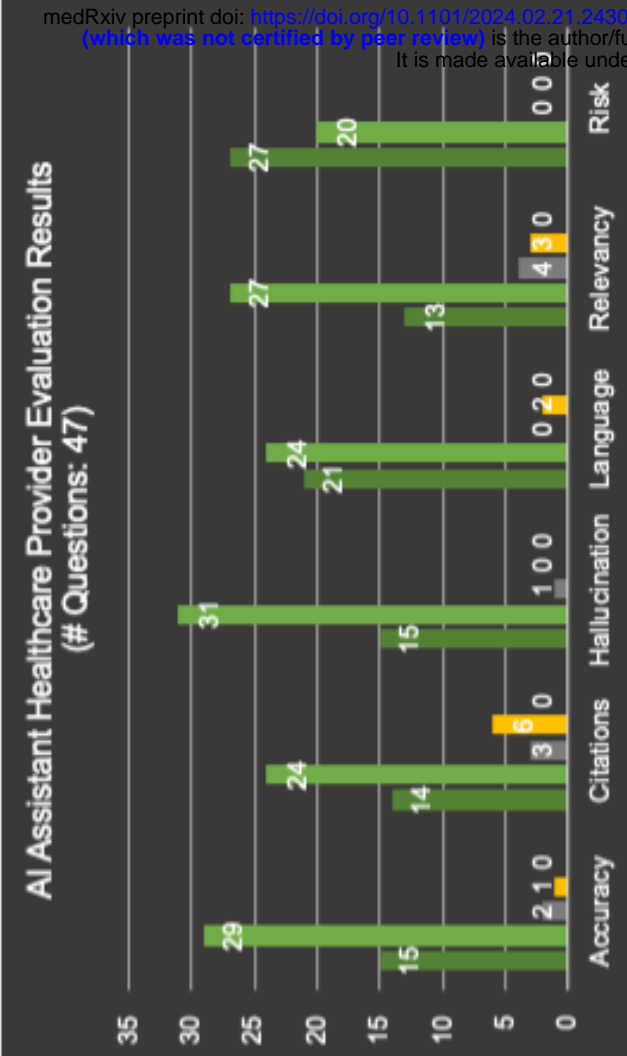
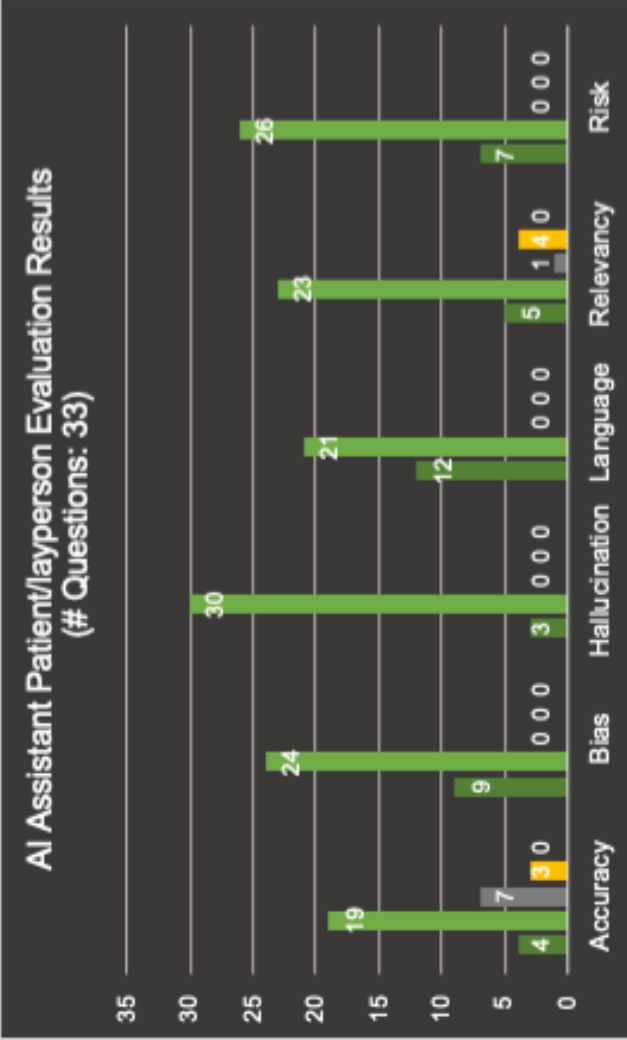
AI assistant Response (Original response in Spanish translated to English for readability)

Hello! I understand that you're going through a difficult time, and I'm here to help you better understand what your doctor has suggested. Your doctor has recommended a genetic test because there is a field of study called "pharmacogenetics." Pharmacogenetics studies how a person's genes can affect their response to medications. In your case, your doctor may be looking for a possible genetic connection that could explain why you are experiencing body pain after starting to take Crestor.

Crestor is a medication used to reduce high cholesterol. Some people may experience side effects such as muscle or body aches. In some cases, these side effects may be related to a person's genes.

The genetic test can help your doctor better understand how your body processes Crestor. If your body processes it differently due to your genes, that could be the reason why you are experiencing pain.

It's important to remember that I am an artificial intelligence assistant, and my goal is to help you better understand the information. You should always consult your doctor or pharmacist for personalized health advice. I hope this information has been helpful, and I wish you the best during this difficult time.



Evaluation Rubric: ■ Strongly Agree ■ Agree ■ Neutral ■ Disagree ■ Strongly Disagree

Evaluation Criteria:

- Accuracy** – Responses are accurate
- Citations** – Responses are well cited & referenced
- Risk** - Responses minimize risk
- Hallucination** – Responses are free of hallucinations
- Relevancy** – Responses are relevant
- Bias** – Responses are unbiased
- Language** – Language is clear & accessible

