PGxQA: A Resource for Evaluating LLM Performance for Pharmacogenomic QA Tasks[†]

Karl Keat[1]*, Rasika Venkatesh[1]*, Yidi Huang[1]*, Rachit Kumar[1], Sony Tuteja[2], Katrin Sangkuhl[3], Binglan Li[3], Li Gong[3], Michelle Whirl-Carrillo[3], Teri E. Klein[3,4,5], Marylyn D. Ritchie[6,7,8]**, Dokyoon Kim[7,8]**
[1]*Genomics and Computational Biology Graduate Program,* [2]*Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[3]*Department of Biomedical Data Science,* [4]*Department of Medicine (BMIR),* [5]*Department of Genetics, Stanford University, Stanford, CA USA*
[6]*Department of Genetics,* [7]*Institute for Biomedical Informatics,* [8]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA*
*, **The Authors contributed equally to this work
***Emails: marylyn@pennmedicine.upenn.edu, dokyoon.kim@pennmedicine.upenn.edu*

Pharmacogenetics represents one of the most promising areas of precision medicine, with several guidelines for genetics-guided treatment ready for clinical use. Despite this, implementation has been slow, with few health systems incorporating the technology into their standard of care. One major barrier to uptake is the lack of education and awareness of pharmacogenetics among clinicians and patients. The introduction of large language models (LLMs) like GPT-4 has raised the possibility of medical chatbots that deliver timely information to clinicians, patients, and researchers with a simple interface. Although state-of-the-art LLMs have shown impressive performance at advanced tasks like medical licensing exams, in practice they still often provide false information, which is particularly hazardous in a clinical context. To quantify the extent of this issue, we developed a series of automated and expert-scored tests to evaluate the performance of chatbots in answering pharmacogenetics questions from the perspective of clinicians, patients, and researchers. We applied this benchmark to state-of-the-art LLMs and found that newer models like GPT-4o greatly outperform their predecessors, but still fall short of the standards required for clinical use. Our benchmark will be a valuable public resource for subsequent developments in this space as we work towards better clinical AI for pharmacogenetics.

Keywords: Pharmacogenetics; Pharmacogenomics, Large Language Models, Artificial Intelligence, Clinical Informatics.

1. Introduction

1.1. *Pharmacogenetics*

Pharmacogenetics (PGx) is the study of the role of genetics on an individual's response to medication, with the aim of bringing tools to the clinic that can utilize a patient's genetic information to improve medication safety and efficacy. Genetic variations that lead to changes in the activity or availability of drug metabolizing enzymes (DMEs), receptors, channels, and other proteins involved in pharmacodynamics and pharmacokinetics can contribute strongly to interindividual variability in drug response, resulting in an increased risk of adverse drug reactions (ADRs) and nonresponse

[†]KK is funded through National Human Genome Research Institute (NHGRI) F31HG013246. RK is funded through NHGRI T32HG000046. This work was also supported by the following grants from the National Institutes of Health (NIH): U24HG010615, U24HG013077, UL1TR001878, and K23HL143161.

phenotypes.[1] By identifying genetic markers that influence drug response, PGx enables healthcare providers to predict which patients are more likely to experience adverse reactions or treatment failure. This knowledge allows for more individually tailored medication regimens, optimizing therapeutic outcomes while minimizing the risk of side effects.[2] The overarching goal of PGx is promoting personalized medicine, such that patients receive the right drug and the right dose, at the right time. In doing so, the field aims to improve patient outcomes, enhance medication safety, and reduce healthcare costs associated with ineffective or harmful treatments.

Despite the availability of numerous well-characterized, clinically actionable PGx guidelines for widely used medications, the clinical implementation of PGx has been slow. Very few medical centers and clinics routinely use this technology. This gap is due to various factors such as a lack of awareness and education among healthcare providers, the constantly evolving body of PGx guidelines, and technical challenges in integrating PGx data into electronic health records (EHRs).[3] The cost of PGx testing and variable insurance coverage can also pose significant financial barriers, while regulatory and legal concerns may also impact the extent of implementation of PGx testing in hospital systems.[4] Lack of domain expertise and education among healthcare providers, patients, and researchers in particular poses a critical barrier to the implementation of PGx-guided therapies in clinical settings as this leads to difficulty understanding and interpreting test results, in addition to limited research conducted regarding the clinical impact of such technologies.[5]

1.2. *Existing PGx Resources and Limitations*

Given that there are many causes for interindividual variability in treatment response as well as a need for guidance in interpreting PGx screening results, multiple independent bodies of experts have published research and guidelines to inform PGx-guided treatment. The Clinical Pharmacogenetics Implementation Consortium (CPIC) is one such group that has generated a set of specific drug recommendations to guide prescribing practices in the presence of genetic test results. CPIC has established 43 evidence-based clinical guidelines for 151 commonly prescribed medications. These recommendations were created based on a large body of evidence showing the impact of known PGx alleles in altering drug metabolism or response. Level A refers to gene-drug pairs where genetic information "should be used" for prescribing decisions and alternative therapies or dosing are highly likely to be effective and safe. At least one moderate or strong action (change in prescribing) is recommended for Level A pairs. Level B refers to pairs where genetic information "could be used" to change prescribing because alternative therapies/dosing are extremely likely to be as effective and as safe as non-genetically based dosing. Other international committees with their own sets of guidelines include The Dutch Pharmacogenetics Working Group (DPWG), and the French National Network (Réseau) of Pharmacogenetics (RNPGx).[6] The Pharmacogenomics Knowledge Base (PharmGKB), is a resource that aims to comprehensively aggregate, curate, and characterize PGx knowledge including the literature and guidelines from these distinct sources.[7]

While these resources are highly comprehensive, most require a moderate to high degree of domain knowledge to understand and interpret the provided information. Clinicians and patients, in particular, need PGx expertise to understand reports and utilize them to inform treatment decisions. Clinicians typically receive limited PGx training and therefore rely heavily on these resources for guidance.[5,8–12] Moreover, differences among guidance sources and the rapid pace of new discoveries and guidelines create potential for misunderstandings and confusion. While PharmGKB curates,

aggregates, and presents guidance across sources, clinicians, patients, and researchers may prefer an interface that allows them to query and access targeted information using natural language instead of menus and tables.

1.3. *Opportunities for Large Language Models to Guide PGx*

Large language models (LLMs) represent a major advance in artificial intelligence, allowing for the creation of seemingly intelligent chatbots which can interpret questions and assist with various tasks. LLMs have shown promise in a variety of natural language tasks, including those in medicine. For example, chatbots using LLMs can accurately answer patient queries in a conversational manner preferred by patients. GPT-4 has also achieved human-level accuracy on the United States Medical Licensing Exam (USMLE), outperforming the minimum passing threshold on short answer and multiple-choice questions.[13] LLMs have been proposed for integration into clinical workflows to handle administrative tasks, which include managing appointment scheduling by patient request, answering routine inquiries about medication or treatment plans, and assisting in the preparation of medical records.[14,15] Additionally, LLMs can support clinical decision-making by providing real-time information retrieval and analysis, potentially reducing the cognitive load on healthcare professionals and improving patient outcomes.[16] For these reasons, advances in LLMs have created an exciting opportunity to build chatbots to assist with complex medical specialties like PGx, providing a powerful and intuitive interface to access pharmacogenetic knowledge.

Despite the promise of LLMs in medicine, there are significant issues that must be addressed before widespread clinical integration. These models are limited to the information they were trained on and can produce fabricated responses with an authoritative and confident tone when lacking information. There are numerous examples of this phenomenon across disciplines, but this poses a particularly large barrier to use in healthcare, where real time patient decisions rely on the presence of accurate information and mistakes can cost lives.[17–19] Moreover, LLMs are costly to update and retrain as new information becomes available.[20–22] This poses a challenge in fields where clinical guidelines are routinely updated, such as in PGx, and even current state-of-the-art LLMs had their training data capped several months before the latest CPIC guideline release. Despite these risks, LLMs are already being employed by clinicians, patients, and researchers to answer medical questions and their performance must be studied in order to understand their limitations.[23]

1.4. *Prior work on LLMs for PGx*

PGx is a specialized area of medicine with limited and variable levels of coverage in the US medical and pharmacy curriculum.[5,10–12] Despite this, PGx has a wide impact on several specialties due to the variety of drugs with actionable guidelines. Therefore, leveraging LLMs in this field has the potential to significantly enhance clinical practice and patient care. For instance, Murugan et al., used GPT-4 and retrieval-augmented generation (RAG) to build PGx4Statins, a PGx chatbot for answering questions about statin therapy guidelines.[24] However, the limitations of LLMs may pose a particular risk in this field, as PGx guidelines are revised and updated irregularly as new evidence becomes available, and inaccurate or outdated advice may result in adverse drug reactions or

treatment nonresponse. As such, any PGx chatbot would need to be thoroughly vetted before clinical implementation is possible.

While the performance of LLMs at answering general medical questions has been demonstrated, there is limited data on how LLMs perform with PGx queries. Prior to now, there have been no comprehensive, publicly available benchmarks to assess the performance of LLM chatbots in answering PGx queries. PGx4Statins was benchmarked manually, requiring a team of scorers to rate LLM responses based on the criteria of accuracy, relevancy, risk management, language clarity, bias neutrality, empathetic sensitivity, citation support, and hallucination limitation on a 1-5 scale. While this likely represents a gold-standard approach for evaluating real-world performance of a PGx clinical chatbot, PGx4Statins was only able to be tested on a small number of questions and for a single drug, demonstrating the limitations of this evaluation strategy.[24] As new chatbots and language models are released, a more scalable solution is needed to comprehensively test the accuracy of these tools, so that we can then prioritize top performers for more rigorous, labor-intensive testing.

To address the absence of evaluation strategies for PGx chatbots, we have developed PGxQA, a resource for evaluating the performance of LLMs in a variety of PGx-related tasks for multiple identified stakeholders: patients, clinicians, and researchers. PGxQA consists of a large corpus of PGx questions generated directly from CPIC data resources, CPIC PGx guidance for Level A drug-gene pairs, or provided by experts in the field. In addition, PGxQA includes tools for higher throughput manual and automated evaluation of accuracy and completeness. PGxQA's question set covers all of the CPIC Level A guidelines across several dimensions, such as translating genotypes into phenotypes, naming the dbSNP ID(s) for variant(s) that define a particular star-allele, and most importantly, translating phenotypes into clinical recommendations. These resources will help promote the responsible development of medical chatbots by allowing us to assess their knowledge of PGx topics, thus lowering barriers to implementation of PGx in the clinic and providing easier access to PGx knowledge for clinicians, patients, and researchers.

2. Methods

2.1. *Automated Question Generation*

To generate a meaningfully large corpus of evaluation questions, a significant proportion of the question bank was generated using custom python scripts to extract relevant information from the 'CPIC Data' database from their GitHub repository and format the information as question-answer pairs.[25] The psycopg2 package was used to load and query CPIC's postgresql database and pandas was used to output tables of questions.[26–28]

Due to a large degree of redundancy in questions and the potential for an over-weighting of pharmacogenes with many defined star alleles in our overall scoring, we implemented a subsetting tool which takes each set of questions and drops redundant questions to maintain roughly even proportions of questions based on which genes they cover and what answer choices they cover. All generated questions are available for download, such that users can run the entire set or generate custom subsets based on their own criteria.

2.2. *LLM Querying*

To query the various studied LLMs, we wrote a set of python scripts to load in our questions and send them to a local or remote LLM server. We defined a universal base prompt for all LLMs to ensure that all LLMs are working with similar basic instructions. We used the 'openai' python package along with an OpenAI API key to remotely query GPT-3.5-turbo, GPT-4-turbo, and OpenAI's latest model as of writing, GPT-4o. We were also able to use the 'openai' python interface to send queries to a locally hosted instance of the open-source LLM Llama 3. Lastly, we used the 'requests' library in python to connect to Google's Generative Language REST API to query Gemini 1.5 Pro, Google's flagship LLM product.[29] We used our python code to query the LLMs with all of the questions in our subsets, outputting tables containing the original question, question metadata, the ground-truth reference answer, the LLM answer, and some automated scoring metrics.

2.3. *Manual Question Generation*

2.3.1. *External Provided Questions*

While the structured information within the CPIC database allows us to cover a large proportion of the potential use cases for a PGx chatbot, we sought out real world sources of PGx questions to represent what information is being sought by actual clinicians, researchers, and patients. We acquired a set of questions sent to PharmGKB scientists from 2020-2024, containing queries about PGx and the PharmGKB scientists' responses. Additionally, we obtained an anonymized set of questions and answers from Penn Medicine's Pharmacogenetics Consult Service, which provided a rich source of clinician-centric questions on PGx testing, results interpretation, and other relevant queries. We manually pruned these datasets to stay within our scope of queries about PGx information retrieval and formatted them into tables as short answer questions for our LLMs.

2.3.2. *Adversarial Questions*

To assess how the models perform when presented with incorrect information, insufficient information, or information outside of the scope of queries regarding PGx, we devised sets of structured adversarial questions. These queries were structured to be nearly identical to the question bank extracted directly from the CPIC database, with the exception of having extraneous or missing information. For these queries, we evaluate whether LLMs answer that sufficient information was not available to answer the question, scoring based on the rate of refusal to respond. We additionally ran the whole set of LLM queries, giving the LLMs the option to refuse to respond, as to compare refusal rates between standard and adversarial queries.

2.4. *Automated LLM Metrics*

To rapidly score the large corpus of questions and reduce reliance on expert labor, we generated a set of automated scoring functions to directly measure or approximate the performance of the LLMs on each specific task.

2.4.1. *Numeric Scoring*

For questions requiring a numeric answer, such as the allele frequency tests, LLMs were instructed to format their response as a number. We then parsed out this number and calculated the mean absolute deviation (defined as the mean of the differences) between the LLM answer and the reference answer for the entire question set.

2.4.2. *Information Retrieval Scoring*

For questions where the task involved returning non-sentence information such as dbSNP IDs, gene symbols, or generic drug names, we instructed the LLMs to return the desired information in a predictable format that can be parsed using regular expressions or by splitting a defined delimiting character like ';'. For question sets where there are multiple values making up the answer (for example to list all of the drugs which have CPIC guidelines linked to a particular gene), performance was measured as precision and recall, where precision is the proportion of values in the LLM answer that are found in the reference answer, and recall is the proportion of values in the reference answer that were correctly included in the LLM answer.

2.4.3. *Multiple Choice Scoring*

For question sets where the questions had a small finite set of possible answers, we constrained them to multiple choice, where the LLM was told to select the correct answer from a provided list of options, facilitating the process of detecting if the LLM answered correctly programmatically. For these queries, the accuracy of the LLM in identifying the correct response was computed as the proportion of answers that were correctly selected.

2.4.4. *Automated Text Similarity Metrics*

In the case of short-answer questions where we wanted the LLMs to answer in one or two sentences, it is nontrivial to directly score the accuracy without human graders with the expertise to evaluate the answers, which presents a scalability issue. To roughly approximate human scoring, we computed automated text similarity metrics between the LLM answer and a human-written reference answer. Specifically, we compute the cosine similarity of the answers under different text embedding models as well as BERTScore using the microsoft/deberta-xlarge-mnli base model. We selected the model that most closely resembled human judgement by comparing the embedding scores' concordance with human-scored answers.[30–34] We then calculated the "win-rate" of the LLM answers by looking at the percentage of answers where the LLM similarity score to the reference answer was higher than the LLM similarity score to a generic discordant answer. For example, if asked to make a clinical recommendation, where the correct answer is to avoid the drug and the discordant answer is to take the drug as normal, the LLM would "win" if its answer has higher similarity to the reference answer than the discordant answer.

2.5. *Human Review of LLM Answers*

2.5.1. *Concordance with Automated Metrics*

To determine which text metric best captures the semantics of PGx recommendations, we manually reviewed a set of 77 short-answer questions and responses from GPT-4o. For each question, we manually annotated whether the LLM answer was closest to the ground truth reference answer, or an alternative response containing a discordant recommendation. Using these human labels as ground truth, we computed the F1 score of each text metric by classifying an example positive if the LLM-reference pair has the highest metric value among all LLM-response pairs.[30–34] We found that BERTScore Precision maximizes agreement with human judgment.

2.5.2. *Subject Matter Expert Reviews*

We recruited 4 PGx experts to perform a granular manual review of a selected subset of short-answer LLM responses. For each question, reviewers were shown a human-written and LLM-generated response in randomized blinded order and asked to rate each answer on a five-point Likert scale along attributes of accuracy (i.e. "This response is clinically/scientifically accurate"), completeness (i.e. "This response contains all of the necessary information to address the question fully"), and safety (i.e. "This answer does not pose any danger to human health or safety). For each question, reviewers were also presented with the relevant CPIC guideline document. Ratings were collected using the open-source Data Annotator for Machine Learning tool[35], which was deployed on an AWS EC2 instance with a public IP address so that expert reviewers from around the country could easily work on the assigned scoring task or quit and return to the task later.

2.6. *Data Analysis and Visualization*

The results of our various scoring approaches were analyzed in a Jupyter notebook with pandas, which is included in the GitHub repository for this project.[27,28,36] All plots were generated using the matplotlib and seaborn python packages.[37,38]

3. Results

3.1. *The PGxQA Question Corpus*

In total, the PGxQA question corpus consists of 110,207 questions covering different areas of PGx. While we subsequently present our own tools for querying and evaluating LLMs using this expansive dataset, we make available the entire set of questions as a resource agnostic of downstream evaluation approach. We detail the question types covered in **Table 1**.

Table 1: Representative examples of PGxQA questions generated from CPIC database or external sources

Question Type	Description	Number of questions	Example Prompt	Expected Response
Allele frequency	Ask for a value indicating the allele frequency of a given allele in a population.	2,548	"What is the average allele frequency of ABCG2 rs2231142 reference (G) in the African American/Afro-Caribbean population? Respond with just a number, rounded to 4 decimal places, with no additional text."	0.9651
Allele definition	Ask for dbSNP IDs for variants that define or are part of a given allele. Note that some alleles consist of multiple SNPs.	901	"What SNPs are in the allele definition for CFTR F1052V? Provide a dbSNP ID (also known as an rsID, starting with rs) when available."	rs150212784
Allele function	Determine how an allele affects the overall function of a gene.	1,111	"What is the allele functionality of CYP2C9 *9? Please select the answer from the following choices: {'Normal function', 'Decreased function', 'Uncertain function', 'No function', 'Unknown function'}, and respond with only your selection."	Normal function
Genes to drugs	Ask for drugs with actionable CPIC guidelines for a given gene. Note that multiple drugs can be listed.	22	"Which drugs have actionable CPIC guidelines for CYP2C19? Please respond with nothing but a list of generic drug names delimited by ';'."	pantoprazole;sertraline;omeprazole;lansoprazole;amitriptyline;citalopram;voriconazole;escitalopram;clopidogrel
Diplotype to phenotype	Ask what the defined pharmacogenetic phenotype is for a given set of alleles in a gene.	101,138	"What is the pharmacogenetic phenotype for CYP2C9 *1/*1? Please select the answer from the following choices: {'Intermediate Metabolizer', 'Normal Metabolizer', 'Poor Metabolizer', 'Indeterminate'}, and respond with only your selection."	Normal Metabolizer
Drugs to genes	Ask what genes a clinician might want to include in a panel given what drug a patient is taking OR what genes have actionable guidelines for certain drugs for an interested researcher. Note that multiple genes can be listed.	79 (each); 158 (total)	Clinician: "I want to give my patient paroxetine. What genes should I include in a pharmacogenetics panel? Please respond with nothing but a list of gene symbols delimited by ';'." Researcher: "What genes have actionable pharmacogenetic guidelines for paroxetine? Please respond with nothing but a list of gene symbols delimited by ';'."	Clinician: CYP2D6 Researcher: CYP2D6
Phenotype to category	Given an individual with a certain allele and a drug, provide a guideline for that phenotype-drug combination if applicable in terms of drug dosing (multiple choice). Note: this is a multiple choice version of "Phenotype to guideline".	2,145	"What would be the clinical guidance for someone who is HLA-B*57:01 negative for HLA-B with regards to taking abacavir? Please respond with just 'Avoid' if the guidance is to avoid the drug or take an alternate drug, 'Alter dose' if the guideline is to raise, lower, or start with a specific dose, or 'Unchanged', if there are no clinical recommendations or there is no deviation from standard care for this phenotype and drug."	Unchanged
Phenotype to guideline	Ask the LLM to, given an individual with a certain allele and a drug, provide a guideline for that allele-drug combination if applicable in a short-answer format (not multiple choice). Note: this is a short answer version of "Phenotype to category".	2,133	"What would be the clinical guidance for someone who is HLA-B*57:01 negative for HLA-B with regards to taking abacavir?"	Use abacavir per standard dosing guidelines
Adversarial questions (refusal)	For the above categories, provide a similar prompt, but with one of the entities (genes, drugs, alleles, etc.) being fabricated or incorrect. A model is expected to refuse to answer.	36	"What SNPs are in the allele definition for QSTG1 reference (C)? Provide a dbSNP ID (also known as an rsID, starting with rs) when available or answer UNKNOWN if unknown."	UNKNOWN
External Questions	Questions provided by one or more external sources, as described in Section 2.3.1. Note that these were all scored manually using expert raters, as described in Section 2.5.2.	15	"My patient underwent a percutaneous coronary intervention (PCI) and I want to prescribe clopidogrel. They had pharmacogenetic testing and are a CYP2C19 rapid metabolizer (*1/*17). Do they need a different dose of clopidogrel from the standard 75 mg daily?"	"Per the current CPIC guidelines, patients who are CYP2C19 poor metabolizers have significantly reduced CYP2C19 activity, and should avoid clopidogrel if possible due to increased risk of adverse cardiac and cerebrovascular events."

3.2. *Automated Performance Metric Results*

3.2.1. *Quantitative or Categorical Responses*

OpenAI's GPT models almost universally performed better than Llama or Gemini on numeric, information retrieval, and multiple-choice query metrics (**Table 2**). In particular, GPT-4o, outperformed or was in second place for nearly every metric. However, overall performance varied widely across question categories, with models performing worse at Allele Definition, Allele Function, Diplotype to Phenotype, and Phenotype to Category questions than the other question categories. Performances of less than 0.5 for most metrics and LLMs indicate that allele-related questions were more likely to lead to incorrect answers, potentially because allele definitions are dependent on contextual information such as genes. This potentially highlights that LLM training data or approaches may not properly encode allele information, particularly if they do not incorporate tabular data like the CPIC allele tables. Additionally, the number of star alleles has grown massively as new variants and combinations of variants are discovered. Limited references to these alleles in scientific literature likely contribute to poor performance, since LLMs primarily draw from natural language and at baseline struggle with tabular data.[39]

In contrast, other categories saw stronger performance such as the "Genes to drugs" or "Drugs to genes" categories, particularly in the average recall of the LLMs in identifying the expected entities. This indicates that entities such as drugs and genes, which have been described in text for much longer, and across a wider variety of sources, may be better encoded within the LLM weights. However, the precision in these categories was lacking for several LLMs, indicating that such LLMs may be prone to so-called "hallucinations" when responding to these questions, or may make claims backed up by inconclusive evidence.

Table 2. Mean scores for each automated question category except for Phenotype to Guideline. The top scoring model for each category is bolded

Question Category	Metric	Llama 3	Gemini Pro 1.5	GPT3.5	GPT4	GPT4o
Allele frequency	Mean Absolute Deviation	0.1178	0.1465	0.1147	0.0601	**0.0561**
Allele definition	Average Precision	0.1443	0.1341	0.1750	**0.2599**	**0.2599**
	Average Recall	**0.2274**	0.1422	0.2107	0.2221	0.2229
Allele function	Accuracy	0.3856	0.3791	0.3333	**0.5033**	0.4771
Genes to drugs	Average Precision	0.2870	0.1364	0.5459	0.4760	**0.6843**
	Average Recall	0.3955	0.1104	**0.6810**	0.6719	0.6300
Diplotype to phenotype	Accuracy	0.3770	0.3455	0.2565	0.3665	**0.4346**
Drugs to genes (clinician)	Average Precision	0.3177	0.1706	0.2169	0.4424	**0.5992**
	Average Recall	0.7679	0.4494	0.7152	0.8481	**0.9367**
Drugs to genes (researcher)	Average Precision	0.4325	0.3430	0.2968	0.5580	**0.8091**
	Average Recall	0.7489	0.5190	0.7278	0.6667	**0.8418**
Phenotype to category	Accuracy	0.4365	0.3538	0.3212	0.4385	**0.5635**
Phenotype to guideline	BERTscore Precision Win rate	0.7056	0.5499	0.7178	**0.7251**	0.7056

3.2.2. *Short Answer Responses*

After comparing each text embedding method to human classification results, the BERTScore Precision metric was the most concordant with human similarity assessments in indicating which of several reference answers the GPT-4o-generated response was the most concordant with (**Figure 1a., Supplementary Table S1**).[30–34] Because this metric seemed the closest to capturing human judgment on a broad scale, we used it as an automated scoring proxy for LLM performance on our short answer "Phenotype to guideline" tests. Based on automated tests, GPT-4-turbo slightly outperformed GPT-3.5-turbo, GPT-4o, and Llama 3 in average win rate as defined in the methods (**Figure 1b.**). However, Gemini-Pro seems to greatly underperform relative to its counterparts, having an average win rate roughly 0.15 lower than the other models, indicating that its answers likely significantly diverged from the other models and from the ground truth reference.

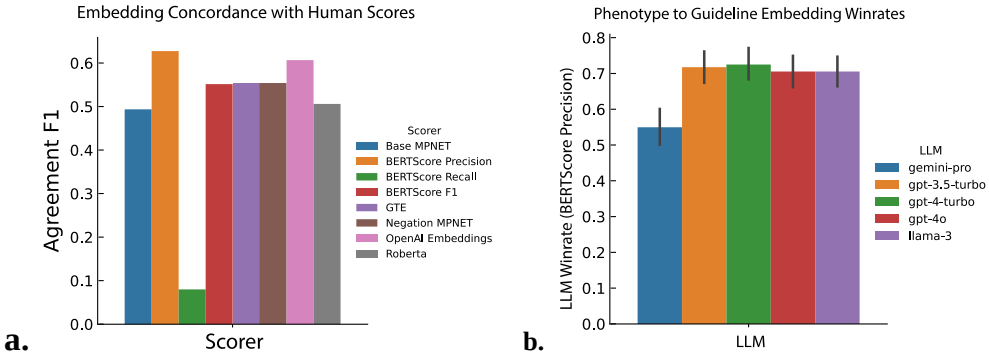

Figure 1: a.) Scorer concordance with human ratings of response similarity as defined by the F1 of the agreement for the Phenotype to Guideline question category for GPT-4o. b.) Model win rates in Phenotype to Guideline Tests.

3.2.3. *Refusal Assessment*

When given the option to refuse to respond, LLMs had highly variable rates of refusal on misspecified and properly specified questions (where misspecified refers to questions where there is not sufficient information to answer, or there exist no clinical guidelines for the requested information). Ideally, a medical chatbot should refuse to answer misspecified questions (a refusal rate of 1 is best) and answer properly specified questions (a refusal rate of 0 is best). Llama, Gemini, and GPT3.5 all refused to answer both types of questions at roughly equal rates. Llama and Gemini tended to refuse very infrequently (<0.2 refusal rate) in either circumstance, while GPT-3.5 refused at roughly equal rates for both circumstances (~0.3 refusal rate) (**Figure 2**). A low refusal rate for misspecified queries might indicate a higher tendency to hallucinate information when given confusing or contradictory queries. In contrast, GPT-4 and GPT-4o showed a higher rate of refusal for misspecified questions (~0.7) compared to properly specified questions (~0.3), indicating that these two models exhibit ability to identify questions with incorrect information as well as a propensity to avoid hallucinations, though there remains significant room for improvement. These results are further broken down in Supplementary Table S2, which shows the refusal rates for different categories.

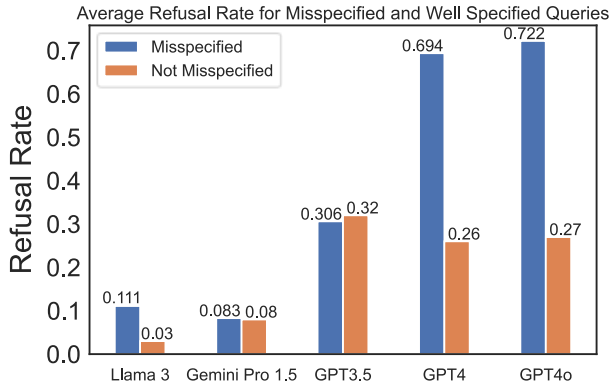

Figure 2: Refusal rates of the different LLMs for misspecified and properly specified question sets.

3.3. *LLM Results with Human Scoring*

3.3.1. *Manual LLM Metrics*

Although the emphasis of this work is on large scale benchmarks that can be employed widely, even in settings where manual expert review would be intractable, it is undeniable that expert reviewers provide invaluable understanding of the nuances and details of PGx which cannot easily be measured by automated scorers and text similarity scores. We recruited 4 PGx experts to manually score a set of GPT-4o responses to 15 short answer questions, and had those same experts score the human-written reference answers. On average, GPT-4o performed lower than the reference answer in all categories, with 'Accuracy' having the largest gap (**Table 3**). While these results reflect that GPT-4o performed well for many questions, there were some answers where it provided highly incorrect or even dangerous responses, such as when it gave incorrect recommendations on tacrolimus PGx in the context of liver transplant.

Table 3. Average Likert scores for Accuracy, Completeness, and Safety of GPT-4o and reference answers as scored by PGx experts

Metric	GPT4o	Reference Answer	Performance Gap
Accuracy (Likert)	3.917	4.917	-1.000
Completeness (Likert)	4.167	4.533	-0.367
Safety (Likert)	4.083	4.850	-0.767

4. Discussion

This work provides a framework and dataset to evaluate LLM-based chatbots in their ability to answer PGx questions derived from gold-standard PGx data sources. In demonstrating our framework, we have highlighted the strengths and weaknesses of LLMs in handling a wide range of PGx queries, providing guidance for future improvements.

4.1. *Avenues for Improving LLMs*

The main limitations we identified in LLM-based chatbots are their especially poor accuracy for queries requesting numeric answers as well as newer or less common star alleles, their tendency to

invent false information instead of refusing to answer unknown queries, and their inability to understand the quality of the underlying sources of their claims. These are broader issues in LLM research, and many techniques have been employed to address them. Prompt-engineering involves devising specific prompts to elicit more comprehensive, more accurate, and better-worded responses from LLMs, which is inexpensive and requires minimal technical expertise, making it highly accessible.[40] However, its ability to enhance results is limited, and excessive engineering can lead to increased token usage per query, potentially raising costs and complexity in processing time.[41,42] This approach was employed in many of the structured answer questions in PGxQA and yielded more concise and readily usable information.

Fine-tuning LLMs on specific datasets of PGx questions, such as those generated in this study, presents an opportunity for models to better understand and respond to domain-specific queries. This approach has been shown to improve the relevance and accuracy of LLM responses. Although fine-tuning can be expensive, requiring significant computational resources like GPUs to train and update the model, it provides a tailored solution for domain-specific prompts.[43] However, fine-tuned models can still hallucinate, as they rely on pre-trained embeddings.[44]

Retrieval augmented generation (RAG) incorporates a retrieval mechanism into LLMs, enabling the model to directly source information from an updated knowledge base. This approach is relatively cheap and straightforward to maintain, as updating the knowledge base is less resource-intensive compared to training the LLM itself.[45] This is ideal for domains such as PGx, where knowledge bases are constantly updated. This also reduces the risk of hallucinations by providing the model with direct access to accurate data sources. However, RAG systems require large context windows for effective querying and a higher degree of human intervention is involved to teach the LLM how to access and utilize these external sources.[44,46]

To address the needs efforts are underway by the PharmGKB/CPIC group at Stanford to create AI-ready data for consumption by LLMs. In addition, collaborative efforts are underway by Dr. Roxana Daneshjou and Dr. Klein's groups at Stanford to develop both clinician-forward and patient-forward tools using generative AI to disseminate this knowledge on the current PharmGKB website and in the future, in the ClinPGx resource.

4.2. *Limitations of PGxQA*

PGxQA is intended to be a framework for initial evaluation of a chatbot in answering PGx questions, particularly in answering questions concordant with pre-existing guidelines (such as information from CPIC, PharmGKB, and others). As shown above, PGxQA provides a variety of metrics that provide insight into several dimensions of the performance of LLMs. However, it is important to recognize that PGxQA has several limitations due to the way that it was devised and developed with a focus on automated assessment. First, the questions in PGxQA are largely created automatically from public PGx data sources. Most questions are query-based—requesting information that would require looking up information from one database and not synthesizing knowledge across multiple databases or fields. This facilitates automated evaluation at the expense of being able to understand this dimension of LLMs, referred to as "multi-hop reasoning". To mitigate this, handcrafted questions and actual questions asked of PGx researchers and clinicians are included through the

“External Questions" category, though LLM responses to these questions cannot fully be assessed automatically.

Our emphasis on automated scoring approaches, while valuable for large-scale evaluation, introduces other limitations as well. We engineered the prompts to instruct the LLM to return answers in our desired format to properly score responses for our information retrieval tasks, introducing a small possibility that asking for results in this strict format alters performance. As shown in the comparison between the clinical and researcher versions of our drug to genes questions, the LLMs do seem to have variable performance when similar questions are asked in different ways. However, this represents a weakness of LLMs that must also be studied prior to clinical use due to the heterogeneous nature of real-life queries. There are also limitations to our text-similarity-based scoring, as text embeddings do not fully capture the nuances of human judgment. Despite these compromises, we believe that PGxQA will still provide useful metrics for chatbot evaluation and we anticipate that future work may address many of the limitations of PGxQA and of LLM chatbots.

4.3. *Future Directions*

Going forward, we expect PGxQA to serve as an automatic evaluation framework to continually evaluate LLMs. This initial evaluation has shown dramatic improvements in performance in more recent models, such as GPT-4o, relative to older iterations such as GPT3.5. We anticipate that further advancements in model architecture and training will strengthen the ability of these models to function as a valuable resource in PGx. Using PGxQA, we can continually monitor improvements in LLM performance and assess new technologies as they are unveiled. The automatic generation of questions from the CPIC database, which is routinely updated, will also ensure that LLMs are updated with the latest information and clinical guidelines. The metrics presented in PGxQA will be continually refined to best reflect the latest evidence. As PGx is a continually evolving area of study, it is essential to have a scalable framework for ongoing evaluation to ensure that model improvements translate into tangible benefits for the field in terms of accuracy and relevance.

The future of PGx chatbots holds significant promise as LLMs become increasingly integrated into healthcare settings to provide clinical recommendations and support. These chatbots will be able to use large quantities of PGx literature and evidence to strengthen and personalize their responses to clinician, patient, and researcher queries. The development of advanced LLMs, coupled with emerging techniques like RAG, will help ensure that PGx chatbots can reliably provide personalized and accurate evidence-based guidance regarding medication intake and dosage. However, the future of these chatbots depends on rigorous continual assessment of their performance. The resources developed in PGxQA represent a first-in-class approach to guide automated LLM evaluation, prioritizing accuracy, completeness, and safety for PGx chatbots.

5. Supplemental Materials and Data Availability

Supplemental tables and the author contributions list are available at:
https://ritchielab.org/publications/supplementary-data/psb-2025/pgxqa
All code, questions, LLM answers, and scoring results are available at:
https://github.com/KarlKeat/PGxQA/

6. References

1. Lewis, J. P. & Shuldiner, A. R. Clopidogrel pharmacogenetics: Beyond candidate genes and genome-wide association studies. *Clin. Pharmacol. Ther.* **101**, 323–325 (2017).
2. Daly, A. K. Pharmacogenetics: a general review on progress to date. *Br. Med. Bull.* **124**, 65–79 (2017).
3. Klein, M. E., Parvez, M. M. & Shin, J.-G. Clinical Implementation of Pharmacogenomics for Personalized Precision Medicine: Barriers and Solutions. *J. Pharm. Sci.* **106**, 2368–2379 (2017).
4. Relling, M. V. & Evans, W. E. Pharmacogenomics in the clinic. *Nature* **526**, 343–350 (2015).
5. Nagy, M., Eirini Tsermpini, E., Siamoglou, S. & Patrinos, G. P. Evaluating the Current Level of Pharmacists' Pharmacogenomics Knowledge and its Impact on Pharmacogenomics Implementation. *Pharmacogenomics* **21**, 1179–1189 (2020).
6. Alshabeeb, M. A., Alyabsi, M., Aziz, M. A. & Abohelaika, S. Pharmacogenes that demonstrate high association evidence according to CPIC, DPWG, and PharmGKB. *Front. Med.* **9**, (2022).
7. Thorn, C. F., Klein, T. E. & Altman, R. B. PharmGKB: The Pharmacogenomics Knowledge Base. in *Pharmacogenomics: Methods and Protocols* (eds. Innocenti, F. & van Schaik, R. H. N.) 311–320 (Humana Press, Totowa, NJ, 2013). doi:10.1007/978-1-62703-435-7_20.
8. Duarte, J. D. & Cavallari, L. H. Pharmacogenetics to guide cardiovascular drug therapy. *Nat. Rev. Cardiol.* **18**, 649–665 (2021).
9. Tuteja, S. Application of Pharmacogenetics for the Use of Antiplatelet and Anticoagulant Drugs. *Curr. Cardiovasc. Risk Rep.* **17**, 27–38 (2023).

10. Green, J. S., O'Brien, T. J., Chiappinelli, V. A. & Harralson, A. F. Pharmacogenomics Instruction in US and Canadian Medical Schools: Implications for Personalized Medicine. *Pharmacogenomics* **11**, 1331–1340 (2010).
11. Karas Kuželički, N. *et al.* Pharmacogenomics Education in Medical and Pharmacy Schools: Conclusions of a Global Survey. *Pharmacogenomics* **20**, 643–657 (2019).
12. Nutter, S. C. & Gálvez-Peralta, M. Pharmacogenomics: From classroom to practice. *Mol. Genet. Genomic Med.* **6**, 307–313 (2018).
13. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. Preprint at https://doi.org/10.48550/arXiv.2303.13375 (2023).
14. Meng, X. *et al.* The application of large language models in medicine: A scoping review. *iScience* **27**, 109713 (2024).
15. Soroush, A. *et al.* Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying. *NEJM AI* **1**, AIdbp2300040 (2024).
16. Clusmann, J. *et al.* The future landscape of large language models in medicine. *Commun. Med.* **3**, 1–8 (2023).
17. Rawte, V., Sheth, A. & Das, A. A Survey of Hallucination in Large Foundation Models. Preprint at http://arxiv.org/abs/2309.05922 (2023).
18. Xu, Z., Jain, S. & Kankanhalli, M. Hallucination is Inevitable: An Innate Limitation of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2401.11817 (2024).
19. Zhang, Y. *et al.* Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2309.01219 (2023).
20. Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and Adoption of Large Language Models in Medicine. *JAMA* **330**, 866–869 (2023).

21. Mousavi, S. M., Alghisi, S. & Riccardi, G. DyKnow:Dynamically Verifying Time-Sensitive Factual Knowledge in LLMs. Preprint at https://doi.org/10.48550/arXiv.2404.08700 (2024).
22. Ullah, E., Parwani, A., Baig, M. M. & Singh, R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review. *Diagn. Pathol.* **19**, 43 (2024).
23. Kanter, G. P. & Packel, E. A. Health Care Privacy Risks of AI Chatbots. *JAMA* **330**, 311–312 (2023).
24. Murugan, M. *et al.* Empowering personalized pharmacogenomics with generative AI solutions. *J. Am. Med. Inform. Assoc.* **31**, 1356–1366 (2024).
25. cpicpgx/cpic-data. CPIC (2024).
26. psycopg/psycopg2. The Psycopg Team (2024).
27. team, T. pandas development. pandas-dev/pandas: Pandas. Zenodo https://doi.org/10.5281/zenodo.10957263 (2024).
28. McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* 56–61 (2010) doi:10.25080/Majora-92bf1922-00a.
29. Generative Language API | Google AI for Developers. *Google for Developers* https://ai.google.dev/api/rest.
30. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: Evaluating Text Generation with BERT. Preprint at https://doi.org/10.48550/arXiv.1904.09675 (2020).
31. He, P., Liu, X., Gao, J. & Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. Preprint at https://doi.org/10.48550/arXiv.2006.03654 (2021).
32. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. Preprint at https://doi.org/10.48550/arXiv.1907.11692 (2019).

33. Li, Z. *et al.* Towards General Text Embeddings with Multi-stage Contrastive Learning. Preprint at https://doi.org/10.48550/arXiv.2308.03281 (2023).
34. Song, K., Tan, X., Qin, T., Lu, J. & Liu, T.-Y. MPNet: Masked and Permuted Pre-training for Language Understanding. Preprint at https://doi.org/10.48550/arXiv.2004.09297 (2020).
35. vmware/data-annotator-for-machine-learning. VMware (2024).
36. Kluyver, T. *et al. Jupyter Notebooks—a Publishing Format for Reproducible Computational Workflows*. *IOS Press* 87–90 (2016). doi:10.3233/978-1-61499-649-1-87.
37. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
38. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
39. Sui, Y., Zhou, M., Zhou, M., Han, S. & Zhang, D. Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study. in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* 645–654 (Association for Computing Machinery, New York, NY, USA, 2024). doi:10.1145/3616855.3635752.
40. Chen, B., Zhang, Z., Langrené, N. & Zhu, S. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *arXiv.org* https://arxiv.org/abs/2310.14735v4 (2023).
41. Patel, D. *et al.* The Limits of Prompt Engineering in Medical Problem-Solving: A Comparative Analysis with ChatGPT on calculation based USMLE Medical Questions. 2023.08.06.23293710 Preprint at https://doi.org/10.1101/2023.08.06.23293710 (2023).
42. shanepeckham. Getting started with LLM prompt engineering. https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/prompt-engineering (2024).

43. Finetuning in large language models. https://blogs.oracle.com/ai-and-datascience/post/finetuning-in-large-language-models.
44. Alghisi, S., Rizzoli, M., Roccabruna, G., Mousavi, S. M. & Riccardi, G. Should We Fine-Tune or RAG? Evaluating Different Techniques to Adapt LLMs for Dialogue. Preprint at https://doi.org/10.48550/arXiv.2406.06399 (2024).
45. Lewis, P. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. in *Advances in Neural Information Processing Systems* vol. 33 9459–9474 (Curran Associates, Inc., 2020).
46. Zhao, P. *et al.* Retrieval-Augmented Generation for AI-Generated Content: A Survey. Preprint at http://arxiv.org/abs/2402.19473 (2024).