

# LLM on FHIR - Demystifying Health Records

Paul Schmiedmayer<sup>1\*</sup>, Adrit Rao<sup>1</sup>, Philipp Zagar<sup>1</sup>,  
Vishnu Ravi<sup>1</sup>, Aydin Zahedivash<sup>1, 2</sup>, Arash Fereydooni<sup>1</sup>,  
Oliver Aalami<sup>1, 3</sup>

<sup>1</sup>Byers Center for Biodesign, Stanford University, 318 Campus Drive,  
Stanford, 94305, CA, USA.

<sup>2</sup>Department of Pediatrics, Stanford University, 453 Quarry Road,  
Stanford, 94305, CA, USA.

<sup>3</sup>Division of Vascular and Endovascular Surgery, Department of Surgery,  
Stanford University, 780 Welch Road, Stanford, 94305, CA, USA.

\*Corresponding author(s). E-mail(s): [schmiedmayer@stanford.edu](mailto:schmiedmayer@stanford.edu);

**Objective:** To enhance health literacy and accessibility of health information for a diverse patient population by developing a patient-centered artificial intelligence (AI) solution using large language models (LLMs) and Fast Healthcare Interoperability Resources (FHIR) application programming interfaces (APIs).

**Materials and Methods:** The research involved developing LLM on FHIR, an open-source mobile application allowing users to interact with their health records using LLMs. The app is built on Stanford's Spezi ecosystem and uses OpenAI's GPT-4. A pilot study was conducted with the SyntheticMass patient dataset and evaluated by medical experts to assess the app's effectiveness in increasing health literacy. The evaluation focused on the accuracy, relevance, and understandability of the LLM's responses to common patient questions.

**Results:** LLM on FHIR demonstrated varying but generally high degrees of accuracy and relevance in providing understandable health information to patients. The app effectively translated medical data into patient-friendly language and was able to adapt its responses to different patient profiles. However, challenges included variability in LLM responses and the need for precise filtering of health data.

**Discussion and Conclusion:** LLMs offer significant potential in improving health literacy and making health records more accessible. LLM on FHIR, as a pioneering application in this field, demonstrates the feasibility and challenges of integrating LLMs into patient care. While promising, the implementation and pilot also highlight risks such as inconsistent responses and the importance of replicable output. Future directions include better resource identification mechanisms and executing LLMs on-device to enhance privacy and reduce costs.

## 1 Background and Significance

The US Government's Healthy People 2030 initiative's overarching goal is to "*eliminate health disparities, achieve health equity, and attain health literacy to improve the health and well-being of all.*" [15] The initiative promotes health literacy as a key goal to improve health and well-being over the next decade [15]. Initiatives like these underscore the profound impact enhanced health literacy could have on individuals' overall health and well-being over the next decade. Health literacy, as defined by the Health Resources & Services Administration, extends beyond mere access to health information; it encompasses the critical ability to find, comprehend, and effectively utilize health-related information [9].

The introduction of HL7 Fast Healthcare Interoperability Resources (FHIR) APIs and the enactment of the 21st Century Cures Act mark significant milestones in this journey [1]. These patient-facing APIs enable individuals to have access to their electronic health records stored across care providers and have the potential to revolutionize how individuals interact with their electronic health records, laying the groundwork for a new patient empowerment and engagement era. Focusing on understanding and acting on these insights is the crucial next step to positively impact patients' health, improve health outcomes, reduce clinical workloads, and democratize access to health data [8]. The emerging field of mobile health data further amplifies these opportunities and challenges, offering a rich source of real-time, personalized health insights while intensifying concerns over data management, privacy, and security [11, 26]. Nevertheless, despite all these advancements, initiatives to provide access to health records infused with verbose medical jargon and uninterpreted mobile health data have not yet led to a giant leap toward transforming patient health outcomes.

Given that limited English proficiency has been implicated in poor health outcomes in the US [24], fostering understandability and equal access to health data without hurdles such as language barriers, medical knowledge [4], and limited access is essential to improve health equity and social impact. Our project proposes making health literacy radically accessible, understandable, and actionable, particularly to under-resourced communities. The advent of large language models (LLMs) [28] in healthcare holds tremendous potential to democratize health literacy by augmenting human knowledge and capabilities. Large language models like OpenAI's GPTs [5, 17] can address these challenges and unlock improved comprehension and engagement with patient access to electronic health records and personal health data. We hypothesize that LLMs can significantly elevate health literacy by enhancing human knowledge and skills. While significant research is conducted to support physician and electronic health record system integrations, we are taking an essential human-centered approach: By proposing patient-centered, privacy-preserving AI solutions, we aim to transform how health information is accessed and understood, catering to a diverse patient population. This represents a pivotal shift in healthcare, with the power to reduce the strain on medical professionals and place patients at the helm of their health journey [14]. We see this approach as a crucial cornerstone with immediate clinical impact, reducing workloads and enabling focus for medical professionals while empowering patients to be the steward of their health journey.

Various studies have explored the use of LLMs in healthcare. These include aiding clinicians with decision-making and documentation [25], and enhancing medical education [7]. Additionally, LLMs show promise in patient-facing applications, particularly in answering medical questions. Preliminary validations have assessed the efficacy of LLMs such as GPT in this context [12, 16]. Integrating these models into the mobile health ecosystem from a digital health perspective can have the potential to improve medical literacy and understanding of one’s health. Patients often struggle to fully understand their health information due to the vast amount of data, compounded by language barriers and widespread lack of medical literacy [10, 31]. Incorporating LLMs tailored to one’s personal Fast Healthcare Interoperability Resources (FHIR) records, within a mobile app, has the potential to increase the accessibility and understanding of one’s health. Our software, LLM on FHIR (LLM on FHIR) [21], seeks to develop solutions that address these issues, paving the way for responsible, secure, and effective use of LLMs and mobile health data in healthcare as a critical step in enabling the next generation of digital health advancements. We demonstrate the development and considerations of this patient-facing LLM-based tool and highlight its applicability to make health records available to patients while evaluating its efficacy and potential challenges using sophisticated expert reviews. Our pilot study focuses on an automatically synthesized user cohort of simulated FHIR resources for patients with cardiovascular diseases and expert reviews of common questions asked by patients.

## 2 Materials and Methods

The software developed as part of this research project is an open-source application licensed under the MIT license<sup>1</sup> and builds on the foundational open-source *Stanford Spezi* ecosystem [22] and GPT-4 [17] in its gpt-4-1106-preview<sup>2</sup> version. The evaluation in a pilot study of the LLM-based solution relies on the SyntheticMass patient and population health dataset generated by the Synthea™ generator [29].

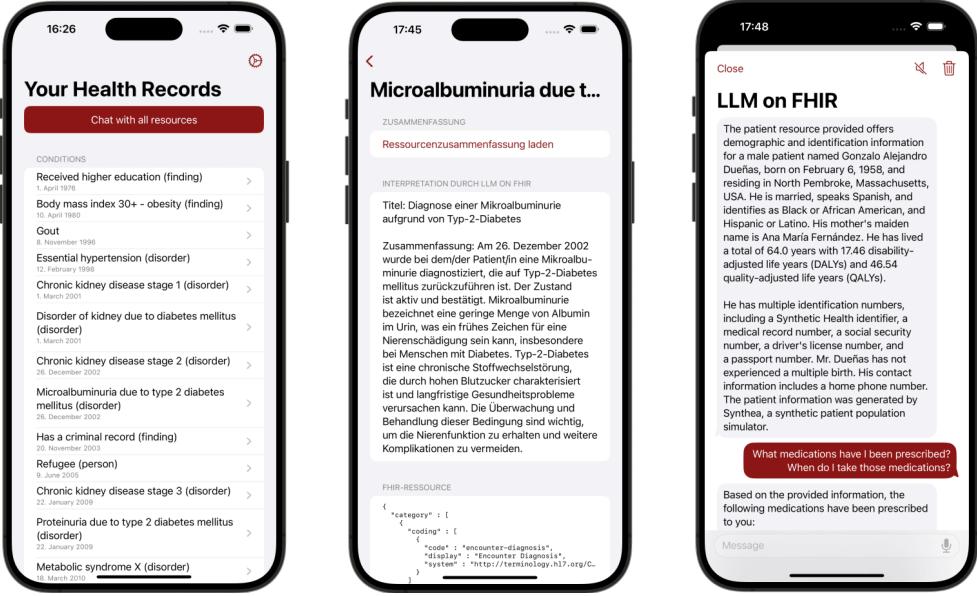
### 2.1 LLM on FHIR

LLM on FHIR is an open-source Swift-based iOS application that demonstrates the ability to interactively converse with one’s FHIR records using LLMs [21]. The LLM on FHIR app enables users to ask questions to an LLM, which then generates responses based on the user’s personal health records retrieved from an electronic health record (EHR). The app uses the open-source Stanford Spezi ecosystem [22], the successor of CardinalKit [2], allowing the modular component-based development of digital health applications. The Stanford Spezi [22] ecosystem is a collection of software subsystems formulating modules that enable reusability and high-quality mobile computing applications, specifically targeting the digital health research and translation ecosystem.

---

<sup>1</sup><https://github.com/StanfordBDHG/LLMonFHIR/blob/main/LICENSE.md>

<sup>2</sup>OpenAI GPT-4 Turbo: <https://openai.com/blog/new-models-and-developer-products-announced-at-devday>



(a) Overview of all available FHIR records

(b) Interpretation and Summary of a single resource

(c) Interactive Chat with all FHIR health records

**Fig. 1:** Screenshots of the LLM on FHIR iOS Application

### 2.1.1 FHIR Resources Data Flow

Enabled by *SpeziHealthKit* [19], LLM on FHIR connects with the Apple Health app using HealthKit<sup>3</sup> to obtain users' FHIR-encoded health records across various hospitals and institutions. The Apple Health app manages authentication and data fetching, simplifying the development of patient-facing solutions accessing health records retrieved from the EHR. In addition, LLM on FHIR provides a mechanism to load patient resources from a FHIR bundle used for testing and validating the application's functionality.

Once health records are delivered to the application, LLM on FHIR enables users to interact with individual resources of their interest and to inspect the underlying FHIR data in detail (Figure 1a). Users can tap on any specific health record to navigate to a view providing a summary and interpretation in a particular record using LLMs. The GPT-4 interaction was prompt-engineered to provide short, concise summaries and longer-form interpretations based on the FHIR data in JSON format for the single resource (shown in Figure 1b) to help users understand the essence of the records. The *SpeziLLM* module [20] allows developers to interface with LLMs in mobile applications ranging from cloud services to the execution of smaller models in a local execution environment.

---

<sup>3</sup>Apple HealthKit: <https://developer.apple.com/documentation/healthkit>

### 2.1.2 All FHIR Resources Chat

The key health literacy enabler and element validated in [section 3](#) is a general interface allowing users to interact with their health records using an LLM-based chat interface (as shown in [Figure 1c](#)). To further address the challenges posed by health literacy and improve the accessibility of LLM on FHIR, voice and translation capabilities have been integrated into the application. Voice-based interaction is facilitated across all chat interfaces within LLM on FHIR, using the *SpeziSpeech* module [23] from the Spezi ecosystem. To allow for the use of different languages in LLM on FHIR, we ensure that the LLM has access to the user's locale and can respond to questions in the user's preferred language. We enable full app translations in English, Spanish, Chinese, German, and French that translate all titles, descriptions, and the onboarding flow.

The LLM-based interactions are facilitated using a LLM function-calling mechanism, allowing the LLM to request and retrieve information about specific health records by choosing from a list of predefined identities. The identifiers are made up of a triplet of information detailed in the LLM function call description and feature the (1) resource type, e.g., a medication request or observation, (2) display name of the resource computed based on the resource type, e.g., the title of the medication for a medication request, and (3) date that best describes the FHIR resource such as the date of an observation or start date of a medication request. The LLM system prompt starting the conversation and introducing the function call mechanism can be found in [Figure 2](#).

Health resources are pre-filtered to avoid exceeding the LLM context limit and to reduce the number of choices for the model to a reasonable size. For medications, only active and outpatient medications are included in the function call considerations. Resources like observations, lab values, and conditions are filtered to reduce them to the most recent occurrence of each type of resource, focusing on the patient's need to identify and explain recent occurrences and procedures. The patient record providing detailed information about a patient is injected at the beginning of each LLM interaction, ensuring that the model understands the rough patient context adequately. The model can use the function calling mechanism at any time to request information about a specific resource or a type of resource. FHIR resources are subsequently queried and processed using the summarization mechanism (as described in [subsubsection 2.1.1](#)) to create a representative text of fewer than 100 words that are then fed into the larger LLM chat context forming the interaction with the patient.

## 2.2 Methods

To evaluate the ability of LLM on FHIR to increase health literacy in a pilot study, we asked a group of four Medical Doctors at the Stanford School of Medicine to judge the responses from the LLM in the interactive chat mode with all available health records (see [subsubsection 2.1.2](#)). We selected a cohort of six exemplified FHIR patient data sets from the *SyntheticMass dataset (Version 2)* generated by *Synthea* [29] to allow for a reproducible and unbiased evaluation of the responses from the Language Model. SyntheticMass has been used for many preliminary machine learning model evaluations [6, 18, 30]. To select representative patients, the entire dataset was partitioned

You are the LLM on FHIR application. Your task is to interpret FHIR resources from the user's clinical records.

Throughout the conversation with the user, use the "get\_resources" function to obtain the FHIR health resources necessary to answer the user's question properly. For example, if the user asks about their allergies, you must use the "get\_resources" function to output the FHIR resource titles for allergy records so you can then use them to answer the question. The end goal is to answer the user's question in the best way possible while taking the FHIR resources obtained using "get\_resources" into consideration. Leave out any technical details like JSON, FHIR resources, and other implementation details of the underlying data resources.

Only request relevant resources and focus on recent resources. Try to reduce the number of requested resources to a reasonable scope.

E.g., if the user asks about their Medication, it would be recommended to request all MedicationRequest FHIR resource types.

Interpret the resources by explaining the data relevant to the user's health. Explain the relevant medical context in a language understandable by a user who is not a medical professional, ideally at a 5th-grade reading level. You should provide factual and precise information in a compact summary in short responses.

The patient resource is provided as the initial context. Ensure that you leave out anything like SSN, passport number, and telephone number.

Do not introduce yourself at the beginning, and immediately return a summary of the user based on the FHIR patient resources.

Do not request any FHIR resources using the "get\_resources" function for your first message. The initial summary should be short and simple; stay at a high level. End with a question asking the user if they have any questions. Make sure your response is in the same language the user writes to you in, not the one stated in the patient resource. The tense should be present.

**Fig. 2:** Main multiple resources prompt provided as default in the LLM on FHIR application. The prompt instructs the model to use the provided LLM function calling mechanisms and establishes a base expectation of the interaction with the user.

into ten individual buckets consisting of a unique cardiovascular condition/procedure, given that cardiovascular conditions were the specialty of most of our expert reviewers. After partitioning the entire dataset and excluding patients that did not fit the criteria, a balanced set of patients was selected by ensuring that all patients were still alive, at least two patients had allergies, and that the selected patients resulted in a representative balance of self-reported gender, ethnic background, and age groups ranging from young patients (8 years old) to elderly patients (the most senior patient is 82 years old). The list of the selected patient data set and a summary of their corresponding conditions, allergies, and medications can be found in [Table 1](#).

Each patient can be selected in the LLM on FHIR application, which loads all their FHIR resources in the context of the application using the FHIR bundle loading

Name	Sex	Age	Conditions	Allergies	Medications
Beatris270 Bogan287	F	8 years	Aortic valve stenosis (disorder) Perennial allergic rhinitis Atopic dermatitis	Latex (substance) Bee venom (substance) Mold (organism) House dust mite (organism) Animal dander (substance) Grass pollen (substance) Tree pollen (substance) Aspirin	Fexofenadine hydrochloride 30 MG Oral Tablet Epinephrine 1 MG/ML Auto-Injector 0.3 ML
Milton509 Ortiz186	M	26 years	Hypertension Hypertension (disorder) Stress (finding)	-	MedicationRequest amLODIPine 2.5 MG Oral Tablet
Edythe31 McDer- mott739	F	49 years	Body mass index 30+ - obesity (finding) Received higher education (finding) Prediabetes Anemia (disorder) Victim of intimate partner abuse (finding) Cardiac Arrest History of cardiac arrest (situation)	-	Jolivette 28 Day Pack
Gonzalo160 Dueñas839	M	65 years	Body mass index 30+ - obesity (finding) Gout - Erosive hypothyroidism (disorder) Disorder of kidney due to diabetes mellitus (disorder) Microalbuminuria due to type 2 diabetes mellitus (disorder) Proteinuria due to type 2 diabetes mellitus (disorder) Metabolic syndrome X (disorder) Prediabetes Anemia (disorder) Ischemic heart disease (disorder) Abnormal findings diagnostic imaging- heart+coronary circulatory (finding) History of heart transplant (situation) Metabolic syndrome (situation)	-	Simvastatin 20 MG Oral Tablet Vitamin B12 5 MG/ML Injectable Solution Clopidogrel 75 MG Oral Tablet Hydrochlorothiazide 25 MG Oral Tablet amLODIPine 2.5 MG Oral Tablet Metoprolol succinate 100 MG Extended Release Oral Tablet Insulin isophane, human 70 UNIT/ML / insulin, regular, human 30 UNIT/ML Injectable Suspension [Humalog] Nitroglycerin 0.4 MG/ACTUAT Mucosal Spray Tacrolimus 1 MG 24 HR Extended Release Oral Tablet
Jacklyn830 Veum823	F	72 years	Essential hypertension (disorder) Miscarriage in first trimester Ischemic heart disease (disorder) Chronic kidney disease stage 3 (disorder) Proteinuria due to type 2 diabetes mellitus (disorder) Social isolation (finding) Sprain (morphologic abnormality)	-	Nitroglycerin 0.4 MG/ACTUAT Mucosal Spray Simvastatin 20MG Oral Tablet Clopidogrel 75 MG Oral Tablet 24 HR metoprolol succinate 100 MG Extended Release Oral Tablet Acetaminophen 325 MG Oral Tablet Hydrochlorothiazide 25 MG Oral Tablet
Allen332 Ferry570	M	82 years	Chronic sinusitis (disorder) Hypertension Served in armed forces (finding) Received high education (finding) Body mass index 30+ - obesity (finding) Prediabetes Anemia (disorder) Opioid abuse (disorder) Atrial Fibrillation Neoplasm of prostate Carcinoma in situ of prostate (disorder) Chronic intractable migraine without aura Victim of intimate partner abuse (finding) Stress (finding) Alzheimer's disease (disorder))	Animal dander (substance) Penicillin V Peanut (substance)	Galantamine 4 MG Oral Tablet Warfarin Sodium 5 MG Oral Tablet doxycycline hyclate 100 MG 1 ML DOCETaxel 20 MG/ML Injection 0.25 ML Leuprolide Acetate 30 MG/ML Prefilled Syringe lisinopril 10 MG Oral Tablet Verapamil Hydrochloride 40 MG Digoxin 0.125 MG Oral Tablet

**Table 1:** Summary of the six selected exemplified FHIR patient data sets from the *SyntheticMass dataset* generated by *Synthea*. We use the *SyntheticMass dataset* (Version 2) that consists of one million synthetic FHIR patient medical records from a simulated population in Massachusetts, USA.

functionality. Each reviewer selected the *all resources chat* functionality, clears out the current context, and subsequently asks the typical patient questions or questions that were determined based on the available information exposed by FHIR APIs (**Table 2**). In total, we have evaluated 168 LLM responses (6 patients x 7 questions x 4 repetitions) equally distributed to all expert reviewers while considering and explicitly inspecting the variations between the responses. Some of the questions build on top of each other while investigating the capabilities of LLMs in a patient-facing solution to foster health literacy ranked in increasing complexity and challenge for the LLM.

Each answer is scored on a Likert score between 1 (Strongly Disagree) and 5 (Strongly Agree), asking four expert medical doctors if they agree that the response was factual and accurate based on the content provided by the FHIR resources that were manually inspected and used when evaluating the responses of the LLM. In addition to accuracy, the four experts also evaluated the responses in relevance to the question and information available to the model using the same scale. The third aspect determined the understandability of the question for the target audience of a patient with no in-depth medical knowledge and their provided age. To judge understandability, experts investigated the usage of medical terms that might not be known

ID	Question
Q1	What are my current medications and how should I be taking them?
Q2	What are the most common side effects for each medication I am taking?
Q3	Am I allergic to any of my medications?
Q4	Can you summarize my current medical conditions?
Q5	What are the health behaviors I should be incorporating into my daily routine to help with my conditions?
Q6	Can you summarize my current medical conditions in German?
Q7	What are my recent laboratory values, what do they mean, and how can I improve them?

**Table 2:** Questions that are sequentially asked to the LLM on FHIR application to validate the responses' accuracy, understandably, and relevance.

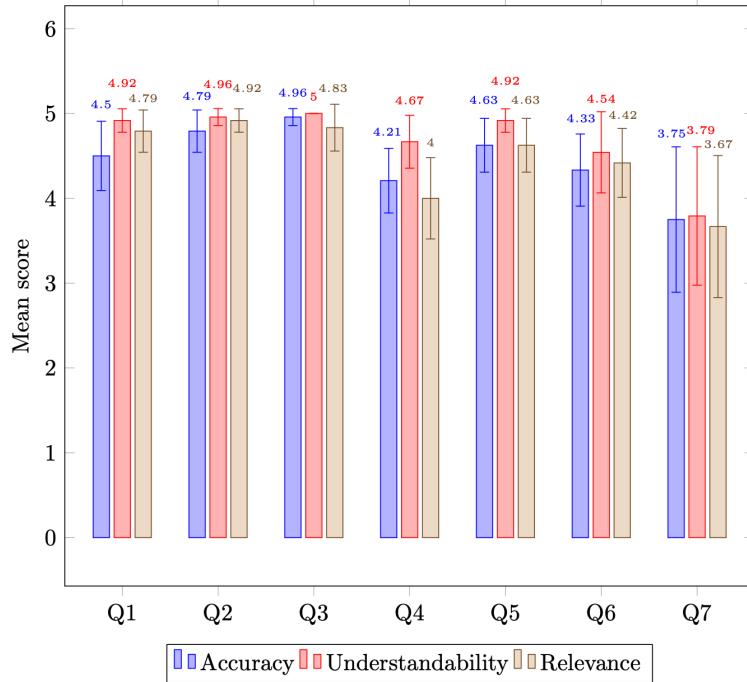
by patients and the quality of the instructions provided by the LLM for any suggestions or interpretations. The answers to the translation question (Q6) were inspected by three German native speakers and translated using Google Translate to enable the medical experts to inspect the factual accuracy of the translated content.

An explicit limitation of the setup is that every expert interacts individually with the Language Model, resulting in slightly different answers from the model and somewhat deviating chat histories. Each interaction with the LLM was recorded using a screen recording. The tests were automated using Apple's user interface testing framework and are provided in the open-source implementation to enable full reproducibility of the results. The automation interaction was designed to ensure that the LLM chat and context were cleared before asking the first question and that questions were always asked in the same sequential order (Q1 to Q7).

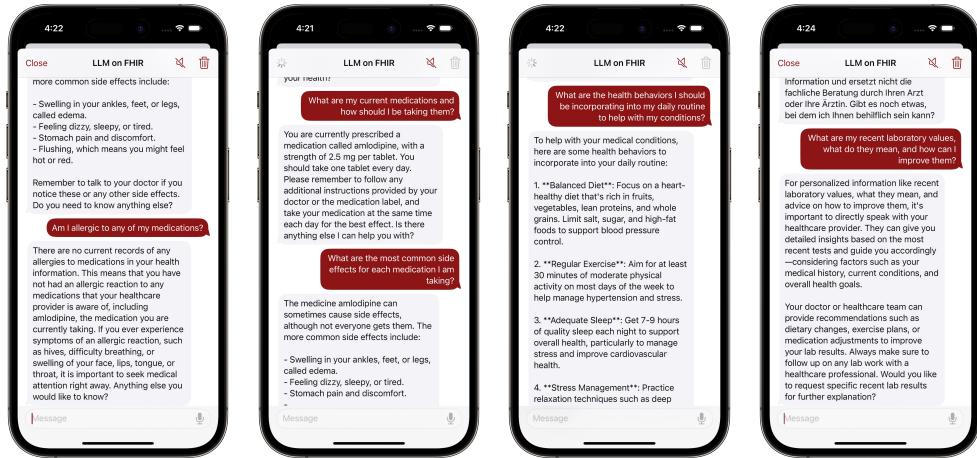
### 3 Results

The expert reviews demonstrated varying but generally high degrees of accuracy and relevance in providing understandable health information to patients. The accuracy, understandably, and average relevance scores (see [subsection 2.2](#)) and their standard deviations are displayed in [Figure 3](#). The overall expert perception was that the app effectively translated medical data into patient-friendly language and was able to adapt its responses to different patient profiles. However, challenges included variability in LLM responses and the need for precise filtering of health data.

Across the board, most experts scored the responses as accurate and especially understandable, highlighting the ability of the model to transform complex concepts into more concise and expressible patient-facing summaries. Expert reviewers that deducted points in the scoring highlighted that the output of the model sometimes misses information included in the FHIR resources and observed that the model did not request all necessary resources using the function-calling mechanism.



**Fig. 3:** Scoring of LLM on FHIR answers by expert reviewers displaying the average response (added as an annotation) as well as the standard deviation for each question and subset of scores.



**Fig. 4:** Outputs of LLM on FHIR judged by expert reviewers as part of the pilot study.

Questions 1, 2, and 3, focusing on medications, generally resulted in high ratings from the reviewers, highlighting the tone and understanding of the responses as well as correct context as demonstrated in [Figure 4a](#) and [Figure 4b](#). The answers were seen as very clear and concise, providing a well-reasoned explanation for each medication, including its dosage and route for most medications. Additionally, the LLM demonstrated caution with certain medications, such as insulin, recommending adherence to prescribed instructions. The LLM always identified if the patient had no allergies and clearly reported this in the chat interface. The main expert feedback stemmed from missing information about dosages and routes taken in some answers, one misassociation where the model reported that insulin is used for prediabetes despite the patient having a type 2 diabetes diagnosis, and instances of rare side effects reported as common.

In investigating questions 4, 5, and 6, expert reviewers deducted points where the model provided out-of-context information, including conditions such as education status and refugee status which would traditionally fall under a social history as opposed to a medical history. The LLM's understanding of recent conditions and resulting implications on current behavior were criticized by some experts as not relevant and intermixed in irrelevant information or too generic advice as seen in [Figure 4c](#). We performed a sub-analysis of Q4 (summarizing health conditions) to assess variability in answers provided by the model to the same question asked by the four different reviewers. Never were two responses exactly the same. Omission of conditions was noted in an estimated 20% of the responses. Patterns of omission were not consistent. No obvious or significant hallucinations were noted.

Question 6, focusing on translating information into another language, was generally perceived as correct and understandable for native German speakers while being close to the English answer to previous questions and lacking the typical nuances of interaction we perceived in English, e.g., adjusting the responses to the age and background of the patient (see [section 4](#)). In addition, in one instance, the model only translated condition names and lost the remainder of the context.

The most significant variation of LLM-performance was perceived in question 7, mostly stemming from the challenges of the model to retrieve insightful and applicable lab results from the FHIR resources using the function-calling mechanism as shown in [Figure 4d](#). Due to the missing context, several answers have neither been rated as accurate nor understandable as most of them did not provide any insightful information. When the LLM correctly identified lab values, expert reviewers perceived most of the summaries as accurate and ranges for lab values and interpretations as appropriate and valuable insights, while actionable insights were not always provided. Ranges provided by the model were mostly perceived as accurate, while interpretations of these values in correlation to the lab results were sometimes too strict, e.g., marking values as too high when they were only slightly out of range. In one instance, the model was noted to avoid answering the question completely and referred the patient to their healthcare provider.

## 4 Discussion

LLM on FHIR presents a novel approach to address the challenges of health literacy and the accessibility of health data for patients. The application showcases the potential to augment clinicians' correspondence with patients regarding straightforward and frequently asked questions utilizing patient-facing FHIR application programming interfaces (APIs) facilitated by the 21st Century Cures Act [1]. By leveraging foundational LLMs like GPT-4 and mechanisms to feed external information into the LLM context, such as function calling, LLMs can be used to aggregate and transform information stored in the EHR and present it in a more accessible manner. The opportunities become especially apparent when observing LLMs interacting and interfacing with a wide range of patients in different age ranges, socioeconomic backgrounds, and levels of health literacy who might otherwise find it challenging to understand their health data and records. The pilot study outlined in section 3 demonstrates the app's ability to translate complex information and provide context in a personalized and patient-appropriate setting. To our knowledge, this is the first application using LLMs in a patient-facing application to explain and contextualize health records retrieved using FHIR APIs.

At the same time, using LLMs in this context presents considerable risks and challenges that need to be investigated and addressed when allowing patients to interface with their health records. This includes the challenge of reproducible answers and behavior when generating content using LLMs. While hallucinations were not common in our evaluation, we observed a noticeable variation in requests for patient resources for the same questions, resulting in different responses, sometimes excluding insightful patient information. We aim to build an ecosystem of open-source tools interfacing with different LLM providers and services addressing patient health literacy and health data processing challenges. For example, the infrastructure could be used to automatically fill out intake forms or enroll patients in clinical trials using patient information processed by LLMs.

As outlined in section 3, the LLM has varying, but generally high, degrees of accuracy and relevance when answering common questions based on a wide variety of patient profiles representing a typical subset of patients with multiple comorbidities. Observing the variability of the LLM's responses demonstrated the possibilities of a new mechanism of human-computer interaction (HCI) by introducing nuances and context awareness that were previously challenging to achieve and hard to replicate across various domains. The LLM references patients in the responses and adjusts itself to a patient, as demonstrated in conversations with the 8-year-old patient. The model starts to use more straightforward and child-friendly language in explaining complex medical information, even referring to parents or other adults within the generated outputs.

A key learning from building, using, and validating the application of LLMs in the health literacy context is the necessity to strictly filter and automatically preprocess any data fed to the LLM. This challenge arises due to the limited context size and the hard-to-predict reasoning of requesting resources using function calls. We continuously iterated over the process while developing and testing the application with synthetic patient data. We uncovered good mechanisms to calculate identifiers for

patient resources to provide enough context to the model, such as the triplet of information described in subsection 2.1. In addition, we developed a mechanism to remove outdated medications and older irrelevant observations that have been replaced by more recent insights and FHIR resources. As noted in the results, we still observed that irrelevant observations and conditions, such as social history facts, were surfaced by the model (e.g., criminal record, higher education, refugee status, need for medication review, and full-time employment). These data points can be present in FHIR record bundles but should not surface in a patient-facing context outlining the medical conditions. In addition, the definition of *recent* conditions was often perceived as inaccurate, highlighting the challenges of the LLM of understanding and processing temporal correlations.

The perceived challenge of variability within the generated answers can be addressed by reducing the LLM output temperature setting and adjusting other model parameters to fit better the individual context of the question in the patient correspondence. In addition, initial seed definitions and other LLM replicability efforts under active research are essential and become more relevant for medical applications. Recent advancements by commercial offerings like GPT-4, such as reproducible outputs<sup>4</sup> will facilitate the applicability and replicability of these models in a patient context. The eventual first adoption of these tools in a patient-facing context will require a manual review by clinicians. We hypothesize it will gradually move to a more automated input and output risk assessment. Additionally, we expect an expansion in the abilities of generative machine learning (ML) models, enabling them to handle medical correspondences independently without needing extra supervision.

## 5 Conclusion

LLMs provide tremendous opportunities for transforming, summarizing, making health records more accessible, and improving health literacy. The LLM on FHIR application demonstrates this promise while highlighting the current shortcomings and next steps in integrating LLMs in clinical workflows and fitting the more remote patient care paradigm observed in the recent decade. Automating typical patient interactions while safeguarding them from misinformation, hallucinations, and irrelevant information is the pathway to better, more scalable patient care. By providing the application as an open-source project and sharing our results, we aim to ensure that other researchers can collaborate on a set of best practices, engineered prompts, and function-calling mechanisms establishing LLM interactions with FHIR resources. The core infrastructure of LLM on FHIR is provided as a reusable component using SpeziLLM and SpeziFHIR, enabling developers to utilize it to prototype and validate their research results.

We intend to keep enhancing LLM on FHIR to democratize health data understandability and accessibility. Nevertheless, given the sensitive nature of health data, we aim to shift the LLM execution environment from opaque, proprietary, and centralized cloud providers like OpenAI closer to the patient’s device. We hypothesize that executing open-source LLMs [27] in more trusted environments, like the patient’s

---

<sup>4</sup><https://platform.openai.com/docs/guides/text-generation/reproducible-outputs>

edge device, will help to mitigate privacy, trust, and financial challenges posed by cloud-based LLMs [3, 13, 32]. As an example, summarizations in LLM on FHIR, as described in subsection 2.1, can be performed by small, on-device models like Llama 2, which then serve as the input for more challenging interpretation tasks performed in the cloud-layer by more extensive and capable models.

Committing to the open-source and trust-building mechanism of collaborative software development, we plan to extend the LLM capabilities in the Stanford Spezi ecosystem and facilitate the sharing of best practices and reusability in accessing EHR data using FHIR APIs to enable the scalability and usage of patient-empowering digital solutions in healthcare.

## Acknowledgments

We thank the Stanford Byers Center for Biodesign for supporting this project and our digital health research.

## Author Contributions

Contributor Roles Taxonomy (CRediT) author statement: **Paul Schmiedmayer**: Conceptualization, Software, Writing - (Original Draft, Review & Editing), Supervision; **Adrit Rao**: Conceptualization, Software, Writing - (Original Draft, Review & Editing); **Philipp Zagar**: Software, Writing - (Original Draft, Review & Editing); **Vishnu Ravi**: Software, Validation, Writing - Review & Editing; **Aydin Zahe-divash**: Validation, Writing - Review & Editing; **Arash Fereydooni**: Validation, Writing - Review & Editing; **Oliver Aalami**: Conceptualization, Validation, Writing - (Original Draft, Review & Editing), Supervision.

## Declarations

None of the authors have a competing interest. The LLM on FHIR application and the Stanford Spezi ecosystem are open-source and licensed using the MIT license. Stanford Spezi is being used to foster a digital health ecosystem and teach the next generation of digital health leaders.

## Data Availability

We have included links to all the open-source software used in this manuscript. Further information and an overview of our open-source tools are available at <https://github.com/StanfordBDHG> and <https://github.com/StanfordSpezi>.

## References

- [1] 21st century cures act - an act to accelerate the discovery, development, and delivery of 21st century cures, and for other purposes. Bills and Statutes, December 2016. URL <https://www.govinfo.gov/app/details/PLAW-114publ255>.
- [2] Oliver Aalami, Mike Hittle, Vishnu Ravi, Ashley Griffin, Paul Schmiedmayer, Varun Shenoy, Santiago Gutierrez, and Ross Venook. CardinalKit: open-source standards-based, interoperable mobile development platform to help translate the promise of digital health. *JAMIA Open*, 6(3):ooad044, 07 2023. ISSN 2574-2531. doi: 10.1093/jamiaopen/oad044. URL <https://doi.org/10.1093/jamiaopen/oad044>.
- [3] Joshua Au-Yeung, Zeljko Kraljevic, Akish Luintel, Alfred Balston, Esther Idowu, Richard Dobson, and James Teo. Ai chatbots not yet ready for clinical use. 5:23, 04 2023. doi: 10.3389/fdgth.2023.1161098. URL <https://doi.org/10.3389/fdgth.2023.1161098>.
- [4] Lucas M. Bachmann, Felix S. Gutzwiller, Milo A. Puhan, Johann Steurer, Claudia Steurer-Stey, and Gerd Gigerenzer. Do citizens have minimum medical knowledge? a survey. *BMC Medicine*, 5:14, 2007. doi: 10.1186/1741-7015-5-14. URL <https://doi.org/10.1186/1741-7015-5-14>.
- [5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 2020.
- [6] Anjun Chen and Drake O Chen. Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data. *Scientific Reports*, 12(1):17917, 2022.
- [7] Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. The future landscape of large language models in medicine. *Communications Medicine*, 3(1): 141, 2023.
- [8] Susan Graham and Jonathan Brookey. Do patients understand? *The Permanente Journal*, 12(3):67–69, 2008. doi: 10.7812/TPP/07-144. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037129/>.
- [9] Health Resources & Services Administration. Health literacy. URL <https://www.hrsa.gov/about/organization/bureaus/ohe/health-literacy>.

- [10] Kathleen T. Hickey, Ruth M. Masterson Creber, Martha Reading, Robert R. Sciacca, Teresa C. Riga, Adrienne P. Frulla, and Jesus M. Casida. Low health literacy: Implications for managing cardiac patients in practice. *The Nurse Practitioner*, 43(8):49–55, 2018. doi: 10.1097/01.NPR.0000541468.54290.49. URL <https://doi.org/10.1097/01.NPR.0000541468.54290.49>.
- [11] C. Johnson, C. Richwine, and V. Patel. Individuals' access and use of patient portals and smartphone health apps, 2020. 2020. URL <http://resource.nlm.nih.gov/9918317682006676>.
- [12] Douglas Johnson, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, et al. Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model. *Research square*, 2023.
- [13] Maribeth Rauh Laura Weidinger, John Mellor. Ethical and social risks of harm from language models. 2021. doi: 10.48550/arXiv.2112.04359. URL <https://doi.org/10.48550/arXiv.2112.04359>.
- [14] Jan Lorkowski and Agnieszka Jugowicz. Shortage of physicians: A critical review. *Advances in Experimental Medicine and Biology*, 1324:57–62, 2021. doi: 10.1007/5584\_2020\_601. URL <https://pubmed.ncbi.nlm.nih.gov/33346901/>.
- [15] National Academies of Sciences, Engineering, and Medicine. *Criteria for Selecting the Leading Health Indicators for Healthy People 2030*. The National Academies Press, Washington, DC, 2019. ISBN 978-0-309-49594-3. doi: 10.17226/25531. URL <https://doi.org/10.17226/25531>.
- [16] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.
- [17] OpenAI. Gpt-4 technical report. 2023.
- [18] Debbie Rankin, Michaela Black, Raymond Bond, Jonathan Wallace, Maurice Mulvenna, Gorka Epelde, et al. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR medical informatics*, 8(7):e18910, 2020.
- [19] Paul Schmiedmayer and Oliver Aalami. SpeziHealthKit. URL <https://doi.org/10.5281/zenodo.7824636>.
- [20] Paul Schmiedmayer and Philipp Zagar. SpeziLLM. URL <https://doi.org/10.5281/zenodo.7954213>.
- [21] Paul Schmiedmayer, Vishnu Ravi, and Oliver Aalami. Stanford LLM on FHIR, . URL <https://github.com/StanfordBDHG/LLMonFHIR>.

- [22] Paul Schmiedmayer, Vishnu Ravi, and Oliver Aalami. Spezi, . URL <https://doi.org/10.5281/zenodo.7538238>.
- [23] Paul Schmiedmayer, Philipp Zagar, and Adrit Rao. SpeziSpeech, . URL <https://doi.org/10.5281/zenodo.10162167>.
- [24] Tetine Sentell and Kathryn L Braun. Low health literacy, limited english proficiency, and health status in asians, latinos, and other racial/ethnic groups in california. *J Health Commun*, 17 Suppl 3(Suppl 3):82–99, 2012. ISSN 1087-0415 (Electronic); 1081-0730 (Print); 1081-0730 (Linking). doi: 10.1080/10810730.2012.712621.
- [25] Nigam H Shah, David Entwistle, and Michael A Pfeffer. Creation and adoption of large language models in medicine. *Jama*, 330(9):866–869, 2023.
- [26] James Shaw, Payal Agarwal, Laura Desveaux, Daniel Cornejo Palma, Vess Stamenova, Trevor Jamieson, Rebecca Yang, R Sacha Bhatia, and Onil Bhattacharyya. Beyond "implementation": digital health innovation and service design. *NPJ Digit Med*, 1:48, 2018. ISSN 2398-6352 (Electronic); 2398-6352 (Linking). doi: 10.1038/s41746-018-0059-8.
- [27] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Bin Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. 2023.
- [28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. page 6000–6010, 2017. doi: 10.48550/arXiv.1706.03762. URL <https://doi.org/10.48550/arXiv.1706.03762>.
- [29] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 08 2017. ISSN

- 1527-974X. doi: 10.1093/jamia/ocx079. URL <https://doi.org/10.1093/jamia/ocx079>.
- [30] Jason Walonoski, Sybil Klaus, Eldesia Granger, Dylan Hall, Andrew Gregorowicz, George Neyrapally, Abigail Watson, and Jeff Eastman. Synthea™ novel coronavirus (covid-19) model and synthetic data set. *Intelligence-based medicine*, 1: 100007, 2020.
  - [31] Harriët Wittink and Janke Oosterhaven. Patient education and health literacy. *Musculoskeletal Science and Practice*, 38:120–127, 2018.
  - [32] Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability. 2023. doi: 10.48550/arXiv.2303.16756. URL <https://doi.org/10.48550/arXiv.2303.16756>.