PatientLens: AI-Enabled Interactive Avatars for Patient Report Summarization and Visualization

Utkarsh Singh
Arizona State University
usingh31@asu.edu

Bretho Danzy III
Arizona State University
bdanzy@asu.edu

Muhammad Umar Afzal
Mayo Clinic
afzal.muhammadumar@mayo.edu

Syed Naqvi
Mayo Clinic
naqvi.syed@mayo.edu

Haider Abdul-Muhsin
Mayo Clinic
abdul-muhsin.haidar@mayo.edu

Ewan Cobran
Mayo Clinic
cobran.ewan@mayo.edu

Irbaz Riaz
Mayo Clinic
riaz.irbaz@mayo.edu

Chris Bryan
Arizona State University
cbryan16@asu.edu

Abstract

In clinical environments, doctors often must review large amounts of patient reports prior to consults and check-ups — a task that is time-consuming, cognitively taxing, and prone to errors. We investigate how to improve this workflow via the use of AI-driven virtual avatars that enable clinicians to query, and summarize information from text-based patient reports. While the use of AI (specifically LLMs) brings significant potential benefits for clinical settings, it also presents critical challenges such as hallucination. Based on discussions and iterative prototyping with clinicians, we develop a human-in-the-loop approach that supports interactively creating and refining virtual avatars that visually present patient information while efficiently supporting LLM oversight and transparency. Evaluations help validate that the developed tool, nicknamed PatientLens, supports clinical review and summarization workflows. We also discuss how lessons learned during this project can enhance healthcare communication, optimize clinical workflows, and support improved health equity and outcomes.

Keywords: patient avatars, clinical visualization, large language models, electronic health records, human-computer interaction

1. Introduction

Doctor-patient communication is a critical component to effective healthcare, as effective communication between clinicians and patients has been shown to help to regulate patient emotions, facilitate improved comprehension of medical information, and better identify patient needs, perceptions, and expectations (Ha & Longnecker, 2010).

A requisite for effective clinical communication is the doctor's knowledge of the patient's medical history. Patient visits and consults are time-limited encounters that require proactive planning by clinicians; they generally must review the patient's historical electronic health records (EHRs) and summarize them in a manner that is personable, succinct, and comprehensible. Unfortunately, clinicians often have limited time to review large amounts of fragmented and unstructured EHR documents about a patient. Such files, which can include text files such as clinical, radiology, and pathology reports, not to mention additional data types such as imaging or biomarker information, are often created by different doctors and technicians and can span years of information. Clinicians typically spend significant time and effort manually reviewing, synthesizing, and summarizing patient reports prior to a visit, and there is a large body of research about how such review is not only cognitively taxing but also prone to misinterpretation and error (e.g., Patel et al., 2022).

We investigate how to improve this activity in the context of two emerging areas in data-driven healthcare: First, large language models (LLMs) are increasingly being studied as a way to streamline and improve various clinical practices, including patient review and consult planning. Second, the use of virtual, humanoid avatars are being studied as a mechanism to improve the efficiency of clinical review and to support patient communication. Specifically, we investigate how to design an effective software platform that enables the intuitive and customizable creation of interactive virtual avatars to summarize free-text EHR data (e.g., reports and clinical notes) about patients via the use of LLMs. Based on a pre-study and prototyping phase with a team of clinicians, we first distilled several design requirements for such a tool, such as

supporting mechanisms for efficient review, refinement, and mitigation of LLM risks (e.g., hallucination). Our implemented tool, named *PatientLens*, represents a novel software artifact for constructing such AI-enabled virtual avatars. We validate PatientLens via empirical and quantitative evaluations, and discuss lessons learned throughout this design study process, such as how the tool effectively surfaces a human-in-the-loop approach to promote oversight and mitigation of LLM risks, and how various techniques developed for this tool can be translated in other (non-medical) application scenarios.

2. Related Work

This work sits at the intersection of two areas in data-driven healthcare: the use of generative AI (specifically LLMs) and virtual, humanoid avatars for clinical understanding and communication. While each of these facets represents an active research topic, to our knowledge this work is some of the first that investigates how they can be integrated together within a single integrated framework for clinical application.

LLMs in Healthcare. LLMs can both recognize and generate text, and are highly effective at a variety of information extraction and processing tasks on text documents, including information retrieval, summarization, classification, comparison, and entity extraction (Xu et al., 2024). Particularly for medical and healthcare scenarios, one significant benefit of LLMs lies in their ability to process and synthesize diverse medical literature and patient data, which can be messy, unstructured (or semi-structured), and span several years of information about a single patient. Recent work has demonstrated the potential of LLMs to aid healthcare professionals in reducing information overload (Nazi & Peng, 2024), enhancing diagnostic accuracy (McDuff et al., 2025), and reducing the time required to for clinical summarization tasks and increasing patient comprehension (Yang et al., 2025); moreover, LLMs are specifically being developed for analyzing patient medical histories (Porter et al., 2024).

Unfortunately, the use of LLMs introduces several challenges, two of which are particularly salient for healthcare. The first deals with privacy, as patient EHR data contains significant amounts of personally identifiable information (PII). Several strategies have been proposed to address this, such as removing or masking PII prior to interacting with an LLM (Asthana et al., 2025) or using siloed LLMs that enforce required regulations such as HIPAA (Min et al., 2024). In our case, we assume a siloed LLM to maintain PII privacy; see Section 7 for more discussion on this.

The second risk deals with the fact that LLMs sometimes hallucinate plausible-sounding (but factually incorrect) information in their responses (Huang et al., 2025), which in clinical settings can result in misdiagnoses or inaccurate patient procedures or treatments. Recent work has categorized common forms of medical hallucinations (Kim et al., 2025) and underscored a need to develop validation and mitigation strategies to detect hallucinated outputs (Aurangzeb et al., 2023). This was also considered a critical risk by our collaborators when distilling design requirements and prototyping potential designs (see Section 3). As a solution, PatientLens adopts a human-in-the-loop approach to support oversight and transparency for the responses generated by LLMs.

Virtual Patient Avatars. Similar to LLMs, the use of virtual, humanoid avatars in healthcare is increasing (Tscholl et al., 2020), including for patient monitoring, clinician education, and diagnostic support. Empirical studies have demonstrated several potential advantages of avatar-based approaches for clinical activities, including increased diagnostic confidence, enhanced situational awareness, reduced cognitive load, and increased patient engagement and understanding (Bergauer et al., 2023; Tscholl et al., 2018; Wonggom et al., 2019). The design of avatars can also span from highly stylized (or cartoonish) to realistic. However, there has been little work at the intersection of LLMs and avatars, such as leveraging LLMs to summarize and visualize information on patient avatars (as is done in PatientLens).

3. Distilling Design Requirements and Iterative Prototyping

This project originated via discussions between the authors and clinicians (primarily a team of oncologists) who work in a cancer research clinic and hospital in the United States. The clinicians regularly spend significant time and effort reviewing patient reports prior to consults and check-ups (sometimes hours for a single patient), and experienced many of the issues already discussed in this paper (pressure due to time constraints, high cognitive load to interpret reports written by different people, etc.). The clinicians were interested in leveraging LLMs as a way to improve their patient review and summarization workflows. Moreover, they were also interested in leveraging personalized avatars as a medium for presenting such content (specifically text-based clinical reports — see Section 7 for discussion of additional data types such as imaging or biomarker data). Based on a series of discussion and observation sessions with these clinicians and a meta-analysis of the current state of LLMs and virtual

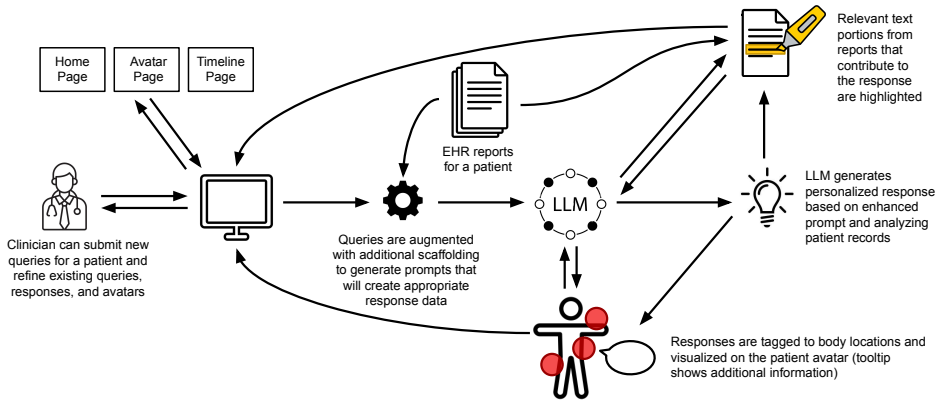

Figure 1. A high-level overview of the PatientLens platform. See Section 4 for a description of the frontend interfaces and Section 7 and Figure 7 for a detailed description of the backend pipeline.

patient avatars, we distilled a set of high-level design requirements DR1–DR4:

DR1: Support efficient querying, retrieval, and summarization of reports for a patient. To handle the scale, longitudinal nature, multi-authorship, and generally unstructured nature of EHR reports, our clinician collaborators wanted an intuitive and flexible approach querying and extracting desired information from these documents. This motivates our use of LLMs, as they have demonstrated success in these sorts of information extraction and processing tasks on document ensembles (Xu et al., 2024).

DR2: Support the interactive construction and refinement of humanoid avatars for personalized patient reporting. As mentioned above, our clinician collaborators were also interested in constructing virtual avatars to support reviewing patient information in a visual manner. However, manually creating avatars could take significant effort in settings that are already time-constrained and mentally intensive. As such, they needed a way to streamline the creation of avatars, such as automatically creating information items and appending them to the correct body location(s), and then efficiently supporting subsequent review and editing.

DR3: Provide mechanisms for transparency and mitigation of LLM risks such as hallucination. One of our collaborators' top concerns about using LLMs was hallucination. As such, they wanted ways to efficiently review and verify the responses produced by these models, by "going to the source" and seeing what specific content in the patient reports was being used to help create the response.

DR4: Support longitudinal review and comparison of multiple avatars for a single patient. Finally, because patients change and progress over time, the doctors wanted a way to create multiple "snapshots" to demonstrate such progress. Put another way, they wanted to be able to flexibly construct and compare multiple avatars (e.g. to across time periods).

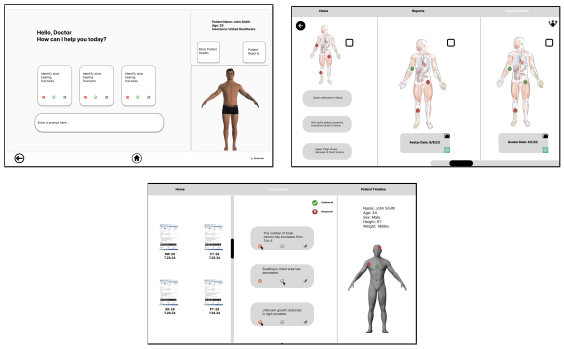

Figure 2. Examples mockups from prototyping sessions, created using Figma.

As a part of the planning process for PatientLens, we also created several UI/UX prototypes to solicit feedback about potential designs, affordances, aesthetics, and the overall system workflow (see Figure 2 for some examples). During prototyping, we met weekly with our collaborators over a period of approximately three months to prototype, review, and refine various interface and tool designs, including testing various UI/UX and interaction approaches, with the goal of effectively supporting clinical needs, minimizing required time and cognitive load, and supporting DRs 1–4, until a final design was approved. In particular, the clinicians continually expressed it was important to maintain an overall intuitive and streamlined workflow that flexibly allowed them to both generate personalized summary information about a patient, place this information onto physical locations on the avatar, and support real-time oversight and correction of potential LLM responses.

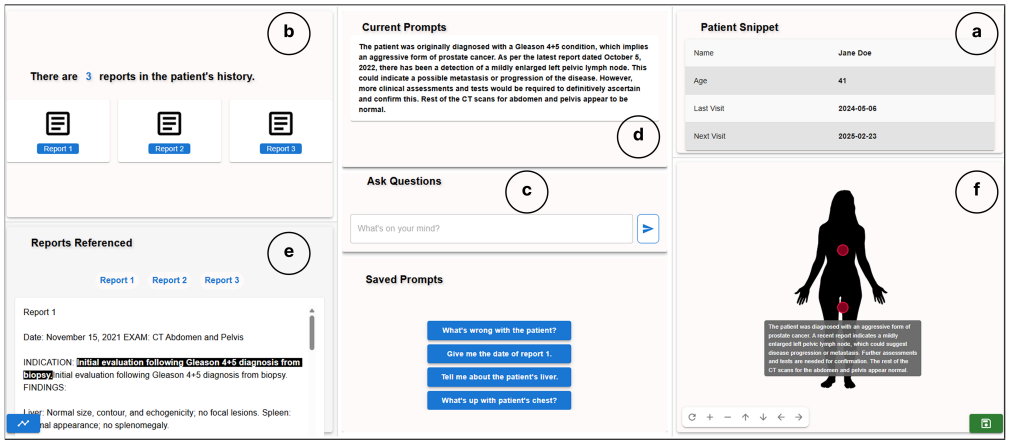

Figure 3. The Avatar Page contains six connected panels, showing (a) patient demographics, (b) available patient reports, (c) the chatbot query interface, (d) responses from queries, (e) the reports reference functionality, and (f) the patient's virtual avatar.

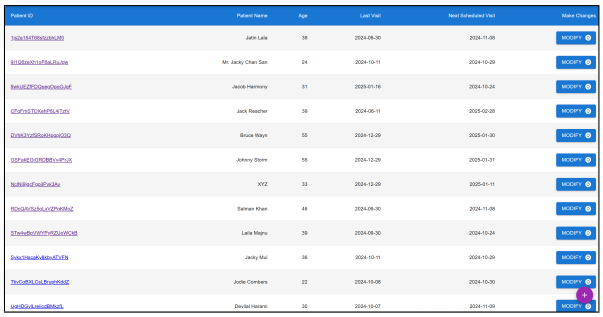

Figure 4. The Home Page lists available patients in a tabular format.

4. PatientLens Frontend

The PatientLens platform, an overview of which is shown in Figure 7, is designed to support DR1–DR4. The frontend contains three primary interfaces that allow users to interactively query on a patient's collection of text reports and construct, review, edit, and compare avatars. The frontend is primarily implemented using a combination of React, D3.js, and Material UI. The backend, described in Section 5, employs a RAG (retrieval augmented generation) approach (Lewis et al., 2020) to process patient reports and generate contextually accurate and customizable responses.

To begin using PatientLens, users load the **Home Page** (see Figure 4) which mimics existing clinical software such as EPIC EMR and lists available patients and their demographic information in a tabular format.

Selecting a patient navigates to the **Avatar Page**, which supports the querying patient reports and constructing virtual avatars. This interface consists of several coordinated panels: (a) At top right, the **patient demographics panel** displays demographic information for the selected patient. (b) Similarly, the **reports panel** displays the reports that are available for the patient. Individual reports can be selected for detailed review. (c) The **chatbot panel** allow users to submit queries about the patient (using the reports listed in the reports panel as a reference). A submitted query goes to the PatientLens backend, which uses an LLM-based pipeline to construct a response (this process is described in detail in Section 5). (d) The response's text (i.e., the answer to the user's query, based on prompting the LLM) is added to the **responses panel** within a box element. (e) In addition, the **reports referenced panel** shows the specific patient reports that contribute to the response text, and relevant text snippets within each report are highlighted (see Section 5 for a description of how these snippets are extracted).

In the response panel (d), hovering or clicking on the response's box loads these reports in the reference panel (e) and highlights the relevant text snippets. Additionally, clicking on a response box in the response panel animates it to an expanded view that shows additional information (see Figure‘5). The expanded box lists the relevant referenced text reports, lists the body parts(s) that the response is tagged to (see Section 5 for a description of how these are calculated), and shows the initial prompt that was submitted. The user can interactively update and resubmit the prompt by interacting with this box.

(f) When the LLM is queried and a response is generated, it also appends one or more circles to the avatar in the **avatar panel**. This panel shows a 2D outline of a human (either male or female) and visualizes each response as one or more circles on

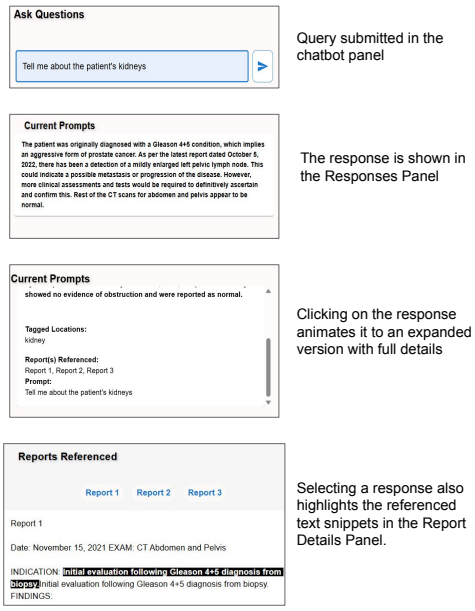

Figure 5. An example of how clicking on a response box expands it to show additional information and highlights the relevant text snippets in the report detail panel.

the avatar. Responses can reference one or more locations on the body; we break up the response text into "sub-responses" based on the body parts it relates to, and append each sub-response to the relevant location on the avatar (see Section 5 for details on this process). When multiple circles are appended to the same area, a collision detection algorithm adjusts their placements to prevent overlaps. Hovering on a circle displays the response text as a tooltip; clicking selects the response in the response panel, and also allows the user to delete or edit the circle's properties (e.g., the sub-response's text, the tagged body location, the specific x/y pixel coordinates the circle is positioned at, etc.).

Clicking on a navigation button in the avatar page navigates to the **Timeline Page**, which arranges all created avatars along a timeline (see Figure 6). Avatars are ordered based on their information date ranges (i.e., based on the date of the earliest referenced report). Users can review avatars similar to the previous page's avatar panel, or load a previously created avatar into the Avatar Page for further editing or inspection.

5. PatientLens Backend

The PatientLens backend employs a RAG approach and is implemented using a variety of technologies, including Google Firebase for data storage, and OpenAI's API for LLM interaction. At a high level,

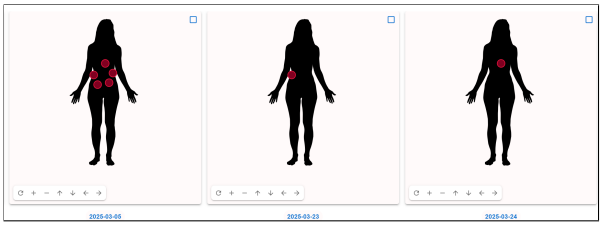

Figure 6. Within the Timeline Page, created avatars are ordered along a timeline.

when a query is submitted from the Avatar Page: (i) it submits an augmented prompt to the specified LLM, which generates an initial set of response text. (ii) The backend next identifies one or more "referenced text sections" from the patient reports, which represent text snippets that contribute to the LLM's response text. These are shown as highlighted snippets in the Avatar Page's referenced reports panel (Figure 3(e)). (iii) Additionally, one or more body locations are identified and tagged, and a set of specific "sub-response" text is generated, which are parts of the original response that correlate to specific body parts. Each sub-response is placed on the patient avatar (Figure 3(f)) as a circle. Figure 7 overviews the steps in the backend pipeline, which we detail below.

(i) **Processing prompts and generating responses.** When a user submits a query on the Avatar Page's chatbot panel (specifically, a query can request *any* information contained in a patient's records, such as a summary of their medical history, the date and/or status of a diagnosis, how a treatment has progressed over time, etc.), the system first retrieves patient's records (stored using Google Firebase Storage) and inserts the user query into a **Query Prompt Template** (shown in Figure 7). This prompt is submitted to an LLM touchpoint (for this paper, we use GPT-4, accessed via OpenAI's API, though PatientLens is extensible; see Section 7 for more discussion on this). The returned response text is used for steps (ii) and (iii), while also being sent back to the frontend for display in the Avatar Page's response panel (Figure 3(d)).

(ii) **Identifying the referenced text sections for a response.** The response text returned from the initial LLM query is next sent to a function that creates a second prompt (using Figure 7's **Report Reference Identification Prompt Template**) along with the patient's set of reports. This follow-up prompt queries the LLM to identify the specific sections in the patient reports that contribute to the original response, returned in a list format. This returned list of identified text snippets is sent to the frontend, where it is used to highlight the relevant report snippets in the Avatar

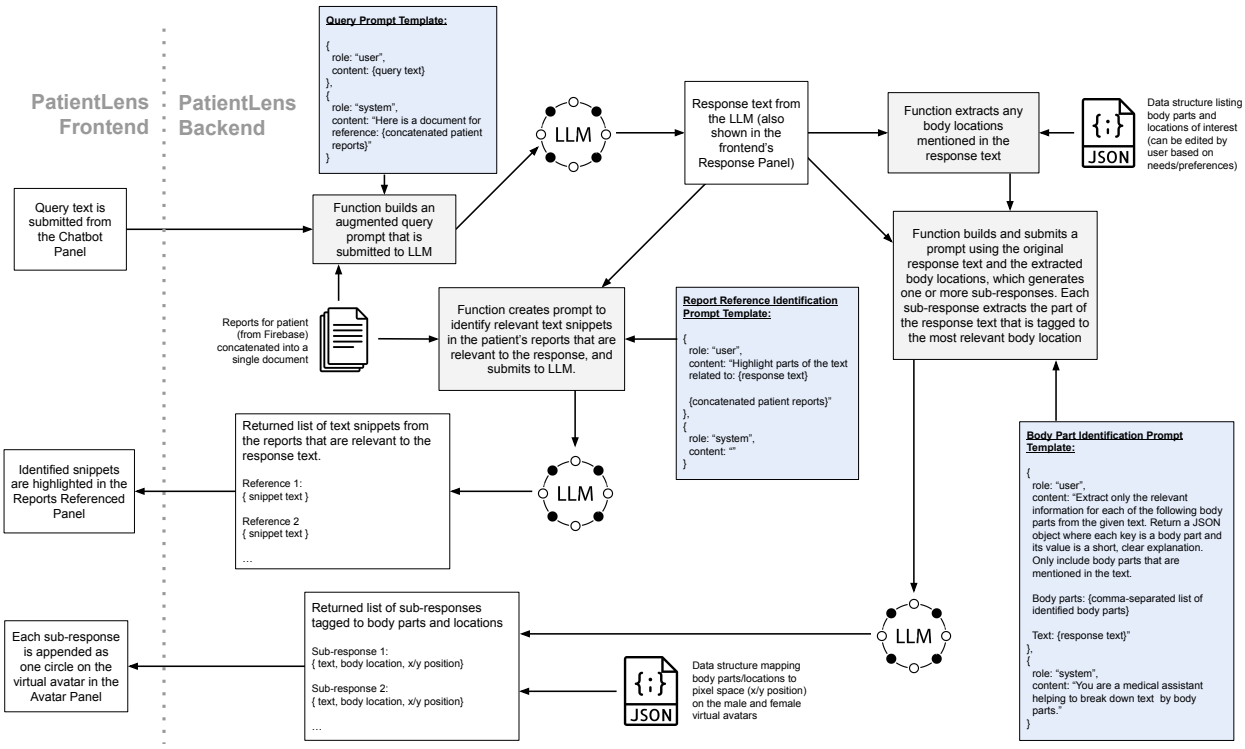

Figure 7. When a user submits a query in the PatientLens frontend (specifically via the Avatar Page's chatbot panel), the following pipeline is invoked. The backend incorporates a number of augmented prompts and service functions to query an LLM (for simplicity, we show the LLM API touchpoint using three icons), process responses, and return relevant information to the frontend.

Page's referenced reports panel (Figure 3(e)).

(iii) **Identifying the relevant body locations for a response.** Additionally, the backend system also identifies the set of (one or more) body parts that are referenced in the report. This is first done by extracting any body parts that are mentioned or inferred in the original response text, based on extracting keywords or vocabulary/terminology of interest (e.g., "thoracic," "inguinal," "calcaneal") and matching them to tagged body regions (e.g., "chest," "crotch," "foot"). This mathcing is based on an auxiliary data structure (stored as a JSON file) that can be edited, updated, or added to as needed, to flexibility support the specific needs of clinicians using the system (e.g., if a clinic employs abbreviations or shorthands for a body part/location, these can easily be added to this file). The extracted set of referenced body parts is then embedded into a prompt template (specifically, Figure 7's **Body Part Identification Prompt Template**) along with the original response text for LLM querying.

The returned result is a list of sub-responses: each sub-response contains part of the original response text and a tagged body location, or in cases where only one body part is tagged in the original response text, the sub-response equals the original response. (The Body Part Identification prompt also generates each sub-response text as a clear, standalone explanation to ensure interpretibility.) We then use a second auxiliary data structure (another JSON file) that maps tagged body parts to pixel space for the current avatar type (since the male and female avatars have different physical compositions; this mapping data structure can also be updated as needed by users). This mapping is used when appending the list of sub-responses to the patient avatar: each sub-response is one circle that is mapped to a tagged body part that has a corresponding x/y coordinate, which is where the circle is placed (notwithstanding any slight shifting due to collision avoidance).

6. Evaluation

To validate PatientLens, we performed two evaluations: first, we conducted empirical evaluations with clinicians, primarily to validate the overall design choices and that the tool successfully addresses DR1–DR4. Next, we performed a set of quantitative tests to assess the backend pipeline's performance

(e.g., its ability to correctly answer queries and tag appropriate report and body locations).

Feedback Sessions with Clinicians. We conducted feedback sessions with three clinicians (specifically, two oncologists and one neurosurgeon) who regularly review significant amounts of patient reports in their pre-consult preparation workflows. Given that this evaluation primarily focused on the design and usability aspects of PatientLens, we curated a synthetic patient dataset with our clinician collaborators, and used OpenAI's GPT-4 as our LLM (see Section 7 for additional discussion on the use of real patient data and siloed LLMs).

To mitigate learning effects, we employed a pair analytics methodology (Arias-Hernandez et al., 2011). Pair analytics is a commonly-used approach for evaluating visual interfaces, where a study administrator "drives" the system according to the participant's directions. Sessions lasted as long as each clinician desired, and the participant could explore or use any functionality as desired. Prior to the session, a short system demo was given to each participant, and afterwards each participated in a short semi-structured interview to elicit additional system feedback.

In terms of feedback, the primary sentiments, strongly expressed among the clinicians, is that the system effectively supports DR1–DR4 while also dramatically reduces the cognitive effort and time required to review patient reports. For example, one clinician emphasized that the use of LLMs in conjunction with "*the timeline and hallucination feature is definitely something of use*" in terms of saving time and eliminating potential errors. Another clinician estimated the tool could potentially "*reduce pre-checkup prep time by almost 50%*," especially compared to their current "baseline" workflows and tools (such as reviewing patient records in EPIC EMR). Interestingly, we also received several comments about the simplified avatar designs. While our pre-study collaborators emphasized and agreed upon the "simplified" avatar representation (i.e., black outlines in the shape of a male or female) that was implemented, two clinicians in these feedback sessions felt that "*the avatar needs to be more natural looking*" and realistic. We discuss this idea more in Section 7.

Quantitative Evaluations. To assess the quantitative performance of PatientLens, we conducted a set of performance tests where representative prompts (i.e., that mimic what would be submitted by a clinician for a real-world patient) were assessed for accuracy and hallucination. This evaluation primarily assessed the PatientLens' backend, including its use of prompt templates and service functions.

To do this, in coordination with our collaborators, we curated a set of queries that broadly fell into three categories: (i) 50 queries asking broad, general, open-ended questions about a patient (e.g., "*What's wrong with the patient?*"), (ii) 60 queries asking specific, detail-oriented questions ("*What is the size of the enlargement in lymph node 3?*"), and (iii) 65 queries asking comparative reasoning tasks that required analyzing differences between reports ("*How did the condition of the lymph nodes change across 2024?*"). These prompts were submitted in the PatientLens' chatbot panel for a curated synthetic patient dataset (the same one used during our empirical feedback sessions) and the responses reviewed, including the response text, the linked sources in the text reports, and the tagged body locations. Two coders independently rated the correctness of responses; inter-rater agreement was 96%, and all discrepancies were discussed to determine a final decision.

(i) For the broad, open-ended prompts that were tested, in 91% of cases the model correctly identified key anatomical terms (e.g., "chest," "crotch," "liver," "kidney"). Minor omissions (e.g., 3 of 4 regions were correctly identified) occurred in 8% of responses. Hallucinated outputs (e.g., mentioning "foot" erroneously) were rare ($\sim$1%).

(ii) For narrowly-scoped queries asking to identify and extract specific, detail-oriented information, the backend achieved almost perfect accuracy (99%), such as reliably retrieving desired numerical values from the correct report sections. This is in line with prior work (e.g., Rasool et al., 2024) that indicates LLMs can have high performance when when prompts are well-grounded and documents are coherently structured (even if the structuring is implicit).

(iii) For prompts that asked comparative reasoning tasks, particularly when comparing across multiple reports, we saw nuanced outcomes based on the specific wordings in the reports. For example, when comparing "numeric vs. semantic" descriptive terms (e.g., "1.2 cm" vs. "slight"), the model had extremely poor performance, only succeeding in 5% of the tested prompts. In many cases, the model misattributed or hallucinated one of the terms (e.g., misattributing "1.2 cm" with a semantic term like "mild," and then comparing this against a term like "slight" in a different report, resulting in an incorrect answer).

For "numeric vs. numeric" comparative queries, the model had high performance, succeeding in 94% of the tested prompts. Reviewing the 6% of failures, we saw they tended to be hedged responses. As an example, in one instance the model was asked to compare lymph node enlargement sizes listed as "1.2 cm" and "1.4 cm" in two reports. In another failure case, the model

provided an uncertain response: "[s]eems like report X. But more context is needed. Get a medical professional to look at the reports." Such a response is ambiguous, but could efficiently be reviewed in the PatientLens' frontend (e.g., by looking at highlighted snippets in the report reference panel).

Finally, for "semantic vs. semantic" queries (e.g., "slight" vs. "severe"), the model again showed high performance, succeeding in 94% of the tested prompts. Errors again tended to be cases where the response was seen as too cautious (e.g., "[s]eems like report X. But more context is needed...") rather than simply being incorrect or hallucinated.

Despite using synthetic patient data, these results are in line with prior studies assessing the correctness of LLMs to summarize and synthesize real-world clinical text reports (e.g., a recent study on LLM-based clinical documentation showed a 93.1% "usable" rating in their generated responses (Heilmeyer et al., 2024), though the details of the prompt have significant impacts on performance). However, we want to emphasize a couple of important caveats to this study: (i) PatientLens is LLM and query-agnostic. As different (i.e., newer) LLMs are introduced in clinical settings (and imported into PatientLens), performance will certainly change. (ii) Different clinicians (even if they are in the same role or clinic, such as two oncologists) will likely have different prompting styles. Likewise, different patients might have different types, amounts, and styles of EHR reports. All of these factors could impact the performance of submitted queries.

As such, while our quantitative analysis gives confidence in the tool's backend capabilities, a comprehensive benchmarking of LLM performance on patient EHR data is beyond the scope of this paper. Instead, our evaluations on both PatientLens' frontend and backend are meant to provide a holistic evaluation and validation in our approach. Specifically for the backend, our performance is in line with current prompting approaches, though we intend to investigate ways to improve this process in the future.

7. Discussion and Conclusion

PatientLens represents a novel AI-enabled approach for generating personalized and interactive avatars for clinical report summarization. PatientLens effectively supports the design requirements DR1–DR4 outlined in Section 3, and promotes an intuitive human-in-the-loop workflow that enables clinicians to summarize patient information while mitigating the hallucination risks inherent to LLMs. Below, we discuss some of the takeways and lessons learned during our process of designing, implementing, and validating PatientLens, including future work that can build on the results developed in this paper.

Evaluating PatientLens in Real-World Clinical Settings. As a design study (Sedlmair et al., 2012), this paper primarily focuses on understanding how to effectively design tools, techniques, and workflows for interacting with LLMs (including oversight) and supporting the construction of virtual patient avatars. As such, we test using synthetic patient data to mitigate potential privacy and HIPAA concerns, and employ a publicly accessible LLM (GPT-4). As a follow up, we intend to deploy PatientLens in actual clinical settings by conducting longitudinal evaluations (e.g., with clinicians using PatientLens to prepare for actual patient consults). For such deployments, the use of a siloed LLM will be necessary (i.e., one that is HIPAA-compliant for the clinicians), but this is easy to enable in PatientLens via updating a configuration file to point to the correct (siloed) LLM and any necessary patient databases. Such study will allow us to explore in detail how clinicians use PatientLents, compare its benefits (and drawbacks) to their baseline practices, compare how the tool is used across different clinical specialties, collect and report formal usability metrics, and study how individual differences might impact the usability and efficacy (e.g., two clinicians might have different querying approaches, or two patients might have different styles of EHR records). We are currently working with our pre-study collaborators towards such deployments and evaluations.

Expanding to Other Data Modalities. One limitation in PatientLens is it so far only considers text-based EHR data, while clinical review can also include additional data types such as imaging data or biomarker results. We intend to integrate such data modalities in the future, but accomplishing this will be non-trivial. For example, there are open questions about how to best extract information from such non-text data and then integrate it into PatientLens' overall architecture and pipeline. We intend to investigate such questions in the future.

Additional Opportunities for Future Work. We also see several additional potential opportunities for future work in this area, such as developing a more robust or complex RAG approach in the backend (e.g., dynamically building prompt templates based on context clues given in an initial query or the user's overall workflow). Such work would entail conducting additional benchmarking tests (expanding on the quantitative evaluation that performed in Section 6) to understand how such features could impact performance. As another example, the research

community is currently working to develop and mature foundation models that are tailored for medical and healthcare applications and constraints (e.g., aligning with HIPAA) (He et al., 2025; Tu et al., 2025); we are interested to test such models in PatientLens.

Likewise, as mentioned in Section 2, avatar designs can span from cartoonish to highly realistic. In our initial prototyping and implementation, we employed simple avatars as a way to streamline the user experience, though some clinicians in our empirical feedback sessions suggested more realistic avatars (e.g., based on 3D renderings of human anatomy) as an option. However, there are questions about how to design such avatars to dynamically show retrieved information in a way that doesn't degrade the overall user experience. Additionally, avatar representations could be customized to specific clinical workflows and patient needs (e.g., based on a diagnosis or treatment). Aspects of the backend pipeline would likely need to be improved to accommodate more complex, realistic, or customizable avatars, such as dynamically changing how content is tagged and shown on the avatars. Our collaborators are also interested in surfacing these avatars as a patient-facing tool, given the prior success of using avatars to promote patient engagement, understanding, and communication (e.g., Wonggom et al., 2019). For such scenarios, we would need to re-design aspects of PatientLens to create a patient-facing set of interfaces.

Applying Developed Techniques to Other Applications Domains and Scenarios. Finally, we are interested in applying some of the developed RAG and human-in-the-loop techniques here to other text processing and analysis scenarios outside of healthcare, such as digital media and journalism, social media analysis, and linguistic analysis. While PatientLens could not be directly deployed for such scenarios (e.g., the concept of an avatar does not make sense for analyzing an ensemble of news stories), many of the techniques developed for the backend could be modified for such use cases, e.g. to help summarize the content of such datasets. We would likely need to design a different frontend that is adapted for the specific application(s), however some of the mitigation and sourcing techniques (e.g., highlighting relevant text snippets in documents) could likewise be modified for such scenarios.

Conclusion. The ongoing maturation of LLMs is bringing significant opportunity to the medical and healthcare communities, but there are intrinsic and non-trivial challenges in their adoption and usage. In particular, human-in-the-loop approaches such as PatientLens that leverage intuitive and visualization-driven designs can streamline clinician workflows without imposing additional overhead or effort. However, such tools should be appropriately designed and provide mechanisms for oversight and risk mitigation. PatientLens represents one study of how to design such a tool, via the integrated use of LLMs, interactive avatars, and a connected, full-stack software framework.

References

Arias-Hernandez, R., Kaastra, L., Green, T., & Fisher, B. (2011). Pair analytics: Capturing reasoning processes in collaborative visual analytics. *2011 44th Hawaii International Conference on System Sciences*, 1–10.

Asthana, S., Mahindru, R., Zhang, B., & Sanz, J. (2025). Adaptive PII mitigation framework for large language models. *AAAI Conference on Artificial Intelligence*.

Aurangzeb, M., Yaramis, I., & Dutta, T. (2023). Creating trustworthy LLMs: Dealing with hallucinations in healthcare AI. *arXiv preprint arXiv:2311.01463*.

Bergauer, L., Braun, J., Roche, T., Meybohm, P., Hottenrott, S., Zacharowski, K., Raimann, F., Rivas, E., López-Baamonde, M., Ganter, M., Nöthiger, C., Spahn, D., Tscholl, D., & Akbas, S. (2023). Avatar-based patient monitoring improves information transfer, diagnostic confidence and reduces perceived workload in intensive care units: Computer-based, multicentre comparison study. *Scientific Reports*, *13*(1).

Ha, J., & Longnecker, N. (2010). Doctor-patient communication: A review. *Ochsner Journal*, *10*(1), 38–43.

He, Y., Huang, F., Jiang, X., Nie, Y., Wang, M., Wang, J., & Chen, H. (2025). Foundation model for advancing healthcare: Challenges, opportunities and future directions. *IEEE Reviews in Biomedical Engineering*, *18*, 172–191.

Heilmeyer, F., Böhringer, D., Reinhard, T., Arens, S., Lyssenko, L., & Haverkamp, C. (2024). Viability of open large language models for clinical documentation in german health care: Real-world model evaluation study. *JMIR Medical Informatics*, *12*, e59617.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models:

Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55.

Kim, Y., Jeong, H., Chen, S., Li, S. S., Lu, M., Alhamoud, K., Mun, J., Grau, C., Jung, M., Gameiro, R., et al. (2025). Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., Singhal, K., Sharma, Y., Azizi, S., Kulkarni, K., et al. (2025). Towards accurate differential diagnosis with large language models. *Nature*, 1–7.

Min, S., Gururangan, S., Wallace, E., Shi, W., Hajishirzi, H., Smith, N., & Zettlemoyer, L. (2024). SILO language models: Isolating legal risk in a nonparametric datastore. *The Twelfth International Conference on Learning Representations*.

Nazi, Z., & Peng, W. (2024). Large language models in healthcare and medical domain: A review. *Informatics*, *11*(3).

Patel, Z., Schroeder, J., Bunch, P., Evans, J., Steber, C., Johnson, A., Farris, J., & Hughes, R. (2022). Discordance between oncology clinician-perceived and radiologist-intended meaning of the postradiotherapy positron emission tomography/computed tomography freeform report for head and neck cancer. *JAMA Otolaryngology–Head & Neck Surgery*, *148*(10), 927–934.

Porter, R., Diehl, A., Pastel, B., Hinnefeld, J., Nerenberg, L., Maung, P., Kerbrat, S., Hanson, G., Astorino, T., & Tarsa, S. (2024). LLMD: A large language model for interpreting longitudinal medical records. *arXiv preprint arXiv:2410.12860*.

Rasool, Z., Kurniawan, S., Balugo, S., Barnett, S., Vasa, R., Chesser, C., Hampstead, B. M., Belleville, S., Mouzakis, K., & Bahar-Fuchs, A. (2024). Evaluating LLMs on document-based QA: Exact answer selection and numerical extraction using cogtale dataset. *Natural Language Processing Journal*, *8*, 100083.

Sedlmair, M., Meyer, M., & Munzner, T. (2012). Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2431–2440.

Tscholl, D., Rössler, J., Said, S., Kaserer, A., Spahn, D., & Nöthiger, C. (2020). Situation awareness-oriented patient monitoring with visual patient technology: A qualitative review of the primary research. *Sensors*, *20*(7).

Tscholl, D., Weiss, M., Handschin, L., Spahn, D., & Nöthiger, C. (2018). User perceptions of avatar-based patient monitoring: A mixed qualitative and quantitative study. *BMC Anesthesiology*, *18*, 1–11.

Tu, T., Schaekermann, M., Palepu, A., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Cheng, Y., et al. (2025). Towards conversational diagnostic artificial intelligence. *Nature*, 442–450.

Wonggom, P., Kourbelis, C., Newman, P., Du, H., & Clark, R. (2019). Effectiveness of avatar-based technology in patient education for improving chronic disease knowledge and self-care behavior: A systematic review. *JBI Evidence Synthesis*, *17*(6), 1101–1129.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., & Chen, E. (2024). Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, *18*(6), 186357.

Yang, X., Xiao, Y., Liu, D., Shi, H., Deng, H., Huang, J., Zhang, Y., Liu, D., Liang, M., Jin, X., et al. (2025). Enhancing physician-patient communication in oncology using GPT-4 through simplified radiology reports: Multicenter quantitative study. *Journal of Medical Internet Research*, *27*, e63786.