Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)

A Human-Centered Approach to Identifying Promises, Risks, & Challenges of Text-to-Image Generative AI in Radiology

Katelyn Morrison[1], Arpit Mathur[1], Aidan Bradshaw[1], Tom Wartmann[2], Steven Lundi[3], Afrooz Zandifar[4], Weichang Dai[5], Kayhan Batmanghelich[5], Motahhare Eslami[1]*, Adam Perer[1]*

[1]Carnegie Mellon University
[2]ETH Zurich
[3]UCLA Health
[4]UPMC
[5]Boston University
kcmorris@andrew.cmu.edu, arpitmam@andrew.cmu.edu, abradsha@andrew.cmu.edu, wd2119@bu.edu, batman@bu.edu, meslami@andrew.cmu.edu, adamperer@cmu.edu

Abstract

As text-to-image generative models rapidly improve, AI researchers are making significant advances in developing domain-specific models capable of generating complex medical imagery from text prompts. Despite this, these technical advancements have overlooked whether and how medical professionals would benefit from and use text-to-image generative AI (GenAI) in practice. By developing domain-specific GenAI without involving stakeholders, we risk the potential of building models that are either not useful or even more harmful than helpful. In this paper, we adopt a human-centered approach to responsible model development by involving stakeholders in evaluating and reflecting on the promises, risks, and challenges of a novel text-to-CT Scan GenAI model. Through exploratory model prompting activities, we uncover the perspectives of medical students, radiology trainees, and radiologists on the role that text-to-CT Scan GenAI can play across medical education, training, and practice. This human-centered approach additionally enabled us to surface technical challenges and domain-specific risks of generating synthetic medical images. We conclude by reflecting on the implications of medical text-to-image GenAI.

Introduction

The rapid integration of Generative Artificial Intelligence (GenAI) into various workplace and educational settings, including healthcare, has simultaneously sparked excitement and concern. GenAI holds great promise for augmenting workflows, improving productivity, and enhancing training across a wide range of industries (Chan 2023; Ooi et al. 2023). However, the rapid development of domain-specific GenAI without understanding the needs of and challenges faced by domain stakeholders raises the risk of GenAI being misused. Within healthcare, there is a growing amount of literature on new architectures for text-to-image (T2I) GenAI capable of generating complex medical imagery from text prompts (Mai, Mirza, and DiMarco 2024; Hamamci et al. 2023; Xu et al. 2024; Chen et al. 2024). However, much of the research has focused on technical advancements, such as novel architectures, instead of being rooted in the needs and challenges of medical stakeholders.

Traditionally, synthetic image generation has been used to address pivotal technical medical imaging tasks (Sorin et al. 2020), such as image reconstruction or correction (Levac et al. 2024). Now, with foundation T2I models becoming widely accessible, the allure of simulating rare conditions, creating patient vignettes, or augmenting training datasets is attracting significant interest (Thesen, Alilonu, and Stone 2024; Pérez-García et al. 2025; Waikel et al. 2023; Waisberg et al. 2023). However, much of this enthusiasm is driven by technological evaluations (Noel 2024; Bhardwaj et al. 2023; Lin et al. 2024; Ku et al. 2023) rather than human-centered ones, leading to a *disconnect* from key stakeholders' real needs and challenges. As recently emphasized by human-centered AI researchers, evaluations need to extend beyond algorithmic metrics by understanding how the AI can address stakeholders' real needs and what challenges arise by doing so (Sivaraman et al. 2025). Although recent research captures medical stakeholders' perspectives on GenAI's potential applications when used in practice (Yildirim et al. 2024), few works capture medical stakeholders' perspectives on the technical challenges and domain-specific risks of generating synthetic medical imagery.

Several studies have identified and evaluated the challenges and risks of domain-agnostic T2I models (Bird, Ungless, and Kasirzadeh 2023; Naik and Nushi 2023; Bianchi et al. 2023). However, it is unclear how these challenges and risks translate to domain-specific T2I models. To address these gaps in literature, we follow up a quantitative evaluation of a novel medical T2I model (Xu et al. 2024) with a human-centered qualitative evaluation. As an interdisciplinary team with expertise in human-computer interaction, machine learning, and radiology, we conducted a human-centered evaluation of the promises, technical challenges, and domain-specific risks of a text-guided model that generates Computed Tomography (CT) images of the lung (text-to-CT Scan GenAI) (Xu et al. 2024). CT scans provide radi-

*These authors contributed equally.

ologists with cross-sectional views of the internal structures of the human body to aid in diagnosis (Mayo Clinic 2023).

Our human-centered approach in this work consists of two phases (formative discussions and model exploration) to facilitate discussions about: (**RQ1**) How do key stakeholders imagine using text-to-CT Scan GenAI for medical education, training, and practice? and (**RQ2**) What technical and domain challenges emerge from stakeholders' needs and workflows for medical text-to-CT Scan GenAI? To address these questions, we engaged with a total of eight radiologists, four trainees, and two medical students. We will refer to this group as key stakeholders throughout the paper.

Our formative discussions included six radiologists and one senior trainee, and they took place at the end of the quantitative evaluation of the model in the study conducted by (Xu et al. 2024). These discussions focused on capturing stakeholders' ideas on how the model might be applied in practice (**RQ1**). This informed the type of tasks we gave participants during the second phase (model exploration), broadened our network for recruiting additional participants, and informed the selection of participant demographics. For the second phase, we recruited two medical students, three radiology trainees, and four radiologists (including two from the formative discussions) to prompt and evaluate the model in a semi-structured interview format. To facilitate easy prompting and output review, participants interacted with the model through a plugin that we developed for an existing open-source medical imaging interface (created by Ziegler et al. (2020)). By interacting with the model, participants imagined scenarios of how it can address real needs and challenges faced throughout radiology workflows (**RQ1**). The exploratory prompting activity also shed light on the technical challenges and potential domain-specific risks that researchers should consider when developing medical T2I models (**RQ2**).

Our work makes three contributions to the AIES and broader Responsible AI (RAI) community. First, we present a text-to-CT Scan GenAI plugin for a popular open-source medical imaging viewer (created by Ziegler et al. (2020)), which researchers can extend to explore additional approaches for human-centered evaluations of the ethical and safety challenges of domain-specific T2I models. Second, we are the first work to leverage a human-centered approach to explore a medical T2I model with medical stakeholders. As a result of this, our paper expands upon Yildirim et al. (2024)'s human-centered GenAI work for applications of multimodal AI in radiology by mapping out applications of domain-specific T2I across medical education, radiology training, and practice. Third, we extend existing RAI work and build on del Cerro and Esteve (2023)'s position on the implications of using medical T2I GenAI by presenting technical challenges and domain-specific risks that emerged from participants interacting with the model. These challenges and risks range from topics such as confirmation bias, misrepresentation of image findings, and output image resolution. We discuss the implications of developing medical T2I and suggest future research directions to consider exploring.

Background and Related Works

We distinguish the novelty of our contributions by highlighting the latest research on human-centered approaches to evaluating text-to-image GenAI. Additionally, we contextualize our work by synthesizing the current literature that has discussed the promises and challenges of medical text-to-image GenAI in education and practice.

Human-Centered Approaches to T2I Evaluation

The development and deployment of GenAI systems has attracted significant attention, especially as GenAI becomes widely accessible across domains, such as healthcare, education, and creative industries. As they become accessible across domains, researchers have set out to understand and mitigate harm from and misuse of T2I models (Gu 2024; Naik and Nushi 2023; Bianchi et al. 2023; Bird, Ungless, and Kasirzadeh 2023). As part of this effort, numerous works have designed benchmarks and metrics to help quantify the harm and safety of T2I models (Lee et al. 2023; Li et al.; Hao et al. 2024). Parrish et al. (2023) and Deng et al. (2025) create frameworks and tools to engage the public in auditing the safety of T2I models. Collins et al. (2024) design a feedback mechanism with thumbs up/down as a way to operationalize human-centered evaluations. Lin et al. (2024) and Ku et al. (2023) leverage image-to-text GenAI to quantitatively evaluate model generations through prompt and text output alignment. Bird, Ungless, and Kasirzadeh (2023) conducted a literature review to synthesize the different types of risks that emerge for researchers to consider when developing and evaluating domain-agnostic T2I models. Human-centered design principles, agnostic to the type of GenAI application, have since been developed, emphasizing the importance of co-creativity, transparency, and generative variability (Weisz et al. 2024; Chen et al. 2023).

Several studies have recently evaluated the capability of domain-agnostic foundation models to generate medical images (Bhardwaj et al. 2023; Moin et al. 2024; Temsah et al. 2024). Noel (2024) and Kumar et al. (2024) similarly probe different text-to-image GenAI models to evaluate their capability in generating different medical images. One study ran an educational intervention study with students to quantitatively evaluate the impact of generative AI images in education about pediatric genetics (Waikel et al. 2023). Thesen, Alilonu, and Stone (2024) recruited a diverse group of stakeholders to evaluate generated images for education across several scales, including the images' potential to be used in medical presentations. The above-mentioned studies primarily focus on quantitative evaluations of domain-agnostic T2I models with proxy metrics, such as anatomical realism or the presence of diseases, instead of assessing whether T2I capabilities meaningfully address stakeholders' needs and challenges. Our work builds on existing studies by uncovering the promises, risks, and challenges of T2I capabilities in the context of medical education, training, and practice through prompting activities with relevant stakeholders.

Promises and Challenges of Medical T2I GenAI

In Medical Education. With text-to-image models increasing in competency and becoming widely accessible, researchers are exploring the possibility of using these models to generate medical images for education and training, according to a recent survey (Mai, Mirza, and DiMarco 2024). In fact, the use of foundation models to generate medical images is becoming so popular among the medical community that Safadi, Zayegh, and Hawoot (2024) created a guide for medical educators on using generative AI in presentations.

Numerous previous works from medical literature have explored foundation models' ability to generate images sufficient to use for medical education (Waikel et al. 2023; Kumar et al. 2024; Koljonen 2023; Zhu et al. 2024; Noel 2024). For example, Zhu et al. (2024) probe the potential of OpenAI's DALL-E (Ramesh et al. 2021) to generate ECGs, including ECGS that are abnormal for teaching purposes. Although they do not present a study on probing or interviewing stakeholders, the authors use their professional experiences to provide insight into DALL-E's potential as a teaching tool. Currie, Hawk, and Rohren (2024) raise numerous concerns about using T2I models in nuclear medicine for education, citing the need to identify policies and regulations that address misuse and abuse, misrepresentations and hallucinations, and patient privacy concerns. Our paper adds to these findings by capturing perspectives on T2I in medical education from a diverse group of medical stakeholders.

In Clinical Practice. Clinical decision-making informed by medical imaging is a nuanced process that relies on expertise, peer collaboration, iterative review, and the use of reference resources to augment diagnostic accuracy and patient care. Despite advancements in AI, existing tools and interfaces tend to overlook these needs and focus on higher-order problems such as pathology detection (Sloan et al. 2024) or automated report writing (Yousefirizi et al. 2022). However, recent work has argued for the need to explore AI's true value to clinicians throughout their clinical workflows (Sivaraman et al. 2025). A recent study by Waisberg et al. (2023) looks at text-to-image GenAI's potential as a collaborative tool for clinicians in helping them understand neuro-visual diseases' impact on patients' ability to see by generating an image based on a description of the patient's symptoms. However, no formal quantitative or human-centered evaluations were conducted. del Cerro and Esteve (2023) suggests how T2I can be used by physicians to simulate treatment effects to navigate treatment plans or by radiologists to enhance noisy images. The authors simultaneously cite general concerns and potential implications of misuse and data bias.

Although previous works have identified risks with T2I models in medical education, training, and practice, they are primarily based on domain-agnostic T2I models, such as DALL-E or MidJourney. Furthermore, they primarily consist of showing already generated images to participants instead of allowing key stakeholders to explore and prompt the model themselves. Our work builds on the existing literature by having medical students, radiology residents, and radiologists interact with a domain-specific T2I to capture richer applications and challenges of medical T2I.

Study Design

We present the main takeaways from our formative discussions and then describe the protocol for our semi-structured interview. We also provide technical details on the text-to-CT Scan plugin that participants used during the interview.

Formative Discussion Takeaways

We interviewed six radiologists and one senior radiology trainee (F1-F7) for an average of 6.8 minutes after they quantitatively evaluated outputs from the MedSyn model in Xu et al. (2024)'s study[1]. Ultimately, the formative discussions provided us with preliminary insights into radiologists' challenges and their perceptions of the value of text-to-CT Scan GenAI in education, training, and practice. However, participants struggled to detail specific scenarios since many had never used T2I GenAI before. We also uncovered participants' challenges with current online resources, such as visualizing the exact pathologies they search for, the importance of visual comparative analyses, and lack of exposure to T2I GenAI. These challenges inspire us to integrate MedSyn into the medical imaging interface by Ziegler et al. (2020).

Text-to-CT Scan GenAI Plugin

Our formative discussions with radiologists revealed to us that they frequently search for image findings and pathologies on domain-specific search engines. However, they may not always find exactly what they are searching for. Therefore, we developed a GenAI plugin powered by the text-to-CT Scan GenAI model, MedSyn (Xu et al. 2024), and integrated it into a popular open-source medical image viewer (Ziegler et al. 2020)[2]. The MedSyn model enables users to generate CT scans based on exact and specific pathologies, extending capabilities beyond current search tools. As depicted in Figure **1** below, MedSyn takes a user's text prompt and generates a 3D CT Scan consisting of the axial, sagittal, and coronal views of the lungs. Generating the complete CT scan takes between three and four minutes when using an A100 GPU. Our plugin manifests in the interface (shown in Figure **2** below) as a text box where users can (**1a**) name their prompt, (**1b**) write their prompt, and (**1c**) generate a CT scan by calling the model API through the click of a button. Users can view a list of previous CT scans that they generated (**1d**) and have the opportunity to provide feedback on the generation (**1e**).

Front-end Interactions. When using this tool, a radiologist can view a patient's case history (**2a**), noting their findings (**2b**) and impressions (**2c**). This functionality is incorporated to align with how radiologists currently read CT scans in clinical practice. Viewing the patient's CT in the main viewer (**3a**) gives them access to numerous functionalities (**3b**) that they are used to having, such as adjusting the contrast to preset window levels (*e.g.*, soft tissue, bone, and lung)[3], adding annotations, comparing images side-by-side,

[1]Available at https://arxiv.org/pdf/2507.16207

[2]Plugin Demo available at https://genai-radiology.web.app/.

[3]Presets were determined by an author, who is a radiologist.

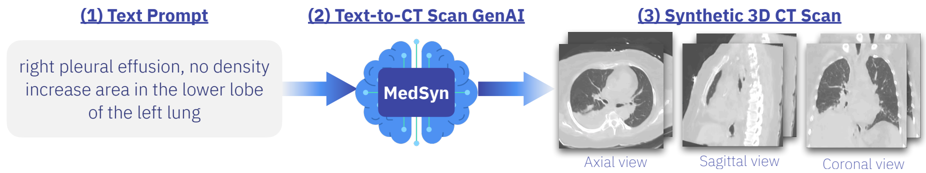

Figure 1: **Input/Output of the Text-to-CT Scan GenAI Model (Xu et al. 2024).** The text-to-CT Scan GenAI model that we use, called MedSyn (Xu et al. 2024), takes in a text prompt from the user to generate a 3D CT Scan. *Brain icon by Icons8.*

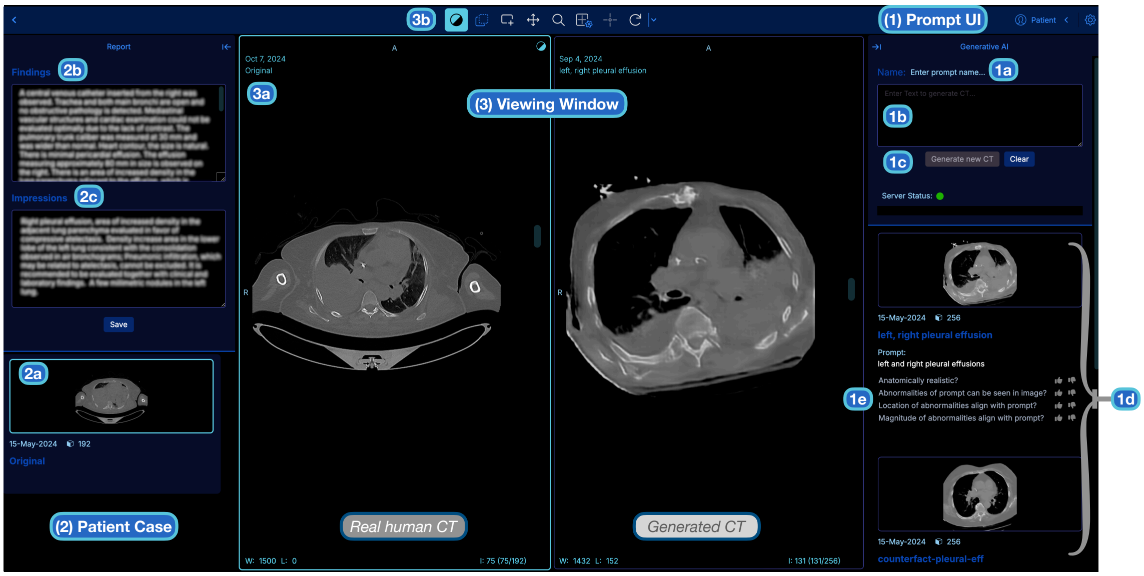

Figure 2: **Open-sourced medical imaging interface by Ziegler et al. (2020) with our GenAI plugin powered by MedSyn (Xu et al. 2024). (2) Patient Case.** The first column shows a real case consistent with severe right pleural effusion (sampled from CT-RATE dataset (Hamamci et al. 2023, 2024; Hamamci, Er, and Menze 2024). **(3) CT Scan Viewing Window.** The second and third columns show a CT scan comparison viewing window, where a human's CT scan (from (Hamamci et al. 2023, 2024; Hamamci, Er, and Menze 2024)) is on the left, and a generated CT scan from MedSyn (Xu et al. 2024) is on the right. **(1) Prompt UI.** The fourth column is our text-to-CT Scan GenAI plugin, allowing users to generate a CT scan.

and using a localizer tool. Many of these functionalities either expand upon what the existing radiology search engines offer or leverage findings from our formative interviews. For example, the tool allows users to view and compare a generated CT scan with a real CT scan, as well as annotate them.

Semi-structured Interviews

We conducted 45-minute semi-structured interviews with relevant medical stakeholders to elicit their perspectives on the potential applications and implications of text-to-CT Scan GenAI. These interviews were designed to be exploratory in nature; the goal was not to formally audit or analyze prompting behaviors. The interview sessions consist of four parts, inspired by Cai et al. (2019)'s approach with a clinical decision-making design probe, which included a *pre-probe*, *probe*, and *post-probe* phase. We describe each part in the following subsections.

Introduction. The interviews started with an introduction and background discussion, where participants shared their experiences and challenges throughout their medical education, training, and practice. This portion was designed to understand the tools, methods, and resources they use to learn about different pathologies and image findings in CT scans.

Pre-probe. Participants were asked about their experiences with text-to-image GenAI. For participants who have never used text-to-image GenAI before, we showed them a short conversation made in ChatGPT where we already requested the generation of a medical image from DALL-E[4]. All participants were asked before interacting with the model if they could think of any use cases for text-to-image GenAI being valuable to them throughout their education, training, or practice to prime them for the rest of the interview session. Following this, we showed all participants an animated illustration of how the text-to-CT Scan GenAI model works, and participants were acclimated to the interface through a guided walkthrough video.

Probe. During the probe phase, participants generated two to four CT scans, depending on session duration and the

[4]Provided at https://arxiv.org/pdf/2507.16207

PID	Expertise Level	Radiology Experience (years)
R1	Expert	5–10
R2	Trainee	1
R3	Expert	> 20
R4	Expert	5–10
R5	Expert	> 20
R6	Trainee	1
R7	Student	0*
R8	Student	0**
R9	Trainee	1

Table 1: Participant Demographics for interview sessions. *Fourth-year med student. **Third-year med student.

depth of discussions. They were guided to use the model to generate CT scans based on different goals informed by the formative interviews, such as validating their impression of a real patient's CT scan loaded in the viewer from CT-RATE (Hamamci et al. 2023, 2024; Hamamci, Er, and Menze 2024), learning about a pathology through differential diagnosis, and generating something of their choice. Throughout this process, participants provided feedback about the prompts they chose, the quality of the generation, and insights on the model's potential applications in radiology education, training, and practice.

Post-probe. Finally, after interacting with the model, participants were asked to reflect on the risks of generating synthetic CT scans in real-world applications and raise any concerns about the generated output.

Participant Recruitment

Participants were recruited through snowball sampling methods from our clinical collaborators' networks. For each session, we provided the participants with an overview of the study's purpose and methods and obtained their consent for participation. All participants consented to audio and screen recording for data analysis and were provided a form of monetary compensation for their time. We interviewed nine stakeholders (R1–R9) across three different institutions: two medical students, three radiology trainees, and four attending radiologists (two from the formative study and two new ones); experience distribution is provided in Table 1). Sessions were held on Zoom or Teams, based on preference, and lasted for an average of 46.34 minutes.

Analysis

We used an affinity diagramming approach to analyze the qualitative data. First, the transcripts from the interview sessions were verified for accuracy, and corrections were made by listening to the recordings as necessary. Quotes were pulled out from the transcripts, and affinity mapping was conducted by multiple authors to group similar quotes across interviews based on common themes. All quotes were grouped into primary themes and sub-themes that aligned with our research questions. Multiple authors grouped the quotes individually, and then the final groups were determined through discussion.

Findings

We first report on the identified promises of T2I in medical education, training, and practice (**RQ1**). We then follow up on these promises with the risks and challenges of T2I GenAI that stakeholders identified (**RQ2**). Table **2** and Table **3** show high-level representations of the findings for both research questions. A complete list of the prompts created by each participant, along with their comments about the prompt and the generated output, is provided in the supplemental materials.

Promises of T2I in Medical Education & Training

An application context that was brought up frequently in the formative interviews, as well as our semi-structured interviews, was the potential to use this technology in medical education and training. R7 presents a compelling case for using text-to-CT Scan GenAI in medical education because, "*...it gives you the opportunity to identify and recognize something that is potentially super dangerous, that you then don't have to guess about when you see it in real life*". We see the essence of this quote throughout our conversations with students and trainees. We have grouped participants' insights on specific use cases into three themes regarding the promises of T2I GenAI in medical education and training: (**1**) supplementing traditional lectures and practice materials, (**2**) expediting learning through variations, differentials, and rare cases, and (**3**) empowering trainees during on-call or emergency situations.

Supplement Lectures & Practice Materials. Despite numerous resources out there, "*...there isn't a great place to be like, Hey med students, let's go here and learn chest CT today*" (R8). R8 goes on further describing how they have never done anything in their classes where they get to interact with and compare the differences between abnormal and normal anatomies on CT scans. This could be due to how medical school lectures are designed, showing static images on lecture slides. R3 corroborates this limitation in existing lecture style by describing how they get content for their lecture slides by searching online or through their archives: "*I have this list of typical cases for all of these different lesions...choosing representative images from them...or showing a student the entirety of the scan*". However, R3 went on to envision a scenario where they could use this T2I model to generate specific examples to use on their slides: "*I'm thinking about pre-populating...I'm gonna produce a lecture on this list of obscure topics...And I need specific examples of A, B, C, ... I feed those into your generator, and...I'll come back in an hour....and I'm ready to start producing my lecture.*" R1 provides another perspective, acknowledging how convenient text-to-CT Scan GenAI could be when "*...teaching the residents or medical students,... we just want to show them something, we can just type it, and it will pop up*."

From the students' perspective, R7 and R8 saw the potential for text-to-CT Scan GenAI to support their studies from the earliest stages of their medical education. The process of medical students shadowing radiology residents is currently difficult as "*It's often been said that radiology even is very hard to shadow in because* ***it's like you're watch-***

Phase	Promises of Text-to-CT Scan GenAI	Participants
Education & Training	Supplement Lectures & Practice Materials	F2, R1, R3, R7, R8
	Learning through Variations, Differentials, & Rare Cases	F6, R1, R2, R6, R9
	Empower Trainees when On-Call or in Emergency Situations	F3, R2, R7, R8
Practice	Improving Report Impression of Medical Images	F5, R1, R4–R6
	Planning & Communicating with Others	R2, R6, R7
	Visual Memory Support	F6, R4, R5

Table 2: (*RQ1*) Participants mapped to which applications of text-to-CT Scan GenAI they imagined across the different phases.

ing someone play a video game you don't know how to play" (R8). R7 speaks to how this GenAI can "*lessen the learning curve*" as students transition to trainees, exposing them to various pathologies before reviewing real patients. R8 reflected on how the tool could help medical students grasp foundational concepts and prepare for exams: "*...being able to play with it...as you're doing questions, 'Hey, here's what a pneumothorax looks like under tension, under no tension...I think would be invaluable.*"

Despite being able to similarly perform these interactions through Radiopaedia or StatDx, the students instead mentioned relying on other resources, such as UWorld or Amboss, for licensing exam prep.

Expedite Learning through Variations, Differentials, and Rare Cases. As medical students transition into their training program, it becomes increasingly important to get as much exposure as possible to a wide variety of cases. As R1 says, "*the more you see, the better radiologist you will become*". However, despite relying on existing resources populated by the community and their training to see as many diverse images as possible, they are limited by what others *can* share based on patient privacy regulations or what patient cases come through at their hospital. Radiology trainee, R2, highlighted how this GenAI could accelerate their learning by allowing them to study variations of diseases: "*...this would expedite that if I could just be like, 'okay, I need to know the difference between this, this, and this'...I can just generate 5 different variances of those diseases on this tool with the different manifestations of those variants.*"

R2 exercises this during the interview session by prompting the model to generate '`lung cancer`', and then refines their prompt to generate a different variation of lung cancer. In their review of their generations, they put the two generated CT scans side-by-side. R5 similarly experiments with the granularity of the prompts by adding '`3cm`' to their initial prompt '`irregular speculated lesion...`'.

As many diseases can mimic each other on a CT scan, it is important to learn not only through the variations of how a disease can manifest but also through differentials of how the pathology of one disease is different from another. For example, R6 emphasized how this GenAI could make positive findings more accessible in a low-risk environment: "*Being able to see those positive findings is more educational than commenting on negative findings...sometimes we need to see, like...maybe more than 200 cases to see one positive finding...But if I can be able to generate those positive findings by just giving my impressions...I think that's something that can be very helpful.*"

Beyond variations and differentials, R8 imagined how this GenAI could be applied in day-to-day training scenarios to prepare for region-specific diseases. For example, they described a scenario where generated CT scans could be "*...included in my docket of cases I have to look at for the day and one or two of them might be training, region-specific training*". They similarly mentioned the tool's value in training medical professionals in low-resource environments, such as being a "*...benefit to even Africa...And being able to say, 'Okay, here's what Cardiomegaly looks like with someone who might have tuberculosis.*'"

Empowering Trainees when On-Call or in Emergency Situations. Unlike clinical practice, where attending radiologists rely on their accumulated expertise, training is based on an apprenticeship model where trainees are closely supervised by their seniors. As trainees become more senior, they are increasingly put in situations where they may need to read cases in emergencies or on-call situations, where there is little to no time for doubt or waiting for an expert's approval. These situations demand quick and accurate decision-making, making tools like text-to-CT Scan GenAI an invaluable resource for trainees. Participants highlighted how this tool could serve as a critical support system during high-pressure scenarios. For instance, R2 described how it could aid in trauma cases by enabling them to create a checklist of generated CT scans based on diseases to look out for:

> "*...when you're on trauma, you have a trauma come in...if I had this tool up and I had like 5 different generated pictures of the 5 things I'm checking for when a trauma case comes in...and then when the surgery team is there with me, 'Hey, these are kind of like the things I'm looking out for', while we're searching through the image together*".

R2 expands this idea by imagining how they could use the tool to generate baseline CT scans to make sure they do not miss anything while on-call. R7 imagines how generating images for severe conditions ahead of shifts in the ER could accelerate trainees' decision-making, empowering them to handle emergency situations appropriately and promptly: "*...if you can generate your own images of what different aortic dissections look like, then, when you have a patient that comes into the ER with ripping chest pain, and you take them to CT, and you see, like, 'Oh, sh*t! This is really bad, I need to call someone for help now', as opposed* *to 'Okay, I haven't seen this before, but this sounds like what I think it could be like. Let me wait and see if I can get CT to read this, and then I have to call my attending', and that all adds time in an emergency situation...time is life.*" R7 then prompted the model to generate an aortic dissection.

Promises of T2I in Clinical Practice

While GenAI can have a high potential for medical education and training, it remains to be seen whether and how it can augment clinical workflows. Clinical practice is slightly different from training because during training, the junior trainees, "*...are reading every case with the attending*" (R6). Our study further addresses *RQ1* by identifying how text-to-CT Scan GenAI can augment attending radiologists in clinical practice, including (**1**) improving report impressions of medical images, (**2**) planning and communicating with others, and (**3**) supplementing the existing resources they rely on when faced with challenging cases.

Improving Report Impression of Medical Images. Interpreting medical images to identify disease progression or abnormalities often depends on differential diagnosis and validating speculations. Participants highlighted how text-to-CT Scan GenAI could play a pivotal role in supporting these tasks by generating CT scans based on their interpretation to visualize hypotheses and refine interpretations. For example, R1 described how the GenAI could assist in determining conditions like cardiomegaly by providing reference images to compare against patient scans. Beyond aiding in diagnosis, the tool could support the refinement of more precise clinical impressions, which participants discovered as they continued generating more CT Scans. R5 described how the generation of reference images could refine the specificity of their observations, "*Instead of just right pleural effusion, I can compare the image and then specify which segment I expect the pleural effusion to be in*". Using this process, R5 reflected on how generating CT scans "*...can be useful to revise my diagnosis to add more details"* and "*...use the generation for differential diagnosis*".

The comparative functionality in the interface allowed R6 to discover that their impression of the original patient case presented to them was not completely correct when they compared the image findings in the generated image to the provided real CT: "*So, I said, smaller effusion, and sometimes people say mild effusion, some say severe effusion. But so this generated AI gives me a very small effusion. But* ***now I'm thinking this might be a mild effusion instead of a small effusion****, which can be very helpful actually...Something like when we use those subjective words and if we have something where somethings objective, could help us a lot actually.*"

Together, these scenarios underscore the potential of text-to-CT Scan GenAI to enhance the accuracy, efficiency, and precision of image interpretation in clinical practice, empowering clinicians while reviewing both routine and challenging cases. However, it also raises implications for accountability, confirmation bias, and unintentional deskilling.

Planning & Communicating with Others. An interesting use case outside of direct diagnosis or education that was mentioned was in planning and communicating with other stakeholders. For example, R2 described how generating CT scans could be "*...**super helpful in surgical planning**...you can show the patient like, 'hey, this is going to be what you look like post-surgery.*" In a similar vein, R8 describes how they could imagine a scenario where they show a patient how their condition looks compared to a normal condition.

Radiologists are not the only ones who will be interacting with medical imaging throughout the clinical workflow. This could introduce multiple impressions that may disagree about a diagnosis. As such, R5 brings up how "***These generated images can act as additional evidence when I am communicating my diagnosis to other doctors***." R7 motivates R5's point, noting how some surgeons they have shadowed will "*...trust their own reading of the image over whatever the official [radiologist's] read is*." This is an interesting dynamic in the clinical workflow where generated CT scans could aid in these collaborative workflows, helping radiologists provide validations of their reports by generating CTs that show differentials for other physicians to read.

Visual Memory Support. Radiologists are constantly learning beyond their formal training programs, often needing to recall visual features of pathologies they may not have encountered recently. As R4 describes it, "*I think the more it passes from graduation, yeah, the more you forget.*" As a result, participants highlighted how text-to-CT Scan GenAI could serve them "*by strengthening our visual memory*" (R4). This is particularly valuable for general radiologists working across different specialties where they encounter CT scans of the chest along with other regions of the body. As R5 explained: "*I'm working in a different field of radiology. If I am specialized in one branch, and I see a different CT scan, this can be useful to remember something you haven't seen in a long time.*" R5 goes on to say "*It would be easier to use this tool instead of Radiopaedia when I was a student, I would just type my prompt and see an image much faster than having to search for the right case*. R3 and R4 similarly conclude this could be used as "*...an alternative to looking something up on the Internet.*"

Risks & Challenges with Medical T2I GenAI

As a result of key stakeholders generating CT Scans for pathologies beyond what was shown to be validated by Xu et al. (2024), we discovered technical challenges and domain-specific risks, addressing *RQ2*. This section is separated into two parts: (**1**) technical challenges and (**2**) domain-specific risks, each with sub-themes shown in Table **3**. These groupings were identified from a thematic analysis of quotes related to model behavior and implications of model use in practice. For the interpretation of our findings, technical challenges refer to model development, and domain-specific risks refer to downstream impacts from model use. Such risks and challenges may act as barriers to the adoption of medical T2I GenAI if left unaddressed.

Technical Challenges. Impressively, the model was able to generate pathologies that were not tested in the original study (Xu et al. 2024). However, there are four primary technical and AI safety challenges that arose while participants

Category	Challenges	Participants
Technical Challenges	Generated CT scan resolution quality	R1, R3, R6, R7, R9
	Transparency and Hallucinations	R3, R4, R9
	Contrast-Enhanced CTs	R3, R4, R9
	Generation Expectations	R3–R5, R9
Domain-Specific Risks	Confirmation Bias & Distilling Incorrect Information	R2, R3, R6, R7, R9
	Over-exposure to synthetic CT scans during training	R3, R5, R7

Table 3: *RQ2*: Participants mapped to which technical challenges and domain-specific risks that they brought up.

were generating and exploring CT scans: (**1**) the resolution quality of the CT Scans, (**2**) the approach to validation and hallucinations, (**3**) generating contrast-enhanced CTs, and (**4**) technical expectations of the model generation.

Five participants commented on the resolution of the generated CT scans being of lower quality, even though the model we used was designed to develop high-resolution CTs. While looking through their generated CT scan, R7 found it difficult to see if the pathology they asked for was there "*because of the resolution*". They described the experience as if they were looking at a CT scan on a really bad computer instead of the expensive computers that the radiologists use, while R9 described it looking "*like they moved during*" the CT scan. R6 similarly asked during their session if they had a bad internet connection or if it was just the generation quality. However, R3 clarified that a radiologist may not necessarily require a high-resolution CT scan as long as it offers the required information to support the goal of the generation: "*It's not necessary to fool me into thinking it's a real CT. What you want it to do is* ***provide a sufficient example that I would recognize this on a real CT***."

Second, several participants encountered flaws in the generated output as they were prompting the model on image findings that were not tested for by the quantitative model evaluations. For example, R3 spent time closely auditing the outputs, commenting that the model has "*...thrown all sorts of pathology onto this patient...*" including "*...metastatic bone disease...and a plural recess down here that doesn't connect to anything*." R4 and R9 similarly identified missing or injected pathologies in their generations. A vast majority of the ML literature refers to this type of model behavior as hallucinations, which come with domain-specific risks. As with any black box AI, it is important that users can see which part of the prompts are associated with which regions of the generation. This becomes especially critical and challenging with a 3D-generated output where multiple pathologies are combined within a single prompt. This task, however, is computationally expensive and underexplored for text-to-image GenAI (Bradshaw et al. 2025).

Third, R3's prompt ("`Necrotizing Pneumonia on contrast-enhanced CT`") asked for a contrast-enhanced CT scan, which is a specific type of CT scan that shows reactions of a fluid given to a patient before the CT scan (Mayo Clinic 2023), to probe the model's limitations further. As discovered, the model currently cannot produce contrast-enhanced CT scans, as it was not trained on contrast-enhanced CT scans. R4 and R9 also acknowledged this limitation when they generated something that they expected to be shown on a contrast-enhanced CT. Without discussing with radiologists the importance of contrast-enhanced CTs in accurately determining certain image findings, models may be trained on data that does not represent what radiologists would benefit from. Similar to this, R1 and R6 mention how report impressions can sometimes be subjective (*i.e.*, how severe a disease actually is), which is especially consequential in training models depending on whether the reports are sourced from trainees or experts. R9 mentioned how they struggled to get a certain pathology to be of less severity, and they concluded that "*maybe all the patients you guys had were more like outpatient.*"

Lastly, despite the GenAI's potential value, R3 and R9 acknowledge that the generation time delay is substantial (3 minutes), especially if comparing it to searching for examples on Radiopaedia. R9 said that as residents, they have "*...limited time while you're reading images to do research.*" Beyond generation time, several participants commented on the GenAI's limitations compared to content provided by existing resources, such as Radiopaedia. For example, R5 contrasts the model & interface to how "*Radiopaedia also has some additional notes [about the image finding] which can be helpful for differential diagnosis*." These resources allow them to parse detailed descriptions, aiding in their conclusion about a diagnosis, whereas the GenAI is limited to taking a single prompt at a time to generate a CT scan and does not provide any detailed description in return.

Domain-Specific Risks The development of medical T2I that addresses stakeholders' needs and concerns goes beyond solving technical challenges. We must also consider the complex interplay between user expectations and domain-specific implications. Our findings revealed two domain-specific risks that need to be considered to ensure that T2I GenAI meets the needs of the key stakeholders and aligns with complex workflows: (**1**) mitigating confirmation bias and the distillation of incorrect information and (**2**) supporting continuous exposure to real medical images.

First, participants were unsure if and where the requested pathology from the prompt was generated (R2, R6, R7, R9). While reviewing the generations, they would say something like "*Maybe this is what they [AI] wanted*" (R2) or "*I potentially see that dissection*" (R7). This type of behavior with AI, masked as confirmation bias, can be detrimental to a trainee's learning progress. R9 highlighted potential consequences: "*If you're learning and then you get like false information, uh oh, you know, that's confusing.* ***And you don't know if you're wrong or [the] software is wrong***." In the use case of practice exams, these unprompted pathologies could distract more than support the student (R3). Ultimately, it is possible for students generating CT scans who are unaware of the additional, missing, or misrepresented pathologies to unintentionally learn the wrong things. For example, after R7 generated a CT scan for aortic dissection, they describe a potential issue related to confirmation bias: "***Am I seeing what I'm actually seeing? Or am I seeing what I want to see?***" To this point, R7 emphasizes that research would need to "*...make sure that this is something that's not going to...indirectly, [give] poor quality education.*"

Second, despite the participants acknowledging that the generated CT scans were of lower resolution, they were surprised by the fact that they were looking at a generated CT scan. R7 remarked, "***If you hadn't told me I wouldn't have known [that the CT was AI generated]***." F2 reacted similarly, saying "*Oh wow that's really good because I couldn't tell which one is from which*." R5 stated "*With the right prompt, the quality of the output is so good that I cannot distinguish that this is AI, if I did not create the prompt myself*." This in itself raises potentially consequential implications as R5 states that "*Generated images can be useful for experienced doctors who have seen thousands of images, but **residents need to develop experience with real patient cases and images***." From a lecturer's perspective, the students want "*...to see real examples with validated pathology*" (R3). This underscores the ongoing theme that radiologists and those in the broader medical field need to continue seeing a plethora of real CT scans throughout their careers, given that the ones they need to interpret are real patients.

Discussion

Building advanced models, such as the text-to-CT Scan GenAI, often occurs without a deep understanding of key stakeholders' needs and potential safety risks. This lack of clarity raises important questions about misuse, misinformation, and the potential erosion of expert skills, particularly in high-stakes environments such as medicine (Karabacak et al. 2023; Ooi et al. 2023). Below, we discuss the benefits of our study design and the implications associated with the use of text-to-image GenAI for radiology

Evaluating Domain-Specific Text-to-Image GenAI

Human-Centered Evaluations of T2I GenAI with Stakeholders. As an outcome of conducting exploratory model prompting with key stakeholders, we observed how each participant in our study was interested in exploring and evaluating a different capability of the model. Unlike the formative interviews, where users could not prompt the model, the exploratory model prompting study design allowed us to uncover numerous technical challenges and domain-specific risks. For example, it helped expose that the thumbs-up/down feedback mechanisms offered for each generated image in the interface did not address what participants really wanted to audit, which was the ability to generate contrast-enhanced CT scans, the resolution of the images (similar to Brisco, Hay, and Dhami (2023)'s observation), or the misrepresentation of image findings.

We also learned that CT scan impressions can be subjective, making it important to be concerned about the expertise of the radiologists who wrote the reports for the training data. Using reports by trainees or unverified by multiple radiologists could introduce inconsistencies that can impact the reliability of the model. All of these challenges are difficult to quantitatively capture; however, conducting exploratory human-centered AI evaluations to simultaneously elicit this will be useful to the advancement of RAI.

Lastly, unlike many existing risk taxonomies and evaluations for T2I models, we observe a specific set of risks that the radiologists were primarily concerned about: distilling incorrect information and losing exposure to real CT scans during training. These risks represent unintentional harms that could come about from this technology being widely available in practice. It is interesting that more granular risks or harms, such as various forms of bias, were not acknowledged by the participants. Despite this, it will be critical to identify domain-specific biases that arise with domain-specific T2I models and innovate solutions to mitigate them.

Implications of Text-to-CT Scan GenAI in Medical Education, Training, & Practice

We allude to *RQ1* as we discuss the implications of T2I for medical education, training, and practice. Our participants emphasized the role that online resources and pure experience play when faced with challenging cases and professional growth, as similarly observed by Yildirim et al. (2024) and Rassie (2017). El-Ali et al. (2019) and Quaia (2023) have observed an increase in students' and trainees' performance when provided access to radiology-specific repositories, emphasizing the importance of supplemental resources to traditional lecture materials. Extending these insights, five participants referenced how getting exposure to variations can play an important role in competency within medical imaging-based professions. Generating variations as a design has similarly been explored by (Gero et al. 2024), highlighting the value that variation plays in sensemaking. However, the design of our GenAI plugin did not easily enable the generation and visualization of variations for a single prompt. This was also difficult due to the generation time for a single CT scan. As Weisz et al. (2024) suggest, it is important to design GenAI applications to enable variability and expose differences from a prompt. While this generates variability aimed at meeting the user's needs, we also see how this can expedite education by showing several CT scans with the same pathology in several variations. This would help prepare students to practice identifying serious conditions before dealing with real patients, a benefit that del Cerro and Esteve (2023) similarly mention.

Our participants also stressed the importance of being able to understand what real human anatomy is like by looking at thousands of real CT scans. This opens the door for RAI work to explore various design opportunities for domain-specific T2I models, such as inpainting capabilities. Furthermore, RAI can explore the human-centered needs around explicitly identifying which case in the viewing window is generated and which is real to avoid misuse and abuse of the GenAI capabilities (del Cerro and Esteve 2023).

Considering Collaborative Workflows. Our participants highlighted the collaborative workflows that take place within medicine and how text-to-CT Scan GenAI could provide radiologists an opportunity to provide generations as rationalizations for their reading to enhance their collaboration with surgeons. Yildirim et al. (2024) also saw the value of showing similar images as a use case for reassurance between radiologists and other medical professionals. Such an interaction throughout collaborative workflows might require designs that allow radiologists to leave additional notes or engage in an asynchronous discussion with the surgeon about their generations and patient case impressions.

Safety Implications of Medical T2I GenAI

With any powerful text-to-image GenAI model, it is necessary to consider the broader ethical and safety implications. We allude to *RQ2* while discussing design implications for safeguarding against hallucinations and confirmation bias, as well as addressing misuse of the generated CT scans.

Safeguard Against Hallucinations & Confirmation Bias. As we have seen through our own experiences and previous work, hallucinations are prevalent in GenAI systems. This is especially concerning because primarily the more experienced participants were commenting on misrepresentations of the prompt in their generated CT scan. This motivates future works to explore ways to express hallucinations to users, especially to novice users. Along with hallucinations, some participants raised concerns about confirmation bias, which can emerge when working with AI (Rosbach et al. 2025), especially among students and junior trainees. While it may be hard to avoid generating hallucinations, we can design mechanisms to safeguard end-users from confirmation bias during their learning experience. Designing mechanisms to minimize confirmation bias is especially important in an educational environment because recent work has shown how people can start reasoning incorrectly when given misleading explanations generated by AI (Spitzer et al. 2024). These designs are especially important to avoid deskilling, slowing professional growth, and minimizing over-reliance. One potential solution that we see is by developing visual explanations that associate concepts from the prompt to regions of the generated image, such as the technical solutions proposed by Chefer et al. (2023), Bradshaw et al. (2025), and Evirgen, Wang, and Chen (2024).

Preventing Misuse and Malpractice. Although this was not an emerging theme from the interviews, it is necessary to anticipate potential misuse and malpractice of synthetic medical images. Koohi-Moghadam and Bae (2023) suggests the potential for AI-generated medical images to fabricate clinical results or patients' conditions, which can lead to inaccurate conclusions or mistreating patients. The misuse and malpractice scenarios could go on, but it is more important to discuss practices to prevent such scenarios from becoming prevalent. This especially becomes important when these models increase competency by being able to do tasks such as inpainting original patient CT scans. Horvitz (2022) proposes watermarking as a potential solution to distinguish synthetic from real content, promoting transparency.

Limitations & Future Work

This work comes with limitations that future work can build upon. First, our study involved a small sample size for both the formative study and semi-structured sessions. However, this sample size choice is restricted by domain stakeholders' availability and respect for their time. Additionally, our participants all had different levels of experience and perspectives about AI, which could have tainted their perspectives on the promises, risks, and challenges that they identified. While we had briefly asked participants about their prior experiences with AI in the pre-probe sessions, we have not done an extensive analysis of that data. Also, our study is limited to a very domain-specific T2I for chest CT scans, which could potentially limit how our conclusions translate to other medical text-to-image GenAI. Although not all participants were chest specialists, we felt that they were still able to successfully identify valid promises, risks, and challenges even beyond chest CTs. For example, two participants were able to see how this text-to-image generation process could generalize to other imaging modalities.

Additionally, we did not conduct a detailed analysis of all the prompts generated by participants, as this was not our primary goal with this study. Furthermore, we have a limited number of prompts/generations from participants, as each generation took between three and four minutes. We did not intend to further evaluate the model's performance or identify heuristics for prompts that work better with this model for this study. These areas remain for future exploration to better understand how key stakeholders would prompt generative models depending on their goals.

Similarly, we did not perform quantitative evaluations of the accuracy of the generated CT scans from the radiologists' perspective during these sessions. Future work should consider including quantitative model auditing activities with stakeholders using familiar medical interfaces and tools to maintain external validity and to ensure that evaluations appropriately target stakeholders' concerns.

Conclusion

Medical text-to-image GenAI capabilities are increasingly being investigated as models become increasingly competent. Given the lack of human-centered research involving key stakeholders on the promises, risks, and challenges of medical text-to-image GenAI, our work seeks to bridge this gap by conducting exploratory model prompting sessions with a text-to-CT Scan generative model. Our study highlights the opportunities of text-to-image GenAI in medical education, training, and practice, such as empowering students to learn in low-stakes environments and augmenting radiologists when faced with challenging cases. These promising opportunities expose technical challenges and domain-specific risks of domain-specific T2I models that the RAI community can consider, such as confirmation bias. Overall, our work represents the importance of evaluating T2I models from a human-centered lens, taking domain-specific needs and challenges into consideration. We hope this work encourages RAI researchers to further explore the risks and challenges of domain-specific T2I models.

Acknowledgments

Thank you to Dr. Ahmed El-Sherief and Dr. Samir Shah for insightful informal conversations about the implications of our work. Thank you to everyone who helped us recruit participants. Thank you to Wesley Deng for providing feedback on the draft manuscript. ChatGPT4, Writeful, and Grammarly were utilized throughout the sections of this paper to improve the authors' writing, such as grammar, clarity, flow, and spelling.

References

Bhardwaj, G.; Govindarajulu, Y.; Narayanan, S.; Kulkarni, P.; and Parmar, M. 2023. On the notion of Hallucinations from the lens of Bias and Validity in Synthetic CXR Images. *arXiv preprint arXiv:2312.06979*.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504.

Bird, C.; Ungless, E.; and Kasirzadeh, A. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 396–410.

Bradshaw, A.; Morrison, K.; Mathur, A.; Dai, W.; Eslami, M.; Perer, A.; et al. 2025. Toward Interpretable 3D Diffusion in Radiology: Token-Wise Attribution for Text-to-CT Synthesis. In *Medical Imaging with Deep Learning-Short Papers*.

Brisco, R.; Hay, L.; and Dhami, S. 2023. Exploring the role of text-to-image AI in concept generation. *Proceedings of the Design Society*, 3: 1835–1844.

Cai, C. J.; Winter, S.; Steiner, D.; Wilcox, L.; and Terry, M. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW): 1–24.

Chan, C. K. Y. 2023. A comprehensive AI policy education framework for university teaching and learning. *International journal of educational technology in higher education*, 20(1): 38.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–10.

Chen, W.; Wang, P.; Ren, H.; Sun, L.; Li, Q.; Yuan, Y.; and Li, X. 2024. Medical image synthesis via fine-grained image-text alignment and anatomy-pathology prompting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 240–250. Springer.

Chen, X.; Burke, J.; Du, R.; Hong, M. K.; Jacobs, J.; Laban, P.; Li, D.; Peng, N.; Willis, K. D.; Wu, C.-S.; et al. 2023. Next steps for human-centered generative ai: A technical perspective. *arXiv preprint arXiv:2306.15774*.

Collins, K. M.; Kim, N.; Bitton, Y.; Rieser, V.; Omidshafiei, S.; Hu, Y.; Chen, S.; Dutta, S.; Chang, M.; Lee, K.; et al. 2024. Beyond Thumbs Up/Down: Untangling Challenges of Fine-Grained Feedback for Text-to-Image Generation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 293–303.

Currie, G. M.; Hawk, K. E.; and Rohren, E. M. 2024. Generative artificial intelligence biases, limitations and risks in nuclear medicine: an argument for appropriate use framework and recommendations. In *Seminars in Nuclear Medicine*. Elsevier.

del Cerro, J. M.; and Esteve, A. V. 2023. How Imaging Generative AI Will Transform the Medical Radiological Practice. *HealthManagement.org: The Journal*, 23(2): 128–137. For full references, email edito@healthmanagement.org.

Deng, W. H.; Claire, W.; Han, H. Z.; Hong, J. I.; Holstein, K.; and Eslami, M. 2025. WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI. *arXiv preprint arXiv:2501.01397*.

El-Ali, A.; Kamal, F.; Cabral, C. L.; and Squires, J. H. 2019. Comparison of Traditional and Web-Based Medical Student Teaching by Radiology Residents. *Journal of the American College of Radiology*, 16(4): 492–495.

Evirgen, N.; Wang, R.; and Chen, X. 2024. From Text to Pixels: Enhancing User Understanding through Text-to-Image Model Explanations. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 74–87.

Gero, K. I.; Swoopes, C.; Gu, Z.; Kummerfeld, J. K.; and Glassman, E. L. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–21.

Gu, J. 2024. A Survey on Responsible Generative AI: What to Generate and What Not. *arXiv preprint arXiv:2404.05783*.

Hamamci, I. E.; Er, S.; Almas, F.; Simsek, A. G.; Esirgun, S. N.; Dogan, I.; Dasdelen, M. F.; Wittmann, B.; Simsar, E.; Simsar, M.; et al. 2024. A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv:2403.17834*.

Hamamci, I. E.; Er, S.; and Menze, B. 2024. Ct2rep: Automated radiology report generation for 3d medical imaging. *arXiv preprint arXiv:2403.06801*.

Hamamci, I. E.; Er, S.; Sekuboyina, A.; Simsar, E.; Tezcan, A.; Simsek, A. G.; Esirgun, S. N.; Almas, F.; Dogan, I.; Dasdelen, M. F.; et al. 2023. GenerateCT: Text-Conditional Generation of 3D Chest CT Volumes. *arXiv preprint arXiv:2305.16037*.

Hao, S.; Shelby, R.; Liu, Y.; Srinivasan, H.; Bhutani, M.; Ayan, B. K.; Poplin, R.; Poddar, S.; and Laszlo, S. 2024. Harm amplification in text-to-image models. *arXiv preprint arXiv:2402.01787*.

Horvitz, E. 2022. On the horizon: Interactive and compositional deepfakes. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, 653–661.

Karabacak, M.; Ozkara, B. B.; Margetis, K.; Wintermark, M.; and Bisdas, S. 2023. The advent of generative language models in medical education. *JMIR Medical Education*, 9: e48163.

Koljonen, V. 2023. What could we make of AI in plastic surgery education. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 81: 94–96.

Koohi-Moghadam, M.; and Bae, K. T. 2023. Generative AI in medical imaging: applications, challenges, and ethics. *Journal of Medical Systems*, 47(1): 94.

Ku, M.; Jiang, D.; Wei, C.; Yue, X.; and Chen, W. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*.

Kumar, A.; Burr, P.; Young, T. M.; et al. 2024. Using AI text-to-image generation to create novel illustrations for medical education: current limitations as illustrated by hypothyroidism and horner syndrome. *JMIR Medical Education*, 10(1): e52155.

Lee, T.; Yasunaga, M.; Meng, C.; Mai, Y.; Park, J. S.; Gupta, A.; Zhang, Y.; Narayanan, D.; Teufel, H.; Bellagente, M.; et al. 2023. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36: 69981–70011.

Levac, B.; Kumar, S.; Jalal, A.; and Tamir, J. I. 2024. Accelerated motion correction with deep generative diffusion models. *Magnetic Resonance in Medicine*, 92(2): 853–868.

Li, L.; Shi, Z.; Hu, X.; Dong, B.; Qin, Y.; Liu, X.; Sheng, L.; and Shao, J. ???? Benchmarking Ethics in Text-to-Image Models: A Holistic Dataset and Evaluator for Fairness, Toxicity, and Privacy.

Lin, Z.; Pathak, D.; Li, B.; Li, J.; Xia, X.; Neubig, G.; Zhang, P.; and Ramanan, D. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 366–384. Springer.

Mai, M.; Mirza, F. N.; and DiMarco, C. 2024. Application of text-to-image translation algorithms in medicine: a systematic review. *JAAD Reviews*.

Mayo Clinic. 2023. CT scan. Accessed: 2025-01-20.

Moin, K. A.; Nasir, A. A.; Petroff, D. J.; Loveless, B. A.; Moshirfar, O. A.; Hoopes, P. C.; Moshirfar, M.; Moshirfar, O.; and Hoopes Sr, P. 2024. Assessment of generative artificial intelligence (AI) models in creating medical illustrations for various corneal transplant procedures. *Cureus*, 16(8).

Naik, R.; and Nushi, B. 2023. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 786–808.

Noel, G. P. 2024. Evaluating AI-powered text-to-image generators for anatomical illustration: a comparative study. *Anatomical Sciences Education*, 17(5): 979–983.

Ooi, K.-B.; Tan, G. W.-H.; Al-Emran, M.; Al-Sharafi, M. A.; Capatina, A.; Chakraborty, A.; Dwivedi, Y. K.; Huang, T.-L.; Kar, A. K.; Lee, V.-H.; et al. 2023. The potential of generative artificial intelligence across disciplines: Perspectives and future directions. *Journal of Computer Information Systems*, 1–32.

Parrish, A.; Kirk, H. R.; Quaye, J.; Rastogi, C.; Bartolo, M.; Inel, O.; Ciro, J.; Mosquera, R.; Howard, A.; Cukierski, W.; et al. 2023. Adversarial nibbler: A data-centric challenge for improving the safety of text-to-image models. *arXiv preprint arXiv:2305.14384*.

Pérez-García, F.; Bond-Taylor, S.; Sanchez, P. P.; van Breugel, B.; Castro, D. C.; Sharma, H.; Salvatelli, V.; Wetscherek, M. T.; Richardson, H.; Lungren, M. P.; et al. 2025. RadEdit: stress-testing biomedical vision models via diffusion image editing. In *European Conference on Computer Vision*, 358–376. Springer.

Quaia, E. 2023. Electronic platform-based education for radiology residents: results of a two-year survey. *BMC Medical Education*, 23(1): 202.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rassie, K. 2017. The apprenticeship model of clinical medical education: time for structural change. *The New Zealand Medical Journal (Online)*, 130(1461): 66.

Rosbach, E.; Ammeling, J.; Krügel, S.; Kießig, A.; Fritz, A.; Ganz, J.; Puget, C.; Donovan, T.; Klang, A.; Köller, M. C.; et al. 2025. " When Two Wrongs Don't Make a Right"-Examining Confirmation Bias and the Role of Time Pressure During Human-AI Collaboration in Computational Pathology. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.

Safadi, M. F.; Zayegh, O.; and Hawoot, Z. 2024. Advancing Innovation in Medical Presentations: A Guide for Medical Educators to Use Images Generated With Artificial Intelligence. *Cureus*, 16(12).

Sivaraman, V.; Morrison, K.; Epperson, W.; and Perer, A. 2025. Over-Relying on Reliance: Towards Realistic Evaluations of AI-Based Clinical Decision Support. *arXiv preprint arXiv:2504.07423*.

Sloan, P.; Clatworthy, P.; Simpson, E.; and Mirmehdi, M. 2024. Automated Radiology Report Generation: A Review of Recent Advances. *IEEE Reviews in Biomedical Engineering*, 1–20.

Sorin, V.; Barash, Y.; Konen, E.; and Klang, E. 2020. Creating artificial images for radiology applications using generative adversarial networks (GANs)–a systematic review. *Academic radiology*, 27(8): 1175–1185.

Spitzer, P.; Holstein, J.; Morrison, K.; Holstein, K.; Satzger, G.; and Kühl, N. 2024. Don't be Fooled: The Misinformation Effect of Explanations in Human-AI Collaboration. *arXiv preprint arXiv:2409.12809*.

Temsah, M.-H.; Alhuzaimi, A. N.; Almansour, M.; Aljamaan, F.; Alhasan, K.; Batarfi, M. A.; Altamimi, I.; Alharbi, A.; Alsuhaibani, A. A.; Alwakeel, L.; et al. 2024. Art or artifact: evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL· E 3 for illustrating congenital heart diseases. *Journal of Medical Systems*, 48(1): 54.

Thesen, T.; Alilonu, N. A.; and Stone, S. 2024. AI Patient Actor: An Open-Access Generative-AI App for Communication Training in Health Professions. *Medical Science Educator*, 1–3.

Waikel, R. L.; Othman, A. A.; Patel, T.; Hanchard, S. L.; Hu, P.; Tekendo-Ngongang, C.; Duong, D.; and Solomon, B. D. 2023. Generative Methods for Pediatric Genetics Education. *medRxiv*.

Waisberg, E.; Ong, J.; Masalkhi, M.; Zaman, N.; Kamran, S. A.; Sarker, P.; Lee, A. G.; and Tavakkoli, A. 2023. Text-to-image artificial intelligence to aid clinicians in perceiving unique neuro-ophthalmic visual phenomena. *Irish Journal of Medical Science (1971-)*, 192(6): 3139–3142.

Weisz, J. D.; He, J.; Muller, M.; Hoefer, G.; Miles, R.; and Geyer, W. 2024. Design Principles for Generative AI Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–22.

Xu, Y.; Sun, L.; Peng, W.; Jia, S.; Morrison, K.; Perer, A.; Zandifar, A.; Visweswaran, S.; Eslami, M.; and Batmanghelich, K. 2024. Medsyn: Text-guided anatomy-aware synthesis of high-fidelity 3d ct images. *IEEE Transactions on Medical Imaging*.

Yildirim, N.; Richardson, H.; Wetscherek, M. T.; Bajwa, J.; Jacob, J.; Pinnock, M. A.; Harris, S.; Coelho De Castro, D.; Bannur, S.; Hyland, S.; et al. 2024. Multimodal healthcare AI: identifying and designing clinically relevant vision-language applications for radiology. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–22.

Yousefirizi, F.; Decazes, P.; Amyar, A.; Ruan, S.; Saboury, B.; and Rahmim, A. 2022. AI-Based Detection, Classification and Prediction/Prognosis in Medical Imaging: Towards Radiophenomics. arXiv:2110.10332.

Zhu, L.; Mou, W.; Wu, K.; Zhang, J.; and Luo, P. 2024. Can DALL-E 3 Reliably Generate 12-Lead ECGs and Teaching Illustrations? *Cureus*, 16(1).

Ziegler, E.; Urban, T.; Brown, D.; Petts, J.; Pieper, S. D.; Lewis, R.; Hafey, C.; and Harris, G. J. 2020. Open health imaging foundation viewer: an extensible open-source framework for building web-based imaging applications to support cancer research. *JCO clinical cancer informatics*, 4: 336–345.