Artificial Intelligence in the Life Sciences 8 (2025) 100145

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Hallucinations in medical devices☆

Jason Granstedt *, Prabhat Kc, Rucha Deshpande, Victor Garcia, Aldo Badano

Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD 20993, United States of America

ARTICLE INFO

Keywords:
Hallucinations
Deep learning
Artificial intelligence
Generative models
Medical imaging

ABSTRACT

Computer methods in medical devices are frequently imperfect and are known to produce errors in clinical or diagnostic tasks. However, when deep learning and data-based approaches yield output that exhibit errors, the devices are frequently said to hallucinate. Drawing from theoretical developments and empirical studies in multiple medical device areas, we introduce a practical and universal definition that denotes hallucinations as a type of error that is plausible and can be either impactful or benign to the task at hand. The definition aims at facilitating the evaluation of medical devices that suffer from hallucinations across product areas. Using examples from imaging and non-imaging applications, we explore how the proposed definition relates to evaluation methodologies and discuss existing approaches for minimizing the prevalence of hallucinations.

1. Introduction

The phenomenon of hallucinations within AI systems can adversely affect the efficacy of algorithmic applications by diminishing user trust and introducing safety hazards in critical contexts. In other contexts, hallucinations may offer advantages, such as in the creation of innovative content or the production of synthetic data for model training. Hallucinations pose substantial challenges particularly in high-stakes applications where accuracy is imperative. Within AI applications in medical devices, hallucinations may influence clinical decision-making and potentially jeopardize patient outcomes through diagnostic or therapeutic errors. Despite the concept of hallucination having been introduced to the scholarly community about a decade ago, a definitive and universally recognized definition pertaining to hallucinations in medical devices is currently absent. This article delineates an approach designed to provide a clear context for referring to hallucinations in outputs of AI medical applications, thereby aiding in the assessment and prevention of such phenomena within the methodologies for medical device evaluation.

Recently, Xu et al. offered a more pragmatic approach to defining hallucinations with a theoretical framework in which hallucinations are delineated as the discrepancies between generated outputs and a ground truth function [1]. By leveraging learning theory, they elucidate that hallucination is inherently unavoidable and that the complete eradication of hallucinations from real-world large language models (LLMs) is not feasible. Expanding upon Xu's framework, we propose to subset errors as either hallucinations or non-hallucinations. Hallucinations are identified as plausible errors with two distinct subtypes: (1) impactful hallucinations and (2) benign hallucinations. Plausible errors refer to device outputs which are erroneous but may be visually or linguistically perceived as truth such that readers may not recognize them as errors. Impactful hallucinations negatively impact device performance, whereas benign hallucinations have no significant effect. As an example, consider a reconstructed image for a patient with a cough where the model adds a structure that can be perceived as a connection between two organs, such as a tracheoesophageal fistula. Such an error would be an impactful hallucination as it may lead to a change in patient diagnosis and management. However, if the model instead added a small gas bubble within the small intestine, the hallucination would be benign as it is unlikely to be perceived by a clinician or change patient management. Additionally, there exist non-hallucination errors, which are characterized by their obviousness and traceability to device artifacts, such as Gibbs ringing or aliasing, or pre-specified failure modes. To determine whether an error is a hallucination or a non-hallucination error, we consider the nature of both the assessment task and device user. This definition does not specify the type of user, and the determination of whether an error is plausible or subtle is contingent upon the nature and level of expertise of the user which, according to the intent of the evaluation framework, could be a domain expert, a naive user, or in some cases, an algorithmic interpreter. This approach to defining hallucinations is consistent with other work by [2] where hallucinations are defined as "false outputs or answers that are not substantiated by evidence", which is equivalent to Xu's definition linked to ground truth functions in certain cases. A

☆ This article is part of a Special issue entitled: 'Good Practices in AI' published in Artificial Intelligence in the Life Sciences.

* Corresponding author.
E-mail address: jason.granstedt@fda.hhs.gov (J. Granstedt).

https://doi.org/10.1016/j.ailsci.2025.100145
Received 6 August 2025; Received in revised form 21 October 2025; Accepted 22 October 2025
Available online 24 October 2025

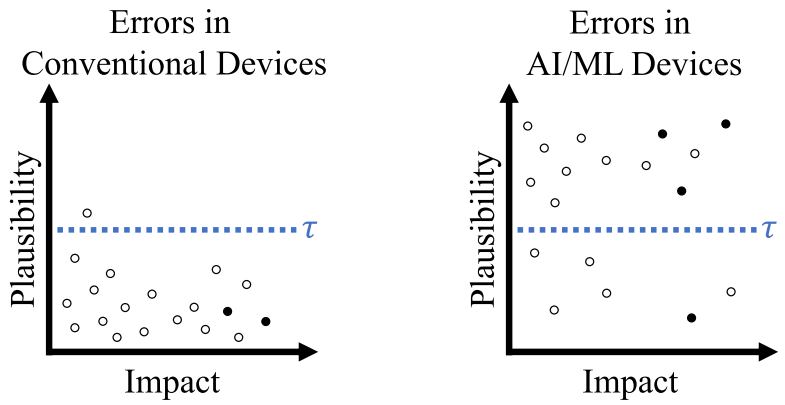

Fig. 1. Mock diagram of errors from a conventional and AI-enabled device, plotted against axes of impact and plausibility. Unmitigated impactful errors are indicated by filled circles. Plausibility introduces another risk vector, as such errors may lie outside the domain of conventional risk mitigation strategies and clinician intuition. Thus, an AI model may lead to worse patient outcomes even if it produces fewer impactful errors, as the plausibility of such errors may circumvent the traditional guardrails of medical devices. Errors above a certain plausibility threshold τ are labeled as hallucinations per our definition to identify such risks.

diagram of the considered axes for our definition is included in Fig. 1. The proposed definition requires three components: a method for identifying an error, a metric for assessing the impact of an error, and a measure for the plausibility of an error. Metrics for error are specific to the considered task and can take various forms, depending on the device and use case. A thorough discussion is outside the scope of this work, but substantial explorations of the topic have been performed [3].

The primary departure from our definition in other proposals is the incorporation of plausibility. Plausibility is a continuum and the threshold at which an error becomes sufficiently plausible to be labeled a hallucination is likely observer and task-specific. The concept of plausibility has been previously established in the medical domain: a 2013 Pew survey revealed that doctors disagreed with patients' self-diagnoses informed by online resources approximately a third of the time [4]. Such resources provided answers that were plausible to the patient, but were readily discernible as incorrect by a medical professional. The danger with AI methods is that errors may be so plausible that they may fool even experts. There are several recent occurrences of such errors in the legal profession [5–7]; it is likely that health professionals will also be susceptible. For instance, it has been demonstrated in practice that AI-reconstructed medical images have higher subjective quality scores but inferior detection performance for metastatic liver lesions [8]. These hallucinations herald a new risk vector that can circumvent professional intuition and bypass current risk mitigation strategies. It may be possible to conduct studies to empirically determine the plausibility threshold, represented by τ in Fig. 1, for relevant use cases. It would likely require a multi-reader, multi-case study with a group of comparably proficient readers assessing different hallucinations on the same task. Though, the specifics of these studies are outside the scope of this work, the role of plausibility and the different proficiencies of observers underlay the identification of hallucinations.

An additional departure from previous definitions is our introduction of the concepts of benign and impactful hallucinations. While other work focuses on the taxonomy of a hallucination, what is more important in the medical field is the impact on patient care. Human judgment has been demonstrated to be susceptible to AI errors [9–11] and clinicians can inherit AI model biases [12]. A complicating factor is that what may originally be perceived as a benign hallucination may become impactful if the information is later used to affect patient care decisions. If the output may be referred to later, then there is always a risk of downstream error propagation.

Due to the high-risk nature of medical applications, tolerance for hallucinations is frequently low. Minor modifications to factual details can impact patient management, so systems that produce hallucinations may degrade clinician trust and decrease utility even when the hallucination is benign [13]. Every poor-quality system deployed further degrades trust in AI as a whole and leads to an increasing skepticism towards future applications.

While there are many types of AI-enabled medical devices, in this work we will focus on three areas: imaging devices, generative-based synthetic medical images, and large language models. These three areas have seen explosive growth and extensive implementation of the types of AI models that lead to the proliferation of hallucinations. We will begin by describing the types of hallucinations in these device types in Section 2 and then discuss how these hallucinations may be quantified and mitigated in Sections 3 and 4, respectively. Finally, we will conclude with suggestions for the usage of our definition in Section 5 and a summary of the impact of hallucinations on the future of AI-enabled medical device development in Section 6.

2. Hallucinations in medical devices

2.1. Imaging devices

Inverse problems in imaging have been an active area of investigation for AI methods [14–20]. A critical task in many applications in medical, scientific, and industrial applications is the recovery of an image from a set of measurements, which are frequently noisy and incomplete. Improving the quantity or quality of these measurements often has an associated cost. Thus, it can be desirable to explore computational techniques to improve the utility of the image. One such method is regularization, which encodes desired attributes for an image into a mathematical formula that is applied during the reconstruction process to recover a more applicable image.

Recent research works have been predominantly focused on strategies that learn a prior distribution from a dataset via neural networks. Instead of a handcrafted term employed by conventional regularization strategies [17,21–26], these methods are data-driven. Though these methods have achieved state-of-the-art results in several areas [16,27,28], there are rising concerns about generalization performance and robustness [29,30]. This is of particular concern in medical imaging, where even a large dataset may lack rare abnormalities.

These stability issues sometimes result in the generation of false structures in reconstructed images [31–33], which have been referred to as hallucinations. Studies have warned of the potential for misdiagnoses from these hallucinated structures [34,35]. These concerns have recently been validated with clinical samples [36].

The robustness of neural networks has been investigated in many fields [37–41]. Some of these approaches consider worst-case small permutations to the input of the network [30,42–44], while others consider alternative adversarial methods [45–47]. A recently developed tool for analyzing neural network reconstructions for these phenomena is hallucination maps, which allow the isolation of artifacts associated with imperfect priors [48]. Various approaches have been proposed to incorporate information about an imaging system into neural network reconstruction methods and demonstrated resilience to these adversarial approaches [24,29,49]. Adding noise to the dataset has also demonstrated to be effective at increasing stability, albeit at the expense of performance [29].

Regularization techniques can improve human observer performance, but they cannot add any additional information to a reconstruction [50]. This is a fundamental limitation of imaging systems — information is always lost during the imaging process, and no post-processing can recover diagnostic details if a device does not measure the relevant details [48,50,51]. Effectively, the relative increase in image quality comes with the trade-off of instability and the resulting hallucinations [29]. This is an inherent flaw of data-driven approaches.

Exploring hallucinations in imaging devices presents a unique opportunity due to the accessibility of ground truth that can lead to the certain identification of errors. Nevertheless, what makes an error "plausible" remains frustratingly elusive. Plausibility can vary depending on the downstream task and whether the image is employed by a human or an algorithm. There can also be significant differences in plausibility between humans, based on level of training or simply sheer variability. Nonetheless, plausibility is one of the most complex aspects of hallucinations in a clinical environment. Neural network reconstructions can lead to an overestimation of the diagnostic utility of an image disconnected from the quality of the underlying measurements, which can subvert the intuition of a reader [34,35].

The predictable behavior of conventional regularizers enables clinicians to recognize and adapt to the various errors that arise from their use. For instance, a radiologist can turn off an image enhancement-based smoothing option in a radiological image acquisition system if the radiologist deems that a lesion in the acquired radiological image has been over-smoothed.

Task-based evaluation through reader studies, both human and computational, is one method for evaluating performance on downstream tasks [51–53]. However, images are sometimes employed for multiple tasks and improvements in one area may come at the expense of another. Some new datasets have begun to bridge this gap by providing diagnostic information [54], but access to larger datasets and deployment of multi-task evaluations will likely be necessary to assess the utility of neural network reconstructions.

The driving force in the technological advancement of medical imaging has been less radiation[1] and saving scan time in the last two decades. AI-based methods are being proposed to supplant conventional physics-based methods (like the Filtered BackProjection [56] and inverse Fourier transform [57]) such that one can faithfully recover internal organs corresponding to the person using measurements acquired at very low-dose [58,59] or under-sampled rate [60]. Recently, domain-transfer-based applications have also been proposed such that images acquired using a given modality (or a sub-modality) can be seamlessly transformed into a different modality (or sub-modality) [61]. For instance, cycleGAN has been proposed to translate an MRI image to its CT counterpart [62]. However, due to the "curse of data processing inequality" [63], an AI-based method might compensate for the information lost due to hardware-based less radiation, undersampled-acquisition, or lack of the imaging domain-specific properties with the data priors that are not specific to the person being scanned [64,65]. Simply put, as measurement quality deteriorates, AI models become more unstable.[2] This subsequently leads to imaging errors [66] that cannot be distinguished as conventional artifacts, either in terms of their obviousness from our past use of imaging devices or their traceability to imaging system-based shortcomings. We refer to them as hallucinations.

An essential hallmark of hallucinations in medical imaging is that — unlike conventional artifacts such as distortions, line artifacts, beam hardening, Gibbs ringing, aliasing, etc. [67,68] — it may not be possible to identify all the hallucinated features without the corresponding reference image. Only in the presence of a reference image with a thorough review of AI-based super-resolution do all the factually incorrect features resolved by the AI become evident. An example of such an instance is illustrated in Fig. 2 by the addition of plaques and change in the anatomy of the bowels in the AI-enhanced image. In contrast, the line artifacts are readily discernible to human eyes and can be traced to the limited angular tilt of the imaging system when acquiring the data. Per our definition, the CT image in Fig. 2(a) would constitute a non-hallucinatory or conventional artifact while Fig. 2(b) would constitute a hallucination. Further, both images in the would likely qualify as impactful errors as the utility of the images in both cases is compromised. The particular risk the hallucination poses is that the clinician is less likely to perceive the error due to the apparent quality — the AI-enhanced image appears to be of diagnostic quality and is thus plausible.

From the perspective of information theory, a typical 512×512 image encodes much more information than a single page with 500 words. A typical medical imaging-based denoising or reconstruction problem incorporates the raw data acquired from a patient (or is a conditional problem; more information in Section 2.2.2). Hence, a large number of hallucinations in the denoising and reconstruction domain may be more subtle and impactful than nonsensical or benign compared to what we may observe in language-based or unconditional domains. However, the nature of benign versus impactful hallucinations also depends on the imaging problem. For instance, consider a case whereby only half of a patient's internal body part is scanned and AI is used to predict the remaining half. This might yield highly perturbed/nonsensical outputs that experts may easily be able to categorize as errors. Overall, AI in medical imaging may yield a range of errors that may be subtle to obvious and may have benign to impactful harm. As such, it is critically important to use benchmarked imaging datasets (with patient-based diseased labels from patient follow-up data) [69–71] and perform various downstream evaluations [72] (such as pathology-based classification, quantification, detection, discrimination, etc.) to understand the nature of AI hallucinations for a given imaging application.

2.2. Synthetic data

Generative AI for data augmentation holds great promise for learning-based methods in the medical domain as it may address data scarcity issues while maintaining patient privacy. It has been employed for generating synthetic data in various imaging modalities [74,75] such as ultrasound imaging [76], mammography [77] and histopathology [78]. Typically, in data augmentation applications, the generation task is one of two kinds: (1) "unconditional generation" or generation with no prompts (only random noise as an input), e.g., given a dataset of chest radiographs, generate a similar synthetic dataset, and (2) conditional generation where the prompt may be a class-label, feature value, or another image, e.g., generate a mammogram from a given

[1] As Low as Reasonably Achievable (ALARA) has been the guiding principle of radiation safety when using imaging modalities like CT. ALARA advocates for dose optimization while maintaining the image quality required to perform the diagnostic task at hand accurately. As such, increasing the dose level — for large patients — would be consistent with ALARA's principle [55].

[2] A model is unstable when a small perturbation in input to the model leads to a large fluctuation in the model output.

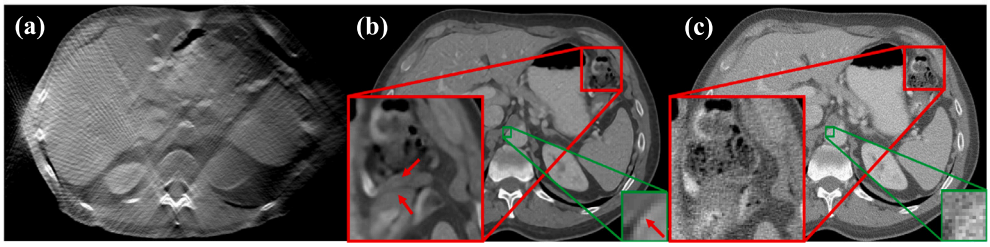

Fig. 2. An illustration of artifacts that are readily discernible in (a) and non-discernible in (b) to human eyes (adapted from [73]). The CT image in (c) is obtained by applying a physics-based analytical algorithm (i.e., filtered backprojection) on its full view projection data (i.e., 0° to 360°). The image in (b) is obtained by applying an AI-based super-resolution model on the four times downsampled version of (c). The AI-enhanced output adds two loops of bowels and plaque-like features, indicated by the red arrows. These hallucinations only become evident after comparison against its reference image in (c). The image in (a) is a reconstruction of its measurements with an imposed missing wedge acquisition (i.e., using projections from 30° to 150°).

breast type (class-label), or generate a T2-weighted magnetic resonance (MR) image given the corresponding T1-weighted MR image of the same patient. Although conditional generation may ensure consistency with the input condition (for a well-trained generative model), it does not preclude hallucinations in features that are uncorrelated with the conditioning input.

2.2.1. Unconditional generation

A distinctive aspect of unprompted/unconditional generation of images is that each generated image is entirely synthetic and does not correspond to any individual in the real world. These synthetic images can still have defined ground truth functions and hallucinations, but the "hallucination is no longer related to correctness or factualness in the real world" [1]. Specifically, ground truth functions describe the anatomical knowledge represented in the *entire* training data and can be considered as a mapping between hidden variables and images in the training set. In unconditional generation, an AI model generates new content by seeking to learn the underlying patterns of the training data without receiving explicit guidance, human labels, instructions, or a priori constraints. Inconsistencies or errors with respect to the ground truth function might still exist if the generative model function fails to learn the ground truth function. These inconsistencies may manifest as network artifacts and/or hallucinations. Recall that the difference between the two is highly subjective and based on perceptual plausibility and that lower plausibility does *not* necessarily imply lower downstream clinical impact.

In literature, hallucinations have been reported in various attributes such as per-image feature prevalence, feature-specific intensity distributions, and relative feature locations, both in domain-agnostic [79,80] and domain-specific studies [81]. Furthermore, some works report network artifacts and hallucinations under the same terminology and both are commonly known to occur in generative tasks of images [82,83] in practice. Some examples of hallucinations reported in literature according to the proposed definition are multiple optical disks instead of one in eye fundus images (as expected from the training data) and unexpected locations of medical devices in chest radiographs [82]. Examples of network artifacts include checkerboard artifacts in histopathology images [82] and nipple artifacts in mammography images [77,84].

2.2.2. Conditional generation

In prompted or conditional image generation, the generated image may be (1) entirely synthetic (e.g., when the prompt is a class-label) or (2) partially synthetic (e.g., when a patient image in one imaging modality is to be transformed to another), i.e., a domain transfer task. In the first case, the ground truth function and hallucinations are defined similarly to unconditional generation, only with the assumption that the generative model function will be consistent, i.e., not hallucinate with respect to the conditioning input and correlated attributes.

In the second case (domain transfer task), a unique ground truth function may be computable from the training data when assumptions of data sufficiency and relevance are met. Here, the ground truth function encompasses logical consistencies and relative anatomical mappings between domains, which can intuitively be understood as a bijective mapping between the domains. If the generative model function fails to learn this ground truth function, the resulting inconsistencies or errors between the two will lead to hallucinations.

However, the ground truth function may not be computable if the training dataset does not contain relevant and sufficient information for the generation task. In that case, hallucinations will occur (assumptions for definition 4 in [1]). One scenario when the ground truth function is not uniquely computable is when the physics of the input and output domains differs vastly for a given anatomy. For example, in a generation task where computed tomography (CT) is to be generated from positron emission tomography (PET) image inputs, hallucinations *must be expected* in the generated images as a unique ground truth mapping cannot be computed from the training data for the two imaging domains. Thus, the use of generative models in such problems is not advisable. Similarly, when the dataset is not relevant for a generation task, e.g., domain transfer of a diseased patient when the disease case was absent in the training data (detailed demonstration in [85]), a ground truth function does not exist and hallucinations must be expected.

Examples of hallucination in conditional generation tasks in literature include the addition of tumors in T1 MRI that did not exist in FLAIR MRI [85] and the unexpected addition of realistic histopathological features in virtual staining [86].

In both unconditional and conditional generation, errors in the generative model function may arise from various factors such as: (i) ineffective latent encoding [87], (ii) distribution-matching loss functions [85], and (iii) insufficient receptive field in the network architecture [86]. As generative models continue to evolve, so do the manifestations of their hallucinations and network artifacts.

However, it is possible to make a case for correctly using this methodology. One such example is using synthetic CT (sCT) from MRI-only scans employing a conditional cycleGAN for radiotherapy re-planning. Such usage of conditional generation may be permissible for the radiotherapy treatment of cancerous conditions that are a priori known to yield minimum dose variation compared to the original CT [88] from previously established methods like Atlas- or segmentation-based sCT [89]. The radiotherapy team can also perform various checks and balances by comparing the current sCT with the previous CT scans from initial rounds of treatment planning to mitigate hallucinations. Prior knowledge and pre-established clinical utility are important considerations when employing conditional generation, as opposed to directly employing the technology for arbitrary domain transfers without any objective assessment to validate clinical efficacy.

2.3. Language and multimodal devices

Large language models have been applied to many natural language tasks, from text summarization [90] to question answering systems [91]

and machine translation [92]. An interesting property of these tasks is the relative sensitivity to errors based on the application — while errors in summarization quickly become impactful due to the concern of fidelity [93], question answering systems may permit more errors to achieve the secondary objective of user engagement [94]. The medical versions of these tasks are likewise varied, which leads to a corresponding spread in health risks; a model for generating radiology impressions has a notably different risk profile than one performing physician note summarization.

Many of the errors that LLMs make can appear to be plausible, in part due to the ability of LLMs to correctly mimic grammar. For instance, a patient record summarization summary with an erroneously inserted diagnosis is likely to be plausible to all but a doctor who is intimately familiar with that patient's medical history. It is important to remember, however, that LLMs are trained to produce the most likely outputs, not the most accurate ones [95]. Indeed, it can be argued that the outputs of such models are persuasive because they have been specifically trained to produce plausible answers to convince humans [96]. While the likelihood of answers can be correlated to truth [95,97,98], the two are not the same. This issue becomes especially prevalent in long-tailed domains where low probability events are critical for understanding the complex systems involved [99], such as law and medicine. One could consider truthful, accurate, and modern data itself as inherently long-tailed [100].

A complication with LLMs is that for any given task, there is usually more than one output that satisfies the query [101,102]. Human evaluation [103–108] continues to remain as the preferred means of developing a reference standard for LLMs because humans can assess the various application of LLMs, including objective and subjective evaluation metrics. Unfortunately, generating a reference standard using humans is resource intensive, costing both time and money. Many public datasets have been created to evaluate LLMs for pre-specified tasks [109–111], and newer evaluation approaches of an LLM-as-a-Judge framework leverage LLMs to replace human assessments [112]. However, such methods are limited in reasoning capability and it is unclear if they will generalize to long-tailed domains such as medicine.

There are various approaches for increasing the utility of LLMs for particular applications [113]. In reinforcement learning techniques, humans analyze model responses and indicate preferential answers [96, 114]. Models may also be fine-tuned trained on specialized datasets to impart knowledge on a particular area [105] and to implement safety guardrails [114]. However, one challenge with fine-tuning approaches for models is that model performance may improve in one area but degrade in another potentially relevant areas; this is referred to as an alignment tax [115]. In deployed models that continually learn, this performance trade-off may occur unintentionally during retraining leading to unintended performance drift. Additionally, fine tuning on additional training samples may remove fine-tuned weights [116], even when such a result is unintentional [117].

Assembling a robust dataset for medical tasks for such fine-tuning is also frequently not trivial. Ensuring a clean dataset is also essential for model performance [118,119]. For foundation models, this can come both in upsampling high-quality data during pretraining [114] or fine-tuning to a specific task [120–122]. The ground truth can be unknown due to patient drop-out, lack of follow-up data on disease-based outcomes at the patient level, or disagreements among doctors. Additionally, even the data employed to train the models may be suspect. Models may rely on documentation that is either misleading, contains wrongly imputed/extrapolated/interpolated labels when compensating for missing data, or is out of date.

In addition to LLMs, visual language models (VLMs) are now being explored for applications in the medical domain [123]. While these models can generate much more detailed responses to visual inputs than previous visual question answering systems, this comes with a corresponding increase in hallucination risks. At the basic level, such models have demonstrated vulnerability to object hallucination, i.e., a description from VLM that is inconsistent with the target image [124]. Furthermore, commonly-used representations for the language-visual alignment training have demonstrated shortcomings in representations for object counts, viewpoints, and orientations [125]. Finally, errors in the alignment between the vision and language can result in an observable gap between the visual backbone of the model and the visual recognition capabilities of the LLM [126].

3. Quantifying hallucinations

The determination of an artifact as a hallucination is difficult to formally define, as it relies on the artifact being "plausible". Thus, what may appear as an obvious error to one observer (human or mathematical) may instead fool another. The severity of hallucinations varies a great deal depending on context. Minor artifacts in irrelevant portions of an output seldom have an effect on plans of care, but the same distortion in another location can lead to a misdiagnosis.

Thus, a hallucination can be impactful for one task and benign for another, and may be plausible for one observer (such as a patient) and obvious to another (such as a doctor). As a result, there is a great deal of subjectivity inherent in identifying hallucinations relevant to clinical care, and hallucinations are highly application and user dependent. Nevertheless, some algorithmic approaches have been proposed to quantify and mitigate hallucinations.

One connection that has been made across multiple fields is that hallucinations are connected to stability. While the implementation may differ, the core concept is the same — AI devices that produce substantially different outputs from a small perturbation in the input are more likely to produce hallucinations. This connection was first made in imaging, where a trade-off was observed between global metrics (e.g., mean squared error) and stability [29]. This observation has more recently been repeated in image generation [127] and LLMs [128,129]. Many developed methods exploit this relationship to measure and adjust the trade-offs between hallucinations and performance.

When the ground truth is known, as in many inverse problem simulations, a straightforward method for measuring stability is worst-case small permutations [30]. The input is modified by a small amount and iteratively optimized to maximize the change in the output. However, while this method is sufficient for demonstrating that neural network reconstruction techniques are unstable [29], it does not provide the necessary measure of plausibility to evaluate if the alteration is a hallucination. Hence, as previously mentioned in Section 2.1, rather than relying on anecdotal accounts of the efficacy of a new AI reconstruction model, a preferred methodology is to objectively quantify the error of the model for imaging tasks such as quantification, detection, discrimination, or prediction to assess the impact of hallucinations.

Likewise, scanning Fourier Ring Correlation (sFRC) has been proposed for detecting hallucinated regions-of-interest (ROIs) when reference images are available for image reconstruction [73,130]. sFRC scans and performs Fourier Ring Correlation (FRC)-based analysis over small patches between images from AI-based methods and their reference counterparts to objectively and automatically detect hallucinations. The method calls for tuning the hallucination threshold, which differentiates between hallucinated ROI and faithfully reconstructed ROI, using prior clinical knowledge of hallucinated anatomical features by the given AI-based method, ROIs conclusively identified as hallucinations by experts, and imaging theory-based limitations for a given image restoration problem. Subsequently, for the AI-based CT super-resolution problem, the sFRC paper reported an array of hallucinations including underfitting of HU attenuation (e.g., transferring fatty attenuation to air), distortion of small organelles, addition of minute indentation-like, vessel-like, plaque-like structures (depicted in Fig. 2(b)), and unwarranted folding. Similarly, for AI-enabled brain MR reconstruction, sFRC is able to detect hallucinations related to contrast migration, the omission of clinically important dark targets, thickening of gray matter, and the loss of subtle sulcus features. It should be noted

that sFRC does not provide uncertainty estimates or direct diagnostic inferences related to true positives, false positives, true negatives, and false negatives on the ROIs sFRC detects as hallucinations. The difficulty in developing a Receiver Operating Characteristic (ROC) curve-like diagnostic measure for hallucination analysis can be partially attributed to the scope of hallucinations not defined within the binary signal detection paradigm. Studies to grade hallucinations (beyond a binary level of presence versus absence) allowing for ROC-type analysis in medical image reconstruction requires significant effort.

For generative tasks in both imaging and language, the ground truth is less defined. To measure hallucinations in these domains, special datasets have been constructed. For generative models, these take the form of purposefully created stochastic models that encode attributes of interest in medical imaging [79,80]. However, while such methods are valuable for demonstrating relative performance of generative models on the specific task, it remains unknown how the results generalize to the broader medical imaging domain. When the ground truth is accessible, such as in a generative reconstruction task, another method of evaluating hallucinations is the hallucination index [127]. This method computes the Hellinger distance between the distributions of the ground truth and the reconstructed images. Still, even after such computation, determining the cut-off that dichotomizes faithful and hallucinated reconstruction may not be trivial.

For evaluating LLM performance, datasets have been constructed to specifically probe for hallucinations in the medical domain [131–133]. However, LLMs have been demonstrated to be highly unstable with performance metrics varying dramatically depending on the instruction set [134] and permutations in the input prompt [135–137]. This results in the LLMs not performing as well in practice as the obtained metrics would suggest. Thus, understanding the impact of this stability is crucial to understand how an LLM may perform in a medical environment.

4. Minimizing hallucinations

As with hallucination detection, similar concepts have been applied across fields to reduce hallucinations. Such methods involve incorporating either a measure of truth or a concept of uncertainty into the training or evaluation method.

Truth in imaging is governed by the acquired measurements and the properties of the imaging system. One such approach in imaging is the null-space shuttle method [24,26]. This modification prevents the neural network from modifying the information in the measurement data and only permits reconstruction on unknown parameters of the image. Thus, it can prevent hallucinations from the captured data. However, the risk of hallucinations remains for components of the image that are measured. Furthermore, some modifications to the measurement data, such as mitigating measurement noise, are beneficial. The null space shuttle procedure prevents the network from learning these tasks, necessitating the inclusion of additional elements in the pipeline if such features are desired.

To permit the learning of such features, other network architectures for reconstruction consider softer constraints [138,139]. Thus, data fidelity is incorporated but it is not as binding as null-space shuttle procedures. While this method may reduce hallucinations by mitigating some divergences from the ground truth, it remains unknown if the trade-off is worthwhile in many medical imaging cases. An alternative approach instead considers incorporating uncertainty into the measurements by injecting noise during the training process [29]. This method is able to improve the stability of neural network methods, but comes with a corresponding reduction in performance.

For generative imaging applications, a corresponding implementation is the AmbientGAN [140]. This modification to the typical GAN framework includes a measurement function that encodes information about an imaging system.

Other methods are useful for LLMs to incorporate relevant knowledge for queries. Retrieval augmented generation (RAG) is one of these techniques, which searches for relevant documents to append to the query [141–143]. However, RAG remains vulnerable to many hallucination vectors. In many modern implementations, an LLM is responsible for generating the RAG call from the original query. This process remains susceptible to some of the baseline hallucinations observed in LLMs. Additionally, retrieving substandard or out of date evidence may degrade model performance further [144]. Finally, such methods are also susceptible to the alignment tax. Knowledge graphs are also being explored for hallucination detection by providing explicit facts and reasoning [145], but likewise suffer from many of the above issues.

Another way of mitigating hallucinations in LLMs is post-processing of the response to remove potentially erroneous information. Some proposed method employ conformal probability to redact portions of the LLM's response [128,129]. However, to obtain a high confidence of factual information, such methods frequently prune substantial amounts from the response which potentially limits the utility of the method.

Finally, many prompt-engineering based strategies have been employed in an attempt to reduce hallucinations. Some approaches prompt a chain of reasoning for the models [146]. Others use ensemble methods, either with a collection of models [147] or a single model with multiple personas [148], to produce and analyze multiple responses. In many ways, such approaches are reminiscent of averaging across several random samples to increase confidence in the output. However, it has been convincingly argued that such approaches are unable to prevent hallucinations [1].

While many of the methods discussed in this section may improve performance on physical metrics or even tasks, none of them prevent hallucinations from occurring. Thus, there always remains a risk when employing AI for medical tasks.

5. Usage

Measures of plausibility and impact are necessary to implement the proposed definition. It is worth restating that plausibility and impact can be highly domain- and task-specific, with considerable variability even among populations of experts. As such, multi-reader multi-case studies remain the gold standard for evaluation and much of this information may already be present in current medical device study designs.

Expert elicitations of qualitative image quality have been demonstrated to not necessarily correlate with performance for data-driven algorithms [8]. However, such measures may find use in identifying plausibility. Further research may identify additional methods for determining plausibility metrics, such as investigating inter-rater agreements or adversarial reader studies.

Impact is here defined as a measure of task-based performance. Consider a simple binary detection task of an image reconstruction evaluated by an observer. In this case, error is any deviation or artifact in the reconstructed image from the acquired measurement data. In this case, impact also becomes binary — deviations that do not impact the observer's performance are benign, while those that affect the observer's output are impactful. For deployed devices, sophisticated measures exist for the tracking and reporting of device failures. In these cases, a more granular impact evaluation that includes additional resources employed to treat the patient would be more indicative of the impact of the error. Assigning partial blame to multiple devices in a chain will also facilitate the identification of failure states among increasingly interconnected devices.

The last remaining task to apply the definition is to determine the cutoff threshold τ. This threshold is also specific to the device and task. One approach includes developing a utility function for cost–benefit analysis of errors, given the mitigation strategies employed. Other approaches include estimating the trade-offs performed by clinicians in

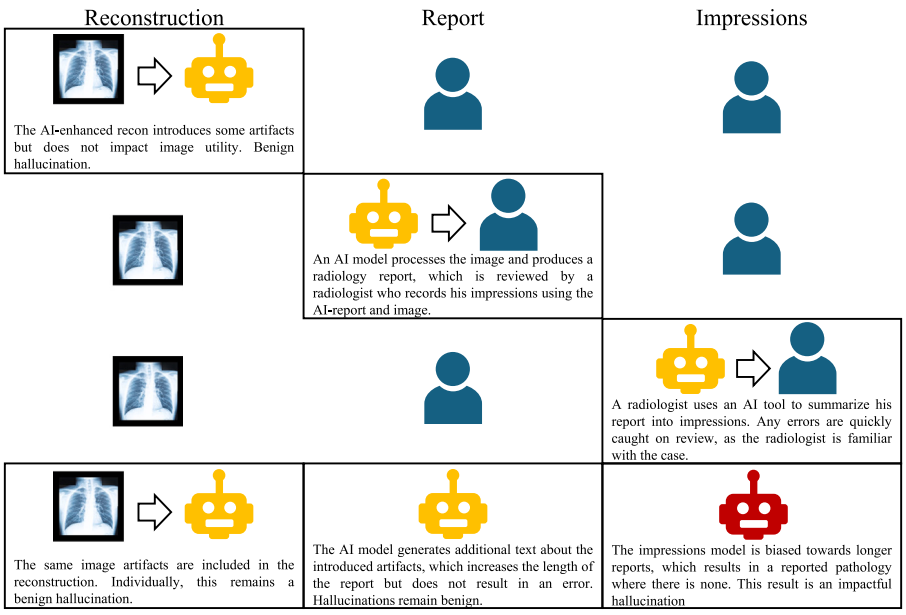

Fig. 3. Consider a radiology report workflow where an image is developed and read by a radiologist, who generates a report and returns impressions to the primary care physician. The first three rows represent integrating an AI device into one stage of the workflow. While the AI models produce errors, they are benign or mitigated. However, a concatenation of all three devices can generate unmitigated impactful errors due to instabilities within the AI models. While each individual device performs fine in isolation, the full stack presents additional risks.

everyday tasks and conducting specialized reader studies to determine what errors may "fool" experts.

Adopting the proposed definition enables a more granular analysis that may prove beneficial as the medical device ecosystem becomes increasingly interconnected, as illustrated by the example in Fig. 3. Further research into the phenomena of hallucinations is anticipated to develop more sophisticated and less burdensome evaluation measures for plausibility and impact to be incorporated earlier in device development. Note that medical experts' ability to efficiently discern conventional non-hallucinatory artifacts (like aliasing, ringing, and metal artifacts as previously explained in Section 2.1) has progressively increased over the years with increasing awareness of the limits different imaging modalities across various applications and patient populations. It is possible a similar learning curve will emerge as experts become more adept at distinguishing between benign and impactful hallucinations.

6. Summary

Medical devices are employed in many applications that impact patient care, both directly and indirectly. The incorporation of AI methods into these devices contains both benefits and risks. It is important to emphasize that AI models in devices do not need to be perfect to be useful, especially when such models demonstrate performance improvements over existing standards of care. However, hallucinations pose novel challenges to the existing medical ecosystem. By focusing the discussion of hallucinations on downstream impacts to patient care, meaningful progress can be made for the safe and effective integration of AI-enabled medical devices.

Nonetheless, hallucinations cannot be fully removed as they are intrinsic to neural network-based methods and attempts to reduce hallucinations may come at the cost of decreased performance. This inherent instability introduces unique risks due to injecting errors in the chain of care that may not manifest until much later. Furthermore, the risk belongs to the entire stack of deployed AI models. While each model may be individually low-risk, the combined system may become high-risk due to cascading instabilities.

CRediT authorship contribution statement

Jason Granstedt: Conceptualization, Writing – original draft, Writing – review & editing. **Prabhat Kc:** Conceptualization, Writing – original draft, Writing – review & editing. **Rucha Deshpande:** Conceptualization, Writing – original draft, Writing – review & editing. **Victor Garcia:** Conceptualization, Writing – review & editing. **Aldo Badano:** Conceptualization, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: An innate limitation of large language models. 2024, arXiv preprint arXiv:2401.11817.
[2] Wei J, Karina N, Chung HW, Jiao YJ, Papay S, Glaese A, Schulman J, Fedus W. Measuring short-form factuality in large language models. 2024, arXiv preprint arXiv:2411.04368.
[3] Barrett HH, Myers KJ. Foundations of image science. John Wiley & Sons; 2013.
[4] Fox S, Duggan M. Health Online 2013. Tech. rep., Washington, D.C.: Pew Research Center; 2013.
[5] Mata v Avianca, Inc. No. 54 civ. 1461. 2023, https://storage.courtlistener.com/recap/gov.uscourts.nysd.575368/gov.uscourts.nysd.575368.54.0_3.pdf.
[6] Ko v Li. ONSC 2766. 2025, https://www.canlii.org/en/on/onsc/doc/2025/2025onsc2766/2025onsc2766.html.
[7] Lacey v State Farm. No. 2:24-cv-05205-FMO-maa. 2025, https://www.lawnext.com/wp-content/uploads/2025/05/C.D.-Cal.-24-cv-05205-dckt-000119_000-filed-2025-05-06.pdf.
[8] Jensen CT, Gupta S, Saleh MM, Liu X, Wong VK, Salem U, Qiao W, Samei E, Wagner-Bartak NA. Reduced-dose deep learning reconstruction for abdominal CT of liver metastases. Radiology 2022;303(1):90–8.
[9] Agudo U, Liberal KG, Arrese M, Matute H. The impact of AI errors in a human-in-the-loop process. Cogn Res: Princ Implic 2024;9(1):1.

[10] Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision making. NPJ Digit Med 2023;6(1):206.
[11] Jacobs M, Pradier MF, McCoy Jr. TH, Perlis RH, Doshi-Velez F, Gajos KZ. How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. Transl Psychiatry 2021;11(1):108.
[12] Vicente L, Matute H. Humans inherit artificial intelligence biases. Sci Rep 2023;13(1):15737.
[13] Kim Y, Jeong H, Chen S, Li SS, Lu M, Alhamoud K, Mun J, Grau C, Jung M, Gameiro RR, et al. Medical hallucination in foundation models and their impact on healthcare. 2025, arXiv preprint arXiv:2503.05777.
[14] Arridge S, Maass P, Öktem O, Schönlieb C-B. Solving inverse problems using data-driven models. Acta Numer 2019;28:1–174.
[15] Liang D, Cheng J, Ke Z, Ying L. Deep magnetic resonance image reconstruction: Inverse problems meet neural networks. IEEE Signal Process Mag 2020;37(1):141–51.
[16] McCann MT, Jin KH, Unser M. Convolutional neural networks for inverse problems in imaging: A review. IEEE Signal Process Mag 2017;34(6):85–95.
[17] Ongie G, Jalal A, Metzler CA, Baraniuk RG, Dimakis AG, Willett R. Deep learning techniques for inverse problems in imaging. IEEE J Sel Areas Inf Theory 2020;1(1):39–56.
[18] Wang G, Ye JC, De Man B. Deep learning for tomographic image reconstruction. Nat Mach Intell 2020;2(12):737–48.
[19] Reader AJ, Corda G, Mehranian A, da Costa-Luis C, Ellis S, Schnabel JA. Deep learning for PET image reconstruction. IEEE Trans Radiat Plasma Med Sci 2020;5(1):1–25.
[20] Wang G, Ye JC, Mueller K, Fessler JA. Image reconstruction is a new frontier of machine learning. IEEE Trans Med Imaging 2018;37(6):1289–96.
[21] Smith B. Null-space smoothing of tomographic images using TV norm minimization. In: 2016 IEEE nuclear science symposium, medical imaging conference and room-temperature semiconductor detector workshop. IEEE; 2016, p. 1–4.
[22] Hahn BN. Null space and resolution in dynamic computerized tomography. Inverse Problems 2016;32(2):025006.
[23] Kelly B, Matthews TP, Anastasio MA. Deep learning-guided image reconstruction from incomplete data. 2017, arXiv preprint arXiv:1709.00584.
[24] Schwab J, Antholzer S, Haltmeier M. Deep null space learning for inverse problems: Convergence analysis and rates. Inverse Problems 2019;35(2):025008.
[25] Rowbotham PS, Pratt RG. Improved inversion through use of the null space. Geophysics 1997;62(3):869–83.
[26] Deal MM, Nolet G. Nullspace shuttles. Geophys J Int 1996;124(2):372–80.
[27] Wang G. A perspective on deep imaging. IEEE Access 2016;4:8914–24.
[28] Ravishankar S, Ye JC, Fessler JA. Image reconstruction: From sparsity to data-adaptive methods and machine learning. Proc IEEE 2020;108(1):86–109.
[29] Gottschling N, Antun V, Adcock B, Hansen AC. The troublesome kernel: Why deep learning for inverse problems is typically unstable. 2020, ArXiv, arXiv:2001.01258.
[30] Antun V, Renna F, Poon C, Adcock B, Hansen AC. On instabilities of deep learning in image reconstruction and the potential costs of AI. Proc Natl Acad Sci 2020;117(48):30088–95.
[31] Belthangady C, Royer LA. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. Nature Methods 2019;16(12):1215–25.
[32] Hoffman DP, Slavitt I, Fitzpatrick CA. The promise and peril of deep learning in microscopy. Nature Methods 2021;18(2):131–2.
[33] Varoquaux G, Cheplygina V. Machine learning for medical imaging: Methodological failures and recommendations for the future. NPJ Digit Med 2022;5(1):48.
[34] Knoll F, Murrell T, Sriram A, Yakubova N, Zbontar J, Rabbat M, Defazio A, Muckley MJ, Sodickson DK, Zitnick CL, et al. Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge. Magn Reson Med 2020;84(6):3054–70.
[35] Muckley MJ, Riemenschneider B, Radmanesh A, Kim S, Jeong G, Ko J, Jun Y, Shin H, Hwang D, Mostapha M, et al. Results of the 2020 fastMRI challenge for machine learning MR image reconstruction. IEEE Trans Med Imaging 2021;40(9):2306–17.
[36] Bosbach WA, Merdes KC, Jung B, Montazeri E, Anderson S, Mitrakovic M, Daneshvar K. Deep learning reconstruction of accelerated MRI: False-positive cartilage delamination inserted in MRI arthrography under traction. Top Magn Reson Imaging 2024;33(4):e0313.
[37] Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C, Prakash A, Kohno T, Song D. Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 1625–34.
[38] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text. In: 2018 IEEE security and privacy workshops. IEEE; 2018, p. 1–7.
[39] Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019;363(6433):1287–9.
[40] Heaven D, et al. Why deep-learning AIs are so easy to fool. Nature 2019;574(7777):163–6.
[41] Bastounis A, Hansen AC, Vlačić V. The mathematics of adversarial attacks in AI–why deep learning is unstable despite the existence of stable neural networks. 2021, arXiv preprint arXiv:2109.06098.
[42] Huang Y, Würfl T, Breininger K, Liu L, Lauritsch G, Maier A. Some investigations on robustness of deep learning in limited angle tomography. In: Medical image computing and computer assisted intervention–mICCAI 2018: 21st international conference, granada, Spain, September 16-20, 2018, proceedings, part i. Springer; 2018, p. 145–53.
[43] Darestani MZ, Chaudhari AS, Heckel R. Measuring robustness in deep learning based compressive sensing. In: International conference on machine learning. PMLR; 2021, p. 2433–44.
[44] Genzel M, Macdonald J, März M. Solving inverse problems with deep neural networks–robustness included? IEEE Trans Pattern Anal Mach Intell 2022;45(1):1119–34.
[45] Raj A, Bresler Y, Li B. Improving robustness of deep-learning-based image reconstruction. In: International conference on machine learning. PMLR; 2020, p. 7932–42.
[46] Morshuis JN, Gatidis S, Hein M, Baumgartner CF. Adversarial robustness of MR image reconstruction under realistic perturbations. In: International workshop on machine learning for medical image reconstruction. Springer; 2022, p. 24–33.
[47] Alaifari R, Alberti GS, Gauksson T. Localized adversarial artifacts for compressed sensing MRI. SIAM J Imaging Sci 2023;16(4):SC14–26.
[48] Bhadra S, Kelkar VA, Brooks FJ, Anastasio MA. On hallucinations in tomographic image reconstruction. IEEE Trans Med Imaging 2021;40(11):3249–60.
[49] Colbrook MJ, Antun V, Hansen AC. The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem. Proc Natl Acad Sci 2022;119(12):e2107151119.
[50] Zhang X, Kelkar VA, Granstedt J, Li H, Anastasio MA. Impact of deep learning-based image super-resolution on binary signal detection. J Med Imaging 2021;8(6). 065501–065501.
[51] Wagner RF, Brown DG. Unified SNR analysis of medical imaging systems. Phys Med Biol 1985;30(6):489.
[52] Vennart W. ICRU report 54: Medical imaging-the assessment of image quality-isbn 0-913394-53-X. April 1996, Maryland, USA. Radiography 1997;3(3):243–4.
[53] Barrett HH, Yao J, Rolland JP, Myers KJ. Model observers for assessment of image quality. Proc Natl Acad Sci 1993;90(21):9758–65.
[54] Zhao R, Yaman B, Zhang Y, Stewart R, Dixon A, Knoll F, Huang Z, Lui YW, Hansen MS, Lungren MP. fastMRI+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. Sci Data 2022;9(1):152.
[55] McCollough CH, Primak AN, Braun N, Kofler J, Yu L, Christner J. Strategies for reducing radiation dose in CT. Radiol Clin 2009;47(1):27–40.
[56] Kak AC, Slaney M. Algorithms for reconstruction with nondiffracting sources. In: Principles of computerized tomographic imaging. Philadelphia: SIAM; 2001, p. 49–112.
[57] Santiago A-F, Gonzalo V-S-F. Acquisition and reconstruction of magnetic resonance imaging. In: Statistical analysis of noise in MRI modeling, filtering and estimation. Switzerland: Springer International Publishing; 2016, p. 9–29.
[58] Chen H, Zhang Y, Kalra MK, Lin F, Chen Y, Liao P, Zhou J, Wang G. Low-dose CT with a residual encoder-decoder convolutional neural network. IEEE Trans Med Imaging 2017;36(12):2524–35.
[59] Yang Q, Yan P, Zhang Y, Yu H, Shi Y, Mou X, Kalra MK, Zhang Y, Sun L, Wang G. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. IEEE Trans Med Imaging 2018;37(6):1348–57.
[60] Sun J, Li H, Xu Z, et al. Deep ADMM-net for compressive sensing MRI. Adv Neural Inf Process Syst 2016;29.
[61] Vemulapalli R, Nguyen HV, Zhou SK. Deep networks and mutual information maximization for cross-modal medical image synthesis. In: Zhou SK, Greenspan H, Shen D, editors. Deep learning for medical image analysis. Oxford: Academic Press; 2023, p. 381–403.
[62] Hiasa Y, Otake Y, Takao M, Matsuoka T, Takashima K, Carass A, Prince JL, Sugano N, Sato Y. Cross-modality image synthesis from unpaired data using cycleGAN: Effects of gradient consistency loss and training data size. In: Simulation and synthesis in medical imaging: Third international workshop, SASHIMI 2018, held in conjunction with MICCAI 2018, granada, Spain, September 16, 2018, proceedings 3. Springer; 2018, p. 31–41.
[63] McDonnell MD, Stocks NG, Pearce CE, Abbott D. The data processing inequality and stochastic resonance. In: Noise in complex systems and stochastic dynamics. vol. 5114, SPIE; 2003, p. 249–60.
[64] Bhadra S, Kelkar VA, Brooks FJ, Anastasio MA. On hallucinations in tomographic image reconstruction. IEEE Trans Med Imaging 2021;40(11):3249–60.
[65] Tam LK, Stockmann JP, Galiana G, Constable RT. Null space imaging: Nonlinear magnetic encoding fields designed complementary to receiver coil sensitivities for improved acceleration in parallel imaging. Magn Reson Med 2012;68(4):1166–75.

[66] Cohen JP, Luck M, Honari S. Distribution matching losses can hallucinate features in medical image translation. In: Medical image computing and computer assisted intervention–mICCAI 2018: 21st international conference, granada, Spain, September 16-20, 2018, proceedings, part i. Springer; 2018, p. 529–36.
[67] Hsieh J. Image artifacts: Appearances, causes, and corrections. In: Computed tomography: Principles, design, artifacts, and recent advances. Bellingham, Washington: SPIE Press; 2003, p. 207–300.
[68] Krupa K, Bekiesińska-Figatowska M. Artifacts in magnetic resonance imaging. Pol J Radiol 2015;80:93.
[69] Zhao R, Yaman B, Zhang Y, Stewart R, Dixon A, Knoll F, Huang Z, Lui YW, Hansen MS, Lungren MP. Fastmri+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. Sci Data 2022;9(1):152.
[70] Armato III SG, Drukker K, Li F, Hadjiiski L, Tourassi GD, Engelmann RM, Giger ML, Redmond G, Farahani K, Kirby JS, et al. LUNGx challenge for computerized lung nodule classification. J Med Imaging 2016;3(4). 044506–044506.
[71] Yan K, Wang X, Lu L, Summers RM. DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. J Med Imaging 2018;5(3). 036501–036501.
[72] Udandarao V, Prabhu A, Ghosh A, Sharma Y, Torr P, Bibi A, Albanie S, Bethge M. No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In: The thirty-eighth annual conference on neural information processing systems. 2024.
[73] Kc P, Zeng R, Soni N, Badano A. sFRC for assessing hallucinations in medical image restoration. 2025, TechRxiv Preprints.
[74] Chen Y, Yang X-H, Wei Z, Heidari AA, Zheng N, Li Z, Chen H, Hu H, Zhou Q, Guan Q. Generative adversarial networks in medical image augmentation: A review. Comput Biol Med 2022;144:105382.
[75] Garcea F, Serra A, Lamberti F, Morra L. Data augmentation for medical imaging: A systematic literature review. Comput Biol Med 2023;152:106391.
[76] Xu J, Hua Q, Jia X, Zheng Y, Hu Q, Bai B, Miao J, Zhu L, Zhang M, Tao R, et al. Synthetic breast ultrasound images: A study to overcome medical data sharing barriers. Research 2024;7:0532.
[77] Lee J, Nishikawa RM. Analyzing GAN artifacts for simulating mammograms: Application towards finding mammographically-occult cancer. In: Medical imaging 2022: Computer-aided diagnosis. vol. 12033, SPIE; 2022, p. 78–84.
[78] Xue Y, Ye J, Zhou Q, Long LR, Antani S, Xue Z, Cornwell C, Zaino R, Cheng KC, Huang X. Selective synthetic augmentation with histogan for improved histopathology image classification. Med Image Anal 2021;67:101816.
[79] Deshpande R, Anastasio MA, Brooks FJ. A method for evaluating deep generative models of images for hallucinations in high-order spatial context. Pattern Recognit Lett 2024;186:23–9.
[80] Deshpande R, Özbey M, Li H, Anastasio MA, Brooks FJ. Assessing the capacity of a denoising diffusion probabilistic model to reproduce spatial context. IEEE Trans Med Imaging 2024.
[81] Kelkar VA, Gotsis DS, Brooks FJ, Prabhat K, Myers KJ, Zeng R, Anastasio MA. Assessing the ability of generative adversarial networks to learn canonical medical image statistics. IEEE Trans Med Imaging 2023;42(6):1799–808.
[82] Müller-Franzes G, Niehues JM, Khader F, Arasteh ST, Haarburger C, Kuhl C, Wang T, Han T, Nolte T, Nebelung S, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. Sci Rep 2023;13(1):12098.
[83] Deshpande R, Kelkar VA, Gotsis D, Kc P, Zeng R, Myers KJ, Brooks FJ, Anastasio MA. Report on the AAPM grand challenge on deep generative modeling for learning medical image statistics. Med Phys 2025;52(1):4–20.
[84] Deshpande R, Lago M, Subbaswamy A, Kahaki S, Delfino JG, Badano A, Zamzmi G. A knowledge-based method for detecting network-induced shape artifacts in synthetic images. In: Medical imaging with deep learning. 2025.
[85] Cohen JP, Luck M, Honari S. Distribution matching losses can hallucinate features in medical image translation. In: Medical image computing and computer assisted intervention–mICCAI 2018: 21st international conference, granada, Spain, September 16-20, 2018, proceedings, part i. Springer; 2018, p. 529–36.
[86] Vasiljević J, Nisar Z, Feuerhake F, Wemmert C, Lampert T. Cyclegan for virtual stain transfer: Is seeing really believing? Artif Intell Med 2022;133:102420.
[87] Bond-Taylor S, Leach A, Long Y, Willcocks CG. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. IEEE Trans Pattern Anal Mach Intell 2021;44(11):7327–47.
[88] Ranta I, Wright P, Suilamo S, Kemppainen R, Schubert G, Kapanen M, Keyriläinen J. Clinical feasibility of a commercially available MRI-only method for radiotherapy treatment planning of the brain. J Appl Clin Med Phys 2023;24(9):e14044.
[89] Bahloul MA, Jabeen S, Benoumhani S, Alsaleh HA, Belkhatir Z, Al-Wabil A. Advancements in synthetic CT generation from MRI: A review of techniques, and trends in radiation therapy planning. J Appl Clin Med Phys 2024;25(11):e14499.
[90] Zhang T, Ladhak F, Durmus E, Liang P, McKeown K, Hashimoto TB. Benchmarking large language models for news summarization. Trans Assoc Comput Linguist 2024;12:39–57.
[91] Tan Y, Min D, Li Y, Li W, Hu N, Chen Y, Qi G. Evaluation of ChatGPT as a question answering system for answering complex questions. 2023, arXiv preprint arXiv:2303.07992.
[92] Zhu W, Liu H, Dong Q, Xu J, Huang S, Kong L, Chen J, Li L. Multilingual machine translation with large language models: Empirical results and analysis. 2023, arXiv preprint arXiv:2304.04675.
[93] Pagnoni A, Balachandran V, Tsvetkov Y. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. 2021, arXiv preprint arXiv:2104.13346.
[94] Huang M, Zhu X, Gao J. Challenges in building intelligent open-domain dialog systems. ACM Trans Inf Syst (TOIS) 2020;38(3):1–32.
[95] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI Blog 2019;1(8):9.
[96] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 2022;35:27730–44.
[97] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers). 2019, p. 4171–86.
[98] Roberts A, Raffel C, Shazeer N. How much knowledge can you pack into the parameters of a language model?. 2020, arXiv preprint arXiv:2002.08910.
[99] Taleb NN. Black swans and the domains of statistics. Amer Statist 2007;61(3):198–200.
[100] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. ACM Comput Surv 2023;55(12):1–38.
[101] Su H, Shen X, Zhao S, Zhou X, Hu P, Zhong R, Niu C, Zhou J. Diversifying dialogue generation with non-conversational text. 2020, arXiv preprint arXiv:2005.04346.
[102] Guan J, Huang M. Union: An unreferenced metric for evaluating open-ended story generation. 2020, arXiv preprint arXiv:2009.07602.
[103] Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval augmentation reduces hallucination in conversation. 2021, arXiv preprint arXiv:2104.07567.
[104] Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, Wang A, Li B, Amin M, Tomasev N, et al. Towards conversational diagnostic AI. 2024, arXiv preprint arXiv:2401.05654.
[105] Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang P-C, Carroll A, Lau C, Tanno R, Ktena I, et al. Towards generalist biomedical AI. Nejm Ai 2024;1(3). AIoa2300138.
[106] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, Hou L, Clark K, Pfohl SR, Cole-Lewis H, et al. Toward expert-level medical question answering with large language models. Nature Med 2025;1–8.
[107] Li C-Y, Chang K-J, Yang C-F, Wu H-Y, Chen W, Bansal H, Chen L, Yang Y-P, Chen Y-C, Chen S-P, et al. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. Nat Commun 2025;16(1):2258.
[108] Tanno R, Barrett DG, Sellergren A, Ghaisas S, Dathathri S, See A, Welbl J, Lau C, Tu T, Azizi S, et al. Collaboration between clinicians and vision–language models in radiology report generation. Nature Med 2025;31(2):599–608.
[109] Wang A. Glue: A multi-task benchmark and analysis platform for natural language understanding. 2018, arXiv preprint arXiv:1804.07461.
[110] Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, Yasunaga M, Zhang Y, Narayanan D, Wu Y, Kumar A, et al. Holistic evaluation of language models. 2022, arXiv preprint arXiv:2211.09110.
[111] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J. Measuring massive multitask language understanding. 2020, arXiv preprint arXiv:2009.03300.
[112] Zheng L, Chiang W-L, Sheng Y, Zhuang S, Wu Z, Zhuang Y, Lin Z, Li Z, Li D, Xing E, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. Adv Neural Inf Process Syst 2023;36:46595–623.
[113] Patil R, Gudivada V. A review of current trends, techniques, and challenges in large language models (LLMs). Appl Sci 2024;14(5):2074.
[114] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, et al. Llama 2: Open foundation and fine-tuned chat models. 2023, arXiv preprint arXiv:2307.09288.
[115] Lin Y, Lin H, Xiong W, Diao S, Liu J, Zhang J, Pan R, Wang H, Hu W, Zhang H, et al. Mitigating the alignment tax of RLFH. 2023, arXiv preprint arXiv:2309.06256.
[116] Shen X, Chen Z, Backes M, Shen Y, Zhang Y. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In: Proceedings of the 2024 on ACM SIGSAC conference on computer and communications security. 2024, p. 1671–85.
[117] Qi X, Zeng Y, Xie T, Chen P-Y, Jia R, Mittal P, Henderson P. Fine-tuning aligned language models compromises safety, even when users do not intend to!. 2023, arXiv preprint arXiv:2310.03693.
[118] Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, Huang X, Zhao E, Zhang Y, Chen Y, et al. Siren's song in the AI ocean: A survey on hallucination in large language models. 2023, arXiv preprint arXiv:2309.01219.

[119] Penedo G, Malartic Q, Hesslow D, Cojocaru R, Cappelli A, Alobeidli H, Pannier B, Almazrouei E, Launay J. The RefinedWeb dataset for falcon LLM: Outperforming curated corpora with web data, and web data only. 2023, arXiv preprint arXiv:2306.01116.

[120] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, Hou L, Clark K, Pfohl SR, Cole-Lewis H, et al. Toward expert-level medical question answering with large language models. Nature Med 2025;1–8.

[121] Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li Y, Wang X, Dehghani M, Brahma S, et al. Scaling instruction-finetuned language models. J Mach Learn Res 2024;25(70):1–53.

[122] Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-llama: Toward building open-source language models for medicine. J Am Med Inform Assoc 2024;31(9):1833–43.

[123] Hartsock I, Rasool G. Vision-language models for medical report generation and visual question answering: A review. Front Artif Intell 2024;7:1430984.

[124] Li Y, Du Y, Zhou K, Wang J, Zhao WX, Wen J-R. Evaluating object hallucination in large vision-language models. 2023, arXiv preprint arXiv:2305.10355.

[125] Tong S, Liu Z, Zhai Y, Ma Y, LeCun Y, Xie S. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 9568–78.

[126] Zhai Y, Tong S, Li X, Cai M, Qu Q, Lee YJ, Ma Y. Investigating the catastrophic forgetting in multimodal large language models. 2023, arXiv preprint arXiv:2309.10313.

[127] Tivnan M, Yoon S, Chen Z, Li X, Wu D, Li Q. Hallucination index: An image quality metric for generative reconstruction models. In: International conference on medical image computing and computer-assisted intervention. Springer; 2024, p. 449–58.

[128] Mohri C, Hashimoto T. Language models with conformal factuality guarantees. 2024, arXiv preprint arXiv:2402.10978.

[129] Cherian J, Gibbs I, Candes E. Large language model validity via enhanced conformal prediction methods. Adv Neural Inf Process Syst 2024;37:114812–42.

[130] US Food and Drug Administration. sFRC for detecting hallucinations in medical image restoration (RST24MD16.01). 2025, https://cdrh-rst.fda.gov/sfrc-detecting-hallucinations-medical-image-restoration.

[131] Pal A, Umapathi LK, Sankarasubbu M. Med-halt: Medical domain hallucination test for large language models. 2023, arXiv preprint arXiv:2307.15343.

[132] Xia P, Chen Z, Tian J, Gong Y, Hou R, Xu Y, Wu Z, Fan Z, Zhou Y, Zhu K, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. Adv Neural Inf Process Syst 2024;37:140334–65.

[133] Chen J, Yang D, Wu T, Jiang Y, Hou X, Li M, Wang S, Xiao D, Li K, Zhang L. Detecting and evaluating medical hallucinations in large vision language models. 2024, arXiv preprint arXiv:2406.10185.

[134] Mizrahi M, Kaplan G, Malkin D, Dror R, Shahaf D, Stanovsky G. State of what art? A call for multi-prompt LLM evaluation. Trans Assoc Comput Linguist 2024;12:933–49.

[135] Wang B, Xu C, Wang S, Gan Z, Cheng Y, Gao J, Awadallah AH, Li B. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. 2021, arXiv preprint arXiv:2111.02840.

[136] Wang J, Hu X, Hou W, Chen H, Zheng R, Wang Y, Yang L, Huang H, Ye W, Geng X, et al. On the robustness of chatGPT: An adversarial and out-of-distribution perspective. 2023, arXiv preprint arXiv:2302.12095.

[137] Wang B, Chen W, Pei H, Xie C, Kang M, Zhang C, Xu C, Xiong Z, Dutta R, Schaeffer R, et al. DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. In: NeurIPS. 2023.

[138] Hammernik K, Klatzer T, Kobler E, Recht MP, Sodickson DK, Pock T, Knoll F. Learning a variational network for reconstruction of accelerated MRI data. Magn Reson Med 2018;79(6):3055–71.

[139] Chen H, Zhang Y, Chen Y, Zhang J, Zhang W, Sun H, Lv Y, Liao P, Zhou J, Wang G. LEARN: Learned experts' assessment-based reconstruction network for sparse-data CT. IEEE Trans Med Imaging 2018;37(6):1333–47.

[140] Bora A, Price E, Dimakis AG. Ambientgan: Generative models from lossy measurements. In: International conference on learning representations. 2018.

[141] Lee K, Chang M-W, Toutanova K. Latent retrieval for weakly supervised open domain question answering. 2019, arXiv preprint arXiv:1906.00300.

[142] Guu K, Lee K, Tung Z, Pasupat P, Chang M. Retrieval augmented language model pre-training. In: International conference on machine learning. PMLR; 2020, p. 3929–38.

[143] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-t, Rocktäschel T, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst 2020;33:9459–74.

[144] Xu R, Qi Z, Guo Z, Wang C, Wang H, Zhang Y, Xu W. Knowledge conflicts for llms: A survey. 2024, arXiv preprint arXiv:2403.08319.

[145] Lavrinovics E, Biswas R, Bjerva J, Hose K. Knowledge graphs, large language models, and hallucinations: An NLP perspective. J Web Semant 2025;85:100844.

[146] Ji Z, Yu T, Xu Y, Lee N, Ishii E, Fung P. Towards mitigating hallucination in large language models via self-reflection. 2023, arXiv preprint arXiv:2310.06271.

[147] Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I. Improving factuality and reasoning in language models through multiagent debate. In: Forty-first international conference on machine learning. 2023.

[148] Wang Z, Mao S, Wu W, Ge T, Wei F, Ji H. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. 2023, arXiv preprint arXiv:2307.05300.