Radiology

REVIEWS AND COMMENTARY • REVIEW

Generative AI and Foundation Models in Radiology: Applications, Opportunities, and Potential Challenges

Neda Tavakoli, PhD[1] • Zahra Shakeri, PhD[2] • Vrushab Gowda, MD, JD[3] • Konrad Samsel, MPH[2] • Arash Bedayat, MD[4] • Ahmadreza Ghasemiesfe, MD[5] • Ulas Bagci, PhD[1] • Albert Hsiao, MD, PhD[6] • Tim Leiner, MD, PhD[7] • James Carr, MD[1] • Daniel Kim, PhD[1] • Amir Ali Rahsepar, MD[1]

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

Radiology 2025; 317(2):e252961 • https://doi.org/10.1148/radiol.242961 • Content code: AI

Foundation models (FMs) represent a transformative advancement in artificial intelligence (AI), with growing applications in medical imaging. These models leverage self-attention mechanisms and are capable of processing multimodal data, such as images, text, audio, and video, across multiple scales. Although FMs require large datasets for initial training, they can be adapted to specific medical imaging tasks using smaller labeled datasets through techniques such as transfer learning, fine-tuning, prompt engineering, few-shot learning, and zero-shot learning, making them especially valuable in data-scarce settings. Many FMs also incorporate generative AI capabilities that support the creation of synthetic medical images to further address annotation limitations. Current applications span various imaging modalities in radiology, where FMs have shown potential to improve diagnostic accuracy and streamline workflows. However, clinical integration remains challenging due to issues such as limited interpretability, potential bias, privacy concerns, regulatory constraints, high computational costs, and domain shifts between training data and real-world clinical environments. Addressing these barriers will require coordinated efforts among technical developers, health care providers, and regulatory bodies. This review explores the evolving role of FMs and generative AI in radiology, highlighting recent research advances, clinical applications, and the key challenges that must be addressed for responsible deployment.

An earlier incorrect version appeared online. This article was corrected on November 19, 2025.

Artificial intelligence (AI) is rapidly transforming the field of medical imaging, with foundation models (FMs) and generative AI driving a paradigm shift in how imaging data are analyzed, interpreted, and used in clinical care. As shown in Figure 1, traditional machine learning and deep learning approaches rely on task-specific models and typically require large amounts of manually labeled (ie, simple categorical or numerical tags) or annotated (ie, detailed markup, including boundary boxes and segmentation masks) data created by human annotators for each individual application. In contrast, FMs represent a more flexible and scalable approach. They are capable of integrating generative AI techniques and leveraging self-supervised learning, a method that allows models to learn from inherent structures within unlabeled data by predicting parts of the input from other parts, without requiring manual annotation. Self-supervised learning differs from unsupervised learning by creating its own training targets from the data itself, rather than simply discovering patterns without any guidance. When combined with the well-established technique of transfer learning, this enables FMs to extract generalizable features from large volumes of medical imaging data, allowing them to adapt to various clinical tasks, including segmentation, classification, and report generation, often with minimal fine-tuning. In addition, FMs can apply few-shot (1) and zero-shot (2) learning approaches, allowing them to perform tasks using only a few labeled examples, or even none. This makes them especially valuable for rare diseases and emerging imaging protocols, where annotated data are scarce or unavailable.

Generative AI, a broader class of models that predates FMs, has recently evolved through large-scale transformer-based architectures (3) and is now being integrated into FMs. This integration has enabled advanced capabilities such as synthetic image generation and corresponding annotations, data augmentation, and image enhancement (4,5), supporting model performance in data-scarce or heterogeneous imaging environments.

Although FMs reduce dependence on large-scale labeled datasets, annotated data remain important for downstream tasks such as fine-tuning, model testing, and semisupervised learning, for which small amounts of labeled data are used alongside larger unlabeled datasets to improve accuracy and reliability (6).

Despite their promise, integrating FMs into clinical workflows remains challenging. Although several technical reviews have detailed FM architectures and categorizations (7), practical guidance for workflow integration, real-world external testing, interpretability, and ethical implementation is still lacking. Moreover, issues such as fairness, domain generalization, and regulatory compliance must be addressed before widespread deployment.

This review offers an overview of FMs and generative AI in radiology, highlighting their transformative potential and key challenges. Given the early stage of FM development in radiology, this review provides essential grounding and guidance for radiologists preparing for clinical implementation and real-world external testing.

FM Description

In this review, the term *foundation models* or *FMs* refers to advanced deep learning models that are pretrained on broad datasets and can be adapted for various tasks, as defined by Bommasani et al (8). These models are characterized by their versatility to new tasks with minimal additional training, supported by their exposure to vast and diverse datasets during pretraining. FMs use self-attention mechanisms, a core feature of the transformer architecture (3), allowing them to capture global context, making them ideal for health care applications like diagnosis and treatment planning. Many of these models can process multimodal data, including text, images, video, and audio (8), although some are designed as unimodal systems focusing on single data types.

Abbreviations

AI = artificial intelligence, BERT = bidirectional encoder representations from transformers, CLIP = Contrastive Language–Image Pretraining, FDA = Food and Drug Administration, FM = foundation model, GPT = generative pretrained transformer, SAM = Segment Anything Model, ViT = vision transformer

Summary

Foundation models can streamline imaging workflows, improve diagnostic accuracy, enable personalized treatment, and, via integrated generative artificial intelligence, create synthetic images to augment scarce real-world data; however, successful adoption requires explainability, trust, and workflow integration.

Essentials

- Foundation models (FMs) can be adapted for specific medical imaging tasks through multiple approaches, including transfer learning, fine-tuning, prompt engineering, zero-shot learning, or few-shot learning, depending on available data and specific clinical requirements.
- Multimodal FMs that integrate medical imaging and text, such as clinical reports and patient history, may offer more comprehensive diagnostic support and improve clinical decision-making.
- Generative artificial intelligence (AI) capabilities can be integrated into FMs to enable synthetic image generation and corresponding annotations as well as image augmentation and enhancement, addressing critical data scarcity challenges in medical imaging.
- Future development should focus on specialized models tailored to specific medical conditions and expand their capabilities to integrate multimodal data, including real-time inputs from wearable devices, to enhance personalized care.
- The successful adoption of generative AI and FMs in medical imaging will require explainability, clinical trust, and thoughtful integration into existing health care systems.

FM Development and Training

FMs are developed through a multiphase process beginning with large-scale dataset preparations from diverse sources to ensure adequate quality and usability, which involves selecting, organizing, cleaning, enriching, and preserving data. In the initial phase, FMs undergo training on massive unlabeled datasets using self-supervised learning to capture generalizable patterns and representations, then are adapted to specific tasks with smaller labeled datasets (5,9). This initial training requires substantially more data and computational resources than task-specific adaptation, often using high-memory, multiple GPU or tensor processing unit clusters (10). FM development has primarily been driven by well-resourced institutions with substantial technical expertise due to high infrastructure and data costs (11). However, these technologies have become more accessible through open-source tools such as Medical Open Network for AI (12), which enable health care systems with limited resources to fine-tune existing models.

Types of FMs

FMs are categorized by architecture and modality into text-based, image-based, multimodal, generalist, and agentic and multiagent FMs to improve clarity and understanding, with various examples of FMs used in health care summarized in Tables 1 and 2.

Text-based FMs

Text-based FMs, particularly large language models, are AI systems used in natural language processing to understand and generate humanlike text. In health care, they are used to analyze and summarize large volumes of medical content, such as clinical notes, radiology reports, and research articles, supporting tasks like documentation, decision support, and information retrieval (3). The following subsections highlight different types of text-based FMs and their applications in health care (Table 1).

Bidirectional encoder representations from transformers

Bidirectional encoder representations from transformers (BERT), which use an encoder-only architecture, are pretrained on large text collections and fine-tuned for specific tasks such as extracting clinical information and classifying medical texts (13). In this context, an encoder refers to the part of the model responsible for processing and understanding input text by converting words into contextualized numerical representations, similar to how a radiologist interprets images to extract key clinical information. BioBERT (14), a variant of BERT pretrained on biomedical text (eg, PubMed articles), excels in biomedical text mining, information extraction, and question answering, outperforming general BERT models in health care applications. Additional examples of BERT-based FMs used in health care are summarized in Table 1.

Text-based generative pretrained transformers

Generative pretrained transformers (GPTs) use a decoder-only architecture, which is highly effective for generating coherent contextually appropriate text, making them particularly useful for tasks such as medical report generation (15) and clinical dialogue systems. In this architecture, the decoder is the component responsible for generating text output based on an input prompt, similar to how a radiologist might dictate a structured report after reviewing imaging findings, using prior context to guide each sentence. Models like GatorTronGPT (16), adapted for health care, excel in biomedical text generation. GatorTronGPT, trained on a combination of clinical notes and general English texts, is especially strong in relation extraction and the creation of synthetic clinical documentation, advancing the use of large language models in health care. BioMedLM (17), a language model trained on biomedical literature, advances biomedical text analysis and supports health care research. Although models like OpenAI's ChatGPT have enhanced natural language processing with advanced generative capabilities, their training details remain undisclosed (18).

Pathways Language Models

Med-PaLM (19), based on Google's PaLM (Pathways Language Model) architecture and fine-tuned on medical question-answering tasks, supports clinical decision-making and report generation. Pathways Language Models use Google's Pathways system, which enables efficient training and deployment across diverse tasks in health care applications.

Llama models

Meta's Llama (11) is an open-source general large language model designed for high-performance language tasks, with applications like summarizing medical literature and clinical dialogues. Llama-2 (20) improves on this with instruction-based tuning, and its 70 billion–parameter model is a top performer in natural language tasks (18). Med42 (21), built on Llama-2's 70 billion–parameter model, is aimed at increasing access to medical

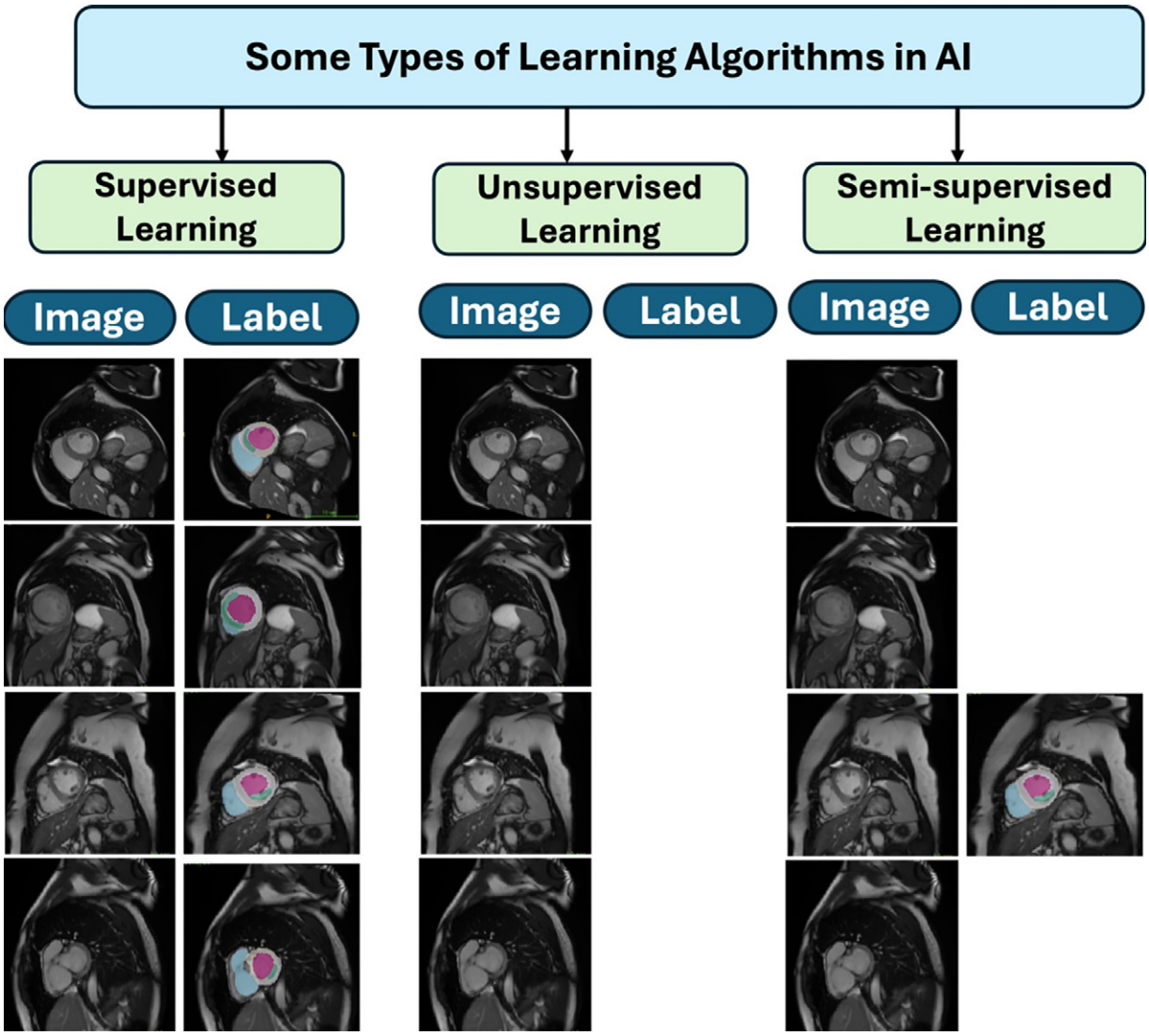

Figure 1: Cardiac imaging examples illustrate the three common types of learning algorithms in artificial intelligence (AI): supervised, unsupervised, and semi-supervised learning. Supervised learning relies on labeled input-output pairs, such as segmentation masks for myocardial tissue. It is commonly used for tasks like regression (eg, predicting ejection fraction) or classification (eg, detecting myocardial infarction). A major limitation of supervised learning is the cost and time required to generate large annotated datasets. Unsupervised learning, on the other hand, works without labels, uncovering patterns and structures within the data, such as clustering similar patient groups. This approach helps identify latent features but generally requires large datasets for meaningful insights. Semi-supervised learning combines a small amount of labeled data with a larger set of unlabeled data, reducing the need for manual annotations while still training effective models. This approach is particularly useful in medical imaging, where labeled data are scarce and expensive to acquire.

knowledge. Additional examples of Llama-based FM applications in health care are summarized in Table 1.

Image-based FMs

Image-based FMs are AI models initially trained on large collections of medical images. They can be fine-tuned for specific tasks in different specialties, including radiology, pathology, and cardiology (22,23). These models help with applications such as image classification, segmentation, and diagnosis. The following subsections describe several examples of image-based FMs and how they can be used in health care (Table 2).

Vision transformers

Vision transformers (ViTs) (Fig 2) analyze medical images by dividing them into patches and capturing global context, making them effective for tasks like image segmentation (Fig 3) (22,24,25). Although early ViTs had high computational demands (26), models like DeiT (data-efficient image transformer) reduce data requirements with strong performance (27), and radiology-specific studies have confirmed their effectiveness (28).

The FlashAttention technique (29) improves the efficiency of transformer models, including those used in medical imaging, by reducing memory and computational overhead (29,30). ViT-based FMs include BrainSegFounder for neuroimage segmentation (31) and ScarNet (32), which was developed for cardiac scar quantification. Additional examples of ViT-based FMs in health care are summarized in Table 2.

U-Net

U-Net (33) is a widely adopted architecture in medical imaging, known for its effectiveness in detailed image segmentation. Originally developed for biomedical image analysis, U-Net excels at localizing fine anatomic structures. Combining U-Net (Fig 4) with transformer architectures enhances performance by uniting

Table 1: Text-based FMs for Medical Applications

Model Variant	Developer or Organization	Description
BERT based		
BioBERT (14)	Korea University	BERT-based model that undergoes pretraining activities on biomedical text, including PubMed articles. Optimized for biomedical text mining, information extraction, and question answering. Improves data extraction and analysis in health care.
ClinicalBERT (105)	MIT and Microsoft with initialization from BioBERT, version 1.0	Specialized for clinical texts like notes and discharge summaries, improving medical inference and entity recognition.
PubMedBERT (106)	Microsoft	Undergoes pretraining activities on PubMed abstracts, improves performance in biomedical tasks like question answering and text classification.
BioLinkBERT (107)	Stanford University	Integrates biomedical knowledge from linked text and structured data, enhancing entity recognition and relation extraction.
DRAGON (108)	Stanford University and EPFL	BERT-based model that integrates medical knowledge from multiple text sources to enhance the understanding of complex biomedical data.
NYUTron (109)	Multi-institutional collaboration (NYU Langone Health, NYU, NVIDIA, Genentech, and CIFAR)	BERT-based model trained on clinical notes, excels as a universal predictor, showcasing the value of domain-specific pretraining activities for medical tasks.
GPT based		
BioMedLM (17)	Stanford University and Databricks	Language model trained on biomedical literature, excels in tasks related to information extraction and biomedical text analysis.
GatorTronGPT (16)	University of Florida and NVIDIA	Tailored for biomedical natural language processing, excels in relation extraction and clinical text generation.
PaLM based		
Med-PaLM (19)	Multi-institutional collaboration (Google Research, DeepMind, National Library of Medicine)	Fine-tuned for encoding clinical knowledge and assisting in medical report generation and decision-making.
Llama based		
Med42 (21)	G42 Healthcare, Core42, and Cerebras Systems	An open-source clinical language model built on the Llama-2 framework with 70 billion parameters. Med42 is designed to enhance medical knowledge accessibility and is fine-tuned for clinical language tasks.
MEDITRON-70B (110)	Multi-institutional collaboration (EPFL, Idiap Research Institute, Open Assistant, Yale University)	A 70 billion–parameter medical language model based on Llama-2. It has been further trained on medical literature and specialized instruction datasets, making it highly effective for health care and biomedical tasks.
PMC-Llama (18)	Shanghai Jiao Tong University and Shanghai AI Laboratory	Fine-tuned on medical books and papers using the Llama model, excels in medical question-answering tasks, enhanced with medical-specific instructions for improved performance in health care.
Me-Llama (111)	Multi-institutional collaboration (Yale University, University of Florida, University of Texas Health Science Center)	A large language model specifically designed for medical applications through continual pretraining activities and instruction tuning.

Note.—Models are categorized according to their primary foundation architecture. BioMedLM could also be considered a generative language model. AI = artificial intelligence, BERT = bidirectional encoder representations from transformers, CIFAR = Canadian Institute for Advanced Research, CRFM = Center for Research on Foundation Models, EPFL = École Polytechnique Fédérale de Lausanne, FM = foundation model, GPT = generative pretrained transformer, MIT = Massachusetts Institute of Technology, ML = machine learning, NYU = New York University, PaLM = Pathways Language Model.

U-Net's ability to capture precise image features with transformers' strength in modeling broader contextual information. This integration is particularly useful for complex segmentation tasks that require both local and global understanding. For instance, TransUNet (34) embeds transformer blocks within the U-Net structure, while Swin Transformer (35) adopts a hierarchical approach to process both fine-grained and high-level image features. These models typically undergo large-scale pretraining activities followed by fine-tuning for specific medical applications. Additional examples of U-Net–based FMs in health care are summarized in Table 2.

Segment Anything Model

The Segment Anything Model (SAM) (36) is an image segmentation model originally trained on a vast and diverse dataset of general images. This extensive training allows these models to generalize well across various tasks and modalities. Once trained, SAM can be fine-tuned for medical applications, such as MRI, CT, and US, with minimal labeled data. It generates segmentations based on simple prompts, such as points, clicks, or bounding boxes, making it flexible and efficient for a wide range of clinical scenarios. MedSAM (23) is a version fine-tuned specifically for medical images and

Table 2: Image-based FMs for Medical Applications

Model Variant	Developer or Organization	Description
ViT based		
BrainSegFounder (31)	University of Florida and NVIDIA	A three-dimensional FM for neuroimage segmentation, designed for versatility across neurological applications such as brain tumor segmentation, stroke detection, and Alzheimer disease diagnosis.
ScarNet (32)	Multi-institutional collaboration (Northwestern University, University of Miami, Cedars-Sinai Medical Center)	A novel FM developed for cardiac scar quantification from left ventricular late gadolinium enhancement imaging.
DinoV2 (112)	Multi-institutional collaboration (Meta AI Research, Inria)	Self-supervised vision model enabling superior medical image feature extraction with minimal labeled data.
Virchow (113)	Multi-institutional collaboration (Paige, Microsoft Research, Memorial Sloan-Kettering Cancer Center, NSW Health Pathology, University of Rochester)	Specialized FM for clinical-grade pathology image analysis and rare cancer detection, built on DINOv2 (112) architecture.
MedImageInsight (114)	Multi-institutional collaboration (Microsoft Health and Life Sciences, Microsoft Research, Microsoft Azure AI, University of Wisconsin-Madison, University of Washington)	Open-source medical imaging embedding model trained on diverse medical images and associated text, achieving state-of-the-art performance across various classification and retrieval tasks.
RadDino (115)	Multi-institutional collaboration (Microsoft Research, Microsoft Health and Life Sciences, Microsoft Azure AI, University of Cambridge)	Self-supervised ViT tailored for radiology applications, focusing on robust medical image analysis.
RayDino (116)	Multi-institutional collaboration (Meta AI, CentraleSupélec Université Paris-Saclay, Hôpital Cochin AP-HP, Université Paris Cité)	Self-supervised ViT designed for x-ray analysis through holistic self-supervised learning approaches.
U-Net based		
TransUNet (34)	Multi-institutional collaboration (Johns Hopkins University, University of Electronic Science and Technology of China, Stanford University, East China Normal University, PAII)	Integrates transformer architecture with U-Net to enhance medical image segmentation, effectively capturing both local and global image features.
Swin Transformers (35)	Microsoft Research	Uses a hierarchical ViT approach with shifted windows, enabling the model to capture local and global context for tasks like medical imaging.
MEDT (117)	Multi-institutional collaboration (Tianjin University of Science and Technology, Tianjin Sino-German University of Applied Sciences)	Combines transformers with U-Net by replacing convolutional layers, allowing for global context understanding while maintaining precise segmentation.
UNETR (118)	Multi-institutional collaboration (NVIDIA, Vanderbilt University)	Leverages transformers within the U-Net framework for improved global feature capture and accurate medical image segmentation.
U-Mamba (119)	Multi-institutional collaboration (University Health Network, University of Toronto, Vector Institute)	A U-Net–based hybrid image segmentation model combining CNNs and attention mechanisms for segmenting complex structures with improved accuracy. Multiscale learning and custom loss functions address class imbalance and boundary precision.
SAM based		
SAM (36)	Meta AI Research	Adaptable for segmentation across modalities like MRI, CT, and US, reducing the need for large labeled datasets.
MedSAM (23)	Multi-institutional collaboration (University Health Network, University of Toronto, Vector Institute, Western University, New York University, Yale University)	SAM-based fine-tuned for medical image segmentation across MRI, CT, and US.
MedSAM2 (37)	Multi-institutional collaboration (University Health Network, Vector Institute, University of Toronto)	The latest iteration of MedSAM (23), further enhances segmentation capabilities by treating medical images as video sequences, allowing it to handle more complex tasks and modalities.

(Table 2 continues)

Table 2 *(continued)*: Image-based FMs for Medical Applications

Model Variant	Developer or Organization	Description
Mammo-SAM (38)	Multi-institutional collaboration (Sun Yat-sen University, Peking University, Tianjin University)	Leverages SAM for automatic breast mass segmentation in mammograms, outperforming traditional models like U-Net.
SAM-Adapter (103)	Multi-institutional collaboration (KOKONI Tech, Zhejiang University, Huzhou University, Singapore University of Technology and Design, Beihang University)	Extends SAM's functionality for polyp identification in colonoscopy images; in radiology, this approach can be used to enhance accuracy of virtual colonoscopy.
ONCOPILOT (120)	Multi-institutional collaboration (Raidium, AP-HP, Université Paris Cité, INRIA, Center d'Imagerie du Nord)	A promptable FM adopted from SAM, specifically designed for CT-based solid tumor evaluation.
MONAI framework (12)	Extensive multi-institutional collaboration (NVIDIA, King's College London, and more than 15 institutions)	Integrates SAM2 across a broad range of applications, including radiology and pathology, showcasing its versatility in clinical practice.
VISTA3D (121)	Multi-institutional collaboration (NVIDIA, University of Arkansas for Medical Sciences, National Institutes of Health, University of Oxford)	A unified segmentation FM for three-dimensional medical imaging, though its current application remains limited to CT scans (32).
Generative AI based		
CycleGAN (39)	University of California, Berkeley	Uses cycle-consistency loss for accurate, reversible medical image domain translation, enabling consistent mapping between different imaging modalities.
UVCGAN (40)	Brookhaven National Laboratory	Combines ViT with the CycleGAN framework (39) to enhance nonlocal pattern capture in images, excelling in tasks like data augmentation and image reconstruction.
LeFusion (100)	Multi-institutional collaboration (USTC, Beihang, EPFL, Stanford University)	A diffusion-based generative AI used in cardiac imaging to synthesize myocardial pathology data, improving segmentation and MRI reconstruction by generating realistic lesion textures.

Note.—Models are categorized according to their primary architectural foundation. TransUNet, MEDT, UNETR, and U-Mamba combine transformers with U-Net architectures; UVCGAN combines ViT with GAN; Swin Transformers could be categorized as either ViT based or U-Net based depending on application. The Medical Open Network for AI framework is included under SAM-based models due to its prominent integration of SAM2 capabilities across medical imaging applications. However, Medical Open Network for AI functions as a comprehensive open-source framework that supports multiple model architectures beyond SAM, including various deep learning models for medical image analysis. It serves as a platform that can incorporate different FMs rather than being a single-architecture model itself. AI = artificial intelligence, AP-HP = Assistance Publique–Hôpitaux de Paris, CNN = convolutional neural network, EPFL = École Polytechnique Fédérale de Lausanne, FM = foundation model, GAN = generative adversarial network, MIT = Massachusetts Institute of Technology, MONAI = Medical Open Network for AI, PAII = Ping An Insurance (Group) Company of China, SAM = Segment Anything Model, USTC = University of Science and Technology of China, ViT = vision transformer.

performs well even when only a small amount of annotated data are available, reducing the cost and time needed for training (Fig 5). MedSAM2 (37) extends this approach by treating medical images as video sequences, which helps improve segmentation accuracy in more complex, multiframe imaging tasks. In breast imaging, for example, Mammo-SAM has been shown to outperform U-Net in automatically segmenting breast masses (38). One important limitation, however, is that SAM currently depends on user input (prompts) to generate segmentations, which may limit its use in fully automated workflows. Additional examples of SAM-based FMs in health care are summarized in Table 2.

Generative AI

Generative AI, when combined with transformer architectures, has enabled FMs with strong generative ability. These models can generate synthetic data to improve performance (eg, synthesizing MRI from CT scans). Models like CycleGAN (39) enable accurate and reversible translation between image domains (eg, CT to MRI), making them valuable for tasks such as modality conversion. UVCGAN (40) integrates ViT to better capture long-range anatomic relationships.

Diffusion-based models, such as Medical AI for Synthetic Imaging, or MAISI (41), use a step-by-step process to generate realistic synthetic images. In simple terms, a diffusion model starts with random noise and gradually denoises it to form a detailed image. This approach mimics how high-spatial-resolution images might be constructed over time, making it well suited for generating three-dimensional medical images like CT scans from scratch. Diffusion models are especially helpful when high-quality annotated data are limited. Additional examples of generative AI–based models in health care are summarized in Table 2.

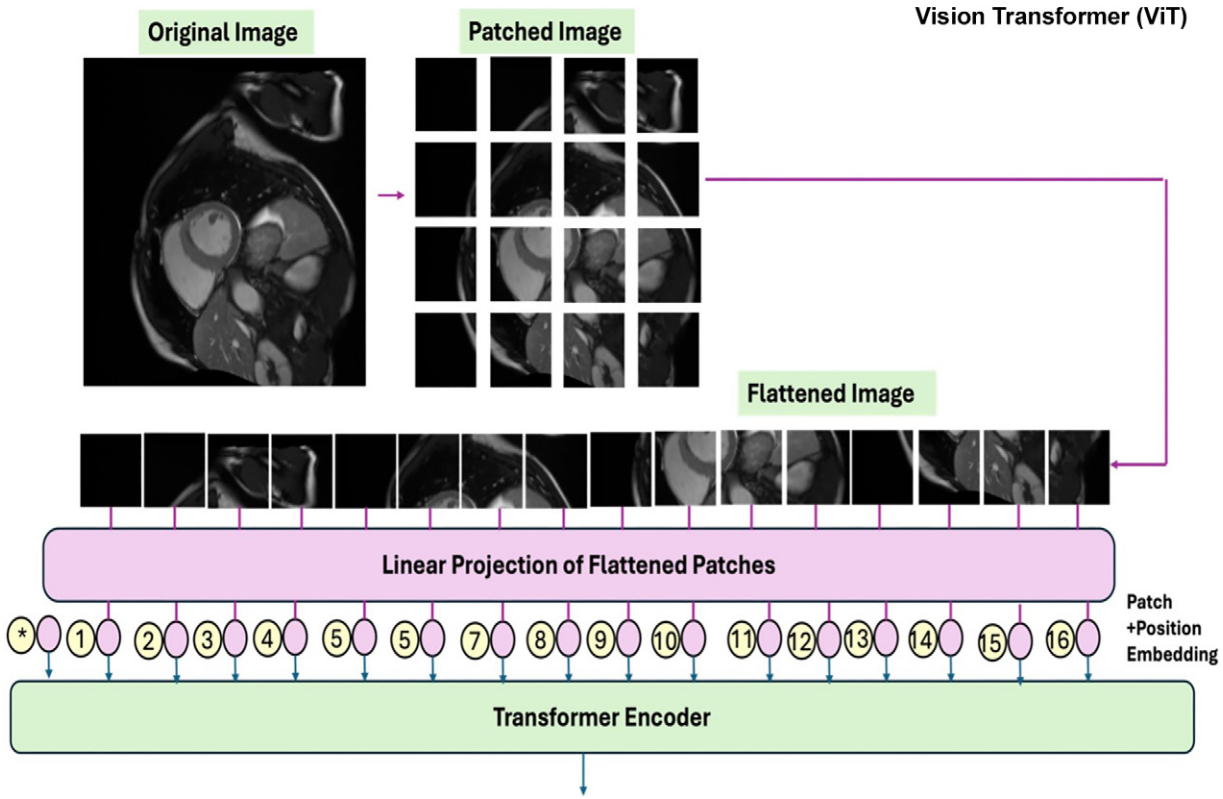

Figure 2: Diagram illustrates the process of transforming cardiac MRI scans for analysis using a vision transformer (ViT) model. The original image is divided into smaller, uniformly sized patches (grid of squares). These patches are then flattened (converted into linear sequences of pixels) and processed through a linear projection layer. This step converts the patches into a format that can be fed into the transformer model, where each patch is treated like a sequence token. Positional embeddings are added to maintain spatial relationships between patches. The transformer encoder processes these tokens, enabling the model to learn the global relationships between different regions in the image. This capability is crucial for analyzing medical images, such as identifying structures or abnormalities in cardiac MRI.

Multimodal FMs

Multimodal FMs combine information from multiple sources (ie, modalities), such as text, image, audio, and video, to perform complex tasks that require both visual and textual understanding (42). These models are particularly relevant in radiology, where integrating imaging data with clinical context is essential. The following subsection provides examples of such models and their applications in health care.

Contrastive Language–Image Pretraining models

In contrastive learning models, an image encoder and a text encoder are used to align medical images with clinical reports through contrastive learning, enabling improved multimodal understanding in health care applications. Contrastive Language–Image Pretraining (CLIP) models (42) associate images with textual descriptions, supporting cross-modal retrieval and classification in health care. MedCLIP (43) and CXR-CLIP (44) have been used to link medical images with clinical reports. Additional examples of multimodal CLIP-based FMs in health care are summarized in Table 3.

Multimodal GPTs

BiomedGPT (45) is a multimodal FM that processes both medical images and clinical text, enabling tasks like visual question answering and report generation. Its ability to handle diverse biomedical data makes it effective for radiology interpretation and clinical summarization. Additional examples of multimodal GPT-based FMs in health care are summarized in Table 3.

Multimodal Gemini-based models

Med-Gemini (46) is an advanced multimodal model for medical imaging that can enhance diagnostics by integrating images and clinical data, excelling in visual question answering and diagnostic support.

Multimodal fusion

Multimodal fusion refers to the integration of different types of data, such as imaging and text, to improve clinical interpretation and decision-making. In radiology, this means combining visual information from scans with associated clinical text (eg, indications, prior reports) to improve diagnostic accuracy and report generation. RadFM (47) is an FM that combines radiology images and clinical text to generate accurate reports. Trained on 16 million paired scans and descriptions (MedMD) and fine-tuned on a radiology-specific dataset (RadMD), RadFM outperforms models like OpenAI's GPT-4V in producing precise, clinically relevant radiology reports (47).

Vision language models

Flamingo (48), Open-Flamingo (49), and MedFlamingo (50) are advanced vision language models designed to integrate visual and textual data, making them powerful tools in medical imaging. Flamingo (48) is a versatile model excelling in few-shot learning by integrating pretrained vision and language models. Open-Flamingo (49) enables training autoregressive vision language

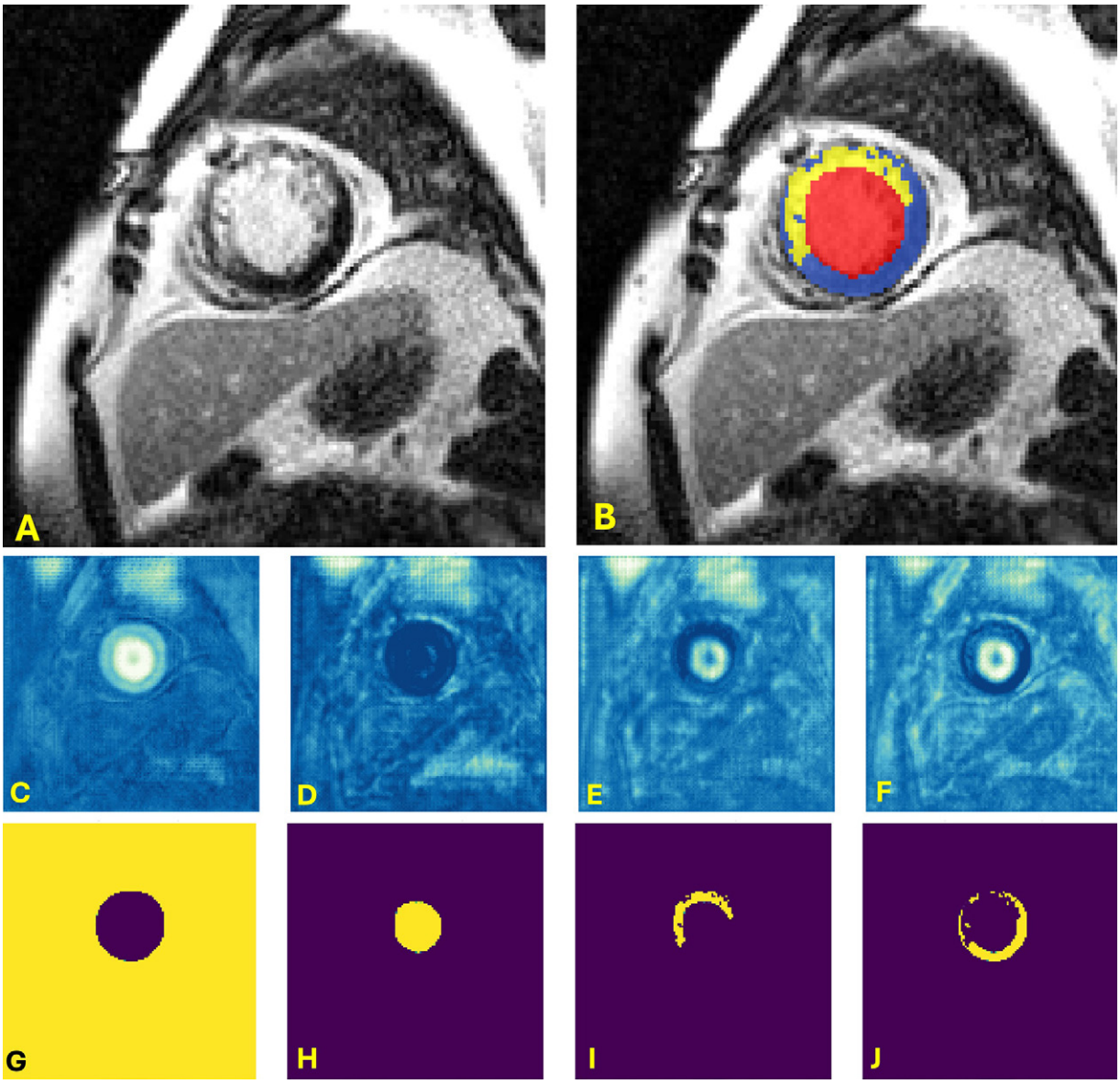

Figure 3: Multiclass cardiac MRI segmentation visualization using a transformer-based foundation model. **(A)** Original short-axis cardiac MRI scan of the left ventricle. **(B)** Predicted segmentation mask overlaid on the input image, where red represents the blood pool, blue indicates myocardium, and yellow denotes scar tissue. **(C–F)** Raw logits maps generated by the transformer model show unnormalized predictions before softmax activation for **(C)** background, **(D)** blood pool, **(E)** scar, and **(F)** myocardium. The logits represent the model's learned discriminative features for each cardiac structure. **(G–J)** Corresponding probability maps obtained after softmax normalization of the logits, visualized with yellow indicating high probability (1.0) and purple indicating low probability (0.0) for each class (background, blood pool, scar, and myocardium, respectively). The distinct separation in the probability maps, particularly evident in the blood pool **(H)** and myocardium **(J)** regions, demonstrates the model's robust capability in delineating different cardiac structures. The progression from raw logits to probability maps illustrates the model's decision-making process, highlighting its effectiveness in multiclass cardiac segmentation while maintaining clear boundary distinctions between different cardiac tissues.

models on tailored datasets for clinical customization. MedFlamingo (50), which undergoes pretraining activities on medical datasets, specializes in tasks like medical text and image interpretation. Flamingo-CXR (51) is a report generation system for chest radiographs. Additional examples of multimodal vision language models in health care are summarized in Table 3.

Zero-shot learning models

Zero-shot learning allows a model to perform a new task, even one it is not specifically trained for, by using knowledge it learned during general pretraining. For example, CT-CLIP (52) can detect abnormalities on chest CT scans without needing additional labeled training data, often outperforming fully supervised models. However, not all models claiming zero-shot ability are truly generalizable. For instance, ChexZero (53) works well for chest radiographs but was specifically trained for that purpose, so it is not considered a true zero-shot model. Additional examples of multimodal zero-shot learning FMs are summarized in Table 3.

BERT models

BERT models, including AlphaBERT (54) and BEHRT (55), are based on BERT (13), a language model that learns from large

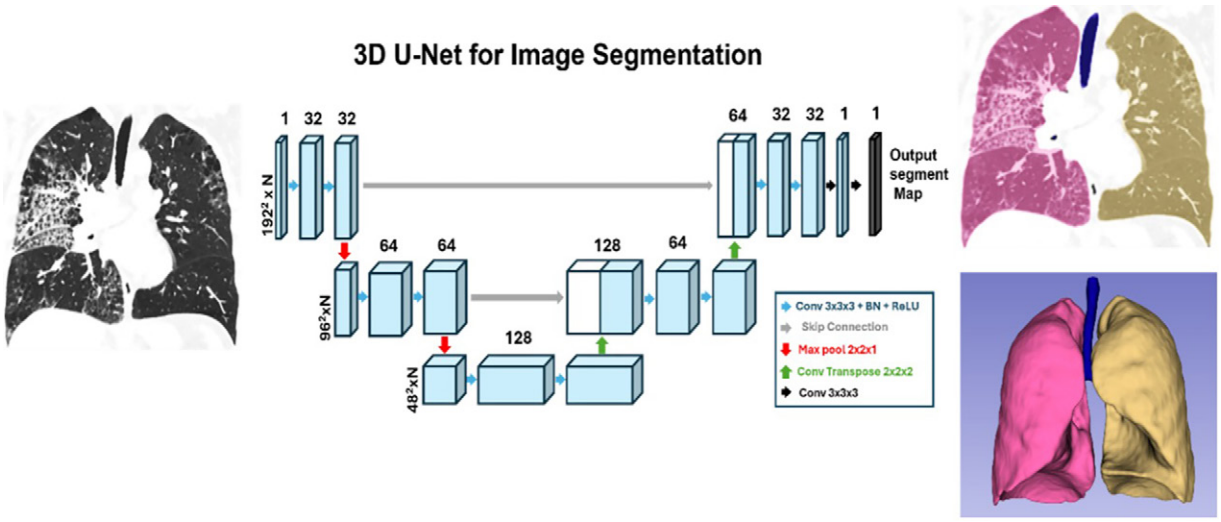

Figure 4: Diagram shows the architecture of a three-dimensional (3D) U-Net model for volumetric lung segmentation on chest CT scans. The network follows an encoder-decoder structure with skip connections, processing three-dimensional volumes through multiple resolution levels. The encoder path (left to right) consists of repeated blocks of three-dimensional convolutions (Conv 3 × 3×3), batch normalization (BN), and rectified linear unit (ReLU) activation, followed by max pooling operations (2×2×1) for downsampling. The decoder path (right to left) uses transposed convolutions (2×2×2) for upsampling, combined with skip connections from the encoder path to preserve fine spatial details. Feature maps are shown with their respective sizes (height × width × channels), starting from the input resolution of $192^2 \times N$ to generate the final segmentation map. The numbers above and below the blue convolutional blocks indicate the number of feature channels at each layer. The model demonstrates effective segmentation of left and right lungs (shown in pink and yellow, respectively) in both two-dimensional sections and three-dimensional visualizations, highlighting its ability to capture complex anatomic structures while maintaining spatial consistency. The model also enables accurate quantification of lung volumes, serving as a critical imaging biomarker for patients with pulmonary fibrosis.

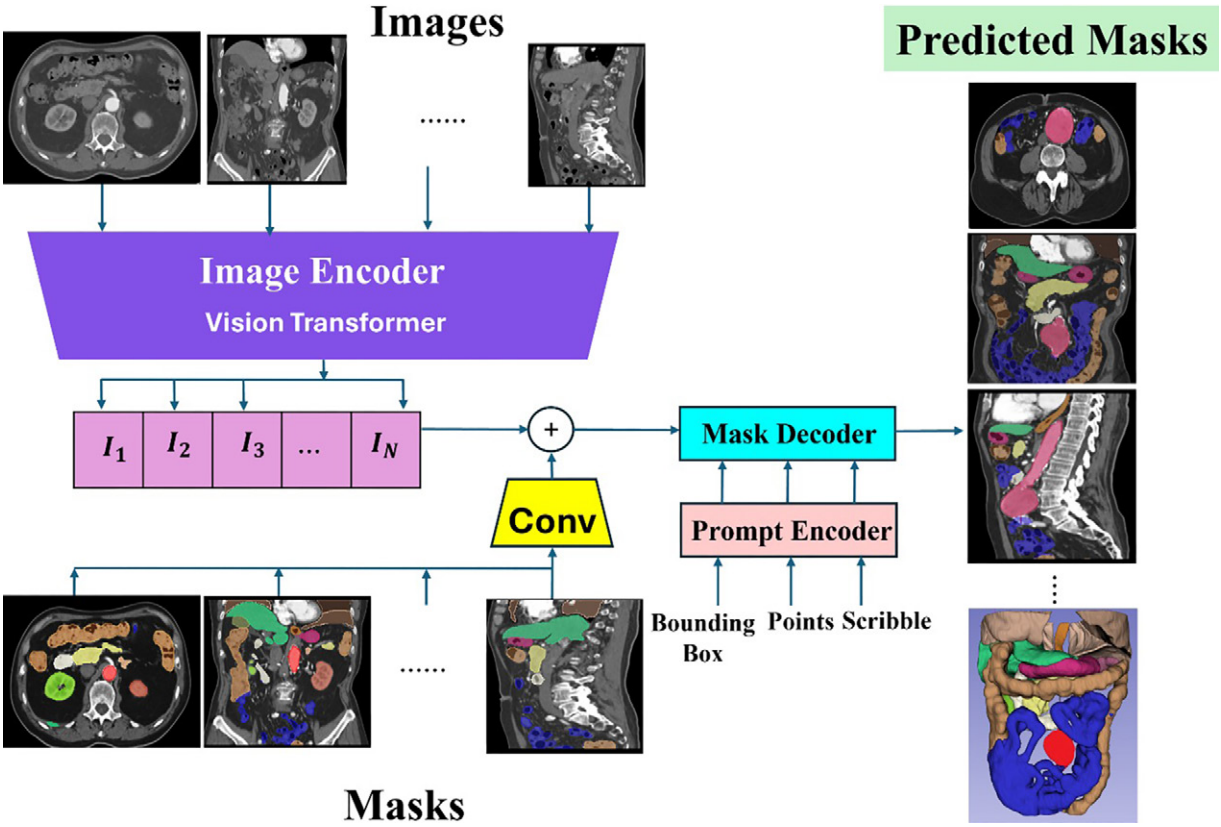

Figure 5: Diagram illustrates the architecture of a foundation model (FM) adaptation framework using Segment Anything Model (SAM) for medical image segmentation. The pipeline consists of three main components. First, an image encoder based on a vision transformer processes input CT and MRI scans and generates feature representations (I_1 to I_N). Second, a parallel processing stream incorporates prior knowledge through segmentation masks, which are processed by a convolutional layer (Conv) and combined with the encoder features through element-wise addition (+). Third, a mask decoder integrates information from a prompt encoder processing auxiliary inputs (bounding box, points, and scribble) to generate the final segmentation masks. The right panel demonstrates example results showing accurate multiorgan segmentation across different anatomic views (axial, coronal, and sagittal), with each organ denoted by distinct colors. This architecture uses SAM's transformer-based FM capabilities to capture long-range dependencies (ie, relationships between distant parts of an image that are important for understanding the overall structure). It incorporates interactive guidance through prompts, enabling robust and accurate organ segmentation across various medical imaging scenarios.

Table 3: Multimodal FMs for Medical Applications

Model Variant	Developer or Organization	Description
CLIP based		
CLIP (42)	OpenAI	Learns to associate images with textual descriptions, enabling cross-modal retrieval and classification tasks in health care.
MedCLIP (43)	Multi-institutional collaboration (University of Illinois Urbana-Champaign, Adobe)	Tailored for medical imaging applications, leveraging unpaired medical images and text through contrastive learning.
CXR-CLIP (44)	Kakaobrain	Combines image-label data with image-text pairs using prompt-based text generation, enhancing chest radiograph analysis.
Mammo-CLIP (104)	Multi-institutional collaboration (Boston University, University of Pittsburgh)	A VLM trained on mammogram-report pairs, demonstrating strong performance in classifying and localizing key findings for breast cancer detection across multiple datasets.
AnatCL (99)	Multi-institutional collaboration (University of Turin, CEA Paris-Saclay)	An innovative anatomic FM that leverages features like cortical thickness and gray matter volume to create robust representations for brain MRI analysis.
BiomedCLIP (122)	Multi-institutional collaboration (Microsoft Research, University of Washington, Providence Genomics)	Extends CLIP architecture to the biomedical domain, enhancing image-text pairing for medical applications.
PMC-CLIP (123)	Multi-institutional collaboration (Shanghai Jiao Tong University, Shanghai AI Laboratory)	Tailored for the biomedical domain using PubMed Central literature and medical images, enhancing image-text retrieval and understanding.
GPT based		
BiomedGPT (45)	Extensive multi-institutional collaboration (more than 12 institutions, including Lehigh University, Harvard Medical School, Stanford University, Mayo Clinic)	Handles both medical images and text, enabling visual question answering and medical report generation.
GPT-4 and GPT-5	OpenAI	The multimodal versions of GPT-4 and GPT-5, capable of handling text and visual data, ideal for medical image analysis and multimodal reasoning.
Gemini based		
Med-Gemini (46)	Multi-institutional collaboration within Alphabet (Google Research, DeepMind, Google Cloud, Verily)	Advanced multimodal models for medical imaging enhance diagnostics by integrating images and clinical data, excelling in visual question answering and diagnostic support.
Fusion based		
RadFM (47)	Multi-institutional collaboration (Shanghai Jiao Tong University, Shanghai AI Laboratory)	Integrates radiology images with text for report generation, improving diagnostic accuracy.
VLM based		
Flamingo (48)	DeepMind	Advanced VLM designed for few-shot learning in medical imaging tasks.
MedFlamingo (50)	Multi-institutional collaboration (Stanford University, Stanford Medicine, Hospital Israelita Albert Einstein, Harvard Medical School)	Undergoes pretraining activities on medical datasets for medical visual question-answering tasks.
Flamingo-CXR (51)	Multi-institutional collaboration (Google DeepMind, Google Research, OpenAI, GlaxoSmitheKlineAI, Apollo Radiology International)	A report-generation system for chest radiographs.
PMC-VQA and MedVINT (124)	Multi-institutional collaboration (Shanghai Jiao Tong University, Shanghai AI Laboratory)	Visual instruction-tuned model for interpreting medical images and textual information in radiology report generation.
ViLBERT (125)	Multi-institutional collaboration (Georgia Institute of Technology, Oregon State University, Facebook AI Research)	Extension of BERT designed for vision-and-language tasks like visual question answering and image captioning.
Llava-Med (126)	Microsoft	Adapted for medical applications, excels in medical image captioning and question-answering tasks.

(Table 3 continues)

Table 3 *(continued)*: Multimodal FMs for Medical Applications

Model Variant	Developer or Organization	Description
Zero-shot based		
CT-CLIP (52)	Multi-institutional collaboration (University of Zurich, ETH Zurich, Istanbul Medipol University, Boston University)	A FM using chest CT volumes and radiology reports for true zero-shot detection of abnormalities without requiring task-specific training data.
EchoCLIP (101)	Multi-institutional collaboration (Cedars-Sinai Medical Center, UCLA, UCSF, San Francisco Veterans Affairs Medical Center)	VLM featuring zero-shot learning for echocardiogram interpretation, assessing cardiac function and predicting clinical outcomes.
BERT based		
AlphaBERT (54)	Multi-institutional collaboration (National Taiwan University, National Taiwan University Hospital)	Combines medical and general text for implementing pretraining activities for large FMs, enhancing capabilities in diverse text processing tasks.
BEHRT (55)	University of Oxford	Follows the same approach as AlphaBERT (54), focusing on medical and general text integration.

Note.—Models are categorized according to their primary approach. BiomedGPT could fit in generalist models (Table 4); CLIP variants (MedCLIP, CXR-CLIP) could be categorized as both contrastive learning and VLM; AlphaBERT and BEHRT could fit in text-based models (Table 1); RadFM could overlap with VLM categories; GPT-4 and GPT-5 represent general-purpose models that could be categorized in generalist models (Table 4). Zero-shot learning allows models to perform new tasks without requiring task-specific training data. ADNI = Alzheimer's Disease Neuroimaging Initiative, AI = artificial intelligence, BERT = bidirectional encoder representations from transformers, CLIP = Contrastive Language–Image Pretraining, FM = foundation model, GPT = generative pretrained transformer, UCLA = University of California, Los Angeles, UCSF = University of California, San Francisco, VA = Veterans Affairs, VLM = Vision Language Model, VQA = visual question answering.

collections of text. These models are trained using both medical and general language data, helping them better understand and process clinical information such as patient histories, diagnoses, and radiology reports.

Generalist FMs

Generalist FMs are highly versatile, capable of processing multiple modalities, including images, text, video, and audio, across disease areas. They differ from multimodal FMs in their scope and versatility. Although multimodal FMs typically integrate specific data types (eg, images, text, and video) for particular applications, generalist FMs can handle diverse modalities across multiple domains and disease areas. They integrate data from various domains like radiology, pathology, and other medical domains (Fig 6). Llama-3 (56) is a generalist model capable of integrating image, video, and speech. For example, in cardiology, generalist FMs like Perceiver IO (57) can analyze diverse data, including cardiac imaging, clinical notes, videos, and heart sounds, offering a unified diagnostic approach. However, clinical use still requires substantial task-specific training, fine-tuning, and testing. Additional examples of generalist FMs are summarized in Table 4.

Agentic and Multiagent FMs

Recent advances have introduced agentic and multiagent FMs for complex health care tasks. Agentic FMs involve a single autonomous agent capable of setting goals, making decisions, and executing multistep actions, supporting tasks like diagnosis, treatment planning, and patient monitoring.

In contrast, multiagent FMs coordinate multiple specialized agents. These multiagent systems (58) enable agents to collaborate via natural language, combining clinical and operational tasks such as diagnostics, scheduling, and resource management. Multiagent frameworks have been described as the next paradigm in FMs, offering modularity, privacy, and scalability by allowing assistive or autonomous agents to operate within specific domains (58). These models can advance distributed, adaptive intelligence in health care.

Challenges, Limitations, and Mitigating Strategies for Successful Integration of AI in Medical Imaging

General Challenges

The successful integration of AI in medical imaging requires addressing a range of practical, ethical, and sociotechnical challenges. Key concerns include model bias (59), data privacy (60), hallucination (61), and sycophancy (62), where models tend to agree with user input regardless of whether it is correct, which can reinforce clinical mistakes. The spread of misinformation due to weak safeguards or intentional attacks like data poisoning—the deliberate manipulation of training data to corrupt or bias an AI model—is also a major risk (63,64). In addition, techniques like hypernudging (65), which subtly influence user decisions through personalized prompts or interfaces, raise concerns about transparency and clinician autonomy (Fig 7).

Despite progress in FM development, clinical adoption remains limited due to regulatory uncertainty around generalist AI systems, challenges integrating with existing picture archiving and communication system infrastructure, high computational costs, workflow disruptions, and limited real-world external testing for specific patient populations and imaging protocols. Further concerns include the reasoning reliability of large language models, which may appear to understand medical tasks without true comprehension (66), and the need for careful training, fine-tuning, and testing of FMs across broader biomedical applications (58).

Human-Computer Interaction

The human-computer interaction implications of FM integration in radiology remain underexplored, particularly regarding their

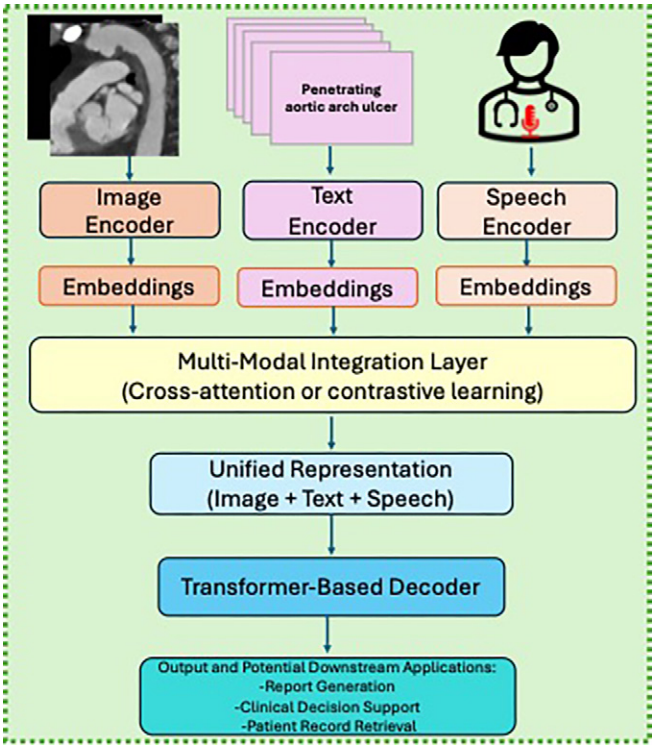

Figure 6: Generalist foundation models (FMs) handle multiple modalities (images, text, video, audio) and integrate data across domains like radiology, gastroenterology, pathology, and ophthalmology, showing promise in disease management. This diagram illustrates an example of a generalist FM for radiology, combining medical images, clinical text, and spoken dictation. The image encoder, a vision transformer, processes radiologic images like MRI and CT scans and generates embeddings that capture key visual features. The text encoder, based on models like bidirectional encoder representations from transformers or generative pretrained transformers, converts clinical notes or patient records into text embeddings, encapsulating relevant contextual information. Meanwhile, the speech encoder, using models such as wave to vector, or wav2vec, transcribes and processes spoken input such as radiologists' dictations into speech embeddings. These embeddings are integrated within a multimodal layer, enabling the model to comprehend and merge information from images, text, and speech. The unified representation enables the model to perform various tasks, such as generating radiology reports, retrieving patient information, or supporting clinical decisions by incorporating and synthesizing information from all input types.

impact on workflow, diagnostic confidence, and clinician well-being. Although FMs can streamline tasks across imaging modalities within a single system, their multitask nature complicates workflow integration and trust calibration, as performance may vary across applications.

Emerging reports suggest a potential link between AI usage and radiologist burnout (67). Seamless integration into existing systems such as the picture archiving and communication system is essential, but developers must also address risks like alert fatigue and dismissal bias, when clinicians may overlook true-positive results due to cognitive overload (68). Although AI models offer benefits (69–72), they further raise concerns around automation bias, when users may rely too much on AI and neglect their own clinical judgment (73). In one study (72), both radiologists and nonradiologists showed decreased diagnostic accuracy when exposed to incorrect AI advice, especially among less experienced users.

To mitigate these risks, FM-based tools should be designed to signal uncertainty and support critical reflection. Unlike traditional AI models, FMs use attention mechanisms that produce different forms of explanation, requiring new strategies for effective human-AI collaboration. The authors agree with previous recommendations (72,74) that FMs should serve as a supplemental second opinion, supported by organizational policies that promote cautious and informed use.

Interpretability and Explainability

A major challenge with FMs in medical imaging is their black box nature, which limits interpretability and may reduce trust in clinical use (75,76). Traditional explainability tools like saliency maps and Grad-CAM, designed for convolutional neural networks, are less effective for transformer-based FMs due to architectural differences (77,78). Instead, attention visualization methods better capture how FMs focus on image regions during decision-making. Although more appropriate, these approaches require new interpretation frameworks and user training. Developing effective explainability for FMs remains an active research area to support clinical evaluation and patient communication (79).

Data Dependency and Privacy

Unlike traditional deep learning models that require large labeled datasets, FMs reduce this dependency by using self-supervised learning on unlabeled data, followed by fine-tuning on smaller labeled sets (9,80). When additional data are needed, techniques like data augmentation and synthetic data generation can enhance performance, especially in limited-resource settings (81,82). However, FM performance can degrade over time due to evolving imaging protocols, scanner changes, or population shifts (83). Ongoing clinical oversight by teams of physicians and computer scientists is essential to audit outputs and retrain models as needed.

Data privacy is also critical, requiring secure handling of medical data and compliance with regulations such as the General Data Protection Regulation in the European Union and the Health Insurance Portability and Accountability Act in the United States (60). Removing low-utility attributes that could risk reidentification is recommended. For multi-institutional studies, federated learning offers a way to collaborate without centralizing data (60), though it poses technical and coordination challenges, making them less practical for real-world scenarios. In some cases, deidentification protocols and business associate agreements are more practical.

Computational Demands

FMs are computationally intensive, requiring substantial energy for training and deployment (10). Doo et al (84) explored AI's role in improving environmental sustainability in radiology by optimizing protocols, reducing scan times, and eliminating redundant imaging. Techniques like low-rank adaptation reduce GPU memory requirements through parameter-efficient fine-tuning, which decreases the number of trainable parameters that need to be stored and updated while maintaining performance (85).

Liability and Regulatory Considerations

AI-powered clinical tools are fallible, with risks including hallucinations, inaccuracies, and user misuse. A robust risk management framework is essential, and regulatory compliance is a key component. The U.S. Food and Drug Administration (FDA) currently oversees most AI tools in imaging, evaluating risk based on intended use under traditional medical device pathways (86). However, FMs present new challenges because of their generality,

Table 4: Generalist FMs for Medical Applications

Model Variant	Developer or Organization	Description
Multimodal integration models		
Perceiver IO (57)	DeepMind	A generalist model that processes various types of inputs (image, text, video, and audio) with a unified architecture. In medical imaging, Perceiver IO can analyze imaging data (eg, MRI or US) along with clinical notes and video data to improve diagnosis and treatment planning.
Gato (127)	DeepMind	A unified model capable of performing various tasks from playing video games to robotic control, combining different modalities like text, images, and video.
MERLOT (128)	University of Washington, Allen Institute for AI	A model designed for video understanding along with associated text. In medical imaging, MERLOT can be used to analyze real-time US video footage and combine it with clinical data to provide actionable insights.
X-CLIP (129)	Multi-institutional collaboration (Xiamen University, Alibaba Group)	An extension of CLIP that handles images, text, and video. It can be applied in medical imaging for tasks such as associating MRI scans with textual reports and video data like US recordings to improve diagnostic accuracy.
VideoBERT (130)	Google Research	A BERT-based model that handles video and associated text. In medical imaging, VideoBERT can analyze US video data, correlate it with textual clinical notes, and assist in diagnoses by learning from multimodal data.
VATT (96)	Multi-institutional collaboration (Google, Columbia University, Cornell University)	A transformer model that integrates visual, audio, and textual data. In cardiac imaging, VATT could assist in analyzing echocardiograms (visual), interpreting heart sounds (audio), and combining this with clinical notes (text) for comprehensive assessments.
Florence (131)	Multi-institutional collaboration (Microsoft Cloud and AI, Microsoft Research)	A multimodal FM that handles both vision and text, excelling in image classification, text-image alignment, and generation tasks.
PANDA (132)	Multi-institutional collaboration (Tsinghua University, Duke University)	A multimodal model capable of understanding and interacting with visual, auditory, and textual data. In cardiac imaging, PANDA could integrate visual imaging data and clinical audio (such as stethoscope sounds) with patient records to assist in diagnosis.
Agentic and multiagent FMs		
Agentic and multiagent FMs (58)	Not applicable	Multiagent systems coordinating specialized agents via natural language for clinical and operational tasks (diagnostics, scheduling, resource management). Offer modularity, privacy, and scalability, representing the next paradigm in distributed health care intelligence.

Note.—Models are categorized according to their primary approach. X-CLIP could be categorized in multimodal models (Table 3); VideoBERT could fit in multimodal models (Table 3) due to its video-text integration; Florence combines vision and text capabilities similar to multimodal models (Table 3) but is included here for its generalist approach; VATT could overlap with multimodal models (Table 3) due to its video-audio-text integration; Gato represents a general-purpose model that could fit in various categories depending on the specific application domain. The generalist models differ from multimodal FMs in their scope and versatility; although multimodal FMs typically integrate specific data types (eg, images, text, and video) for particular applications, generalist FMs can handle diverse modalities across multiple domains and disease areas. AI = artificial intelligence, BERT = bidirectional encoder representations from transformers, CLIP = Contrastive Language–Image Pretraining, FM = foundation model, VATT = Video-Audio-Text Transformer.

real-time adaptability, and lack of explainability. As of February 2025, the FDA has cleared at least one clinical application built on an FM: Aidoc's rib fracture triage tool, developed using the CARE1 Foundation Model (Clinical AI Reasoning Engine, version 1) (87).

The FDA has proposed a total product life cycle approach emphasizing input data quality, platform security, generalizability, and good machine learning practices over purely algorithmic traits (88). The FDA currently regulates AI-enabled software under its existing medical device framework. It has issued guidance, including the final Predetermined Change Control Plan (December 2024) and draft life cycle guidance (January 2025) (89,90). There is no separate law that gives the FDA specific powers to regulate FMs, and new legislation would be needed to create such authority.

Minimizing liability risk offers a second pillar. In the absence of specific federal regulations, judicial systems may serve as de facto guardrails for clinical AI use. Liability could fall anywhere along the AI supply chain, from physicians and health care systems to product developers, depending on the use case (91). This remains speculative due to the lack of test cases. Determining liability is complex, influenced by factors such as AI autonomy, physician expertise, on- versus off-label use, decision transparency, and FDA approval status (91). The learned intermediary doctrine will likely be central to determining future liability for manufacturers and clinicians. It positions physicians as intermediaries responsible for communicating product risks and weighing potential harms against benefits for patients (92). Some argue developers might intentionally design tools to involve clinicians and diffuse liability under this doctrine (93). For the time being, self-regulation offers the likeliest road ahead for users of medical imaging AI. These tools require testing on large datasets to

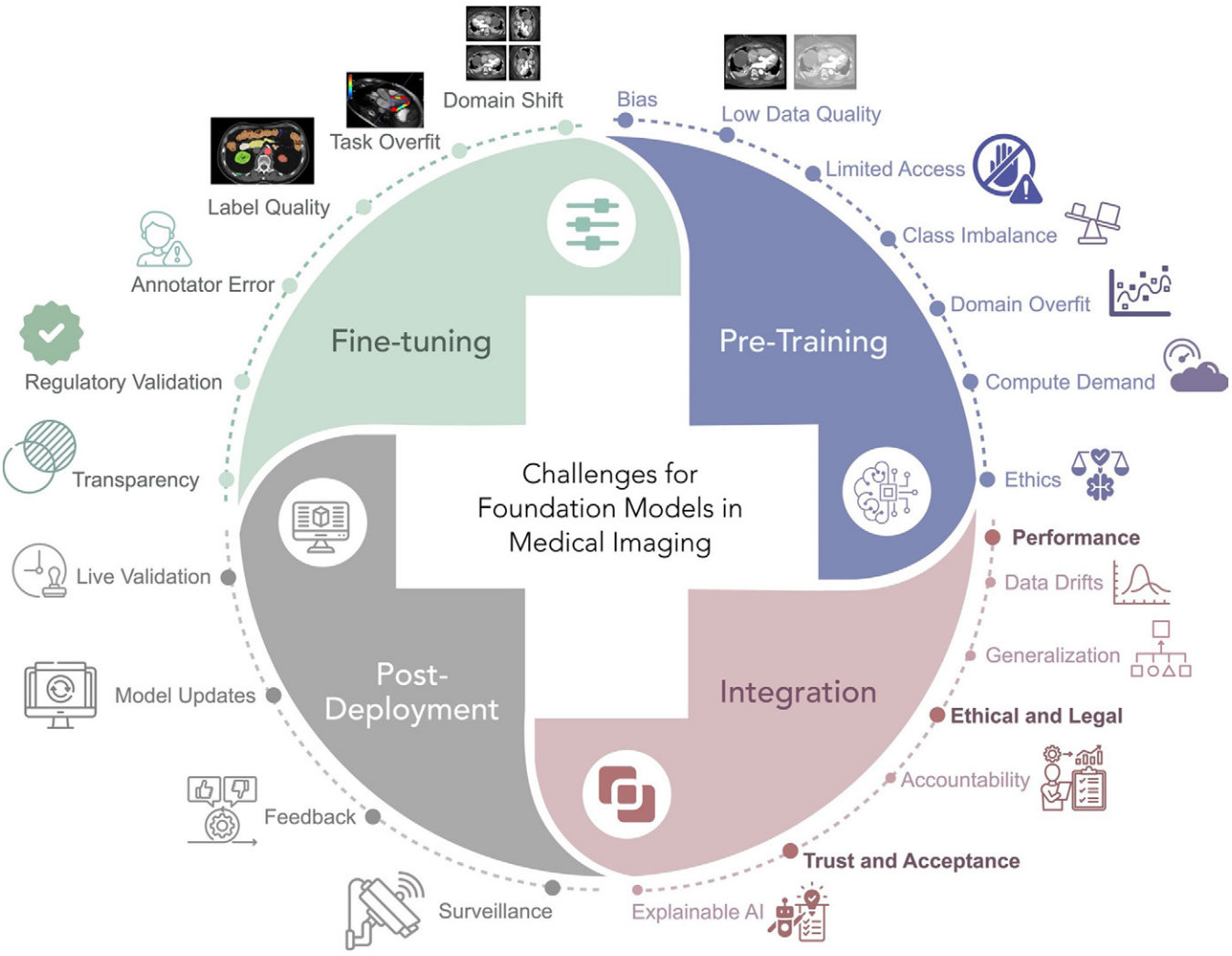

Figure 7: Diagram illustrates the challenges for foundation models (FMs) in medical imaging. FMs in health care face unique challenges throughout their life cycle, spanning pretraining, fine-tuning, postdeployment, and integration stages. During pretraining, issues such as data bias, low data quality, limited access, class imbalance, domain overfit, computational demand, and ethical concerns can hinder the development of robust models. Fine-tuning introduces challenges like domain shift, task overfit, label quality, annotator error, regulatory validation, and transparency requirements. The postdeployment stage poses hurdles, including live validation, model updates, feedback mechanisms, and surveillance. Integration challenges include performance issues, data drifts, generalization problems, ethical and legal considerations, accountability, trust and acceptance, and explainable artificial intelligence requirements. This classification highlights the multifaceted barriers to implementing FMs effectively in health care.

ensure accuracy and unbiased results. Given the potential for model performance degradation due to data drift from evolving imaging protocols, equipment changes, and patient population shifts, periodic postmarketing surveillance is essential to maintain AI accuracy and safety. A joint commission of stakeholders should develop a validation plan that includes randomized clinical trials and postmarket surveillance to ensure ongoing safety and performance.

Evaluating FMs in Medical Imaging: Lessons from Research Studies

Although clinical implementation of FMs in radiology remains limited, research studies have provided valuable insights into their performance and challenges in clinical-like settings. The following examples represent investigational experiments rather than approved clinical tools, as the regulatory landscape for FM deployment continues to evolve.

Several FMs have been evaluated under simulated clinical conditions for different organ systems in medical imaging (Fig 8), highlighting both possibilities and limitations. MedSAM (23), trained on more than 1.5 million diverse image-mask pairs, shows how large, heterogeneous datasets can improve segmentation performance across modalities and various organs. BrainSegFounder (31) improves diagnostic accuracy in brain lesion segmentation but faces challenges such as high computational cost and limited modality generalizability. Its development underscores the value of multistage pretraining and the need to balance model complexity with adaptability.

CXR-CLIP (44) demonstrates strong zero-shot capabilities and dataset generalization, emphasizing the importance of domain-specific design, such as prompt engineering and clinically meaningful data augmentation. Notably, increasing image-label data without careful balancing can harm performance, particularly in image-to-text retrieval tasks.

Together, these research examples illustrate the promise of FMs in radiology while highlighting key challenges, including scalability, generalizability, and clinical testing, that must be addressed for successful deployment.

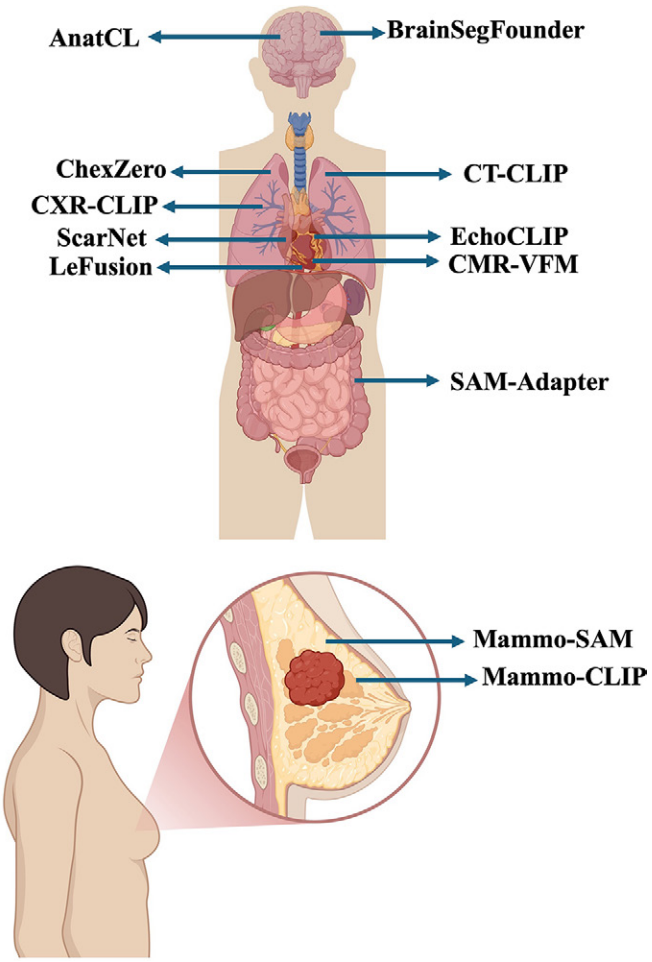

Figure 8: Diagram shows an overview of recent foundation models (FMs) developed for different organ systems in medical imaging. Each arrow points to an FM associated with a specific anatomic region or imaging modality. For example, BrainSegFounder (31) and AnatCL (99) target brain imaging; CT-CLIP (52), ChexZero (53), and CXR-CLIP (44) are applied to chest imaging; ScarNet (32), LeFusion (100), EchoCLIP (101), and CMR-VFM (102) are used for cardiac imaging; SAM-Adapter (103) is applied in abdominal imaging; and Mammo-SAM (38) and Mammo-CLIP (104) focus on breast imaging. AnatCL = Anatomical Contrastive Learning, CT-CLIP = Computed Tomography Contrastive Language–Image Pretraining (CLIP), CXR-CLIP = chest x-ray CLIP, CMR-VFM = Cardiac MR vision FM, Mammo-SAM = Mammography SAM, Mammo-CLIP = Mammography CLIP, SAM-Adapter = Segment Anything Model (SAM) Adapter. Created with BioRender.com (license obtained).

Future Directions

The future of FMs in medical imaging may follow multiple development pathways. One potential direction involves model refinement for specific conditions, imaging modalities, and patient demographics (94). Alternatively, advancing zero-shot and few-shot learning could enable models to handle new tasks without task-specific training. Retrieval-augmented generation approaches (95) enhance model performance by retrieving information from external knowledge databases, potentially improving clinical applications. For instance, task-specific fine-tuned models for tasks like tumor detection may outperform general-purpose models in some scenarios, while zero-shot approaches could be more practical for rare conditions or resource-limited settings (52). Addressing data scarcity and diversity through self-supervised and transfer learning will be crucial in these contexts.

A key research direction is developing multimodal FMs that process images, text, and clinical reports for comprehensive diagnostic insights, enhancing clinical decision-making. Integrating real-time data, like heart rate and electrocardiography, from wearables monitoring devices could further enable personalized, dynamic care through models such as Video-Audio-Text Transformer, or VATT (96). For radiologists preparing for FM adoption, evidence suggests that subspecialty-specific applications will achieve clinical readiness before generalist models. These necessitate strategic planning for gradual implementation across various imaging domains.

Choosing FMs for medical imaging involves assessing the clinical objective to ensure optimal performance. The type of task (image segmentation, disease classification, or risk prediction) guides the selection of the most suitable model, as each task benefits from different architectures and capabilities. Another key factor is the imaging modality, as models are optimized for various imaging types like MRI, CT, PET, and US. For instance, U-Net and its transformer-based variants are particularly effective for high-resolution MRI and CT segmentation (97).

Interpretability and explainability are vital in clinical settings, as clinicians need to understand model predictions for high-stakes decisions. Models with interpretability features, like attention-based transformers or those offering visual explanations (eg, saliency maps), enhance trust and facilitate clinical adoption of AI-driven insights (98).

Assessing a model's ability to generalize to real-world clinical data is vital, as it must handle diverse patient populations, imaging protocols, and varying image quality. Testing models across varied data sources ensures robustness. Multimodal models are especially valuable for comprehensive patient assessments, leading to accurate diagnostic insights. By considering clinical objectives, imaging modalities, data availability, interpretability, and generalizability, clinicians can select the most suitable FMs.

Practical implementation necessitates that radiologists develop institutional evaluation frameworks. These frameworks should include vendor assessment criteria, infrastructure readiness checklists, and protocols tailored to specific patient populations and imaging modalities.

Conclusion

Advancements in computational hardware and software and the introduction of foundation models (FMs) and generative artificial intelligence (AI) have made previously impossible tasks now achievable. FMs can transform radiology by automating complex tasks and improving diagnostic accuracy and efficacy. However, their integration into clinical practice requires careful fine-tuning for specific imaging modalities and medical conditions. Ensuring interpretability and trustworthiness, especially through explainable AI, is crucial for clinical adoption. Additionally, ethical considerations, such as ensuring equitable performance across specific patient populations, must be addressed. Building FMs responsibly requires proactive approaches rather than reactive ones. Collaborative efforts between industry and academia are essential for advancing FMs while tackling key challenges such as fairness, interpretability, and scalability. For radiologists, these necessitate strategic preparation, including staff education and evidence-based FM selection tailored to specific practice environments. Radiologists must develop institutional policies for FM evaluation and maintain

ongoing assessment of FM performance in their practice settings. Moreover, as FM continues to evolve in medicine and radiology, it is essential for radiologists to understand at least the fundamentals of FMs, their applications in radiology, and their associated advantages and limitations. Success will depend on radiologists' ability to make informed decisions about FMs and maintain clinical oversight during their adoption.

Deputy Editor: Linda Moy
Scientific Editor: Sarah Atzen

Author affiliations:

[1] Department of Radiology, Northwestern Memorial Hospital, 676 N Saint Clair St, Arkes Family Pavilion Ste 800, Chicago, IL 60611

[2] Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

[3] Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, Mass

[4] Department of Radiological Sciences, David Geffen School of Medicine at UCLA, Los Angeles, Calif

[5] Department of Radiology, Division of Cardiothoracic Radiology, University of California, Davis, Sacramento, Calif

[6] Division of Cardiothoracic Radiology, University of California, San Diego, San Diego, Calif

[7] Department of Radiology, Mayo Clinic Rochester, Rochester, Minn

Received November 2, 2024; revision requested November 18; final revision received August 10, 2025; accepted August 18.

Address correspondence to: A.A.R.

Supplemental material: Supplemental material is available at *Radiology* online.

Funding: Authors declared no funding for this work.

Disclosures of conflicts of interest: Please see ICMJE form(s) for author conflicts of interest. These have been provided as supplemental materials.

References

1. Pałczyński K, Śmigiel S, Ledziński D, Bujnowski S. Study of the few-shot learning for ECG classification based on the PTB-XL dataset. Sensors (Basel) 2022;22(3):904.
2. Liu C, Wan Z, Ouyang C, et al. Zero-Shot ECG Classification with Multimodal Learning and Test-time Clinical Knowledge Enhancement. arXiv 2024. Preprint posted online March 11, 2024; doi:10.48550/2403.06659.
3. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. arXiv 2017. Preprint posted online June 12, 2017; https://doi.org/10.48550/arXiv.1706.03762.
4. Zhang S, Metaxas D. On the challenges and perspectives of foundation models for medical image analysis. Med Image Anal 2024;91:102996.
5. Yong G, Jeon K, Gil D, Lee G. Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. Comput Aided Civil Eng 2023;38(11):1536–1554.
6. Barragán-Montero A, Javaid U, Valdés G, et al. Artificial intelligence and machine learning for medical imaging: A technology review. Phys Med 2021;83:242–256.
7. Paschali M, Chen Z, Blankemeier L, et al. Foundation Models in Radiology: What, How, Why, and Why Not. Radiology 2025;314(2):e240597.
8. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models. arXiv 2021. Preprint posted online August 16, 2021; doi:10.48550/2108.07258.
9. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal 2019;54:280–296.
10. Patterson D, Gonzalez J, Le Q, et al. Carbon Emissions and Large Neural Network Training. arXiv 2021. Preprint posted online April 21, 2021; doi:10.48550/2104.10350.
11. Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and Efficient Foundation Language Models. arXiv 2023. Preprint posted online February 27, 2023; doi:10.48550/2302.13971.
12. Cardoso MJ, Li W, Brown R, et al. MONAI: An open-source framework for deep learning in healthcare. arXiv 2022. Preprint posted online November 4, 2022; doi:10.48550/2211.02701.
13. Devlin J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2018. Preprint posted online October 11, 2018; doi:10.48550/1810.04805.
14. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–1240.
15. Doshi R, Amin KS, Khosla P, et al. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. Radiology 2024;310(3):e231593.
16. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med 2023;6(1):210.
17. Venigalla A, Frankle J, Carbin M. BioMedLM: a Domain-Specific Large Language Model for Biomedical Text. MosaicML. https://medium.com/@MosaicML/pubmed-gpt-a-domain-specific-large-language-model-for-biomedical-text-567b18e2b11. Published January 27, 2023.
18. Wu C, Lin W, Zhang X, et al. PMC-LLaMA: toward building open-source language models for medicine. J Am Med Inform Assoc 2024;31(9):1833–1843.
19. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–180. [Published correction appears in Nature 2023;620(7973):E19.]
20. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv 2023. Preprint posted online July 18, 2023; doi:10.48550/2307.09288.
21. Christophe C, Gupta A, Hayat N, et al. Med42 - Clinical Large Language Model. https://huggingface.co/m42-health/med42-70b. Published 2023.
22. Han K, Wang Y, Chen H, et al. A Survey on Vision Transformer. IEEE Trans Pattern Anal Mach Intell 2023;45(1):87–110.
23. Ma J, He Y, Li F, et al. Segment anything in medical images. Nat Commun 2024;15(1):654.
24. Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey. ACM Comput Surv 2022;54(10s):1–41.
25. Kim JW, Khan AU, Banerjee I. Systematic Review of Hybrid Vision Transformer Architectures for Radiological Image Analysis. J Imaging Inform Med 2025;38(5):3248-3262.
26. Du D, Gong G, Chu X. Model Quantization and Hardware Acceleration for Vision Transformers: A Comprehensive Survey. arXiv 2024. Preprint posted online May 1, 2024; doi:10.48550/2405.00314.
27. Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention. In: Marina M, Tong Z, eds. Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research. PMLR, 2021; 10347–10357.
28. Murphy ZR, Venkatesh K, Sulam J, Yi PH. Visual Transformers and Convolutional Neural Networks for Disease Classification on Radiographs: A Comparison of Performance, Sample Efficiency, and Hidden Stratification. Radiol Artif Intell 2022;4(6):e220012.
29. Dao T, Fu D, Ermon S, Rudra A, Ré C. Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in Neural Information Processing Systems 2022;35:16344–16359.
30. Dao T. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. arXiv 2023. Preprint posted online July 17, 2023; doi:10.48550/2307.08691.
31. Cox J, Liu P, Stolte SE, et al. BrainSegFounder: Towards 3D foundation models for neuroimage segmentation. Med Image Anal 2024;97:103301.
32. Tavakoli N, Rahsepar AA, Benefield BC, et al. ScarNet: A Novel Foundation Model for Automated Myocardial Scar Quantification from Late Gadolinium-Enhancement Images. J Cardiovasc Magn Reson 2025:101945. doi: 10.1016/j.jocmr.2025.101945.
33. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing, 2015; 234–241.
34. Chen J, Lu Y, Yu Q, et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv 2021. Preprint posted online February 8, 2021; doi:10.48550/2102.04306.
35. Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2021; 10012–10022.
36. Kirillov A, Mintun E, Ravi N, et al. Segment Anything. Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2023; 4015–4026.
37. Ma J, Kim S, Li F, et al. Segment Anything in Medical Images and Videos: Benchmark and Deployment. arXiv 2024. Preprint posted online August 6, 2024; doi:10.48550/2408.03322.

38. Xiong X, Wang C, Li W, Mammo-Sam LG. Adapting Foundation Segment Anything Model for Automatic Breast Mass Segmentation in Whole Mammograms. In: Cao X, Xu X, Rekik I, Cui Z, Ouyang X, eds. Machine Learning in Medical Imaging. Cham: Springer Nature Switzerland, 2024; 176–185.
39. Zhu JY, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. Proceedings of the IEEE International Conference on Computer Vision. IEEE, 2017; 2223–2232.
40. Torbunov D, Huang Y, Yu H, et al. UVCGAN: UNet Vision Transformer cycle-consistent GAN for unpaired image-to-image translation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. IEEE, 2023; 702–712.
41. Guo P, Zhao C, Yang D, et al. MAISI: Medical AI for Synthetic Imaging. arXiv 2024. Preprint posted online September 13, 2024; doi:10.48550/2409.11169.
42. Radford A, Kim JW, Hallacy C, et al. Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning. PMLR, 2021; 8748–8763.
43. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. arXiv 2022. Preprint posted online October 18, 2022; doi:10.48550/2210.10163.
44. You K, Gu J, Ham J, et al. CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2023; 101–111.
45. Zhang K, Yu J, Yan Z, et al. BiomedGPT: A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks. arXiv 2023. Preprint posted online May 26, 2023; doi:10.48550/2305.17100.
46. Saab K, Tu T, Weng WH, et al. Capabilities of Gemini Models in Medicine. arXiv 2024. Preprint posted online April 29, 2024; doi:10.48550/2404.18416.
47. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D&3D Medical Data. arXiv 2023. Preprint posted online August 4, 2023; ; doi:10.48550/2308.02463.
48. Alayrac JB, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems 2022;35:23716–23736.
49. Awadalla A, Gao I, Gardner J, et al. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. arXiv 2023. Preprint posted online August 2, 2023; doi:10.48550/2308.01390.
50. Moor M, Huang Q, Wu S, et al. Med-Flamingo: a Multimodal Medical Few-shot Learner. Machine Learning for Health (ML4H). PMLR, 2023; 353–367.
51. Tanno R, Barrett DGT, Sellergren A, et al. Collaboration between clinicians and vision-language models in radiology report generation. Nat Med 2025;31(2):599–608.
52. Hamamci IE, Er S, Almas F, et al. A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities. arXiv 2024. Preprint posted online March 26, 2024; doi:10.48550/2403.17834.
53. Tiu E, Talius E, Patel P, et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng 2022;6(12):1399–1406.
54. Chen YP, Chen YY, Lin JJ, Huang CH, Lai F. Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT): Development and Performance Evaluation. JMIR Med Inform 2020;8(4):e17787.
55. Li Y, Rao S, Solares JRA, et al. BEHRT: Transformer for Electronic Health Records. Sci Rep 2020;10(1):7155.
56. Dubey A, Jauhri A, Pandey A, et al. The Llama 3 Herd of Models. arXiv 2024. Preprint posted online July 31, 2024; doi:10.48550/2407.21783.
57. Jaegle A, Borgeaud S, Alayrac JB, et al. Perceiver IO: A General Architecture for Structured Inputs & Outputs. arXiv 2021. Preprint posted online July 30, 2021; doi:10.48550/2107.14795.
58. Moritz M, Topol E, Rajpurkar P. Coordinated AI agents for advancing healthcare. Nat Biomed Eng 2025;9(4):432–438.
59. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. Patterns (N Y) 2021;2(10):100347.
60. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digit Med 2020;3:119.
61. Rahsepar AA, Tavakoli N, Kim GHJ, et al. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. Radiology 2023;307(5):e230922.
62. Sharma M, Tong M, Korbak T, et al. Towards Understanding Sycophancy in Language Models. arXiv 2023. Preprint posted online October 20, 2023; doi:10.48550/2310.13548.
63. Chua J, Li Y, Yang S, Wang C, Yao L. AI Safety in Generative AI Large Language Models: A Survey. arXiv 2024. Preprint posted online July 6, 2024; doi:10.48550/2407.18369.
64. Kure HI, Sarkar P, Ndanusa AB, Nwajana AO. Detecting and Preventing Data Poisoning Attacks on AI Models. arXiv 2025. Preprint posted online March 12, 2025; doi:10.48550/2503.09302.
65. Yeung K. 'Hypernudge': Big Data as a mode of regulation by design. Inform Commun Soc 2017;20(1):118–136.
66. Shojaee P, Mirzadeh I, Alizadeh K, et al. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. arXiv 2025. Preprint posted online June 7, 2025; doi:10.48550/2506.06941.
67. Liu H, Ding N, Li X, et al. Artificial Intelligence and Radiologist Burnout. JAMA Netw Open 2024;7(11):e2448714.
68. Sutton RT, Pincock D, Baumgart DC, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17.
69. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. Nat Med 2020;26(8):1229–1234.
70. Rajpurkar P, O'Connell C, Schechter A, et al. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. NPJ Digit Med 2020;3:115.
71. Homayounieh F, Digumarthy S, Ebrahimian S, et al. An Artificial Intelligence-Based Chest X-ray Model on Human Nodule Detection Accuracy From a Multicenter Study. JAMA Netw Open 2021;4(12):e2141096.
72. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021;4(1):31.
73. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121–127.
74. Harvey HB, Gowda V. Regulatory Issues and Challenges to Artificial Intelligence Adoption. Radiol Clin North Am 2021;59(6):1075–1083.
75. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021;3(11):e745–e750.
76. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44–56.
77. Suara S, Jha A, Sinha P, Sekh AA. Is Grad-CAM Explainable in Medical Images? International Conference on Computer Vision and Image Processing. Springer, 2023; 124–135.
78. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1(5):206–215.
79. Borys K, Schmitt YA, Nauta M, et al. Explainable AI in medical imaging: An overview for clinical practitioners - Beyond saliency-based XAI approaches. Eur J Radiol 2023;162:110786.
80. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
81. Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H. Synthetic data augmentation using GAN for improved liver lesion classification. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, 2018; 289–293.
82. Antoniou A. Data Augmentation Generative Adversarial Networks. arXiv 2017. Preprint posted online November 12, 2017; doi:10.48550/1711.04340.
83. Hanneman K, Playford D, Dey D, et al; American Heart Association Council on Cardiovascular Radiology and Intervention; and Council on Lifelong Congenital Heart Disease and Heart Health in the Young. Value Creation Through Artificial Intelligence and Cardiovascular Imaging: A Scientific Statement From the American Heart Association. Circulation 2024;149(6):e296–e311.
84. Doo FX, Vosshenrich J, Cook TS, et al. Environmental Sustainability and AI in Radiology: A Double-Edged Sword. Radiology 2024;310(2):e232030.
85. Hu EJ, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models. arXiv 2021. Preprint posted online June 17, 2021; doi:10.48550/2106.09685.
86. Harvey HB, Gowda V. How the FDA Regulates AI. Acad Radiol 2020;27(1):58–61.
87. Aidoc. Aidoc Secures Landmark FDA Clearance for First Foundation Model-Powered Clinical AI Solution of Its Kind. https://www.aidoc.com/about/news/aidoc-secures-landmark-fda-clearance/. Published 2025. Accessed August 10, 2025.
88. U.S. Food and Drug Administration. Executive Summary for the Digital Heath Advisory Committee Meeting: Total Product Lifecycle Considerations for Generative AI-Enabled Devices. https://www.fda.gov/media/182871/download. Published 2024.
89. U.S. Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence. Published December 2024. Accessed August 10, 2025.

90. U.S. Food and Drug Administration. Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing. Published January 2025. Accessed August 10, 2025.
91. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. Front Med (Lausanne) 2023;10:1305756.
92. Duffourc MN, Giovanniello DS. The Autonomous AI Physician: Medical Ethics and Legal Liability. In: Sousa Antunes H, Freitas PM, Oliveira AL, Martins Pereira C, Vaz de Sequeira E, Barreto Xavier L, eds. Multidisciplinary Perspectives on Artificial Intelligence and the Law. Cham: Springer International Publishing, 2024; 207–228.
93. Harned Z, Lungren MP, Rajpurkar P. Machine Vision, Medical AI, and Malpractice. Zach Harned, Matthew P. Lungren & Pranav Rajpurkar, Comment, Machine Vision, Medical AI, and Malpractice, Harv. J.L. & Tech Dig. (2019). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3442249. Published August 28, 2019.
94. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616(7956):259–265.
95. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems 2020;33:9459–9474.
96. Akbari H, Yuan L, Qian R, et al. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems 2021;34:24206–24221.
97. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203–211.
98. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Trans Neural Netw Learning Syst 2021;32(11):4793–4813.
99. Barbano CA, Brunello M, Dufumier B, Grangetto M. Anatomical Foundation Models for Brain MRIs. arXiv 2024. Preprint posted online August 7, 2024; doi:10.48550/2408.07079.
100. Zhang H, Yang J, Wan S, Fua P. LeFusion: Controllable Pathology Synthesis via Lesion-Focused Diffusion Models. arXiv 2024. Preprint posted online March 21, 2024; doi:10.48550/2403.14066.
101. Christensen M, Vukadinovic M, Yuan N, Ouyang D. Vision–language foundation model for echocardiogram interpretation. Nat Med 2024;30(5):1481–1488.
102. Jacob AJ, Borgohain I, Chitiboi T, et al. Towards a vision foundation model for comprehensive assessment of Cardiac MRI. arXiv 2024. Preprint posted online October 2, 2024; doi:10.48550/2410.01665.
103. Chen T, Zhu L, Ding C, et al. SAM Fails to Segment Anything? -- SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, Medical Image Segmentation, and More. arXiv 2023. Preprint posted online April 18, 2023; doi:10.48550/2304.09148.
104. Ghosh S, Poynton CB, Visweswaran S, Batmanghelich K. Mammo-CLIP: A Vision Language Foundation Model to Enhance Data Efficiency and Robustness in Mammography. International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer, 2024; 632–642.
105. Alsentzer E, Murphy JR, Boag W, et al. Publicly Available Clinical BERT Embeddings. arXiv 2019. Preprint posted online April 6, 2019; doi:10.48550/1904.03323.
106. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) 2021;3(1):1–23.
107. Yasunaga M, Leskovec J, Liang P. LinkBERT: Pretraining Language Models with Document Links. arXiv 2022. Preprint posted online March 29, 2022; doi:10.48550/2203.15827.
108. Yasunaga M, Bosselut A, Ren H, et al. Deep bidirectional language-knowledge graph pretraining. Advances in Neural Information Processing Systems 2022;35:37309–37323.
109. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. Nature 2023;619(7969):357–362.
110. Chen Z, Cano AH, Romanou A, et al. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. arXiv 2023. Preprint posted online November 27, 2023; doi:10.48550/2311.16079.
111. Xie Q, Chen Q, Chen A, et al. Me-LLaMA: Foundation Large Language Models for Medical Applications. Res Sq 2024:rs.3.rs-r4240043.
112. Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning Robust Visual Features without Supervision. arXiv 2023. Preprint posted online April 14, 2023; doi:10.48550/2304.07193.
113. Vorontsov E, Bozkurt A, Casson A, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. Nat Med 2024;30(10):2924–2935.
114. Codella NC, Jin Y, Jain S, et al. MedImageInsight: An Open-Source Embedding Model for General Domain Medical Imaging. arXiv 2024. Preprint posted online October 9, 2024; doi:10.48550/2410.06542.
115. Pérez-García F, Sharma H, Bond-Taylor S, et al. RAD-DINO: Exploring Scalable Medical Image Encoders Beyond Text Supervision. arXiv 2024. Preprint posted online January 19, 2024; doi:10.48550/2401.10815.v1.
116. Moutakanni T, Bojanowski P, Chassagnon G, et al. Advancing human-centric AI for robust X-ray analysis through holistic self-supervised learning. arXiv 2024. Preprint posted online May 2, 2024; doi:10.48550/2405.01469.
117. Qi Q, Lin L, Zhang R, Xue C. MEDT: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis. IEEE Access 2022;10:28750–28759.
118. Hatamizadeh A, Tang Y, Nath V, et al. UNETR: Transformers for 3D Medical Image Segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. IEEE, 2022; 574–584.
119. Ma J, Li F, Wang B. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. arXiv 2024. Preprint posted online January 9, 2024; doi:10.48550/2401.04722.
120. Machado L, Philippe H, Ferreres É, et al. ONCOPILOT: A Promptable CT Foundation Model For Solid Tumor Evaluation. arXiv 2024. Preprint posted online October 10, 2024; doi:10.48550/2410.07908.
121. He Y, Guo P, Tang Y, et al. VISTA3D: Versatile Imaging SegmenTation and Annotation model for 3D Computed Tomography. arXiv 2024. Preprint posted online June 7, 2024; doi:10.48550/2406.05285.v1.
122. Zhang S, Xu Y, Usuyama N, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv 2023. Preprint posted online March 2, 2023; doi:10.48550/2303.00915.
123. Lin W, Zhao Z, Zhang X, et al. PMC-CLIP: Contrastive Language-Image Pre-training Using Biomedical Documents. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2023; 525–536.
124. Zhang X, Wu C, Zhao Z, et al. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. arXiv 2023. Preprint posted online May 17, 2023; doi:10.48550/2305.10415.
125. Lu J, Batra D, Parikh D, Lee S. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems 2019;33:13–23.
126. Li C, Wong C, Zhang S, et al. LLaVA-med: training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems 2024;37:28541–28564.
127. Reed S, Zolna K, Parisotto E, et al. A Generalist Agent. arXiv 2022. Preprint posted online May 12, 2022; doi:10.48550/2205.06175.
128. Zellers R, Lu X, Hessel J, et al. MERIOT: multimodal neural script knowledge models. Advances in Neural Information Processing Systems 2021;34:23634–23651.
129. Ma Y, Xu G, Sun X, et al. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. Proceedings of the 30th ACM International Conference on Multimedia 2022;30:638–647.
130. Sun C, Myers A, Vondrick C, Murphy K, Schmid C. VideoBERT: A Joint Model for Video and Language Representation Learning. Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2019; 7464–7473.
131. Yuan L, Chen D, Chen YL, et al. Florence: A New Foundation Model for Computer Vision. arXiv 2021. Preprint posted online November 22, 2021; doi:10.48550/2111.11432.
132. Wang X, Zhang X, Zhu Y, et al. PANDA: A Gigapixel-Level Human-Centric Video Dataset. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020; 3268–3278.

Radiology

COMMUNICATIONS • ERRATUM

Erratum for: Generative AI and Foundation Models in Radiology: Applications, Opportunities, and Potential Challenges

Originally published in:
https://doi.org/10.1148/radiol.242961

Generative AI and Foundation Models in Radiology: Applications, Opportunities, and Potential Challenges

Neda Tavakoli, Zahra Shakeri, Vrushab Gowda, Konrad Samsel, Arash Bedayat, Ahmadreza Ghasemiesfe, Ulas Bagci, Albert Hsiao, Tim Leiner, James Carr, Daniel Kim, Amir Ali Rahsepar

Erratum in:
https://doi.org/10.1148/radiol.259019

A misplaced citation (typographical error) was removed from the first paragraph of the article.

In Table 2, **University of Oxford** was added to the multi-institutional collaboration listed as the developer for VISTA3D.