From large language models to multimodal AI: A scoping review on the potential of generative AI in medicine

Lukas Buess[1*], Matthias Keicher[2], Nassir Navab[2], Andreas Maier[1], Soroosh Tayebi Arasteh[1]

[1]Pattern Recognition Lab, Friedrich-Alexander-Universit¨at Erlangen-Nu¨rnberg, Erlangen, Germany.

[2]Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany.

*Corresponding author E-mail: lukas.buess@fau.de

Abstract

Generative artificial intelligence (AI) models, such as diffusion models and OpenAI's ChatGPT, are transforming medicine by enhancing diagnostic accuracy and automating clinical workflows. The field has advanced rapidly, evolving from textonly large language models for tasks such as clinical documentation and decision support to multimodal AI systems capable of integrating diverse data modalities, including imaging, text, and structured data, within a single model. The diverse landscape of these technologies, along with rising interest, highlights the need for a comprehensive review of their applications and potential. This scoping review explores the evolution of multimodal AI, highlighting its methods, applications, datasets, and evaluation in clinical settings. Adhering to PRISMA-ScR guidelines, we systematically queried PubMed, IEEE Xplore, and Web of Science, prioritizing recent studies published up to the end of 2024. After rigorous screening, 144 papers were included, revealing key trends and challenges in this dynamic field. Our findings underscore a shift from unimodal to multimodal approaches, driving innovations in diagnostic support, medical report generation, drug discovery, and conversational AI. However, critical challenges remain, including the integration of heterogeneous data types, improving model interpretability, addressing ethical concerns, and validating AI systems in real-world clinical settings. This review summarizes the current state of the art, identifies critical gaps, and provides insights to guide the development of scalable, trustworthy, and clinically impactful multimodal AI solutions in healthcare.

Keywords: Large language models, Generative AI, Multimodal AI, Scoping review

1 Introduction

Generative artificial intelligence (AI), exemplified by models like ChatGPT, has drawn widespread attention for its ability to process and generate human-like text, substantially advancing various domains. In healthcare, these models have rapidly

transformed traditional approaches by offering capabilities beyond conventional data analysis [1, 2]. For instance, large language models (LLMs) have been applied in tasks such as summarizing medical records [3], assisting in diagnostic reasoning [4], and conducting bioinformatics research [5]. These advancements highlight the ability of LLMs to process and interpret complex clinical language, improving efficiency and accuracy across tasks such as radiology reporting. Recent studies further demonstrate their impact, showing that AI-generated draft radiology reports can reduce reporting time by about 25% while maintaining diagnostic accuracy [6], thus addressing workload challenges in clinical practice [7].

However, healthcare data extends far beyond clinical texts, encompassing diverse modalities such as medical images [8, 9], laboratory results [10, 11], and genomic data [12]. To address this diversity, multimodal AI systems have emerged, integrating these data types within a single model. This integration paves the way for comprehensive decision support systems that more closely mimic human clinical reasoning. Recent advancements in multimodal AI represent a significant shift, expanding generative AI applications beyond language-focused tasks to more complex data integration scenarios [13–15]. By unifying text, images, and other clinical data, these systems hold potential for improved diagnostic accuracy and broader applications, from predictive analytics to complex interventional support [16].

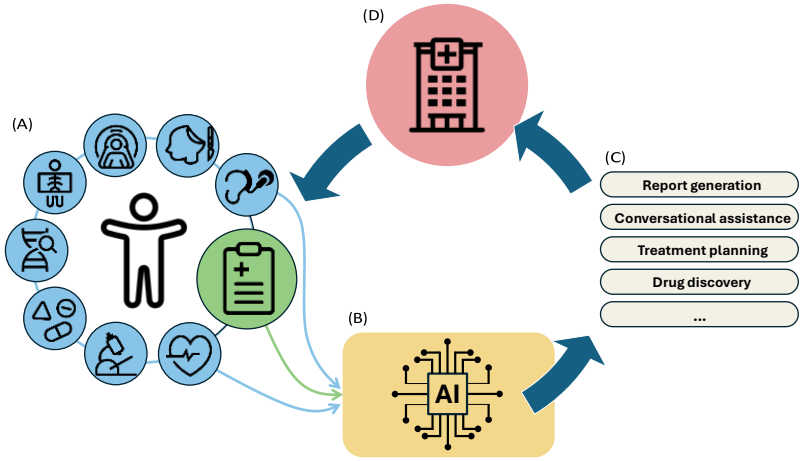

Fig. 1: Multimodal AI pipeline in healthcare: (A) Diverse medical data modalities (e.g., images, genomics, and clinical notes) are collected and processed, (B) transformed into unified representations by AI models, (C) used to generate insights such as reports, conversational assistance, and treatment plans, and (D) refined through iterative feedback to continuously optimize data collection and AI performance.

Several recent review articles have provided valuable overviews of multimodal AI and LLMs. Comprehensive surveys of multimodal large language models (MLLMs) in the broader computer vision domain were presented by Yin et al. [17] and Wang et al. [18], highlighting recent advancements, providing a summary of architectural developments, and identifying key trends in model evolution. A broader perspective on multimodal

approaches in healthcare was provided by Kline et al. [19] and Acosta et al. [1]. He et al. [20] present a comprehensive collection of foundation models, spanning from image-only architectures to advanced multimodal models.

While previous reviews provide essential insights, the dynamic and rapidly evolving nature of this field necessitates an up-to-date and focused exploration of recent developments in LLM-based multimodal AI for medicine. This review aims to fill this gap by providing a comprehensive overview of the evolution from text-only LLMs to multimodal AI systems in medicine, with a particular emphasis on recent advancements. Unlike prior reviews, we also discuss evaluation methods specifically tailored to the challenges and requirements of medical generative AI, ensuring real-world clinical utility and reliability.

To guide this review, we formulated the following research questions: • What methods are commonly used in the development of generative AI for healthcare applications?

- What datasets support the development of generative AI in medical contexts?
- Which evaluation metrics are employed to assess the utility of generative AI models in medical contexts?

In the following sections, we first outline the methodology employed for literature collection and selection, detailing the search strategy, inclusion criteria, and data extraction processes used to ensure a comprehensive review. We then present our findings, emphasizing the shift from text-only LLMs to multimodal AI systems in medicine, with a particular focus on their applications, datasets, model architectures, and evaluation metrics. Our results reveal a significant shift towards multimodal models, which are driving innovation across various areas of healthcare. However, persistent challenges remain, particularly in the evaluation of these models, including the assessment of their reliability, clinical relevance, and generalizability. Finally, we provide an outlook on the future of generative AI in medicine, offering insights to guide further research and development in this rapidly evolving field.

2 Methods

Our scoping review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) [21, 22], which provides a standardized framework for methodological transparency in scoping reviews. This section details the data collection methods used in our review. The complete PRISMAScR checklist is available in Supplementary Table S.1.

2.1 Eligibility criteria

We included studies published between January 2020 and December 2024 to capture recent advancements in the rapidly advancing field of generative AI in medicine. Only original research in English was eligible, as our focus is on primary contributions rather than synthesized findings. Review and meta-analysis papers were therefore excluded.

We included peer-reviewed conference and journal publications, alongside manually selected preprints with high relevance and potential impact. To ensure a comprehensive overview, foundational dataset papers published before 2020 were also included when they were widely used in the selected studies or remained relevant for benchmarking. This approach ensured a focus on current, state-of-the-art developments in multimodal AI applications in medicine.

2.2 Information sources

We performed a systematic search in PubMed, IEEE Xplore, and Web of Science, employing a standardized set of keywords derived from our research objectives. Full search queries are detailed in Supplementary Table S.2. The searches, conducted on October 1, 2024, were imported into Rayyan [23], a web-based tool designed to facilitate literature screening and semi-automated duplicate removal.

2.3 Search strategy

The literature search consisted of a systematic database search structured into two subsearches to capture the development and application of text-only LLMs and multimodal models in medicine. The first subsearch targeted text-only LLMs using the keyword groups ”medical” and ”language model”. The second subsearch focused on multimodal models, using three groups of keywords: ”medical”, ”language model”, and ”multimodal”. The full search queries, including the specific combinations used, are provided in Supplementary Table S.2. Additionally, a manual search was performed to identify recent preprints, datasets, and other resources not captured by the database search, which continued through the end of 2024 to ensure the inclusion of the most current and impactful studies.

2.4 Inclusion and exclusion criteria

The selection process began with structured database queries, followed by duplicate removal, title and abstract screening, and subsequent full-text reviews for potentially relevant papers. We excluded articles that were non-medical or lacked methodological novelty. To ensure balanced representation across application areas, we aimed for proportional inclusion from prevalent fields, such as X-ray report generation.

2.5 Synthesis of results

The selected papers were categorized through a two-step process. First, they were grouped by topics, including text-only LLMs, multimodal models, datasets, and evaluation metrics. Within each topic, papers were further categorized based on their application areas. This dual-layer categorization provides a structured overview of developments in generative AI for medicine, illustrating the progression from textonly LLMs to multimodal models. Key publications are summarized through narrative descriptions and tables, offering insights into methodological approaches, application

domains, datasets, and evaluation frameworks to provide a comprehensive understanding of current trends and challenges. Tables 1 (text-only LLMs), 2 (text-only datasets), 3 (contrastive learning methods), 4 (MLLMs), 5 (multimodal datasets), and 6 (evaluation metrics) summarize the results.

3 Included studies

A total of 4,384 papers were retrieved from three databases. After removing duplicates, 2,656 articles were excluded during the initial screening based on their titles and abstracts, following the predefined inclusion and exclusion criteria. The remaining articles underwent a full-text review, during which both relevance and topic diversity were considered to avoid overrepresentation of similar studies. This step led to the exclusion of an additional 249 papers. Ultimately, 60 papers from the database search were included in the review. Additionally, 83 papers were identified through manual searches to capture the most current and relevant studies not covered in the database queries. Figure 2 provides an overview of the full screening process. In total, 144 papers were included in this review.

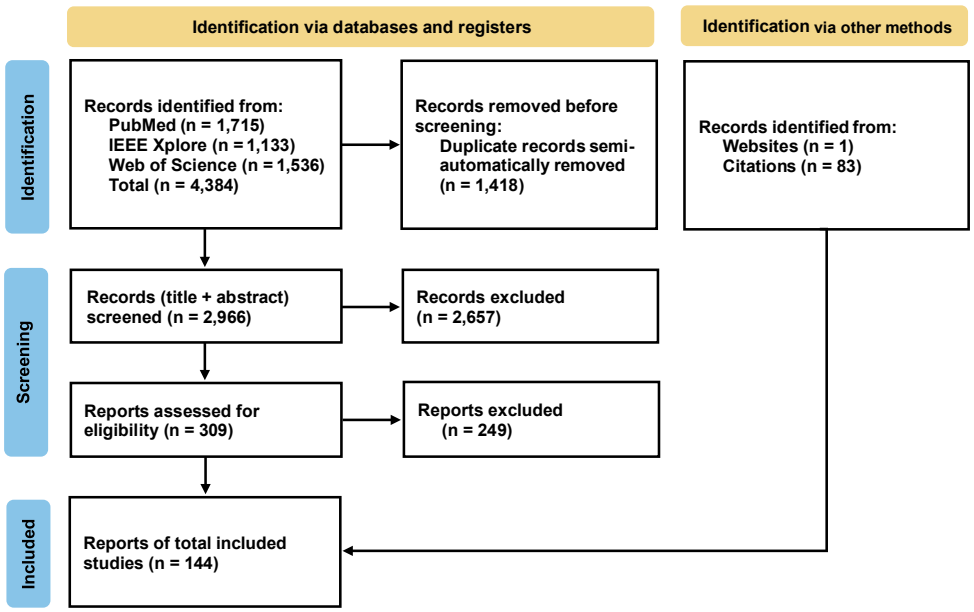

Fig. 2: PRISMA flow diagram illustrating the study selection process for the scoping review. The diagram shows the number of records identified through database searches and manual searches, the removal of duplicates, the screening of titles and abstracts, the review of full-text articles, and the final inclusion of studies in the review.

4 Language models in medicine

Mono-modal LLMs, which process textual data exclusively, have laid the foundation for the development of multimodal systems, demonstrating remarkable capabilities in understanding and generating human-like text. In the medical domain, LLMs demonstrated high effectiveness in processing and analyzing complex clinical data, enabling advancements in applications such as clinical documentation, medical literature summarization, and diagnostic support [3, 24]. Their success is based on the transformer architecture, introduced in the landmark paper "Attention Is All You Need" [25], which employs self-attention mechanisms to effectively capture contextual relationships and long-range dependencies in text. This architecture has enabled LLMs to scale effectively, making them capable of processing medical texts.

4.1 LLM methods

LLMs tailored to medical applications (Table 1) leverage various approaches to adapt general-purpose models for specialized medical tasks. A prevalent method is supervised finetuning (SFT), where general LLMs are finetuned on domain-specific datasets, such as biomedical literature and clinical notes, to enhance their understanding of medical concepts and vocabulary. This approach has been instrumental in models like BioBERT and BioMistral, which adapt general-purpose language models for biomedical applications [26, 27].

In contrast to SFT, prompt engineering techniques have emerged as a lightweight alternative for guiding pretrained models without additional training, relying on carefully designed input prompts to achieve strong task performance in medical text understanding and generation [28].

Advanced alignment techniques such as reinforcement learning from human feedback (RLHF) have been developed to further refine the outputs of LLMs for medical applications. RLHF leverages reward models trained on expert feedback to align model responses with clinical expectations. However, due to the cost of obtaining expert feedback in the healthcare domain, reinforcement learning from AI feedback (RLAIF) has emerged as an alternative [29]. This technique replaces human feedback with evaluations from auxiliary AI models, reducing reliance on scarce human resources while maintaining alignment capabilities. A notable example is HuatuoGPT [30], which uses RLAIF for clinical alignment.

Another recent development in model adaption is chain-of-thought (CoT) prompting, a technique where models generate intermediate reasoning steps before producing a final answer. By breaking down complex tasks into substeps, CoT enhances model explainability and task performance, which is especially valuable in the medical domain as it not only improves accuracy but also increases trust in the model's reasoning process. For example, HuatuoGPT-o1 [31] applies CoT prompting to improve medical response clarity and ensure step-by-step diagnostic reasoning.

An additional adaptation technique is retrieval augmented generation (RAG) [32], which equips LLMs with mechanisms to query external knowledge bases during

inference. This approach enables models to access up-to-date information, such as medical guidelines or recent research findings, without requiring retraining. For instance, Almanac [33], ChatDoctor [4], and RadioRAG [34] combine generative capabilities with retrieval systems. However, maintaining the retrieval database and ensuring its comprehensiveness pose ongoing challenges [4, 35].

4.2 LLM applications

LLMs have revolutionized various applications in biomedical language processing, demonstrating utility across a range of tasks. In named entity recognition (NER), they

Study	Downstream task
Clinical text	
Almanac [33]	QA
BioALBERT [36]	NER
BioBERT [26]	NER, QA
BioGPT [37]	Classification, QA
BioMistral [27]	QA
ChatDoctor [4]	Dialogue
ChestXRayBERT [3]	Summarization
DRG-LLaMA [38]	Classification
GatorTron [39]	QA
HuatuoGPT [30]	Dialogue
HuatuoGPT-o1 [30]	Dialogue
Johnson et al. [40]	Deidentification
Kresevic et al. [41]	Summarization
Mahendran and McInnes [42]	NER
MAPLEZ [43]	Classification
Med-BERT [44]	NER
MedAlpaca [45]	QA
MEDITRON-70B [46]	QA
MED-PaLM [47]	QA
MMed-Llama 3 [48]	QA
Mu et al. [49]	Classification
NYUTron [24]	Clinical outcome prediction
PMC-LLaMA [50]	QA
PodGPT [51]	QA
RadBERT [52]	Classification, Summarization
Schmidt et al. [53]	Error detection
Bioinformatics	
AlphaFold [5]	Structure prediction
BioPhi [54]	Antibody design
CADD v1.7 [55]	Scoring
DNABERT [56]	Structure analysis
Geneformer [57]	Classification
Hie et al. [58]	Antibody design
MSA Transformer [59]	Structure analysis
ProGen [60]	Structure prediction
ProtGPT2 [61]	Protein design
ProtTrans [62]	Structure analysis

scBERT [63]	Classification
ToxinPred 3.0 [64]	Classification

Table 1: Summary of LLM methods, categorized by their application to clinical text and bioinformatics tasks. The table includes method names and target applications. (Abbreviations: NER - named entity recognition, QA - question answering)

enable the extraction of critical medical entities, such as diseases, drugs, and symptoms from unstructured text. This capability supports clinical data annotation, which is crucial for automated clinical decision support systems [36].

Dialogue systems represent another application of LLMs in medicine. Models like ChatDoctor [4] and HuatuoGPT [30] facilitate patient interactions, simulate doctorpatient consultations, and assist in providing medical information and guidance. These systems aim to reduce barriers to medical access by providing instant responses.

In summarization tasks, medical LLMs condense lengthy electronic health records (EHRs) into concise summaries. This application significantly reduces the documentation burden on healthcare providers and aids decision-making by presenting critical patient information in a structured format [3, 65].

Deidentification and privacy-preserving applications are critical areas where LLMs contribute to medical data management by safeguarding patient confidentiality in sensitive clinical texts. LLMs can automate the removal of protected health information from medical documents by anonymizing identifiers such as names and dates while preserving data utility [40, 43].

Text classification tasks also benefit from LLM advancements, with applications such as predicting patient outcomes and categorizing medical literature [24].

In bioinformatics, LLMs have expanded beyond language processing to analyze biological sequences like DNA, RNA, and proteins. Models such as DNABERT [56] have advanced gene annotation, while AlphaFold [5] has achieved groundbreaking success in protein structure prediction.

4.3 LLM datasets

The development of medical LLMs relies on diverse and specialized datasets that capture the complexity of medical language, context, and tasks. These datasets fall into categories such as clinical text, domain-specific literature, conversational data, and bioinformatics resources, each serving distinct purposes in the development of medical LLMs. These datasets enable general-purpose LLMs to align with the medical domain, which is critical for achieving reliable and accurate outputs in clinical settings. Clinical text datasets play a central role in training medical LLMs (see Table 2). For instance, EHR datasets like MIMIC-IV [66] provide a rich source of structured and unstructured clinical data, commonly used for tasks such as summarization and NER, which are both essential for automating documentation and decision-making processes in healthcare. The eICU-

CRD dataset [67], another EHR resource, focuses on intensive care unit patient data, further broadening the scope of potential applications.

To introduce domain-specific knowledge into LLMs, datasets like GAP-Replay [46] and MedC-K [50], composed of biomedical literature and textbooks, are essential. These datasets are designed to equip models with the specialized terminology and reasoning patterns found in biomedical research and education.

For conversational AI in medicine, dialogue datasets are crucial. MedDialog [68] provides examples of doctor-patient interactions, enabling LLMs to learn medical dialogues, including patient concerns, physician responses, and diagnostic reasoning. These datasets are essential for developing medical conversational assistance systems capable of simulating clinical dialogues and supporting in patient education, diagnostic reasoning, and post-treatment follow-ups.

Bioinformatics datasets extend the scope of LLM applications beyond clinical text, supporting tasks in genomics and molecular biology. Resources like AlphaFold DB [5] and UniProtKB [69] provide structured data for protein structure and sequence analysis, making them valuable for drug discovery and molecular research. Similarly, genomic datasets such as GENCODE [12] and GenBank [70] offer data for tasks like gene prediction, helping models to better understand complex biological patterns.

Dataset	Size	Application
Clinical text		
eICU-CRD [67]	200K instances	EHR
GAP-Replay [46]	48.1B tokens	Literature
MedDialog-EN [68]	250K dialogues	Dialogue
MedC-K [50]	4.8M papers, 30K textbooks	Literature
MedC-I [50]	202M tokens	Dialogue, QA
Medical Meadow [45]	160K instances	QA
MIMIC-IV [66]	299K patients	EHR
MMedC [48]	25.5B tokens	Multilingual literature
MultiMedQA [47]	213K instances	QA
Bioinformatics		
AlphaFold DB [5]	200M entries	Protein Design
CPTAC Data Portal [71]	NA	Genomics, Protein Design
GenBank [70]	NA sequences	Genomics
GENCODE [12]	NA	Genomics
UniProtKB [69]	227M sequences	Protein Design

Table 2: Summary of datasets used for training medical LLMs, categorized into clinical text and bioinformatics data. The table includes dataset names, sizes, and primary application areas. (Abbreviations: NA - not available, NER - named entity recognition, QA - question answering, EHR - electronic health record)

5 Multimodal language models in medicine

By showcasing the potential of LLMs in processing clinical text, these models have established a strong foundation for integrating additional modalities, leading to the development of multimodal language models specifically designed for healthcare. Multimodal models combine diverse data types, such as text and medical images, to tackle complex medical tasks, including report generation [72, 73], image-text retrieval [74, 75], and medical consultation [14]. By building on advancements in LLMs, multimodal language models improve the integration and contextual understanding of multimodal medical data. This section provides an overview of recent architectures and methods addressing the unique challenges posed by multimodal medical data.

5.1 Architectures

Before presenting the literature, we briefly outline the two primary architectures in multimodal AI, i.e., the contrastive language-image pretraining (CLIP) and MLLMs (see Figure 4). These architectures serve as foundational frameworks for integrating multiple data modalities in medical AI. Although CLIP [76] is not inherently generative, its ability to align images and text within a shared embedding space makes it a crucial component in multimodal AI systems.

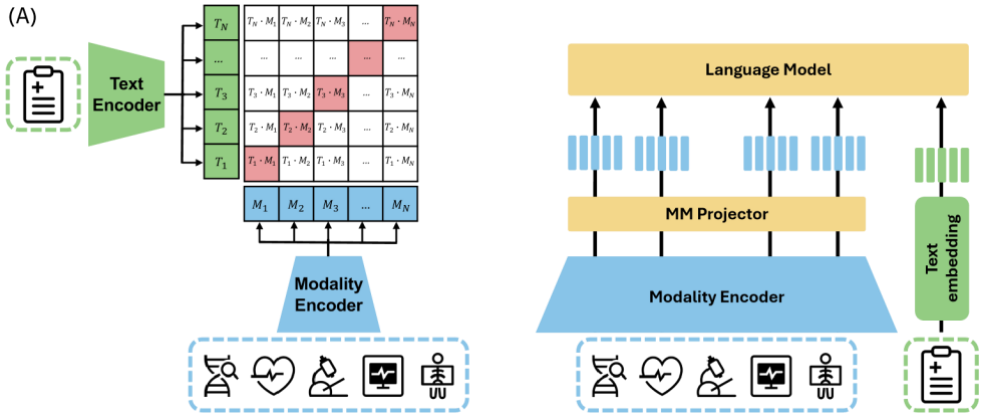

Fig. 3: Multimodal architectures: (A) CLIP-based models, which align embeddings of different modalities in a shared latent space, and (B) LLM-based models, which directly integrate different modality inputs through feature extraction and projection into LLM's ebbedding space.

CLIP [76] is designed to align different modalities, such as image and text, in a shared embedding space. Although originally developed for image-text pairs, its framework can be extended to other modalities, making it a versatile tool for various multimodal

learning tasks. By jointly training on paired modalities data, it excels in tasks like zero-shot image classification [74, 77], where new classes can be recognized without additional training. This makes CLIP particularly useful for situations where annotated medical data is limited.

On the other hand, MLLMs, such as LLaVA [78], extend the capabilities of LLMs by integrating non-textual data directly into their embeddings. This integration allows for a more holistic understanding of complex datasets, combining linguistic context with multimodal features like images or clinical measurements. These models excel in tasks such as radiology report generation [72, 73], question answering (QA) about medical images [79, 80], and decision support in diagnosis [13, 77, 81].

By leveraging complementary strengths, these architectures address the diverse challenges posed by multimodal medical data. CLIP is effective for aligning different data modalities, while MLLMs excel in diagnostic reasoning, together forming a powerful combination for improving multimodal AI in medicine.

5.2 Multimodal LLM methods

Modality alignment is a fundamental step for most MLLMs. Many approaches leverage CLIP-based methods (Table 3), which primarily focus on learning a shared latent space where modalities can be jointly represented for downstream tasks.

For instance, BiomedCLIP [82] uses contrastive learning to align medical images with paired reports, achieving state-of-the-art results in retrieval tasks. Building on this framework, CheXzero [74] adapts CLIP for zero-shot classification of X-ray images, while CT-CLIP [14] extends this approach to computed tomography (CT) scans. Similarly, UniMed-CLIP [103] enhances this paradigm by using classification datasets augmented by LLM-generated captions to train a foundation model capable of handling various medical image modalities.

More recent efforts have focused on large-scale pretrained models developed by industry leaders, aiming to generalize across diverse medical imaging tasks. Models like CT Foundation [87] and MedImageInsight [96], accessible via application programming interfaces (APIs), exemplify this trend by offering robust pretrained embeddings that address data scarcity in medical imaging and support downstream applications.

While many CLIP-based methods focus on aligning text with medical images, recent approaches have extended this to other modalities. For example, ETP [88] aligns electrocardiogram (ECG) signals [105, 106] with clinical reports, while MolLM [100] pairs chemical structures with textual descriptions to support drug discovery.

Study	Modalities	Application
BiomedCLIP [82]	Medical images, Descriptions	Classification, Retrieval, Visual QA
BioViL [83]	X-ray, Reports	Classification, Grounding
BioViL-T [84]	X-ray, Reports	Classification, Grounding, Reporting
CheXzero [74]	X-ray, Reports	Classification, Retrieval
ConVIRT [85]	X-ray, Reports	Classification, Retrieval
CPLIP [86]	Histopathology images, Descriptions	Classification
CT-CLIP [14]	CT, Reports, Labels	Classification, Retrieval
CT Foundation [87]	CT, Reports	Classification, Retrieval
CXR-RePaiR [75]	X-ray, Reports	Reporting
ETP [88]	ECG, Reports	Classification

FairCLIP [89]	SLO fundus images, Clinical notes	Classification
FiVE [90]	Histopathology images, Descriptions	Classification
FlexR [91]	X-ray, Reports	Classification
GLoRIA [92]	X-ray, Reports	Classification, Retrieval, Segmentation
KAD [93]	X-ray, Reports	Classification
MaCo [94]	X-ray, Reports	Classification
MCPL [95]	X-ray, Reports	Classification
MedImageInsight [96]	Medical images, Descriptions	Classification, Retrieval, Reporting
Med-MLLM [97]	CT, X-ray, Descriptions	Classification, Reporting
Merlin [77]	CT, EHR, Reports	Classification, Retrieval, Reporting, Segmentation
MedViLL [98]	X-ray, Reports	Classification, Retrieval, Reporting, Visual QA
MoleculeSTM [99]	Molecule structure, Descriptions	Retrieval
MolLM [100]	Molecule structures, Descriptions	Retrieval, Molecule description
PLIP [101]	Histopathology images, Descriptions	Classification, Retrieval
Prov-GigaPath [102]	Histopathology images, Reports	Classification
UniMed-CLIP [103]	Medical images, Captions	Classification
Xplainer [104]	X-ray, Reports	Classification

Table 3: Summary of multimodal CLIP-based methods. The table includes method names, the modalities utilized (e.g., text and medical images), and the primary tasks addressed, such as image-text retrieval, report generation, and disease classification. (Abbreviations: QA - question answering)

Study	Modalities	Downstream task
Alsharid et al. [107]	US video, Transcriptions, Gaze data	Captioning
AutoRG-Brain [108]	MRI, Reports, Masks	Reporting, Grounding
BiomedGPT [13]	Medical images, Literature, EHR	Reporting, Summarization, Visual QA
BioMed-VITAL [109]	Medical images, Instructions	Visual QA
ChatCAD [110]	X-ray, Reports	Reporting
CheXagent [111]	X-ray, Reports	Classification, Reporting, Grounding
COMG [112]	X-ray, Reports, Masks	Reporting
CT-CHAT [14]	CT, Reports	Reporting, Visual QA
FFA-GPT [113]	Fundus fluorescein angiography, Reports	Reporting, Visual QA
GenerateCT [114]	CT, Reports	Image generation
Huh et al. [115]	X-ray, Reports	Reporting
LLaVA-Med [15]	Medical images, Captions	Visual QA
LViT [116]	CT, X-ray, Masks, Text annotations	Segmentation
M3D-LaMed [117]	CT, Reports, Masks	Reporting, Visual QA, Segmentation
MAIRA-2 [72]	X-ray, Reports, Masks	Reporting, Grounding
MAIRA-Seg [118]	X-ray, Reports, Masks	Reporting
Med-Flamingo [119]	Medical images, Captions	Visual QA
Med-PaLM M [79]	Medical images, Reports, Genomics	Classification, Reporting, Visual QA, Summarization
MedVersa [80]	CT, X-ray, Dermatology images, Reports	Classification, Reporting, Visual QA, Segmentation
MMBERT [120]	Radiology images, Captions	Visual QA
MVG [121]	Medical images, Text	Disease simulation
ORacle [16]	Multi-view images, SSG, Descriptions	OR scene graph prediction
PathChat [122]	Histopathology images, QA-pairs	Visual QA
PathLDM [123]	Histopathology images, Reports	Image generation
QUILT-LLaVA [124]	Histopathology images, QA-pairs	Visual QA
R2GenGPT [125]	X-ray, Reports	Reporting
RaDialog [73]	X-ray, Reports	Reporting, Dialogue
RadFM [81]	Medical images, Reports, Descriptions	Reporting, Visual QA
ReXplain [126]	Video, Reports, Masks	Video report generation
RGRG [127]	X-ray, Reports, Bounding-boxes	Reporting
RoentGen [128]	X-ray, Reports	Image generation
SkinGPT-4 [129]	Dermatology images, Clinical notes	Visual QA, Dialogue
Surgical-VQLA++ [130]	Surgical images, QA-pairs	Visual QA
Universal Model [131]	CT, Masks, Descriptions	Segmentation
Vote-MI [132]	MRI, Reports	Visual QA

Table 4: Summary of multimodal MLLM-based methods. The table includes method names, the modalities utilized (e.g., text and medical images), and the primary tasks

addressed, such as report generation, visual QA, and disease classification. (Abbreviations: QA - question answering)

LLM-based methods, in contrast to CLIP approaches, directly integrate multimodal inputs into the language model's embeddings, enabling more complex reasoning and generative tasks. These approaches rely on modality-specific encoders to process non-textual data, converting them into feature embeddings compatible with the LLM's text-based representation space (Table 4). For instance, SkinGPT-4 [129] and RaDialog [73] integrate features from two-dimensional (2D) images, while models like Merlin [77] and CT-CHAT [14] extend this capability to volumetric three-dimensional (3D) CT data. Some models, such as MAIRA-2 [72] and AutoRG-Brain [108], further ground text predictions by incorporating bounding boxes and segmentation masks, enabling interactive, region-based report generation for enhanced explainability [127].

Current advancements also focus on text-guided segmentation and synthetic medical image generation. Text-guided segmentation models like LViT create segmentation masks from textual prompts, enabling tasks such as tumor detection and organ identification [116]. Beyond segmentation, synthetic image generation has emerged as another multimodal approach for data augmentation and model training. Methods such as GenerateCT [114] for CT volumes and RoentGen [128] for X-rays use text-conditioned diffusion models to produce realistic medical images [133].

Generalist models, such as BiomedGPT [13] and MedVersa [80], unify multiple modalities and tasks through shared representations or mixture-of-experts strategies. These models employ specialized modules to process different modalities while a central language model coordinates their outputs, enabling tasks such as classification, segmentation, retrieval, and visual QA. This approach highlights the scalability and versatility of generalist models in addressing complex multimodal challenges in medicine.

5.3 Multimodal LLM applications

MLLMs have been increasingly applied across diverse medical tasks, showcasing their potential to transform clinical workflows and decision support systems. This section highlights key applications where MLLMs contribute to improving healthcare.

A key advancement in multimodal AI is generalist models capable of handling diverse medical data types and tasks. Models such as BiomedGPT [13] and RadFM [81] support a wide range of imaging modalities and anatomical regions, enabling comprehensive diagnostic assistance across multiple specialties.

Radiology report generation remains one of the most important applications of MLLMs in healthcare, providing detailed textual descriptions directly from medical images. Systems such as MAIRA-2 [72] and RaDialog [73] have demonstrated their ability to generate comprehensive reports from X-rays, while CT-CHAT [14] and AutoRG-Brain [108] extend this capability to CT and magnetic resonance imaging (MRI) scans,

respectively. These tools assist radiologists by automating preliminary reporting and standardizing documentation, potentially reducing reporting delays.

Visual QA systems support clinicians in querying medical images using natural language prompts, supporting real-time decision-making and diagnostic interpretation. For instance, models like LLaVA-Med [15] and Med-Flamingo [119] provide concise, contextually relevant answers to clinical queries, assisting radiologists and physicians in complex cases.

Synthetic medical image generation has become increasingly important for data augmentation and simulating rare pathological conditions. Models like GenerateCT [114] and RoentGen [128] generate realistic CT and X-ray images from textual prompts, enhancing dataset diversity.

Semantic scene modeling is another emerging application where models create structured representations of complex environments, such as the operating room. For example, ORacle [16] generates semantic scene graphs to assist with surgical planning and intraoperative navigation by representing tools, anatomy, and procedural stages in a comprehensive framework.

Finally, systems like ReXplain [126] aim to bridge communication gaps between clinicians and patients. By transforming radiology reports into patient-friendly video summaries, these models provide an accessible way to convey complex clinical information, further highlighting multimodal AI's potential to improve patient care.

5.4 Multimodal LLM datasets

Multimodal datasets integrating images, text, and other clinical information (Table 5) are essential for tasks such as radiology report generation, visual QA, and cross-modal retrieval. These datasets not only enable effective model training but are also crucial for ensuring fairness and generalization in medical AI systems. A range of multimodal datasets has been curated to support various medical imaging and diagnostic tasks.

Dataset	Modalities	Size	Application
2D-image-text			
CheXpert [134]	X-ray, Reports, Labels	224K triplets	Chest X-ray
CheXinstruct [111]	X-ray, Instructions	8.5M instruction triplets	Chest X-ray
Harvard-FairVLMed [89]	SLO fundus images, Demographics, Notes	10K samples	Ophthalmology
MedTrinity-25M [135]	Medical images, Captions, Bounding-boxes	25M pairs	Medical imaging
MedVidQA [136]	Medical videos, Labels, QA-pairs	6K videos, 6K labels, 3K QA	Medical videos
MIMIC-CXR [8]	X-ray, Reports	377K images, 227K reports	Chest X-ray
MS-CXR [83]	X-ray, Descriptions, Bounding-boxes	1K image-sentence pairs, Bounding-boxes	Chest X-ray
OmniMedVQA [137]	Medical images, QA	118K images, 127K QA-pairs	Medical imaging
OpenPath [101]	Histopathology images, Captions	208K pairs	Digital pathology
PadChest [138]	X-ray, Reports	160K images, 109K texts	Chest X-ray
PathVQA [139]	Medical images, QA	5K images, 33K QA	Medical imaging
PMC-15M [82]	Medical images, Captions	15M image-text pairs	Medical imaging
PubMedVision [140]	Medical images, QA	1.3M QA pairs	Medical imaging
Quilt-1M [141]	Histopathology images, Captions	1M pairs	Digital pathology
Rad-ReStruct [142]	X-ray, Structured reports	3720 images, 3597 Reports	Chest X-ray
SLAKE [143]	Medical images, QA	642 images, 14K QA pairs	Medical imaging
UniMed [103]	Medical images, Captions	5.3M image-text pairs	Medical imaging
VQA-RAD [144]	Radiology images, Captions	315 images, 3.5K QA pairs	Radiology
3D-volume-text			
AMOS-MM [145] [146]	CT, Reports, QA	2K image-report pairs, 7K QA	Chest, abdomen, pelvis CT
BrainMD [132]	MRI, Reports, EHR	2.5K cases	Brain MRI
BIMCV-R [147]	CT, Reports	8K image-report pairs	CT
CT-RATE [14]	CT, Reports, Labels	25K triplets	Chest CT

INSPECT [9]	CT, Reports, EHR, labels	23K image-report pairs, EHRs	Chest CT
M3D-Data [117]	CT, Captions, QA, Masks	120K images, 42K captions, 509K QA, 149K masks	CT
RadGenome-Brain MRI [108]	MRI, Reports, Masks	3.4K image-region-report triplets	Brain MRI
RadGenome-Chest CT [148]	CT, Reports, Masks, QA	25K image-report pairs, 665K masks, 1.3M QA	Chest CT
Others			
Duke Breast Cancer MRI [149]	Genomic, MRI images, Clinical data	922 subjects	Breast cancer
PTB-XL [150]	ECG signals, Reports, Labels	21K triplets	ECG
PubChemSTM [99]	Chemical structures, Descriptions	280K chemical structure–text pairs	Drug design
SwissProtCLAP [151]	Protein Sequence, Text	441K sequence-text pairs	Protein design

Table 5: Summary of multimodal datasets used for medical AI, grouped by modality categories. The table lists dataset names, the types of modalities (e.g., text and medical images), dataset sizes, and key applications such as image-text retrieval, report generation, and disease classification. (Abbreviations: QA - question answering.)

A substantial proportion of multimodal datasets focus on pairing vision and text data, as this combination is central to tasks where both visual context and descriptive language are critical for diagnostic interpretation. Notable public datasets like MIMICCXR [8] and CheXpert [134] provide rich resources for training 2D vision-language models in radiology. These datasets include not only radiology reports but also diseasespecific labels, enabling more comprehensive evaluations. For benchmarking report generation, ReXGradient [152], a private benchmark dataset of 10,000 studies collected across 67 medical sites in the United States, offers diverse coverage and serves as a reliable standard for radiology-specific performance evaluation.

Expanding beyond radiology, datasets like Quilt-1M [141] have introduced multimodal resources covering additional domains such as digital pathology [122, 153].

Recent advancements have also led to datasets tailored for 3D imaging modalities such as CT [9, 14, 145, 147] and MRI [108]. Notably, RadMD [81] integrates both 2D and 3D imaging modalities, supporting a broader range of applications.

In addition to image-text pairs, a few datasets now include task-specific annotations to support specialized applications. For instance, RadGenome-Brain MRI [108] and RadGenome-Chest CT [148] provide segmentation masks, while datasets like MedTrinity-25M [135] offer bounding box annotations. These annotations are critical for grounding text predictions to specific regions of interest, enhancing both explainability and diagnostic accuracy in multimodal models.

The data formats of multimodal datasets also vary significantly based on their intended use cases. While datasets like OpenPath [101] present images from publicly available sources in formats such as JPEG, datasets like MIMIC-CXR [8] and CTRATE [14] preserve clinical formats such as Digital Imaging and Communications in Medicine (DICOM) and Neuroimaging Informatics Technology Initiative (NIfTI). These formats are essential for maintaining complete clinical information and enabling compatibility with healthcare systems.

Beyond traditional imaging and text combinations, datasets have also begun exploring additional modalities for specialized biomedical tasks. For example, SwissProtCLAP [151] integrates protein sequence data to support protein design

frameworks, highlighting the potential of multimodal datasets to extend AI applications beyond diagnostic imaging into molecular and genomic research.

6 Evaluation metrics for generative AI in medicine

Evaluating generative AI in medicine is essential to ensure models produce accurate, clinically relevant, and reliable outputs [154]. This section explores evaluation metrics for both text generation, such as radiology report generation, and image generation, emphasizing the importance of clinical validity and utility. As general-purpose metrics often fall short in capturing medical accuracy, domain-specific approaches are required.

As report generation is a key application of generative AI in medicine, research has focused on developing reliable evaluation strategies. While standard lexical metrics such as BLEU [157], ROUGE [158], and METEOR [159] are commonly used, they often fail to reflect clinical accuracy, as high scores can be achieved despite factually incorrect outputs. To address this, specialized clinical metrics tailored for report generation have emerged (Table 6).

For instance, NER-based metrics like RaTEScore [155] evaluate key medical entities extracted from both predicted and ground truth reports, offering a more targeted assessment of clinical relevance. RadFact [72] further incorporates grounding by assessing factual correctness against reference image annotations. The GREEN metric described in [154] goes beyond standard evaluations by integrating error detection with explanations. It provides a clinically grounded score alongside human-interpretable feedback on significant errors, making it a promising tool for both model validation and iterative improvement. ReXrank [152], a benchmark for chest X-ray report generation, combines lexical and clinical metrics for more task-specific assessment.

Additionally, clinical efficacy can be measured using standard classification metrics, such as precision, recall, sensitivity, specificity, and F1-score, particularly when evaluation datasets include labeled disease categories [8, 14]. A text classifier can be trained

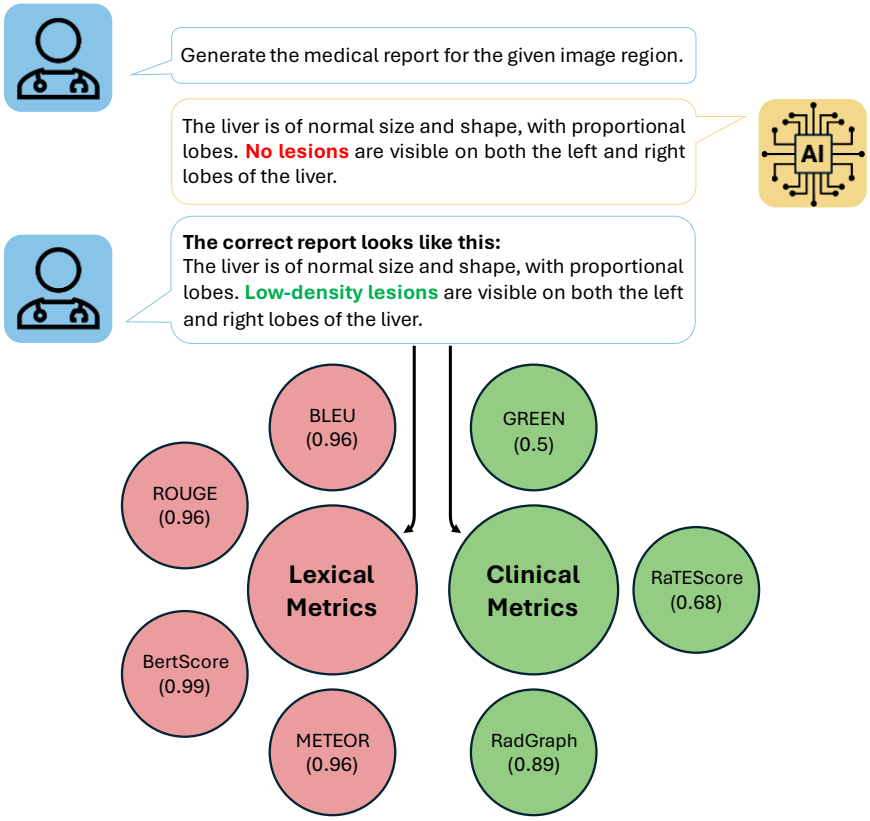

Fig. 4: Evaluation of generative AI in medicine: Lexical metrics from the general domain cannot completely capture the clinical correctness as they mainly cover text similarity. Clinically-relevant metrics like GREEN [154], RaTEScore [155], or RadGraph [156] also evaluate the clinical correctness.

on the generated reports to predict labels, enabling a more structured evaluation of diagnostic accuracy.

Evaluating image generation in medical AI requires considerations beyond standard image quality metrics like Fr´echet Inception Distance [164] and mean squared error. Since synthetic medical images are often used for data augmentation or diagnostic training, their clinical utility must be assessed alongside visual quality. One effective strategy involves generating condition-specific medical images and training a classifier on the synthetic data to evaluate its generalization performance on real clinical cases [114]. This ensures that the generated images are not only visually realistic but also contribute to model performance on downstream tasks, such as disease classification and segmentation.

Despite advancements in specialized evaluation metrics for both text and image generation, challenges remain regarding their generalizability across clinical sites and datasets. Frameworks like ReXamine-Global [165] address this by evaluating the

Metric	Type	Application
CheXbert [160]	Classification	Chest X-ray report labeling
CRAFT-MD [161]	Generative	Conversation evaluation
FineRadScore [162]	Generative	Report evaluation
GREEN [154]	Generative	Report evaluation
Ong Ly et al. [163]	Calibration	Model generalization
RadCliQ [156]	Composite metric	Report evaluation
RadFact [72]	Grounding	Grounded report evaluation
RadGraph-F1 [156]	NER similarity	Report evaluation
RaTEScore [155]	NER similarity	Report evaluation

Table 6: Evaluation metrics for medical report generation. This table summarizes key metrics used to evaluate generative AI systems in medical report generation, categorized by type and primary application. (Abbreviations: NER - named entity recognition)

robustness of metrics across diverse institutions and data distributions. For text generation, a combination of lexical metrics and clinically grounded assessments is essential to ensure factual correctness and clinical relevance. Similarly, for image generation, both visual quality and downstream clinical utility, such as diagnostic performance on real clinical cases, should be jointly evaluated. Ultimately, a multi-dimensional evaluation approach that considers both data diversity and task-specific requirements is crucial for the safe and effective deployment of generative AI in healthcare.

7 Discussion

In this scoping review, we systematically explored the evolution of generative AI in medicine, focusing on LLMs, multimodal LLMs, and their evaluation metrics. Using the PRISMA-ScR framework [21], we collected 144 papers published between January 2020 and December 2024 from PubMed, IEEE Xplore, and Web of Science, complemented by a manual search to ensure comprehensive coverage. Our findings highlight the shift from unimodal LLMs focused on textual tasks to more complex multimodal systems capable of integrating medical images, clinical notes, and structured data. These models have shown promise in enhancing diagnostic support, automating clinical workflows, and reducing the workload of healthcare professionals.

LLMs have advanced biomedical language processing, improving tasks like medical report summarization, named entity recognition, and conversational AI. Adaptation techniques such as supervised finetuning, reinforcement learning, and RAG have further specialized language models for clinical tasks. However, reliance on static datasets like

MIMIC-IV [66] limits the ability to capture evolving medical knowledge. Moreover, privacy issues persist due to the need for extensive data deidentification, and dataset biases can affect fairness by overrepresenting specific populations [166, 167].

Multimodal LLMs extend LLM capabilities by integrating multiple data types, such as medical images and text, to address tasks like report generation, cross-modal retrieval, and clinical question answering. Despite these advancements, data heterogeneity remains a challenge, as clinical datasets often vary significantly in format, quality, and completeness across institutions. Additionally, most widely used datasets, such as MIMIC-CXR and CT-RATE [8, 14], focus heavily on radiology, limiting the generalizability of models to other medical domains.

Evaluating generative AI models in medicine requires specialized metrics that go beyond standard language evaluation metrics. While lexical metrics like BLEU [157] and ROUGE [158] are commonly used, they often fail to capture clinical relevance and factual accuracy. To address this, domain-specific metrics such as RadGraph [156], RaTEScore [155], and GREEN [154] have been developed to assess the clinical validity of generated medical reports. However, challenges remain in standardizing evaluation practices across diverse medical tasks and datasets. Most evaluations are limited to radiology, with less attention given to other specialties. The limited availability of well-annotated multimodal datasets with fine-grained clinical labels further complicates performance benchmarking. Additionally, only a few benchmarking frameworks, such as ReXrank [152], offer the ability to neutrally evaluate models on non-public datasets, limiting comparative performance assessments across different models and data sources. Expanding such benchmarks and ensuring their applicability to a broader range of clinical tasks is essential for developing trustworthy generative models in medicine.

While this scoping review provides a comprehensive overview of generative AI advancements in medicine, it has certain limitations. Despite the systematic search strategy using the PRISMA-ScR framework, the literature search may not have captured all relevant studies due to the rapidly evolving nature of the field. To mitigate this, a manual search was conducted alongside the database queries to ensure the inclusion of recent and high-impact publications. Moreover, while efforts were made to cover multiple clinical specialties, there remains an overrepresentation of radiologyfocused datasets and models, reflecting a broader trend in the literature. We aimed to balance the inclusion of topics and application areas by diversifying the datasets and models included in our analysis, but certain domains such as pathology and genomics remain less represented due to the current availability of multimodal datasets in these fields.

To further advance the development and responsible deployment of generative AI in medicine, several areas need attention [168–170]. First, evaluation frameworks need to evolve beyond lexical metrics by incorporating clinically grounded assessments and domain-specific error analysis. Second, expanding the diversity of training datasets is critical. The current overrepresentation of western institutions and radiology-focused datasets risks introducing biases that limit global applicability [8, 134]. Future datasets should encompass a wider range of medical specialties, imaging modalities, and patient demographics, with careful attention to privacy protection and data fairness. Third, improving model explainability remains a priority [171, 172]. Techniques such as

region-specific grounding can help build clinician trust. Finally, the emergence of generalist models [13, 80] capable of handling multiple modalities and tasks within a unified architecture represents an important step forward, but broader coverage across medical specialties and improved datasets remain essential for widespread adoption.

This scoping review provides a structured analysis of the evolution from unimodal LLMs to multimodal generative AI models in medicine, highlighting their potential for improving diagnostic support, clinical documentation, and decision-making. However, challenges related to data diversity, clinical relevance, model interpretability, and the standardization of evaluation metrics remain critical barriers to widespread adoption. Addressing these challenges through interdisciplinary collaboration, improved datasets, and clinically grounded evaluation strategies will be essential to ensure the responsible deployment of generative AI in healthcare.

Additional information

Acknowledgements This work was partially funded via the EVUK programme ("Next-generation AI for Integrated Diagnostics") of the Free State of Bavaria, the Deutsche Forschungsgemeinschaft (DFG), and Friedrich-Alexander-Universit¨at Erlangen-Nu¨rnberg within the funding program Open Access Publication Funding.

Author contributions The idea for this review article was developed by all authors. L.B. performed the literature search, paper screening, and selection. The first draft of the manuscript was written by L.B. and subsequently refined by L.B. and S.T.A.. S.T.A. provided clinical expertise. L.B., M.K., N.N., A.M., and S.T.A. provided technical expertise. All authors revised the manuscript and approved the final version for submission.

Competing interests The authors declare no competing interests.

Ethical approval No human or animal subjects are involved in this study.

Consent to participate No human or animal subjects are involved in this study.

Consent to publish No human or animal subjects are involved in this study.

List of Abbreviations

- AI - Artificial intelligence
- API - Application programming interface
- CLIP - Contrastive language-image pretraining
- CoT - Chain-of-thought
- CT - Computed tomography
- DICOM - Digital Imaging and Communications in Medicine
- ECG - Electrocardiogram
- EHR - Electronic health record
- LLM - Large language model
- MLLM - Multimodal large language models
- MRI - Magnetic resonance imaging
- NER - Named entity recognition
- NIfTI - Neuroimaging Informatics Technology Initiative
- PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses
- PRISMA-ScR - Preferred Reporting Items for Systematic reviews and MetaAnalyses extension for Scoping Reviews
- QA - Question answering
- RAG - Retrieval augmented generation

- RLHF - Reinforcement learning from human feedback
- RLAIF - Reinforcement learning from AI feedback
- SFT - Supervised finetuning

References

[1] Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical ai. Nature Medicine **28**(9), 1773–1784 (2022)

[2] Tayebi Arasteh, S., Han, T., Lotfinia, M., Kuhl, C., Kather, J.N., Truhn, D., Nebelung, S.: Large language models streamline automated machine learning for clinical studies. Nature Communications **15**(1), 1603 (2024)

[3] Cai, X., Liu, S., Han, J., Yang, L., Liu, Z., Liu, T.: Chestxraybert: A pretrained language model for chest radiology report summarization. IEEE Transactions on Multimedia **25**, 845–855 (2021)

[4] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., Zhang, Y.: Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus **15**(6) (2023)

[5] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Z´ıdek, A., Potapenko, A.,ˇ *et al.*: Highly accurate protein structure prediction with alphafold. nature **596**(7873), 583–589 (2021)

[6] Acosta, J.N., Dogra, S., Adithan, S., Wu, K., Moritz, M., Kwak, S., Rajpurkar, P.: The Impact of AI Assistance on Radiology Reporting: A Pilot Study Using Simulated AI Draft Reports (2024). https://arxiv.org/abs/2412.12042

[7] Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E.P., Seehofnerov´a, A., *et al.*: Adapted large language models can outperform medical experts in clinical text summarization. Nature medicine **30**(4), 1134–1142 (2024)

[8] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)

[9] Huang, S.-C., Huo, Z., Steinberg, E., Chiang, C.-C., Lungren, M.P., Langlotz, C.P., Yeung, S., Shah, N.H., Fries, J.A.: Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis. arXiv preprint arXiv:2311.10798 (2023)

[10] Tayebi Arasteh, S., Siepmann, R., Huppertz, M., Lotfinia, M., Puladi, B., Kuhl, C., Truhn, D., Nebelung, S.: The treasure trove hidden in plain sight: The utility of gpt-4 in chest radiograph evaluation. Radiology **313**(2), 233441 (2024)

[11] Khader, F., Mu¨ller-Franzes, G., Wang, T., Han, T., Tayebi Arasteh, S., Haarburger, C., Stegmaier, J., Bressem, K., Kuhl, C., Nebelung, S., *et al.*: Multimodal deep learning for integrating chest radiographs and clinical parameters: a case

for transformers. Radiology **309**(1), 230806 (2023)

[12] Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., *et al.*: Gencode 2021. Nucleic acids research **49**(D1), 916–923 (2021)

[13] Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B.D., Ren, H., et al.: A generalist vision–language foundation model for diverse biomedical tasks. Nature Medicine, 1–13 (2024)

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

[15] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36** (2024)

[16] Ozsoy, E., Pellegrini, C., Keicher, M., Navab, N.: Oracle: Large vision-language¨ models for knowledge-guided holistic or domain modeling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 455–465 (2024). Springer

[17] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. arXiv preprint arXiv:2306.13549 (2023)

[18] Wang, J., Jiang, H., Liu, Y., Ma, C., Zhang, X., Pan, Y., Liu, M., Gu, P., Xia, S., Li, W., et al.: A comprehensive review of multimodal large language models: Performance and challenges across different tasks. arXiv preprint arXiv:2408.01319 (2024)

[19] Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., Luo, Y.: Multimodal machine learning in precision health: A scoping review. npj Digital Medicine **5**(1), 171 (2022)

[20] He, Y., Huang, F., Jiang, X., Nie, Y., Wang, M., Wang, J., Chen, H.: Foundation model for advancing healthcare: Challenges, opportunities, and future directions. arXiv preprint arXiv:2404.03264 (2024)

[21] Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., Moher, D., Peters, M.D., Horsley, T., Weeks, L., *et al.*: Prisma extension for scoping reviews (prisma-scr): checklist and explanation. Annals of internal medicine **169**(7), 467–473 (2018)

[22] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al.: The prisma 2020 statement: an updated guideline for reporting systematic reviews. bmj **372** (2021)

[23] Ouzzani, M., Hammady, H., Fedorowicz, Z., Elmagarmid, A.: Rayyan—a web and mobile app for systematic reviews. Systematic reviews **5**, 1–10 (2016)

[24] Jiang, L.Y., Liu, X.C., Nejatian, N.P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H.A., Laufer, I., Punjabi, P., *et al.*: Health system-scale language models are all-purpose prediction engines. Nature **619**(7969), 357–362 (2023)

[25] Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)

[26] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pretrained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)

[27] Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., Dufour, R.: Biomistral: A collection of open-source pretrained large language models for medical domains. arXiv preprint arXiv:2402.10373 (2024)

[28] Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, Q., Li, J.: Prompt engineering in consistency and reliability with the evidence-based guideline for llms. npj Digital Medicine **7**(1), 41 (2024)

[29] Lee, H., Phatale, S., Mansoor, H., Lu, K.R., Mesnard, T., Ferret, J., Bishop, C., Hall, E., Carbune, V., Rastogi, A.: Rlaif: Scaling reinforcement learning from human feedback with ai feedback (2023)

[30] Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Li, J., Chen, G., Wu, X., Zhang, Z., Xiao, Q., et al.: Huatuogpt, towards taming language model to be a doctor.
arXiv preprint arXiv:2305.15075 (2023)

[31] Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J., Wang, B.: Huatuogpt-o1, towards medical complex reasoning with llms. arXiv preprint arXiv:2412.18925 (2024)

[32] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Ku¨ttler, H., Lewis, M., Yih, W.-t., Rockt¨aschel, T., *et al.*: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020)

[33] Zakka, C., Shad, R., Chaurasia, A., Dalal, A.R., Kim, J.L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., *et al.*: Almanac—retrieval-augmented language models for clinical medicine. NEJM AI **1**(2), 2300068 (2024)

[34] Arasteh, S.T., Lotfinia, M., Bressem, K., Siepmann, R., Adams, L., Ferber, D., Kuhl, C., Kather, J.N., Nebelung, S., Truhn, D.: RadioRAG: Factual Large Language Models for Enhanced Diagnostics in Radiology Using Online Retrieval Augmented Generation (2024). https://arxiv.org/abs/2407.15621

[35] Gilbert, S., Kather, J.N., Hogan, A.: Augmented non-hallucinating large language models as medical information curators. NPJ Digital Medicine **7**(1), 100 (2024)

[36] Naseem, U., Khushi, M., Reddy, V., Rajendran, S., Razzak, I., Kim, J.: Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2021). IEEE

[37] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.-Y.: Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics **23**(6), 409 (2022)

[38] Wang, H., Gao, C., Dantona, C., Hull, B., Sun, J.: Drg-llama: tuning llama model to predict diagnosis-related group for hospitalized patients. npj Digital Medicine **7**(1), 16 (2024)

[39] Yang, X., Chen, A., PourNejatian, N., Shin, H.C., Smith, K.E., Parisien, C., Compas, C., Martin, C., Costa, A.B., Flores, M.G., *et al.*: A large language model for electronic health records. NPJ digital medicine **5**(1), 194 (2022)

[40] Johnson, A.E., Bulgarelli, L., Pollard, T.J.: Deidentification of free-text medical records using pre-trained bidirectional transformers. In: Proceedings of the ACM Conference on Health, Inference, and Learning, pp. 214–221 (2020)

[41] Kresevic, S., Giuffr`e, M., Ajcevic, M., Accardo, A., Croc`e, L.S., Shung, D.L.: Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. NPJ Digital Medicine **7**(1), 102 (2024)

[42] Mahendran, D., McInnes, B.T.: Extracting adverse drug events from clinical notes. AMIA Summits on Translational Science Proceedings **2021**, 420 (2021)

[43] Lanfredi, R.B., Mukherjee, P., Summers, R.M.: Enhancing chest x-ray datasets with privacy-preserving large language models and multi-type annotations: a data-driven approach for improved classification. Medical Image Analysis **99**, 103383 (2025)

[44] Liu, N., Hu, Q., Xu, H., Xu, X., Chen, M.: Med-bert: A pretraining framework for medical records named entity recognition. IEEE Transactions on Industrial Informatics **18**(8), 5600–5608 (2021)

[45] Han, T., Adams, L.C., Papaioannou, J.-M., Grundmann, P., Oberhauser, T., L¨oser, A., Truhn, D., Bressem, K.K.: Medalpaca–an open-source collection of medical conversational ai models and training data. arXiv preprint arXiv:2304.08247 (2023)

[46] Chen, Z., Cano, A.H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., K¨opf, A., Mohtashami, A., et al.: Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:2311.16079 (2023)

[47] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., *et al.*: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023)

[48] Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., Xie, W.: Towards building multilingual language model for medicine. Nature Communications **15**(1), 8384 (2024)

[49] Mu, Y., Tizhoosh, H.R., Tayebi, R.M., Ross, C., Sur, M., Leber, B., Campbell, C.J.: A bert model generates diagnostically relevant semantic embeddings from pathology synopses with active learning. Communications medicine **1**(1), 11 (2021)

[50] Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., Wang, Y.: Pmc-llama: toward building open-source language models for medicine. Journal of the American Medical Informatics Association, 045 (2024)

[51] Jia, S., Bit, S., Searls, E., Claus, L.A., Fan, P., Jasodanand, V.H., Lauber, M.V., Veerapaneni, D., Wang, W.M., Au, R., et al.: Medpodgpt: A multilingual audioaugmented large language model for medical research and education. medRxiv (2024)

[52] Yan, A., McAuley, J., Lu, X., Du, J., Chang, E.Y., Gentili, A., Hsu, C.-N.: Radbert: adapting transformer-based language models to radiology. Radiology: Artificial Intelligence **4**(4), 210258 (2022)

[53] Schmidt, R.A., Seah, J.C., Cao, K., Lim, L., Lim, W., Yeung, J.: Generative large language models for detection of speech recognition errors in radiology reports. Radiology: Artificial Intelligence **6**(2), 230205 (2024)

[54] Prihoda, D., Maamary, J., Waight, A., Juan, V., Fayadat-Dilman, L., Svozil, D., Bitton, D.A.: Biophi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. In: MAbs, vol. 14, p. 2020203 (2022). Taylor & Francis

[55] Schubach, M., Maass, T., Nazaretyan, L., R¨oner, S., Kircher, M.: Cadd v1. 7: using protein language models, regulatory cnns and other nucleotide-level scores to improve genome-wide variant predictions. Nucleic acids research **52**(D1), 1143–1154 (2024)

[56] Ji, Y., Zhou, Z., Liu, H., Davuluri, R.V.: Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. Bioinformatics **37**(15), 2112–2120 (2021)

[57] Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H., Brydon, E.M., Zeng, Z., Liu, X.S., *et al.*: Transfer learning enables predictions in network biology. Nature **618**(7965), 616–624 (2023)

[58] Hie, B.L., Shanker, V.R., Xu, D., Bruun, T.U., Weidenbacher, P.A., Tang, S., Wu, W., Pak, J.E., Kim, P.S.: Efficient evolution of human antibodies from general protein language models. Nature Biotechnology **42**(2), 275–283 (2024)

[59] Rao, R.M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., Rives, A.: Msa transformer. In: International Conference on Machine Learning, pp. 8844–8856 (2021). PMLR

[60] Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos, J.L., Xiong, C., Sun, Z.Z., Socher, R., *et al.*: Large language models generate functional protein sequences across diverse families. Nature Biotechnology **41**(8), 1099–1106 (2023)

[61] Ferruz, N., Schmidt, S., H¨ocker, B.: Protgpt2 is a deep unsupervised language model for protein design. Nature communications **13**(1), 4348 (2022)

[62] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., *et al.*: Prottrans: Toward understanding the language of life through self-supervised learning. IEEE transactions on pattern analysis and machine intelligence **44**(10), 7112–7127 (2021)

[63] Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., Yao, J.: scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. Nature Machine Intelligence **4**(10), 852–866 (2022)

[64] Rathore, A.S., Choudhury, S., Arora, A., Tijare, P., Raghava, G.P.: Toxinpred 3.0: An improved method for predicting the toxicity of peptides. Computers in Biology and Medicine **179**, 108926 (2024)

[65] Nowak, S., Biesner, D., Layer, Y., Theis, M., Schneider, H., Block, W., Wulff, B., Attenberger, U., Sifa, R., Sprinkart, A.: Transformer-based structuring of freetext radiology report databases. European Radiology **33**(6), 4228–4236 (2023)

[66] Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., *et al.*: Mimic-iv, a freely accessible electronic health record dataset. Scientific data **10**(1), 1 (2023)

[67] Pollard, T.J., Johnson, A.E., Raffa, J.D., Celi, L.A., Mark, R.G., Badawi, O.: The eicu collaborative research database, a freely available multi-center database for critical care research. Scientific data **5**(1), 1–13 (2018)

[68] Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., Zhou, M., Zeng, J., Dong, X., Zhang, R., *et al.*: Meddialog: Large-scale medical dialogue datasets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9241–9250 (2020)

[69] Uniprot: the universal protein knowledgebase in 2023. Nucleic acids research **51**(D1), 523–531 (2023)

[70] Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W.: Genbank. Nucleic acids research **41**(D1), 36–42 (2012)

[71] Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., Ketchum, K.A.: The cptac data portal: a resource for cancer proteomics research. Journal of proteome research **14**(6), 2707–2713 (2015)

[72] Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Bond-Taylor, S., Ilse, M., P´erez-Garc´ıa, F., Salvatelli, V., Sharma, H., Meissen, F., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)

[73] Pellegrini, C., Ozsoy, E., Busam, B., Navab, N., Keicher, M.: Radialog: A¨ large vision-language model for radiology report generation and conversational assistance. arXiv preprint arXiv:2311.18681 (2023)

[74] Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expertlevel detection of pathologies from unannotated chest x-ray images via selfsupervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022)

[75] Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., Rajpurkar, P.: Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In: Machine Learning for Health, pp. 209–219 (2021). PMLR

[76] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language

supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR

[77] Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.-B., Reis, E., Truyts, C., et al.: Merlin: A vision language foundation model for 3d computed tomography. arXiv preprint arXiv:2406.06512 (2024)

[78] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)

[79] Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., *et al.*: Towards generalist biomedical ai. NEJM AI **1**(3), 2300138 (2024)

[80] Zhou, H.-Y., Adithan, S., Acosta, J.N., Topol, E.J., Rajpurkar, P.: A generalist learner for multifaceted medical image interpretation. arXiv preprint arXiv:2405.07988 (2024)

[81] Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology. arXiv preprint arXiv:2308.02463 (2023)

[82] Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI, 2400640 (2024)

[83] Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., *et al.*: Making the most of text semantics to improve biomedical vision–language processing. In: European Conference on Computer Vision, pp. 1–21 (2022). Springer

[84] Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., *et al.*: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15016–15027 (2023)

[85] Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference, pp. 2–25 (2022). PMLR

[86] Javed, S., Mahmood, A., Ganapathi, I.I., Dharejo, F.A., Werghi, N., Bennamoun, M.: Cplip: Zero-shot learning for histopathology with comprehensive vision-language alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11450–11459 (2024)

[87] Yang, L., Xu, S., Sellergren, A., Kohlberger, T., Zhou, Y., Ktena, I., Kiraly, A., Ahmed, F., Hormozdiari, F., Jaroensri, T., et al.: Advancing multimodal medical capabilities of gemini. arXiv preprint arXiv:2405.03162 (2024)

[88] Liu, C., Wan, Z., Cheng, S., Zhang, M., Arcucci, R.: Etp: Learning transferable ecg representations via ecg-text pre-training. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8230–8234 (2024). IEEE

[89] Luo, Y., Shi, M., Khan, M.O., Afzal, M.M., Huang, H., Yuan, S., Tian, Y., Song, L., Kouhana, A., Elze, T., *et al.*: Fairclip: Harnessing fairness in vision-language learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12289–12301 (2024)

[90] Li, H., Chen, Y., Chen, Y., Yu, R., Yang, W., Wang, L., Ding, B., Han, Y.: Generalizable whole slide image classification with fine-grained visual-semantic interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11398–11407 (2024)

[91] Keicher, M., Zaripova, K., Czempiel, T., Mach, K., Khakzar, A., Navab, N.: Flexr: few-shot classification with language embeddings for structured reporting of chest x-rays. In: Medical Imaging with Deep Learning, pp. 1493–1508 (2024). PMLR

[92] Huang, S.-C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3942–3951 (2021)

[93] Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visuallanguage pre-training on chest radiology images. Nature Communications **14**(1), 4542 (2023)

[94] Huang, W., Li, C., Zhou, H.-Y., Yang, H., Liu, J., Liang, Y., Zheng, H., Zhang, S., Wang, S.: Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. Nature Communications **15**(1), 7620 (2024)

[95] Wang, P., Zhang, H., Yuan, Y.: Mcpl: Multi-modal collaborative prompt learning for medical vision-language model. IEEE Transactions on Medical Imaging (2024)

[96] Codella, N.C., Jin, Y., Jain, S., Gu, Y., Lee, H.H., Abacha, A.B., SantamariaPang, A., Guyman, W., Sangani, N., Zhang, S., et al.: Medimageinsight: An open-source embedding model for general domain medical imaging. arXiv preprint arXiv:2410.06542 (2024)

[97] Liu, F., Zhu, T., Wu, X., Yang, B., You, C., Wang, C., Lu, L., Liu, Z., Zheng, Y., Sun, X., *et al.*: A medical multimodal large language model for future pandemics. NPJ Digital Medicine **6**(1), 226 (2023)

[98] Moon, J.H., Lee, H., Shin, W., Kim, Y.-H., Choi, E.: Multi-modal understanding and generation for medical images and text via vision-language pre-training. IEEE Journal of Biomedical and Health Informatics **26**(12), 6070–6080 (2022)

[99] Liu, S., Nie, W., Wang, C., Lu, J., Qiao, Z., Liu, L., Tang, J., Xiao, C., Anandkumar, A.: Multi-modal molecule structure–text model for text-based retrieval and editing. Nature Machine Intelligence **5**(12), 1447–1457 (2023)

[100] Tang, X., Tran, A., Tan, J., Gerstein, M.B.: Mollm: a unified language model for integrating biomedical text with 2d and 3d molecular representations. Bioinformatics **40**(Supplement 1), 357–368 (2024)

[101] Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual– language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)

[102] Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., Gonz´alez, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature, 1–8 (2024)

[103] Khattak, M.U., Kunhimon, S., Naseer, M., Khan, S., Khan, F.S.: Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. arXiv preprint arXiv:2412.10372 (2024)

[104] Pellegrini, C., Keicher, M., Ozsoy, E., Jiraskova, P., Braren, R., Navab, N.:¨ Xplainer: From x-ray observations to explainable zero-shot diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 420–429 (2023). Springer

[105] Ran, A., Liu, H.: Joint spatio-temporal features constrained self-supervised electrocardiogram representation learning. Biomedical Engineering Letters **14**(2), 209–220 (2024)

[106] Kang, Y., Yang, G., Eom, H., Han, S., Baek, S., Noh, S., Shin, Y., Park, C.: Ganbased patient information hiding for an ecg authentication system. Biomedical Engineering Letters **13**(2), 197–207 (2023)

[107] Alsharid, M., Cai, Y., Sharma, H., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Gaze-assisted automatic captioning of fetal ultrasound videos using threeway multi-modal deep neural networks. Medical Image Analysis **82**, 102630 (2022)

[108] Lei, J., Zhang, X., Wu, C., Dai, L., Zhang, Y., Zhang, Y., Wang, Y., Xie, W., Li, Y.: Autorg-brain: Grounded report generation for brain mri. arXiv preprint arXiv:2407.16684 (2024)

[109] Cui, H., Mao, L., Liang, X., Zhang, J., Ren, H., Li, Q., Li, X., Yang, C.: Biomedical visual instruction tuning with clinician preference alignment. arXiv preprint arXiv:2406.13173 (2024)

[110] Wang, S., Zhao, Z., Ouyang, X., Wang, Q., Shen, D.: Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257 (2023)

[111] Chen, Z., Varma, M., Delbrouck, J.-B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J.M.J., Youssef, A., Cohen, J.P., Reis, E.P., et al.: Chexagent: Towards a foundation model for chest x-ray interpretation. arXiv preprint arXiv:2401.12208 (2024)

[112] Gu, T., Liu, D., Li, Z., Cai, W.: Complex organ mask guided radiology report generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 7995–8004 (2024)

[113] Chen, X., Zhang, W., Xu, P., Zhao, Z., Zheng, Y., Shi, D., He, M.: Ffa-gpt: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. npj Digital Medicine **7**(1), 111 (2024)

[114] Hamamci, I.E., Er, S., Sekuboyina, A., Simsar, E., Tezcan, A., Simsek, A.G., Esirgun, S.N., Almas, F., Dogan, I., Dasdelen, M.F., et al.: Generatect: Textconditional generation of 3d chest ct volumes. arXiv preprint arXiv:2305.16037 (2023)

[115] Huh, J., Park, S., Lee, J.E., Ye, J.C.: Improving medical speech-to-text accuracy with vision-language pre-training model. arXiv preprint arXiv:2303.00091 (2023)

[116] Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE transactions on medical imaging (2023)

[117] Bai, F., Du, Y., Huang, T., Meng, M.Q.-H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)

[118] Sharma, H., Salvatelli, V., Srivastav, S., Bouzid, K., Bannur, S., Castro, D.C., Ilse, M., Bond-Taylor, S., Ranjit, M.P., Falck, F., et al.: Maira-seg: Enhancing radiology report generation with segmentation-aware multimodal large language models. arXiv preprint arXiv:2411.11362 (2024)

[119] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H), pp. 353–367 (2023). PMLR

[120] Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D., Jawahar, C.: Mmbert: Multimodal bert pretraining for improved medical vqa. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1033–1036 (2021). IEEE

[121] Cao, X., Liang, K., Liao, K.-D., Gao, T., Ye, W., Chen, J., Ding, Z., Cao, J., Rehg, J.M., Sun, J.: Medical video generation for disease progression simulation. arXiv preprint arXiv:2411.11943 (2024)

[122] Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Zhao, M., Chow, A.K., Ikemura, K., Kim, A., Pouli, D., Patel, A., *et al.*: A multimodal generative ai copilot for human pathology. Nature **634**(8033), 466–473 (2024)

[123] Yellapragada, S., Graikos, A., Prasanna, P., Kurc, T., Saltz, J., Samaras, D.: Pathldm: Text conditioned latent diffusion model for histopathology. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5182–5191 (2024)

[124] Seyfioglu, M.S., Ikezogwo, W.O., Ghezloo, F., Krishna, R., Shapiro, L.: Quiltllava: Visual instruction tuning by extracting localized narratives from opensource histopathology videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13183–13192 (2024)

[125] Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology **1**(3), 100033 (2023)

[126] Luo, L., Vairavamurthy, J., Zhang, X., Kumar, A., Ter-Oganesyan, R.R., Schroff, S.T., Shilo, D., Hossain, R., Moritz, M., Rajpurkar, P.: Rexplain: Translating radiology into patient-friendly video reports. arXiv preprint arXiv:2410.00441 (2024)

[127] Tanida, T., Mu¨ller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7433–7442 (2023)

[128] Bluethgen, C., Chambon, P., Delbrouck, J.-B., Sluijs, R., Pol acin, M., Zambrano Chaves, J.M., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A.S.: A vision–language foundation model for the generation of realistic chest x-ray images. Nature Biomedical Engineering, 1–13 (2024)

[129] Zhou, J., He, X., Sun, L., Xu, J., Chen, X., Chu, Y., Zhou, L., Liao, X., Zhang, B., Afvari, S., *et al.*: Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. Nature Communications **15**(1), 5649 (2024)

[130] Bai, L., Wang, G., Islam, M., Seenivasan, L., Wang, A., Ren, H.: Surgicalvqla++: Adversarial contrastive learning for calibrated robust visual questionlocalized answering in robotic surgery. Information Fusion **113**, 102602 (2025)

[131] Liu, J., Zhang, Y., Chen, J.-N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21152–21164 (2023)

[132] Wang, Y., Dai, Y., Jones, C., Sair, H.I., Shen, J., Loizou, N., Hsu, W.-C., Imami, M.R., Jiao, Z., Zhang, P.J., et al.: Enhancing vision-language models for medical imaging: bridging the 3d gap with innovative slice selection. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track

[133] Khader, F., Mu¨ller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., *et al.*: Denoising diffusion probabilistic models for 3d medical image generation. Scientific Reports **13**(1), 7303 (2023)

[134] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., *et al.*: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)

[135] Xie, Y., Zhou, C., Gao, L., Wu, J., Li, X., Zhou, H.-Y., Liu, S., Xing, L., Zou, J., Xie, C., et al.: Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. arXiv preprint arXiv:2408.02900 (2024)

[136] Gupta, D., Attal, K., Demner-Fushman, D.: A dataset for medical instructional video classification and question answering. Scientific Data **10**(1), 158 (2023)

[137] Hu, Y., Li, T., Lu, Q., Shao, W., He, J., Qiao, Y., Luo, P.: Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22170–22183 (2024)

[138] Bustos, A., Pertusa, A., Salinas, J.-M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)

[139] He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)

[140] Chen, J., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., Yu, G., et al.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. arXiv preprint arXiv:2406.19280 (2024)

[141] Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. Advances in neural information processing systems **36** (2024)

[142] Pellegrini, C., Keicher, M., Ozsoy, E., Navab, N.: Rad-restruct: A novel vqa¨ benchmark and method for structured radiology reporting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 409–419 (2023). Springer

[143] Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., Wu, X.-M.: Slake: A semanticallylabeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1650–1654 (2021). IEEE

[144] Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)

[145] codabench: MICCAI24 AMOS-MM: ABDOMINAL MULTIMODAL ANALYSIS CHALLENGE. Accessed: 2024-11-04 (2024). https://www.codabench.org/competitions/3137/

[146] Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., *et al.*: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022)

[147] Chen, Y., Liu, C., Liu, X., Arcucci, R., Xiong, Z.: Bimcv-r: A landmark dataset for 3d ct text-image retrieval. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 124–134 (2024). Springer

[148] Zhang, X., Wu, C., Zhao, Z., Lei, J., Zhang, Y., Wang, Y., Xie, W.: Radgenomechest ct: A grounded vision-language dataset for chest ct analysis. arXiv preprint arXiv:2404.16754 (2024)

[149] Saha, A., Harowicz, M.R., Grimm, L.J., Kim, C.E., Ghate, S.V., Walsh, R., Mazurowski, M.A.: A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. British journal of cancer **119**(4), 508–516 (2018)

[150] Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T.: Ptb-xl, a large publicly available electrocardiography dataset. Scientific data **7**(1), 1–15 (2020)

[151] Liu, S., Li, Y., Li, Z., Gitter, A., Zhu, Y., Lu, J., Xu, Z., Nie, W., Ramanathan, A., Xiao, C., et al.: A text-guided protein design framework. arXiv preprint arXiv:2302.04611 (2023)

[152] Zhang, X., Zhou, H.-Y., Yang, X., Banerjee, O., Acosta, J.N., Miller, J., Huang, O., Rajpurkar, P.: Rexrank: A public leaderboard for ai-powered radiology report generation. arXiv preprint arXiv:2411.15122 (2024)

[153] Ferber, D., W¨olflein, G., Wiest, I.C., Ligero, M., Sainath, S., Ghaffari Laleh, N., El Nahhas, O.S., Mu¨ller-Franzes, G., J¨ager, D., Truhn, D., *et al.*: In-context learning enables multimodal large language models to classify cancer pathology images. Nature Communications **15**(1), 10104 (2024)

[154] Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Michalson, A.E., Moseley, M., Langlotz, C., Chaudhari, A.S., et al.: Green: Generative radiology report evaluation and error notation. arXiv preprint arXiv:2405.03595 (2024)

[155] Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Ratescore: A metric for radiology report generation. medRxiv, 2024–06 (2024)

[156] Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., et al.: Evaluating progress in automatic chest x-ray radiology report generation. Patterns **4**(9) (2023)

[157] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)

[158] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[159] Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization, pp. 65–72 (2005)

[160] Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.P.: Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. arXiv preprint arXiv:2004.09167 (2020)

[161] Johri, S., Jeong, J., Tran, B.A., Schlessinger, D.I., Wongvibulsin, S., Barnes, L.A., Zhou, H.-Y., Cai, Z.R., Van Allen, E.M., Kim, D., Daneshjou, R., Rajpurkar, P.: An evaluation

framework for clinical use of large language models in patient interaction tasks. Nature Medicine (2025) https://doi.org/10.1038/ s41591-024-03328-5

[162] Huang, A., Banerjee, O., Wu, K., Reis, E.P., Rajpurkar, P.: Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores. arXiv preprint arXiv:2405.20613 (2024)

[163] Ong Ly, C., Unnikrishnan, B., Tadic, T., Patel, T., Duhamel, J., Kandel, S., Moayedi, Y., Brudno, M., Hope, A., Ross, H., *et al.*: Shortcut learning in medical ai hinders generalization: method for estimating ai model generalization without external data. NPJ Digital Medicine **7**(1), 124 (2024)

[164] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)

[165] Banerjee, O., Saenz, A., Wu, K., Clements, W., Zia, A., Buensalido, D., Kavnoudias, H., Abi-Ghanem, A.S., Ghawi, N.E., Luna, C., *et al.*: Rexamine-global: A framework for uncovering inconsistencies in radiology report generation metrics. In: Biocomputing 2025: Proceedings of the Pacific Symposium, pp. 185–198 (2024). World Scientific

[166] Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., Liu, J.: Ethical considerations of using chatgpt in health care. Journal of Medical Internet Research **25**, 48009 (2023)

[167] Li, H., Moon, J.T., Purkayastha, S., Celi, L.A., Trivedi, H., Gichoya, J.W.: Ethics of large language models in medicine and medical research. The Lancet Digital Health **5**(6), 333–335 (2023)

[168] Paschali, M., Chen, Z., Blankemeier, L., Varma, M., Youssef, A., Bluethgen, C., Langlotz, C., Gatidis, S., Chaudhari, A.: Foundation models in radiology: What, how, when, why and why not. arXiv preprint arXiv:2411.18730 (2024)

[169] Bluethgen, C., Van Veen, D., Zakka, C., Link, K., Fanous, A., Daneshjou, R., Frauenfelder, T., Langlotz, C., Gatidis, S., Chaudhari, A.: Best practices for large language models in radiology. arXiv preprint arXiv:2412.01233 (2024)

[170] Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., *et al.*: Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature medicine **30**(9), 2613–2622 (2024)

[171] Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let's verify step by step. arXiv preprint arXiv:2305.20050 (2023)

[172] Chua, M., Kim, D., Choi, J., Lee, N.G., Deshpande, V., Schwab, J., Lev, M.H., Gonzalez, R.G., Gee, M.S., Do, S.: Tackling prediction uncertainty in machine learning for healthcare. Nature Biomedical Engineering **7**(6), 711–718 (2023)

A Supplementary information

A.1 PRISMA-ScR Checklist

Section	Item	PRISMA-ScR checklist item	Section
Title			
Title	1	Identify the report as a scoping review.	Title
Abstract			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	Abstract
Introduction			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	1
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	1
Methods			

Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	2
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	2.1

Section	**Item**	**PRISMA-ScR checklist item**	**Section**
Information sources	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	2.2
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	A.2
Selection of sources of evidence	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	2.3
Data charting process	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	2.4

Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	2.3
Critical appraisal of individual sources of evidence	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	N/A
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	2.5
Results			
Selection of sources 14 Give numbers of sources of evidence 3 of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.			
Section	**Item**	**PRISMA-ScR checklist item**	**Section**
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations	4, 5, 6
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	N/A
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	4, 5, 6
Synthesis of resu	18	Summarize and/or present the charting results as they relate to the review questions and objectives	4, 5, 6
Discussion			

Summary ofevidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	7
Limitations	20	Discuss the limitations of the scoping review process	7
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	7
Funding			
Funding	19	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	Additional information

Table S.1: PRISMA-ScR checklist [21]

A.2 Database search queries

Database	Results	Search string
PubMed	1,325	(”medic*”[TIAB] OR ”healthcare”[TIAB] OR ”clinic*”[TIAB] OR ”diagnosis*”[TIAB] OR ”biomedical”[TIAB]) AND (”language model*”[TIAB] OR ”LLM”[TIAB]) NOT (”ChatGPT”[TIAB]) NOT (review[PT])
	390	(”medic*”[TIAB] OR ”healthcare”[TIAB] OR ”clinic*”[TIAB] OR ”diagnosis*”[TIAB] OR ”biomedical”[TIAB]) AND (”language”[TIAB] OR ”language model*”[TIAB] OR ”LLM”[TIAB]) AND (”multimodal”[TIAB] OR ”multi-modal”[TIAB] OR ”generalist”[TIAB] OR ”CLIP*”[TIAB]) NOT (”ChatGPT”[TIAB]) NOT (review[PT])
IEEE Xplore	893	(”All Metadata”:”medic*” OR ”All Metadata”:”healthcare” OR ”All Metadata”:”clinic*” OR ”All Metadata”:”diagnosis*” OR ”biomedical”) AND (”All Metadata”:”language model*” OR ”All Metadata”:”LLM”) NOT (”All Metadata”:”ChatGPT”) NOT (”All Metadata”:”Review”) NOT (”All Metadata”:”Study”)
	240	(”All Metadata”:”medic*” OR ”All Metadata”:”healthcare” OR ”All Metadata”:”clinic*” OR ”All Metadata”:”diagnosis*” OR ”biomedical”) AND (”All Metadata”:”language” OR ”All Metadata”:”language model*” OR ”LLM”) AND (”All Metadata”:”multimodal” OR ”All Metadata”:”multimodal” OR ”generalist” OR ”CLIP*”) NOT (”All Metadata”:”ChatGPT”) NOT (”All Metadata”:”Review”) NOT (”All Metadata”:”Study”)
Web of Science	1,111	TS=(”medic*” OR ”healthcare” OR ”clinic*” OR ”diagnosis*” OR ”biomedical”) AND TS=(”language model*” OR ”LLM”) NOT TS=(”ChatGPT”) AND DT=(”Article”)

425	TS=(”medic*” OR ”healthcare” OR ”clinic*” OR ”diagnosis*” OR ”biomedical”) AND TS=(”language” OR ”language model*” OR ”LLM”) AND TS=(”multimodal” OR ”multi-modal” OR ”generalist” OR ”CLIP*”) NOT TS=(”ChatGPT”) AND DT=(”Article”)

Table S.2: Search results from different databases