Biomedical Engineering Letters (2025) 15:845–863
https://doi.org/10.1007/s13534-025-00497-1

REVIEW ARTICLE

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]

Received: 31 January 2025 / Revised: 2 June 2025 / Accepted: 17 July 2025 / Published online: 22 August 2025

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 text-only 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, 145 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

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
PLM	Pretrained language model
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
PRISMA-ScR	Preferred reporting items for systematic reviews and meta-analyses 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

✉ Lukas Buess
lukas.buess@fau.de

[1] Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

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

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 to support more comprehensive and clinically relevant decision-making. Recent advancements in this field mark a shift beyond language-focused tasks toward complex, multimodal data integration [13–15]. These systems hold potential for improved diagnostic accuracy and broader applications, from predictive analytics to complex interventional support [16]. Figure 1 illustrates how such models transform heterogeneous medical inputs into clinically meaningful insights through an iterative pipeline.

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

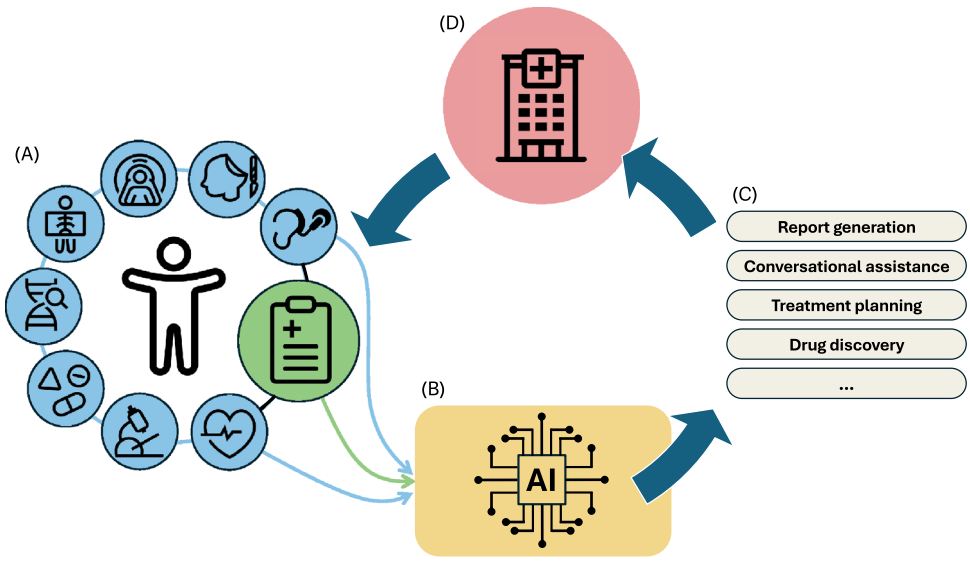

Fig. 1 Multimodal AI pipeline in healthcare: **A** Diverse medical data modalities (e.g., images, genomics, and clinical notes) are collected and processed, **B** transformed into unified representations by AI models, **C** used to generate insights such as reports, conversational assistance, and treatment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

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 PRISMA-ScR 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: "multimodal", "medical" and "language model". 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 text-only 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. Table 1 (text-only LLMs), Table 2 (text-only datasets), Table 3 (contrastive learning methods), Table 4 (MLLMs), Table 5 (multimodal datasets), and Table 6 (evaluation metrics) summarize the results.

Table 1 Summary of language model methods categorized by application to clinical text and bioinformatics

Study	Downstream task
Clinical text	
Almanac [33]	QA
BioALBERT [27]	NER
BioBERT [26]	NER, QA
BioGPT [34]	Classification, QA
BioMistral [28]	QA
ChatDoctor [4]	Dialogue
ChestXRayBERT [3]	Summarization
DRG-LLaMA [35]	Classification
GatorTron [36]	QA
HuatuoGPT [32]	Dialogue
HuatuoGPT-o1 [32]	Dialogue
Johnson et al. [37]	Deidentification
Kresevic et al. [38]	Summarization
Mahendran and McInnes [39]	NER
MAPLEZ [40]	Classification
Med-BERT [41]	NER
MedAlpaca [42]	QA
MEDITRON-70B [43]	QA
MED-PaLM [29]	QA
MMed-Llama 3 [44]	QA
Mu et al. [45]	Classification
NYUTron [24]	Clinical outcome prediction
PMC-LLaMA [46]	QA
PodGPT [47]	QA
RadBERT [48]	Classification, Summarization
Schmidt et al. [49]	Error detection
Bioinformatics	
AlphaFold [5]	Structure prediction
BioPhi [50]	Antibody design
CADD v1.7 [51]	Scoring
DNABERT [52]	Structure analysis
Geneformer [53]	Classification
Hie et al. [54]	Antibody design
MSA Transformer [55]	Structure analysis
ProGen [56]	Structure prediction
ProtGPT2 [57]	Protein design
ProtTrans [58]	Structure analysis
scBERT [59]	Classification
ToxinPred 3.0 [60]	Classification

The table includes method names and target applications

NER, named entity recognition; QA, question answering

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, 84 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, 145 papers were included in this review.

Table 2 Summary of datasets used for training medical LLMs, categorized into clinical text and bioinformatics data

Dataset	Size	Application
Clinical text		
eICU-CRD [67]	200K instances	EHR
GAP-Replay [43]	48.1B tokens	Literature
MedDialog-EN [68]	250K dialogues	Dialogue
MedC-K [46]	4.8M papers, 30K textbooks	Literature
MedC-I [46]	202M tokens	Dialogue, QA
Medical Meadow [42]	160K instances	QA
MIMIC-IV [66]	299K patients	EHR
MMedC [44]	25.5B tokens	Multilingual literature
MultiMedQA [29]	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

The table includes dataset names, sizes, and primary application areas

NA, not available; NER, named entity recognition; QA, question answering; EHR, electronic health record

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 rooted in the transformer architecture [25], which uses self-attention mechanisms to capture contextual relationships and long-range dependencies in text. Language models use self-supervised learning objectives such as masked language modeling or causal language modeling, which allow them to learn from vast unlabeled datasets. This enables scalable pretraining and contributes to their effectiveness when adapted to medical applications.

Table 3 Summary of multimodal CLIP-based methods

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

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

QA, question answering

4.1 Language model methods

Language models for medical applications (Table 1) differ in their architectures and adaptation strategies. Early pretrained language models (PLMs) such as BioBERT [26] and BioALBERT [27] are based on the BERT architecture and pretrained on biomedical corpora using masked language modeling. They are typically finetuned for specific tasks like named entity recognition (NER) or question answering (QA). These models are optimized for understanding and classification rather than open-ended generation, and thus do not fall under the current definition of modern LLMs.

In contrast, modern LLMs are generally generative and adapted for medical use through instruction-tuned supervised finetuning (SFT), enabling capabilities like zero-shot or few-shot generalization across diverse tasks. Prominent examples include BioMistral [28], ChatDoctor [4], and Med-PaLM [29].

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 [30].

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 [31]. 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 [32], which uses RLAIF for clinical alignment.

Another recent development in model adaptation 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 [61] applies CoT prompting to improve medical response clarity and ensure step-by-step diagnostic reasoning.

An additional adaptation technique is retrieval augmented generation (RAG) [62], 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 [63] combine generative capabilities with retrieval systems. However, maintaining the retrieval database and ensuring its comprehensiveness pose ongoing challenges [4, 64].

Table 4 Summary of multimodal MLLM-based methods

Study	Modalities	Downstream task
Alsharid et al. [115]	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 [116]	Medical images, Instructions	Visual QA
ChatCAD [117]	X-ray, Reports	Reporting
CheXagent [118]	X-ray, Reports	Classification, Reporting, Grounding
COMG [119]	X-ray, Reports, Masks	Reporting
CT-CHAT [14]	CT, Reports	Reporting, Visual QA
FFA-GPT [120]	Fundus fluorescein angiography, Reports	Reporting, Visual QA
GenerateCT [111]	CT, Reports	Image generation
Huh et al. [121]	X-ray, Reports	Reporting
LLaVA-Med [15]	Medical images, Captions	Visual QA
LViT [110]	CT, X-ray, Masks, Text annotations	Segmentation
M3D-LaMed [122]	CT, Reports, Masks	Reporting, Visual QA, Segmentation
MAIRA-2 [72]	X-ray, Reports, Masks	Reporting, Grounding
MAIRA-Seg [123]	X-ray, Reports, Masks	Reporting
Med-Flamingo [124]	Medical images, Captions	Visual QA
Med-MoE [114]	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 [125]	Radiology images, Captions	Visual QA
MVG [126]	Medical images, Text	Disease simulation
ORacle [16]	Multi-view images, SSG, Descriptions	OR scene graph prediction
PathChat [127]	Histopathology images, QA-pairs	Visual QA
PathLDM [128]	Histopathology images, Reports	Image generation
QUILT-LLaVA [129]	Histopathology images, QA-pairs	Visual QA
R2GenGPT [130]	X-ray, Reports	Reporting
RaDialog [73]	X-ray, Reports	Reporting, Dialogue
RadFM [81]	Medical images, Reports, Descriptions	Reporting, Visual QA
ReXplain [131]	Video, Reports, Masks	Video report generation
RGRG [109]	X-ray, Reports, Bounding-boxes	Reporting
RoentGen [112]	X-ray, Reports	Image generation
SkinGPT-4 [107]	Dermatology images, Clinical notes	Visual QA, Dialogue
Surgical-VQLA++ [132]	Surgical images, QA-pairs	Visual QA
Universal Model [133]	CT, Masks, Descriptions	Segmentation
Vote-MI [134]	MRI, Reports	Visual QA

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

QA, question answering

4.2 Language model applications

LLMs have revolutionized various applications in biomedical language processing, demonstrating utility across a range of tasks. In named entity recognition, they 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 [27].

Dialogue systems represent another application of LLMs in medicine. Models like ChatDoctor [4] and HuatuoGPT [32] facilitate patient interactions, simulate doctor-patient 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

Table 5 Summary of multimodal datasets used for medical AI, grouped by modality categories

Dataset	Modalities	Size	Application
2D-image-text			
CheXpert [135]	X-ray, Reports, Labels	224K triplets	Chest X-ray
CheXinstruct [118]	X-ray, Instructions	8.5M instruction triplets	Chest X-ray
Harvard-FairVLMed [89]	SLO fundus images, Demographics, Notes	10K samples	Ophthalmology
MedTrinity-25 M [136]	Medical images, Captions, Bounding-boxes	25M pairs	Medical imaging
MedVidQA [137]	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 [138]	Medical images, QA	118K images, 127K QA-pairs	Medical imaging
OpenPath [101]	Histopathology images, Captions	208K pairs	Digital pathology
PadChest [139]	X-ray, Reports	160K images, 109K texts	Chest X-ray
PathVQA [140]	Medical images, QA	5K images, 33K QA	Medical imaging
PMC-15 M [82]	Medical images, Captions	15M image-text pairs	Medical imaging
PubMedVision [141]	Medical images, QA	1.3M QA pairs	Medical imaging
Quilt-1 M [142]	Histopathology images, Captions	1M pairs	Digital pathology
Rad-ReStruct [143]	X-ray, Structured reports	3720 images, 3597 Reports	Chest X-ray
SLAKE [144]	Medical images, QA	642 images, 14K QA pairs	Medical imaging
UniMed [103]	Medical images, Captions	5.3M image-text pairs	Medical imaging
VQA-RAD [145]	Radiology images, Captions	315 images, 3.5K QA pairs	Radiology
3D-volume-text			
AMOS-MM [146, 147]	CT, Reports, QA	2K image-report pairs, 7K QA	Chest, abdomen, pelvis CT
BrainMD [134]	MRI, Reports, EHR	2.5K cases	Brain MRI
BIMCV-R [148]	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 [122]	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 [149]	CT, Reports, Masks, QA	25K image-report pairs, 665K masks, 1.3M QA	Chest CT
Others			
Duke Breast Cancer MRI [150]	Genomic, MRI images, Clinical data	922 subjects	Breast cancer
PTB-XL [151]	ECG signals, Reports, Labels	21K triplets	ECG
PubChemSTM [99]	Chemical structures, Descriptions	280K chemical structure–text pairs	Drug design
SwissProtCLAP [152]	Protein Sequence, Text	441K sequence-text pairs	Protein design

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

QA, question answering

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 [37, 40].

Text classification is another important application area for LLMs in medicine. These models have been used to categorize medical literature and to predict patient outcomes based on clinical text, highlighting their ability to extract structured insights from unstructured data [24].

In bioinformatics, LLMs have expanded beyond language processing to analyze biological sequences like DNA, RNA, and proteins. Models such as DNABERT [52] have advanced gene annotation, while AlphaFold [5] has achieved groundbreaking success in protein structure prediction.

Table 6 Evaluation metrics for medical report generation

Metric	Type	Application
CheXbert [162]	Classification	Chest X-ray report labeling
CRAFT-MD [163]	Generative	Conversation evaluation
FineRadScore [161]	Generative	Report evaluation
GREEN [155]	Generative	Report evaluation
Ong Ly et al. [164]	Calibration	Model generalization
RadCliQ [157]	Composite metric	Report evaluation
RadFact [72]	Grounding	Grounded report evaluation
RadGraph-F1 [157]	NER similarity	Report evaluation
RaTEScore [156]	NER similarity	Report evaluation

This table summarizes key metrics used to evaluate generative AI systems in medical report generation, categorized by type and primary application

NER, named entity recognition

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 [43] and MedC-K [46], 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.

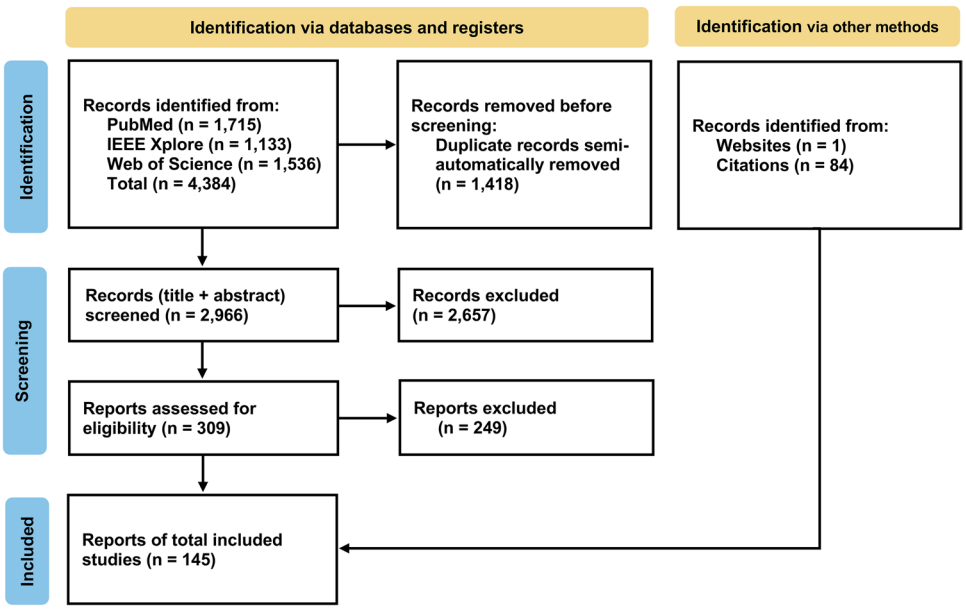

Fig. 2 PRISMA flow diagram illustrating the study selection process for the scoping review. The diagram shows the number of records identified through database searches and manual searches, the removal of duplicates, the screening of titles and abstracts, the review of full-text articles, and the final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

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 outline two distinct architectural paradigms in multimodal AI: contrastive models (e.g., CLIP [76]) and generative multimodal large language models (MLLMs) (see Fig. 3). Contrastive models learn joint embeddings of different modality pairs and are primarily used for tasks such as retrieval or classification. While they do not support language generation or multimodal reasoning, they form the foundation for many downstream applications and are often used to pretrain modality encoders for MLLMs. In contrast, MLLMs directly integrate multimodal inputs into a language model to generate natural language outputs and perform complex reasoning.

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 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 Contrastive multimodal methods

Contrastive models like CLIP and its medical variants align paired inputs in a shared embedding space and are widely used for representation learning. Table 3 summarizes recent CLIP-based approaches across modalities.

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,

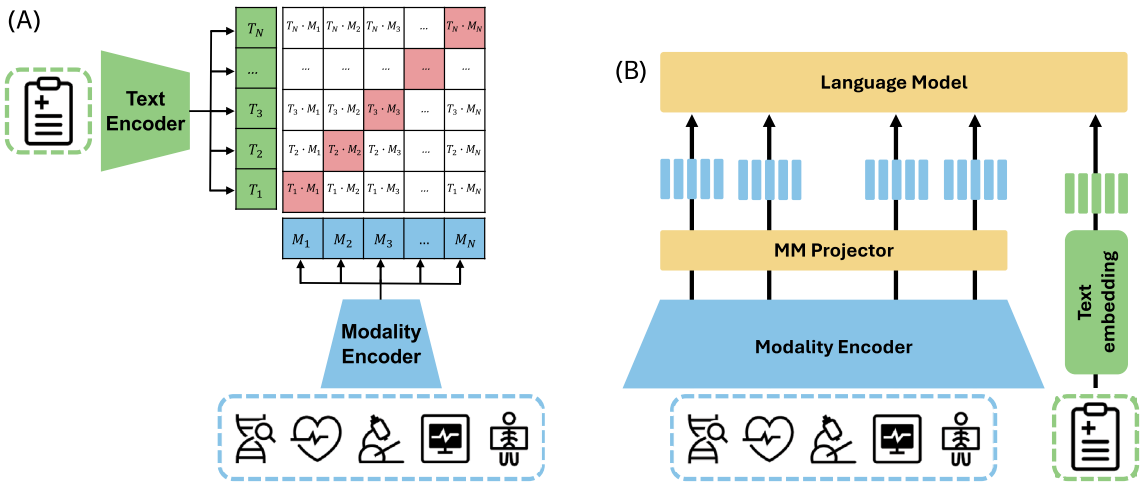

Fig. 3 Multimodal architectures: **A** CLIP-based models, which align embeddings of different modalities in a shared latent space for retrieval or classification; and **B** MLLM-based models, which integrate multimodal inputs directly into the language model for generative tasks such as reporting or reasoning

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.

5.3 Multimodal LLM methods

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 [107] 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 [109].

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 [110]. Beyond segmentation, synthetic image generation has emerged as another multimodal approach for data augmentation and model training. Methods such as GenerateCT [111] for CT volumes and RoentGen [112] for X-rays use text-conditioned diffusion models to produce realistic medical images [113].

Generalist models, such as BiomedGPT [13] and MedVersa [80], unify multiple modalities and tasks either through shared representations, by combining specialized expert models under a common orchestrator, or by employing mixture-of-experts (MoE) architectures with learnable routing mechanisms [114]. 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.

5.4 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 [124] 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 [111] and RoentGen [112] 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 [131] 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.

5.5 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.

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 MIMIC-CXR [8] and CheXpert [135] provide rich resources for training 2D vision-language models in radiology. These datasets include not only radiology reports but also disease-specific labels, enabling more comprehensive evaluations. For benchmarking report generation, ReXGradient [153], 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-1 M [142] have introduced multimodal resources covering additional domains such as digital pathology [127, 154].

Recent advancements have also led to datasets tailored for 3D imaging modalities such as CT [9, 14, 146, 148] 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 [149] provide segmentation masks, while datasets like MedTrinity-25 M [136] 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 CT-RATE [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 [152] 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 [155]. 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 robust evaluation strategies. While standard lexical metrics such as BLEU [158], ROUGE [159], and METEOR [160] are commonly used, they often fail to reflect clinical accuracy, as high scores can be achieved despite factually incorrect outputs. These metrics focus on surface-level similarity between generated and reference texts, employing n-gram precision (BLEU), recall of overlapping segments (ROUGE), and flexible word-level matching that accounts for stemming and synonyms (METEOR). However, they lack sensitivity to clinical nuances such as negation, anatomical detail, and factual correctness (see Fig. 4).

To address these shortcomings, several specialized metrics have been proposed for evaluating medical reports (Table 6). RaTEScore [156] is an entity-aware metric that evaluates the overlap of extracted clinical concepts, including anatomical structures, findings, and their relationships. It captures clinically significant variations in expression, such as synonymous terms and negated findings, providing a more informative measure of clinical relevance.

FineRadScore [161] evaluates radiology reports by prompting powerful language models like GPT-4 to assess clinical aspects such as factuality, temporal consistency, and severity. While effective, it depends on access to strong language models, which may limit reproducibility and practical use.

GREEN [155] addresses this limitation by distilling the evaluation capability of large models into a smaller, instruction-tuned language model. It detects and categorizes

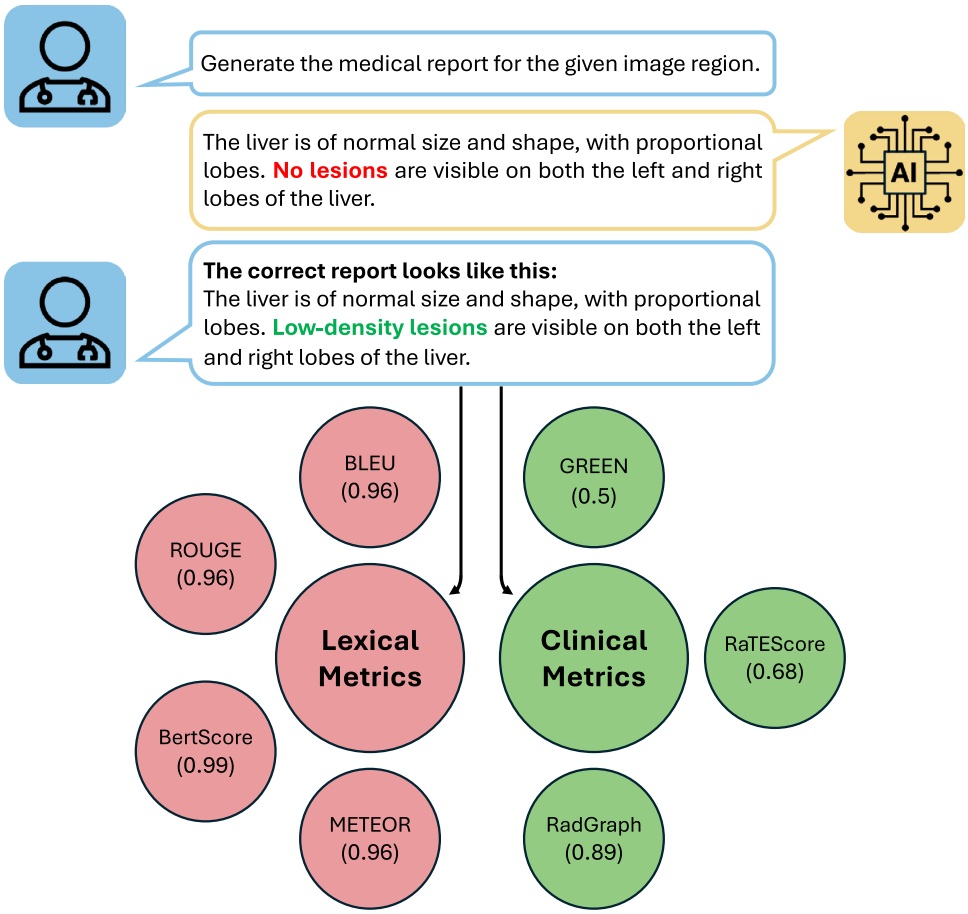

Fig. 4 Evaluation of generative AI in medicine: Lexical metrics from the general domain cannot completely capture the clinical correctness as they mainly cover text similarity. Clinically-relevant metrics like GREEN [155], RaTEScore [156], or RadGraph [157] also evaluate the clinical correctness

inconsistencies between generated and reference reports and produces human-interpretable feedback alongside a numerical score. This makes GREEN a practical and clinically grounded tool for both benchmarking and error analysis.

RadFact [72] uses a LLM to evaluate sentence-level factual consistency by comparing generated text to reference reports. In grounding tasks, it also incorporates image annotations to assess whether model predictions are supported by visual evidence.

In addition to text-based evaluation, clinical efficacy can also be assessed using standard classification metrics such as precision, recall, specificity, and F1-score. This is particularly relevant when using label-based datasets [8, 14], where a classifier is used to assign diagnostic labels to generated reports (e.g. CheXbert [162]), allowing comparison to ground truth annotations.

Evaluating image generation in medical AI requires considerations beyond standard image quality metrics like Fréchet Inception Distance [165] 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 [111]. 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 [166] address this by evaluating the 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 145 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 [167, 168].

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 [158] and ROUGE [159] are commonly used, they often fail to capture clinical relevance and factual accuracy. To address this, domain-specific metrics such as RadGraph [157], RaTEScore [156], and GREEN [155] 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 [153], 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 radiology-focused 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 [169–171]. 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, 135]. 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 [172, 173]. 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.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13534-025-00497-1.

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-Universität Erlangen-Nürnberg 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.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare no conflict of interest.

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.

References

1. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. 2022;28(9):1773–84.
2. Tayebi Arasteh S, Han T, Lotfinia M, Kuhl C, Kather JN, Truhn D, Nebelung S. Large language models streamline automated machine learning for clinical studies. Nat Commun. 2024;15(1):1603.
3. Cai X, Liu S, Han J, Yang L, Liu Z, Liu T. Chestxraybert: a pretrained language model for chest radiology report summarization. IEEE Trans Multimed. 2021;25:845–55.
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. 2023;15(6):65.
5. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with alphafold. Nature. 2021;596(7873):583–9.
6. Acosta JN, 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; arxiv:2412.12042.
7. Van Veen D, Van Uden C, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, Pareek A, Polacin M, Reis EP, Seehofnerová A, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30(4):1134–42.
8. Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C-Y, Mark RG, Horng S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019;6(1):317.
9. Huang S-C, Huo Z, Steinberg E, Chiang C-C, Lungren MP, Langlotz CP, Yeung S, Shah NH, Fries JA. Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis. 2023; arXiv preprint arXiv:2311.10798.
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. 2024;313(2): 233441.
11. Khader F, Müller-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. 2023;309(1): 230806.
12. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, et al. Gencode 2021. Nucleic Acids Res. 2021;49(D1):916–23.
13. Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, Liu Z, Chen X, Davison BD, Ren H, et al. A generalist vision-language foundation model for diverse biomedical tasks. Nat Med. 2024;30(11):1–13.
14. Hamamci IE, Er S, Almas F, Simsek AG, Esirgun SN, Dogan I, Dasdelen MF, 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. 2024; arXiv preprint arXiv:2403.17834 .
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. Adv Neural Inf Process Syst. 2024;8:36.
16. Özsoy 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. Springer; 2024. pp. 455–465.
17. Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, Chen E. A survey on multimodal large language models. 2023; arXiv preprint arXiv:2306.13549.
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. 2024; arXiv preprint arXiv:2408.01319.
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 Digit Med. 2022;5(1):171.

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. 2024; arXiv preprint arXiv:2404.03264.
21. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MD, Horsley T, Weeks L, et al. Prisma extension for scoping reviews (prisma-scr): checklist and explanation. Ann Intern Med. 2018;169(7):467–73.
22. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;6:372.
23. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016;5:1–10.
24. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, Eaton K, Riina HA, Laufer I, Punjabi P, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023;619(7969):357–62.
25. Vaswani A. Attention is all you need. Adv Neural Inf Process Syst. 2017.
26. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.
27. 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). IEEE; 2021. pp. 1–7.
28. 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. 2024; arXiv preprint arXiv:2402.10373.
29. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80.
30. 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 Digit Med. 2024;7(1):41.
31. Lee H, Phatale S, Mansoor H, Lu KR, Mesnard T, Ferret J, Bishop C, Hall E, Carbune V, Rastogi A. Rlaif: scaling reinforcement learning from human feedback with ai feedback; 2023.
32. 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. 2023; arXiv preprint arXiv:2305.15075.
33. Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, Fong R, Phillips C, Alexander K, Ashley E, et al. Almanac-retrieval-augmented language models for clinical medicine. NEJM AI. 2024;1(2):2300068.
34. 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. Brief Bioinform. 2022;23(6):409.
35. Wang H, Gao C, Dantona C, Hull B, Sun J. Drg-llama: tuning llama model to predict diagnosis-related group for hospitalized patients. Npj Digit Med. 2024;7(1):16.
36. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Costa AB, Flores MG, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5(1):194.
37. Johnson AE, Bulgarelli L, Pollard TJ. Deidentification of free-text medical records using pre-trained bidirectional transformers. In: Proceedings of the ACM conference on health, inference, and learning; 2020. pp. 214–221.
38. Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. NPJ Digit Med. 2024;7(1):102.
39. Mahendran D, McInnes BT. Extracting adverse drug events from clinical notes. AMIA Summits Transl Sci Proc. 2021;2021:420.
40. Lanfredi RB, Mukherjee P, Summers RM. Enhancing chest x-ray datasets with privacy-preserving large language models and multi-type annotations: a data-driven approach for improved classification. Med Image Anal. 2025;99: 103383.
41. Liu N, Hu Q, Xu H, Xu X, Chen M. Med-bert: a pretraining framework for medical records named entity recognition. IEEE Trans Ind Inf. 2021;18(8):5600–8.
42. Han T, Adams LC, Papaioannou J-M, Grundmann P, Oberhauser T, Löser A, Truhn D, Bressem KK. Medalpaca—an open-source collection of medical conversational ai models and training data. 2023; arXiv preprint arXiv:2304.08247.
43. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, Pagliardini M, Fan S, Köpf A, Mohtashami A, et al. Meditron-70b: scaling medical pretraining for large language models. 2023; arXiv preprint arXiv:2311.16079.
44. Qiu P, Wu C, Zhang X, Lin W, Wang H, Zhang Y, Wang Y, Xie W. Towards building multilingual language model for medicine. Nat Commun. 2024;15(1):8384.
45. Mu Y, Tizhoosh HR, Tayebi RM, Ross C, Sur M, Leber B, Campbell CJ. A bert model generates diagnostically relevant semantic embeddings from pathology synopses with active learning. Commun Med. 2021;1(1):11.
46. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. Pmc-llama: toward building open-source language models for medicine. J Am Med Inform Assoc. 2024;6:045.
47. Jia S, Bit S, Searls E, Claus LA, Fan P, Jasodanand VH, Lauber MV, Veerapaneni D, Wang WM, Au R, et al. Medpodgpt: a multilingual audio-augmented large language model for medical research and education. medRxiv 2024.
48. Yan A, McAuley J, Lu X, Du J, Chang EY, Gentili A, Hsu C-N. Radbert: adapting transformer-based language models to radiology. Radiol Artif Intell. 2022;4(4): 210258.
49. Schmidt RA, Seah JC, Cao K, Lim L, Lim W, Yeung J. Generative large language models for detection of speech recognition errors in radiology reports. Radiol Artif Intell. 2024;6(2): 230205.
50. Prihoda D, Maamary J, Waight A, Juan V, Fayadat-Dilman L, Svozil D, Bitton DA. Biophi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. In: MAbs, vol 14. Taylor & Francis; 2022. p. 2020203.
51. Schubach M, Maass T, Nazaretyan L, Röner 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 Res. 2024;52(D1):1143–54.
52. Ji Y, Zhou Z, Liu H, Davuluri RV. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. Bioinformatics. 2021;37(15):2112–20.
53. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, Mantineo H, Brydon EM, Zeng Z, Liu XS, et al. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616–24.
54. Hie BL, Shanker VR, Xu D, Bruun TU, Weidenbacher PA, Tang S, Wu W, Pak JE, Kim PS. Efficient evolution of human antibodies from general protein language models. Nat Biotechnol. 2024;42(2):275–83.
55. Rao RM, Liu J, Verkuil R, Meier J, Canny J, Abbeel P, Sercu T, Rives A. Msa transformer. In: International conference on machine learning. PMLR; 2021. pp. 8844–8856.
56. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, Olmos JL, Xiong C, Sun ZZ, Socher R, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol. 2023;41(8):1099–106.

57. Ferruz N, Schmidt S, Höcker B. Protgpt2 is a deep unsupervised language model for protein design. Nat Commun. 2022;13(1):4348.
58. 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 Trans Pattern Anal Mach Intell. 2021;44(10):7112–27.
59. 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. Nat Mach Intell. 2022;4(10):852–66.
60. Rathore AS, Choudhury S, Arora A, Tijare P, Raghava GP. Toxinpred 3.0: an improved method for predicting the toxicity of peptides. Comput Biol Med. 2024;179: 108926.
61. Chen J, Cai Z, Ji K, Wang X, Liu W, Wang R, Hou J, Wang B. Huatuogpt-o1, towards medical complex reasoning with llms. 2024; arXiv preprint arXiv:2412.18925.
62. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-T, Rocktäschel T, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv Neural Inf Process Syst. 2020;33:9459–74.
63. Arasteh ST, Lotfinia M, Bressem K, Siepmann R, Adams L, Ferber D, Kuhl C, Kather JN, Nebelung S, Truhn D. RadioRAG: factual large language models for enhanced diagnostics in radiology using online retrieval augmented generation 2024; arxiv:2407.15621.
64. Gilbert S, Kather JN, Hogan A. Augmented non-hallucinating large language models as medical information curators. NPJ Digit Med. 2024;7(1):100.
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 free-text radiology report databases. Eur Radiol. 2023;33(6):4228–36.
66. Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, et al. Mimic-iv, a freely accessible electronic health record dataset. Sci Data. 2023;10(1):1.
67. Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O. The eicu collaborative research database, a freely available multi-center database for critical care research. Sci Data. 2018;5(1):1–13.
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); 2020. pp. 9241–9250.
69. Uniprot: the universal protein knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):523–531.
70. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. Genbank. Nucleic Acids Res. 2012;41(D1):36–42.
71. Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, Jacob S, Madhavan S, Ketchum KA. The cptac data portal: a resource for cancer proteomics research. J Proteome Res. 2015;14(6):2707–13.
72. Bannur S, Bouzid K, Castro DC, Schwaighofer A, Bond-Taylor S, Ilse M, Pérez-García F, Salvatelli V, Sharma H, Meissen F, et al. Maira-2: grounded radiology report generation. 2024; arXiv preprint arXiv:2406.04449.
73. Pellegrini C, Özsoy E, Busam B, Navab N, Keicher M. Radialog: a large vision-language model for radiology report generation and conversational assistance. 2023; arXiv preprint arXiv:2311.18681.
74. Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nat Biomed Eng. 2022;6(12):1399–406.
75. Endo M, Krishnan R, Krishna V, Ng AY, Rajpurkar P. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In: Machine learning for health. PMLR 2021. pp. 209–219.
76. Radford A, Kim JW, 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. PMLR; 2021. pp. 8748–8763.
77. Blankemeier L, Cohen JP, Kumar A, Van Veen D, Gardezi SJS, Paschali M, Chen Z, Delbrouck J-B, Reis E, Truyts C, et al. Merlin: a vision language foundation model for 3d computed tomography. 2024; arXiv preprint arXiv:2406.06512.
78. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. Adv Neural Inf Process Syst. 2024;6:36.
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. 2024;1(3):2300138.
80. Zhou H-Y, Adithan S, Acosta JN, Topol EJ, Rajpurkar P. A generalist learner for multifaceted medical image interpretation. 2024; arXiv preprint arXiv:2405.07988.
81. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Towards generalist foundation model for radiology. 2023; arXiv preprint arXiv:2308.02463.
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. 2024. p. 2400640.
83. Boecking B, Usuyama N, Bannur S, Castro DC, 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. Springer; 2022. pp. 1–21.
84. Bannur S, Hyland S, Liu Q, Perez-Garcia F, Ilse M, Castro DC, 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; 2023. pp. 15016–15027.
85. Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. In: Machine learning for healthcare conference. PMLR; 2022. pp. 2–25.
86. Javed S, Mahmood A, Ganapathi II, Dharejo FA, 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; 2024. pp. 11450–11459.
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. 2024; arXiv preprint arXiv:2405.03162.
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). IEEE; 2024. pp. 8230–8234.
89. Luo Y, Shi M, Khan MO, Afzal MM, 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; 2024. pp. 12289–12301.
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; 2024. pp. 11398–11407.
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. PMLR; 2024. pp. 1493–1508.
92. Huang S-C, Shen L, Lungren MP, 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; 2021. pp. 3942–3951.
93. Zhang X, Wu C, Zhang Y, Xie W, Wang Y. Knowledge-enhanced visual-language pre-training on chest radiology images. Nat Commun. 2023;14(1):4542.
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. Nat Commun. 2024;15(1):7620.
95. Wang P, Zhang H, Yuan Y. Mcpl: multi-modal collaborative prompt learning for medical vision-language model. IEEE Trans Med Imaging 2024.
96. Codella NC, Jin Y, Jain S, Gu Y, Lee HH, Abacha AB, Santamaria-Pang A, Guyman W, Sangani N, Zhang S, et al. Medimageinsight: an open-source embedding model for general domain medical imaging. 2024; arXiv preprint arXiv:2410.06542.
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 Digit Med. 2023;6(1):226.
98. Moon JH, 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 J Biomed Health Inform. 2022;26(12):6070–80.
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. Nat Mach Intell. 2023;5(12):1447–57.
100. Tang X, Tran A, Tan J, Gerstein MB. Mollm: a unified language model for integrating biomedical text with 2d and 3d molecular representations. Bioinformatics. 2024;40(1):357–68.
101. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual-language foundation model for pathology image analysis using medical twitter. Nat Med. 2023;29(9):2307–16.
102. Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, Wong C, Gero Z, González J, Gu Y, et al. A whole-slide foundation model for digital pathology from real-world data. Nature. 2024;6:1–8.
103. Khattak MU, Kunhimon S, Naseer M, Khan S, Khan FS. Unimed-clip: towards a unified image-text pretraining paradigm for diverse medical imaging modalities. 2024; arXiv preprint arXiv:2412.10372.
104. Pellegrini C, Keicher M, Özsoy 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. Springer 2023. pp. 420–429.
105. Ran A, Liu H. Joint spatio-temporal features constrained self-supervised electrocardiogram representation learning. Biomed Eng Lett. 2024;14(2):209–20.
106. Kang Y, Yang G, Eom H, Han S, Baek S, Noh S, Shin Y, Park C. Gan-based patient information hiding for an ecg authentication system. Biomed Eng Lett. 2023;13(2):197–207.
107. 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. Nat Commun. 2024;15(1):5649.
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. 2024; arXiv preprint arXiv:2407.16684.
109. Tanida T, Müller 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; 2023. pp. 7433–7442.
110. 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 Trans Med Imaging; 2023.
111. Hamamci IE, Er S, Sekuboyina A, Simsar E, Tezcan A, Simsek AG, Esirgun SN, Almas F, Dogan I, Dasdelen MF, et al. Generatect: text-conditional generation of 3d chest ct volumes. 2023; arXiv preprint arXiv:2305.16037.
112. Bluethgen C, Chambon P, Delbrouck J-B, Sluijs R, Połacin M, Zambrano Chaves JM, Abraham TM, Purohit S, Langlotz CP, Chaudhari AS. A vision-language foundation model for the generation of realistic chest x-ray images. Nat Biomed Eng. 2024;6:1–13.
113. Khader F, Müller-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. Sci Rep. 2023;13(1):7303.
114. Jiang S, Zheng T, Zhang Y, Jin Y, Yuan L, Liu Z. Med-moe: mixture of domain-specific experts for lightweight medical vision-language models. 2024; arXiv preprint arXiv:2404.10237.
115. Alsharid M, Cai Y, Sharma H, Drukker L, Papageorghiou AT, Noble JA. Gaze-assisted automatic captioning of fetal ultrasound videos using three-way multi-modal deep neural networks. Med Image Anal. 2022;82: 102630.
116. Cui H, Mao L, Liang X, Zhang J, Ren H, Li Q, Li X, Yang C. Biomedical visual instruction tuning with clinician preference alignment. 2024; arXiv preprint arXiv:2406.13173.
117. Wang S, Zhao Z, Ouyang X, Wang Q, Shen D. Chatcad: interactive computer-aided diagnosis on medical image using large language models. 2023; arXiv preprint arXiv:2302.07257.
118. Chen Z, Varma M, Delbrouck J-B, Paschali M, Blankemeier L, Van Veen D, Valanarasu JMJ, Youssef A, Cohen JP, Reis EP, et al. Chexagent: towards a foundation model for chest x-ray interpretation. 2024; arXiv preprint arXiv:2401.12208.
119. 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; 2024. pp. 7995–8004.
120. 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 Digit Med. 2024;7(1):111.
121. Huh J, Park S, Lee JE, Ye JC. Improving medical speech-to-text accuracy with vision-language pre-training model. 2023; arXiv preprint arXiv:2303.00091.
122. Bai F, Du Y, Huang T, Meng MQ-H, Zhao B. M3d: advancing 3d medical image analysis with multi-modal large language models. 2024; arXiv preprint arXiv:2404.00578.
123. Sharma H, Salvatelli V, Srivastav S, Bouzid K, Bannur S, Castro DC, Ilse M, Bond-Taylor S, Ranjit MP, Falck F, et al. Maira-seg: enhancing radiology report generation with segmentation-aware multimodal large language models. 2024; arXiv preprint arXiv:2411.11362.
124. Moor M, Huang Q, Wu S, Yasunaga M, Dalmia Y, Leskovec J, Zakka C, Reis EP, Rajpurkar P. Med-flamingo: a multimodal medical few-shot learner. In: Machine learning for health (ML4H). PMLR; 2023. pp. 353–367.
125. Khare Y, Bagal V, Mathew M, Devi A, Priyakumar UD, Jawahar C. Mmbert: multimodal bert pretraining for improved medical vqa. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). IEEE; 2021. pp. 1033–1036.

126. Cao X, Liang K, Liao K-D, Gao T, Ye W, Chen J, Ding Z, Cao J, Rehg JM, Sun J. Medical video generation for disease progression simulation. 2024; arXiv preprint arXiv:2411.11943.
127. Lu MY, Chen B, Williamson DF, Chen RJ, Zhao M, Chow AK, Ikemura K, Kim A, Pouli D, Patel A, et al. A multimodal generative AI copilot for human pathology. Nature. 2024;634(8033):466–73.
128. 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; 2024. pp. 5182–5191.
129. Seyfioglu MS, Ikezogwo WO, Ghezloo F, Krishna R, Shapiro L. Quilt-llava: visual instruction tuning by extracting localized narratives from open-source histopathology videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2024. pp. 13183–13192.
130. Wang Z, Liu L, Wang L, Zhou L. R2gengpt: radiology report generation with frozen llms. Meta Radiol. 2023;1(3): 100033.
131. Luo L, Vairavamurthy J, Zhang X, Kumar A, Ter-Oganesyan RR, Schroff ST, Shilo D, Hossain R, Moritz M, Rajpurkar P. Rexplain: Translating radiology into patient-friendly video reports. 2024; arXiv preprint arXiv:2410.00441.
132. Bai L, Wang G, Islam M, Seenivasan L, Wang A, Ren H. Surgical-vqla++: adversarial contrastive learning for calibrated robust visual question-localized answering in robotic surgery. Inf Fusion. 2025;113: 102602.
133. Liu J, Zhang Y, Chen J-N, Xiao J, Lu Y, 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; 2023. pp. 21152–21164.
134. Wang Y, Dai Y, Jones C, Sair HI, Shen J, Loizou N, Hsu W-C, Imami MR, Jiao Z, Zhang PJ, 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.
135. 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; 2019, vol 33, pp. 590–597.
136. 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. 2024; arXiv preprint arXiv:2408.02900.
137. Gupta D, Attal K, Demner-Fushman D. A dataset for medical instructional video classification and question answering. Sci Data. 2023;10(1):158.
138. 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; 2024. pp. 22170–22183.
139. 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. Med Image Anal. 2020;66: 101797.
140. He X, Zhang Y, Mou L, Xing E, Xie P. Pathvqa: 30000+ questions for medical visual question answering. 2020; arXiv preprint arXiv:2003.10286.
141. Chen J, Ouyang R, Gao A, Chen S, Chen GH, Wang X, Zhang R, Cai Z, Ji K, Yu G, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. 2024; arXiv preprint arXiv:2406.19280.
142. Ikezogwo W, Seyfioglu S, Ghezloo F, Geva D, Sheikh Mohammed F, Anand PK, Krishna R, Shapiro L. Quilt-1m: one million image-text pairs for histopathology. Adv Neural Inf Process Syst. 2024;36:65.
143. Pellegrini C, Keicher M, Özsoy 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. Springer; 2023. pp. 409–419.
144. Liu B, Zhan L-M, Xu L, Ma L, Yang Y, Wu X-M. Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). IEEE; 2021. pp. 1650–1654.
145. Lau JJ, Gayen S, Ben Abacha A, Demner-Fushman D. A dataset of clinically generated visual questions and answers about radiology images. Sci Data. 2018;5(1):1–10.
146. codabench: MICCAI24 AMOS-MM: abdominal multimodal analysis challenge; 2024. Accessed 11 Apr 2024. https://www.codabench.org/competitions/3137/.
147. 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. Adv Neural Inf Process Syst. 2022;35:36722–32.
148. 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. Springer; 2024. pp. 124–134
149. Zhang X, Wu C, Zhao Z, Lei J, Zhang Y, Wang Y, Xie W. Radgenome-chest ct: a grounded vision-language dataset for chest ct analysis. 2024; arXiv preprint arXiv:2404.16754 .
150. Saha A, Harowicz MR, Grimm LJ, Kim CE, Ghate SV, Walsh R, Mazurowski MA. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. Br J Cancer. 2018;119(4):508–16.
151. Wagner P, Strodthoff N, Bousseljot R-D, Kreiseler D, Lunze FI, Samek W, Schaeffter T. Ptb-xl, a large publicly available electrocardiography dataset. Sci Data. 2020;7(1):1–15.
152. 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. 2023; arXiv preprint arXiv:2302.04611.
153. Zhang X, Zhou H-Y, Yang X, Banerjee O, Acosta JN, Miller J, Huang O, Rajpurkar P. Rexrank: a public leaderboard for ai-powered radiology report generation. 2024; arXiv preprint arXiv:2411.15122.
154. Ferber D, Wölflein G, Wiest IC, Ligero M, Sainath S, Ghaffari Laleh N, El Nahhas OS, Müller-Franzes G, Jäger D, Truhn D, et al. In-context learning enables multimodal large language models to classify cancer pathology images. Nat Commun. 2024;15(1):10104.
155. Ostmeier S, Xu J, Chen Z, Varma M, Blankemeier L, Bluethgen C, Michalson AE, Moseley M, Langlotz C, Chaudhari AS, et al. Green: generative radiology report evaluation and error notation. 2024; arXiv preprint arXiv:2405.03595.
156. Zhao W, Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Ratescore: a metric for radiology report generation. medRxiv; 2024;2024-06.
157. Yu F, Endo M, Krishnan R, Pan I, Tsai A, Reis EP, Fonseca EKUN, Lee HMH, Abad ZSH, Ng AY, et al. Evaluating progress in automatic chest x-ray radiology report generation. Patterns. 2023;4(9):63.
158. 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; 2002. pp. 311–318.
159. Lin C-Y. Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out; 2004. pp. 74–81.
160. 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; 2005. pp. 65–72.

161. Huang A, Banerjee O, Wu K, Reis EP, Rajpurkar P. Fineradscore: a radiology report line-by-line evaluation technique generating corrections with severity scores. 2024; arXiv preprint arXiv:2405.20613.
162. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. 2020; arXiv preprint arXiv:2004.09167.
163. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, Zhou H-Y, Cai ZR, Van Allen EM, Kim D, Daneshjou R, Rajpurkar P. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med. 2025. https://doi.org/10.1038/s41591-024-03328-5.
164. 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 Med. 2024;7(1):124.
165. 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. Adv Neural Inf Process Syst. 2017;30:65.
166. Banerjee O, Saenz A, Wu K, Clements W, Zia A, Buensalido D, Kavnoudias H, Abi-Ghanem AS, Ghawi NE, Luna C, et al. Rexamine-global: A framework for uncovering inconsistencies in radiology report generation metrics. In: Biocomputing 2025: proceedings of the pacific symposium. World Scientific; 2024. pp. 185–198
167. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using chatgpt in health care. J Med Internet Res. 2023;25:48009.
168. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health. 2023;5(6):333–5.
169. 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. 2024; arXiv preprint arXiv:2411.18730.
170. 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. 2024; arXiv preprint arXiv:2412.01233.
171. 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. Nat Med. 2024;30(9):2613–22.
172. 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. 2023; arXiv preprint arXiv:2305.20050.
173. Chua M, Kim D, Choi J, Lee NG, Deshpande V, Schwab J, Lev MH, Gonzalez RG, Gee MS, Do S. Tackling prediction uncertainty in machine learning for healthcare. Nat Biomed Eng. 2023;7(6):711–8.