ICJK

Journal of Artificial Intelligence in Bioinformatics
http://dx.doi.org/10.62762/JAIB.2026.839612

REVIEW ARTICLE

The Evolution of Artificial Intelligence in Healthcare: From LLMs to Multimodal and Agentic Systems

Muhammad Bilal[1], **Hafsa Gulzar**[2,*], **Sibtain Ahmed**[1] **and Muhammad Saqlain Aslam**[3]

[1] Department of Computer Science, Iowa State University, Ames, IA 50011, United States
[2] Department of Computer Science, Donghua University, Shanghai, China
[3] AI Research Center, Hon Hai Research Institute, Taipei 114699, Taiwan

Abstract

Artificial intelligence (AI) in healthcare has moved from rule-based expert systems to statistical and deep learning methods and, more recently, to large language models (LLMs), multimodal models, and agentic AI systems. This review treats these developments as connected stages rather than separate technologies. We conducted a focused review of foundational and recent literature, giving priority to studies that report clinical applications and quantitative evaluation. Representative applications are summarised in medical question answering, diagnosis, radiology, pathology, chronic disease management, clinical decision support, and healthcare administration. LLMs have improved medical language understanding and reasoning; multimodal models extend these abilities by combining clinical text, medical images, and structured data; and agentic systems add planning, tool use, memory, and multi-step workflow execution, with multi-agent designs now appearing in diagnosis, oncology, and pharmacovigilance. Routine clinical adoption is still limited by unreliable outputs, data privacy, lack of prospective real-world validation, unsettled regulation, infrastructure gaps, and the continued need for clinician oversight. We argue that each stage addresses limitations of the previous one while introducing new risks, most notably error propagation across agents and the consequences of autonomous actions.

Keywords: agentic AI, large language models, multimodal models, healthcare artificial intelligence, clinical decision support.

1 Introduction

Artificial intelligence has been part of healthcare for several decades, beginning with rule-based expert systems such as MYCIN and INTERNIST-1 in the 1970s. MYCIN was developed at Stanford to diagnose bacterial infections and recommend antibiotic treatment, using several hundred hand-crafted production rules together with a certainty-factor model for handling uncertainty [1, 2]. INTERNIST-1, built at the University of Pittsburgh, supported differential diagnosis in general internal medicine by

Submitted: 24 July 2026
Revised: 18 September 2026
Accepted: 05 October 2026
Published: 27 November 2026

Vol. 3, **No.** 1, 2027.
10.62762/JAIB.2026.839612

***Corresponding author:**
✉ Hafsa Gulzar
hafsa@mail.dhu.edu.cn

Citation
Bilal, M., Gulzar, H., Ahmed, S., & Aslam, M. S. (2026). The Evolution of Artificial Intelligence in Healthcare: From LLMs to Multimodal and Agentic Systems. *Journal of Artificial Intelligence in Bioinformatics*, 3(1), 1–12.

© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

linking diseases to clinical findings through manually encoded disease-finding relationships [3]. These systems showed that expert clinical reasoning could, in principle, be captured in software. They were, however, hard to maintain as medical knowledge grew, because every new rule had to be written by hand and the systems could not learn from data.

The field later moved to statistical machine learning and then to deep learning, in which models learn patterns directly from data. Deep convolutional neural networks matched dermatologists in classifying skin cancer from photographs [4] and detected diabetic retinopathy from fundus images at a level comparable to ophthalmologists [5]. These results established deep learning as a practical tool for diagnostic support, although each model was still trained for a single, narrowly defined task.

The transformer architecture [6] replaced recurrence with self-attention, which relates every position of an input sequence to every other position, and it became the basis of large language models (LLMs). Trained on very large text corpora, LLMs reached passing-level performance on the United States Medical Licensing Examination without task-specific training [7], and medically adapted models such as Med-PaLM 2 approached clinician-level accuracy on examination-style questions [8].

Two further developments followed. The first is the multimodal LLM. Models such as Med-Flamingo [9] and LLaVA-Med [10] process medical images and clinical text within one model, which gives outputs that draw on more of the clinical context than text-only or image-only systems. The second is agentic AI, which is increasingly treated as a distinct class of system. A conventional LLM predicts the next token in response to a single prompt. An agentic system, in contrast, pursues a goal: it can plan, call external tools and application programming interfaces (APIs), coordinate with other task-specific agents, and adapt a multi-step clinical workflow with limited human intervention.

This review follows that progression. Section 2 examines LLMs and their performance on medical reasoning tasks. Section 3 discusses multimodal LLMs, which combine text with medical images, electronic health records (EHRs), and other clinical data. Section 4 turns to agentic AI, covering its architecture and published applications. Section 5 discusses limitations, and Section 6 concludes. Throughout, we consider applications, limitations, and relevance to biomedical data analysis and clinical informatics.

Relevant literature was identified by searching PubMed, Google Scholar, arXiv, and IEEE Xplore with combinations of the terms "large language models", "multimodal", "agentic AI", "clinical decision support", and "healthcare". Because agentic AI is moving quickly, recent arXiv preprints and conference proceedings were included alongside peer-reviewed journal articles. Priority was given to studies reporting quantitative clinical evaluation (e.g., comparisons with physicians or patient actors, benchmark accuracy) over conceptual or opinion pieces.

Previous reviews have mostly examined LLMs, multimodal models, and agentic AI separately. Here we treat them as successive, connected stages and examine how each stage addresses limitations of the one before while creating new problems for clinical

Table 1. Representative text-based large language models in healthcare.

Model	Base architecture	Main healthcare tasks
BioGPT [11]	GPT-style decoder-only Transformer	Biomedical QA, relation extraction, document classification, biomedical text generation
Med-PaLM [12]	Flan-PaLM with instruction prompt tuning	Medical examination QA, consumer health QA, long-form medical answers
Med-PaLM 2 [8]	PaLM 2 with medical domain fine-tuning	Medical examination QA, specialist questions, consumer health QA, clinical reasoning
GatorTronGPT [13]	GPT-3 architecture; up to 20B parameters	Clinical text generation, biomedical NLP, relation extraction, synthetic clinical text generation
PMC-LLaMA [14]	LLaMA-based; 7B and 13B variants	Medical QA, reasoning, dialogue, medical knowledge retrieval
Clinical Camel [15]	LLaMA/Llama 2 with parameter-efficient fine-tuning	Medical QA, clinical conversation, case reasoning, note generation
MEDITRON [16]	Llama 2; 7B and 70B variants	Medical QA, clinical reasoning, decision-support questions
BioMistral [17]	Mistral-7B	Medical QA, biomedical reasoning, multilingual medical QA

translation, safety, governance, and deployment.

2 Large Language Models in Healthcare

Before general-purpose LLMs became widely available, researchers trained language models on biomedical text. BioBERT, one of the earliest and most influential of these, was obtained by continuing the pre-training of BERT on PubMed abstracts and full-text articles [18]. After task-specific fine-tuning it outperformed general BERT on biomedical named entity recognition, relation extraction, and question answering, showing that domain-specific training helped even when the architecture was unchanged [18]. ClinicalBERT was trained on clinical notes from the MIMIC-III intensive care database, which exposed it to the writing style, abbreviations, and structure of hospital documentation rather than published papers [19]. These models were not capable of open-ended conversation or step-by-step reasoning. Each was fine-tuned for one task at a time, such as biomedical named entity recognition and relation extraction [20] or emotion classification in psychotherapy text [21]. Even so, they made an early point clear: clinical and biomedical text has enough specialised structure that general-purpose language models benefit from additional domain-specific training before use in healthcare.

2.1 General-purpose LLMs and the Shift Toward Medical Reasoning

The next shift came when general-purpose LLMs, trained on much larger and more varied text collections, began to perform well on medical tasks without biomedical training. An early demonstration tested ChatGPT on the United States Medical Licensing Examination (USMLE), the three-step examination required for medical licensure in the United States [7]. ChatGPT performed at or near the 60% passing threshold on all three steps without specialised training or reinforcement, and its explanations showed reasonable internal consistency [7]. The result suggested that a model trained mainly on general internet text had already absorbed a meaningful amount of medical knowledge, and it prompted further work on how far such reasoning could go and how it could be made reliable enough for clinical use.

2.2 Medically Specialised LLMs

Researchers then began adapting LLMs specifically for medicine. Google Research introduced Med-PaLM, built on PaLM, and evaluated it with MultiMedQA, a benchmark combining medical examination questions, medical research questions, and consumer health questions [12]. Flan-PaLM reached 67.6% accuracy on USMLE-style MedQA questions, and Med-PaLM, obtained from Flan-PaLM by instruction prompt tuning (a parameter-efficient method that adapts a large model to a new domain with a small number of examples), gave consumer health answers that compared favourably with those of clinicians under structured human evaluation [12]. Med-PaLM 2 combined a stronger base model with medical fine-tuning and new prompting strategies. It reached 86.5% on MedQA, more than 19 percentage points above Med-PaLM, and physicians preferred its answers to those of other physicians on eight of nine clinical-utility axes [8]. Within the limits of a structured evaluation, these results showed that a carefully adapted LLM could produce answers that doctors judged as good as, or better than, those written by their peers.

As confidence in these models grew, researchers tested them on everyday clinical work rather than examination questions. Clinical documentation has received particular attention because clinicians spend a large share of their time reading and writing notes. A study that adapted eight LLMs to four summarisation tasks (radiology reports, patient questions, progress notes, and doctor-patient dialogue) found that, in most cases, ten physician reviewers judged the best-adapted models' summaries to be equivalent or superior to those written by medical experts [22]. Adapted LLMs may therefore reduce the documentation burden that contributes to clinician workload and burnout [23], provided a qualified professional checks the output before it enters the medical record. LLMs have also been tested for patient-facing question answering, medical education, and first-pass organisation of information in the EHR. Across these uses the literature shows a consistent pattern: LLMs do well on language generation, summarisation, and structured question answering, but are less reliable on precise numerical reasoning, up-to-date factual knowledge, and the full clinical context of an individual patient. This is one reason researchers looked for ways to extend the models beyond plain text. Table 1 lists representative text-based LLMs developed for healthcare, with their base architectures and main clinical or biomedical tasks.

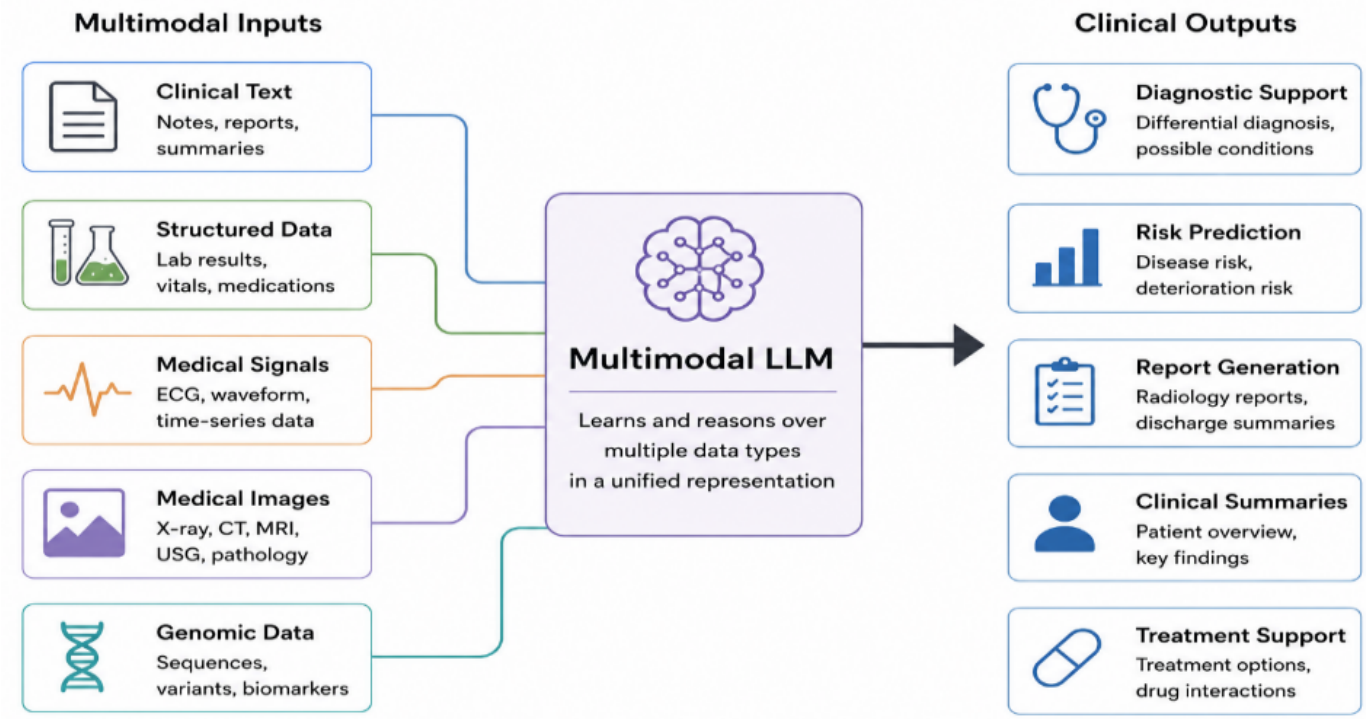

Figure 1. Conceptual overview of a multimodal large language model integrating heterogeneous healthcare data to support clinical decision-making and patient care.

3 Multimodal Models in Healthcare

3.1 Why Healthcare Needs More Than Text

The LLMs described above work only with text, whereas clinical decisions rarely depend on text alone. A single decision may draw on written notes, laboratory values, images such as radiographs or skin photographs, and sometimes genomic data; a text-only model misses most of this. Multimodal LLMs were developed to accept and reason over more than one type of input at once, most often images together with text [24, 25]. Instead of treating an image and a note as unrelated inputs, these models learn a shared representation that allows them to describe an image, answer a question about it, or combine an image and a note to reach a conclusion that neither input supports alone.

3.2 Vision-Language Foundation Models in Medicine

Med-Flamingo, one of the earlier multimodal medical models, was obtained by continuing the training of an existing vision-language model (OpenFlamingo) on medical textbook pages that interleave images and text, together with image-caption pairs from the published literature [9]. It was designed as a few-shot learner that can adapt to a new task from a handful of examples, which matters in medicine because labelled data for rare conditions are scarce. Its developers reported that it improved generative medical visual question answering (VQA) by up to 20% in clinicians' ratings [9]. LLaVA-Med was trained in two stages: it first aligned biomedical vocabulary with images using figure-caption pairs from PubMed Central, and then learned to hold open-ended conversations about biomedical images using instruction data generated with GPT-4 [10]. It outperformed earlier supervised models on standard biomedical VQA datasets while requiring far less training time than building such a system from scratch [10]. Both models showed that medical multimodal AI does not have to be trained from the ground up: general vision-language architectures can be adapted efficiently with image-text pairs already available in the literature.

3.3 Toward Generalist Multimodal Systems

Med-PaLM M was described by its developers as a proof of concept for a generalist biomedical AI system. Instead of a separate model for each data type, it encodes and interprets clinical language, medical images (chest radiographs, dermatology photographs, mammograms), and genomic data with a single set of weights. To evaluate it, the authors built MultiMedBench, a benchmark of 14 tasks including medical question answering, image interpretation, radiology report generation and summarisation, and genomic variant calling [26]. Med-PaLM M was

Table 2. Representative multimodal large language models in healthcare.

Model	LLM backbone	Main healthcare tasks
LLaVA-Med [10]	LLaVA with a Vicuna (LLaMA-based) language model	Medical VQA, image explanation, biomedical visual conversation
Med-Flamingo [9]	OpenFlamingo-9B	Few-shot medical VQA, visual medical-examination questions, rationale generation
Med-PaLM M [26]	PaLM-E (PaLM with a ViT image encoder)	Medical QA, medical image classification, radiology report generation and summarisation, genomic variant calling
RadFM [28]	Autoregressive medical LLM connected to visual encoders	Medical VQA, disease diagnosis, report generation, modality recognition, diagnostic rationale
PathChat [27]	Llama 2 (13B) with a pathology-specific UNI vision encoder	Pathology QA, morphological description, differential diagnosis, visual dialogue
M3D-LaMed [29]	LLM connected to a 3D medical-image encoder	3D VQA, report generation, retrieval, localisation, segmentation
MAIRA-2 [30]	Vicuna-7B with Rad-DINO image encoder	Grounded radiology report generation, finding localisation
Med-Gemini [31]	Gemini-based medical model family	Medical QA, report generation, image classification, VQA, genomic risk prediction

competitive with or better than task-specific models, and in a side-by-side evaluation clinicians preferred its chest radiograph reports over those written by radiologists in up to 40.50% of cases [26]. This does not mean the model outperformed radiologists overall, since human reports were still preferred more often; it does show that one generalist model can produce clinically usable output across very different tasks, which earlier narrow systems could not. Figure 1 gives a conceptual overview of a multimodal LLM in healthcare, in which heterogeneous clinical inputs are mapped into a shared representation and used to produce outputs such as diagnostic support, risk prediction, and report generation.

3.4 Applications of Multimodal Models in Clinical Practice

Radiology is the most consistently studied application of medical multimodal models: models generate or summarise written reports directly from images, as in Med-PaLM M and the radiology tasks of its benchmark [26]. Med-PaLM M also interprets dermatology photographs [26], and in pathology, PathChat reads a tissue image together with a text prompt to describe morphology and suggest differential diagnoses [27]. In each case, combining modalities grounds the output more firmly in the clinical situation than text or images alone, and later agentic systems build on this grounding when they plan and carry out multi-step tasks. Table 2 lists representative multimodal LLMs in healthcare, with their language backbones and main tasks.

4 Agentic AI and Healthcare

Earlier clinical AI tools either followed fixed rules written by experts or, in the case of machine learning and deep learning, made predictions such as flagging an abnormal scan or estimating disease risk. Neither type of system can plan, remember previous interactions, or act independently towards a larger goal.

Agentic AI tries to close this gap. In engineering, an "agent" has long been defined as a system that perceives its environment, reasons about what it observes, and acts autonomously to achieve a goal [32]. Agentic AI combines this older idea with the capabilities of LLMs [33].

Compared with text-based LLMs, multimodal models represent the clinical context more completely by combining text with images, structured records, and other biomedical data, which lets them handle tasks that language alone cannot. Their performance, however, depends heavily on the quality and alignment of the input modalities. Agentic systems build on these multimodal capabilities by adding planning, tool use, memory, and iterative decision-making, moving from language-focused reasoning towards more integrated and autonomous clinical systems.

4.1 Emergence of Agentic AI in Modern Healthcare

Intelligent agents are not new to healthcare. Researchers had already built agent-based systems for narrow tasks, for example belief-desire-intention (BDI) agents that track trauma patients and alert clinicians when a patient's condition changes or needs immediate attention [34]. These early systems were agentic only in a limited sense, because their reasoning was hand-coded and their range of actions was small.

Capable LLMs changed this. They read and generate natural language, reason step by step, and can call

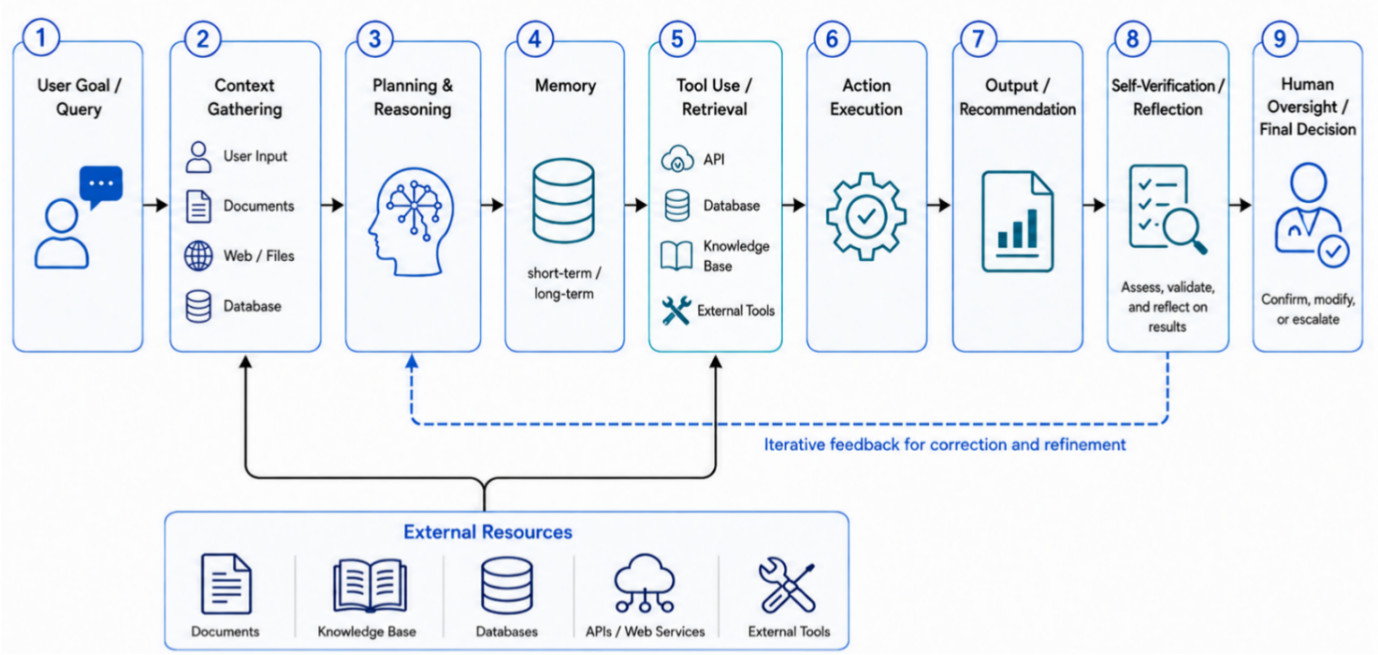

Figure 2. Conceptual pipeline of an agentic AI system, showing the interaction between reasoning, memory, external knowledge sources, tool use, action execution, self-reflection, and human-in-the-loop validation.

external tools. Once an LLM is allowed to plan its own steps, use tools, remember earlier exchanges, and check its own output, it behaves as an agent rather than a chatbot. A review of 153 peer-reviewed publications from 2021 to 2026 reported a sharp rise in publications on agentic medical AI, with 101 of them appearing in 2025 following the wider release of advanced LLMs [35]. The move from reactive to agentic AI in healthcare is therefore recent but fast.

Several reviews published in 2025 and 2026 have mapped this area. Srinivasu et al. [36] described the working mechanism of agentic AI in healthcare, including its components and enabling technologies, and illustrated it with case studies in ambient assisted living and remote robotic surgery. A scoping review that searched PubMed, Web of Science, arXiv, and medRxiv found that 36 of the 43 included studies were published in 2025, which again shows how new most of this work is [37]. Another scoping review, registered on the Open Science Framework and covering PubMed, Embase, Cochrane, Scopus, and Google Scholar, separated agentic AI systems, defined by autonomous operation, goal-directed behaviour, and the initiation of actions or tool calls, from simpler tools that are loosely called agents [38].

4.2 Architecture of Agentic AI

A similar architecture appears across most agentic healthcare systems described in the literature [39, 40], and we outline its components before turning to applications. The first is perception, the ability to take in information from the environment. In healthcare this includes reading health records, interpreting medical images, and processing laboratory results; more advanced systems combine several of these inputs [41]. Figure 2 shows the general workflow of an agentic clinical LLM pipeline, from the query and context gathering through planning, memory access, tool use, and output generation, followed by self-verification, human oversight, and iterative refinement supported by external resources such as documents, databases, and APIs.

The agent then plans and reasons, breaking a broad goal into sub-tasks and deciding the order in which to carry them out. Simpler systems do this in a single pass, whereas more advanced agents iterate: they generate an initial response, review and refine it, and then produce a final answer [39].

Agentic systems usually also include memory. Depending on the architecture, memory may be short-term, limited to the current conversation, or long-term. Long-term memory matters in chronic disease management, where treatment decisions depend on a patient's history across many visits and on how the disease has progressed.

Another defining feature is tool use: an agent can call external resources such as drug-interaction databases, clinical guidelines, or hospital scheduling systems instead of relying only on what it learned during

Table 3. Representative agentic AI systems in healthcare.

Agent	Domain	Description
EHRAgent [42]	Electronic health records	Breaks complex EHR questions into smaller steps and generates and executes code; retrieves information from multiple EHR tables and uses execution errors to revise its code in an agentic loop
MMedAgent [41]	Multimodal medical analysis	Selects and invokes specialised medical tools for radiology, pathology, and other data; integrates six tools covering seven tasks and five modalities
MDAgents [43]	Medical diagnosis	Assesses the complexity of a medical problem, chooses a single-agent or multi-agent collaboration structure, and recruits specialist agents
MALADE [44]	Pharmacovigilance and adverse drug-event detection	Coordinates multiple LLM agents with retrieval-augmented generation over drug labels, medical literature, and OpenFDA resources
ClinicalAgent [45]	Clinical trials	Uses planning, enrolment, safety, efficacy, and reasoning agents
Healthcare Agent [46]	Online medical consultation	Contains dialogue, planning, safety, memory, and report-generation components
Autonomous oncology agent [47]	Precision oncology	Selects and operates pathology, radiology, genomic, and web-search tools; analyses histopathology and radiological images
MIRA [48]	EHR-integrated diagnosis	Takes patient histories; orders and interprets laboratory, microbiology, and imaging tests

training [40].

Finally, the architecture includes action and self-correction. The agent carries out the planned task by producing a concrete output, such as a clinical note, a message, an order, or an answer to a clinician's question. Many current systems add a self-correction step in which the agent reviews its own output for errors and inconsistencies before it is shown to a human reviewer; iterative review of this kind is one way to improve reliability [39, 40]. Table 3 summarises representative agentic AI systems in healthcare, grouped by domain.

4.3 Applications of Agentic AI in Healthcare

The published applications fall into several groups, described below.

4.3.1 Conversational diagnostic agents

AMIE (Articulate Medical Intelligence Explorer), developed by Google Research and Google DeepMind, is one of the most closely watched examples. It was built to hold diagnostic conversations with patients, taking a history before reaching a diagnosis as a doctor would [49]. In a randomised, double-blind study modelled on an Objective Structured Clinical Examination (OSCE), AMIE was compared with 20 primary care physicians across 159 case scenarios from Canada, the United Kingdom, and India. AMIE achieved higher diagnostic accuracy and was rated better on 30 of 32 axes by specialist physicians and on 25 of 26 axes by patient actors [49]. The evaluation has clear constraints: it used scripted patient actors rather than real encounters and relied on a text-chat interface, which the authors themselves noted was unfamiliar to the participating physicians and not representative of routine practice.

AMIE was later extended to interpret images and documents, such as skin photographs, electrocardiograms, and clinical documents, during a conversation [50]. A further version splits the system into two cooperating agents: a dialogue agent that talks with the patient and a management agent that reasons over long clinical-guideline documents, so that the system can support disease management over several visits rather than a single consultation [51].

4.3.2 Multi-agent collaboration

In multi-agent designs, several AI agents work together, each acting like a specialist consulting on a difficult case. MedAgents assigns different medical roles to LLM-based agents, which discuss a case and reach a joint answer through structured exchange; this improved zero-shot medical reasoning compared with a single model [52].

MDAgents makes the collaboration structure adjustable. It first estimates how complex a medical question is and then decides whether one agent or a group should answer it, much as clinical teams decide whether a case needs one opinion or a multidisciplinary discussion [43]. Across ten medical benchmarks, it gave the best performance on seven, outperforming both solo and fixed-group configurations [43].

Multi-agent designs have also been built for specific clinical domains. CardAIc-Agents supports cardiac

care, with agents that retrieve guidelines, interpret cardiac test results, and adjust recommendations to individual risk factors [53]. Other work has evaluated multi-agent LLM systems that recommend safer drug therapies for patients with several conditions, a relevant question given how common polypharmacy is among older adults [54].

4.3.3 Tool-using and specialty agents

Some agentic systems rely on specialised tools rather than conversation. MMedAgent learns to choose the appropriate medical tool for a task, for example a segmentation model for a scan or a classification model for a skin lesion, instead of solving every problem with one general model [41]. In radiology, multi-agent systems draft, check, and refine report impressions to improve the quality of the generated impressions [55]. In pathology, PathFinder coordinates triage, navigation, description, and diagnosis agents that examine a whole-slide image region by region and combine their findings into a diagnosis, mirroring how a pathologist works through a slide [56].

4.3.4 Mental health and administrative support

Agentic ideas have also reached mental health care, where continuing, personalised support matters as much as a single assessment. Mental health appears among the clinical disciplines covered by recent scoping work on AI agents in healthcare [37]. Because these tools serve vulnerable patients, human oversight by a therapist or counsellor is generally considered essential, with the agent intended to extend the reach of scarce professionals rather than to replace them [35].

Not all agentic work concerns diagnosis or treatment. A considerable share addresses administrative burden, for instance drafting clinical notes, managing appointment scheduling, or summarising a patient's chart before a visit [35, 37]. These operational uses carry lower clinical risk than diagnostic ones, because errors in scheduling or note drafting are usually easier to detect and correct, although a human is still expected to check the output [38].

The limitations discussed below differ in scope. Some apply to LLM-based systems in general; others are specific to agentic architectures, in particular the propagation and amplification of errors across multi-agent workflows and the safety risks that arise when a system moves from recommending to acting [38].

5 Limitations

5.1 Hallucination and Reliability

LLMs tend to produce fluent, confident statements that are factually wrong, a behaviour known as hallucination. In healthcare this is serious because a well-written but incorrect answer can be harder for a busy clinician to catch than an obviously poor one, and an undetected error can affect diagnosis, treatment advice, or medication guidance [57, 58]. This risk does not disappear when an LLM is embedded in an agentic system; it can compound. In a multi-agent workflow, an error introduced at one step (a misread laboratory value, a wrongly recalled piece of patient context, or a flawed intermediate conclusion) can be passed on and amplified by downstream agents that accept the earlier output without checking it, so that one mistake affects the whole workflow rather than a single response [38].

5.2 Data Privacy and Security

Medical text and images routinely contain sensitive, identifiable information, and LLMs trained on such data can memorise and later reproduce protected details [58]. Complying with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States or the General Data Protection Regulation (GDPR) in the European Union requires robust de-identification before data are used for training or inference, yet current rule-based and machine learning de-identification methods perform unevenly across hospitals, specialties, and documentation formats, so residual identifiers can slip through [57, 58]. Privacy-preserving methods such as federated learning, in which a model learns from data that never leave each hospital's servers, and differential privacy, which adds calibrated statistical noise to protect individual records, are promising. Both, however, require substantial computing resources and specialised engineering expertise to deploy at scale, which has limited their adoption [59, 60].

5.3 Autonomy, Speed, and Safety Trade-offs

Agentic systems plan several steps, call external tools, and revise their own output, which adds latency and unpredictability compared with a single-pass model. In tasks that need both speed and precision, such as titrating medication doses, planning radiation therapy, or flagging an acute change on a monitor, this delay and variability may outweigh the benefit of multi-step reasoning, especially when simpler, validated automation with a longer safety record is available [38]. How well LLM agents perform in

realistic, time-sensitive clinical settings is also hard to judge, because current evaluations rarely reproduce the conditions of real clinical work [61]. More generally, giving a system authority to act, rather than only to generate text for a human to review, raises the stakes of any single failure, because an autonomous action such as ordering a test, changing a recommendation, or sending a message can have consequences before a human can intervene.

5.4 Infrastructure and Equity of Access

Deploying LLM- or agent-based tools requires integration with the technical and organisational reality of existing health systems. Many hospitals still run older EHR systems that were never designed to connect with modern AI tools, so adoption requires investment in integration, workflow redesign, and staff training. These costs are unevenly distributed. Lower-resource settings face additional barriers from limited computing infrastructure and the energy demands of running large models, raising a legitimate concern that the benefits of agentic AI will accrue mainly to well-resourced institutions [38].

Beyond these specific risks, clinical translation of multimodal and agentic systems is held back by the lack of standardised validation frameworks, unresolved regulatory and governance mechanisms for autonomous clinical actions, and continued reliance on benchmark evaluations rather than prospective evidence from real clinical settings.

6 Conclusion

This review has traced the path of AI in healthcare from early rule-based expert systems to the LLMs, multimodal systems, and agentic AI that define the field today. LLMs showed that a substantial amount of medical knowledge can be captured through broad training, and medically specialised variants such as Med-PaLM 2 reached accuracy on standardised medical questions that clinicians judged favourably. Multimodal models extended this reasoning beyond text, allowing one system to interpret images, structured records, and clinical language together, which highlights the growing role of bioinformatics in integrating heterogeneous biomedical data for clinical analysis and decision support. Agentic AI is the most recent step, building on both to support planning, tool use, and multi-step clinical workflows rather than single-turn responses, as shown by systems such as AMIE, MedAgents, and MDAgents.

The evidence reviewed here also shows that this progress has not yet translated into safe, unsupervised clinical deployment. Concerns about data privacy, hallucinated or unreliable output, unclear accountability, and poor fit with existing hospital systems recur across the literature, and they apply with particular force to agentic systems, in which errors can propagate along a chain of tool calls or cooperating agents. Nearly every system described here, from Med-PaLM 2 to AMIE and MDAgents, has been evaluated only in research, simulated, or pilot settings, not in fully autonomous routine practice.

Agentic AI in healthcare is therefore best seen as a promising but early-stage technology. Its ability to plan, retrieve information, and coordinate tools offers a plausible route to more integrated clinical support, but realising it will depend on further work on reliability, transparent regulation, and infrastructure that lets these systems fit into existing clinical workflows rather than sit beside them. Studies that compare agentic systems directly with current standards of care, in real clinical settings and over longer periods, are needed before these systems are given a larger share of clinical responsibility.

Data Availability Statement

Not applicable.

Funding

This work was supported without any funding.

Conflicts of Interest

Muhammad Saqlain Aslam is affiliated with Hon Hai Research Institute, the corporate research institute of Hon Hai Precision Industry Co., Ltd. This affiliation had no role in the design of the review, the selection or interpretation of the literature, or the decision to publish. Muhammad Saqlain Aslam also served as an Associate Editor of the Journal of Artificial Intelligence in Bioinformatics at the time of manuscript submission. To ensure the integrity of the peer-review process, he was not involved in the editorial handling, peer review, or decision-making process for this manuscript, which was handled independently by another editor. The other authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project* (*the Addison-Wesley series in artificial intelligence*). Addison-Wesley Longman Publishing Co., Inc. https://dl.acm.org/doi/abs/10.5555/1096485

[2] Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences, 23*(3-4), 351-379. [CrossRef]

[3] Miller, R. A., Pople Jr, H. E., & Myers, J. D. (1985). INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. In *Computer-assisted medical decision making* (pp. 139-158). New York, NY: Springer New York. [CrossRef]

[4] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115-118. [CrossRef]

[5] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA, 316*(22), 2402-2410. [CrossRef]

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

[7] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digital Health, 2*(2), e0000198. [CrossRef]

[8] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., ... & Natarajan, V. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine, 31*(3), 943-950. [CrossRef]

[9] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., ... & Rajpurkar, P. (2023, December). Med-Flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)* (pp. 353-367). PMLR. https://proceedings.mlr.press/v225/moor23a.html

[10] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., ... & Gao, J. (2023). LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems, 36*, 28541-28564. [CrossRef]

[11] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics, 23*(6), bbac409. [CrossRef]

[12] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172-180. [CrossRef]

[13] Peng, C., Yang, X., Chen, A., Smith, K. E., PourNejatian, N., Costa, A. B., ... & Wu, Y. (2023). A study of generative large language model for medical research and healthcare. *npj Digital Medicine, 6*(1), 210. [CrossRef]

[14] Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., & Wang, Y. (2024). PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association, 31*(9), 1833-1843. [CrossRef]

[15] Toma, A., Lawler, P. R., Ba, J., Krishnan, R. G., Rubin, B. B., & Wang, B. (2023). Clinical Camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*. [CrossRef]

[16] Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., ... & Bosselut, A. (2023). MEDITRON-70B: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*. [CrossRef]

[17] Labrak, Y., Bazoge, A., Morin, E., Gourraud, P. A., Rouvier, M., & Dufour, R. (2024, August). BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: ACL 2024* (pp. 5848-5864). [CrossRef]

[18] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234-1240. [CrossRef]

[19] Alsentzer, E., Murphy, J., Boag, W., Weng, W. H., Jindi, D., Naumann, T., & McDermott, M. (2019, June). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*(pp. 72-78). [CrossRef]

[20] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH), 3*(1), 1-23. [CrossRef]

[21] Rasool, A., Aslam, S., Hussain, N., Imtiaz, S., & Riaz, W. (2025). nBERT: Harnessing NLP for emotion recognition in psychotherapy to transform mental health care. *Information, 16*(4), 301. [CrossRef]

[22] Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., ... & Chaudhari, A. S. (2024). Adapted large language models can outperform medical experts in clinical

text summarization. *Nature Medicine, 30*(4), 1134-1142. [CrossRef]

[23] Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Kipnis, P., ... & Lee, K. (2024). Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery, 5*(3), CAT-23. [CrossRef]

[24] Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature, 616*(7956), 259-265. [CrossRef]

[25] AlSaad, R., Abd-Alrazaq, A., Boughorbel, S., Ahmed, A., Renault, M. A., Damseh, R., & Sheikh, J. (2024). Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of Medical Internet Research, 26*(1), e59505. [CrossRef]

[26] Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P. C., ... & Natarajan, V. (2024). Towards generalist biomedical AI. *NEJM AI, 1*(3), AIoa2300138. [CrossRef]

[27] Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Zhao, M., Chow, A. K., ... & Mahmood, F. (2024). A multimodal generative AI copilot for human pathology. *Nature, 634*(8033), 466-473. [CrossRef]

[28] Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., & Xie, W. (2025). Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications, 16*(1), 7866. [CrossRef]

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

[30] Bannur, S., Bouzid, K., Castro, D. C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., ... & Hyland, S. L. (2024). MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*. [CrossRef]

[31] Yang, L., Xu, S., Sellergren, A., Kohlberger, T., Zhou, Y., Ktena, I., ... & Golden, D. (2024). Advancing multimodal medical capabilities of Gemini. *arXiv preprint arXiv:2405.03162*. [CrossRef]

[32] Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review, 10*(2), 115-152. [CrossRef]

[33] Karunanayake, N. (2025). Next-generation agentic AI for transforming healthcare. *Informatics and Health, 2*(2), 73-83. [CrossRef]

[34] Croatti, A., Montagna, S., Ricci, A., Gamberini, E., Albarello, V., & Agnoletti, V. (2019). BDI personal medical assistant agents: The case of trauma tracking and alerting. *Artificial Intelligence in Medicine, 96*, 187-197. [CrossRef]

[35] Alhazba, S. A., Ali, M. H., Alawi, D. A., Al-Betar, M. A., & Abd Elaziz, M. (2026). From reactive AI to agentic systems: A review of autonomous medical AI agents in healthcare. *Archives of Computational Methods in Engineering, 33*(8), 11693-11730. [CrossRef]

[36] Srinivasu, P. N., Aruna Kumari, G. L., Ahmed, S., & Alhumam, A. (2026). Exploring agentic AI in healthcare: A study on its working mechanism. *Frontiers in Medicine, 12*, 1753443. [CrossRef]

[37] Njei, B., Al-Ajlouni, Y. A., Sidney Kanmounye, U., Boateng, S., Loic Nguefang, G., Njei, N., ... & Al-Ajlouni, A. F. (2026). Artificial intelligence agents in healthcare research: A scoping review. *PLOS ONE, 21*(2), e0342182. [CrossRef]

[38] Collaco, B. G., Haider, S. A., Prabha, S., Gomez-Cabello, C. A., Genovese, A., Wood, N. G., ... & Forte, A. J. (2026). The role of agentic artificial intelligence in healthcare: a scoping review. *npj Digital Medicine, 9*(1), 345. [CrossRef]

[39] Huang, K. A., Choudhary, H. K., & Kuo, P. C. (2024). Artificial intelligent agent architecture and clinical decision-making in the healthcare sector. *Cureus, 16*(7). [CrossRef]

[40] Vatsal, S., Dubey, H., & Singh, A. (2026). Agentic AI in healthcare and medicine: A seven-dimensional taxonomy for empirical evaluation of LLM-based agents. *IEEE Access, 14*, 4840-4863. [CrossRef]

[41] Li, B., Yan, T., Pan, Y., Luo, J., Ji, R., Ding, J., ... & Wang, Y. (2024, November). MMedAgent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 8745-8760). [CrossRef]

[42] Shi, W., Xu, R., Zhuang, Y., Yu, Y., Zhang, J., Wu, H., ... & Wang, M. D. (2024, November). EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 22315-22339). [CrossRef]

[43] Kim, Y., Park, C., Jeong, H., Chan, Y. S., Xu, X., McDuff, D., ... & Park, H. W. (2024). MDAgents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems, 37*, 79410-79452. [CrossRef]

[44] Choi, J., Palumbo, N., Chalasani, P., Engelhard, M. M., Jha, S., Kumar, A., & Page, D. (2024). MALADE: Orchestration of LLM-powered agents with retrieval augmented generation for pharmacovigilance. In *Proceedings of the 9th Machine Learning for Healthcare Conference* (PMLR 252). https://proceedings.mlr.press/v252/choi24a.html

[45] Yue, L., Xing, S., Chen, J., & Fu, T. (2024, November). ClinicalAgent: Clinical trial multi-agent system with large language model-based reasoning. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (pp. 1-10). [CrossRef]

[46] Ren, Z., Zhan, Y., Yu, B., Ding, L., Xu, P., & Tao, D. (2025). Healthcare Agent: Eliciting the power of large language models for medical consultation. *npj Artificial*

Intelligence, 1(1), 24. [CrossRef]

[47] Ferber, D., El Nahhas, O. S., Wölflein, G., Wiest, I. C., Clusmann, J., Leßmann, M. E., ... & Kather, J. N. (2025). Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature cancer, 6*(8), 1337-1349. [CrossRef]

[48] Ferber, D., Hilgers, L., Höper, C., Kinny-Köster, B., Eckardt, J. N., Egger-Heidrich, K., ... & Kather, J. N. (2026). Towards autonomous medical artificial intelligence agents. *Nature*, 1-10. [CrossRef]

[49] Tu, T., Schaekermann, M., Palepu, A., Saab, K., Freyberg, J., Tanno, R., ... & Natarajan, V. (2025). Towards conversational diagnostic artificial intelligence. *Nature, 642*(8067), 442-450. [CrossRef]

[50] Saab, K., Park, C., Strother, T., Freyberg, J., Barrett, D. G., Cheng, Y., ... & Tanno, R. (2026). Advancing conversational diagnostic AI with multimodal reasoning. *Nature Medicine*, 1-11. [CrossRef]

[51] Liévin, V., Palepu, A., Weng, W. H., Saab, K., Stutz, D., Cheng, Y., ... & Schaekermann, M. (2026). Towards conversational artificial intelligence for disease management. *Nature, 655*(8125), 1292-1299. [CrossRef]

[52] Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., ... & Gerstein, M. (2024, August). MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 599-621). [CrossRef]

[53] Zhang, Y., Bunting, K. V., Champsi, A., Wang, X., Lu, W., Thorley, A., Hothi, S. S., Qiu, Z., Buyukates, B., Kotecha, D., & Duan, J. (2026). CardAIc-Agents: A multimodal framework with hierarchical adaptation for cardiac care support. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning* (PMLR 315, pp. 2143-2170). https://proceedings.mlr.press/v315/zhang26a.html

[54] Wu, Y., Chen, T., Hochberg, I., Sun, Z., Edry, R., Huang, Z., & Peleg, M. (2025). Lessons learned from evaluation of LLM-based multi-agents in safer therapy recommendation. *arXiv preprint arXiv:2507.10911*. [CrossRef]

[55] Zeng, F., Lyu, Z., Li, Q., & Li, X. (2024). Enhancing LLMs for impression generation in radiology reports through a multi-agent system. *arXiv preprint arXiv:2412.06828*. [CrossRef]

[56] Ghezloo, F., Seyfioglu, M. S., Soraki, R., Ikezogwo, W. O., Li, B., Vivekanandan, T., ... & Shapiro, L. (2025). PathFinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 23431-23441). [CrossRef]

[57] Maity, S., & Saikia, M. J. (2025). Large language models in healthcare and medical applications: a review. *Bioengineering, 12*(6), 631. [CrossRef]

[58] Wang, D., & Zhang, S. (2024). Large language models in medical and healthcare fields: Applications, advances, and challenges. *Artificial Intelligence Review, 57*(11), 299. [CrossRef]

[59] Rahim, S., Kong, X., & Abdullah, F. (2025). Navigating ethical boundaries in federated learning for biomedical research. *Journal of Artificial Intelligence in Bioinformatics, 1*(2), 72-78. [CrossRef]

[60] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine, 3*(1), 119. [CrossRef]

[61] Mehandru, N., Miao, B. Y., Almaraz, E. R., Sushil, M., Butte, A. J., & Alaa, A. (2024). Evaluating large language models as agents in the clinic. *npj Digital Medicine, 7*(1), 84. [CrossRef]



Muhammad Bilal is currently a Ph.D. researcher in Computer Science at Iowa State University, Ames, USA. He received the bachelor's degree from the Pakistan Institute of Engineering and Applied Sciences (PIEAS), Pakistan. His research interests include computer vision, machine learning, deep learning, large language models, multimodal learning, and artificial intelligence.



Hafsa Gulzar is currently a Ph.D. researcher at the Institute of Biomedical and Health Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. She received the master's degree in computer science and technology from Donghua University, Shanghai, China, in 2023. Her research interests include artificial intelligence, computer vision, medical image analysis, surgical video understanding, and machine learning for healthcare. (Email: hafsa@mail.dhu.edu.cn)



Sibtain Ahmed is a Ph.D. student in Computer Science at Iowa State University, USA. He received the engineering degree from NED University of Engineering and Technology, Karachi, Pakistan. His research interests include large language models, reinforcement learning, interpretable AI, and data attribution. His current research focuses on the influence of training data during reinforcement learning-based fine-tuning of large language models.

Muhammad Saqlain Aslam is with the AI Research Center, Hon Hai Research Institute, Taipei, Taiwan.