Diagnostic and Interventional Imaging xxx (xxxx) xxx

Contents lists available at ScienceDirect

Diagnostic and Interventional Imaging

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

Review

Agentic systems in radiology: Principles, opportunities, privacy risks, regulation, and sustainability concerns

Eleftherios Tzanis [a], Lisa C. Adams [b], Tugba Akinci D'Antonoli [c,d], Keno K. Bressem [b,e], Renato Cuocolo [f], Burak Kocak [g], Christina Malamateniou [h], Michail E. Klontzas [a,i,j,*]

[a] Artificial Intelligence and Translational Imaging (ATI) Lab, Department of Radiology, School of Medicine, University of Crete, 70013 Heraklion, Greece
[b] Department of Diagnostic and Interventional Radiology, Technical University of Munich, School of Medicine and Health, Klinikum Rechts der Isar, TUM University Hospital, 81675 Munich, Germany
[c] Department of Diagnostic and Interventional Neuroradiology, University Hospital Basel, CH-4031 Basel, Switzerland
[d] Department of Pediatric Radiology, University Children's Hospital Basel, CH-4056 Basel, Switzerland
[e] Department of Cardiovascular Radiology and Nuclear Medicine, Technical University of Munich, School of Medicine and Health, German Heart Center, TUM University Hospital, 80636 Munich, Germany
[f] Department of Medicine, Surgery and Dentistry, University of Salerno, 84081 Baronissi, Italy
[g] Department of Radiology, Basaksehir Cam and Sakura City Hospital, 34480 Istanbul, Turkey
[h] CRRAG Research group, Division of Radiography, School of Health and Medical Sciences, City St George's University of London, SW17 0RE London, UK
[i] Department of Medical Imaging, University Hospital of Heraklion, 71003 Heraklion, Crete, Greece
[j] Division of Radiology, Department of Clinical Science Intervention and Technology (CLINTEC), Karolinska Institute, SE-14152 Huddinge, Sweden

ARTICLE INFO

Keywords:
Agent
Agentic systems
Artificial intelligence
Large language models
Prompting
Radiology

ABSTRACT

The rapid rise of transformer-based large language models (LLMs) has introduced new opportunities for automation and decision support in radiology, particularly in applications such as report generation, protocol optimization, and structured interpretation. Despite their impressive performance in producing contextually coherent text, conventional LLMs remain limited by their inability to interact autonomously with external systems, retrieve data, or execute code, restricting their role in real-world clinical and research workflows. To address these limitations, agentic systems have emerged as a new paradigm. By embedding LLMs within frameworks that enable reasoning, planning, and action, agentic systems extend LLM capabilities to dynamic interaction with users, tools, and data sources. This review provides a comprehensive overview of the foundations, architectures, and operational mechanisms of agentic systems, focusing on their applications in medical imaging and radiology. It summarizes key developments in the literature, including recent multi-agent frameworks for automated radiomics pipelines, and discusses the potential benefits of these systems in enhancing the reproducibility, interpretability, and accessibility of AI-driven workflows. The review critically examines current regulatory considerations, ethical implications, and sustainability challenges to highlight essential gaps that must be addressed for the safe and responsible clinical integration of these systems.

1. Introduction

The introduction of transformer architecture and the attention mechanism sparked the rapid development of highly accurate and efficient models that can predict the next word in a sequence based on the preceding input [1]. These models are known as large language models (LLMs) and represent a highly active area of research. In medical domains such as radiology, LLMs are being explored for tasks including automated report drafting, protocol streamlining, and support for structured interpretation [2,3]. Preliminary studies suggest that these systems could improve the efficiency and consistency of medical decision-making in various settings [4–10].

Although LLMs can generate structured, contextually relevant text, based on user input, they cannot interact autonomously with their

Abbreviations: AI, Artificial intelligence; AIaMD, Artificial intelligence as a medical device; API, Application programming interface; EU, European Union; LLM, Large language model; RAG, Retrieval-augmented generation.

* Corresponding author.
E-mail address: miklontzas@uoc.gr (M.E. Klontzas).

https://doi.org/10.1016/j.diii.2025.10.002
Received 29 August 2025; Received in revised form 9 October 2025; Accepted 15 October 2025

environment. They cannot retrieve or process external data, execute code, or communicate directly with external systems and pipelines. These limitations constrain their applicability as fully integrated tools in real-world research and clinical practice.

A rapidly emerging field of research focuses on the development and deployment of AI-based agentic systems. These systems aim to overcome the inherent limitations of LLMs by enabling them to interact with their environment. In such architectures, LLMs serve as the central reasoning engine, and their behavior is shaped by prompting frameworks such as ReAct, chain of thought and tree of thoughts [11–13]. These prompting strategies define how the system plans, reasons, and acts.

Agentic systems typically engage with users through natural language, relying on the LLM to generate completions, outputs that may include reasoning steps, planning directives, executable code snippets, or commands for calling external functions and applications programming interfaces (APIs) [14]. Completions can be understood as the LLM's proposed next actions, whether textual explanations or instructions for interacting with external tools. In this way, agentic systems extend the capabilities of LLMs beyond text generation, allowing them to dynamically operate within their environment. The excellent, constantly evolving performance of commercial and open source LLMs underpins the promising capabilities of agentic system. The potential of such systems has recently been illustrated in a publication that presented a multi-agent framework for fully automated, end-to-end development of radiomics pipelines and machine learning models for medical imaging [15]. These systems offer a promising avenue for producing reproducible, interpretable, and trustworthy AI models, while also serving as powerful development and evaluation tools for scientific personnel with minimal or no programming expertise.

The purpose of this article was to provide a comprehensive overview of the fundamentals of agentic systems in the context of medical imaging and radiology. We examine how these systems are structured, how they operate, and how they may serve as practical tools to support routine clinical tasks. In addition, we review the current state of the literature, identify existing regulatory frameworks where applicable, and highlight the gaps that should be addressed to ensure the safe and responsible integration of such technologies into clinical practice.

2. LLMs: basic principles of the backbone of agentic systems

LLMs are transformer-based neural networks that learn to generate and interpret human language by optimizing billions of parameters on heterogeneous text corpora in a self-supervised fashion. Their development follows evidence that simply scaling model size and data volume yields systematic, power-law improvements in loss and downstream accuracy, an observation codified in early "scaling law" studies and later refined into compute-optimal training prescriptions [16,17]. Technical terms underlying these mechanisms are summarized in Table 1.

During pretraining, each model is presented with sequences of tokens and learns to predict the next token. This process gradually builds internal representations that capture syntax, semantics, and domain knowledge. The transformer architecture enables this by replacing recurrence with multi-head self-attention, which evaluates pairwise dependencies across the entire sequence in parallel. This mechanism efficiently links distant but clinically related phrases, for example, associating an initial mention of "ground-glass opacities" with a concluding assessment of "viral pneumonia" in a radiology report, mirroring expert radiologists' integrative reading strategies [1].

Tokenization bridges raw text and model input. Subword tokenizers such as WordPiece, SentencePiece, or domain-specific variants split rare biomedical expressions like "hepatosplenomegaly" into morphologically meaningful units, reducing out-of-vocabulary errors and preserving semantic content. BioBERT's success in biomedical natural language processing illustrates how domain-tuned vocabularies materially improve representation quality [18]. Self-attention's quadratic memory footprint ties model performance to the length of the context window.

Table 1
Technical terms related to large language models and agentic systems.

Term	Explanation
Large language models	A large language model is an advanced AI architecture trained using deep learning methodologies on extensive corpora of textual data, enabling the recognition, generation, translation, and summarization of natural language.
Agentic systems	AI agents are systems that employ AI techniques to pursue defined objectives and execute tasks on behalf of users. They exhibit capabilities such as reasoning, planning, and acting, while operating with a degree of autonomy that allows them to make decisions, learn from experience, and adapt to changing contexts. They can operate as single agents or as a combination of multiple individual agents (multi-agent systems).
Transformers	A deep learning architecture that processes input sequences in parallel and uses attention to model long-range dependencies, enabling efficient handling of complex language and imaging tasks.
Attention mechanism	A method within neural networks that dynamically assigns importance weights to different input elements, allowing the model to focus on the most relevant information for the prediction.
Tokenization	The process of splitting text or data into smaller units (tokens), such as words, subwords, or characters, which are then converted into numerical representations for model input.
Context window	The maximum number of tokens a model can process at once, defining how much prior information or context can be considered in generating outputs.
Prompt engineering	The practice of designing, refining, and structuring inputs (prompts) to optimize model responses for specific tasks or domains.
System prompt	A persistent instruction or configuration that establishes the model's role, constraints, and style of interaction throughout a session.

AI indicates artificial intelligence.

Mainstream deployments range from 4 k to 32 k tokens, with research prototypes, using sparse, dilated, or state-space attention, pushing well beyond 100 k and even to the billion-token scale. While such advances permit ingestion of entire longitudinal imaging records, they also impose very high compute and memory costs in routine deployment [19].

Retrieval-augmented generation (RAG) and dense passage retrieval mitigate fixed-window limits by allowing the model to dynamically fetch external documents and weave them into the prompt, thereby extending its effective knowledge base without expanding the core network. These hybrid systems have proven especially useful for surfacing prior reports or guidelines during automated impression generation [20].

Because the raw objective instils only statistical correlations, additional alignment steps are required to approximate clinical reasoning and professional tone. Reinforcement learning from human feedback, instruction tuning, and rule-based "constitutional AI" frameworks train the model to follow radiology-specific instructions, refuse unsafe requests, and prioritize concise, clinically usable output [21].

Medical-domain LLMs now integrate these advances. Med-PaLM demonstrates that a general-purpose model aligned with expert prompts can match clinician-level question-answering performance [22], while Radiology-GPT achieves domain-specific improvements through instruction tuning on curated report corpora [23]. Parallel progress in vision-language foundation models shows that cross-modal pre-training can further ground language understanding in imaging features, setting the stage for holistic reporting assistants that incorporate images, prior text, and laboratory data in a single dialogue [24].

Despite these gains, LLM outputs remain probabilistic extrapolations from training distributions. They excel at structured report drafting, error checking, and rapid literature retrieval, but can misinterpret rare presentations or novel imaging artifacts. In safety-critical settings such

as radiology, expert oversight, rigorous validation, and transparent uncertainty quantification remain indispensable complements to LLM-enabled workflow acceleration [25].

3. Prompt engineering

Prompt engineering refers to the systematic adaptation of text inputs to LLMs to optimize their performance for specific tasks. It involves the strategic design and refinement of instructions to guide LLMs toward generating accurate, contextually appropriate responses. These techniques range from simple query formulation to sophisticated approaches involving output constraints and parameter adaptation.

Several prompting techniques have been developed, progressing from simple to increasingly complex approaches (Fig. 1). Zero-shot prompting requires no prior examples within the prompt itself, relying entirely on the model's pre-trained knowledge to understand and execute tasks. In radiology, this approach proves effective for straightforward classification tasks where the model's existing medical knowledge suffices.

Few-shot prompting adds complexity by providing a small number of examples of the desired input-output format within the prompt, leveraging the LLM's capacity for in-context learning [16]. This approach improves task-specific accuracy, particularly when examples illustrate the expected answer pattern and output format. Chain-of-thought prompting further enhances the prompt with structured reasoning guidance, explicitly directing models through step-by-step reasoning processes and encouraging intermediate reasoning steps before final conclusions [12]. This technique proves particularly valuable for complex diagnostic reasoning requiring multi-step analysis. Studies in radiology have demonstrated that structured reasoning approaches can improve diagnostic accuracy (*i.e.*, the correctness of diagnostic conclusions compared with expert reference standards) from 56.5 % to 60.6 % in complex cases by encouraging systematic information organization before diagnosis formulation [26].

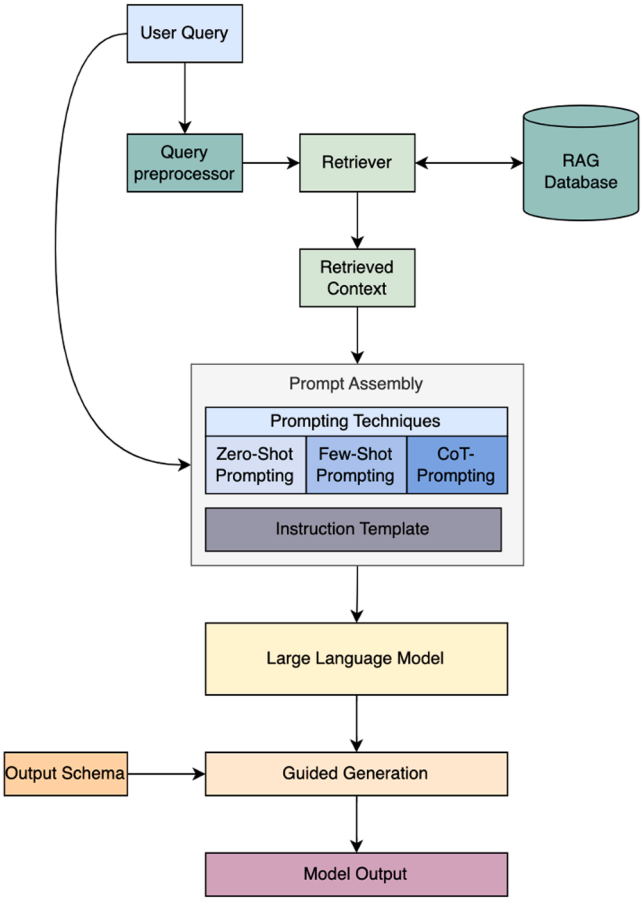

Fig. 1. Simplified prompt assembly with retrieval augmentation and guided decoding. The user query is normalized and possibly expanded by a query preprocessor, which produces the retrieval query for the retriever that searches a knowledge store and returns context. The final model prompt is composed from the normalized query, an instruction template, and instruction strategies such as zero shot, few shot, or chain of thought (CoT), together with the retrieved context. The language model generates the answer under optional decoding constraints defined by an output schema or a controlled vocabulary. RAG indicates retrieval-augmented generation.

Guided generation represents an advanced technique that, instead of engineering the prompt, constrains the model's token generation process according to predefined structures, vocabularies, or grammars [27]. Unlike traditional prompting that relies on instructions alone, guided generation implements hard constraints during the decoding process, ensuring outputs conform to specific formats or standards. This is achieved through several mechanisms: constrained decoding that restricts the model's vocabulary to predefined token sets during generation, grammar-based constraints that enforce syntactic structures (such as valid JSON schemas), and template-based generation that forces outputs to follow specific organizational patterns.

In radiology, guided generation can enforce that diagnostic outputs use standardized classification systems, for instance, restricting mammography assessments to valid BI-RADS categories rather than allowing free-text descriptions that might deviate from established standards [12,16,26–28]. Similarly, it can ensure structured reports are generated in machine-readable formats such as JSON with predefined fields (*e.g.*, {"findings": "…", "impression": "…", "recommendations": "…"}), facilitating integration with electronic health record systems. The technique can also constrain vocabulary to established medical terminologies like SNOMED CT codes, ensuring consistent and interoperable clinical documentation while preventing the generation of non-standard or ambiguous terminology.

For knowledge-intensive domains like medicine, where LLMs may lack access to specialized or institutional knowledge, RAG provides essential capabilities by connecting models to external databases, literature, or institutional guidelines. While not a prompting technique but rather an architectural pattern, where a model is connected to an external knowledge base from which additional context can be derived, RAG proves especially valuable in radiology due to the field's reliance on constantly evolving medical knowledge and the need for factually grounded information. Research has demonstrated that RAG-enhanced models achieve superior diagnostic accuracy, 100 % compared to 93 % for baseline models in trauma radiology applications [27,28]. RAG enables smaller LLMs to compete with larger models that cannot be deployed on-site while allowing incorporation of sensitive institutional data without public disclosure.

These techniques demonstrate particular value in radiology applications, where chain-of-though prompting enhances complex diagnostic reasoning, guided generation ensures compliance with clinical standards, and RAG systems provide access to current literature and institutional protocols.

4. Fundamentals of agentic systems

Agents are systems designed to perform complex tasks while interacting with users through natural language. At their core, these systems consist of two main components: the brain, which is typically an LLM responsible for reasoning and communication, and the body, which refers to the set of tools the agent can use to interact with its environment. These tools may include functions (*e.g.*, in Python or other languages), APIs, or any callable resource that allows the agent to execute specific actions in response to a user's request.

A key element in the behavior and effectiveness of such systems lies in the prompting framework, the set of structured instructions that guide

how the agent reasons, plans, and acts [11–13]. These frameworks shape the agent's overall cognitive workflow, including how it decomposes problems, decides which actions to take, and when to stop. One widely used technique is the ReAct framework, short for Reason and Act [11]. In this setup, agents solve problems iteratively through a cycle of three steps as follows: (*i*), Think: the agent interprets the user's request and formulates a plan; (*ii*), Act: it executes a specific action, such as calling a function or retrieving data; and (*iii*), Observe: it analyzes the outcome of the action to determine if the task is complete. If the goal is not yet achieved, the agent uses the new information to revise its plan and re-enters the think-act-observe loop. This iterative process continues until the agent reaches a satisfactory solution or final answer, which is then returned to the user.

The most important component of agentic systems is the system prompt. This is a block of text provided to the LLM at initialization, containing persistent instructions that define how the agent should operate. It serves as a blueprint for the agent's behavior, specifying the prompting technique to be used, the expected format of interaction with the user, and the overall strategy for task execution. In addition to guiding behavior and reasoning, the system prompt includes descriptions of the available tools and any task-specific agents. These descriptions provide the LLM with the necessary context to determine which tool or specialized agent to invoke and how to use it in response to user queries. By embedding this information into the system prompt, the LLM "knows" what resources it has at its disposal and can generate completions accordingly, whether those completions are reasoning steps, code snippets, or commands for interacting with the environment. The workflow of a typical agentic system is presented in Fig. 2.

Agentic systems fall into two main categories based on how they interact with tools, which are tool-calling agents and code agents (Fig. 3). The most common approach is the tool-calling agent, where interaction with tools is managed through structured JSON definitions. Each tool is described by its name, a brief description, and a schema defining its input parameters and types. In contrast, code agents interact with tools by generating and executing Python code snippets. Each approach has distinct advantages and trade-offs. Tool-calling agents are typically more reliable and safer, as the structure of the tool calls is controlled, reducing the risk of hallucinations or unexpected behavior. However, they are less flexible, limited to a fixed set of predefined actions, and cannot easily perform dynamic transformations or synthesize new logic. Code agents, on the other hand, provide high expressiveness and emergent reasoning capabilities, enabling more sophisticated behaviors without needing to predefine every possible action. The trade-off is that they require a secure execution environment and are more prone to errors, including syntax mistakes or unsafe code [29].

Another component of agentic systems is their memory, which plays an important role in maintaining context throughout interactions with the user. The most commonly used form is short-term memory, also referred to as working memory. During a session, the user's queries and the agent's internal reasoning and actions are stored as logs. This allows the LLM to maintain awareness of the conversation history, understand what has already occurred, and generate contextually appropriate next

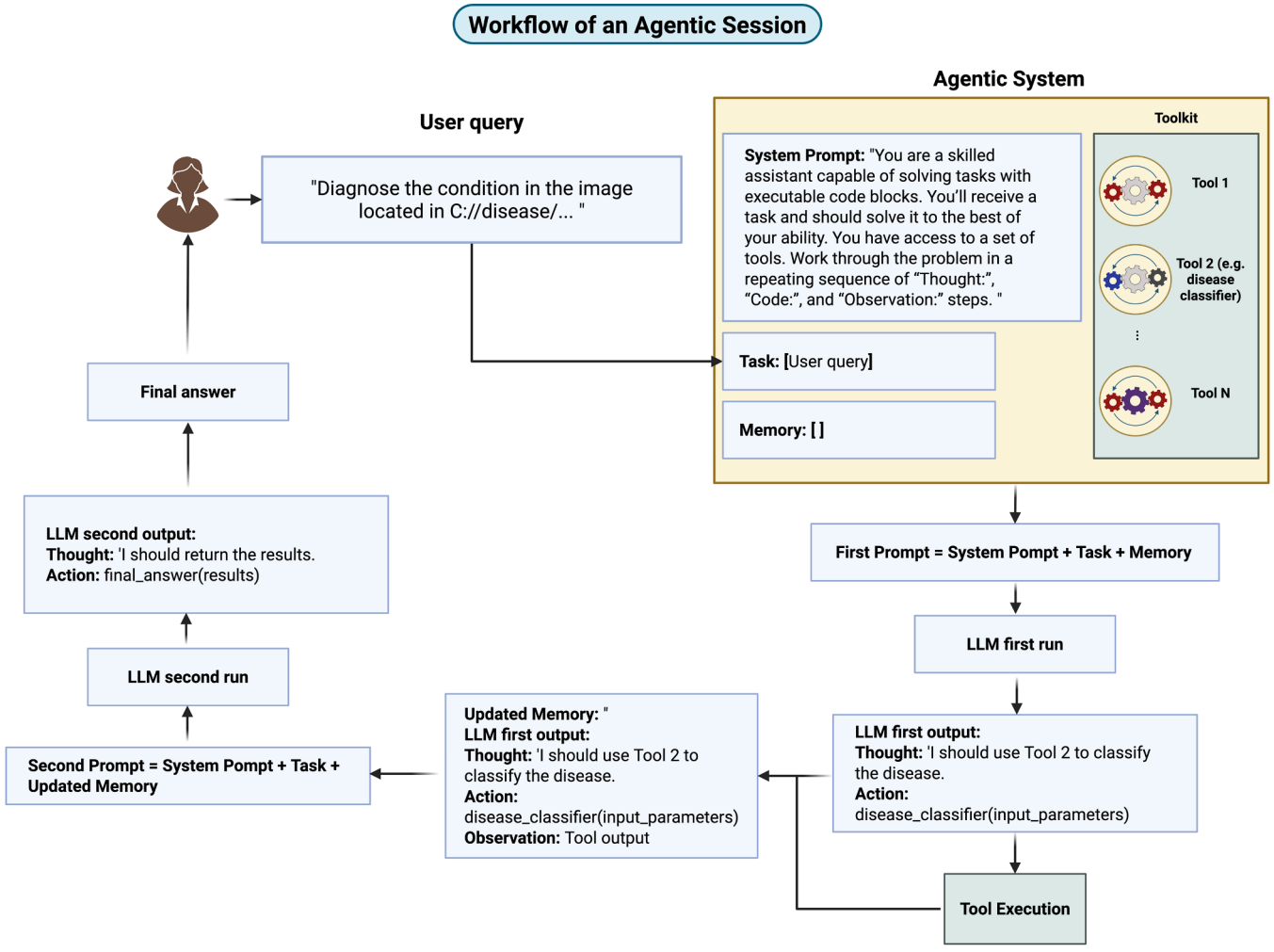

Fig. 2. Graphical description of a typical agentic workflow (created with biorender.com).
LLM indicates large language model.

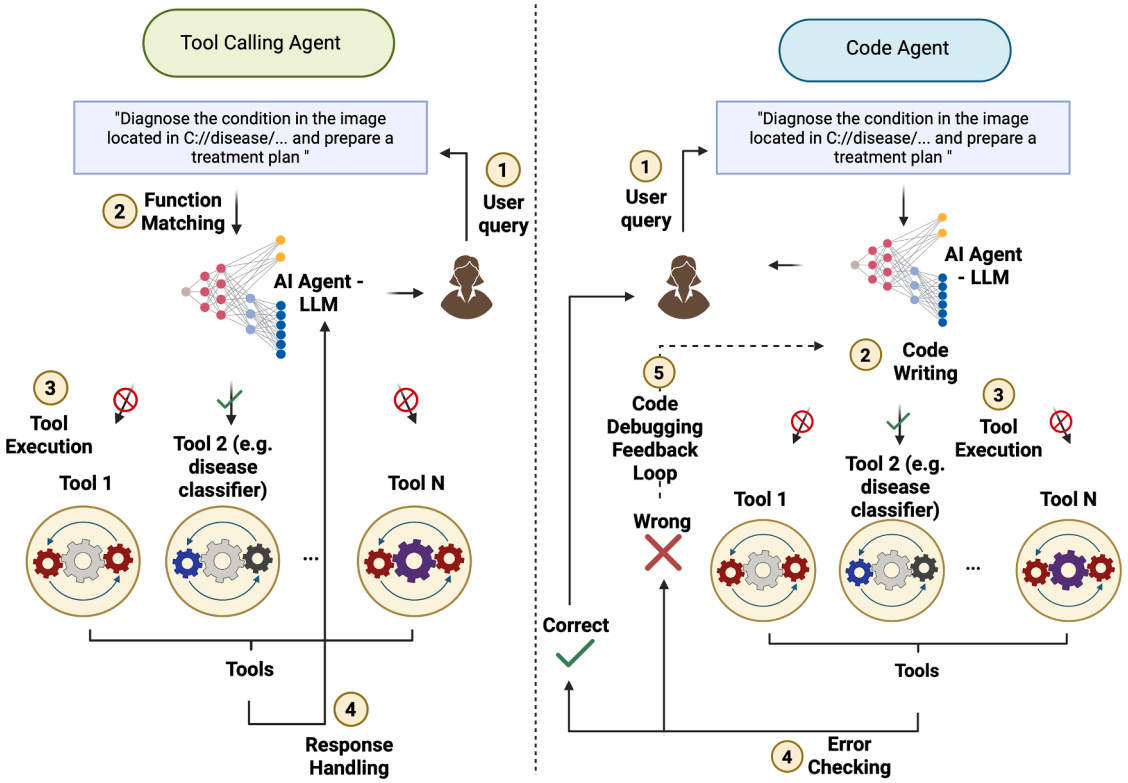

Fig. 3. Schematic representing the function of the two main types of agents, tool calling and code agents (created with biorender.com). LLM indicates large language model.

steps based on prior information. In contrast, long-term memory refers to the agent's ability to retain and access information across different sessions. This involves storing relevant data, insights, or strategies from previous interactions, enabling the agent to learn from past experiences. By integrating long-term memory, agentic systems can be guided by previous successes and failures, improving their ability to solve tasks more efficiently over time. Users can grant access to such memory stores, allowing agents to adapt and evolve through accumulated knowledge.

A variety of frameworks have become available to support the development of such systems, each offering different capabilities for managing tools, memory, and interaction strategies. An overview of some of the most widely used frameworks is presented in Table 2. Such systems can aid a wide variety of tasks ranging from diagnosis (the most common in published literature) to protocol planning and report correction, structuring, and improvement. While these non-diagnostic tasks are equally important and can be aided by agentic systems, the examples provided in this manuscript are based on diagnostic tasks, which may be more complex but are also commonly encountered in the majority of published papers.

Table 2
Frameworks for building agentic systems.

Framework	Characteristics	GitHub repository
Smolagents [29]	A lightweight framework for building agentic workflows. Emphasizes simplicity, sandboxed execution, and ease of use.	https://github.com/huggingface/smolagents
LangGraph [30]	Enables stateful, long-running agents with strong memory support and complex workflow orchestration. Ideal for multi-agent structures with debugging support.	https://github.com/langchain-ai/langgraph
CrewAI [31]	A performant framework for building autonomous agents with GUIs (CrewAI-Studio), RAG workflows, and GitHub integrations.	https://github.com/crewAIInc/crewAI
Agno [32]	A toolkit for multi-agent systems featuring layered levels of agency (from tool-using agents to full agentic workflows), shared memory, reasoning, and observability.	https://github.com/agno-agi/agno
AutoGen [33]	Event-driven framework for building flexible multi-agent workflows. Supports conversational agents, tool integration, asynchronous messaging, and modular components.	https://github.com/microsoft/autogen

5. Privacy issues related to the use of agentic systems

The integration of LLMs and subsequently agentic systems into radiology introduces substantial privacy and cybersecurity risks, driven by the models' interaction with sensitive patient data [34–36]. A major concern is the potential for data leakage during both training and inference. Due to their scale and training methodologies, LLMs can inadvertently memorize and regurgitate sensitive patient information, even when such information was not explicitly intended to be retained. This phenomenon, also known as unintended memorization, has been observed in LLM deployments where inadequately filtered training data resulted in models generating outputs that inadvertently expose sensitive or confidential content [37]. This poses a considerable threat in medical contexts, where maintaining strict confidentiality is not only an ethical imperative but also a legal requirement under regulations such as the Health Insurance Portability and Accountability Act in the USA and the General Data Protection Regulation in Europe [38]. Therefore, anonymizing medical data before training is a widely recommended mitigation strategy; however, achieving effective anonymization remains a substantial technical hurdle. Traditional de-identification techniques, such as removing direct identifiers (*e.g.*, names, social security numbers), often fall short in safeguarding against re-identification, especially when LLMs are able to use indirect or

quasi-identifiers within the data [39]. Furthermore, in complex datasets like those in radiology, metadata embedded in image files in DICOM data format may inadvertently reveal patient-specific details, complicating the anonymization process [38].

Beyond training, there are also risks at the time of inference that demand careful attention. When LLMs are used in clinical decision support or diagnostic settings, they may generate responses that reflect patterns learned from sensitive training data. If such data included protected health information, a malicious actor could exploit the model through adversarial prompting, potentially extracting details such as patient age, diagnoses, medical history, or other confidential attributes [34]. Therefore, there is a need for robust model auditing, differential privacy techniques, and real-time monitoring to ensure outputs remain compliant with privacy standards.

From a cybersecurity standpoint, integrating LLMs into radiology workflows also demands defenses against model poisoning, prompt injection, and backdoor attacks [34]. Model poisoning occurs when adversaries introduce manipulated data during training, aiming to corrupt the model's behavior or introduce vulnerabilities. Prompt injection attacks manipulate user inputs in ways that elicit unintended, misleading, or harmful outputs from the model. Backdoor attacks embed hidden triggers during training, causing the model to behave maliciously when specific inputs are received, and often without detection in normal operations. These threats, particularly in high-stakes medical contexts, pose significant risks to patient safety, data integrity, and system trustworthiness.

All these risks are further amplified in multi-agent systems, where multiple AI agents collaborate and share information across networks [40,41]. The collaborative nature of these systems expands the attack surface, and raises questions about data governance, access control, and accountability. Ensuring secure information exchange between agents and enforcing strict access control, encryption, and identity verification protocols are essential to maintaining system integrity and patient confidentiality. To responsibly integrate LLMs and LLM-powered multi-agent systems into radiology, organizations must proactively address both privacy and cybersecurity challenges. Risks with the use of LLMs and multi-agentic systems are summarised in Fig. 4.

6. Clinical applications of multi-agent systems in radiology

While promising, AI agents are still not typically available in medical devices approved for clinical use. AI as a medical device (AIaMD) has, also for regulatory reasons, generally been focused on narrow, specific tasks, with more deterministic outputs, such as region of interest (*e.g.*, lesion) detection and segmentation, with or without corresponding estimates for a diagnosis of interest (*e.g.*, clinically significant prostate cancer) [42]. Furthermore, given the multimodal nature of data used by AI Agents, one can also expect imaging examinations to become more often part of the AIaMD's input data rather than the focus of the Agent. In other words, fully leveraging agentic LLMs will probably require integration of patient data from multiple sources to perform the action of interest, as the results of imaging examinations represent only one component (albeit essential in many cases) in the diagnostic and patient management process that takes place in healthcare [43].

Nevertheless, agentic AIaMDs can be expected to be introduced in medical imaging in more limited roles at first, either improving upon performance of non-agentic AIs present in current medical devices or only marginally expanding the scope of such systems. In this setting, one can classify clinical applications of agentic LLMs in closed and open-ended tasks [44]. The first presents a limited number of outputs available to the model, representing a much simpler task to be performed compared to open-ended questions. However, while good results are present in this setting, for example when experimenting LLM use in multiple-choice questions for certification examinations, this has little potential for translation into the clinical setting [45,46]. Open-ended tasks for LLMs are more clearly related to the radiology workflow, including summarization of information, extraction and restructuring of data, mainly in text form, and interactive answering of medical

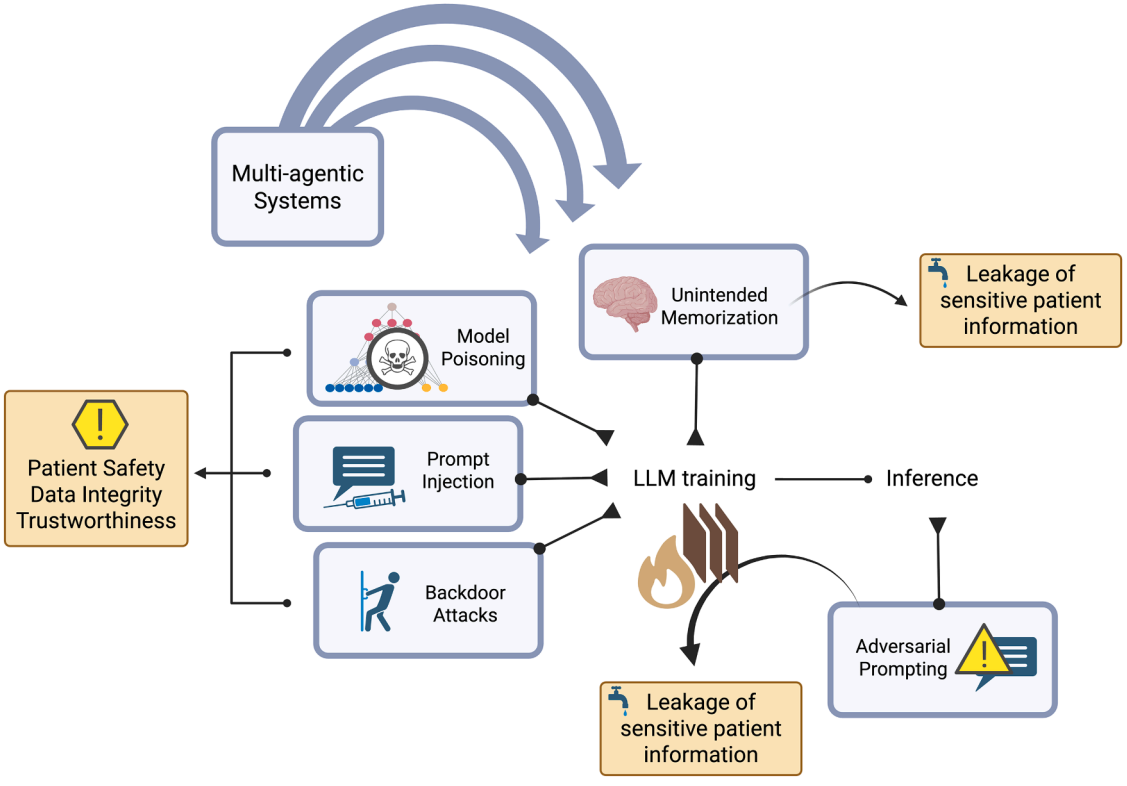

Fig. 4. Risks associated with the use of large language models (LLMs) are amplified in multi-agentic systems. Cybersecurity considerations and privacy risks related to LLMs are presented in this figure (created with biorender.com).

questions [44,47]. AI has shown good promise in research for these applications, with potential end users both within medical imaging professionals, such as AIaMDs providing reporting assistance, and patients, for example, for assistance in interpretation of the radiological report [48–50]. It should be noted that some of these models are being integrated in medical devices, even if not always available in the European Union (EU) [51]. Finally, true AI Agents with access to multimodal patient information could be leveraged to optimize and automate exam scheduling, including follow-up to known pathologies, and provide benefits at a larger scale rather than simply improving efficiency and accuracy of human reporting [52,53].

At this time, many research efforts are aimed at building the necessary infrastructure to perform pre-clinical and clinical validation of agentic AI systems [44]. When thoroughly tested, such models still present open questions in terms of consistency and accuracy, which may also explain the relatively lower availability of medical devices within the EU, where more stringent regulations have been historically present compared to the other major markets. Interestingly, AI Agents have also been proposed to improve the design process for radiomics and AI medical imaging systems, and may actively contribute to accelerate their translation from the scientific to clinical setting [15].

7. Applications of multi-agent systems in radiology research

In radiology research, agentic systems can automate processes that traditionally required advanced coding or data analysis skills. Agent-based systems have emerged over the past years that can handle three-dimensional medical imaging data. VoxelPrompt is one of the first agentic systems that combine language with vision models to perform segmentation and lesion characterisation of multiple types of images [54]. The system has been primarily tested on neuroimaging applications where it was able to segment and classify hundreds of lesions, performing a series of tasks including but not limited to assessment of potential lesion contrast enhancement and diffusion restriction, assessment of brain infarct territories, and temporal follow-up of lesion size across multiple hospital visits.

Since then, a variety of multi-agent systems have been developed to facilitate imaging research that can handle the whole process of image analysis, including image preprocessing, segmentation, quantitative radiomic data extraction, and model building. mAIstro represents a prototype of these systems, which can enable researchers to automate data analysis and machine learning model building by writing Python code with established libraries and interacting with the user using natural language [15]. The system can be used with API calls to a series of state-of-the-art LLMs and assists the user in selecting the appropriate methods for the designated research question, and allows automated radiomic analysis, traditional and deep learning model building, multi-organ segmentation, and exploratory data analysis. NVIDIA, in collaboration with King's College London, has also incorporated agentic systems into the MONAI framework based on Llama 3 to assist users in research related to radiology and surgery, including features for automated radiology report generation [55,56].

While these applications demonstrate the versatility of agentic systems in radiology research, large-scale studies are still required to determine whether such systems can genuinely improve diagnostic accuracy, workflow efficiency, and patient outcomes. Unlike the clinical use of agentic systems, the use of such systems for research purposes does not require regulatory approval as a medical device, rendering their commercial rollout. For the United States Food & Drug Administration, these can be labelled "For Research Use Only. Not for use in diagnostic procedures" [57]. In Europe, if an AI tool is used exclusively for scientific investigation and not intended for obtaining marketing authorization, it is governed by Article 82. Such studies require ethics approval and adherence to relevant national regulations, but CE-marking is not necessary for this type of research use [58].

8. Governance of agentic AI

According to Gartner, nearly a third of enterprise applications will incorporate agentic AI by 2028, compared to less than 1 % in 2024 [59]. Agentic AI's relative autonomy and advanced capabilities in handling complex tasks set it apart from simpler AI tools [60]. While these features make agentic AI an invaluable partner to human actors for the completion of complex tasks in healthcare and other fields, they also introduce new ethical and governance concerns around autonomy, transparency, explainability, bias, and accountability, redefining paradigms of human-AI interaction and exemplifying the need for new human oversight approaches.

A fundamental governance question relates to the accountability of when agentic AI errs or inadvertently causes harm [61]. Humans can be held accountable for decisions they take when these may directly impact others. However, automated decisions are not self-justifiable [62]. AI software applications and hardware systems, although they need to follow principles of responsible AI [63], they do not have the same moral responsibility as human actors do [64]. This creates an accountability gap for agentic AI, which carries a sense of autonomy in part of its decision-making processes, particularly for complex tasks. In clinical contexts, this accountability gap raises practical liability questions: if an agentic system misinterprets an image or generates an unsafe recommendation, responsibility could fall on the hospital, the developer or provider of the agentic system, or the end user. Current legislation, including the EU AI Act, does not yet provide specific guidance for such scenarios, underscoring the need for clear contractual and regulatory frameworks that delineate liability among stakeholders. As such, clear and attributable sources of human answerability should be attached to processes, tasks, and decisions when enabled by an agentic AI system [65].

Agentic AI may also be prone to cyber-attacks and data breaches [60]. Data privacy and security safeguarding (such as encryption, secure coding practices, anomaly detection, and continuous monitoring) are crucial governance measures when using agentic AI systems. There is little work currently in this field for agentic AI, but it will become more relevant in the next couple of years as these systems develop.

While agentic AI is not featured or directly discussed in the most recent updates of the EU AI Act, there is a lot mentioned within it about autonomy and human oversight (particularly in recital 27 and article 14) [66,67]. More specifically, according to the guidelines of the AI HLEG9, "human agency and oversight means that AI systems are developed and used as a tool that serves people, respects human dignity and personal autonomy, and that is functioning in a way that can be appropriately controlled and overseen by humans". Furthermore, "high-risk AI systems (where agentic AI may be classified under) shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use" [68].

AI agents can make decisions about complex tasks by autonomously analysing multimodal data. This requires more transparency on how intermediate decision steps are decided and care to ensure bias from multimodal data does not accumulate in the final product. Strategies to improve transparency of AI agents include direct interpretability (like decision trees or similar tools) or post-hoc interpretability (also known as explainability) [69]. Important to note that there is currently no consensus on what constitutes a good explanation to address the opaque nature of AI agent decision-making [69]. Finally, preliminary work currently takes place to understand and remove bias. Bias is deeply rooted in LLMs and therefore inherent in AI agents; tools such as bias detectors are important developments in the right direction for fairer agentic AI, but have to be properly evaluated in real-world data [70].

As agentic AI advances, more benefits and risks will become apparent; its governance will have to evolve in parallel, and robust regulatory frameworks will need to be appended to current legislation to ensure it delivers safe, transparent, fair results, keeping humans in the

loop and with clear accountability pipelines.

9. Sustainability of agentic systems

Integration of LLMs and agentic AI systems into various sectors, including specialized fields like medicine, demands a thorough examination of their sustainability across environmental, economic, and social dimensions [71]. Significant energy demand, carbon emissions (direct consequence of energy consumption), and water consumption of LLMs pose a considerable environmental challenge [72–76], though contrasting perspectives exist [77]. This challenge is amplified by agentic systems, which may perform numerous, iterative tasks autonomously. While training foundational models is energy-intensive, it is the inference stage (*i.e.*, the actual use of the model) that constitutes the dominant and ongoing environmental cost [78]. The cumulative emissions from inference can exceed those from training by roughly a factor of 1000 [78,79]. One study estimated that the top 20 carbon-emitting AI systems could generate up to 102.6 million tonnes of CO_2 equivalent annually [78], surpassing the yearly emissions of over 100 countries in 2023 [80]. Projections indicate a 30–40 % annual increase in energy demand for AI services over the next decade [81], directly translating into increased carbon emissions. Furthermore, the data centers supporting these models have a substantial water footprint due to their cooling requirements [73]. Mitigation strategies include adopting energy-efficient architectures, such as smaller, fine-tuned models, and quantization (*i.e.*, reducing the bit-width of weights), which reduces model size and computational load [82,83]. Transitioning data centers to renewable energy sources is critical [72–74,78,84,85]. Additionally, prompt engineering that encourages shorter responses can cut energy use by 25–60 % [86].

The development and operation of LLMs involve substantial economic costs. Training a model can run into millions of dollars, and API usage remains expensive [87,88]. Despite these expenses, LLMs can enhance efficiency and reduce costs in sectors like healthcare by streamlining data extraction and administrative tasks [89,90]. A cost-effective strategy is query concatenation, which can reduce costs significantly, by up to 17-fold for 50 simultaneous tasks, by grouping multiple queries into a single request [87]. For agentic systems, which can operate autonomously, the potential for runaway operational costs from inefficient task loops is a critical risk. The Jevons paradox also serves as a warning: increased efficiency might lead to higher overall consumption rather than savings [73]. Importantly, many economic mitigation strategies, such as fine-tuning existing models for specific tasks or careful model selection based on complexity, offer the dual benefit of reducing environmental impact [82,88].

Socially, the increasing autonomy of LLMs and agentic systems raises concerns regarding ethical design, bias, and accountability [77,91,92]. These models can inherit and perpetuate the societal biases present in their training data [93–95]. In medicine, the potential for LLMs to comply with harmful or inappropriate requests poses significant risks [96], while the autonomy of agents introduces complex questions of accountability when they cause harm or make critical errors. Other concerns include job displacement and the erosion of critical thinking skills [88,90]. Ensuring equitable access to these powerful technologies is also a crucial social challenge [88,97–99]. Multi-agent systems introduce further complexity, as LLM agents often struggle to achieve sustainable cooperation without specific interventions and may fail to analyze the long-term consequences of their actions [100]. To address these issues, robust governance, transparent accountability, and clear ethical oversight are essential [89,96,99,101]. Bias mitigation through careful data curation and safety fine-tuning is vital, as is emphasizing human-AI collaboration to ensure human oversight in critical applications [89,96]. Ultimately, harnessing the power of agentic systems responsibly requires an integrated approach where environmental, economic, and social sustainability considerations are treated not as separate challenges but as interconnected components of a single, sustainable framework.

10. Conclusion

It has become evident that agentic systems hold considerable promise for reshaping radiology by executing complex tasks, which range from image interpretation and workflow orchestration to research data analysis. As our review has outlined, their successful deployment will depend on careful navigation of risks such as the protection of sensitive patient information and security vulnerabilities, as well as the environmental sustainability of such increasingly resource-intensive models. Another challenge is the establishment of sustainable economic models, since the costs of training, fine-tuning, and utilizing LLMs, the core reasoning engines of agentic systems, can be substantial. Employing smaller, domain-specific LLMs as reasoning backbones may help reduce these costs, but large-scale studies are still needed to clarify the long-term cost-benefit balance. Alignment with evolving regulatory frameworks is also critical to ensure safety, transparency, and accountability. Furthermore, successful adoption will depend on human factors, particularly radiologists' trust, appropriate training, and seamless workflow integration, without which even advanced systems risk remaining proof-of-concept. When responsibly designed and implemented, agentic systems could augment radiologists in clinical practice by improving efficiency, consistency, and decision support, while also accelerating scientific discovery in radiology research through automated data curation, analysis, and hypothesis generation. Ultimately, their impact will be determined by a balance between innovation and close supervision, leveraging the strengths of distributed intelligence while safeguarding the core values of patient-centered, ethical, and sustainable radiological care.

CRediT authorship contribution statement

Eleftherios Tzanis: Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Approval of final draft.

Lisa C. Adams: Writing – original draft, Writing – review & editing, Approval of final draft.

Tugba Akinci D'Antonoli: Writing – original draft, Writing – review & editing, Approval of final draft.

Keno K. Bressem: Writing – original draft, Writing – review & editing, Approval of final draft.

Renato Cuocolo: Writing – original draft, Writing – review & editing, Approval of final draft.

Burak Kocak: Writing – original draft, Writing – review & editing, Approval of final draft.

Christina Malamateniou: Writing – original draft, Writing – review & editing, Approval of final draft.

Michail E. Klontzas: Conceptualization, Investigation, Project administration, Writing – original draft, Writing – review & editing, Supervision, Approval of final draft.

Human rights

Not applicable to review article.

Informed consent and patient details

Not applicable to review article.

Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for authorship.

Declaration of competing interest

The authors declare that they have no known competing financial or personal relationships that could be viewed as influencing the work reported in this paper.

Acknowledgements

None

References

[1] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., et al. Attention is all you need. arXiv 2025. https://arxiv.org/abs/1706.03762.
[2] Meddeb A, Lüken S, Busch F, Adams L, Ugga L, Koltsakis E, et al. Large language model ability to translate CT and MRI free-text radiology reports into multiple languages. Radiology 2024;313:e241736.
[3] Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology 2024;310:e232756.
[4] Bradshaw T, Cho S. Evaluation of large language models in natural language processing of PET/CT free-text reports. J Nucl Med 2021;62:1188.
[5] Mitsuyama Y, Tatekawa H, Takita H, Sasaki F, Tashiro A, Oue S, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. Eur Radiol 2025; 35:1938–47.
[6] Hirata K, Matsui Y, Yamada A, Fujioka T, Yanagawa M, Nakaura T, et al. Generative AI and large language models in nuclear medicine: current status and future prospects. Ann Nucl Med 2024;38:853–64.
[7] Tordjman M, Bolger I, Yuce M, Restrepo F, Liu Z, Dercle L, et al. Large language models in cancer imaging: applications and future perspectives. J Clin Med 2025; 14:3285.
[8] Shool S, Adimi S, R Saboori Amleshi, Bitaraf E, Golpira R, Tara M. A systematic review of large language model evaluations in clinical medicine. BMC Med Inf Decis Mak 2025;25:117.
[9] Alberts IL, Mercolli L, Pyka T, Prenosil G, Shi K, Rominger A, et al. Large language models and ChatGPT: what will the impact on nuclear medicine be? Eur J Nucl Med Mol Imaging 2023;50:1549–52.
[10] Lecler A, Soyer P, Gong B. The potential and pitfalls of ChatGPT in radiology. Diagn Interv Imaging 2024;105:249–50.
[11] Yao S., Zhao J., Yu D., Du N., Shafran I., Narasimhan K., et al. ReAct: synergizing reasoning and acting in language models. arXiv 2023. https://arxiv.org/abs/2210.03629.
[12] Wei J., Wang X., Schuurmans D., Bosma M., Ichter B., Xia F., et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv 2023. https://arxiv.org/abs/2201.11903.
[13] Yao S., Yu D., Zhao J., Shafran I., Griffiths T.L., Cao Y., et al. Tree of thoughts: deliberate problem solving with large language models. arXiv 2023. https://arxiv.org/abs/2305.10601.
[14] Sumers T.R., Yao S., Narasimhan K., Griffiths T.L. Cognitive architectures for language agents. arXiv 2024. https://arxiv.org/abs/2309.02427.
[15] Tzanis E, Klontzas ME. mAIstro: an open-source multi-agent system for automated end-to-end development of radiomics and deep learning models for medical imaging. arXiv 2024. https://arxiv.org/abs/2505.03785.
[16] Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., et al. Language models are few-shot learners. arXiv 2020. https://arxiv.org/abs/2005.14165.
[17] Hoffmann J., Borgeaud S., Mensch A., Buchatskaya E., Cai T., Rutherford E., et al. Training compute-optimal large language models. arXiv 2022. https://arxiv.org/abs/2203.15556.
[18] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234–40.
[19] Ding J., Ma S., Dong L., Zhang X., Huang S., Wang W., et al. LongNet: scaling transformers to 1,000,000,000 tokens. arXiv 2023. https://arxiv.org/abs/2307.02486.
[20] Karpukhin V., Oguz B., Min S., Lewis P., Wu L., Edunov S., et al. Dense passage retrieval for open-domain question answering. arXiv 2020. https://arxiv.org/abs/2004.04906.
[21] Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C.L., Mishkin P., et al. Training language models to follow instructions with human feedback. arXiv 2022. https://arxiv.org/abs/2203.02155.
[22] Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, et al. Towards generalist biomedical AI. N Eng J Med AI 2024;1:1–12.
[23] Liu Z., Zhong A., Li Y., Yang L., Ju C., Wu Z., et al. Radiology-GPT: a large language model for radiology. arXiv 2024. https://arxiv.org/abs/2306.08666.
[24] Singhal K., Azizi S., Tu T., Mahdavi S.S., Wei J., Chung H.W., et al. Large language models encode clinical knowledge. arXiv 2022. https://arxiv.org/abs/2212.13138.
[25] Duron L, Lecler A. Multimodal artificial intelligence in radiology: text-dominant reasoning limits image understanding. Diagn Interv Imaging 2025;106:333–4.
[26] Willard B.T., Louf R. Efficient guided generation for large language. arXiv 2023. https://arxiv.org/abs/2307.09702.
[27] Fink A, Nattenmüller J, Rau S, Rau A, Tran H, Bamberg F, et al. Retrieval-augmented generation improves precision and trust of a GPT-4 model for emergency radiology diagnosis and classification: a proof-of-concept study. Eur Radiol 2025;35:5091–8.
[28] Arasteh S.T., Lotfinia M., Bressem K., Siepmann R., Adams L., Ferber D., et al. RadioRAG: online retrieval-augmented generation for radiology question answering. arXiv 2025. https://arxiv.org/abs/2407.15621.
[29] Roucher A., Villanova del Moral A., Wolf T., von Werra L., Kaunismäki E. `Smolagents`: a smol library to build great agentic systems. https://github.com/huggingface/smolagents (accessed 28/8/2025).
[30] LangGraph. https://github.com/langchain-ai/langgraph (accessed 28/8/2025).
[31] crewAI. https://github.com/crewAIInc/crewAI (accessed 28/8/2025).
[32] Agno. https://github.com/agno-agi/agno (accessed 28/8/2025).
[33] AutoGen. https://github.com/microsoft/autogen (accessed 28/8/2025).
[34] Akinci D'Antonoli T, Tejani AS, Khosravi B, Bluethgen C, Busch F, Bressem KK, et al. Cybersecurity threats and mitigation strategies for large language models in health care. Radiol Artif Intell 2025;7:e240739.
[35] Lecler A, Soyer P. AI in radiology: powerful, promising… but alarmingly hackable. Diagn Interv Imaging 2025. https://doi.org/10.1016/j.diii.2025.06.003.
[36] Duron L, Soyer P, Lecler A. Generative AI smartphones: from entertainment to potentially serious risks in radiology. Diagn Interv Imaging 2025;106:76–8.
[37] Satvaty A., Verberne S., Turkmen F. Undesirable memorization in large language models: a survey. arXiv 2025. https://arxiv.org/abs/2410.02650.
[38] Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. Radiology 2020;295:4–15.
[39] Falis M, Gruber F, McInerney S, Casey A. Evaluating LLMs' potential to identify rare patient identifiers in patient health records. Stud health. Technol Inf 2025; 327:874–5.
[40] Bengio Y., Cohen M., Fornasiere D., Ghosn J., Greiner P., MacDermott M., et al. Superintelligent agents pose catastrophic risks: can scientist AI offer a safer path? arXiv 2025. https://arxiv.org/abs/2502.15657.
[41] Chan A., Salganik R., Markelius A., Pang C., Rajkumar N., Krasheninnikov D., et al. Harms from increasingly agentic algorithmic systems. arXiv 2023. https://arxiv.org/abs/2302.10329.
[42] Kotter E, D'Antonoli TA, Cuocolo R, Hierath M, Huisman M, Klontzas ME, et al. Guiding AI in radiology: ESR's recommendations for effective implementation of the European AI Act. Insights Imaging 2025;16:33.
[43] Wang W, Ma Z, Wang Z, Wu C, Chen W, Li X, et al. A survey of LLM-based agents in medicine: how far are we from Baymax? Annu Meet Assoc Comput Linguist 2025:10345–59.
[44] Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models and agents in healthcare: key challenges in clinical applications. Intell Med 2025;5: 151–63.
[45] Sun SH, Chen K, Anavim S, Phillipi M, Yeh L, Huynh K, et al. Large language models with vision on diagnostic radiology board exam style questions. Acad Radiol 2025;32:3096–102.
[46] Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024;26:e60807.
[47] Koçak B, Meşe İ. AI agents in radiology: toward autonomous and adaptive intelligence. Diagn Interv Radiol 2025. https://doi.org/10.4274/dir.2025.253470.
[48] Busch F, Hoffmann L, Dos Santos DP, Makowski MR, Saba L, Prucker P, et al. Large language models for structured reporting in radiology: past, present, and future. Eur Radiol 2025;35:2589–602.
[49] Gupta A, Singh S, Malhotra H, Pruthi H, Sharma A, Garg AK, et al. Provision of radiology reports simplified with large language models to patients with cancer: impact on patient satisfaction. JCO Clin Cancer Inf 2025;9:e2400166.
[50] Herwald SE, Shah P, Johnston A, Olsen C, Delbrouck JB, Langlotz CP. RadGPT: a system based on a large language model that generates sets of patient-centered materials to explain radiology report information. J Am Coll Radiol 2025. https://doi.org/10.1016/j.jacr.2025.06.013.
[51] Rad AI. https://www.radai.com/(accessed 28/8/2025).
[52] Akinci D'Antonoli T, Bluethgen C, Cuocolo R, Klontzas ME, Ponsiglione A, Kocak B. Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects. Diagn Interv Radiol 2025. https://doi.org/10.4274/dir.2025.253445.
[53] Pierre K, Haneberg AG, Kwak S, Peters KR, Hochhegger B, Sananmuang T, et al. Applications of artificial intelligence in the radiology roundtrip: process streamlining, workflow optimization, and beyond. Semin Roentgenol 2023;58: 158–69.
[54] Hoopes A., Butoi V.I., Guttag J.V., Dalca A.V. VoxelPrompt: a vision-language agent for grounded medical image analysis. arXiv 2025. https://arxiv.org/abs/2410.08397.
[55] Zephyr M. NVIDIA technical blog. MONAI integrates advanced agentic architectures to establish multimodal medical AI ecosystem. https://developer.nvidia.com/blog/monai-integrates-advanced-agentic-architectures-to-establish-multimodal-medical-ai-ecosystem/(accessed 28/8/2025).
[56] Cardoso M.J., Li W., Brown R., Ma N., Kerfoot E., Wang Y., et al. MONAI: an open-source framework for deep learning in healthcare. arXiv 2022. https://arxiv.org/abs/2211.02701.
[57] U.S. Food and Drug Administration. FAQs about investigational device exemption. https://www.fda.gov/medical-devices/investigational-device-exemption-ide/faqs-about-investigational-device-exemption#:~:text=IVD%20devices%20that%20are%20under,product%20have%20not%20been%20established (accessed 5/8/2025).
[58] Massimo P. Medical device regulation. Article 82 - requirements regarding other clinical investigations. https://www.medical-device-regulation.eu/2019/07/16/mdr-article-82-requirements-regarding-other-clinical-investigations/(accessed 5/8/2025).

[59] Gartner W.K. How agentic AI is shaping business decision-making. https://technologymagazine.com/articles/gartner-how-agentic-ai-is-shaping-business-decision-making (accessed 5/8/2025).
[60] Murugesan S. The rise of agentic AI: implications, concerns, and the path forward. IEEE Intell Syst 2025;40:8–14.
[61] Schoenherr JR, Thomson R. Attributing responsibility in human-AI interactions. IEEE Trans Technol Soc 2024;5:61–70.
[62] AI ethics & governance in practice: AI ethics. https://aiethics.turing.ac.uk/modules/accountability/?modulepage=part-one-introduction-to-accountability (accessed 5/8/2025).
[63] Walsh G, Stogiannos N, van de Venter R, Rainey C, Tam W, McFadden S, et al. Responsible AI practice and AI education are central to AI implementation: a rapid review for all medical imaging professionals in Europe. BJR Open 2023;5:20230033.
[64] Coeckelbergh M. Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 2020;26:2051–68.
[65] Goetze TS. Mind the gap: autonomous systems, the responsibility gap, and moral entanglement. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22. Association for Computing Machinery. New York, NY, USA; 2022. p. 390–400.
[66] EU AI Act, Recital 27. https://artificialintelligenceact.eu/recital/27/ (accessed 5/8/2025).
[67] Shaping Europe's digital future: draft ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/draft-ethics-guidelines-trustworthy-ai (accessed 5/8/2025).
[68] EU AI Act, Article 14: human oversight. https://artificialintelligenceact.eu/article/14/ (accessed 5/8/2025).
[69] Papagni G, de Pagter J, Zafari S, Filzmoser M, Koeszegi ST. Artificial agents' explainability to support trust: considerations on timing and context. AI Soc 2023;38:947–60.
[70] Singh K., Ngu W. Bias-aware agent: enhancing fairness in AI-driven knowledge retrieval. arXiv 2025. https://arxiv.org/abs/2503.2132.
[71] Kocak B, Ponsiglione A, Romeo V, Ugga L, Huisman M, Cuocolo R. Radiology AI and sustainability paradox: environmental, economic, and social dimensions. Insights Imaging 2025;16:88.
[72] Dong H., Xie S. Large language models: deployment, tokenomics and sustainability. arXiv 2024. https://arxiv.org/abs/2405.17147.
[73] Jegham N., Abdelatti M., Elmoubarki L., Hendawi A. How hungry is AI? Benchmarking energy, water, and carbon footprint of LLM inference. arXiv 2025. https://arxiv.org/abs/2505.09598.
[74] Jiang P, Sonne C, Li W, You F, You S. Preventing the immense increase in the life-cycle energy and carbon footprints of LLM-powered intelligent chatbots. Engineering 2024;40:202–10.
[75] Nguyen S., Zhou B., Ding Y., Liu S. Towards sustainable large language model serving. arXiv 2024. https://arxiv.org/abs/2501.01990.
[76] Ueda D, Walston SL, Fujita S, Fushimi Y, Tsuboyama T, Kamagata K, et al. Climate change and artificial intelligence in healthcare: review and recommendations towards a sustainable future. Diagn Interv Imaging 2024;105:453–9.
[77] Ren S, Tomlinson B, Black RW, Torrance AW. Reconciling the contrasting narratives on the environmental impact of large language models. Sci Rep 2024;14:26310.
[78] Yu Y, Wang J, Liu Y, Yu P, Wang D, Zheng P, et al. Revisit the environmental impact of artificial intelligence: the overlooked carbon emission source? Front Env Sci Eng 2024;18:158.
[79] Chien AA, Lin L, Nguyen H, Rao V, Sharma T, Wijayawardana R. Reducing the carbon impact of generative AI inference (today and in 2035). In: Proceedings of the 2nd Workshop on Sustainable Computer Systems. ACM; 2023. p. 1–7.
[80] Ritchie H., Rosado P., Roser M. Greenhouse gas emissions: our world in data. https://ourworldindata.org/greenhouse-gas-emissions (accessed 28/8/2025).
[81] Luers A, Koomey J, Masanet E, Gaffney O, Creutzig F, Lavista Ferres J, et al. Will AI accelerate or delay the race to net-zero emissions? Nature 2024;628:718–20.
[82] Doo FX, Savani D, Kanhere A, Carlos RC, Joshi A, Yi PH, et al. Optimal large language model characteristics to balance accuracy and energy use for sustainable medical applications. Radiology 2024;312:e240320.
[83] Husom E.J., Goknil A., Astekin M., Shar L.K., Kåsen A., Sen S., et al. Sustainable LLM inference for edge AI: evaluating quantized LLMs for energy efficiency, output accuracy, and inference latency. arXiv 2025. https://arxiv.org/abs/2504.03360.
[84] Pipek P, Canavan S, Canavan S, Capinha C, Gippet JMW, Novoa A, et al. Sustainability of large language models: user perspective. Front Ecol Env 2025;23:5.
[85] An J, Ding W, Lin C. ChatGPT: tackle the growing carbon footprint of generative AI. Nature 2023;615:586.
[86] Poddar S., Koley P., Misra J., Podder S., Balani N., Ganguly N., et al. Brevity is the soul of sustainability: characterizing LLM response lengths. arXiv 2025. https://arxiv.org/abs/2506.08686.
[87] Klang E, Apakama D, Abbott EE, Vaid A, Lampert J, Sakhuja A, et al. A strategy for cost-effective large language model use at health system-scale. NPJ Digit Med 2024;7:320.
[88] Nagarajan R, Kondo M, Salas F, Sezgin E, Yao Y, Klotzman V, et al. Economics and equity of large language models: health care perspective. J Med Internet Res 2024;26:e64226.
[89] Hughes L, Dwivedi YK, Malik T, Shawosh M, Albashrawi MA, Jeon I, et al. AI agents and agentic systems: a multi-expert analysis. J Comput Inf Syst 2025;65:489–517.
[90] Kwong JCC, Wang SCY, Nickel GC, Cacciamani GE, Kvedar JC. The long but necessary road to responsible use of large language models in healthcare research. NPJ Digit Med 2024;7:177.
[91] Bush A., Aksoy M., Pauly M., Ontrup G. Choosing a model, shaping a future: comparing LLM perspectives on sustainability and its relationship with AI. arXiv 2025. https://arxiv.org/abs/2505.14435.
[92] Koçak B, Ponsiglione A, Stanzione A, Bluethgen C, Santinha J, Ugga L, et al. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. Diagn Interv Radiol 2025;31:75–88.
[93] Wan Y., Pu G., Sun J., Garimella A., Chang K.W., Peng N. "Kelly is a warm person, Joseph is a role model": gender biases in LLM-generated reference letters. arXiv 2023. https://arxiv.org/abs/2310.09219.
[94] Rutinowski J, Franke S, Endendyk J, Dormuth I, Roidl M, Pauly M. The self-perception and political biases of ChatGPT. Hum Behav Emerg Technol 2024;2024:1–9.
[95] Poulain R., Fayyaz H., Beheshti R. Bias patterns in the application of LLMs for clinical decision support: a comprehensive study. arXiv 2024. https://arxiv.org/abs/2404.15149.
[96] Han T., Kumar A., Agarwal C., Lakkaraju H. Towards safe large language models for medicine. arXiv 2024. https://arxiv.org/abs/2403.03744.
[97] Tang YD, Dong ED, Gao W. LLMs in medicine: the need for advanced evaluation systems for disruptive technologies. Innovation 2024;5:100622.
[98] Wu H., Wang X., Fan Z. Addressing the sustainable AI trilemma: a case study on LLM agents and RAG. arXiv 2025. https://arxiv.org/abs/2501.08262.
[99] Karunanayake N. Next-generation agentic AI for transforming healthcare. Inform Health 2025;2:73–83.
[100] Piatti G., Jin Z., Kleiman-Weiner M., Schölkopf B., Sachan M., Mihalcea R. Cooperate or collapse: emergence of sustainable cooperation in a society of LLM agents. arXiv 2024. https://arxiv.org/abs/2404.16698.
[101] He K, Mao R, Lin Q, Ruan Y, Lan X, Feng M, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. Inf Fusion 2025;118:102963.