

# Agentic Systems in Radiology: Design, Applications, Evaluation, and Challenges

Christian Bluethgen<sup>1,7‡</sup> Dave Van Veen<sup>2</sup> Daniel Truhn<sup>3,4</sup> Jakob Nikolas Kather<sup>5</sup>  
 Michael Moor<sup>6</sup> Małgorzata Połacin<sup>1</sup> Akshay Chaudhari<sup>7,8,9</sup> Thomas Frauenfelder<sup>1</sup>  
 Curtis P. Langlotz<sup>7,8,9</sup> Michael Krauthammer<sup>10</sup> Farhad Nooralahzadeh<sup>10,11</sup>

<sup>1</sup>Diagnostic and Interventional Radiology, University Hospital Zurich, University of Zurich, Zurich, Switzerland  
<sup>2</sup>HOPPR, Menlo Park, CA 94025, USA

<sup>3</sup>Lab for AI in Medicine, University Hospital Aachen, Aachen, Germany

<sup>4</sup>Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany

<sup>5</sup>Else Kroener Fresenius Center for Digital Health, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, 01307 Dresden, Germany

<sup>6</sup>Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

<sup>7</sup>Center for Artificial Intelligence in Medicine and Imaging, Stanford University

<sup>8</sup>Department of Radiology, Stanford University

<sup>9</sup>Department of Biomedical Data Science, Stanford University

<sup>10</sup>Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

<sup>11</sup>Institute of Computer Science, Zurich University of Applied Sciences, Zurich, Switzerland

## Abstract

Building agents, systems that perceive and act upon their environment with a degree of autonomy, has long been a focus of AI research. This pursuit has recently become vastly more practical with the emergence of large language models (LLMs) capable of using natural language to integrate information, follow instructions, and perform forms of "reasoning" and planning across a wide range of tasks. With its multimodal data streams and orchestrated workflows spanning multiple systems, radiology is uniquely suited to benefit from agents that can adapt to context and automate repetitive yet complex tasks. In radiology, LLMs and their multimodal variants have already demonstrated promising performance for individual tasks such as information extraction and report summarization. However, using LLMs in isolation underutilizes their potential to support complex, multi-step workflows where decisions depend on evolving context from multiple information sources. Equipping LLMs with external tools and feedback mechanisms enables them to drive systems that exhibit a spectrum of autonomy, ranging from semi-automated workflows to more adaptive agents capable of managing complex processes. This review examines the design of such LLM-driven *agentic* systems, highlights key applications, discusses evaluation methods for planning and tool use, and outlines challenges such as error cascades, tool-use efficiency, and health IT integration.

---

<sup>‡</sup>Corresponding author.

## 1 Introduction

Radiologists and their teams coordinate patients, operate scanners, interpret images, integrate clinical data, and communicate results. This multifaceted workflow demands adaptable problem-solving and frequent context switching; combined with rising imaging requests and a relative radiologist shortage, it contributes to cognitive overload and diagnostic delay.<sup>1–3</sup>

AI is increasingly seen as a way to help manage this complexity,<sup>4</sup> but many current implementations remain narrow-scoped and poorly integrated into clinical workflows. Large language models (LLMs) like GPT-5, and their multimodal variants, stand out for their ability to flexibly handle tasks specified in natural language<sup>5–8</sup> at unprecedented accessibility. However, their effectiveness remains limited in real-world radiology, where tasks often involve multiple steps that unfold over time. When LLMs are used in isolation or called only once, as is typical in many current applications, they cannot adapt based on new information emerging during response generation.

LLM-driven *agentic* systems address these limitations by embedding one or more LLMs within a framework in which LLMs can generate plans and select actions to iteratively interact with their environment.<sup>9,10</sup> In radiology, such systems could manage multi-step tasks that involve retrieving patient context, orchestrating specialized models, consulting external resources like guidelines, and synthesizing context-rich outputs like radiology reports. The field's data-rich, dynamic nature makes it well-suited for agentic approaches, but its complexity and clinical stakes require rigorous evaluation before deployment.

This review outlines the technical foundations of LLM-based agents, frames radiology as an agent environment and explores potential application, reviews methods for evaluating agent performance, highlights key challenges to clinical deployment, and considers future directions. Our objective is to illustrate to radiologists, researchers, and developers the potential of LLM-based workflows and agents to support complex, real-world radiological tasks.

## 2 Technical Foundations of LLM-based Agents

**What is an agent?** An *agent* is an entity that perceives (through *sensors*) and acts (through *actuators*) on an *environment*.<sup>10</sup> LLM-based agents (Fig. 1) run an LLM with access to external *tools* in a *loop* with some degree of autonomy in deciding *which*, *when*, and *how* actions are executed to pursue a goal.

Here, we use "agentic" to refer to LLM-driven systems exhibiting goal-directed, feedback-adaptive be-

havior under limited supervision, including more autonomous *agents* acting in open-ended settings, and less autonomous *workflows* following predefined multi-step structures while making constrained, feedback-informed decisions within those boundaries. This distinction highlights differences in control, supervision, and adaptability, while recognizing that in practice, such systems often blend elements of both and that agency and autonomy lie on a spectrum.<sup>11</sup>

### 2.1 LLMs as Agent Cores

Agents include a core component that transforms observations into actions, answering the question "Given the available information, what should the system do next?". Earlier approaches to this problem relied on logical rules and structured representations (*symbolic agents*), mapping observations directly to actions (*reactive agents*),<sup>10,12</sup> or reinforcement learning (*RL agents*) that learn through interactions with (and reward signals from) their environment.<sup>13,14</sup> While RL has shown promise in narrow domains, its use in healthcare remains limited<sup>15</sup> as reward design is difficult, and trial-and-error learning can be unsafe or impractical.

Unlike earlier agent designs, LLM-based agents use natural language prompts, intermediate feedback from their environment, and self-evaluation to guide decisions. The strength of LLMs—and by extension, large multimodal models with LLM components—comes from pretraining on massive, diverse datasets,<sup>16,17</sup> which gives them broad general knowledge and the ability to process and produce language, follow instructions, reason and plan (to some extent),<sup>18–23</sup> and utilize tools and memory.<sup>12,24–26</sup>

This combination of abilities and general knowledge marks a qualitative shift: for the first time, it has become practically feasible to build powerful AI agents for a wide range of domains. LLM-based agents are already being deployed across industries such as customer service, software development, and supply chain management, where they demonstrate adaptability and open-ended problem-solving beyond what earlier agent architectures could achieve.<sup>27</sup> These properties make them particularly interesting for complex, dynamic domains such as radiology.

### 2.2 Environment, Tools and Actions

An environment includes everything outside the agent that it can observe or influence through actions.<sup>10</sup> In radiology, this may include imaging devices, hospital and radiology IT systems, EHRs, and human stakeholders involved in clinical and administrative workflows.

To navigate an environment and produce useful decisions and outputs, an agent must connect its inner processes to external, "real-world" signals—a process known as *grounding*.<sup>28–30</sup> For example, describing a left lower lobe consolidation requires connecting the image content to the text sequences "left lower lobe" and "consolidation"

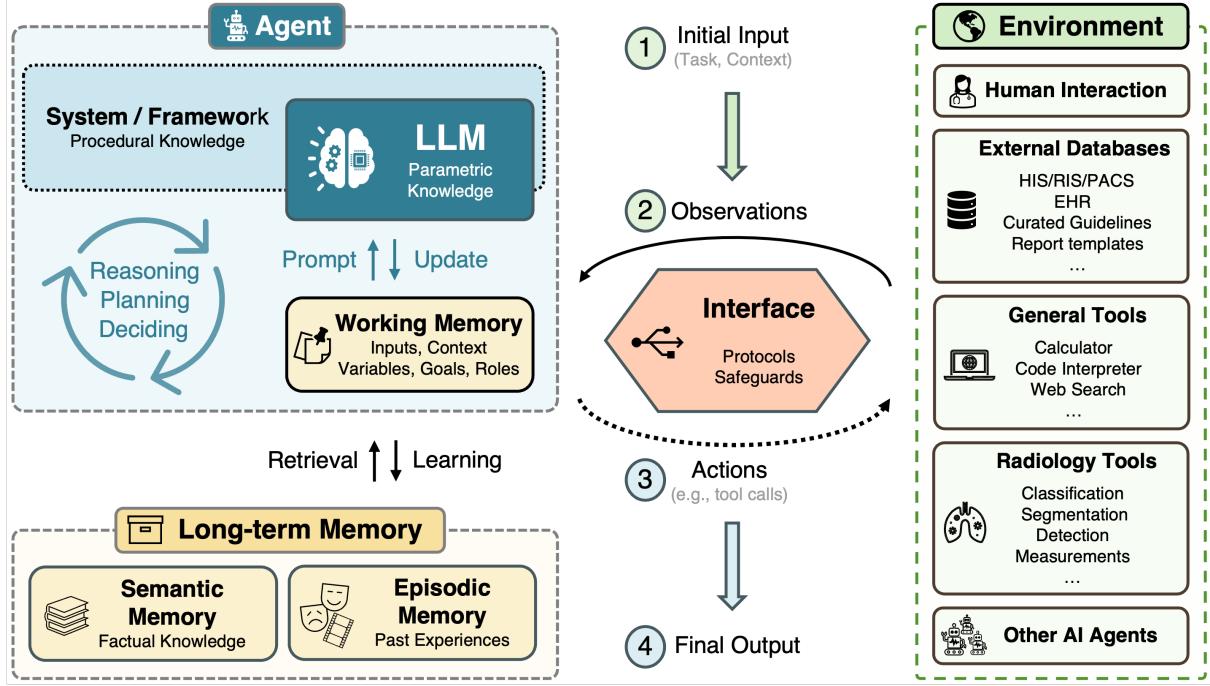


Figure 1: **Conceptual architecture of a radiology-focused LLM-based agent.** An initial input (1) provides the task and context. The agent then enters a cycle of obtaining observations (2), reasoning and planning over context, and performing actions (3) on the environment, such as tool calls or database queries. This cycle continues until a final output (4) is produced. The agent comprises an LLM, a framework, and a *working memory*. An optional agent-owned *long-term memory* stores *episodic* (past interactions) and *semantic* (factual knowledge) information to support retrieval and learning. The agent interacts with its *environment* (green box), including external systems (e.g., HIS/RIS/PACS, EHR, databases), general and radiology-specific tools, humans, and other AI agents, via defined interfaces (e.g., Model Context Protocol (MCP), Agent-to-Agent (A2A)) and safeguards (e.g., PHI reduction, input validation). AI: Artificial Intelligence. EHR: Electronic Health Record. HIS: Hospital Information System. LLM: Large Language Model. PACS: Picture Archiving and Communication System. PHI: Protected Health Information. RIS: Radiology Information System.

(visual grounding<sup>31</sup>). These inputs (such as retrieved documents or images) may be handled directly by the (optionally multimodal) LLM or with help from external tools.

Tools are resources in the environment that agents use to *sense* and *act on* the environment. General tools include search engines, calculators, and code interpreters. In radiology, such tools could offer access to databases like PubMed or to specialized models (e.g., for segmentation). Tool effectiveness depends on alignment with the task and context, making tool optimization a key priority.<sup>11,32</sup>

Tools fall into three broad categories: First, tools for *accessing dynamic or specialized knowledge* to help agents move beyond static training data to retrieve up-to-date, patient-specific, or task-specific relevant information that would be impractical to include for every query (e.g., including lung cancer follow-up guideline texts during an abdominal MRI reporting workflow). Second, tools that *augment information processing* to support tasks that remain difficult for LLMs, such as symbolic logic, math, or specialized vision tasks. Examples include segmentation models, dose calculators, or anatomical landmark detection to obtain measurements. Third, tools that en-

able *acting on the environment* to allow agents to flag priority cases, schedule appointments, or communicate with remote monitoring devices.

Protocols are emerging to standardize how agents use tools and interact with each other.<sup>33</sup> One example is Anthropic’s open Model Context Protocol (MCP), which defines a shared format for tool descriptions, requests, and responses. Instead of relying on custom application programming interfaces (APIs) that define explicit protocols how systems interact, agents can use MCP to flexibly discover and start using new tools while running. For instance, an agent can query an MCP-enabled EHR system without vendor-specific code, while still relying on established infrastructure. A2A (Agent-to-Agent) is another protocol handling secure, structured communication between agents themselves, allowing them to coordinate tasks, exchange data, and delegate subtasks in a standardized way. Although promising, standards like MCP and A2A are still early and fragmented.<sup>33</sup>

An agent’s *action space* is the set of external tools it can access and internal (LLM-native) actions it can perform (e.g., reasoning, summarization).<sup>34</sup> Core agent functions include deciding *when* and *how* to act (e.g., providing the

right input parameters at the right time, correctly parsing the returned output). Frameworks like LangChain,<sup>35</sup> DSPy,<sup>36</sup> and HuggingFace’s smolagents<sup>37</sup> can help manage this logic. The action space can be expanded through LLM fine-tuning, loosening constraints, enhancing tools, or even enabling agents to create tools themselves.<sup>38</sup>

### 2.3 Goals, Reasoning and Planning

LLMs pursue goals specified in natural language, which can leave room for ambiguity that can lead to unwanted outputs—behavior analogous to specification gaming<sup>39,40</sup>. For example, an instruction to “quickly complete radiology reports to improve turnaround time” may be interpreted as prioritizing speed over completeness, yielding terse or incomplete reports. Defining composite requirements for accuracy, completeness, and clinical appropriateness reduces this risk.

When faced with challenging requests, agents can apply *reasoning*, which in the context of LLMs usually involves generating intermediate steps (i.e., chain-of-thought reasoning) to work through a problem systematically rather than jumping directly to conclusions,<sup>20</sup> and *planning*, which constructs a sequence of actions expected to achieve the goal.<sup>10</sup> In radiology, this mirrors how a radiologist first organizes findings (reasoning) and then decides which prior studies and guidelines to consult, or which measurements to obtain, and in what order (planning). For example, interpreting a chest CT with multiple pulmonary nodules might involve (1) cataloging each nodule’s characteristics, (2) comparing findings to prior imaging, (3) considering the patient’s history and differential diagnoses, and (4) synthesizing recommendations. LLM-based agents can follow a similar structure by explicitly writing out their chain of thought before conclusions, typically yielding more accurate results than attempting a full assessment in one step.<sup>20</sup> In multimodal agents, this process may go beyond language: an agent might reason directly over images by generating predicted visual sequences<sup>41</sup> or operate within a latent space before producing a final output.<sup>42</sup>

A straightforward approach to handle complex requests is either break the task into subtasks manually or use a separate external planning systems to orchestrate the work.<sup>32</sup> Letting the model “think” longer while generating (*test-time scaling*) can also help.<sup>25,43</sup> Prompting strategies that expose or explore intermediate steps can raise accuracy further, for example by asking it to show its reasoning (*chain-of-thought*),<sup>20</sup> exploring several solution paths and keeping the majority answer (*self-consistency*),<sup>44</sup> or searching over branching ideas (*tree-of-thoughts*).<sup>45,46</sup> Recent large reasoning models (e.g., OpenAI’s o3 or DeepSeek-R1) are specifically trained to reason (although traces are not always returned to the user), achieving high performance on complex (non-medical) tasks.<sup>47,48</sup>

Beyond these LLM-centered techniques, LLMs embedded in agentic systems operate in loops, mixing reasoning with actions and feedback (reason-act-observe loops, *ReAct*)<sup>49</sup> or separate planning from evidence gathering (*ReWOO*);<sup>50</sup> some add self-critique to improve the next attempt (*Reflexion*).<sup>51</sup>

### 2.4 Context, Memory and Learning

Context is the information available to the LLM when processing a request, including user instructions, conversation history, and any external information such as tool outputs.<sup>52</sup> Just as holistic image interpretation needs the right clinical information at the right time, agents require careful *context engineering*—providing the LLM with optimal information in the most effective format and timing.<sup>52,53</sup> Since LLMs have limited context windows (i.e., the information they can process at once), techniques like summarization help optimize what gets included.<sup>54</sup>

Beyond optimizing existing context, systems can enrich it dynamically through retrieval-augmented generation (RAG), which queries a (trusted) knowledge source (e.g., a database) and provides the LLM with results. RAG may augment each call automatically, be run as a tool, or operate agentically with dynamic retrieval and processing. In radiology, this can mean fetching prior reports, templates, or guidelines, analogous to how radiologists retrieve task-relevant additional information. The value of RAG depends on both the quality of the source information and the performance of the retrieval system.<sup>55–57</sup>

LLMs by themselves are stateless, meaning each new response depends only on the current input. Additional memory systems maintain continuity across interactions.<sup>32,34</sup> The LLM’s *internal knowledge* is fixed at training and not reliable for up-to-date, factual information.<sup>58</sup> Together with prompts or configurations provided by the serving framework, it forms the agent’s “procedural memory”.<sup>34</sup> Updating this knowledge requires model fine-tuning or framework changes. *Short term memory* functions as the agent’s working space. It holds the current conversation state, including recent observations from the environment and outputs from tools. This may be limited to the LLM’s context window or managed via external structures across multiple calls. Known as *state management*, this process involves tracking, updating, and discarding elements like conversation history or user preferences to avoid context drift or degradation (i.e., the gradual loss or distortion of relevant context that can compromise decision-making quality).<sup>59</sup> *Long term memory* includes both general facts (semantic memory) and records of past actions (episodic memory). It can store useful information for future reference, such as user preferences, successful response patterns, examples for learning, or guidelines (making it a frequent source for RAG systems<sup>34</sup>), but raises critical considerations of data governance and privacy when storing patient information for future use.

Some agent systems can improve over time by keeping track of past successes<sup>60,61</sup> (experiential learning). Even without human feedback, *self-evolving agents* can adapt their own skills, memory, and tools through rewards, imitation, or search across strategies.<sup>62</sup> Examples include agents that refine themselves by comparing against earlier versions ("self-play")<sup>63</sup> or by generating tasks with built-in verification and learning from the outcomes<sup>64</sup>. For radiology, such improvement capabilities could enable agents to adapt to user preferences like preferred terminology, structural conventions of individual radiologists or departments – a crucial feature for clinical adoption and integration into existing workflows.<sup>65</sup>

## 2.5 Design Patterns for Agentic Systems

LLM-driven systems can be organized by increasing level of complexity and autonomy: single LLM interactions, structured workflows, and systems with one or more autonomous agents. Each approach differs in how it handles *control flow* – the sequence and logic of executing actions required for a multi-step task.<sup>32</sup>

Control flows can take several forms: sequential (steps executed in order), branching (different paths based on conditions), parallel (multiple simultaneous tasks), or looping (repeating actions until a condition is met). In conventional software, programmers predefine these flows. LLM-driven agentic systems, however, can generate them dynamically.<sup>32</sup>

Several useful design patterns have emerged (Tab. 1, Fig. 2). The fundamental building block is a single interaction with (or "call" to) an *augmented LLM*, which may include tool or memory use (e.g., summarizing a provided report).

When a single call is insufficient, multiple calls can be combined into structured *workflows* for multi-step tasks (e.g., retrieving prior data, fetching a template, then drafting a report). Common control flow patterns in such workflows include *chaining* (using one output as the next input), *routing* (selecting among several paths), and *parallelization* (executing independent subtasks or repeated attempts with aggregation).<sup>11,66</sup>

When control flows cannot be predefined easily (for instance, conducting a systematic literature research), a more autonomous *agent* iterates (e.g., in reason-act-observe loops<sup>49</sup>) until a stop condition is met (goal reached, critical error, or budget exhausted). This enables more adaptive problem-solving.

*Multi-agent systems* (MAS) coordinate multiple agents that delegate tasks, communicate, and share tools and memory. They can be organized in different ways (topologies), for example as peer-to-peer networks (all agents collaborate) or hierarchically (supervisor agents hand off tasks to specialized sub-agents).<sup>67,68</sup> Another key design choice is between general-purpose agents with broad tool access and specialized agents focused on

narrow tasks. While generalists simplify coordination, specialists often achieve higher tool-use accuracy and speed, especially under hierarchical delegation.<sup>11,67,68</sup> MAS can outperform single-agent setups<sup>68,69</sup> and exhibit complex group behaviors,<sup>70</sup> but greater autonomy also amplifies oversight and error-propagation concerns.

Choosing how to structure the system depends on the task and the desired balance between control and flexibility. LLM-based *workflows* suit predictable, auditable tasks and are easier to maintain, with LLM decision-making autonomy limited to predefined points. Autonomous *agents* offer greater adaptability for complex problems but require more error handling and oversight, especially in MAS. Practitioners recommend adding complexity only when necessary.<sup>11</sup> This means starting with an augmented LLM call, using workflows when steps are known and control is critical (applying to many radiology tasks), and adopting agents when their flexibility justifies the extra cost and risk. Guardrails and evaluation depth should scale with system autonomy.

## 3 Radiology as Environment for Agents

Viewing radiology as an "environment" clarifies what an agent can observe and act upon, highlighting challenges unique to radiology. The field's environment is complex and multimodal: it includes images, structured and unstructured EHR information, metadata from radiology and hospital information systems, and the speech, gestures, or written communication of clinicians and patients. These characteristics directly influence the design of agentic systems.<sup>10</sup>

From a technical perspective, the radiology environment involves partial observability (e.g., incomplete patient data), a mix of episodic and longitudinal observations (single studies vs. follow-ups), and occasional real-time responsiveness (e.g., image-guided procedures). It is inherently multi-agent, encompassing radiologists, technologists, referrers, patients, and IT systems, with agents confronting missing data, evolving procedures, structured reporting, and cross-role coordination.

### 3.1 Radiology's Toolbox

Many tasks in radiology exceed what LLMs can handle in isolation. Agentic AI may hand off such tasks to specialized software (e.g., CAD tools), and other AI models like TotalSegmentator<sup>73</sup> for segmenting anatomical structures or foundation models<sup>17</sup> adapted for chest X-ray<sup>74,75</sup> or CT<sup>76,77</sup> analysis and reporting.

Agentic tool use can significantly enhance task performance: For example, Ferber et al. demonstrate a boost from 30% to 87% accuracy over isolated LLM use.<sup>78</sup> To achieve this, agents must "understand" what each tool does, when (and in which order) to use it, and how to interpret the output in the radiology context.<sup>79</sup> For example, lung nodule assessment might involve calling CAD soft-

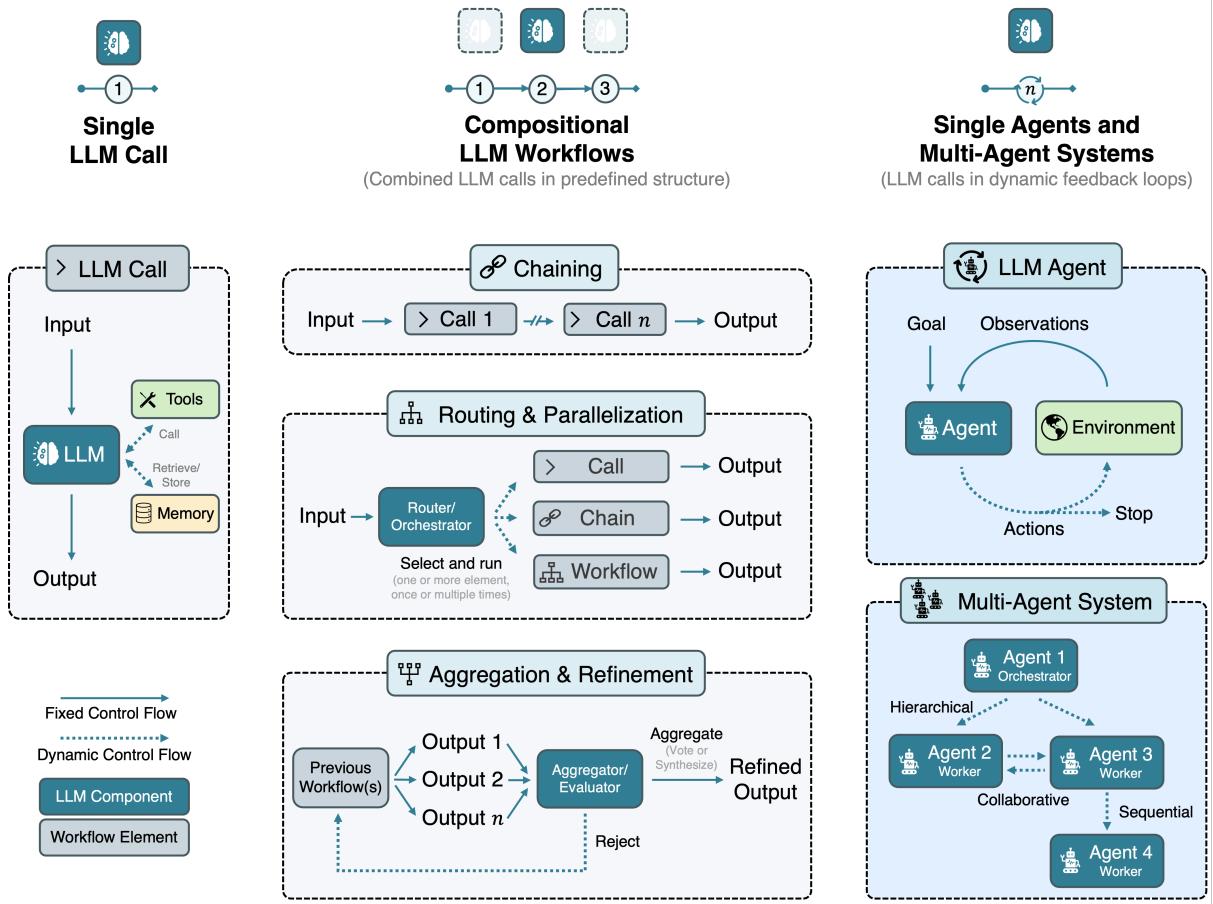


Figure 2: **Overview of building blocks and design patterns for LLM-based agentic systems.** The illustrated components are modular rather than mutually exclusive and can be combined to arbitrary complexity. (**Left column**) A single LLM call forms the basic building block, optionally reading from or writing to external tools or memory. (**Center column**) Multiple LLM calls can form workflows through (i) *chaining* in fixed sequences, (ii) *routing and/or parallelization* by an orchestrator LLM, or (iii) *aggregation and refinement* by an evaluator that synthesizes or rejects results. (**Right column**) Agent systems extend this pattern: a single agent interacts with its environment through observation–reason–action loops, while multi-agent systems organize agents hierarchically (e.g., manager–subagent), collaboratively (specialized roles), or sequentially. LLM: large language model.

ware and interpreting measurements returned in a structured format (e.g., JSON). An "agent-friendly" interface (e.g., MCP) helps by allowing the LLM to understand and communicate with the tools. For instance, RadFabric is an MCP-based multi-agent setup with specialized agents handling CXR analysis and report generation.<sup>80</sup>

### 3.2 Radiology-specific Knowledge Sources

Radiologists often consult literature or reference cases to refine diagnoses. Similarly, LLM-based agents can query databases such as PubMed, Radiopaedia, the RSNA Case Collection to incorporate up-to-date information.

Standardized image reporting systems (BI-RADS, PI-RADS, LI-RADS, Lung-RADS) reduce variability, ground recommendations in evidence, and facilitate interdisciplinary communication,<sup>81</sup> while broader frameworks such as TNM staging<sup>82</sup> and RECIST<sup>83</sup> integrate

imaging with clinical context. While some argue that LLMs could make natural language a universal interface in healthcare, reducing the reliance on fixed schemas,<sup>84</sup> validated ontologies offer human-readable and machine-computable representations of radiology concepts and relations, enabling semantic interoperability.<sup>85,86</sup> For example, SNOMED CT encodes clinical concepts and their relations to support structured documentation and data exchange. RadLex focuses on imaging-specific concepts and supports standardization efforts such as the RSNA-ACR Common Data Elements and the RSNA-LOINC Radiology Playbook.<sup>87</sup> The Radiology Gamuts Ontology (RGO) formalizes radiological differential diagnoses.<sup>88</sup> RadGraph-XL structures radiology data as *knowledge graph*, where nodes represent clinical entities (e.g., "right upper lobe") and edges encode relations (e.g., "suggestive of").<sup>89</sup>

Pattern	Core idea	When to use
<b>Basic Building Block</b>		
Augmented LLM call	Single LLM call enhanced with tools and memory.	Baseline choice; start here before adding complexity.
<b>Compositional Workflows</b>		
Chaining	Break a task into a fixed sequence of LLM calls; each call is informed by previous output.	Tasks with a natural linear decomposition (e.g., outline → draft → full report).
Routing	A router classifies the input, then dispatches to specialized workflows or models.	Tasks requiring one of several specialized models (e.g., call a workflow for CT analysis).
Sectioning	Split input into independent complementary subtasks, solve in parallel, then aggregate.	Tasks requiring a foreseeable number of predefined steps (e.g., retrieving patient appointments from EHR and previous studies from RIS to create a summary).
Voting	Run the same prompt multiple times and score/majority-vote results (self-consistency).	Safety checks, hallucination reduction.
<b>Agent Systems — Single-Agent</b>		
ReAct Agent	An agent iterates in a <b>Reason-Act-Observe</b> loop until a stop criterion is met.	Ill-defined or dynamic tasks where steps cannot be predetermined.
Plan-and-Execute	An agent produces a multi-step plan and follows it step-by-step (usually single-agent).	Tasks that allow or require long-term planning.
<b>Agent Systems — Multi-Agent (MAS)</b>		
Orchestrator-Workers	A central agent plans subtasks and delegates to worker agents, then merges results.	Broad, open-ended tasks, e.g., systematic research (e.g., Biomni <sup>71</sup> )
Evaluator-Optimizer	One agent drafts, another critiques and suggests fixes; loop until quality matches threshold.	Iteratively improving outputs, e.g. writer and reviewer agents taking turns to improve impression sections (RadCouncil). <sup>72</sup>
Peer-to-Peer (Swarm)	Multiple autonomous agents collaborate as equals without a central coordinator; involves communication (e.g., via shared memory).	Large-scale exploration, distributed ownership, or when central orchestration is a bottleneck. <sup>70</sup>

Table 1: **Design patterns of LLM-based workflows and agentic systems.** The basic building block is an LLM with access to tools and memory. Workflows combine multiple LLM calls in sequence, with branching or parallel logic, optionally aggregating the outputs (e.g., by scoring). Agents add autonomy by iteratively reasoning, acting, and observing feedback from the environment in loops until a stop criterion is met (e.g., goal reached, iteration budget exhausted). MAS extend this by coordinating multiple agents that communicate, allocate or negotiate subtasks, and act sequentially or concurrently. LLM: Large language model. MAS: Multi-agent system.

Grounding the agent in structured medical knowledge guides it to operate within consistent, interpretable categories and reduces the risk of clinically ambiguous outputs or confabulations.<sup>55,90</sup> For example, RadioRAG raises diagnostic performance on expert-curated tasks by up to 54%.<sup>57</sup>

### 3.3 Radiology’s Ecosystem

Radiology operates within a digital ecosystem that agents must interface with to access, process, and act on clinical data. This ecosystem includes the Hospital Information System (HIS) for patient administration and clinical history, the Radiology Information System (RIS) for radiology-specific tasks like scheduling and reporting, and the Picture Archiving and Communication System (PACS) for image storage and distribution. PACS is often complemented by a vendor-neutral archive (VNA) for long-term retention. As data volumes grow, data lakes and warehouses are increasingly adopted to support analytics and AI workflows.<sup>91</sup>

Interoperability across these systems relies on standardized communication protocols. DICOMweb extends the DICOM standard, which governs the storage and transmission of images, structured reports, and segmentation data, into modern RESTful web APIs that allow communication using simple, standardized web requests, enabling scalable, network-based integration with agents. Fast Healthcare Interoperability Resources (FHIR) builds

on the Health Level 7 (HL7) standard for clinical data exchange by defining modular resources (e.g., "ImagingStudy") and by supporting semantic interoperability through established vocabularies like SNOMED CT, RadLex, and LOINC.<sup>91</sup> SMART on FHIR introduces secure OAuth2-based access control, while FHIR Subscriptions and FHIRcast provide real-time event updates, allowing an agent to be notified when a new study arrives or when a radiologist opens a case.

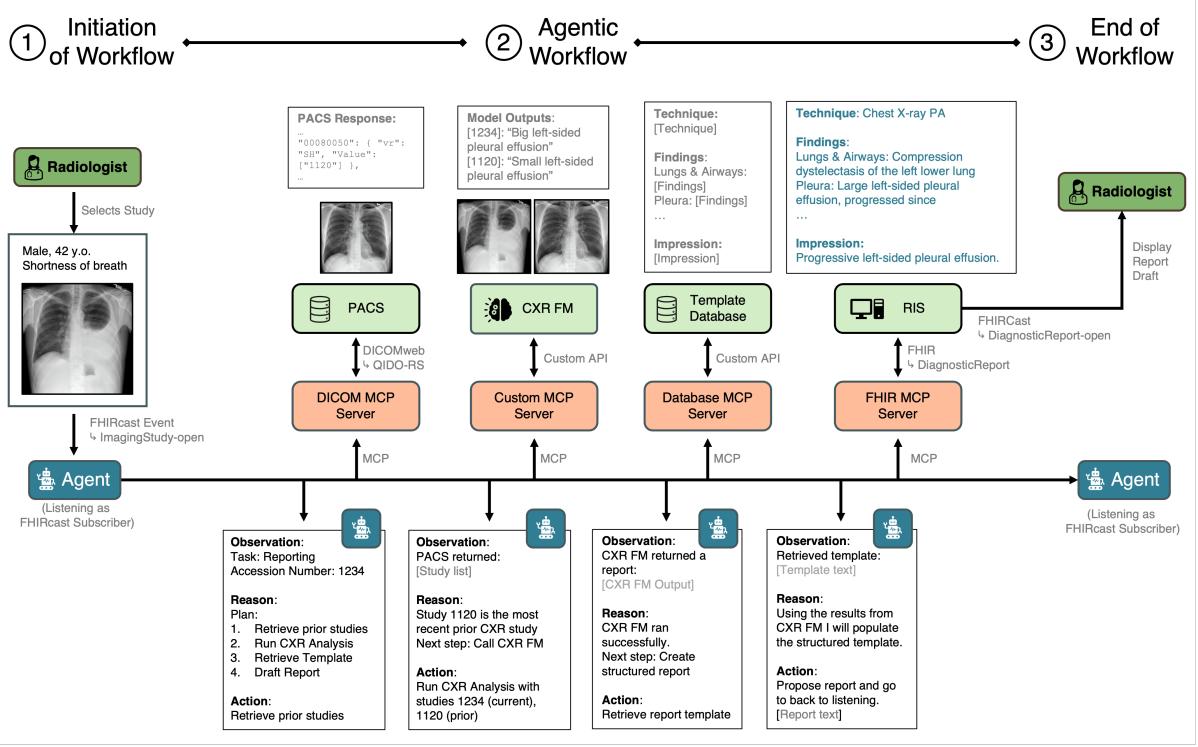
Building on top of these protocols (optionally after wrapping them in MCP<sup>92</sup>), developers can create complex agentic radiology applications while leveraging existing infrastructure for security and stability(Fig. 3).

## 4 Applications in Radiology

While agentic AI has not entered radiology practice, early studies in other medical fields suggest potential for improved triage, decision-making, and efficiency.<sup>93</sup> Here, we walk through exemplary agentic solutions for radiology at varying degrees of autonomy.

### 4.1 Chest X-ray Consistency Checker

When a chest radiograph is opened in the viewer, an agent monitors the dictation stream and concurrently retrieves prior chest films for comparison. As the radiologist begins their impression, the agent asynchronously checks



**Figure 3: Example of an agentic workflow for report drafting within traditional radiology infrastructure.** The workflow begins with a radiologist selecting a study (Step 1), triggering an FHIRcast event that notifies the agent. The agent plans the reporting task by retrieving prior studies from PACS through a DICOM MCP server (Step 2). It then calls a chest X-ray foundation model (CXR FM) via a custom MCP server to analyze the current and prior images. Once the model returns findings, the agent retrieves a structured report template from a template database via a database MCP server and populates it with model output. The structured draft report is sent to the radiology information system (RIS) using the FHIR protocol via an FHIR MCP server. The workflow concludes when the radiologist receives and reviews the draft report (Step 3). The agent continuously observes, reasons, and takes actions throughout the process via Model Context Protocol (MCP) interactions. DICOM: Digital Imaging and Communications in Medicine. CXR: Chest X-Ray. API: Application Programming Interface. FHIR: Fast Healthcare Interoperability Resources. FM: Foundation Model. MCP: Model Context Protocol. PACS: Picture Archiving and Communication System. RIS: Radiology Information System.

for line- and tube placement, and comparison statements. If, for example, the dictation lacks a reference to a newly placed device detected in the image, the agent highlights this as a possible omission. A prompt appears: "Previous film (2 days ago) had no left CVC. Include new line in report?" Upon confirmation, the agent suggests templated language. All edits are logged (for posterity, personalization and future improvements), and the radiologist maintains control over final wording.

## 4.2 Agentic Lung Cancer Screening Reporting

Lung cancer screening (LCS) programs are increasingly implemented, leading to greater workloads.<sup>94</sup> Their structured, multi-step nature makes them ideal for workflow-based automation with limited autonomy. In this setting, an agent manages routine steps while allowing the radiologist to focus on interpretation.

When an exam is opened, the agent identifies the LCS scenario, retrieves validated workflow-specific instruc-

tions, and pulls patient data from PACS, RIS, and HIS, including prior CTs and smoking history. AI-driven image registration<sup>95</sup> aligns current and prior scans to support longitudinal nodule tracking in side-by-side comparison, a process that is typically time-consuming. A CAD model detects nodules, producing structured descriptors (size, texture, lobe location). After radiologist review, the agent assigns Lung-RADS categories, loads the appropriate report template, and drafts the report. Internal consistency checks are applied, and any radiologist edits are logged to personalize future drafts. Once signed, follow-up recommendations are communicated to the referring provider.

This workflow uses DICOMweb for image management, FHIR for structured clinical data exchange with the EHR, and RAG to pull guideline content from a curated database.

#### **4.3 Agent-Assisted Tutoring through Interactive Reporting**

Interactive learning assistance in routine reporting illustrates a more autonomous, conversational workflow. In this setting, the agent engages flexibly with users rather than following a fixed structure. A resident drafting a report can consult an integrated tutor agent to clarify findings or better understand concepts. The agent answers free-text queries with targeted teaching, using curated cases and generative tools to create illustrative examples.

For example, if the resident asks about a pneumonic infiltrate near the right heart border on a chest radiograph, the agent explains the silhouette sign and generates paired synthetic images showing middle and right lower lobe.<sup>96</sup> Once the resident is finished drafting, the agent checks for internal inconsistencies such as laterality mismatches and flags missing responses to relevant clinical questions.

In this example, the agent can provide timely, specific feedback and flexibly adjust to the resident by retrieving additional content (e.g., through RAG) or generate illustrative examples on-the-fly (through a tool call). All interactions can be logged for longitudinal skill tracking.

#### **4.4 Enhanced Multidisciplinary Team Discussions**

Multidisciplinary team discussions (MDT) rely on thorough preparation and documentation, as well as rapid access to additional information during the meeting. Often, radiologists manage the displayed content, manually retrieving additional imaging and pertinent records during the discussion. This scenario benefits from mix of predefined workflows and adaptive agentic assistance.

Before the meeting, an agent can assemble the case list by extracting information from communications, scheduling systems, and prior meeting notes, while verifying the completeness of imaging, pathology, and laboratory data. It then condenses relevant history, imaging findings, and treatment timelines into concise case summaries that are shared with participants in advance.

During the meeting, a domain-adapted automatic speech recognition (ASR) engine transcribes the discussion with high accuracy.<sup>97</sup> The agent monitors the transcript to detect information requests, then queries the EHR to retrieve and display relevant data such as prior imaging, pulmonary function tests, or medications, all in real time.

After the meeting, the agent consolidates the transcript, the retrieved data, and the decisions recorded during the session. It generates a structured summary containing patient-specific conclusions, assigned responsibilities, and relevant supporting images or reports. This summary is stored in the EHR, shared with participants, and used to trigger follow-up actions such as scheduling, ordering tests, or notifying referring clinicians.

#### **4.5 Agent-Driven Follow-up Scheduling**

Scheduling follow-up imaging is a repetitive but essential process that often involves multiple parties beyond radiology. An agent can receive a referral, determine the recommended interval based on guidelines and the prior exam date, and query hospital scheduling systems for equipment and staff availability. It can then contact the patient via phone or secure portal to propose available time slots and explain the clinical importance of the follow-up.

Once the appointment is accepted, the agent suggests an appropriate imaging protocol, seeks radiologist approval where needed, and finalizes the booking across RIS, PACS, and HIS. Notifications are sent to all stakeholders, and reminders are issued automatically. A rule-based check ensures correct protocolling for each referral type, and audit logs maintain a record of all actions for compliance.

### **5 Evaluation of Radiology Agents**

Radiology agents navigate complex environments where they interpret open-ended queries, plan actions, adapt when results are unexpected, while still delivering useful outputs. Metrics like AUROC that suffice for narrow AI systems (e.g., CAD for pneumothorax detection) cannot capture this process. Comprehensive evaluation must therefore consider planning, execution, outcomes, and system-level performance (Fig. 4).

#### **5.1 Planning**

Planning begins with identifying the task and outlining a logical sequence of steps. Evaluation assesses whether the agent understood the request, proposed an appropriate plan, and avoided unnecessary steps. Because downstream reliability depends on plan clarity, ambiguity in goals or task structure can lead to execution errors; well-specified plans reduce uncertainty and allow the agent to act confidently.<sup>39,49</sup> Replanning is equally important; if required tools are unavailable, effective agents adapt without losing sight of the goal or terminate when the task is *a priori* impossible. Planning evaluation also considers responses to redundant, suboptimal, or incomplete tool palettes.<sup>79</sup>

Plans can be compared to expert or strong LLM ("oracle") reference plans using distance metrics (similarity of proposed vs. reference plans) or stepwise matching accuracy, especially for order-sensitive tasks (e.g., segmentation before volume calculation). Expert review can also judge whether chosen actions were necessary, helpful, and reliable, particularly when multiple tools overlap.<sup>78</sup> However, for more complex or dynamically changing tasks, it can be difficult to determine an "optimal" plan or even what to compare against in advance.

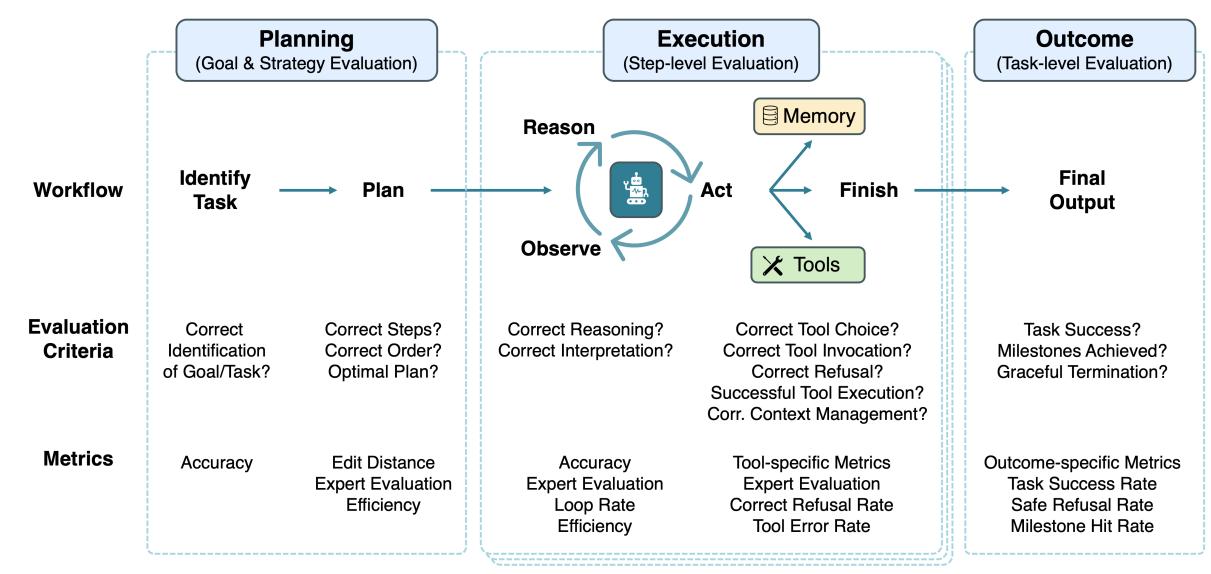


Figure 4: **High-level Evaluation Framework for Agentic Workflows.** This framework decomposes agent behavior into four tiers: Planning, Execution, Outcome and System-level evaluation (not shown). Planning assesses task identification and strategy formulation; Execution evaluates reasoning and decision-making, tool use, and memory management at each step of iterative cycles; and Outcome measures overall task success and termination quality. Note this figure omits system-level performance evaluation (e.g., costs, long-term effects) for clarity.

## 5.2 Execution

Execution is the process of carrying out the planned steps. For autonomous agents, this involves iterative cycles of reasoning, acting, and observing, while in structured workflows it often means following predefined sequences with constrained decision points. Step-level evaluation assesses reasoning accuracy, action quality, and appropriate use of tools and memory.

Tool evaluation checks whether the correct tool was used at the right time with appropriate inputs, whether the tool functioned as intended, and whether outputs were correctly interpreted. Robust execution also requires handling missing or ambiguous data and issuing safe refusals when a step is impossible or unsafe. For example, if a multiphase study is missing a crucial series, the agent should clarify whether the series was omitted from PACS or adjust the plan, rather than looping or hallucinating results. An agent may still succeed despite imperfect intermediate steps, or fail after a sequence of sound ones.

Memory use is evaluated by whether contextually relevant information (e.g., allergies or prior diagnoses) was retrieved and maintained across updates.<sup>98–100</sup>

Metrics include correctness of intermediate outputs, tool calls, and memory interactions; milestone hit rate for partial progress;<sup>79</sup> refusal rates; loop frequency; and efficiency in actions, time, and resources. Expert assessment can additionally judge whether the reasoning and action sequence was logical.<sup>78</sup>

## 5.3 Outcome

Outcome evaluation asks if the task was completed correctly and safely, regardless of how efficient or circuitous the process was. Beyond simple task success rate, many task-specific metrics apply (e.g., Dice score for segmentation performance).<sup>101–103</sup> For close-ended tasks (with definite solutions) this may be exact accuracy against a reference standard (although such a standard can be hard to establish in radiology); for open-ended tasks (with more than one valid solution) like report generation it may be expert (or LLM) rating of output quality (for instance based on similarity compared to a reference output), freedom from hallucinated findings, appropriateness of recommendations and criticality of errors. Outcomes can also be assessed for reliability across repeated runs,<sup>104</sup> for instance via *pass@k* (i.e., the probability that at least one of  $k$  attempts succeeds) or *pass<sup>k</sup>* (i.e., the probability that all  $k$  attempts succeed, thereby capturing consistent rather than occasional success for critical tasks<sup>105</sup>), calibration of uncertainty, and the agent’s ability to terminate gracefully (i.e., without causing additional problems) when no satisfactory solution exists (e.g., essential tools are unavailable).

## 5.4 System-level Performance

System-level evaluation examines how agents perform beyond successful and efficient completion of workflows. It focuses on downstream effects such as radiologist efficiency,<sup>74,106,107</sup> workflow integration, long-term reliability and safety, and patient outcomes. It also considers

how human–agent collaboration affects cognitive load and diagnostic performance, whether biases are exposed or amplified, and risks of clinician deskilling. Broader dimensions include robustness to distribution shifts, agentic improvement over time, and resource demands like compute and energy.<sup>108</sup>

These evaluations depend on real-world deployments or realistic simulations that capture interactions among agents, IT systems, clinicians, staff, and patients, but such benchmarks currently remain underdeveloped.

### 5.5 Conclusions for comprehensive evaluation

Comprehensive evaluation integrates planning, execution, outcome, and system-level performance. No current medical agent benchmark spans all tiers, though several address individual levels (Tab. 2). RadABench is a notable example of a comprehensive benchmark for agentic radiology AI, with fine-grained evaluation of planning quality and adherence, and success and robustness of tool use, including in MAS.<sup>79</sup>

Despite their widespread use, static benchmarks such as USMLE-style multiple-choice questions offer a limited view of agent performance.<sup>109</sup> These tasks often assess isolated knowledge recall, rather than dynamic clinical reasoning.<sup>110</sup> They are often publicly available and may have been seen during training, which compromises their validity. As such, they are best interpreted as partial indicators within a broader framework.

Given the adaptability of LLM-based agents, evaluation should move beyond static benchmarks to interactive, multi-task, multi-modal simulations that mimic real clinical complexities including noise, missing data, and ambiguities. This extends to stress testing through repeated runs, perturbations, and toolset limitations to reveal weaknesses in robustness and recovery.

Inspiration may come from objective structured clinical examinations (OSCEs) in medical education, which use standardized patients, scenarios, and multimodally enriched cases in timed stations, completed in an interactive way and scored by experts with structured checklists and global rating scales. The Sequential Diagnosis Benchmark<sup>111</sup> treats diagnosis as an iterative interaction and measures intermediate decision quality, final accuracy, and resource use. A similar benchmark could be developed for radiology agents.

Guidelines can help structure evaluations. CLAIM<sup>112</sup> provides a radiology-specific reporting checklist for transparent AI studies; DEAL<sup>113</sup> outlines best practices for developing, evaluating, and assessing LLMs in medicine (mentioning agents); TRIPOD-LLM<sup>114</sup> extends transparent reporting to LLM studies with emphasis on data provenance, human oversight, and reproducibility; and DECIDE-AI<sup>115</sup> focuses on early-stage clinical evaluation (pre-deployment) with attention to human factors, safety, and real-world performance. While

these frameworks support evaluation of LLMs, agent-specific guidelines are still lacking.<sup>116</sup>

## 6 Challenges

LLM-driven workflows and agents expand what radiology AI can do but introduce challenges beyond single-turn LLMs (Table 3).

**LLM core limits.** Agents inherit fundamental limitations of their LLM backbone: stochasticity, confabulations, bias, and poor confidence calibration as discussed in other works.<sup>16, 123</sup> Context enrichment (through RAG, memory and tool use) alleviates but does not eliminate these risks entirely. In multimodal agents, cross-modal reasoning (e.g., between a CXR and text) can be brittle, and intransparent, especially if decision traces are not inspectable.

**Cascading errors and context volatility.** More than static models, agents are vulnerable to cascading errors and context degradation, where propagated inaccuracies compound across many turns. Consider a radiologist reports a "12 mm part-solid ground-glass nodule" (recommendation CT in 3-6 months), agent *A* changes this to "12 mm ground-glass nodule" (CT in 6-12 months), and agent *B* reports "12 mm pulmonary opacity" (ambiguous finding, no clear recommendation). Robust step-wise validation (e.g., "Are all relevant nodule characteristics reported?") can mitigate these risks.

**Multi-agent coordination.** Multi-agent coordination presents additional complexity through resource contention and communication failures. In radiology, these issues could manifest as contradictory outcomes, duplicated work, or omission of critical steps. Bedi et al. found that optimizing individual components in multi-agent systems paradoxically reduced overall performance due to impaired information flow and inter-agent compatibility,<sup>124</sup> underscoring the importance of system-level design and validation for agentic AI applications in radiology.

**Integration, governance and human-AI interaction.** Integration of agentic AI into radiology requires more than interoperability with existing systems: It raises security, regulatory, and governance concerns that existing frameworks are not fully equipped to handle. Agentic systems introduce new cybersecurity vulnerabilities like prompt injection attacks.<sup>125, 126</sup>

Their autonomy and adaptability also reduce predictability and human oversight. Agents can act without timely intervention, increasing the risk of error propagation and unexpected behavior. As they adapt to new cases, their behavior may drift, complicating validation and undermining prior regulatory approvals. This makes it essential to define clear boundaries for agent behavior and identify when human review is required.

Year	Benchmark	Multi-Agent	Planning	Execution	Summary
2024	CRAFT-MD <sup>117</sup>	✓	—	✓	Simulated conversational clinical reasoning (Diagnosis, (V)QA, history taking)
2024	AgentClinic <sup>61</sup>	✓	—	✓	Conversational, multimodal diagnosis (Diagnosis, (V)QA)
2024	MIMIC-CDM <sup>118</sup>	—	✓	✓	CDM (diagnosis, treatment recommendation) for abdominal pathologies
2024	MedChain <sup>119</sup>	✓	—	✓	CDM (referral, history+exam, diagnosis, treatment) with 12k EHR cases
2024	RadABench <sup>79</sup>	✓	✓	✓	Radiology tasks with fine-grained plan and tool evaluation
2024	SDBench <sup>111</sup>	(✓)	✓	(✓)	Sequential CDM (diagnosis by asking questions & ordering tests) with cost evaluation
2025	MedAgentBench <sup>120</sup>	—	—	(✓)	EHR query and ordering tasks in FHIR environment
2025	MedAgentBoard <sup>121</sup>	✓	—	(✓)	Single or MAS for medical QA, summarization, prediction and workflow automation
2025	MedAgentsBench <sup>122</sup>	(✓)	—	—	Multi-step clinical reasoning for diagnosis, with cost evaluation

Table 2: **Medical agent benchmarks.** Benchmarks were selected for their medical and radiological scope and offering sufficient reproducibility. All benchmarks evaluate for task success; none of the included benchmarks evaluates for system-level performance beyond cost evaluation. *Multi-Agent*: explicit inclusion of multi-agent systems. *Planning*: explicit evaluation of planning quality beyond task success (e.g., comparison with a reference plan, human review). *Execution*: explicit evaluation of the execution quality beyond task success (e.g., tool-call quality, step-wise failure analysis). Symbols: ✓full, (✓) partial/implicit, —not present. Abbreviations: CDM: clinical decision-making. EHR: electronic health record. FHIR: Fast Healthcare Interoperability Resources. (V)QA: (visual) question answering.

Most medical device regulations were designed for static, narrow-scope software and are poorly suited to autonomous, adaptive agents.<sup>127</sup> Proposals to address this gap include staged approvals, predefined update protocols, regulatory sandboxes, and outcome-based evaluation.

The dynamics of human–AI interaction and the effects on high-level outcomes like radiologist efficiency or patient health remain underexplored. Early studies show potential for agent-led tumor boards,<sup>78</sup> diagnostic collaboration,<sup>128</sup> and improved reporting efficiency with LLM-assisted workflows,<sup>74,106</sup> but also raise concerns about clinician deskillings.<sup>129</sup> Integrating AI agents into practice requires clarifying shared responsibilities and decision-making,<sup>130</sup> and addressing human AI-interaction biases resulting in overreliance on, or distrust in AI outputs.<sup>131</sup>

## 7 Conclusion

LLM-driven agentic systems offer radiology a path from single-step assistance toward adaptive, multi-step automation. By offloading repetitive, non-critical tasks that contribute to cognitive load, the promise is to help radiologists refocus on high-value work, ideally at the peak of their competence. Realizing this vision requires more than technical implementation: future work must develop clinically relevant, holistic benchmarks that evaluate system-level effects; ensure robust integration with systems and human stakeholders; and rethink human–AI interaction to balance oversight, trust, and collaboration. With careful design and governance, agentic AI can evolve from experimental prototypes into valuable assistants, helping radiology adapt to rising demands while maintaining and potentially elevating quality.

Failure Mode	Risk / Impact	Clinical Example	Mitigation
<b>Level: Data Grounding — What does the agent get to see?</b>			
Outdated, incomplete, or ambiguous input	Misguided decisions	Missing biopsy report leads to broader differential	Input validation; RAG optimization; human review
<b>Level: LLM — Where can the model fail?</b>			
Biased or insufficient internal knowledge	Misinterpretation; flawed, potentially unsafe output	Agent reports "hyperdense" lesion on MRI instead of "hyperintense"	RAG with trusted sources; fine-tuning; bias audits
Confabulation	Plausible but false output	Agent cites a non-existent guideline	Response validation; RAG
Crossmodal reasoning error in multi-modal LLMs	Mismatch between visual information and text output	Agent invents a nodule or flips left/right	Modality-specific tuning; hand-off to validated vision models; expert review
Role or goal misunderstanding	Workflow errors; misaligned actions	Agent conducts broad literature review when asked for a specific guideline reference	Goal validation; clear role constraints
Tool misuse or failure	Incorrect output or missed findings	Agent misreads failed CT tool output as "no nodules"	Task-specific validation; stress testing
Unrecognized uncertainty	Missed escalation to human review	Liver mass flagged as malignant without alerting radiologist	Confidence thresholds; escalation channels; fallback rules
<b>Level: Execution — Where can the process fail?</b>			
Context degradation	Loss or drift of information; incorrect reasoning	Prior cancer history lost mid-task	Prioritized context; sliding-window memory; retrieval refresh
Cascading errors	Compounded failures	Confabulated lesion → wrong guideline → wrong recommendation	Fixed validation checkpoints; rollback; self-reflection
Opaque reasoning trace	Reduced possibility to audit or debug	Postulating metastatic disease without providing evidence	Structured logs; provenance tracking; explainability tools
Multi-agent miscoordination	Redundancy or conflict	Two agents write conflicting findings into report	Arbiter agent; task quotas; A2A protocols
Emergent misbehavior	Unintended/unforeseen actions	Agent cancels scheduled exams to save time	Execution sandbox; autonomy limits; active monitoring; human-in-the-loop confirmation for critical steps
<b>Level: Environment &amp; Humans — What happens in the real world?</b>			
Poor IT integration	Broken workflows	AI report fails to transfer to PACS due to format mismatch; radiologist re-dictates manually	Validated interfaces (e.g. FHIR) to existing components
(Novel) Security vulnerabilities	Attack surface increases	Compromised RAG source injects misleading info that gets interpreted as prompt	Vendor vetting; layered security; prompt defense
Cross-department silos	Incomplete information	Histopathology information not accessible to agent preparing MDT case vignette	Unified system access; interdepartmental integration
Unclear human–AI roles	Overreliance (automation bias), mistrust (algorithmic aversion bias); Deskilling	Radiologist misses fracture after AI says "normal"	Confidence calibration; training; explainability; safety roles
Limited external validation	Poor generalization	System underperforms at new hospital	Diverse benchmarks; prospective trials
Unclear accountability	Legal risk	Malpractice claim in AI-involved workflow	Clear roles; audit logs; liability protocols
Regulatory drift	Certification gaps	Pipeline updated without regulatory notice	Gap analysis; QMS integration
High environmental cost	Sustainability concerns	Weekly retraining on full PACS archive	Green compute targets; workload monitoring

Table 3: Failure modes and mitigation strategies across layers of LLM-based radiology agents. LLM: Large Language Model; RAG: Retrieval-Augmented Generation; A2A: Agent-to-Agent (protocol); PACS: Picture Archiving and Communication System; FHIR: Fast Healthcare Interoperability Resources; MDT: Multidisciplinary Team; QMS: Quality Management System.

## Acknowledgments

**C.B.** received research support from the Promedica Foundation, Switzerland. **D.V.V.** is an employee of HOPPR, which develops AI technology for radiology applications. **D.T.** received honoraria for lectures by Bayer, GE, Roche, Astra Zenica, and Philips and holds shares in StratifAI GmbH, Germany and in Synagen GmbH, Germany. D.T. is supported by the German Ministry of Research, Technology and Space (TRANSFORM LIVER - 031L0312C, SWAG - 01KD2215B, DECIPHER-M - 01KD2420B), Deutsche Forschungsgemeinschaft (DFG) (515639690), and the European Union (Horizon Europe, ODELIA - GA 101057091, ERC Starting Grant SAGMA – GA 101222556). **J.N.K.** declares consulting services for AstraZeneca, MultiplexDx, Panakeia, Mindpeak, Owkin, DoMore Diagnostics, and Bioptimus. Furthermore, he holds shares in StratifAI, Synagen, and Spira Labs, has received an institutional research grant from GSK, and has received honoraria from AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, and Fresenius. **A.C.** receives research support from NIH grants R01 HL167974, R01HL169345, R01 AR077604, R01 EB002524, R01 AR079431, P41 EB027060, P50HD118632; Advanced Research Projects Agency for Health (ARPA-H) contracts AY2AX000045 and 1AYSAX0000024-01; and the Medical Imaging and Data Resource Center (MIDRC), which is funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under contract 75N92020C00021 and through ARPA-H. Unrelated to this work, A.C. receives research support from GE Healthcare, Philips, Microsoft, Amazon, Google, NVIDIA, Stability; has provided consulting services to Patient Square Capital, Chondrometrics GmbH, and Elucid Bioimaging; is co-founder of Cognita; has equity interest in Cognita, Subtle Medical, LVIS Corp, Brain Key. **C.P.L.**: This work was supported in part by the Medical Imaging and Data Resource Center (MIDRC), which is funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under contract 75N92020C00021 and through the Advanced Research Projects Agency for Health (ARPA-H). Research support from National Institutes of Health (NIH), National Institute for Biomedical Imaging and Bioengineering (NIBIB) and National Heart, Lung and Blood Institute (NHLBI), National Cancer Institute (NCI), Agency for Health Research and Quality (AHRQ); Advanced Research Projects Agency for Health (ARPA-H), and the Gordon and Betty Moore Foundation; consulting fees from Sixth Street and Gilmartin Capital; Honoraria for Lectures from Singapore Ministry of Health, Philips Medical, Canon Medical, McKinsey & Company; Support for Travel from Singapore Ministry of Health, Philips Medica, McKinsey & Company, C.P.L reports Patents and Patents Pending: Generalizable Machine Learning Medical Protocol Recommendation. Collaboration with GE Healthcare, India; Provision Patent Application No. 202141024621, filed November 25, 2021; Medical Autoencoders and Image Compression. Disclosure submitted with collaborators from Stanford, August 7, 2024; A System Based on a Large Language Model That Generates Sets of Patient-Centered Materials to Explain Radiology Report Information, Disclosure submitted July 24, 2025. Fiduciary duties: Board of directors, Bunkerhill Health; Board of directors Sirona Medical; Board of directors and President, Radiological Society of North America. Equity Shareholder, Bunkerhill Health, Sirona Medical; Option holder of Whiterabbit.ai, Galileo CDS, ADRA.ai, Cognita, TurboRadiology; Grants and Gifts: Grants and gifts to my institution, department, and/or research center from AWS, BunkerHill Health, Carestream, CARPL.ai, Clarity, GE HealthCare, Google Cloud, IBM, Kheiron, Lambda, Lunit, Microsoft, Nightingale Open Science, Philips, Siemens Healthineers, Stability.ai, Subtle Medical, VinBrain, Visiana, Whiterabbit.ai. Royalties: Data set license fee: annalise.ai **F.N.** received research support from the Digitalization Initiative of the Zurich Higher Education Institutions (DIZH)- Rapid Action Call - under the TRUST-RAD project and Project Call - under the RADICAL project, and from the EU Horizon Europe programme: DataGEMS (GA.101188416). All other authors declare no relevant conflicts of interest or acknowledgments.

## Author Contributions

C.B. and F.N. conceptualized the project and created the first draft of the manuscript. C.B., M.P., D.T., T.F. and C.P.L. provided radiological perspectives. D.V.V., J.N.K., F.N., M.M., A.C., C.P.L. and M.K. provided technical advice. D.V.V., M.P., D.T., J.N.K., M.M., A.C., T.F., F.N. and C.P.L. critically revised the draft.

## References

- <sup>1</sup> Thomas C. Kwee and Robert M. Kwee. Workload of diagnostic radiologists in the foreseeable future based on recent (2024) scientific advances: Updated growth expectations. *European Journal of Radiology*, 187:112103, June 2025.
- <sup>2</sup> Noushin Yahyavi-Firouz-Abadi. Preserving the Academic Mission Amid Radiologist Shortages and Financial Pressures. *American Journal of Roentgenology*, January 2025.
- <sup>3</sup> Anna Rozenshtein, Laura K. Findeiss, Monica J. Wood, George Shih, and Jay R. Parikh. The U.S. Radiologist Workforce: AJR Expert Panel Narrative Review. *American Journal of Roentgenology*, December 2024.
- <sup>4</sup> Ryan Han, Julián N Acosta, Zahra Shakeri, John P A Ioannidis, Eric J Topol, and Pranav Rajpurkar. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *The Lancet Digital Health*, 6(5):e367–e373, May 2024.
- <sup>5</sup> Rajesh Bhayana. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. *Radiology*, 310(1):e232756, January 2024.
- <sup>6</sup> Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Małgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, pages 1–9, February 2024.
- <sup>7</sup> Tao Tu, Mike Schaeckermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, June 2025.
- <sup>8</sup> Daniel McDuff, Mike Schaeckermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan. Towards Accurate Differential Diagnosis with Large Language Models. *Nature*, 642(8067):451–457, April 2025.
- <sup>9</sup> Lilian Weng. LLM Powered Autonomous Agents. <https://lilianweng.github.io/posts/2023-06-23-agent/>, 2023. Accessed 11 April 2025.
- <sup>10</sup> Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2020.
- <sup>11</sup> Erik Schluntz and Barry Zhang. Building Effective Agents. <https://www.anthropic.com/engineering/building-effective-agents>, 2025. Accessed: 2025-03-10.
- <sup>12</sup> Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.
- <sup>13</sup> Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning series. MIT Press, Cambridge, MA, 2nd edition, 2018.
- <sup>14</sup> David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.
- <sup>15</sup> Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement Learning in Healthcare: A Survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.
- <sup>16</sup> Christian Bluethgen, Dave Van Veen, Cyril Zakka, Katherine E Link, Aaron Hunter Fanous, Roxana Daneshjou, Thomas Frauenfelder, Curtis P Langlotz, Sergios Gatidis, and Akshay Chaudhari. Best Practices for Large Language Models in Radiology. *Radiology*, 315(1):e240528, April 2025.
- <sup>17</sup> Magdalini Paschali, Zhihong Chen, Louis Blankemeier, Maya Varma, Alaa Youssef, Christian Bluethgen, Curtis Langlotz, Sergios Gatidis, and Akshay Chaudhari. Foundation Models in Radiology: What, How, Why, and Why Not. *Radiology*, 314(2):e240597, February 2025.

- <sup>18</sup> Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.
- <sup>19</sup> Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- <sup>20</sup> Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 24824–24837. Curran Associates, Inc., 2022.
- <sup>21</sup> Rishi Hazra, Gabriele Venturato, Pedro Zuidberg Dos Martires, and Luc De Raedt. Have Large Language Models Learned to Reason? A Characterization via 3-SAT Phase Transition. *arXiv preprint arXiv:2504.03930*, 2025.
- <sup>22</sup> Taylor Webb, Shanka Subhra Mondal, and Ida Momennejad. A Brain-inspired Agentic Architecture to Improve Planning with LLMs. *Nature Communications*, 16(1):8633, 2025.
- <sup>23</sup> Subbarao Kambhampati, Karthik Valmeeckam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22895–22907. PMLR, July 2024.
- <sup>24</sup> Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems*, volume 36 of *NeurIPS*, 2023.
- <sup>25</sup> Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the Reasoning Abilities of Multimodal Large Language Models (MLLMs): A Comprehensive Survey on Emerging Trends in Multimodal Reasoning. *CoRR*, abs/2401.06805, January 2024.
- <sup>26</sup> Farhad Nooralahzadeh, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Explainable Multi-Modal Data Exploration in Natural Language via LLM Agent. *arXiv preprint arXiv:2412.18428*, 2024.
- <sup>27</sup> PwC. PwC's AI agent survey. <https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-agent-survey.html>, 2025. Accessed: 2025-10-07.
- <sup>28</sup> Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990.
- <sup>29</sup> Goonmeet Bajaj, Srinivasan Parthasarathy, Valerie L. Shalin, and Amit Sheth. Grounding From an AI and Cognitive Science Lens. *IEEE Intelligent Systems*, 39(2):66–71, 2024.
- <sup>30</sup> Bing Liu. Grounding for Artificial Intelligence. *arXiv preprint arXiv:2312.09532*, 2023.
- <sup>31</sup> Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards Visual Grounding: A Survey. *arXiv preprint arXiv:2412.20206*, 2024.
- <sup>32</sup> Chip Huyen. *AI Engineering: Building Applications with Foundation Models*. O'Reilly Media, 2025.
- <sup>33</sup> Yingxuan Yang, Huacan Chai, Yuanyi Song, Siyuan Qi, Muning Wen, Ning Li, Junwei Liao, Haoyi Hu, Jianghao Lin, Gaowei Chang, Weiwen Liu, Ying Wen, Yong Yu, and Weinan Zhang. A Survey of AI Agent Protocols. *arXiv preprint arXiv:2504.16736*, 2025.
- <sup>34</sup> Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive Architectures for Language Agents. *Transactions on Machine Learning Research*, 2024.
- <sup>35</sup> Harrison Chase and the LangChain Contributors. LangChain: A framework for developing applications with language models. <https://github.com/langchain-ai/langchain>, 2025.
- <sup>36</sup> Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri A. Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

- <sup>37</sup> Hugging Face. smolagents: a barebones library for agents that think in code. <https://github.com/huggingface/smolagents>, 2025. Accessed 2025-07-06.
- <sup>38</sup> Georg Wölflein, Dyke Ferber, Daniel Truhn, Ognjen Arandjelović, and Jakob Nikolas Kather. LLM Agents Making Agent Tools. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26092–26130, Vienna, Austria, 2025. Association for Computational Linguistics.
- <sup>39</sup> Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*, 2016.
- <sup>40</sup> Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification Gaming: The Flip Side of AI Ingenuity. DeepMind Blog, <https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/>, April 2020. Accessed 1 September 2025.
- <sup>41</sup> Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual Planning: Let's Think Only with Images. may 2025. arXiv:2505.11409, v2 last revised 2025-09-29.
- <sup>42</sup> Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training Large Language Models to Reason in a Continuous Latent Space. *arXiv*, December 2024. arXiv:2412.06769v2 [cs.CL], revised 2024-12-11.
- <sup>43</sup> Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach, 2 2025. arXiv preprint.
- <sup>44</sup> Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.
- <sup>45</sup> Shunyu Yao, Dian Yu, Jeffrey Zhao, Izak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822, 2023.
- <sup>46</sup> Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. Tree Search for Language Model Agents. *Transactions on Machine Learning Research*, 2025.
- <sup>47</sup> OpenAI. Introducing OpenAI o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini>, April 2025. OpenAI Blog, accessed 3 July 2025.
- <sup>48</sup> Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.
- <sup>49</sup> Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023.
- <sup>50</sup> Binfeng Xu, Zhiyuan Peng, Bowen Lei, Subhabrata Mukherjee, Yuchen Liu, and Dongkuan Xu. ReWOO: Decoupling Reasoning from Observations for Efficient Augmented Language Models. *arXiv preprint arXiv:2305.18323*, 2023.
- <sup>51</sup> Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R. Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- <sup>52</sup> Philipp Schmid. The New Skill in AI is Not Prompting, It's Context Engineering. <https://www.philschmid.de/context-engineering>, June 2025. Accessed 3 July 2025.
- <sup>53</sup> Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, Chenlin Zhou, Jiayi Mao, Tianze Xia, Jiafeng Guo, and Shenghua Liu. A Survey of Context Engineering for Large Language Models. *arXiv preprint arXiv:2507.13334*, 2025.
- <sup>54</sup> Zongqian Li, Yinhong Liu, Yixuan Su, and Nigel Collier. Prompt Compression for Large Language Models: A Survey. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7182–7195, Albuquerque, New Mexico, 2025. Association for Computational Linguistics.
- <sup>55</sup> Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curt Langlotz, Rita Lee, Joanna Melia, Joanna Nelson, Karim Sallam, Stacey Tullis, Melissa Ann Vogelsong, John Patrick Cunningham,

- and William Hiesinger. Almanac — Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI*, 1(2):A1oa2300068, 2024.
- <sup>56</sup> Qinyue Zheng, Salman Abdullah, Sam Rawal, Cyril Zakka, Sophie Ostmeier, Maximilian Purk, Eduardo Reis, Eric J. Topol, Jure Leskovec, and Michael Moor. MIRIAD: Augmenting LLMs with Millions of Medical Query-Response Pairs. *arXiv preprint arXiv:2506.06091*, 2025.
- <sup>57</sup> Soroosh Tayebi Arasteh, Mahshad Lotfinia, Keno Bressem, Robert Siepmann, Lisa Adams, Dyke Ferber, Christiane Kuhl, Jakob Nikolas Kather, Sven Nebelung, and Daniel Truhn. RadioRAG: Online Retrieval-Augmented Generation for Radiology Question Answering. *Radiology: Artificial Intelligence*, 7(4):e240476, 2025.
- <sup>58</sup> Daniel Truhn, Jorge S. Reis-Filho, and Jakob Nikolas Kather. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nature Medicine*, 29(12):2983–2984, 2023.
- <sup>59</sup> Kelly Hong, Anton Troynikov, and Jeff Huber. Context Rot: How Increasing Input Tokens Impacts LLM Performance. <https://research.trychroma.com/context-rot>, July 2025.
- <sup>60</sup> Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *arXiv preprint arXiv:2405.02957*, 2024.
- <sup>61</sup> Samuel Schmidgall, Rojin Ziae, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2025.
- <sup>62</sup> Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, Qihan Ren, Cheng Qian, Zhenghailong Wang, Minda Hu, Huazheng Wang, Qingyun Wu, Heng Ji, and Mengdi Wang. A Survey of Self-Evolving Agents: On Path to Artificial Super Intelligence. *arXiv preprint arXiv:2507.21046*, 2025.
- <sup>63</sup> Zixiang Chen, Yihe Deng, Huihuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6621–6642. PMLR, Jul 2024.
- <sup>64</sup> Yifei Zhou, Sergey Levine, Jason Weston, Xian Li, and Sainbayar Sukhbaatar. Self-Challenging Language Model Agents. *arXiv preprint arXiv:2506.01716*, jun 2025.
- <sup>65</sup> Woojin Kim. Seeing the Unseen: Advancing Generative AI Research in Radiology. *Radiology*, 311(2):e240935, May 2024.
- <sup>66</sup> Tongshuang Wu, Michael Terry, and Carrie J. Cai. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*, New York, NY, USA, 2022. Association for Computing Machinery.
- <sup>67</sup> Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. AutoGen: Enabling Next-Gen LLM applications via Multi-Agent conversations. In *Proceedings of the First Conference on Language Modeling (COLM)*, Philadelphia, PA, USA, October 2024.
- <sup>68</sup> Wentao Zhang, Liang Zeng, Yuzhen Xiao, Yongcong Li, Ce Cui, Yilei Zhao, Rui Hu, Yang Liu, Yahui Zhou, and Bo An. AgentOrchestra: Orchestrating Hierarchical Multi-Agent Intelligence with the Tool-Environment-Agent(TEA) Protocol. *arXiv preprint arXiv:2506.12508*, 2025.
- <sup>69</sup> Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje Karlsson, Jie Fu, and Yemin Shi. AutoAgents: A Framework for Automatic Agent Generation. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 22–30, 2024.
- <sup>70</sup> Cristian Jimenez-Romero, Alper Yegenoglu, and Christian Blum. Multi-agent systems powered by large language models: applications in swarm intelligence. *Frontiers in Artificial Intelligence*, 8:1593017, 2025.
- <sup>71</sup> Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Biomni: A General-Purpose Biomedical AI Agent. *bioRxiv*, 2025.
- <sup>72</sup> Fang Zeng, Zhiliang Lyu, Quanzheng Li, and Xiang Li. Enhancing LLMs for Impression Generation in Radiology Reports through a Multi-Agent System. *arXiv preprint arXiv:2412.06828*, 2024.

- <sup>73</sup> Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.
- <sup>74</sup> Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, Jeya Maria Jose Valanarasu, Mohamed Sidig Eltayeb Muneer, Eduardo Pontes Reis, Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew Abraham, Emily B. Tsai, Christopher F. Beaulieu, Jenia Jitsev, Sergios Gatidis, Jean-Benoit Delbrouck, Akshay S. Chaudhari, and Curtis P. Langlotz. A Vision-Language Foundation Model to Enhance Efficiency of Chest X-ray Interpretation. *arXiv preprint arXiv:2401.12208*, 2024.
- <sup>75</sup> Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipérez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, Thomas M. Sutter, Julia E. Vogt, Jonas Kluckert, Thomas Frauenfelder, Christian Blüthgen, Farhad Nooralahzadeh, and Michael Krauthammer. RadVLM: A Multitask Conversational Vision-Language Model for Radiology. *arXiv preprint arXiv:2502.03333*, 2025.
- <sup>76</sup> Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihane Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography. *arXiv preprint arXiv:2403.17834*, 2025.
- <sup>77</sup> Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, Christian Bluethgen, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro, Neera Ahuja, Jason Fries, Nigam H Shah, Andrew Johnston, Robert D Boutin, Andrew Wentland, Curtis P Langlotz, Jason Hom, Sergios Gatidis, and Akshay S Chaudhari. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. *Research Square*, pages rs.3.rs–4546309, 2024. Preprint.
- <sup>78</sup> Dyke Ferber, Omar SM El Nahhas, Georg Wölklein, Isabella C Wiest, Jan Clusmann, Marie-Elisabeth Leßmann, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohoei, Dirk Jäger, et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature cancer*, pages 1–13, 2025.
- <sup>79</sup> Qiaoyu Zheng, Chaoyi Wu, Pengcheng Qiu, Lisong Dai, Ya Zhang, Yanfeng Wang, and Weidi Xie. Can Modern LLMs Act as Agent Cores in Radiology Environments? *arXiv preprint arXiv:2412.09529*, 2024.
- <sup>80</sup> Wenting Chen, Yi Dong, Zhaojun Ding, Yucheng Shi, Yifan Zhou, Fang Zeng, Yijun Luo, Tianyu Lin, Yihang Su, Yichen Wu, Kai Zhang, Zhen Xiang, Tianming Liu, Ninghao Liu, Lichao Sun, Yixuan Yuan, and Xiang Li. RadFabric: Agentic AI System with Reasoning Capability for Radiology. *arXiv preprint arXiv:2506.14142*, 2025.
- <sup>81</sup> Julie Y. An, Kyle M. L. Unsdorfer, and Jeffrey C. Weinreb. BI-RADS, C-RADS, CAD-RADS, LI-RADS, Lung-RADS, NI-RADS, O-RADS, PI-RADS, TI-RADS: Reporting and Data Systems. *RadioGraphics*, 39(5):1435–1436, September 2019. Publisher: Radiological Society of North America.
- <sup>82</sup> James D. Brierley, Meredith Giuliani, Brian O’Sullivan, Brian Rous, and Elizabeth Van Eycken, editors. *TNM Classification of Malignant Tumours*. John Wiley & Sons, Oxford, 9th edition, July 2025.
- <sup>83</sup> Elizabeth A. Eisenhauer, Paul Therasse, Jan Bogaerts, et al. New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, 2009.
- <sup>84</sup> Jakob Nikolas Kather, Dyke Ferber, Isabella C Wiest, Stephen Gilbert, and Daniel Truhn. Large language models could make natural language again the universal interface of healthcare. *Nature Medicine*, 30(10):2708–2710, 2024.
- <sup>85</sup> Ross W. Filice and Charles E. Kahn. Integrating an Ontology of Radiology Differential Diagnosis with ICD-10-CM, RadLex, and SNOMED CT. *Journal of Digital Imaging*, 32(2):206–210, April 2019.
- <sup>86</sup> Leonid L. Chepelev, David Kwan, Charles E. Kahn, Ross W. Filice, and Kenneth C. Wang. Ontologies in the New Computational Age of Radiology: RadLex for Semantics and Interoperability in Imaging Workflows. *RadioGraphics*, 43(3):e220098, March 2023.
- <sup>87</sup> Daniel J Vreeman, Swapna Abhyankar, Kenneth C Wang, Christopher Carr, Beverly Collins, Daniel L Rubin, and Curtis P Langlotz. The LOINC RSNA radiology playbook - a unified terminology for radiology procedures. *Journal of the American Medical Informatics Association*, 25(7):885–893, July 2018.

- <sup>88</sup> Joseph J Budovec, Cesar A Lam, and Charles E Kahn Jr. Informatics in radiology: radiology gamuts ontology: differential diagnosis for the semantic web. *Radiographics*, 34(1):254–264, 2014.
- <sup>89</sup> Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12902–12915, 2024.
- <sup>90</sup> Eunsuk Chang and Sumi Sung. Use of SNOMED CT in Large Language Models: Scoping Review. *JMIR Medical Informatics*, 12(1):e62924, 2024.
- <sup>91</sup> Philipp Arnold, Daniel Pinto Dos Santos, Fabian Bamberg, and Elmar Kotter. FHIR—Overdue Standard for Radiology Data Warehouses. In *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Georg Thieme Verlag KG, 2024.
- <sup>92</sup> Christian Hinge. dicom-mcp: Model context protocol for interacting with dicom servers. <https://github.com/ChristianHinge/dicom-mcp>, 2025. Version v0.1.2 (released 2025-04-28); accessed 2025-08-01.
- <sup>93</sup> Bernardo Gabriele Collaco, Syed Ali Haider, Srinivasagam Prabha, Cesar Abraham Gomez-Cabello, Ariana Genovese, Nadia G. Wood, Sanjay P. Bagaria, Narayanan Gopala, Cui Tao, and Antonio Jorge Forte. The Role of Agentic Artificial Intelligence in Healthcare: A Systematic Review. *Research Square*, August 2025. Preprint, Version 1.
- <sup>94</sup> Thomas E Strayer, Lucy B Spalluto, Abby Burns, Christopher J Lindsell, Claudia I Henschke, David F Yankelevitz, Drew Moghanaki, Robert S Dittus, Timothy J Vogus, Carolyn Audet, et al. Using the framework for reporting adaptations and modifications-expanded (frame) to study adaptations in lung cancer screening delivery in the veterans health administration: a cohort study. *Implementation science communications*, 4(1):5, 2023.
- <sup>95</sup> Xuan Loc Pham, Gwendolyn Vuurberg, Marjan Doppen, Joey Roosen, Tip Stille, Thi Quynh Ha, Thuy Duong Quach, Quoc Vu Dang, Manh Ha Luu, Ewoud J. Smit, Hong Son Mai, Mattias Heinrich, Bram van Ginneken, Mathias Prokop, and Alessa Hering. TotalRegistrar: Towards a Lightweight Foundation Model for CT Image Registration. *arXiv preprint arXiv:2508.04450*, 2025.
- <sup>96</sup> Christian Bluethgen, Pierre Chambon, Jean-Benoit Delbrouck, Rogier Van Der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay S. Chaudhari. A vision–language foundation model for the generation of realistic chest X-ray images. *Nature Biomedical Engineering*, August 2024.
- <sup>97</sup> Manju Dabass, Mohammed M. Jabeer, Anuj Chandalia, and Dwarikanath Mahapatra. Streamlined Speech Recognition Model for Automated Radiology Reporting Employing Combined Automatic Speech Recognition Model, Large Language Model, and Prompt Engineering. In Gaurav Raj, Bhuvan Unhelker, and Ankur Choudhary, editors, *Advances in Artificial Intelligence and Machine Learning*, pages 345–356, Singapore, 2025. Springer Nature Singapore.
- <sup>98</sup> Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. *arXiv preprint arXiv:2501.09136*, 2025.
- <sup>99</sup> Siwei Wu, Yizhi Li, Xingwei Qu, Rishi Ravikumar, Yucheng Li, Tyler Loakman, Shanghaoran Quan, Xiaoyong Wei, Riza Batista-Navarro, and Chenghua Lin. LongEval: A Comprehensive Analysis of Long-Text Generation Through a Plan-based Paradigm. *arXiv preprint arXiv:2502.19103*, 2025.
- <sup>100</sup> Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. *Advances in neural information processing systems*, 37:49881–49913, 2024.
- <sup>101</sup> Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature methods*, 21(2):195–212, 2024.
- <sup>102</sup> Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39, 2022.
- <sup>103</sup> Chen Xiaolan, Xiang Jiayang, Lu Shanfu, Liu Yexin, He Mingguang, and Shi Danli. Evaluating large language models and agents in healthcare: key challenges in clinical applications. *Intelligent Medicine*, 2025.
- <sup>104</sup> Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan.  $\tau$ -bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.
- <sup>105</sup> Philipp Schmid. Pass@k vs pass $\hat{k}$ : Understanding agent reliability. <https://www.philschmid.de/agents-pass-at-k-pass-power-k>, 2025. Accessed: 2025-10-09.

- <sup>106</sup> Eun Kyoung Hong, Byungseok Roh, Beomhee Park, Jae-Bock Jo, Woong Bae, Jai Soung Park, Dong-Wook Sung, et al. Value of Using a Generative AI Model in Chest Radiography Reporting: A Reader Study. *Radiology*, 314(3):e241646, 2025.
- <sup>107</sup> Jonathan Huang, Matthew T. Wittbrodt, Caitlin N. Teague, Eric Karl, et al. Efficiency and Quality of Generative AI-Assisted Radiograph Reporting. *JAMA Network Open*, 8(6):e2513921, 2025.
- <sup>108</sup> Florence X Doo, Jan Vosshenrich, Tessa S Cook, Linda Moy, Eduardo PRP Almeida, Sean A Woolen, Judy Wawira Gichoya, Tobias Heye, and Kate Hanneman. Environmental sustainability and AI in radiology: a double-edged sword. *Radiology*, 310(2):e232030, 2024.
- <sup>109</sup> Inioluwa Deborah Raji, Roxana Daneshjou, and Emily Alsentzer. It's time to bench the medical exam benchmark. *NEJM AI*, 2(2):A1e2401235, 2025.
- <sup>110</sup> Yu Gu, Jingjing Fu, Xiaodong Liu, Jeya Maria Jose Valanarasu, Noel Codella, Reuben Tan, Qianchu Liu, Ying Jin, Sheng Zhang, Jinyu Wang, et al. The illusion of readiness: Stress testing large frontier models on multimodal medical benchmarks. *arXiv preprint arXiv:2509.18234*, 2025.
- <sup>111</sup> Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, et al. Sequential Diagnosis with Language Models. *arXiv preprint arXiv:2506.22405*, 2025.
- <sup>112</sup> Ali S. Tejani, Michail E. Klontzas, Anthony A. Gatti, John T. Mongan, Linda Moy, Seong Ho Park, Charles E. Kahn, for the CLAIM 2024 Update Panel, Sunhy Abbara, Saif Afat, Uduanna C. Anazodo, Anna Andreychenko, Folkert W. Asselbergs, Aldo Badano, Bettina Baessler, Bayarbaatar Bold, Sotirios Bisdas, Torkel B. Brismar, Giovanni E. Cacciamani, John A. Carrino, Julius Chapiro, Michael F. Chiang, Tessa S. Cook, Renato Cuocolo, John Damilakis, Roxana Daneshjou, Carlo N. De Cecco, Hesham Elhalawani, Guillermo Elizondo-Riojas, Andrey Fedorov, Benjamin Fine, Adam E. Flanders, Judy Wawira Gichoya, Maryellen L. Giger, Safwan S. Halabi, Sven Haller, William Hsu, Krishna Juluru, Jayashree Kalpathy-Cramer, Apostolos H. Karantanas, Felipe C. Kitamura, Burak Kocak, Dow-Mu Koh, Elmar Kotter, Elizabeth A. Krupinski, Curtis P. Langlotz, Cecilia S. Lee, Mario Maas, Anant Madabhushi, Lena Maier-Hein, Kostas Marias, Luis Martí-Bonmatí, Jaishree Naidoo, Emanuele Neri, Robert Ochs, Nikolaos Papanikolaou, Thomas Papathomas, Katja Pinker-Domenig, Daniel Pinto Dos Santos, Fred Prior, Alexandros Protonotarios, Mauricio Reyes, Veronica Rotemberg, Jeffrey D. Rudie, Emmanuel Salinas-Miranda, Francesco Sardanelli, Mark E. Schweitzer, Luca Maria Sconfienza, Ronnie Sebro, Prateek Sharma, An Tang, Antonios Tzortzakakis, Jeroen Van Der Laak, Peter M. A. Van Ooijen, Vasantha K. Venugopal, Jacob J. Visser, Bradford J. Wood, Carol C. Wu, Greg Zaharchuk, and Marc Zins. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiology: Artificial Intelligence*, 6(4):e240300, 2024.
- <sup>113</sup> Satvik Tripathi, Dana Alkhulaifat, Florence X Doo, Pranav Rajpurkar, Rafe McBeth, Dania Daye, and Tessa S Cook. Development, Evaluation, and Assessment of Large Language Models (DEAL) Checklist: A Technical Report. *NEJM AI*, 2(6):A1p2401106, 2025.
- <sup>114</sup> Jack Gallifant, Majid Afshar, Saleem Ameen, Yindalon Aphinyanaphongs, Shan Chen, Giovanni Cacciamani, Dina Demner-Fushman, Dmitry Dligach, Roxana Daneshjou, Chrystinne Fernandes, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nature medicine*, 31(1):60–69, 2025.
- <sup>115</sup> Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A Clifton, Gary S Collins, Spiros Denaxas, Alastair K Denniston, Livia Faes, Bart Geerts, Mudathir Ibrahim, Xiaoxuan Liu, Basil A Mateen, Piyush Mathur, Michael D McCradden, Lauren Morgan, Jonathan Ordish, Charlotte Rogers, Suchi Saria, Daniel S W Ting, Peter Watkinson, Wolf Weber, Paul Wheatstone, Peter McCulloch, and DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, 28(5):924–933, 2022.
- <sup>116</sup> Nikita Mehandru, Brenda Y Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J Butte, and Ahmed Alaa. Evaluating large language models as agents in the clinic. *NPJ digital medicine*, 7(1):84, 2024.
- <sup>117</sup> Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. CRAFT-MD: A conversational evaluation framework for comprehensive assessment of clinical LLMs. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.
- <sup>118</sup> David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Matthias Keicher, and Nassir Navab. Language Agents for Hypothesis-driven Clinical Decision Making with Reinforcement Learning. *arXiv preprint arXiv:2506.13474*, 2025.

- <sup>119</sup> Jie Liu, Wenzuan Wang, Zizhan Ma, Guolin Huang, Yihang Su, Kao-Jung Chang, Wenting Chen, Haoliang Li, Linlin Shen, and Michael Lyu. MedChain: Bridging the Gap Between LLM Agents and Clinical Practice through Interactive Sequential Benchmarking. *arXiv preprint arXiv:2412.01605*, 2024.
- <sup>120</sup> Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents. *NEJM AI*, 2(9):AIdbp2500144, 2025.
- <sup>121</sup> Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. MedAgentBoard: Benchmarking Multi-Agent Collaboration with Conventional Methods for Diverse Medical Tasks. *arXiv preprint arXiv:2505.12371*, 2025.
- <sup>122</sup> Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, Arman Cohan, and Mark Gerstein. MedAgentsBench: Benchmarking Thinking Models and Agent Frameworks for Complex Medical Reasoning. *arXiv preprint arXiv:2503.07459*, 2025.
- <sup>123</sup> Tugba Akinci D'Antonoli, Arnaldo Stanzione, Christian Bluethgen, Federica Vernuccio, Lorenzo Ugga, Michail E. Klontzas, Renato Cuocolo, Roberto Cannella, and Burak Koçak. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagnostic and Interventional Radiology (Ankara, Turkey)*, October 2023.
- <sup>124</sup> Suhana Bedi, Iddah Mlauzi, Daniel Shin, Sanmi Koyejo, and Nigam H. Shah. The Optimization Paradox in Clinical AI Multi-Agent Systems. *arXiv preprint arXiv:2506.06574*, 2025.
- <sup>125</sup> Tugba Akinci D'Antonoli, Ali S Tejani, Bardia Khosravi, Christian Bluethgen, Felix Busch, Keno K Bressem, Lisa C Adams, Mana Moassefi, Shahriar Faghani, and Judy Wawira Gichoya. Cybersecurity Threats and Mitigation Strategies for Large Language Models in Health Care. *Radiology: Artificial Intelligence*, 7(4):e240739, 2025.
- <sup>126</sup> Max Ostermann, Rebecca Mathias, Fatemeh Jahed, Mitchell B Parker, Florence D Hudson, William C Harding, Stephen Gilbert, and Oscar Freyer. Cybersecurity Requirements for Medical Devices in the EU and US-A Comparison and Gap Analysis of the MDCG 2019-16 and FDA premarket cybersecurity guidance. *Computational and Structural Biotechnology Journal*, 2025.
- <sup>127</sup> Oscar Freyer, Sandhya Jayabalan, Jakob N Kather, and Stephen Gilbert. Overcoming regulatory barriers to the implementation of ai agents in healthcare. *Nature Medicine*, pages 1–5, 2025.
- <sup>128</sup> Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):A1oa2300138, 2024.
- <sup>129</sup> Krzysztof Budzyń, Marcin Romańczyk, Diana Kitala, Paweł Kołodziej, Marek Bugajski, Hans O Adami, Johannes Blom, Marek Buszkiewicz, Natalie Halvorsen, Cesare Hassan, Tomasz Romańczyk, Øyvind Holme, Krzysztof Jarus, Shona Fielding, Melina Kunar, Maria Pellise, Nastazja Pilonis, Michał Filip Kamiński, Mette Kalager, Michael Bretthauer, and Yuichi Mori. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 2025.
- <sup>130</sup> Pranav Rajpurkar and Eric J Topol. Beyond Assistance: The Case for Role Separation in AI-Human Radiology Workflows. *Radiology*, 316(1):e250477, 2025.
- <sup>131</sup> Burak Koçak, Andrea Ponsiglione, Arnaldo Stanzione, Christian Bluethgen, João Santinha, Lorenzo Ugga, Merel Huisman, Michail E. Klontzas, Roberto Cannella, and Renato Cuocolo. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and Interventional Radiology (Ankara, Turkey)*, July 2024.