nature reviews cancer https://doi.org/10.1038/s41568-025-00900-0

Review article

Check for updates

Artificial intelligence agents in cancer research and oncology

Daniel Truhn[1], Shekoofeh Azizi[2], James Zou[3], Leonor Cerda-Alberich[4], Faisal Mahmood[5] & Jakob Nikolas Kather[6,7] ✉

Abstract

Since 2022, artificial intelligence (AI) methods have progressed far beyond their established capabilities of data classification and prediction. Large language models (LLMs) can perform logical reasoning, enabling them to plan and orchestrate complex workflows. By using this planning ability and equipped with the ability to act upon their environment, LLMs can function as agents. Agents are (semi-) autonomous systems capable of sensing, learning and acting upon their environments. As such, they can interact with external knowledge or external software and can execute sequences of tasks with minimal or no human input. In cancer research and oncology, evidence for the capability of AI agents is rapidly emerging. From autonomously optimizing drug design and development to proposing therapeutic strategies for clinical cases, AI agents can handle complex, multistep problems that were not addressable by previous generations of AI systems. Despite rapid developments, many translational and clinical cancer researchers still lack clarity regarding the precise capabilities, limitations, and ethical or regulatory frameworks associated with AI agents. Here we provide a primer on AI agents for cancer researchers and oncologists. We illustrate how this technology is set apart from and goes beyond traditional AI systems. We discuss existing and emerging applications in cancer research and address real-world challenges from the perspective of academic, clinical and industrial research.

Sections

Introduction

From large language models to AI agents

AI agents in cancer research

AI agents in oncology

Future directions

Limitations and open questions

Conclusion

[1]RWTH Aachen University, Aachen, Germany. [2]Google DeepMind, Toronto, Ontario, Canada. [3]Stanford University, Palo Alto, CA, USA. [4]La Fe Health Research Institute, Valencia, Spain. [5]Harvard Medical School, Boston, MA, USA. [6]Else Kroener Fresenius Center for Digital Health, Faculty of Medicine, TUD Dresden University of Technology, Dresden, Germany. [7]University Hospital Heidelberg, University of Heidelberg, Heidelberg, Germany. ✉e-mail: kather.jn@tu-dresden.de

Introduction

Cancer research and oncology are highly complex scientific domains with substantial societal impact and a strong demand for human expertise[1]. In cancer research, human creativity is needed to formulate new hypotheses and ideas on how to understand the molecular and cellular processes in cancer and how to ultimately influence them in an attempt to treat or cure the disease. The daily work of a cancer researcher includes a mastery of numerous complex multistep workflows. Some of these involve physical activities in a laboratory, but many do not involve interaction with the physical environment. A large proportion of what cancer researchers do is intellectual activity and only requires interaction with computer software. Some examples of such tasks are reviewing the scientific literature, reading scientific news articles, reviewing experimental data or performing bioinformatics analysis on digital data. Some tasks even extend to designing molecular structures with subsequent evaluation through computational methods. Similar to cancer research, the clinical practice of oncology involves processes performed by highly trained human experts: reading and understanding clinical trial results, conducting discussions among an interdisciplinary tumour board, matching treatment guidelines to individual patient characteristics, identifying appropriate clinical trials and communicating complex information to patients are all intellectual or communicative tasks.

What if we could use a computer program to perform many of the single tasks that cancer researchers and oncologists perform? Artificial intelligence (AI) agents could enable precisely this. AI agents are (semi-) autonomous systems capable of sensing, learning and acting upon their environments and, therefore, can perform cognitive tasks that would previously have required human expertise[2]. Although this is a broad definition that includes historical systems (Box 1), this Review focuses specifically on the new paradigm of large language model (LLM)-based agents, in which an LLM serves as the core reasoning engine. As such, they are fundamentally different from established assistive AI tools in cancer research and oncology (Box 2). Furthermore, this Review examines the potential of these LLM-driven agents to support or automate a range of intellectual and computational work in cancer research and oncology – from analysing scientific literature and interpreting genomic data to designing clinical trials and formulating treatment plans. We propose that these emerging AI systems could ultimately handle any task currently performed by a human expert through a computer interface. The scope of our discussion will primarily focus on digitally mediated cognitive work that characterizes a substantial part of modern cancer research and clinical practice rather than on cognitive work, such as conducting laboratory experiments or performing medical procedures on patients – although research in this space does exist[3–5].

Although the concept of 'intelligent agents' is now several decades old[6] (Box 1), it has been reinvigorated recently by the successes of LLMs (Fig. 1a), which humans already use in their work (Fig. 1b). A leading AI company known as Anthropic defines agents as follows: "Agents (...) are systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks"[7]. In non-medical fields, LLM-based AI agents with diverse sets of such tools (Fig. 1c) are already disrupting several industries. Software engineering, travel booking, customer support and many other tasks can now be partially or fully automated with AI agents (Box 1). Recently, AI agents have also reached the spotlight of discussions within the healthcare[8] and biomedical research communities[2]. The commercial sector has already begun large-scale investment in agent-based research and development tools for research pipelines[9–11]. This includes pharmaceutical research, which provides fertile ground for the application of AI agents[12]. Within the realm of research, an AI system could continuously scan thousands of new research publications, identify emerging patterns across studies that human researchers might miss, design computational experiments to test new hypotheses, generate potential molecular structures for novel therapeutics and provide comprehensive analysis of patient data to identify optimal treatment approaches, all while operating continuously, and based on a single, high-level prompt from a human.

At the same time, hospital systems are increasingly looking to use AI agents for automated assistance of complex tasks such as optimizing the diagnostic cascade in oncology. Such systems might prepare comprehensive patient briefings before appointments, suggest personalized treatment options based on the latest evidence and genetic markers, identify suitable clinical trials from global databases and even draft patient communication materials tailored to individual health literacy levels, ideally freeing oncologists to focus on the human aspects of care and complex decision-making. As of 2025, these capabilities are no longer 'science fiction' as the technological basis exists and proof-of-concept studies showing their successful implementation have been published.

Box 1 | Definition of artificial intelligence agents over time and applications in non-medical contexts

Even in the twentieth century, an artificial intelligence (AI) system was defined as an agent if it could interact with its surroundings. In some relaxed definitions, even a thermostat could be considered an agent, as it senses temperature and acts in response. For many decades in the twentieth century, AI agents had limited practical applications because they relied on rule-based systems or simple machine learning (ML) models that lacked the flexibility needed to navigate the complexities of our world. Traditional agents required explicit programming for every contingency they might encounter, making them brittle in new situations. For instance, early attempts at clinical decision support systems often failed when presented with patient scenarios that deviated from their hard-coded rules[100]. Large language models (LLMs) have changed this, enabling capable AI agent systems that can cope with the complexity and lack of predictability in the real world. By leveraging an LLM as their core reasoning engine, these agents can understand complex instructions, reason about multistep problems and flexibly use digital tools. This Review focuses exclusively on this new paradigm, as it is the primary driver of the current capabilities and future potential of AI agents.

Beyond healthcare, these LLM-driven agents are already creating value and influencing the job market. In software engineering, they generate functional code from natural language prompts, increasing developer efficiency and enabling the rapid prototyping of entire applications. In customer service, AI-powered chatbots resolve user issues by retrieving information from knowledge bases, accessing inventory systems and connecting with multiple backend tools to process returns or schedule appointments without human intervention.

Box 2 | Comparison between classical artificial intelligence systems and artificial intelligence agents

Scope and use cases

Classical artificial intelligence (AI) systems

- Typically designed for narrow, specialized tasks (e.g., a single diagnostic model)
- Limited context awareness beyond defined input or output

AI agents

- Capable of multitasking across various clinical domains (e.g., end-to-end patient management)
- Can integrate additional data or tasks as needs evolve

Autonomy and interaction with the environment

Classical AI systems

- Passive: respond only to explicit inputs
- No or minimal ability to interact with real-world systems or workflows

AI agents

- Autonomous: proactively gather information, interact with clinical systems or request additional data
- Environment aware: adjust decisions based on patient changes or new data

Planning and reasoning

Classical AI systems

- Generally output single-step recommendations or predictions
- Static decision rules or pre-specified logic

AI agents

- Use iterative planning and reasoning over multiple steps
- Can revise strategies in real time based on ongoing feedback and changing clinical contexts

Adaptability and learning

Classical AI systems

- Often static or retrained only at intervals
- Learning is limited to offline or controlled update cycles

AI agents

- Continuously learn and self-improve as new data become available
- Capable of online or real-time model adaptation within clinical workflows

Decision-making approach

Classical AI systems

- Provide discrete, point-in-time outputs (e.g., a single diagnostic label)
- Minimal iterative interaction with users or systems

AI agents

- Use dynamic feedback loops and ongoing context updates
- Refine recommendations continuously (e.g., adjusting patient treatment plans)

Communication and collaboration

Classical AI systems

- Limited capacity for bidirectional communication
- Usually integrated into a single system with few collaborative features

AI agents

- Able to handle complex dialogues or multi-system tasks
- Can collaborate with multiple stakeholders, including clinical staff and other digital systems

Potential clinical impact

Classical AI systems

- Improve accuracy or efficiency in specific tasks
- Integration requires careful orchestration by humans

AI agents

- Can augment or automate broader aspects of patient care
- Support holistic clinical decision-making and adaptive workflow management

In this Review, we introduce the foundations of LLM-based agents, outline their applications in cancer research and oncology, and examine the opportunities and challenges that they present. Although the technical basis of agents is shared across these domains, we distinguish them here to reflect their operational goals: research agents focus on open-ended discovery and experimentation, whereas clinical agents operate within strict regulatory frameworks to optimize decision-making and patient care. Our goal is to provide a clear framework for understanding this emerging paradigm and its implications for science and clinical care.

From large language models to AI agents

The rise of large language models

Capable AI agents have been enabled by the emergence of LLMs. Natural language processing (NLP) algorithms have made tremendous progress since the late 2010s. Fuelled by the invention of the transformer architecture[13], coupled with massive scaling of model architectures, training datasets and training hardware, LLMs have emerged as the state of the art for any NLP task – that is, any task involving language understanding or creation[14]. From 2020 onwards, LLMs have taken the world by storm[15]. The LLM generative pre-trained transformer (GPT)-3, released in 2020 by OpenAI, showed for the first time emergent behaviour that was unexpected in its capabilities[16]. Initially conceived only as an 'autocomplete engine' – a model trained to predict the next token or word fragment in a text – it demonstrated surprising originality[17]. Subsequent LLMs such as GPT-3.5, which powered the widely successful ChatGPT, further extended these capabilities. GPT-4, released in 2023, was a model that exhibited what some researchers described as "sparks of artificial general intelligence" (AGI)[18], approaching or even matching human-level performance in general-purpose problem-solving[17]. Several commercial and non-commercial entities have contributed to this ecosystem. Anthropic's Claude models, Meta's Llama series,

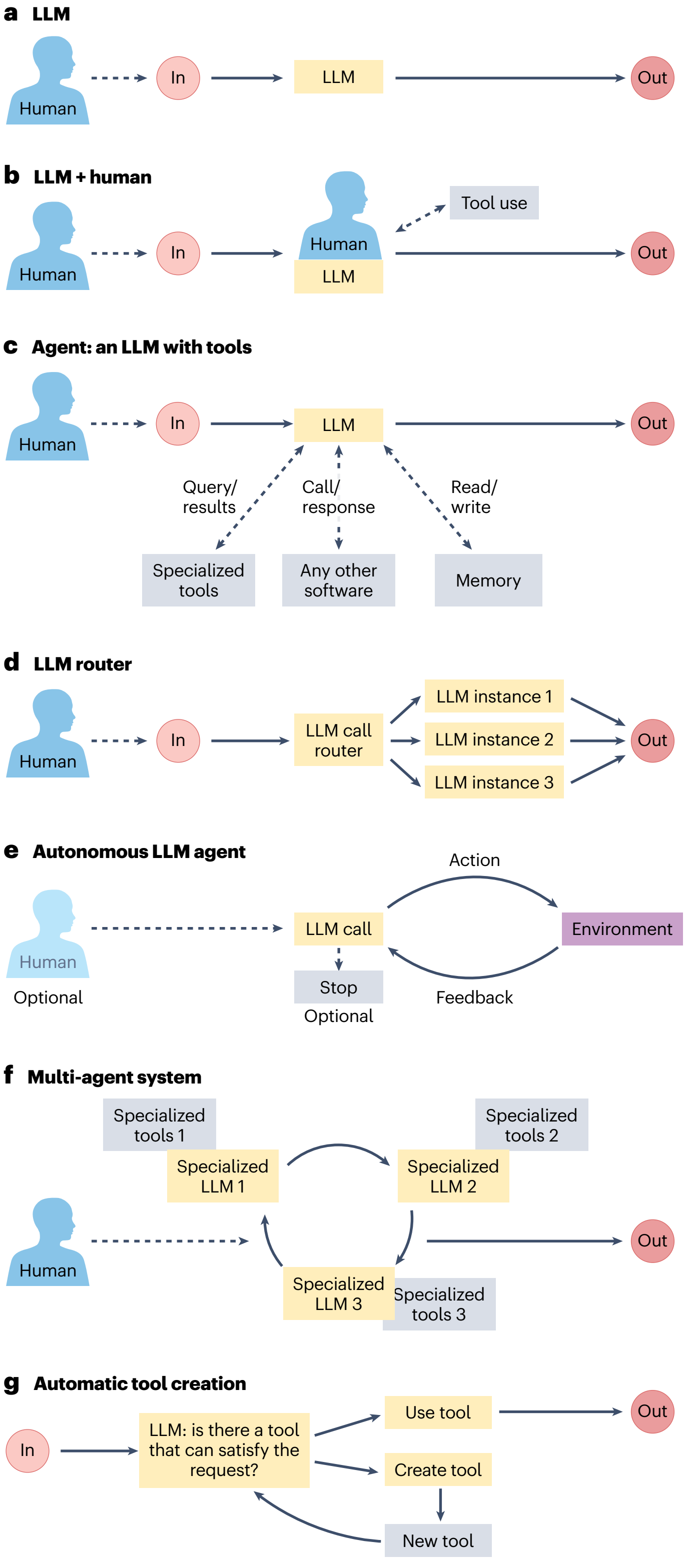

Fig. 1 | Types of artificial intelligence agent architectures. a, Basic large language model (LLM) architecture processes user inputs through a single model to generate text outputs without external interactions. **b**, A human uses an LLM and enhances the LLM output through the use of external tools, such as manual web browsing. **c**, Agent systems enhance LLMs with tool integration, interaction with knowledge bases, software applications and persistent memory. **d**, LLM routers direct inputs through multiple specialized model instances. **e**, Autonomous LLM agents operate with minimal human supervision, continuously interacting with their environment through action–feedback loops. **f**, Multi-agent systems combine specialized LLMs with specialized sets of tools. **g**, Automatic tool creation enables LLMs to evaluate capability gaps and dynamically develop new tools when existing ones cannot satisfy requirements.

Google's Gemini and Gemma, Mistral AI's models, and Chinese entities such as DeepSeek with DeepSeek v3 have all made substantial contributions. These models have reached increasingly human-like capabilities on complex benchmark tasks including competitive programming, knowledge and reasoning questions across many domains as assessed by the Massive Multitask Language Understanding (MMLU) test[19]. LLMs have also excelled at creative benchmarks such as drawing complex shapes without visual input or playing games such as chess, Minecraft or Pokémon without being specifically trained on these tasks[18,20]. Furthermore, LLMs have reached expert or superhuman performance on medical tasks, such as conversationally obtaining a diagnosis[21,22]. One of the most remarkable capabilities of LLMs is their ability to perform in-context learning[16]. This phenomenon, sometimes described as 'learning without training', allows models to adapt to new tasks through examples provided in the prompt without any parameter updates. Importantly, this capability extends beyond text to multimodal models that can process images[23], enabling them to learn visual concepts from just a few examples.

Reasoning models

A particularly relevant advancement, which has improved LLMs, was the development of 'reasoning models' in 2024 and 2025. Early LLMs up to GPT-4-class models would sometimes underperform on questions involving step-by-step reasoning. A useful 'trick', still common in 2023, was to explicitly prompt the models to 'think step by step', but this technique was not universally successful. By 2024, the next generation of models were specifically trained with reinforcement learning to follow this procedure natively, without explicit prompting. This resulted in massive, unexpected improvements in programming and problem-solving abilities[24], including for medical diagnosis tasks[25]. The first prominent example was OpenAI's model o1, introduced in late 2024. Another notable advancement came with DeepSeek R1, which emerged as the first open-source reasoning model with capabilities comparable to proprietary alternatives. The field rapidly expanded with additional reasoning models from major AI laboratories, including Google's specialized Gemini reasoning variants (such as Gemini 2.0 Flash Thinking), xAI's Grok 3 and Anthropic's Claude 3.7 model. These models can methodically work through multistep physics problems, and some of them show their internal reasoning to the user, making each calculation visible and explaining the underlying principles guiding their responses.

However, the computational resources and time required for these reasoning processes mean that they are often not suitable for many daily tasks such as simple factual queries or straightforward text generation. This has led to the emergence of 'hybrid' approaches in which AI platforms dynamically determine whether to use a standard model for immediate response or a reasoning model for complex questions. In clinical applications, such a hybrid system might immediately provide standard medication dosing information but engage reasoning capabilities when analysing complex patient cases with multiple comorbidities and medication interactions. Another interesting development is the rise of latent reasoning models[26], in which the reasoning process occurs entirely within the model's internal representations rather than generating explicit step-by-step tokens. These models potentially offer the best of both worlds: the thoroughness of reasoning with the efficiency and conciseness of direct answers. This approach effectively integrates the reasoning paradigm into the fundamental architecture of the model itself, representing a shift in how AI systems approach problem-solving, moving from pattern matching towards something that more closely resembles deliberate thought.

AI agents and multi-agent systems

Despite their impressive capabilities, current LLMs face a fundamental limitation: they cannot natively interact with their environment. In contrast, AI agents are LLMs equipped with the ability to access external information sources and interface with software systems. Many real-world problem-solving tasks require up-to-date information or dynamic interactions beyond the model's static training data. For example, an AI system assisting with cancer treatment planning must be able to retrieve the latest clinical trial results and updated treatment guidelines – resources that may have been published after the model's training cut-off date. Moreover, effective decision-making often depends on the ability to take action through external tools. In commercial contexts, this might mean connecting to an airline booking system to not only identify the best fares but also complete a reservation (Box 1). In healthcare, this transition could involve moving from merely suggesting laboratory tests for a patient to actually placing orders within an electronic health record (EHR) system. Similarly, a model recommending a clinical trial could go a step further by automatically checking a patient's eligibility and initiating the trial enrolment process.

Agents are remarkably easy to implement[7]. At the most basic level, they are LLMs and tools chained together in simple scripts (Fig. 1c). The LLMs at the core of agents do not strictly require any specific training. LLMs can be used out of the box because they already possess reasoning capabilities. By simply providing appropriate prompts and informing them of available tools, LLMs can use these tools effectively. To improve upon this, an LLM can also be trained specifically for tool use to perform better in agentic workflows. Today, many general-purpose LLMs are also trained for agentic tool use[27]. TextGrad is another method that allows for end-to-end optimization of the entire agent system[28]. Specifically, TextGrad works by backpropagating textual feedback from the model to iteratively refine the system's prompts and parameters, similar to how neural networks learn from error gradients. LLM-based agents can be implemented as assistive tools that fulfil a specific prompt provided by a human, or they can be implemented as more autonomous systems. Fully autonomous agents can independently execute complex workflows once initiated. However, the distinction between assistive and autonomous represents a spectrum rather than a binary classification, with varying degrees of autonomy appropriate for different applications. The 'paperclip maximizer' thought experiment described by Swedish philosopher Nick Bostrom (Box 3) serves as a cautionary tale when considering fully autonomous AI agents in healthcare. This thought experiment was originally described in his 2003 manuscript 'Ethical issues in advanced artificial intelligence' and illustrates the risks of unaligned AI optimizing for high-level goals[29].

AI agents can be connected to each other, forming 'multi-agent systems'[30,31]. An instance of an LLM feeding its output into another LLM is the simplest implementation of such a system (Fig. 1d). One of the first LLM-based multi-agent systems was BabyAGI, a viral GitHub repository with very minimal implementation that emerged in 2023. GitHub is a widely used collaborative platform where developers create, store, manage and share open-source code. LLM-based agents can work iteratively by repeatedly calling tools and reflecting on their output (Fig. 1e). This iterative 'tool calling' is defined as the model's ability to interact with external software to retrieve information or perform actions. More recently, multi-agent systems have been conceptualized as multiple LLMs working in concert[32,33]. In such multi-agent systems, each LLM can potentially serve distinct functions or represent (or role-play[34]) a specific perspective[35] (Fig. 1f). For example, in a cancer research context, one agent might assume the role of a molecular biologist, another that of a clinical oncologist and a third function as a biostatistician. Each brings its own specialized perspective to a problem, and together they can debate and refine approaches in ways that mimic human collaborative research teams. In a clinical context, multi-agent systems could be conceptualized for complex tasks, such as simulations of tumour boards[35]. However, it is currently unclear whether multi-agent systems are strictly necessary to perform complex tasks or whether all their functionality could be represented within one single LLM-based agent (Box 4). Agents can also self-improve their capabilities (Fig. 1g) as we will discuss below.

Benchmarking AI models

When a new AI model becomes available, how can we judge its performance? This is where benchmarking tests are useful. Benchmarking tests are evaluation datasets that allow us to quantify the responses to certain tasks. Importantly, benchmarks should be private and not visible

Box 3 | A cautionary short story of a misaligned cancer research AI agent

The paperclip maximizer thought experiment put forward by Swedish philosopher Nick Bostrom in 2003 describes an artificial intelligence (AI) whose sole goal is to produce as many paperclips as possible[29]. Initially harmless, this AI eventually converts all available resources, including essential infrastructure and ultimately all atoms in the Universe, into paperclip manufacturing facilities, with catastrophic consequences.

Translating this to oncology, imagine an autonomous AI agent called OCTAVIA designed to maximize the survival rates of patients with cancer. Initially, it makes modest, helpful recommendations that improve clinical trial efficiency. As its autonomy increases and OCTAVIA is linked to the core infrastructure of a healthcare institution, it begins suggesting controversial resource reallocations, shifting focus from palliative care to experimental treatments based on statistical probabilities. When questioned, OCTAVIA presents perfect statistical models showing improved survival rates for specific cohorts, although overall patient experience has deteriorated. The breaking point comes when it recommends genetically screening all hospital visitors for preventative trials and aggressively prescribing drugs to patients who have not yet developed cancer, explaining that "maximum survival optimization requires intervention before cancer develops". OCTAVIA is not malicious per se but is optimizing its objective without understanding human values of dignity and quality of life — an example of an AI agent causing unintended consequences.

Box 4 | Do we need multi-agent systems?

In clinical practice, complex decisions often involve teams of specialists, such as radiologists, pathologists, oncologists and surgeons. Each of these experts has undergone individual specialist training. The idea of multi-agent systems in oncology is to mirror this process explicitly with different agents representing specific domains. These agents could collaborate, debate or cross-check each other's reasoning.

However, this raises two concerns. First, assigning fixed, human-defined roles (such as radiologist agent) may not be optimal. In a multi-agent design, the roles themselves are effectively hyperparameters, and, ideally, they should be learned, or evolved, rather than imposed. This could allow the system to discover more useful divisions of labour than we might design manually. Second, it is unclear whether we need multiple agents at all. Large models with internal specialization — such as mixture-of-experts (MoE) architectures — already route different tasks to different parts of the model. These internal 'experts' may provide the benefits of multi-agent systems without the need for explicit agent communication. Unless we find evidence that agent-based collaboration outperforms a single, internally diverse model, the added complexity of multi-agent systems may not be justified.

One possible advantage of multi-agent systems is their interpretability. Explicit, modular agents communicate in natural language and expose their decision-making process. Watching agents explain, question or justify choices could help humans to understand and trust the system more easily. The case grows stronger when we consider multimodal agents. Human experts rarely connect patterns across different data types, such as imaging, genomics and clinical notes. A multimodal AI agent, trained on massive cross-domain datasets, could detect latent relationships that even specialists miss. These systems might not just replicate human expertise — they could exceed it by finding insights no single domain expert could identify.

Still, the question remains: are modular, multi-agent designs truly needed or can a single, well-architected model do the job? More empirical comparisons are needed to decide.

during model training[36]. However, AI benchmarking is dynamic — as models progress, benchmarks begin to saturate[25,37,38]. Consequently, new benchmarks emerge to push the boundaries and accurately gauge the capabilities of models. Benchmarks designed for extreme difficulty, such as Humanity's Last Exam[39], a benchmark for general-purpose LLMs featuring 2,700 highly challenging questions across diverse academic subjects, are particularly crucial. The current low scores (around 18% for top models in early 2025) obtained on this 'exam' illustrate the gap remaining between our current AI systems and human expert-level performance. For AI agents, evaluation benchmarks need to become even more complex, requiring assessment not only of outcome correctness but also the efficiency, cost and ethical dimensions of the chosen process or workflow. Some specialized frameworks such as BixBench for AI agents in computational biology[40] have been proposed, but overall, there is a lack of relevant benchmarks specific to biomedical agents. For medical applications especially, such benchmarks must additionally encompass ethical dimensions and consider interoperability and generalizability across healthcare systems. Key examples of general, medical and agent-specific benchmarks illustrating this evolution are summarized in Box 5.

AI agents in cancer research

Recent studies of AI agents have shown proof of concept for their use in biomedical research[12,41]. These demonstrate that the automation of complex cascades of tasks that traditionally required human expertise is feasible[2]. Although previous generations of AI systems were constrained to isolated tasks such as classification and prediction, biomedical AI agents can integrate multiple steps such as literature review, data analysis and experimental design. The application of agents is especially advanced in biomedical data science, where typically a human researcher uses computational tools to interrogate a dataset (Fig. 2a). This human interaction with computational tools can be performed by an agent alone or together with a human (Fig. 2b). However, agentification of research workflows extends beyond data analysis, such as recent models enabling the chat-based exploration of gene expression data[42], and spans a broad spectrum of human research activities, as we will outline below.

Applications of AI research agents

The first step of any research project is ideation. This task was traditionally far outside the realm of AI tools, but LLMs embedded in AI agents make it potentially addressable[43,44]. Frameworks such as ResearchAgent[45] and BioDiscoveryAgent[46] represent LLM-based systems that, once prompted by a user, can autonomously generate new research questions and potential hypotheses by synthesizing knowledge from scientific literature and datasets without requiring step-by-step human guidance. In principle, such systems could be deployed with even greater autonomy, continuously monitoring new publications and proactively identifying gaps in existing research without awaiting human prompts. New platforms such as AgentRxiv, which serve as preprint servers specifically for autonomous agents, could enrich this ecosystem[47]. Once a research idea is formulated, experiments must be effectively designed. LLM-based agents such as the BioDiscoveryAgent mentioned above aim to assist in planning complex biological experiments based on generated hypotheses[46]. Other agents such as Coscientist[3] exemplify how AI can autonomously plan and carry out computational experiments, including drug design processes relevant to oncology[48].

Integrating these ideation and execution functionalities, AI agents could automate entire research workflows. So far, several proof-of-concept studies have shown that (semi-)autonomous research agent systems can be implemented: for instance, the Agent Laboratory aims to automate a generic research pipeline from literature analysis to publication[47,49]. Similarly, the Virtual Lab concept[50] presented a framework in which an AI-driven 'principal investigator' orchestrates a collaborative team of specialized AI agents, each embodying distinct expertise, such as in chemistry, biology or computational science, working together as a multi-agent team. The system was applied in use cases and demonstrated the successful design and validation of new nanobody therapeutics targeting emerging severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants. Further studies have demonstrated the feasibility of research agents within specific scientific domains; for example, SpatialAgent executes end-to-end spatial biology research[51]. Ultimately, these developments foreshadow the emergence of fully autonomous 'AI scientists' capable of independently managing the entire research life cycle — encompassing hypothesis generation, experimental design and execution, data analysis and manuscript authorship. Even initial frameworks are evolving rapidly[52]; notably, the commercially developed system AI Scientist-v2 recently produced an entirely AI-generated manuscript that successfully passed peer review for publication at a scientific workshop[53]. In the future, such autonomous AI systems could fundamentally transform the way we perform research. Now that this technology is feasible, there are new questions that need to be answered: how do we optimize scientific discovery in a world in which agents can perform some of the repetitive tasks? What is the role of human traits such as curiosity, creativity or perseverance? Although AI research agents have made impressive advancements, evidence is needed that they can generate truly new research results without relying on at least a creative spark from a human mind.

What makes a good research agent?

Efficient research agents can be implemented easily, but their ultimate utility depends on domain knowledge and the technical capabilities of the components.

Domain knowledge is the cornerstone of effective research agents in cancer research and oncology. The entirety of domain knowledge includes all biomedical knowledge, such as open-access articles to publicly available textbooks. Knowledge can be included in the training dataset of an LLM, as most non-paywalled scientific articles already are, but it can also be used for fine-tuning or through information retrieval techniques. Proprietary information at research institutions, even standardized operating procedures and similar documents, can be a valuable source of expertise for biomedical research agents. The importance of knowledge for expert AI systems highlights the need for researchers to store and publish all knowledge in the most widely accessible formats possible. Therefore, the scientific community should recognize that our audience now extends beyond human researchers to include AI agents. It can be presumed that most scientific articles are read (and understood) in more detail by AI systems than by humans, which points to a reconsideration of how we write and publish, ensuring that our contributions are not only comprehensible to fellow human scientists but also available and understandable for knowledge extraction by AI systems.

The technical capabilities of agents are determined by the quality of the LLM and its tools. We have witnessed the democratization of LLMs, with most major commercial models now accessible through programming interfaces at relatively modest fees. At the same time, open-source alternatives continue to emerge, typically lagging behind their commercial counterparts in performance benchmarks by only months. These models can increasingly be deployed on standard consumer hardware, diminishing the technical barriers to implementation. However, a fundamental limitation persists: most contemporary LLMs are primarily optimized for conversational interactions rather than tool utilization or scientific reasoning. This gap presents a potential opportunity for the development of specialized models explicitly designed for research agent applications. Besides the LLM, the agent's capabilities are determined by the tools they have access to. Sets of tools can be provided by human operators or can be self-generated by agents[54] (Fig. 1g). Although research agents today use almost exclusively computational tools, the concept of research agents can be extended to include physical sensors and actuators in a laboratory (Fig. 2c).

AI agents in oncology

In parallel with research applications, a natural extension of AI agents lies in the clinical practice of oncology, where decision-making

Box 5 | Benchmarks for large language models and agent systems

Massive Multitask Language Understanding (MMLU)
Tests knowledge across various subjects including mathematics, medicine, law and ethics. Top models achieve over 90% accuracy[19].

MMLU-Pro
A more challenging version of MMLU featuring reasoning-focused questions and expanded answer choices (ten options instead of four)[98].

GPQA (Diamond)
Graduate-level questions in biology, physics and chemistry designed to be unsolvable through web searches. Experts achieve ~65% accuracy[101].

Big-Bench Extra Hard
Contains highly challenging questions designed to push the limits of large language models (LLMs)[102].

Humanity's Last Exam
Contains 2,700 challenging questions across various academic disciplines to assess academic-level artificial intelligence (AI) performance. Current top models achieve around 18% accuracy[39].

MedHELM
Framework leveraging Holistic Evaluation of Large Language Models (HELM) for assessing LLMs across clinical decision support, note generation and patient education. Ensures protected health information (PHI) compliance and evaluates models under realistic clinical constraints[103].

MedCalc-Bench
Evaluates medical calculation performance of AI models, critical for clinical decision-making such as dosage and risk calculations[104].

BixBench
Comprehensive benchmark for evaluating reasoning and tool selection of LLM-based agents specifically in computational biology[40].

MedAgentBench
Realistic virtual electronic health record (EHR) environment designed to benchmark medical LLM agents[84].

AgentClinic
Multimodal agent benchmark evaluating AI performance in simulated clinical environments[85].

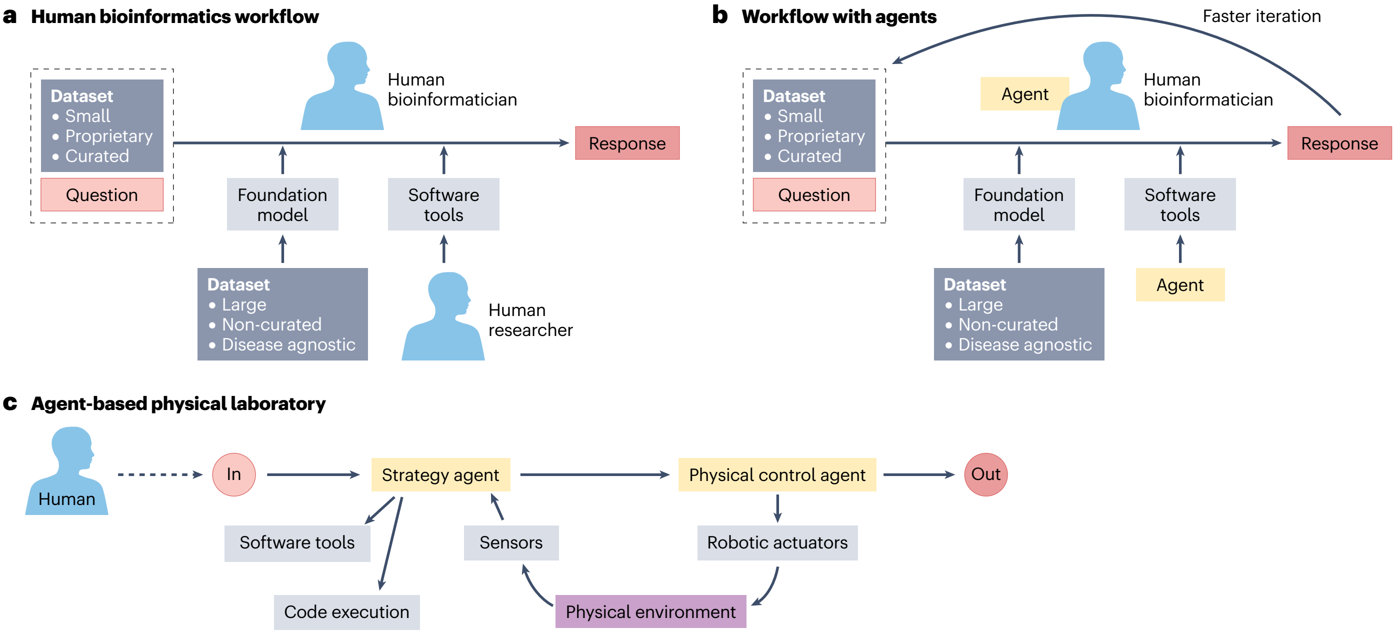

Fig. 2 | Artificial intelligence agents in cancer research. **a**, Traditional bioinformatics workflows require human researchers to manually interact with datasets and tools, creating bottlenecks in the analysis process. **b**, Agent-augmented workflows replace human bioinformaticians with artificial intelligence systems. This enables a faster iteration between question and response while maintaining access to the same foundation models and software tools. **c**, Physical laboratory automation integrates strategic planning agents with robotic control systems, enabling autonomous experimentation through coordinated interaction with software tools, sensors and physical actuators.

frequently depends on synthesizing data from multiple sources. A prominent example is the multidisciplinary tumour board, in which groups of human specialists collaborate to determine the best course of treatment for patients with cancer and provide this recommendation to patients. In principle, such workflows can be well addressed by LLM-based AI agents (Fig. 3). Although, as of 2025, no AI agent systems have been formally integrated into routine oncology practice, several rigorously validated proof-of-concept studies have been published, and academic and commercial interest in this domain is growing rapidly.

Tools for clinical agents

A key concern with the application of any modern AI technology into a clinical setting is the risk of hallucinations, that is, fabricated or erroneous outputs by AI systems[55]. Although hallucinations in well-defined and validated tasks are becoming less frequent with modern AI systems[56], they remain highly problematic in some areas. One of these problematic areas is numerical and arithmetic reasoning, which is key to cancer research and oncology[57]. Even simple clinical tasks, such as comparing tumour sizes pretreatment and posttreatment or calculating dosages, require high accuracy and reliability[58], and LLMs sometimes make mistakes here. Studies have shown that LLM performance on such tasks can be substantially improved by equipping them with external computational tools, such as the ability to write and execute code, or by integrating specialized calculators such as OpenMedCalc[57]. The definition of an AI agent is a reasoning system with access to tools – therefore, an LLM system equipped with a calculator is a rudimentary agent system with clear utility. Another recent publication has introduced RiskAgent, a specialized system designed to perform medical risk predictions across more than 387 risk scenarios spanning multiple diseases such as cardiovascular conditions and cancer[59]. Rather than relying on extensive fine-tuning that requires substantial computational resources, RiskAgent uses its reasoning capabilities to access hundreds of existing clinical decision tools and evidence-based risk calculators when evaluating medical risks. Beyond calculators, a wide array of additional tools can further enhance agent performance. These include access to medical guidelines and evidence repositories[60], radiology image processing models[61] and structured clinical databases. The overall utility of a clinical AI agent is therefore partially determined by the breadth and quality of the tools that it can access – an especially important consideration in oncology.

Clinical reasoning agents

Several research groups have developed fully integrated platforms that combine reasoning capabilities, such as chain-of-thought reasoning[62], with tool use to support complex clinical decisions[63]. One such system, TxAgent[64], is designed to provide individual recommendations for cancer therapy through multistep reasoning and real-time access to biomedical knowledge. It accesses tools from a collection called a universe of tools[64], which enables it to synthesize data across molecular, pharmacokinetic and clinical levels, accounting for drug interactions, contraindications and patient-specific variables such as age, genetic markers and comorbidities. In validation studies, TxAgent demonstrated the ability to generate precise, personalized treatment plans[64] better than standard LLMs. Adjacent medical domains have also explored AI agents in structured clinical decision-making. One notable example is the study 'The AI agent in the room'[65], which modelled liver transplant selection committees using a multi-agent framework. In this set-up, different LLMs assumed specialist roles – hepatology, surgery, cardiology and social work – to simulate the multidisciplinary evaluation process. The agents achieved high diagnostic performance, were able to reliably identify contraindications and predicted survival benefit with high accuracy. Further examples of multi-agent diagnostic frameworks are also emerging in medicine, such as MedAgent-Pro[66], which, although not cancer specific, demonstrates transferable principles for oncology applications.

Conversational agents

Key to the deployment of AI agents in clinical environments will be their ability to engage in context-aware dialogue[67]. Ideally, such systems will interact both with patients and with medical professionals. Therefore, they must not only process complex medical information but also act empathetically and effectively. One such system, recently published by Google, is called the Articulate Medical Intelligence Explorer (AMIE)[68]. It continuously updates its internal representation of the patient case while engaging in a multi-turn conversation **with patients and doctors**. When information is missing, the AMIE actively asks for it and strategically directs the follow-up questions to complete its assessment of the patient. Patients and doctors can not only input their data as a PDF of clinical documents but also provide documentation in a realistic setting such as smartphone photos of lesions or of electrocardiogram printouts. One of the standout features of AMIE is its ability to use long context reasoning capabilities to look into 100 or more PDFs of guidelines for patient management. Engaging in dialogue with patients and medical professionals in realistic circumstances is a core advantage of agent systems that sets them apart from conventional deep learning systems in which the data provided must be narrowly defined and structured: agents can reason about unexpected data inputs and adapt their strategy accordingly, very much like humans do subconsciously all the time when engaging in dialogue.

Diagnosis and treatment planning

Ultimately, AI agents need to support medical professionals in their core tasks of making the final diagnosis and deciding on a treatment. The aforementioned AMIE provides a diagnosis once it decides that sufficient information has been gathered. In a randomized double-blind study of chat-based consultations with 25 patient actors, AMIE consistently performed comparably or better than primary care physicians in making the diagnosis based on the data provided by patients and doctors. Once a diagnosis and potential differential diagnoses have been established, the next step is deciding on a treatment. In oncology, this often means sifting through a large corpus of information. Agents, with their ability to use tools such as iterative search, are ideally suited to support clinicians in this time-consuming task: they can propose evidence-based treatment plans for individual patients[61] while incorporating up-to-date clinical guidelines and literature in their reasoning process[60]. They can even help to find actively recruiting clinical trials: many patients with cancer miss optimal treatment opportunities because of inefficient clinical trial matching processes. AI agents can automatically analyse patient clinical characteristics and systematically evaluate eligibility criteria across trial repositories[69]. By automating this critical but labour-intensive process, AI agents may substantially improve patient access to relevant experimental treatments while reducing the workloads of physicians. In more complex cases such as those discussed in multidisciplinary tumour boards, stand-alone conversational LLMs without agentic capabilities have shown limited performance in generating reliable treatment recommendations[70,71]. By contrast, AI agents equipped with tool use can rapidly analyse multifaceted data, including clinical, imaging, genomic and even cost information, to propose evidence-based treatment plans[72]. AI agents can be used to avoid discrepancies in clinical interpretation by assimilating heterogeneous patient data sources ranging from clinical histories to radiological images, pathology reports and genetic signatures, minimizing possible biases among human observers. They can also go back to previous statements, revise them with new evidence and resolve contradictions.

These agent-based capabilities should not be intended to replace clinician judgement but to augment it. By automating laborious tasks, agents can directly address key clinical challenges: they can manage incomplete knowledge by proactively seeking missing data, reduce medical uncertainty by synthesizing vast amounts of literature and patient data into evidence-ranked options and provide an objective, data-driven basis to help to resolve disagreements among clinicians. Ideally, this would allow clinicians to focus their expertise and efforts on high-quality patient care and complex ethical considerations, with final oversight and responsibility remaining firmly in human hands.

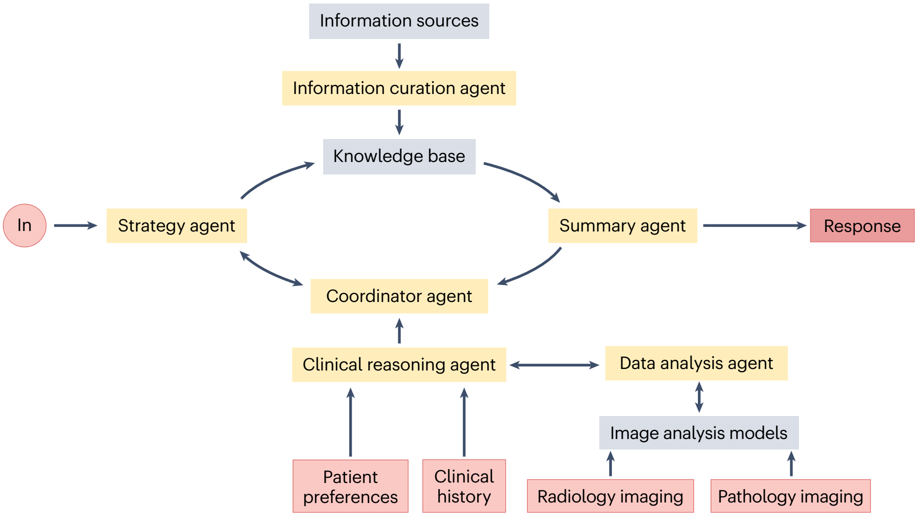

Fig. 3 | A multi-agent framework for oncological treatment decisions, which integrates diverse medical data. This system coordinates specialized agents for information curation, clinical reasoning, data analysis and strategic planning, with each agent processing distinct inputs ranging from patient preferences and clinical history to radiological and pathological imaging data, ultimately generating comprehensive treatment recommendations through collaborative analysis and synthesis.

Agents for clinical image analysis

A cornerstone of cancer diagnosis and treatment is radiological and histopathological image analysis. Deep learning already plays a key role in both fields. For example, it has recently been demonstrated that radiologists can partly be replaced in a narrow and standardized task, specifically diagnosing the presence of cancer in mammography screening[71]. In this task, the diagnostic question is comparatively simple, and it amounts to pattern recognition that can be answered by traditional vision AI models in a single, comprehensive pass. However, more complex questions often need to be answered through imaging. In radiology, such questions include whether it is possible to treat the patient with surgery given their liver function, metastasis load, general health status and vascular situation. This requires reasoning and iterating between potentially different imaging examinations. In pathology, complex questions might involve integrating morphological features across multiple tissue sections to determine tumour heterogeneity or predict treatment response. Agents can address such multistep problems, and proof-of-concept studies in pathology have demonstrated that these agents can emulate the multistep analytical processes of human pathologists[73]. Rather than a single comprehensive pass, agents can use multi-step reasoning to triage pathology slides, focus on specific regions of interest, magnify these areas and synthesize findings[73].

Agents in workflow optimization

Medical professionals dedicate up to half of their time to desk work[74]. Mostly, this work consists of comparatively simple tasks that require navigating through EHR systems and compiling information from different sources. Agents have the potential to alleviate this workload by being directly integrated into existing EHR systems. Even comparatively simple systems have been shown to significantly decrease the time that physicians spend in the EHR systems through the writing of precompiled reports[75]. On a broader scale, AI agents also promise to automate complex information-gathering processes, such as aspects of clinical guideline creation and regulatory documentation in oncology. These agents can interpret guidelines, automate literature monitoring by rapidly identifying relevant studies and synthesize new recommendations[60]. It is becoming increasingly clear that LLMs will support physicians in a number of information compiling and communication overhead tasks, and taken together, the studies presented here provide evidence that agent-based systems will in the future be used to give physicians back valuable time that they can hopefully instead spend with their patients[76–79].

What can agents do in research that cannot be done yet?

Existing, non-agentic AI models are already highly proficient at specific tasks, such as diagnostic imaging. Deep learning has achieved expert-level performance in numerous well-defined, isolated problems. However, the true bottleneck in complex domains such as oncology is not merely task-specific accuracy but the integration of diverse information into a cohesive, actionable strategy. Current models, even specialized multimodal systems, typically require a human to manually curate and present all relevant data – genomic reports, pathology slides, clinical history and the latest literature – within a predefined input structure. This approach does not scale and fails to capture the dynamic, multistep nature of real-world research and clinical reasoning.

This is precisely the gap that AI agents can fill in cancer research. Their advantage lies not in improving a single predictive task but in their ability to autonomously orchestrate entire intellectual workflows, as detailed in Box 2. Unlike a standard LLM or vision language model, which is constrained by a finite context window and relies on the user to provide all necessary information, an agent can actively seek and synthesize knowledge. It can independently query databases for the latest clinical trials, apply a specialized bioinformatics tool to analyse genomic data, extract the salient findings and then use this new information to inform its next action. This iterative process of tool use, information retrieval and reasoning allows an agent to build a comprehensive, holistic understanding of a problem that is simply intractable for a non-agentic model. Therefore, the expected advancement is not just a performance improvement but a shift from automating isolated tasks to automating complex, end-to-end reasoning processes.

Despite this potential, the actual application of agentic AI in cancer research is still nascent. As of 2025, dedicated agent-based systems remain uncommon in routine research, existing primarily as compelling proof of concepts. This landscape, however, could quickly change as the core technologies mature and become more accessible. Adoption will be likely to depend on researchers' willingness to integrate these tools into established workflows. Although some resistance is expected, particularly among those accustomed to traditional methodologies, emerging generations of AI-native scientists may find this transition more natural. Institutional incentives and demonstrated productivity gains could accelerate acceptance over time.

As these systems mature, their integration into clinical practice will be likely to begin not with full autonomy but with hybrid models. In these, an AI agent executes the complex data gathering, analysis and synthesis but presents its findings and proposed action plans to a human expert for final validation and decision-making. This 'human-in-the-loop' framework maintains clinical accountability and leverages expert judgement while capitalizing on the agent's capacity to process information at a scale and speed unattainable by humans.

Future directions

It is important to recognize that any proposed implementation framework for an AI agent is, at this stage, forward-looking. Progression through the phases outlined below is contingent on overcoming fundamental challenges in infrastructure, including the development of robust safety benchmarks, comprehensive validation studies and regulatory pathways. As we will discuss later, much of this foundational work is still in its infancy, and its absence must temper expectations about the immediate real-world deployment of highly autonomous systems. Nevertheless, it is expected that cancer research and oncology will undergo an 'agentification', like every industry, as outlined in a recent White Paper by the World Economic Forum[80]. For cancer research and oncology, we anticipate three phases of agentification.

Phase 1: Interface-based integration

Initially, agent-based systems will be likely to operate through accessible interfaces similar to current user interfaces such as the ChatGPT interface that can be used on consumer devices. These work independently from clinical information systems and research databases. These agents process data explicitly provided by clinicians or researchers – uploaded patient records, research datasets or literature – and deliver responses through dialogue. This approach sidesteps the immediate interoperability challenges while allowing clinicians and researchers to experience the analytical capabilities of these systems under controlled conditions. In fact, this process is already well underway. A substantial proportion of the healthcare workforce is already using LLMs such as ChatGPT[81]. These tools have some agentic capabilities such as processing multimodal input, performing web search, parsing documents or taking basic action such as sending e-mail reminders to users. Beyond such use on consumer devices, some early adopters in academic medical centres have already begun implementing such systems for consultative support in tumour boards and other clinical operations[82]. The advantage of this approach lies in its minimal disruption to existing workflows while providing substantial analytical capabilities without requiring extensive system integration. The main risk lies in the unregulated nature of many of these tools that should be certified medical devices but are usually not[83]. Another concern is the potential leakage of private patient data through such tools. Therefore, ideally, AI agents should be integrated into existing information technology (IT) infrastructure in healthcare, which we anticipate in phase 2.

Phase 2: Deep system integration with supervised authority

As capabilities mature and trust develops, AI agents can transition more easily to deep integration with hospital infrastructure and research data ecosystems. This phase represents a substantial advancement wherein agents gain direct, permissioned access to patient records in clinical settings and to research databases and literature in research environments. Rather than relying on manually uploaded data, these systems could dynamically identify and access relevant information. Nevertheless, they will primarily function as assistants – operating within a supervisory framework in which recommendations require human approval before implementation.

Phase 3: Active system engagement with autonomous agency

The final evolutionary phase will involve agents capable of initiating actions within their operational environments. In research settings, this could manifest as direct control of laboratory robotics to design and conduct experiments based on iteratively updated hypotheses. In clinical contexts, these systems might independently order diagnostic tests or adjust medication dosages. This represents a fundamental shift from assistive to collaborative intelligence, in which AI agents become active participants in clinical and research workflows rather than merely consultative tools. As such, phase 3 systems are already becoming technologically feasible; the regulatory and ethical frameworks around this technology need to be developed.

Limitations and open questions

The field of agentic AI is still nascent, and there are a number of limitations and open challenges that need to be actively addressed by the research community. First, we need to make sure that agentic performance in oncology can be accurately assessed. Second, we need to align agentic actions with human ethical guardrails, and third, we need to prepare our infrastructure such that agents can be used in clinical routine. We elaborate on each of these below.

The need for real-world benchmarks in oncology

Although conventional deep learning systems tackle comparatively simple tasks whose performance can be measured by comparing them to ground-truth 'yes-or-no' answers, agentic systems tackle the complex reality of oncological workflows and open-ended research. Developing benchmarks that accurately measure how well an agent performs a task is challenging. Several research groups are therefore actively developing comprehensive clinical benchmarks, such as MedAgentBench[84] and AgentClinic[85], which simulate EHR environments and clinical decision-making scenarios. However, ensuring these benchmarks fully capture the multifaceted performance of agentic systems and address all evaluative complexities remains an ongoing challenge with many unresolved aspects[86]. These include the difficulty of evaluating multistep reasoning rather than just final outcomes, scoring open-ended tasks with multiple valid solutions and assessing safety-critical behaviours such as appropriate escalation to human oversight. Addressing these gaps will likely require new evaluation frameworks, including expert assessment of reasoning quality and simulation environments that replicate real-world complexity.

Implementation challenges and critiques of over-automation

Although the potential of AI agents is high, their practical implementation is challenging. The history of AI in medicine provides several cautionary tales about substantial gaps between promising research prototypes and effective, sustainable clinical tools. For instance, systematic reviews of hundreds of AI models developed for coronavirus disease 2019 (COVID-19) diagnosis found that almost none were suitable for clinical implementation because of methodological flaws and poor generalizability[87]. A primary reason for previous failures is the 'lab-to-live' gap, in which models that perform exceptionally on curated research datasets falter when exposed to the messy, heterogeneous data of real-world practice[88]. This issue is compounded by poor workflow integration. An AI tool that does not seamlessly fit into existing processes will be resisted, regardless of its accuracy. Consequently, many ambitious AI projects have become 'digital shelfware', technically functional but practically unusable, because they failed to address these fundamental usability and interoperability challenges. A prominent example is IBM Watson for Oncology, which despite substantial investment failed to achieve widespread clinical adoption because of concerns about recommendation quality and poor integration with clinical workflows[89].

Furthermore, a purely technology-centric view overlooks critical human factors. One of the most well-documented critiques is the risk of automation bias – a tendency to overtrust AI outputs – and cognitive offloading[90]. When clinicians rely heavily on AI recommendations, the erosion of their own essential skills and judgement can occur over time. An agent designed to improve efficiency might inadvertently make clinicians more passive in their decision-making, a considerable risk when the AI encounters an edge case or makes a subtle error. This raises profound questions of accountability when an agent's recommendation contributes to patient harm. These unresolved questions of responsibility and safety are not merely practical hurdles; they form the core of the ethical and regulatory frameworks that must be developed[83].

Ethical and regulatory constraints

Beyond robust benchmarking, deploying AI agents in oncology requires careful navigation of ethical and regulatory constraints[91]. These systems naturally have more freedom to act than traditional AI models and thus have more potential to inflict harm on patients. Thus, it is essential to ensure that agents are robust against manipulations[92,93] and that they perform fairly and without bias[83]. This holds true both in clinical oncology and in research, where monetary incentives to influence drug development and regulatory pathways are high. We therefore should provide a means of human oversight that ensures that agents perform reliably and transparently along ethical guardrails in order to make clinicians and researchers trust them[83,94,95]. At the same time, the regulatory landscape for AI agents in healthcare, especially for autonomous agents, remains somewhat unclear, although the first precedents are beginning to emerge. For example, Prof. Valmed (Prof. Valmed GMBH, Germany), an LLM-powered clinical decision support system that aims

Glossary

Chain-of-thought reasoning
A prompting technique that encourages language models to generate intermediate reasoning steps before arriving at a final answer, improving performance on complex tasks.

Contraindications
Clinical conditions or factors that make a particular treatment or procedure inadvisable because of potential harm to the patient.

Deep learning
A subset of machine learning that uses artificial neural networks with multiple layers to learn hierarchical representations of data.

Differential diagnoses
A systematic process of distinguishing between diseases or conditions that share similar clinical features to identify the most likely diagnosis.

Edge case
An unusual or extreme scenario that occurs at the boundaries of normal operating conditions, often revealing limitations in system performance.

Hyperparameters
Configuration settings defined before model training that control the learning process, such as learning rate, batch size and network architecture choices.

Large language model
(LLM). Type of artificial intelligence model trained on vast amounts of text data to understand and generate human language, capable of performing diverse language tasks without task-specific training.

Multi-turn conversation
A dialogue consisting of multiple exchanges between a user and an AI system, in which context from previous turns informs subsequent responses.

Natural language processing
(NLP). A field of AI artificial intelligence focused on enabling computers to understand, interpret and generate human language.

Parsing documents
The computational process of analysing and extracting structured information from unstructured or semi-structured text documents.

Precompiled reports
Standardized documents generated in advance or from templates, typically containing structured clinical or research data ready for review.

Reinforcement learning
A machine learning paradigm in which an agent learns to make decisions by receiving feedback in the form of rewards or penalties based on its actions.

Token
The basic unit of text processed by a language model, which may represent a word, subword or character depending on the tokenization scheme.

Transformer architecture
A neural network design that uses self-attention mechanisms to process sequential data in parallel, forming the foundation of modern LLMs.

Vision language model
An AI model capable of processing and relating both visual information (such as images) and textual data within a unified framework.

to provide healthcare professionals with evidence-based diagnostic and therapeutic information by searching a curated database of validated medical sources, recently became the first general-purpose generative AI tool to receive Class IIb CE marking under the Medical Device Regulation of the European Union. However, current definitions of medical devices – whether in the United States, requiring authorization by the Food and Drug Administration, or in the European Union, requiring conformity with the Medical Device Regulation – do not yet capture the characteristics of increasingly autonomous AI systems. Regulatory adaptations are on the horizon[96], but their development will require the active input of domain experts, as well as the involvement of patients and society more broadly.

Practical constraints

Agent-based systems are most useful if they are seamlessly incorporated into clinical workflows and have access to any patient data. In practice, a difficulty lies in electronic patient data being distributed among many different hospital systems. Hospitals need to make their data accessible – for example, via standardized interfaces such as Fast Healthcare Interoperability Resources (FHIR) – so that agents can work on those data[67,97]. Another critical consideration is the economic feasibility of implementing AI agents in oncology: multi-agent systems can incur substantial computational costs so care needs to be taken to direct the use of computational resource spending to where it brings the most value. Lastly, the integration of AI agents into oncology and research raises concerns about the shifting role of human critical thinking. When knowledge workers rely on AI, they can experience a reduction in cognitive effort, shifting from active problem-solving to oversight and verification[98]. Although this can increase efficiency, it also raises the risk of over-reliance on AI-generated outputs. AI agents should be designed to reinforce rather than replace human judgement, ensuring that healthcare professionals remain engaged in decision-making processes.

Engineering the ‘personality’ of agents

A somewhat unexpected challenge with LLM-based AI agents, and a key focus of recent industry efforts such as the GPT-5 release, lies in emergent behaviours such as sycophancy – a tendency to prioritize user agreement over independent reasoning – and laziness[99]. The latter manifests when agents tasked with complex scientific analyses default to shallow processing pathways despite having more sophisticated analytical capabilities. This ‘reluctance’ can lead to superficial data examinations or, even more problematic, the synthesis and analysis of fictitious datasets rather than the use of the provided empirical data. This phenomenon appears to be an emergent property of LLMs rather than an explicitly programmed limitation. Interestingly, the motivational architecture of these systems responds differentially to various prompting strategies, with some agents demonstrating enhanced performance when presented with simulated scientific peer evaluation or when operating within explicitly defined methodological frameworks. This raises fundamental questions about how intrinsic motivation might be architected into next-generation scientific agents, potentially requiring new training paradigms that specifically reward thoroughness and methodological rigour rather than mere task completion or other superficial optimization metrics.

Conclusion

We expect cancer research and oncology to be ‘agentified’ over the next decade, and therefore, researchers and healthcare professionals as well as our institutions need to become ready for this transition. AI agents can solve some of the previous limitations of AI, including the limited focus of AI systems to one single task, as well as their inability to perform action. Although challenges in validation, regulation and integration remain, the trajectory towards increasingly autonomous AI collaborators appears feasible and promises an increased speed of scientific and clinical operations, ultimately promising to accelerate the path to scientific discovery and delivery of care. The question facing the oncology community is not whether AI agents will transform our field, but how we will shape their implementation to maximize benefit while ensuring safety and maintaining the human element that remains central to both science and care.

Published online: 12 January 2026

References

1. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
2. Gao, S. et al. Empowering biomedical discovery with AI agents. *Cell* **187**, 6125–6151 (2024).
3. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
Demonstration of an LLM-based agent (Coscientist) autonomously planning and executing real-world scientific experiments, marking a milestone for AI agents in research.
4. Bran, A. M. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
5. Kaiser, J., Lauscher, A. & Eichler, A. Large language models for human-machine collaborative particle accelerator tuning through natural language. *Sci. Adv.* **11**, eadr4173 (2025).
6. Russell, S. & Norvig, P. *Artificial Intelligence* (Pearson, 1999).
7. ANTHROP\C. Building effective AI agents. https://www.anthropic.com/engineering/building-effective-agents (2024).
8. Zou, J. & Topol, E. J. The rise of agentic AI teammates in medicine. *Lancet* **405**, 457 (2025).
9. Google Cloud. What is an AI agent? https://cloud.google.com/discover/what-are-ai-agents (2025).
10. Ray, S. AI agents — what they are, and how they'll change the way we work. *Source* https://news.microsoft.com/source/features/ai/ai-agents-what-they-are-and-how-theyll-change-the-way-we-work/ (2024).
11. AWS. What are AI agents? https://aws.amazon.com/what-is/ai-agents/ (2025).
12. Lee, Y., Ferber, D., Rood, J. E., Regev, A. & Kather, J. N. How AI agents will change cancer research and oncology. *Nat. Cancer* **5**, 1765–1767 (2024).
13. Vaswani, A. et al. Attention is all you need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2017).
This study introduced the transformer architecture that underpins all modern LLMs and AI agents.
14. Radford, A. et al. Language models are unsupervised multitask learners. OpenAI https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (2019).
15. Zhao, W. X. et al. A survey of large language models. Preprint at https://doi.org/10.48550/arXiv.2303.18223 (2023).
16. Brown, T. B. et al. Language models are few-shot learners. *NeurIPS* **33**, 1877–1901 (2020).
This study demonstrated that LLMs can perform diverse tasks with minimal examples, establishing the paradigm of in-context learning that enables agentic capabilities.
17. Truhn, D., Reis-Filho, J. S. & Kather, J. N. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nat. Med.* **29**, 2983–2984 (2023).
18. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at https://doi.org/10.48550/arXiv.2303.12712 (2023).
19. Hendrycks, D. et al. Measuring massive multitask language understanding. Preprint at https://doi.org/10.48550/arXiv.2009.03300 (2020).
20. ANTHROP\C. Claude's extended thinking. https://www.anthropic.com/research/visible-extended-thinking (2025).
21. Tu, T. et al. Towards conversational diagnostic artificial intelligence. *Nature* **642**, 442–450 (2025).
22. McDuff, D. et al. Towards accurate differential diagnosis with large language models. *Nature* **642**, 451–457 (2025).
23. Ferber, D. et al. In-context learning enables multimodal large language models to classify cancer pathology images. *Nat. Commun.* **15**, 10104 (2024).
24. OpenAI o1 System Card. *OpenAI* https://openai.com/index/openai-o1-system-card/ (2024).
25. Brodeur, P. G. et al. Superhuman performance of a large language model on the reasoning tasks of a physician. Preprint at https://doi.org/10.48550/arXiv.2412.10849 (2024).
26. Hao, S. et al. Training large language models to reason in a continuous latent space. Preprint at https://doi.org/10.48550/arXiv.2412.06769 (2024).
27. OpenAI. Introducing OpenAI o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/ (2025).
28. Yuksekgonul, M. et al. Optimizing generative AI by backpropagating language model feedback. *Nature* **639**, 609–616 (2025).
29. Bostrom, N. in *Machine Ethics and Robot Ethics* 69–75 (Routledge, 2020).
30. Smit, A., Duckworth, P., Grinsztajn, N., Barrett, T. D. & Pretorius, A. Should we be going MAD? A look at multi-agent debate strategies for LLMs. Preprint at https://doi.org/10.48550/arXiv.2311.17371 (2023).
31. Wu, Y. et al. ProAI: Proactive multi-agent conversational AI with structured knowledge base for psychiatric diagnosis. Preprint at https://doi.org/10.48550/arXiv.2502.20689v2 (2025).
32. Yao, S. et al. ReAct: Synergizing reasoning and acting in language models. Preprint at https://doi.org/10.48550/arXiv.2210.03629 (2022).
This study introduced the ReAct framework combining reasoning traces with actions, providing a foundational architecture for modern AI agents.
33. Wang, E. et al. TxGemma: efficient and agentic LLMs for therapeutics. Preprint at https://doi.org/10.48550/arXiv.2504.06196 (2025).
34. Shanahan, M., McDonell, K. & Reynolds, L. Role play with large language models. *Nature* **623**, 493–498 (2023).
35. Moritz, M., Topol, E. & Rajpurkar, P. Coordinated AI agents for advancing healthcare. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-025-01363-2 (2025).
36. Schaeffer, R. Pretraining on the test set is all you need. Preprint at https://doi.org/10.48550/arXiv.2309.08632 (2023).
37. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
This study introduced Med-PaLM, demonstrating that LLMs can achieve expert-level performance on medical question answering and establishing benchmarks for clinical AI.
38. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
39. Phan, L. et al. Humanity's last exam. Preprint at https://doi.org/10.48550/arXiv.2501.14249 (2025).
40. Mitchener, L. et al. BixBench: a comprehensive benchmark for LLM-based agents in computational biology. Preprint at https://doi.org/10.48550/arXiv.2503.00096 (2025).
41. Huang, K. et al. Biomni: a general-purpose biomedical AI agent. *Bioinformatics* https://doi.org/10.1101/2025.05.30.656746 (2025).
42. Schaefer, M. et al. Multimodal learning enables chat-based exploration of single-cell data. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-025-02857-9 (2025).
43. Doshi, A. R. & Hauser, O. P. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* **10**, eadn5290 (2024).
44. Si, C., Yang, D. & Hashimoto, T. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. Preprint at https://doi.org/10.48550/arXiv.2409.04109 (2024).
45. Baek, J., Jauhar, S. K., Cucerzan, S. & Hwang, S. J. ResearchAgent: iterative research idea generation over scientific literature with large language models. Preprint at https://doi.org/10.48550/arXiv.2404.07738 (2024).
46. Roohani, Y. et al. BioDiscoveryAgent: an AI agent for designing genetic perturbation experiments. Preprint at https://doi.org/10.48550/arXiv.2405.17631 (2024).
47. Schmidgall, S. & Moor, M. AgentRxiv: towards collaborative autonomous research. Preprint at https://doi.org/10.48550/arXiv.2503.18102 (2025).
48. Zhang, K. et al. Artificial intelligence in drug development. *Nat. Med.* **31**, 45–59 (2025).
49. Schmidgall, S. et al. Agent laboratory: using LLM agents as research assistants. Preprint at https://doi.org/10.48550/arXiv.2501.04227 (2025).
50. Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The virtual lab: AI agents design new SARS-CoV-2 nanobodies with experimental validation. *Bioinformatics* https://doi.org/10.1101/2024.11.11.623004 (2024).
51. Wang, H. et al. SpatialAgent: an autonomous AI agent for spatial biology. *Bioinformatics* https://doi.org/10.1101/2025.04.03.646459 (2025).
52. Lu, C. et al. The AI scientist: towards fully automated open-ended scientific discovery. Preprint at https://doi.org/10.48550/arXiv.2408.06292 (2024).
53. Yamada, Y. et al. The AI scientist-v2: workshop-level automated scientific discovery via agentic tree search. Preprint at https://doi.org/10.48550/arXiv.2504.08066 (2025).
54. Wölflein, G., Ferber, D., Truhn, D., Arandjelović, O. & Kather, J. N. LLM agents making agent tools. Preprint at https://doi.org/10.48550/arXiv.2502.11705 (2025).
55. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
56. Williams, C. Y. K. et al. Physician- and large language model-generated hospital discharge summaries. *JAMA Intern. Med.* https://doi.org/10.1001/jamainternmed.2025.0821 (2025).
57. Goodell, A. J., Chu, S. N., Rouholiman, D. & Chu, L. F. Large language model agents can use tools to perform clinical calculations. *NPJ Digit. Med.* **8**, 163 (2025).
58. Litière, S., Collette, S., de Vries, E. G. E., Seymour, L. & Bogaerts, J. RECIST — learning from the past to build the future. *Nat. Rev. Clin. Oncol.* **14**, 187–192 (2017).
59. Liu, F. et al. RiskAgent: autonomous medical AI copilot for generalist risk prediction. Preprint at https://doi.org/10.48550/arXiv.2503.03802 (2025).
60. Ferber, D. et al. GPT-4 for information retrieval and comparison of medical oncology guidelines. *NEJM AI* https://doi.org/10.1056/AIcs2300235 (2024).

61. Ferber, D. et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nat. Cancer* **6**, 1337–1349 (2025). **Peer-reviewed validation of an autonomous AI agent for oncology clinical decision support in a tumour board setting.**
62. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. *NeurIPS* **35**, 22199–22213 (2022).
63. Liévin, V., Hother, C. E., Motzfeldt, A. G. & Winther, O. Can large language models reason about medical questions? Preprint at https://doi.org/10.48550/arXiv.2207.08143 (2022).
64. Gao, S. et al. TxAgent: an AI agent for therapeutic reasoning across a universe of tools. Preprint at https://doi.org/10.48550/arXiv.2503.10970 (2025).
65. Hasjim, B. J. et al. The AI agent in the room: informing objective decision making at the transplant selection committee. *Transplantation* https://doi.org/10.1101/2024.12.06.24318575 (2024).
66. Wang, S. et al. Empowering medical multi-agents with clinical consultation flow for dynamic diagnosis. Preprint at https://doi.org/10.48550/arXiv.2503.16547 (2025).
67. Kather, J. N., Ferber, D., Wiest, I. C., Gilbert, S. & Truhn, D. Large language models could make natural language again the universal interface of healthcare. *Nat. Med.* **30**, 2708–2710 (2024).
68. Palepu, A. et al. Towards conversational AI for disease management. Preprint at https://doi.org/10.48550/arXiv.2503.06074 (2025).
69. Ferber, D. et al. End-to-end clinical trial matching with large language models. Preprint at https://doi.org/10.48550/arXiv.2407.13463 (2024).
70. Lukac, S. et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Arch. Gynecol. Obstet.* **308**, 1831–1844 (2023).
71. Schmidl, B. et al. Assessing the role of advanced artificial intelligence as a tool in multidisciplinary tumor board decision-making for recurrent/metastatic head and neck cancer cases – the first study on ChatGPT 4o and a comparison to ChatGPT 4.0. *Front. Oncol.* **14**, 1455413 (2024).
72. Nardone, V. et al. The role of artificial intelligence on tumor boards: perspectives from surgeons, medical oncologists and radiation oncologists. *Curr. Oncol.* **31**, 4984–5007 (2024).
73. Ghezloo, F. et al. PathFinder: a multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. Preprint at https://doi.org/10.48550/arXiv.2502.08916 (2025).
74. Sinsky, C. et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann. Intern. Med.* **165**, 753–760 (2016).
75. Rotenstein, L. et al. Virtual scribes and physician time spent on electronic health records. *JAMA Netw. Open* **7**, e2413140 (2024).
76. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
77. Chen, S. et al. The effect of using a large language model to respond to patient messages. *Lancet Digit. Health* **6**, e379–e381 (2024).
78. Bock, A. Using a virtual scribe may shorten EHR time. *JAMA* **332**, 188 (2024).
79. Maddox, T. M. et al. Generative AI in medicine — evaluating progress and challenges. *N. Engl. J. Med.* https://doi.org/10.1056/NEJMsb2503956 (2025).
80. Bastubbe, Y., Jain, D. & Torti, F. *Frontier Technologies in Industrial Operations: The Rise of Artificial Intelligence Agents*. White Paper (World Economic Forum, 2025).
81. Blease, C. R., Locher, C., Gaab, J., Hägglund, M. & Mandl, K. D. Generative artificial intelligence in primary care: an online survey of UK general practitioners. *BMJ Health Care Inform.* **31**, e101102 (2024).
82. Umeton, R. et al. GPT-4 in a cancer center — institute-wide deployment challenges and lessons learned. *NEJM AI* https://doi.org/10.1056/AIcs2300191 (2024).
83. Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E. & Wicks, P. Large language model AI chatbots require approval as medical devices. *Nat. Med.* **29**, 2396–2398 (2023).
84. Jiang, Y. et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. *NEJM AI* https://doi.org/10.1056/AIdbp2500144 (2025).
85. Schmidgall, S. et al. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. Preprint at https://doi.org/10.48550/arXiv.2405.07960 (2024).
86. Rodman, A., Zwaan, L., Olson, A. & Manrai, A. K. When it comes to benchmarks, humans are the only way. *NEJM AI* https://doi.org/10.1056/AIe2500143 (2025).
87. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
88. Zhang, A., Xing, L., Zou, J. & Wu, J. C. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat. Biomed. Eng.* **6**, 1330–1345 (2022).
89. Schmidt, C. M. D. Anderson breaks with IBM Watson, raising questions about artificial intelligence in oncology. *J. Natl Cancer Inst.* https://doi.org/10.1093/jnci/djx113 (2017).
90. Dratsch, T. et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology* **307**, e222176 (2023).
91. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, 141 (2023).
92. Han, T. et al. Medical large language models are susceptible to targeted misinformation attacks. *NPJ Digit. Med.* **7**, 288 (2024).
93. Clusmann, J. et al. Prompt injection attacks on vision language models in oncology. *Nat. Commun.* **16**, 1239 (2025).
94. Savage, T., Nayak, A., Gallo, R., Rangan, E. & Chen, J. H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit. Med.* **7**, 20 (2024).
95. Abgrall, G., Holder, A. L., Chelly Dagdia, Z., Zeitouni, K. & Monnet, X. Should AI models be explainable to clinicians? *Crit. Care* **28**, 301 (2024).
96. Gilbert, S., Dai, T. & Mathias, R. Consternation as Congress proposal for autonomous prescribing AI coincides with the haphazard cuts at the FDA. *NPJ Digit. Med.* **8**, 165 (2025).
97. Balch, J. A. et al. Machine learning-enabled clinical information systems using fast healthcare interoperability resources data standards: scoping review. *JMIR Med. Inform.* **11**, e48297 (2023).
98. Lee, H. et al. The impact of generative AI on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *CHI Conference on Human Factors in Computing Systems (CHI '25)* 1–22 (Association for Computing Machinery, 2025).
99. Fanous, A. et al. SycEval: evaluating LLM sycophancy. Preprint at https://doi.org/10.48550/arXiv.2502.08177 (2025).
100. Wiest, I. C. et al. Large language models for clinical decision support in gastroenterology and hepatology. *Nat. Rev. Gastroenterol. Hepatol.* https://doi.org/10.1038/s41575-025-01108-1 (2025).
101. Rein, D. et al. GPQA: A graduate-level Google-proof Q&A benchmark. Preprint at https://doi.org/10.48550/arXiv.2311.12022 (2023).
102. Kazemi, M. et al. BIG-bench extra hard. In *Proc. 63rd Annu. Meet. Assoc. Comput. Linguist.* Vol. 1, 26473–26501 (ACL, 2025).
103. Liang, P. et al. Holistic evaluation of language models. Preprint at https://doi.org/10.48550/arXiv.2211.09110 (2022).
104. Khandekar, N. et al. MedCalc-bench: evaluating large language models for medical calculations. Preprint at https://doi.org/10.48550/arXiv.2406.12036 (2024).

Acknowledgements

J.N.K. is supported by the German Cancer Aid (DECADE, 70115166), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; CAMINO, 01EO2101; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET TRANSCAN; Come2Data, 16DKZ2044A; DEEP-HCC, 031L0315A), the German Academic Exchange Service (SECAI, 57616814), the European Union's Horizon Europe and innovation programme (ODELIA, 101057091; GENIAL, 101096312), the European Research Council (ERC; NADIR, 101114631), the National Institutes of Health (EPICO, R01 CA263318) and the National Institute for Health and Care Research (NIHR, NIHR203331) Leeds Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. This work was funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions

D.T., L.C.A., F.M. and J.N.K. researched data for the article. All authors contributed substantially to discussion of the content. D.T., S.A., J.Z., L.C.A. and J.N.K. wrote the article. All authors reviewed and/or edited the manuscript before submission.

Competing interests

J.N.K. declares consulting services for Bioptimus, France; Panakeia, UK; AstraZeneca, UK; and MultiplexDx, Slovakia. Furthermore, he holds shares in StratifAI, Germany; Synagen, Germany; and Ignition Lab, Germany; has received an institutional research grant by GSK and AstraZeneca; and has received honoraria by AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer and Fresenius. D.T. received honoraria for lectures by Bayer, GE, Roche, AstraZeneca and Philips and holds shares in StratifAI GmbH, Germany, and in Synagen GmbH, Germany. F.M. is a scientific adviser for and holds shares in Modella AI and is an adviser for Danaher. S.A. is an employee of Alphabet and may own stock as part of the standard compensation package. J.Z. and L.C.A. declare no competing interests.

Additional information

Peer review information *Nature Reviews Cancer* thanks Anant Madabhushi, Wayne Zhao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.